

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

Jamie Davies
22nd April 2014

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 improve the quality of the information we can gain from hashtags. However, whilst such a system could improve 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	Introduction	5
1.1	Preliminary Data Analysis	6
2	Background Literature & Related Work	9
2.1	Traditional Recommendation Systems	9
2.2	Assigning Labels with Machine-Learning Techniques	10
2.3	Other Approaches to Hashtag Recommendation	11
3	Requirements	14
3.1	Functional Requirements	14
3.2	Non-Functional Requirements	15
4	Design	16
4.1	Architecture	16
4.1.1	Processing Framework	16
4.1.2	Classifier	16
4.2	Interface	16
4.2.1	API Endpoints	16
4.2.2	Web Interface	16
5	Implementation	17
5.1	Stream Processing	17
5.2	Classification System	17
5.2.1	Training	17
5.2.2	Classifying	17
5.2.3	Optimisation	17
5.2.4	RESTful Interface	17
5.3	Web Application	17
5.3.1	Framework	17
5.3.2	Authentication	17
5.3.3	Connecting to Classifier	17
6	Testing	18
6.1	Strategy	18
6.2	Unit Testing	18
6.3	Classification Testing	18
6.4	User Interface Testing	18

6.5	Requirements Validation	18
7	Conclusion	19
7.1	System Evaluation	19
7.2	Project Management	19
7.3	Future Work	19
A	Time Management	20
B	Testing Plans & Results	22
C	REST API Interface	23
D	Original Project Brief	24
E	References	26

Acknowledgements

I would like to thank my supervisor, Dr. Nick Gibbins, for providing inspiration and guidance throughout the course of the project.

I would also like to thank Dr. Jonathon Hare and Dr. Sina Samangooei for their continued support and advice.

Statement of Originality

The work contained within this report has not been previously submitted for a degree or diploma at any other higher education institution. To the best of my knowledge and belief, this project contains no material previously published or written by another person except where due references are made.

Chapter 1

Introduction

Collaborative tagging (sometimes referred to as social tagging) systems are designed to allow users to navigate, organise and manage online resources. The central idea behind these systems is the concept of an annotation: the action of a user applying a personalised tag to a resource. On a large scale, many of these annotations can be taken together to form a complex network of users, resources and tags — commonly referred to as a *folksonomy* (Xu et al., 2008).

One of the most popular applications of a folksonomy is Twitter¹. ‘Hashtags,’ which are simply any combination of letters or digits preceded by a hash character (#), afford users the ability to create free-form tags to describe the content of their posts (tweets). Tweets can be categorized with several hashtags, thereby creating networks of tweets and users, and making it easy to find other related tweets, users and hashtags. This renders hashtags as a powerful tool in aiding the search and navigation of the billions of tweets contained within Twitter.

Despite its numerous benefits, the hashtag system presents new challenges to overcome before it can become truly useful. Due to the open nature of folksonomies, it is important that users have the freedom to create and use exactly the tags they wish to use. However, this unsupervised tagging can result in vast numbers of hashtags for users to choose from — often including redundant or ambiguous hashtags. When posting a tweet, there is nothing stopping a user from creating an entirely new hashtag to describe something with exactly the same meaning as a collection of other hashtags. This tag redundancy can confuse users and fragment the true meaning behind the synonymous hashtags.

There is also the converse issue of tag ambiguity to consider, in which a single hashtag could be associated with many meanings and connotations. The use of these hashtags can incorrectly imply that tweets are similar and related to one another, when in fact they are from distinct unrelated categories.

Because these problems are complicating the already vast hashtag information space, an intelligent recommendation system is needed to aid users in their choice of hashtags. For example, when a user writes a tweet such as: “*The golden gate bridge*

¹www.twitter.com

looks so nice today! I LOVE SF in this weather!”, the system should recommend hashtags such as `#SanFrancisco`, `#California` or even `#sohappy`, but should not recommend hashtags such as `#ScienceFiction`, `#amreading` or `#fantasy`.

Such a recommendation system would also need to aid users in the navigational and exploratory aspects of folksonomies. For example, when a user performs a search on Twitter for the hashtag `#BGT`, the system should display not only those tweets that contain `#BGT`, but also tweets that contain other relevant hashtags such as `#BritainsGotTalent`, `#VoteAttraction` or `#VoteJackCarroll`.

1.1 Preliminary Data Analysis

To investigate the usage of hashtags and the problems associated with them further, a dataset of 500,000 tweets was collected from the sample Twitter stream over a time period of approximately 4.5 days. The tweets collected were filtered to ensure that they were in English, and contained at least one hashtag.

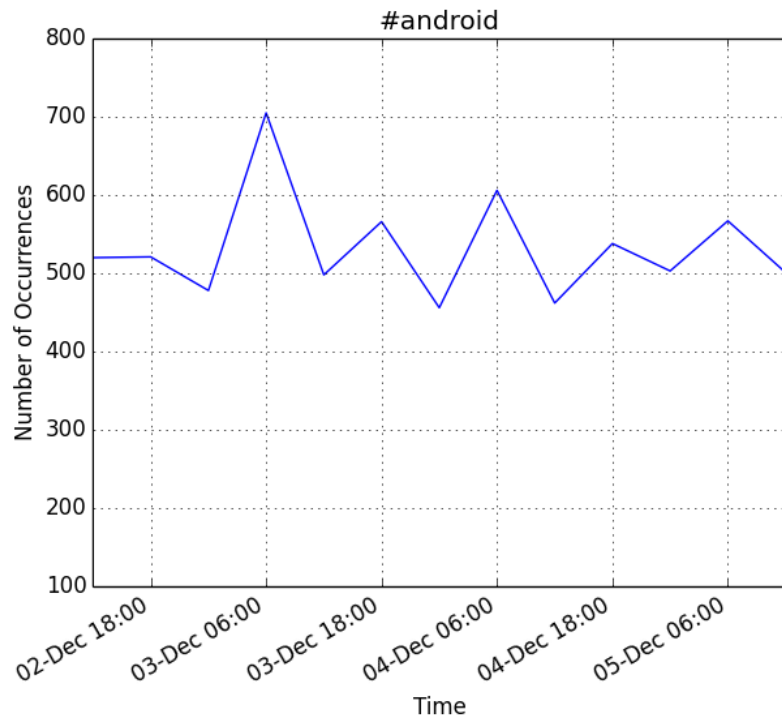
Through the use of some trivial Python scripts, it is easy to find evidence of poor hashtag choices. In particular, the most overwhelming observation is the quantity of redundant hashtags in use throughout Twitter. Table 1.1 shows several hashtags taken from the dataset and their number of occurrences. Despite the tags all having exactly the same semantic meaning, the use of the tags is fragmented and spread across several tags.

Hashtag	Number of Occurrences
<code>#peopleschoice</code>	94849
<code>#peopieschoice</code>	2043
<code>#peoplechoice</code>	439
<code>#peoplesch</code>	287
<code>#peopleschoi</code>	269
<code>#peoplesc</code>	230
<code>#peoplescho</code>	219
<code>#peolpeschoice</code>	164
<code>#peopleschioce</code>	137
<code>#pca</code>	94

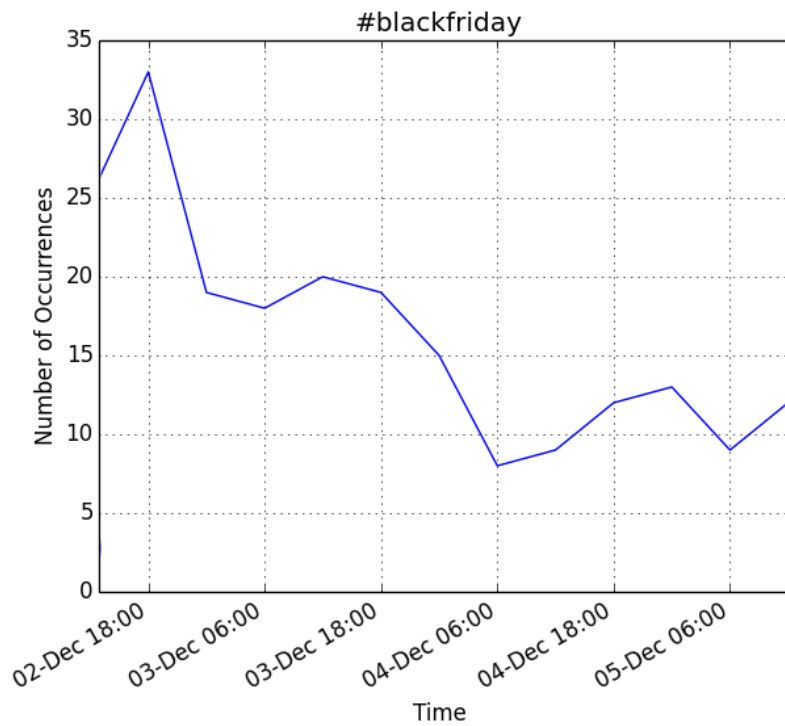
Table 1.1: Occurrences of hashtags referring to the People’s Choice Awards

Another interesting observation is that of the times that different hashtags are used, which varies greatly from hashtag to hashtag. Some tags are used in tweets at a fairly uniform rate (Fig. 1.1(a)), whilst others feature large spikes of usage over a short period of time (Fig. 1.1(b)). This demonstrates the communities behind the hashtags

— rather than a hashtag being used at a particular time for a particular event, some hashtags are used consistently by users to join in on a large-scale conversation.



(a) #android



(b) #blackfriday

Figure 1.1: Usage of hashtags over the period that the dataset was collected.

Chapter 2

Background Literature & Related Work

The main design goals behind hashtags are to categorise tweets and allow them to show up more easily in searches¹. 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 Traditional Recommendation Systems

Traditional recommendation systems are in place all over the web today. From music discovery services (such as Last.fm²) to suggested purchases on retail sites (like on the front page of Amazon³), these systems are all personalised recommendation engines that take an individual user's preferences and use them to provide suggestions tailored to that user.

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).

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

¹<https://support.twitter.com/articles/49309-using-hashtags-on-twitter>

²www.last.fm

³www.amazon.co.uk

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 a content-based description of the user’s preferences with the descriptions of the items (Pazzani and Billsus, 2007).

A key issue with content-based filtering is that the recommendations can only be as accurate as the algorithm used to derive a user’s profile. There are a number of algorithms available to build user profiles, depending upon the context, but essentially a content-based profile is created using a weighted vector of item features. The weights mark the importance of each feature to the user, and can be computed from individually rated content vectors.

Cantador et al. (2010) studied and evaluated a number of content-based recommendation models based upon the premise of user and item profiles being described in terms of weighted lists and tags. Through their experiments they found that models that focused on user profiles outperformed the models oriented towards item profiles in nearly every test. They go on to suggest that a better way of profiling users would be through the use of tag clustering.

Instead of limiting recommendation systems to the accuracy of their classifiers, a common approach is incorporate relevance feedback techniques. 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 into 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, 1990).

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 Assigning Labels with Machine-Learning Techniques

Although there exists a long history of research into traditional recommendation algorithms, the data structure of a folksonomy is distinct from those found in traditional problems. This has led to a new generation of supervised learning classification algorithms which can be applied to the problem of assigning labels to resources.

In order to facilitate the comparison of resources, tags and users, aggregate projections of the data can be constructed. This reduces the dimensionality of the data by sacrificing some information (Schmitz et al., 2006). There are two main classes of algorithms that perform this task: feature *selection*, which reduces the size of a dataset by selecting a subset of the data, and feature *extraction*, which uses the existing data to generate entirely new features. Despite the improved results that feature extraction can provide, their computational cost and memory requirements can often make them impractical for use on the extremely large datasets common

with Folksonomies (Gemmell et al., 2009).

One compromise to this trade-off is to use an algorithm such as Hebbian Deflation, which approximates the outputs of the intensive feature extraction techniques, but with lower computational cost and memory requirements (Oja and Karhunen, 1985). Gemmell et al. (2009) used this approach coupled with a K-Nearest Neighbour classifier to find appropriate tags for a given resource. This was done by finding ‘neighbouring’ users: “Neighbours are selected only if they have tagged the query resource and tags are selected for the recommendation set only if they have been applied by the neighbour to the resource.” It was tested on data from the popular bookmarking site Delicious⁴ and from Citeulike⁵, a tool that enables researchers to manage and discover scholarly references, and it was found to be a successful approach with giving results.

Niwa et al. (2006) took a different approach to overcome the limitations of collaborative filtering. Using a modified version of the TF-IDF (Salton and Buckley, 1988), they calculate the affinity level between users and tags (where affinity level is a measure of how frequently a user uses a particular tag), and then use this to calculate the similarity between tags. These similarity measures are then clustered using a custom clustering algorithm to provide a set of recommended tags for a given user and resource. This enabled them to adjust the level at which a user’s preferences were taken into account when determining recommendations, as well as providing a simple solution to the tag redundancy issue found across most Folksonomies.

Another popular machine-learning approach to the folksonomy tag recommendation problem is the use of a Naive Bayes classifier. Naive Bayes has been an accepted tool in information retrieval for a long time, with some early literature dating back over 50 years (Maron and Kuhns, 1960), and has recently been seeing a renaissance within the machine-learning community.

As shown by De Pessemier et al. (2010), Naive Bayes can be easily adapted and extended to work in the context of providing recommendations to match a query or a set of user preferences. In particular, their algorithm relied upon the general user preferences for cases where the classifier produced an uncertain outcome, and became more context specific on a sliding scale as the classifier gave more certain probabilities.

2.3 Other Approaches to Hashtag Recommendation

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.

The current hashtag system on Twitter (Figure 2.1) 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

⁴<http://delicious.com/>

⁵<http://www.citeulike.com/>

letters that the user has typed so far. These suggestions are chosen from a tiny subset of hashtags, taken from a mixture of those currently trending⁶ and from the user’s history. Whilst better than not having suggestions at all, this system is only truly useful in one of two specific use cases: when the user knows the starting letters of a trending hashtag they want to use, or are trying to recall a hashtag that they have previously used. This system does not help the user choose the correct hashtag for the content of their tweet.

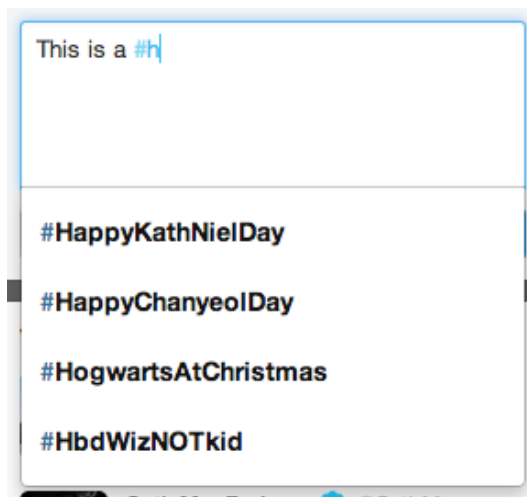


Figure 2.1: Twitter’s current hashtag suggestion system.

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 hashtags from other tweets that are textually similar to the query. The similarity between tweets is calculated with the TF-IDF 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.

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 users 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.

Shepitsen et al. (2008) used a hierarchical agglomerative clustering algorithm to profile users and provide personalised recommendations in collaborative tagging systems. They found that clusters of tags can be effectively used to ascertain a user’s

⁶Trending hashtags are those with the highest rise in usage within a given time period.

interests, which could then be used in a traditional content-based recommendation approach. This technique worked well, particularly for dense folksonomies such as Last.fm.

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.

Chapter 3

Requirements

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

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

3.1 Functional Requirements

1. The system must allow the user to log in and publish tweets to their Twitter account.
2. The system must provide hashtag recommendations as the user is creating a tweet.
3. 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.
4. The system must use information from a large dataset of tweets to generate a model representing each hashtag.
5. The system must be able to compare tweets against its representational hashtag models.
6. *Optional:* The system must be able to update its classification models using information from the live Twitter stream.
7. *Optional:* The system must provide probabilities for how likely a hashtag is to be related to a tweet.

3.2 Non-Functional Requirements

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

Chapter 4

Design

4.1 Architecture

4.1.1 Processing Framework

4.1.2 Classifier

4.2 Interface

4.2.1 API Endpoints

4.2.2 Web Interface

Chapter 5

Implementation

5.1 Stream Processing

5.2 Classification System

5.2.1 Training

5.2.2 Classifying

5.2.3 Optimisation

5.2.4 RESTful Interface

5.3 Web Application

5.3.1 Framework

5.3.2 Authentication

5.3.3 Connecting to Classifier

Chapter 6

Testing

6.1 Strategy

6.2 Unit Testing

6.3 Classification Testing

6.4 User Interface Testing

6.5 Requirements Validation

Chapter 7

Conclusion

7.1 System Evaluation

7.2 Project Management

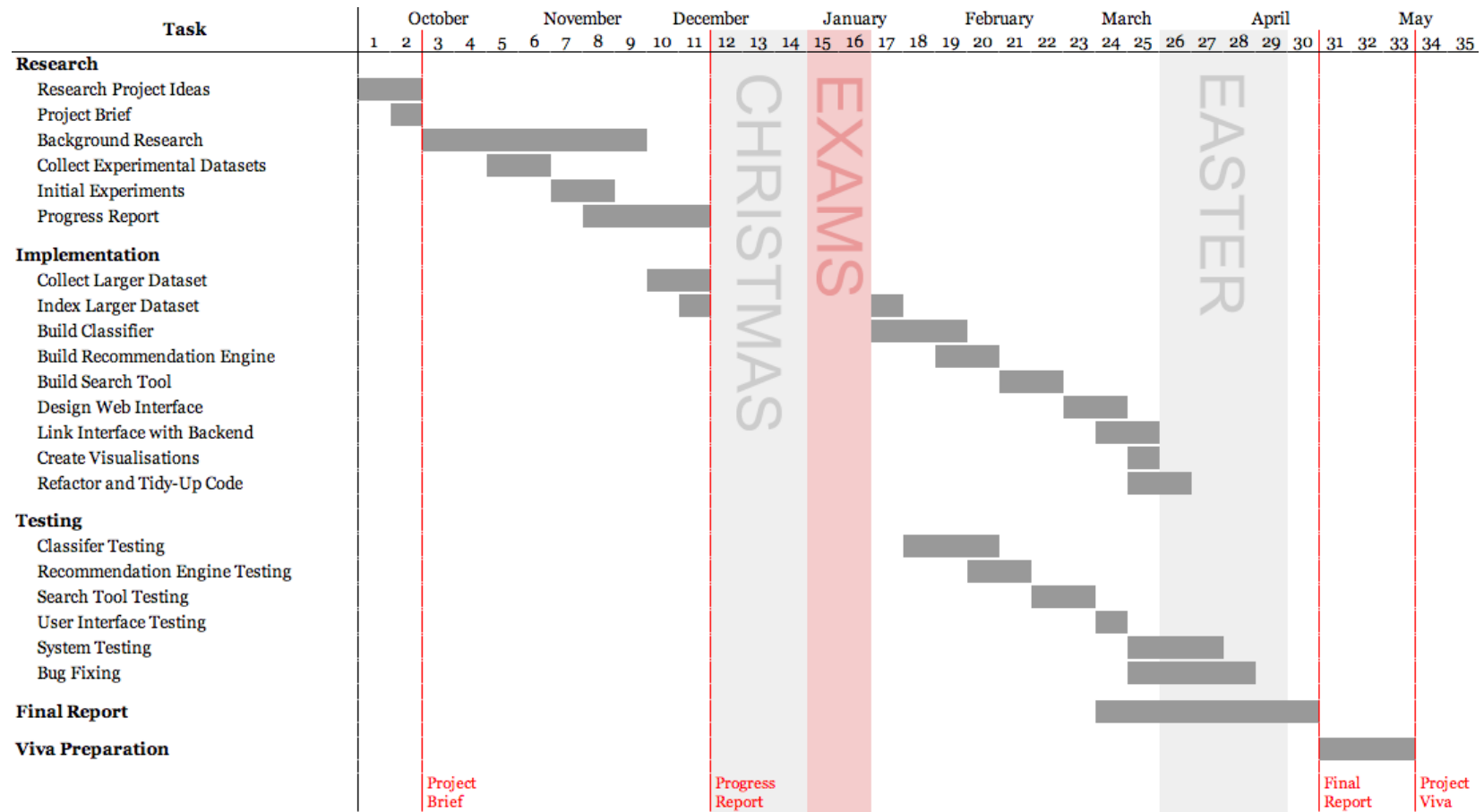
7.3 Future Work

Appendix A

Time Management

20

This is a Gantt chart showing the scheduled progression through the different aspects of the project.



Appendix B

Testing Plans & Results

Appendix C

REST API Interface

Appendix D

Original Project Brief

Part III Individual Project Brief

Enhanced Content Analysis and Information Retrieval Using Twitter Hashtags

Author:

Jamie Davies

jagd1g11@ecs.soton.ac.uk

Supervisor:

Dr. Nicholas Gibbins

nmg@ecs.soton.ac.uk

Microblogging services such as Twitter¹ have become very popular in recent years. One of the key characteristics of such services is the use of ‘hashtags’ — topical/categorical annotations provided by the authors of the posts (tweets) themselves. These hashtags have been proven to be useful and information-rich in previous microblogging research projects; from using them to crowd-source real-time event detection [2], to using them to train sentiment classifiers [1]. However due to their nature, there is a vast and diverse collection of hashtags for users to choose from, resulting in the information provided by hashtags not being as accurate or as complete as it could be.

For my project, I will create a system that aims to support and enrich the information provided by hashtags. I will use the latest machine-learning techniques to examine and classify the topics and concepts behind the hashtags and in doing so, be able to suggest suitable new hashtags for tweets that are relevant to their content. This will allow users to make a better choice of hashtags when writing their tweets, and therefore refine the information they provide. Furthermore, I will create an extension of this system that will provide a context-aware tweet search facility. This will enable a user 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.

Finally, I will test the success and importance of my research by implementing a recent research experiment that depends upon hashtags (such as those conducted by Davidov et al. [1] or Sakaki et al. [2]). I will compare the results of the experiment both with and without my enhanced system, and provide validation for the project.

References

- [1] D. Davidov, O. Tsur, and A. Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters, COLING '10*, pages 241–249. Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

Appendix E

References

- Cantador, I., Bellogín, A. and Vallet, D. (2010). ‘Content-based Recommendation in Social Tagging Systems’. *Proceedings of the Fourth ACM Conference on Recommender Systems*. RecSys ’10. Barcelona, Spain: ACM, pp. 237–240.
- De Pessemier, T., Deryckere, T. and Martens, L. (2010). ‘Extending the Bayesian classifier to a context-aware recommender system for mobile devices’. *Proceedings of the Fifth International Conference on Internet and Web Applications and Services*. ICIW 2010. Barcelona, Spain: IEEE Computer Society, pp. 242–247.
- Gemmell, J., Schimoler, T., Ramezani, M. and Bamshad, M. (2009). ‘Adapting K-Nearest Neighbor for Tag Recommendation in Folksonomies’. *Proceedings of the 7th Workshop on Intelligent Techniques for Web Personalization & Recommender Systems*. ITWP ’09. Pasadena, CA, USA: CEUR Workshop Proceedings, pp. 51–62.
- Goldberg, D., Nichols, D., Oki, B. M. and Terry, D. (1992). ‘Using Collaborative Filtering to Weave an Information Tapestry’. *Communications of the ACM* 35.12, pp. 61–70.
- Kywe, S. M., Hoang, T.-A., Lim, E.-P. and Zhu, F. (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.
- Maron, M. E. and Kuhns, J. L. (1960). ‘On Relevance, Probabilistic Indexing and Information Retrieval’. *Journal of the ACM (JACM)* 7.3, pp. 216–244.
- Niwa, S., Doi, T. and Honiden, S. (2006). ‘Web Page Recommender System Based on Folksonomy Mining for ITNG ’06 Submissions’. *Proceedings of the Third International Conference on Information Technology: New Generations*. ITNG ’06. Las Vegas, Nevada, USA: IEEE Computer Society, pp. 388–393.

- Oja, E. and Karhunen, J. (1985). ‘On Stochastic Approximation of the Eigenvectors and Eigenvalues of the Expectation of a Random Matrix’. *Journal of Mathematical Analysis and Applications* 106.1, pp. 69–84.
- Pazzani, M. J. and Billsus, D. (2007). ‘The Adaptive Web’. Vol. 4321. Berlin, Heidelberg: Springer-Verlag. Chap. Content-based Recommendation Systems, pp. 325–341.
- Salton, G. and Buckley, C. (1990). ‘Improving Retrieval Performance by Relevance Feedback’. *Journal of the American Society for Information Science* 41.4, pp. 288–297.
- Salton, G. and Buckley, C. (1988). ‘Term-weighting Approaches in Automatic Text Retrieval’. *Information Processing and Management* 24.5, pp. 513–523.
- Schafer, J. B., Konstan, J. A. and Riedl, J. (2001). ‘E-Commerce Recommendation Applications’. *Data Mining and Knowledge Discovery* 5.1-2, pp. 115–153.
- Schmitz, C., Hotho, A., Jäschke, R. and Stumme, G. (2006). ‘Mining Association Rules in Folksonomies’. *Proceedings of the 2006 International Federation of Classification Societies Conference*. IFCS ’06. Ljubljana, Slovenia: Springer, pp. 261–270.
- Shepitsen, A., Gemmell, J., Mobasher, B. and Burke, R. (2008). ‘Personalized Recommendation in Social Tagging Systems Using Hierarchical Clustering’. *Proceedings of the 2008 ACM Conference on Recommender Systems*. RecSys ’08. Lausanne, Switzerland: ACM, pp. 259–266.
- 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.
- Xu, S., Bao, S., Fei, B., Su, Z. and Yu, Y. (2008). ‘Exploring Folksonomy for Personalized Search’. *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR ’08. Singapore, Singapore: ACM, pp. 155–162.
- Yang, L., Sun, T., Zhang, M. and Mei, Q. (2012). ‘We Know What @You #Tag: Does the Dual Role Affect Hashtag Adoption?’ *Proceedings of the 21st International Conference on World Wide Web*. WWW ’12. Lyon, France: ACM, pp. 261–270.
- Zangerle, E., Gassler, W. and Specht, G. (2011). ‘Recommending #-Tags in Twitter’. *Proceedings of the Workshop on Semantic Adaptive Social Web*. SASWeb ’11. Girona, Spain: CEUR Workshop Proceedings, pp. 67–78.