Towards Tagging and Categorisation for Micro-blogs

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@ucdconnect.ie,
{michael.p.omahony,barry.smyth}@ucd.ie
http://www.clarity-centre.org/

Abstract. Micro-blogging services are becoming very popular among users who want to share local or global news, their knowledge or their opinions on the real-time web. Lately, users are also using these services to search for information, and some services include *tag* or *category* information to better facilitate search. However, these tags are typically free-form in nature with users permitted to adopt their own conventions without restriction, which can make the set of tags noisy and sparse. A solution to this problem is to recommend tags (or categories) to users. Our work represents an initial study in the recommendation of categories for short-form messages in order to provide for better search and message filtering. In particular, we describe how such real-time web data can be used as a source of indexing and retrieval information for category recommendation. An evaluation performed on two different micro-blogging datasets indicates that promising performance is achieved by our approach.

Keywords: Micro-blogs, categorisation, tagging, recommendation

1 Introduction

Micro-blogs are short textual messages written by users to share information including comments, opinions and personal viewpoints. In the last few years researchers and industry have expressed their interest in the power of micro-blogs for different purposes, such as sentiment analysis[7, 8], hot topics (or trend) discovery[1] and social analysis [2, 17].

The creation of Twitter[1] in 2006 represented a boom for micro-blogging, with almost 106 million users sending some 55 million messages (or *tweets*) every day using this service[2]. Users have also been using micro-blogging services to search for information. For instance, Twitter has its own search engine[3] where

[1] http://www.twitter.com/
[2] http://www.businessinsider.com/twitter-stats-2010-4. Accessed: 20/06/2010.
[3] http://search.twitter.com/

users can find information about the latest movies, opinions on a certain product or real-time updates on an ongoing football match or music concert. As such, these micro-blog search engines can be seen as complimentary to more traditional search engines, being especially useful when it comes to finding real-time information, a task not readily supported by traditional search engines.

While undoubtably plentiful, micro-blog messages are, however, inherently noisy and unstructured. Micro-blog authors can comment or express their opinions on an unlimited variety of topics which makes this data difficult to organise and process. Some services like Blippr[4], 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 preceeded by the 'hash' symbol (#) and it is used to assist in the categorisation of tweets. Examples of hashtags are: *#travel*, *#movies*, *#transformers*, *#fifa*, etc.

The use of such free-form tags introduce problems like sparsity, temporality, ambiguity (a single tag has many meanings, e.g: *#paris* can refer to the French capital or to the celebrity, Paris Hilton) and redundancy (several tags have the same meaning, e.g: *#nowreading* and *#reading*) [5]. In order to mitigate these problems, recommender systems have been used to recommend tags to users for specific resources. For example, previous work has investigated tag or category recommendations for blogs [3, 16]. However, less attention has been focused on micro-blogs, where the text is much shorter and noisier. The aim of this paper is to develop a tag 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 mirco-blog content can be used as a source of indexing and retrieval information for tag (category) recommendation. We apply our approach to messages taken from the Blippr and Twitter services, and evaluate the performance of our approach in respect of five product categories (those currently supported by the Blippr service). In addition, we consider the potential for cross-domain recommendation, by training our model using messages from one domain and recommending tags 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 years, both researchers and industry have expressed interest in the power of micro-blogs. One of the most popular areas of research deals with

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

the problem of tag recommendation and category recognition for such content. To facilitate the indexing and retrieval of content, the Twitter community, for example, has adopted the convention of annotating messages with hashtags as a means of grouping and identifying related messages. 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. In other domains, tag recommender systems have been developed in order to assist users when tagging and searching for content. For example, Delicious was one of the earliest systems that introduced social tagging and tag recommendations, although their algorithms have not been published. In [14], different strategies are proposed to recommend tags for Flickr photos. This work is based on the assumption that some tags already exist for a given photo and new tags are identified using tag co-ccurrence data. In [6], the authors investigate the utility of tags to recommend documents. In order to perform these recommendations the system learns a semantic space for users, tags and documents (objects) where similar objects are close to each other. This semantic space allows for the computation of similarity between instances and therefore for the most similar instances to be recommended.

The problem of tag recommendation is closely linked to that of category recommendation where, for example, categories can be considered to represent generic tags. One such example is text categorisation in the area of information retrieval [12]. The most popular uses of text categorisation are text filtering, hierarchical categorisation of documents (e.g. grouping web pages by topics or organising emails into personal folders) [10] and word sense disambiguation [18]. These applications have been used on different occasions for tag recommendation. 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, [13] 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.

Most of the work related to tag recommendation and text categorisation has been applied to long-form text. Categorising micro-blogs, however, poses additional challenges given the short length of these messages and the practically unlimited range of topics that can be discussed. Some solutions to this problem use external knowledge to make the data less sparse and to discover relationships in the data. For instance, in [9] large amounts of data are collected from external sources which is applied to classify short and sparse text and Web segments. In [15] the authors 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 the message itself to perform categorisation, rather than on the particular intention of the message author.

3 Micro-blog Categorisation

In this paper we consider messages from two services. The first is the well known micro-blogging service, Twitter. Since its creation in 2006, Twitter has gained worldwide popularity and allows users to share their knowledge and opinions on any topic they care to tweet about in a 140-character, or less, message.

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 1 shows a typical Bippr review for the movie *'The Matrix'*.

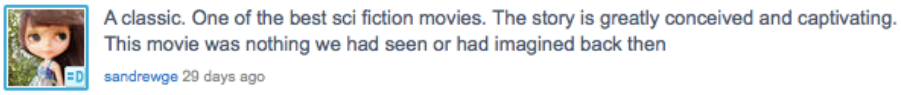

Fig. 1. A Blippr review of the movie *'The Matrix'*.

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 used to assign single or multi-label to messages. In single-label categorisation, only one category is assigned to the text (assuming non-overlapping categories), while in multilabel categorisation, any number of categories can be assigned [12]. In this paper we are interested in the categorisation of short-form messages using single-label categorisation with non-overlapping categories.

We now consider how micro-blog messages can be used as a source of indexing and retrieval information. Our approach involves the creation of an index, representing categories, from which (single) category recommendations can subsequently be made for target messages. Consider a category C_i which is associated with a set of messages as per Equation 1. In turn, each message is made up of a set of terms and so each category can be represented as a set of terms (drawn from messages) using a bag-of-words style approach [11] according to Equation 1.

$$C_i = \{m_1, ..., m_k\} = \{t_1, ..., t_n\} \tag{1}$$

In this way individual categories can be viewed as documents made up of the set of terms (words) contained in their associated messages. We can create an index of these documents so that we can retrieve documents (that is categories) based on the terms that are present in their categories. Using techniques from the information retrieval community, 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 [11] to term weighting (Equation 2). Briefly, the weight of a term t_j in a category C_i, with respect to some collection of categories $\mathbf{C}$, is proportional to the frequency of occurrence of t_j in C_i (denoted by n_{t_j,C_i}), but inversely proportional to the frequency of occurrence of t_j in $\mathbf{C}$ overall, thereby giving preference to terms that help to discriminate C_i from the other categories in the collection.

$$\text{TFIDF}(C_i, t_j, \mathbf{C}) = \frac{n_{t_j,C_i}}{\sum_{t_k \in C_i} n_{t_k,C_i}} \times \log\Big(\frac{|\mathbf{C}|}{|\{C_k \in \mathbf{C} : t_j \in C_k\}|}\Big) \tag{2}$$

Thus we can create a term-based index of categories $\mathbf{C}$, such that each entry $\mathbf{C_{ij}}$ encodes the importance of term t_j in category C_i; see Equation 3. In this work we use Lucene[5] to provide this indexing and term-weighting functionality.

$$\mathbf{C_{ij}} = \text{TFIDF}(C_i, t_j, \mathbf{C}) \tag{3}$$

Once the above index has been created using a training set of messages from each of the categories under consideration, 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. In this paper, we only consider the most similar category to the target message (i.e. a top-1 approach) and present this as the target message's category recommendation.

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 Blippr Dataset

For the Blippr dataset, we focused on strong-positive blips only (i.e. where users have expressed the highest sentiment toward items using the *love it* rating). We collected data from the website using the Blippr API in April 2010, capturing blips written by users prior to that date and after November 2007 (some data

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

had to be scraped from the website due to the limitations of the API) . We performed some preprocessing on the extracted blips such as removing stopwords, special symbols, digits and multiple repetitions of characters in words (e.g. we reduce *faaaaabulous* to *fabulous*). Then we considered only those blips that are written in the English language and with a minimum of three words. 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.

4.2 Twitter Dataset

We used tweets collected using the Twitter API between November 2008 and June 2010. In order to compare performance on the Twitter and Blippr datasets, we manually selected hashtag-to-category mappings for the Twitter dataset so as to select tweets that correspond to each of the five Blippr categories. Then we selected only those tweets that contained the given hashtags. The hashtags used to map tweets to each category are shown in Table 1 (we chose only one hashtag for *music* because there was a large number of tweets available for this category). We performed the same preprocessing steps for tweets as described above for blips. Further, because users can retweet (resend an existing tweet), some tweets may be repeated; all repeated tweets were removed from our dataset. Since our focus is to assign non-overlapping categories to messages, we also removed those blips that belonged to more than one category. Finally, we removed all hashtags present in the tweets; including those that are used to select the tweet category and those that are not, since these latter hashtags may also contain category information. As with blips, we randomly selected the same number (1,600) of tweets for each category to perform our evaluation. Over all categories, 3,203 distinct tweet authors were represented.

Table 1. Lists of hashtags used to map tweets to each category.

Category	Hashtags
Movies	movie, movies, film, films, cinema
Books	book, books, comic, comics, bookreview, graphicnovel, reading, readingnow, literature, 100booksin2010
Music	music
Apps	app, apps, application, software, androidapps, appstore, googleapps, mobileapps2010
Games	pcgames, videogames, videogame, gaming, gamer, xbox, playstation, ps3, psp, wii, nintendo

4.3 Results

For both datasets, we randomly selected 1,500 messages from each category as training data and used the remaining 100 messages from each category 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 average categorisation accuracy across each of the categories considered. Results are shown in Table 2, where it can be seen that good performance was achieved by the approach, with categorisation accuracies ranging from a minimum of 69% for *applications* to a maximum of 85% for *games* (Blippr dataset). While the relative ordering of accuracy across categories was different for the two domains — for example, the highest accuracy was achieved for *games* (85%) in the Blippr dataset and for *applications* (84%) in the Twitter datset — similar performance was seen for the *movies* (75%) and *music* (72%) categories for both datasets. In addition, the mean accuracy across all categories was very similar for the two datasets, with mean accuracies of 75% and 76% being achieved for the Blippr and Twitter datasets, respectively. Overall, these results are promising given the relative simplicity of the approach and its ability to provide good performance in both of the domains evaluated.

Table 2. Categorisation accuracy (%) for Blippr and Twitter datasets.

	Movies	Books	Music	Apps	Games	Mean
Blippr	75	76	72	69	85	75
Twitter	75	71	72	84	80	76

An important consideration in our approach is the number of training messages used to create the category index. In the above analysis, we used all available messages (1,500) 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, 100 test messages were used to evaluate the performance for each category and training set size combination.

Results are shown in Figures 2 and 3 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 an analysis of this question of topic drift to future work.

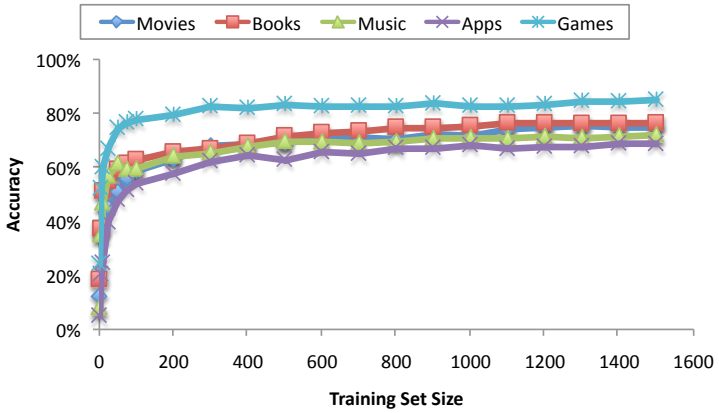

Fig. 2. Blippr dataset: accuracy vs. category training set size.

Finally, we investigate the potential for cross-domain categorisation using our approach. Here, we create the category index using training messages from one domain and examine the accuracy provided on test messages from another domain. In this case, we randomly selected 1,500 Blippr messages from each category as training data and 100 Twitter messages from each category as test data. We repeated this procedure five times and computed the average accuracy achieved for the five categories. Results are presented in Table 3. Clearly, the accuracies achieved for all categories were significantly poorer that those recorded when training and testing using messages from the same domain. For example, the mean accuracy across all categories was 49% for cross-domain categorisation, compared to a mean accuracy of 75% (resp. 76%) for training and testing using Blippr (resp. Twitter) messages (Table 2). These findings indicate that the vocabulary used in messages is quite different in both domains. Due to limitations of space, we leave to future work an analysis of the differences between the domain vocabularies.

Table 3. Cross domain categorisation: accuracy (%) achieved categorising Twitter test messages using Blippr training messages.

Movies	Books	Music	Apps	Games	Mean
43	43	41	62	59	49

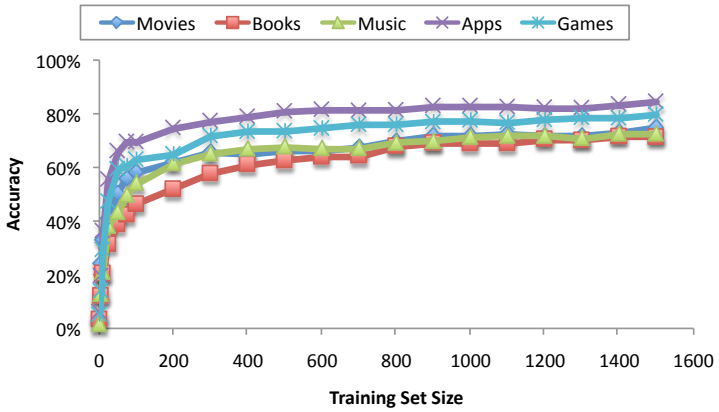

Fig. 3. Twitter dataset: accuracy vs. category training set size.

5 Conclusions and Future Work

In this paper, we have presented an initial study of category recommendation for micro-blog messages in order to provide for better search and message filtering. In particular, we have described how such messages can be used as a source of indexing and retrieval information. An 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 inconsistent use of language and short length. In future work, we plan on expanding on our bag-of-words indexing approach to include additional message features; for example, by parsing messages and extracting grammatical relations [4] which can also be added to the category index as 'virtual words'. In addition, we will examine a broader range of categories and also consider multi-label category recommendation for messages.

6 Acknowledgements

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

References

1. A. Angel, N. Koudas, N. Sarkas, and D. Srivastava. What's on the grapevine? In *SIGMOD '09: Proceedings of the 35th SIGMOD international conference on Management of data*, pages 1047–1050, New York, NY, USA, 2009. ACM.
2. N. Banerjee, D. Chakraborty, K. Dasgupta, S. Mittal, A. Joshi, S. Nagar, A. Rai, and S. Madan. 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*, pages 1823–1826, New York, NY, USA, 2009. ACM.

3. C. H. Brooks and N. Montanez. Improved annotation of the blogosphere via autotagging and hierarchical clustering. In *Proceedings of the 15th international conference on World Wide Web*, pages 625–632, New York, NY, USA, 2006. ACM.
4. M.-C. de Marneffe, B. MacCartney, and C. D. Manning. Generating typed dependency parses from phrase structure parses. In *In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 449–454, Genoa, Italy, May 24–26 2006.
5. J. Gemmell, M. Ramezani, T. Schimoler, L. Christiansen, and B. Mobasher. The impact of ambiguity and redundancy on tag recommendation in folksonomies. In *RecSys '09: Proceedings of the third ACM conference on Recommender systems*, pages 45–52, New York, NY, USA, 2009. ACM.
6. Z. Guan, C. Wang, J. Bu, C. Chen, K. Yang, D. Cai, and X. He. Document recommendation in social tagging services. In *WWW '10: Proceedings of the 19th international conference on World wide web*, pages 391–400, New York, NY, USA, 2010. ACM.
7. B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Micro-blogging as online word of mouth branding. *Proceedings of the 27th international conference extended abstracts on Human factors in computing systems - CHI EA '09*, page 3859, 2009.
8. V. Pandey and C. Iyer. Sentiment analysis of microblogs, 2009.
9. X.-H. Phan, L.-M. Nguyen, and S. Horiguchi. 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*, pages 91–100, New York, NY, USA, 2008. ACM.
10. M. E. Ruiz and P. Srinivasan. Hierarchical text categorization using neural networks. *Inf. Retr.*, 5(1):87–118, 2002.
11. G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986.
12. F. Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47, 2002.
13. A. Shepitsen, J. Gemmell, B. Mobasher, and R. Burke. Personalized recommendation in social tagging systems using hierarchical clustering. In *RecSys '08: Proceedings of the 2008 ACM conference on Recommender systems*, pages 259–266, New York, NY, USA, 2008. ACM.
14. B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 327–336, New York, NY, USA, 2008. ACM.
15. B. Sriram, D. Fuhry, E. Demir, and H. Ferhatosmanoglu. Short text classification in twitter to improve information filtering. In *SIGIR '10: Proceeding of the 33rd international conference on Research and Information Retrieval*, 2010.
16. A. Sun, M. A. Suryanto, and Y. Liu. Blog classification using tags: an empirical study. In *ICADL'07: Proceedings of the 10th international conference on Asian digital libraries*, pages 307–316, Berlin, Heidelberg, 2007. Springer-Verlag.
17. T. Turner, P. Qvarfordt, J. T. Biehl, G. Golovchinsky, and M. Back. Exploring the workplace communication ecology. In *CHI '10: Proceedings of the 28th international conference on Human factors in computing systems*, pages 841–850, New York, NY, USA, 2010. ACM.
18. C.-m. A. Yeung, N. Gibbins, and N. Shadbolt. 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*, pages 3–6, Washington, DC, USA, 2007. IEEE Computer Society.