Further experiments in Micro-blog Categorization

Sandra Garcia Esparza, Michael P. O'Mahony, and Barry Smyth

CLARITY: Centre for Sensor Web Technologies,
School of Computer Science and Informatics,
University College Dublin, Belfield, Dublin 4, Ireland
{sandra.garcia-esparza,michael.p.omahony,barry.smyth}@ucd.ie
http://www.clarity-centre.org/

Abstract. Since the creation of Twitter in 2008, micro-blogging services have received a lot of attention among users who wish to share news items, opinions and information with friends and colleagues. However, these services typically provide for only limited organisation of content, with the main ranking criterion being post time with perhaps some basic message filtering accommodated. Given the substantial and increasing volume of posts that micro-blogging services attract, there is a clear need to assist users when it comes to effectively consuming this content. In this regard, categorisation offers one approach to organise content by grouping related messages together. In this paper we present a study in the recommendation of categories for short-form messages in order to provide for better search and message filtering. In particular, we present an index-based approach where real-time web data can be used as a source of knowledge for category recommendation. Further, we evaluate our approach on two different micro-blogging datasets and results show that micro-blog messages in sufficient quantities provide a useful recommendation signal for category recommendation.

Keywords: Micro-blogs, categorisation, information retrieval, recommendation

1 Introduction

Micro-blogs are short textual messages written by users who wish to to share information, comments or opinions on virtually any topic of their choosing. Currently, there is a wide variety of these services that users can choose from, such as Tumblr, Plurk, Jaiku and of course Twitter, which since its creation in 2006 has ushered in a new era of micro-blogging. Further, users have also been using micro-blogging services to search for information. For instance, Twitter has its own search engine[1] where users can search for social-generated content such as users opinions on a particular product or the progress of a football game. These search engines have introduced a new concept of search which is different from

[1] http://search.twitter.com/

traditional web-search [19]. However, one of the limitations of current micro-blog search engines is the lack of sophisticated ways to filter and rank information. At present, messages are ranked primarily by posting time, with perhaps some basic filtering techniques also provided. Given the volume of message posts and the virtually unlimited variety of topics that messages relate to, there is a clear need for additional mechanisms that allow for better organisation and retrieval of real-time content.

Some services like Blippr[2], which allows users to review products using short-form messages (called *blips*), provide structure by restricting blips to 5 product categories (*movies*, *music*, *books*, *applications* and *games*). While Twitter doesn't facilitate such a categorisation, it has introduced the use of *hashtags* in order to allow for more structured data. A hashtag is a tag preceded by the 'hash' symbol (#) and can be considered as a course form of message categorisation. Examples of hashtags are: *#travel*, *#movies*, *#Inception*, *#fifa*, etc. However, users are not obliged to tag messages and nor are they restricted to using a predefined set of tags when they choose to do so. Thus, categorisation of these messages is still needed. Previous work has investigated categorisation of blogs [3, 18]; however, less attention has been focused on micro-blogs.

The aim of this paper is to develop a category recommender for such short-form messages in order to provide for better organisation and retrieval of this type of information. In particular, we describe how micro-blog content can be used as a source of indexing and retrieval information for category recommendation. We apply our approach to two micro-blogging services, Blippr and Twitter, and evaluate the performance of our approach in respect of five product categories (those currently supported by the Blippr service). We also provide a comparison with different machine learning approaches. In addition, we consider the potential for cross-domain recommendation, by training our model using messages from one domain and recommending categories for messages from another domain.

The paper is organised as follows. In Section 2, we describe related work that has been performed in the area of tag recommendation and text categorisation. A description of the Blippr and Twitter services are presented in Section 3 along with our approach to recommend categories for micro-blog messages. An evaluation of the approach is presented in Section 4 and finally concluding remarks are presented in Section 5.

2 Related Work

In recent times, micro-blog services have been the subject of much academic and industrial research interest. For example, the utility of micro-blogs from the perspectives of sentiment analysis [8, 10], interests and activities discovery [2] and hot topics discovery [1] have been investigated. In this paper, our focus is on the automatic categorisation of micro-blog messages and here we provide an overview of some of the related work that has been conducted in this regard.

[2] `http://www.blippr.com`

One of the most popular uses of text categorisation is the categorisation of web documents. Typically, these documents are websites from portal directories (e.g. Yahoo) [12], blogs [3] or search results [4], which can be grouped into categories to allow for better search and organisation of the content. Sometimes categories are extracted from a hierarchy of tags. For instance, in [3] a hierarchy of tags is created by grouping blog entries using similarity metrics. Their results show that tags so-derived are useful to group blog entries into broad categories, but less so when it comes to selecting tags to indicate more specific blog themes. Similarly, [16] use an agglomerative clustering approach to build a personalised recommender system that can cluster tags in order to extract the resource topic and user's interests. Their approach is evaluated on Delicious and Last.fm datasets and shows a significative improvement over a k-means clustering approach. Text categorisation is also commonly used in text filtering (e.g. email and SMS spam filtering [5, 7, 14]) and in word sense disambiguation [21].

One of the typical approaches to text categorisation lies in the application of machine learning techniques [15]. In past work, Naïve Bayes or Support Vector Machines have shown to be very effective for this task [20, 9]. To date, much of the work related to text categorisation has been applied to long-form text. In this work, we focus on the categorisation of micro-blogs, which poses additional challenges given the short length of these messages and the practically unlimited range of topics that can be discussed. Further, misspellings, irony and use of informal language (e.g. slang, incorrect use of punctuations or elongation of words such as "boooooringggg") are also common in micro-blogs.

Some previous solutions to this problem use external knowledge to make the data less sparse and to discover relationships in the data. For instance, in [11] large amounts of data are collected from external sources which is applied to classify short and sparse text and Web segments. In [17] the authors take a different approach to categorise Twitter messages in order to improve message filtering for users. In this work, messages are classified into five categories (*news*, *events*, *opinions*, *deals* and *private messages*) depending on the communication intention of the tweet author. A total of 8 features are extracted from messages to distinguish between categories, such as whether a message contains a date and location (considered likely to be an *event*), whether it contains currency symbols (indicative of a *deal*), etc. The work presented in this paper is similar in the sense that we also wish to categorise micro-blogs; however, in our approach we focus on the textual content of messages to perform categorisation, rather than on the particular intention of message authors.

3 Micro-blog Categorisation

In this work we consider messages from two services. The first is the well known micro-blogging service Twitter, which allows users to share their knowledge and opinions on any topic in a 140-character message. Figure 1 shows a tweet which expresses a positive opinion on the movie *'Inception'*.

Fig. 1. A tweet representing a review of the movie *'Inception'*.

The second domain we consider is Blippr. Blippr is a product review service which allows registered users to express their views on products from five different categories. These reviews (or *blips*) are in the form of 160-character text messages, and users must also supply an accompanying rating on a 4-point rating scale: *love it*, *like it*, *dislike it* or *hate it*. Figure 2 shows a typical Blippr review for the movie *'The Matrix'*.

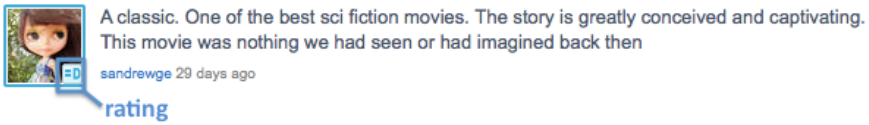

Fig. 2. A Blippr review of the movie *'The Matrix'*.

For both services, we consider the categorisation of messages relating to five different product types: *applications*, *music*, *movies*, *books* and *games* (these are the product types supported by the Blippr service). Text categorisation can be single-label or multi-label. In single-label categorisation, only one category is assigned to the text (assuming non-overlapping categories), while in multi-label categorisation, any number of categories can be assigned [15].

In this paper we are interested in the categorisation of short-form messages using single-label categorisation with non-overlapping categories. We use the same approach to categorise micro-blog messages as in our past work [6]. Basically, the approach involves the creation of an index, representing categories, from which category recommendations can subsequently be made for target messages. In this index individual categories can be viewed as documents made up of the set of terms contained in their associated messages. We can then retrieve documents (that is categories) based on the terms that are present in their categories. Further, we apply *weights* to the terms that are associated with a given category based on how representative or informative these terms are with respect to the category in question. Here we use the well known TFIDF approach [13] to term weighting; see Equation 1. In this work we use Lucene[3] to provide this indexing and term-weighting functionality.

$$\mathbf{C_{ij}} = \text{TFIDF}(C_i, t_j, \mathbf{C}) \tag{1}$$

[3] `http://lucene.apache.org/`

Thus, the category of a target message m_T can be determined as follows. By using the term-vector representation for m_T as a query, we can retrieve the most similar categories from the index using Lucene, ranked according to their similarity to m_T. For a detailed explanation of this approach refer to [6].

In this paper, we consider only the most similar category to the target message and present this as the target message's category recommendation. Of course, multiple categories can be recommended for messages by simply selecting the top-k most similar categories; however, we leave an analysis of such an approach to future work.

4 Evaluation

In this section, we evaluate the categorisation performance provided by the our approach as described above. We begin by describing the datasets used in our evaluation and the experimental methodology employed.

4.1 Datasets

We collected Blippr data using the Blippr API in April 2010, capturing blips authored by users prior to that date and after November 2007 (some data had to be scraped from the website due to the limitations of the API) . We performed the same preprocessing on the extracted blips as in [6]. Further, messages consisting of default text ("your opinion of this product in 160 characters or less") were also discarded. Finally, we randomly selected the same number (1,600) of blips for each of the five categories (where category labels provide the classification ground truth), equal to the number of blips present for the smallest category (*music*). In total, 3,887 distinct blip authors were represented in our dataset.

For the Twitter dataset, we used the Twitter API to collect tweets posted between November 2008 and June 2010. In order to compare performance on the Twitter and Blippr datasets, we used hashtags to select tweets corresponding to each of the five Blippr categories. The hashtag used to select the tweets for each category is equivalent to the category name, in singular and plural (e.g. for the category *movie* we used the hashtags *#movie* and *#movies*). We performed the same preprocessing steps for tweets as in [6]. As with blips, we randomly selected the same number (1,600) of tweets for each category to perform our evaluation. Overall categories 3,203 distinct tweet authors were represented.

Due to slight differences in the selection of the data with respect to our past work, some variations can be seen in the results compared with past results [6].

4.2 Results

For both datasets, we randomly selected 80% of the messages from each category as training data and used the remaining 20% as test messages. Each test message was categorised in turn using the approach described in Section 3 and the percentage of times that the approach produced the correct categorisation was

recorded for each of the five categories. We repeated this procedure five times and calculated the mean categorisation accuracy across each of the categories considered. Results are shown in Table 1. Examining overall performance, it can be seen that the mean accuracy performance across all categories is higher for Blippr (80.9%) than for Twitter (72.6%). This finding is to be expected, given the more structured nature of the Blippr domain (i.e. all blips posted are restricted to comments on one of five product types); in contrast, Twitter is a much more heterogeneous and noisy domain (which, for example, required the additional step of having to identify dataset tweets by hashtag), and thus the relatively high categorisation accuracy of 72.6% achieved for this dataset is an encouraging result. The accuracy across individual categories observed is also higher for the Blippr dataset. Overall, we believe that these results are promising given the relative simplicity of our categorisation approach and its ability to provide good performance in both of the domains evaluated.

Table 1. Categorisation accuracy (%) for Blippr and Twitter datasets.

	Movies	Books	Music	Apps	Games	Mean
Blippr	76.3	76.6	81.4	81.4	89.1	80.9
Twitter	65.6	69.7	68.5	81.5	77.6	72.6

Index Size. An important consideration is the number of training messages used to create the category index. In the above analysis, we used all available messages (1,300) from each category to create the index. Here, we consider the number of training messages that are required to provide good coverage for each of the five categories examined. In the following experiments, we employed the same procedure as outlined above but consider a range of training sizes; as before, 300 test messages were used to evaluate the performance for each category and training set size combination.

Results are shown in Figures 3a and 3b for the Blippr and Twitter datasets, respectively. Although for most categories the rate of accuracy improvement declined significantly for training set sizes above 400 messages per category, we note that accuracy continued to increase in all cases with the addition of new training data, even beyond training set sizes of 1,000 examples. This indicates that significant numbers of messages are required to characterise the vocabulary of each category (although this is unlikely to be a problem given the plentiful nature of micro-blogs). Further, and perhaps of greater significance, it is likely that in a real-world deployment the category index would need to be updated on an ongoing basis with new training messages in order to capture emerging vocabulary relating to different categories (e.g. new actors, games etc.); we leave a detailed analysis of this question of topic drift to future work.

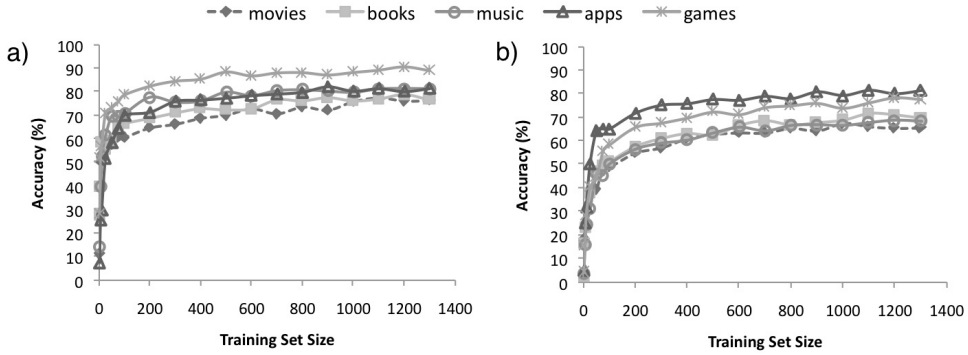

Fig. 3. Accuracy vs. category training set size for Blippr(a) and Twitter(b)

Comparison with Other Methods. In order to benchmark the performance of our categorisation approach, here we repeat the previous categorisation tests using machine learning techniques. In particular, we consider multinomial Naïve Bayes (NB) and Support Vector Machines (SVM), which were found to provide good categorisation performance in [9, 20]. We tested different SVM kernels and chose a linear kernel, which achieved the best performance on our task. In addition, we consider the Random Forest (RF) classifier using 30 trees, which is an ensemble approach consisting of multiple decision trees. In all cases, Weka[4] is used to perform classification.

Further, since we use all message terms to create the category indices using our approach, we also use all terms in the machine learning methods. We leave to future work the application of feature selection in respect of the classifiers and also of our index-based approach. In addition, we note the difference between the our index-based approach and the classification approaches regarding data representation. In the case of the classifiers, each instance represents a message, while in our index-based approach each instance represents an entire category (i.e. the terms from all the messages of that category).

Tables 2 and 3 show the performance achieved by the different classifiers and our index-based approach on the Blippr and Twitter datasets, respectively. The results indicate that the mean accuracy of our indexing approach equals or outperforms SVM and RF for both Blippr and Twitter, and our approach performs best for the *games* category in both domains. Although NB provides the best mean accuracy, the overall differences are relatively small, which demonstrates the inherent potential of our approach. In addition, we note that further extensions to our approach are possible. For example, our approach currently uses Lucene's default parameters for indexing and retrieval (e.g. query-document similarity, term weighting scheme etc.). In future work, we will consider alternatives to these parameters and investigate the effect of such parameters on categorisation performance.

[4] `www.cs.waikato.ac.nz/ml/weka/`

Table 2. Categorisation accuracy (%) vs. method (Blippr dataset). The best performing method for each category is indicated in **boldface**.

	Movies	Books	Music	Apps	Games	Mean
Index-Based	76.3	76.6	81.4	81.4	**89.1**	80.9
NB	**78.9**	**77.0**	**86.4**	83.9	88.0	**82.9**
SVM	73.1	72.5	82.3	**86.9**	79.1	78.8
RF	75.2	72.8	78.3	85.8	78.9	78.2

Table 3. Categorisation accuracy (%) vs. method (Twitter dataset). The best performing method for each category is indicated in **boldface**.

	Movies	Books	Music	Apps	Games	Mean
Index-Based	65.6	69.7	68.5	81.5	**77.6**	72.6
NB	69.0	70.5	69.8	**82.9**	73.6	**73.2**
SVM	**70.3**	**71.6**	69.7	79.4	72.0	72.6
RF	66.3	66.6	**71.5**	79.5	70.3	70.8

Cross-Domain Analysis. Finally, we study the potential for cross-domain categorisation using our approach. In order to do this, we create the category index using training messages from one domain and examine the accuracy provided on test messages from another domain. As before, we randomly selected 80% of the messages per category from one domain as training data and 20% of the messages per category from the other domain as test data. We repeated this procedure five times and computed the average accuracy achieved for the five categories. Results are presented in Table 4.

Interestingly, using Twitter as training data to categorise blips achieves a much better performance than using Blippr data to categorise tweets (65.2% vs. 49.3%). In addition, we observe that the cross-domain mean accuracy achieved when training with tweets (65.2%) is not much worse than when Twitter data is used for both training and testing (72.6%). This contrasts sharply with the Blippr dataset, where cross-domain mean accuracy using Blippr messages as training data is 49.3%, compared to a mean accuracy of 80.9% when training and testing on Blippr data alone. Our preliminary analysis suggests that messages from the Twitter domain are more heterogeneous in nature and contain more diverse terms compared to Blippr messages, allowing for better generalisation and categorisation of unseen data from other sources. A detailed analysis of the vocabulary statistics from these domains is, however, required in order to prove this hypothesis; such an analysis we leave to future work.

5 Conclusions and Future Work

In this paper, we have presented an approach to categorise micro-blog messages in order to provide for better search and message filtering. In particular, we

Table 4. Cross-domain categorisation accuracy (%) (training domain indicated in **boldface**).

	Movies	Books	Music	Apps	Games	Mean
Blippr/Twitter	45.2	47.3	44.7	56.9	52.4	49.3
Twitter/Blippr	67.0	67.2	71.1	45.5	75.1	65.2

have described how such messages can be used as a source of indexing and retrieval information. A preliminary evaluation performed using data from two micro-blogging domains indicates that our approach shows promising performance, suggesting that micro-blog messages in sufficient quantities provide a useful recommendation signal, despite their often inconsistent use of language and short length. We have also compared our index-based approach with classification techniques commonly used in text categorisation and found that our approach performed well, achieving accuracies close to the best performing classifier. Moreover, there is scope for expanding our bag-of-words style indexing approach by modifying Lucene's default parameters and by including additional message features. In future work, we will also study topic drift and the rate at which retraining the index is required to accurately categorise emerging category terms and topics.

6 Acknowledgements

Based on work supported by Science Foundation Ireland, Grant No. 07/CE/I1147.

References

1. Angel, A., Koudas, N., Sarkas, N., Srivastava, D.: What's on the grapevine? In: SIGMOD '09: Proceedings of the 35th SIGMOD international conference on Management of data. pp. 1047–1050. ACM, New York, NY, USA (2009)
2. Banerjee, N., Chakraborty, D., Dasgupta, K., Mittal, S., Joshi, A., Nagar, S., Rai, A., Madan, S.: User interests in social media sites: an exploration with micro-blogs. In: CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management. pp. 1823–1826. ACM, New York, NY, USA (2009)
3. Brooks, C.H., Montanez, N.: Improved annotation of the blogosphere via autotagging and hierarchical clustering. In: WWW '05: Proceedings of the 15th international conference on World Wide Web. pp. 625–632. ACM, New York, NY, USA (2006)
4. Chen, H., Dumais, S.: Bringing order to the web: automatically categorizing search results. In: CHI '00: Proceedings of the SIGCHI conference on Human factors in computing systems. pp. 145–152. ACM, New York, NY, USA (2000)
5. Cormack, G.V.: Email spam filtering: A systematic review. Foundations and Trends in Information Retrieval 1(4), 335–455 (2007)

6. Garcia Esparza, S., P. O'Mahony, M., Smyth, B.: Towards tagging and categorization for micro-blogs. In: AICS '10: Proceedings of the 21st National Conference on Artificial Intelligence and Cognitive Science. Galway, Ireland (2010)
7. Gómez Hidalgo, J.M., Bringas, G.C., Sánz, E.P., García, F.C.: Content based sms spam filtering. In: DocEng '06: Proceedings of the 2006 ACM symposium on Document engineering. pp. 107–114. ACM, New York, NY, USA (2006)
8. Jansen, B.J., Zhang, M., Sobel, K., Chowdury, A.: Micro-blogging as online word of mouth branding. CHI EA '09: Proceedings of the 27th international conference extended abstracts on Human factors in computing systems p. 3859 (2009), http://portal.acm.org/citation.cfm?doid=1520340.1520584
9. Joachims, T.: Text categorization with support vector machines: learning with many relevant features. In: ECML '98: European Conference on Machine Learning. Springer-Verlag, London, UK (1998)
10. Pandey, V., Iyer, C.: Sentiment analysis of microblogs (2009)
11. Phan, X.H., Nguyen, L.M., Horiguchi, S.: Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In: WWW '08: Proceeding of the 17th international conference on World Wide Web. pp. 91–100. ACM, New York, NY, USA (2008)
12. Qi, X., Davison, B.D.: Web page classification: Features and algorithms. ACM Computing Surveys 41(2), 1–31 (2009)
13. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill, Inc., New York, NY, USA (1986)
14. Sarah Jane Delan, P.C., Smyth, B.: Ecue: A spam filter that uses machine leaming to track concept drift. In: ECAI '06: European Conference on Artificial Intelligence. pp. 627–. Riva del Garda, Italy (2006)
15. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys 34(1), 1–47 (2002)
16. Shepitsen, A., Gemmell, J., Mobasher, B., Burke, R.: Personalized recommendation in social tagging systems using hierarchical clustering. In: RecSys '08: Proceedings of the 2008 ACM conference on Recommender systems. pp. 259–266. ACM, New York, NY, USA (2008)
17. Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H.: Short text classification in twitter to improve information filtering. In: SIGIR '10: Proceeding of the 33rd international conference on Research and Information Retrieval (2010)
18. Sun, A., Suryanto, M.A., Liu, Y.: Blog classification using tags: an empirical study. In: ICADL'07: Proceedings of the 10th international conference on Asian digital libraries. pp. 307–316. Springer-Verlag, Berlin, Heidelberg (2007)
19. Teevan, J., Ramage, D., Morris, M.R.: #twittersearch: a comparison of microblog search and web search. In: WSDM '11: Proceedings of the fourth ACM international conference on Web search and data mining. pp. 35–44. ACM, New York, NY, USA (2011)
20. Yang, Y., Liu, X.: A re-examination of text categorization methods. In: SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. pp. 42–49. ACM, New York, NY, USA (1999)
21. Yeung, C.m.A., Gibbins, N., Shadbolt, N.: Tag meaning disambiguation through analysis of tripartite structure of folksonomies. In: WI-IATW '07: Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Workshops. pp. 3–6. IEEE Computer Society, Washington, DC, USA (2007)