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

Jamie Davies
27th 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

The main design goals behind hashtags are to categorise tweets and allow them to show up more easily in searches[2]. Whilst the task that this project is aiming to complete is novel and fairly unexplored, it is well connected with other experiments, systems and projects within the research community.

2.1 Recommendation Systems

Traditional recommendation systems are in place all over the web today. From music discovery services (such as Last.fm[3]) to suggested purchases on retail sites (like that in place at Amazon[4], 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

[2] `https://support.twitter.com/articles/49309-using-hashtags-on-twitter`
[3] `www.last.fm`
[4] `www.amazon.co.uk`

items for a user by comparing their preferences with the descriptions of the items (Pazzani and Billsus, 2007).

2.1.3 Relevance Feedback

Relevance feedback is a process that was originally designed for information retrieval, and works on the assumption that a user can not always correctly encapsulate int a query what it is they are searching for. It works by allowing a user to create an initial query to which an initial set of results is returned. Out of these initial results, the user can then mark certain results as relevant or irrelevant, and this information is then submitted and used to refine the original query and return more relevant results to the user (Salton and Buckley, 1997).

Instead of limiting recommendation systems to the accuracy of their classifiers, a common approach is incorporate relevance feedback techniques. Utiyama and Yamamoto (2006) showed that it is possible to combine collaborative filtering, content-based filtering and relevance feedback techniques into one system to provide better recommendations.

2.2 Hashtag Recommendation Research

Even though providing hashtag recommendations and suggestions is still a new and largely unexplored field, there have been several efforts to improve the hashtag experience for Twitter users.

2.2.1 Current Twitter Hashtag Implementation

The current hashtag system on Twitter (figure 1) uses a non-personalised auto-complete tool to provide suggestions to the user. Whenever a hash symbol (#) is typed in the tweet composer, the system simply suggests hashtags starting with the letters that the user has typed so far. These suggestions are chosen from a tiny subset of hashtags, taken from a mixture of the currently trending[5] and from the user's history. Whilst better than not having suggestions at all, this system is only truly useful in a specific use case: when the user knows the starting letters of a trending hashtag they want to use, or are trying to recall a hashtag from they have previously used. This system does not help the user choose the correct hashtag for their tweet.

[5]Trending hashtags are those with the highest rise in usage within a given time period.

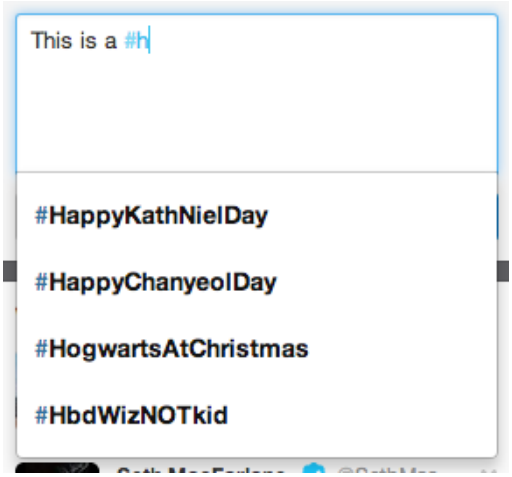

Figure 1: Twitter's current hashtag suggestion system.

2.2.2 Using Tweet Content to Suggest Hashtag Probabilities

Mazzia and Juett (2011) used an application of a naive Bayes model to suggest hashtags by using just the content of the tweet as input. The Bayes model allows the system to calculate the probabilities of the tweet using different hashtags. This gave them suggestion correctness rates of up to 72%, using various cross-validation data sets. They remark that a much larger training set would be required for real-world use, however.

2.2.3 Comparing Tweets To Other Tweets

By assuming that the primary purpose of hashtags is to categorise tweets and improve searching (as Twitter envisioned), Zangerle et al. (2011) created a system that recommends hashtags for a tweet by taking tweets from other tweets that are textually similar to the query. The similarity between tweets is calculated with the TF-IDF (term frequency – inverse document frequency) model. The hashtags are then extracted from the similar tweets, ranked according to how similar the tweets were to the original query, and returned as a list of suggestions to the user. A number of different ranking algorithms were tested, but this was found to be the most successful.

2.2.4 Creating Personalised Recommendations

After studying the advantages of providing personalised recommendations in retail situations on a per-user basis, Kywe et al. (2012) realised that a similar approach towards hashtags could prove fruitful. Hashtag use varies

from user to user, with some users using the latest trending hashtags, other uses only using a specific set or type of tag, and with some users barely using them at all. They proposed a personalised hashtag recommendation system that considers both user preferences and the query tweet content: the system creates a ranked list of hashtags from both the most similar users and most similar tweets. This gave promising results, although it was noted that this may not be the best recommendation system for all types of tweets and hashtags.

2.2.5 Overcoming Hashtag Duality

Observers of social media have realised that hashtags play a dual role within the online microblogging communities, such as Twitter. On one hand, hashtags fulfil the design goals that Twitter created them to accomplish (bookmarking and improving search); on the other hand, however, they serve as a badge of community membership, connecting users together. Yang et al. (2012) took this duality into account when attempting to create a hashtag recommendation system by training a SVM (support vector machine) classifier with a variety of features taken from the tweet metadata to overcome the duality and suggest relevant hashtags.

2.3 Broader Classification in Twitter

Twitter is a thriving metropolis of users expressing themselves on a daily (and often more frequent) basis, and has grown exponentially in size since its inception in 2006. Due to this, the data that it contains has caught the attention of researchers throughout computer science and even other disciplines. Whilst the concept of recommending hashtags is relatively unexplored, there have been many other classification experiments run with Twitter data.

2.3.1 Categorising Tweets

Sriram et al. (2010) proposed an approach to classify tweets into 5 general categories: *news*, *opinions*, *deals*, *events* and *private messages*. This was achieved by using a small set of specific features from each tweet, instead of using the traditional "Bag-Of-Words" (BOW) text classification method. The BOW approach is centered around counting occurences of words in the text, but in the case of Twitter and its 140 character limit, it is very rare that words are actually repeated in a tweet.

2.3.2 Categorising Users

Another approach to deciphering the vast quantity of data on Twitter is to classify the users themselves. Twitter has become a powerful platform for people posting content about events, and as such it would be useful to automatically establish *who* is participating in these events. By taking a number of features from each user account and passing them through a K-Nearest Neighbours (KNN) algorithm, De Choudhury et al. (2012) developed a system that would classify a user's behaviour into one of three categories: organisations, journalists/media bloggers and ordinary individuals.

3 Report on Technical Progress

12pt text for main body.

4 Plan of Remaining Work

12pt text for main body.

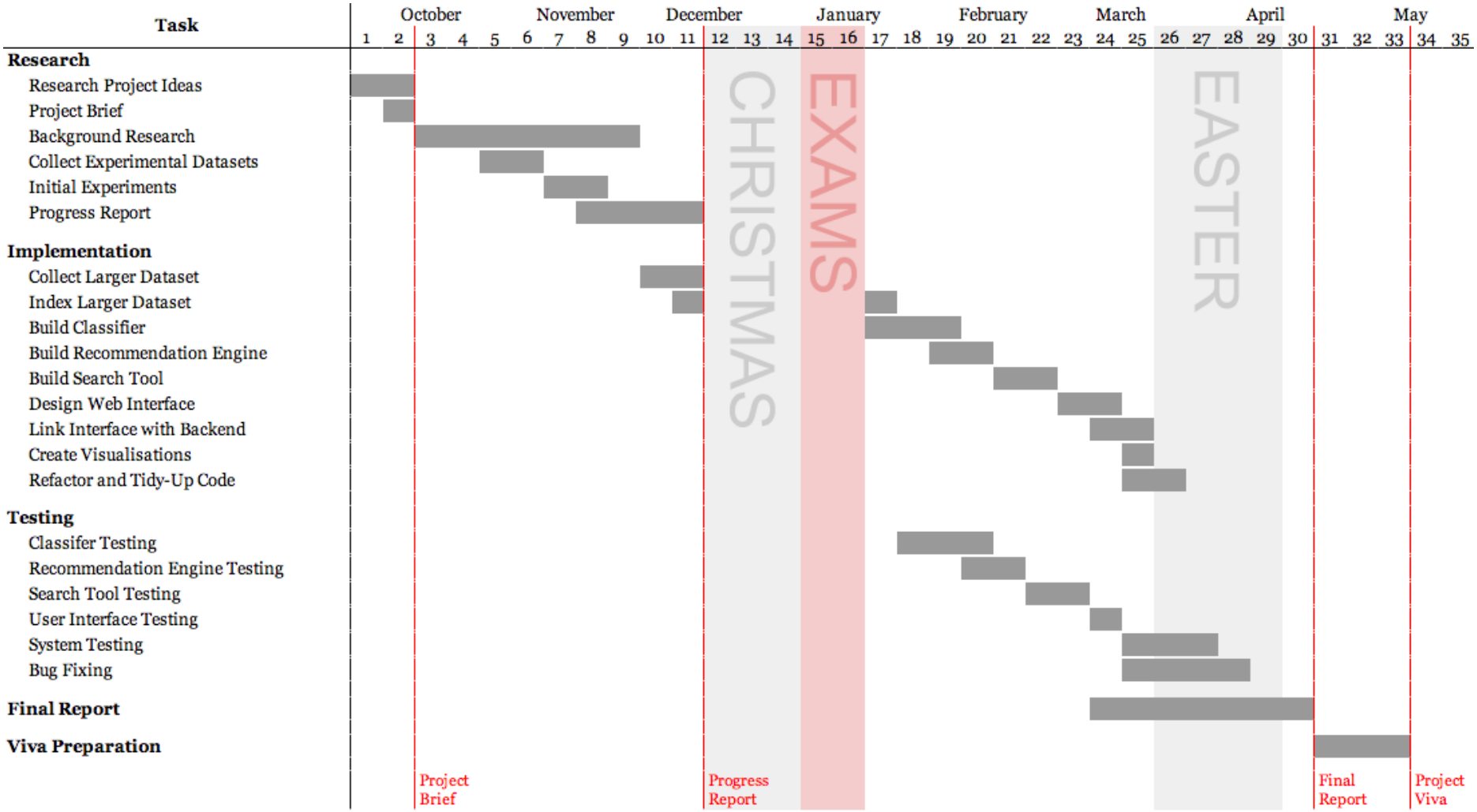

Figure 2: A Gantt chart showing the planned progression through the project.

References

De Choudhury, M. et al. (2012). 'Unfolding the Event Landscape on Twitter: Classification and Exploration of User Categories'. *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*. CSCW '12. Seattle, Washington, USA: ACM, pp. 241–244.

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

Kywe, S. M. et al. (2012). 'On Recommending Hashtags in Twitter Networks'. *Proceedings of the 4th International Conference on Social Informatics*. SocInfo'12. Lausanne, Switzerland: Springer-Verlag, pp. 337–350.

Mazzia, A. and Juett, J. (2011). 'Suggesting Hashtags on Twitter'. *EECS 545, Machine Learning*.

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.

Salton, G. and Buckley, C. (1997). 'Improving Retrieval Performance by Relevance Feedback'. *Readings in Information Retrieval*. Ed. by K. Sparck Jones and P. Willett. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 355–364.

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

Sriram, B. et al. (2010). 'Short Text Classification in Twitter to Improve Information Filtering'. *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '10. Geneva, Switzerland: ACM, pp. 841–842.

Utiyama, M. and Yamamoto, M. (2006). 'Relevance Feedback Models for Recommendation'. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. EMNLP '06. Sydney, Australia: Association for Computational Linguistics, pp. 449–456.

Yang, L. et al. (2012). 'We Know What @You #Tag: Does the Dual Role Affect Hashtag Adoption?' *Proceedings of the 21st International Confer-*

ence on World Wide Web. WWW '12. Lyon, France: ACM, pp. 261–270.

Zangerle, E. et al. (2011). 'Recommending #-Tags in Twitter'. *Proceedings of the CEUR Workshop*. CEUR WS '11.