Topic Classification in Social Media using Metadata from Hyperlinked Objects*

Sheila Kinsella[1], Alexandre Passant[1], and John G. Breslin[1,2]

[1] Digital Enterprise Research Institute, National University of Ireland, Galway
{firstname.lastname}@deri.org
[2] School of Engineering and Informatics, National University of Ireland, Galway
john.breslin@nuigalway.ie

Abstract. Social media presents unique challenges for topic classification, including the brevity of posts, the informal nature of conversations, and the frequent reliance on external hyperlinks to give context to a conversation. In this paper we investigate the usefulness of these external hyperlinks for determining the topic of an individual post. We focus specifically on hyperlinks to objects which have related metadata available on the Web, including Amazon products and YouTube videos. Our experiments show that the inclusion of metadata from hyperlinked objects in addition to the original post content improved classifier performance measured with the F-score from 84% to 90%. Further, even classification based on object metadata alone outperforms classification based on the original post content.

1 Introduction

Social media such as blogs, message boards, micro-blogging services and social-networking sites have grown significantly in popularity in recent years. By lowering the barriers to online communication, social media enables users to easily access and share content, news, opinions and information in general. However navigating this wealth of information can be problematic due to the fact that contributions are often much shorter than a typical Web documents, and the quality of content in social media is highly variable [1].

A potential source of context to conversations in social media is the hyperlinks which are frequently posted to refer to related information. These objects are often an integral part of the conversation. For example, a poster may recommend a book by posting a link to a webpage where you can buy the book, rather than employing the traditional method of providing the title and the name of the author. Yet there still remains the question of identifying which snippets of content from these links are most relevant to the conversation at hand.

Recently, there has been a trend towards publishing structured information on the Web, resulting in an increasing amount of rich metadata associated with Web resources. Facebook have launched their Open Graph Protocol[3], which

* The work presented in this paper has been funded in part by Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2).

[3] http://opengraphprotocol.org/ visited January 2011

allows external site owners to markup their content using Facebook-defined schemas, such that these enriched content items can then be used for metadata import into news feeds, profiles, etc. The Linking Open Data [5] project is a community effort to make structured datasets from diverse domains available on the Web, some of which we use as data sources in this work. Such rich representations of objects can be a useful resource for information retrieval and machine learning. In social media in particular, structured data from hyperlinked websites can provide important context in an otherwise chaotic domain.

In this paper, we investigate the potential of improving topic classification in social media by using metadata retrieved from external hyperlinks in user-generated posts[4]. We compare the results of topic classification based on post content, based on metadata from external websites, and based on a combination of the two. The usage of structured metadata allows us to include only specific, relevant external data. Our experiments show that incorporating metadata from hyperlinks can significantly improve the task of topic classification in social media posts. Our approach can be applied to recommend an appropriate forum in which to post a new message, or to aid the navigation of existing, uncategorised posts.

Related work has been carried out in the field of Web document classification, proving that the classification of webpages can be boosted by taking into account the text of neighbouring webpages ([2], [9]). This work differs in that we incorporate not entire webpages but only metadata which is directly related to objects discussed in a post. There has also been previous work using tags and other textual features to classify social media. Our work is closely related to that of Figueiredo et al. [6], who assess the quality of various textual features in Web 2.0 sites such as YouTube for classifying objects within that site. Berendt and Hanser [4] investigated automatic domain classification of blog posts with different combinations of body, tags and title. Sun et al. [10] showed that blog topic classification can be improved by including tags and descriptions. Our paper differs from these because we use metadata from objects on the Web to describe a post which links to those objects. Thus our approach combines the usage of neighbouring pages in Web search, and of metadata in social media search.

2 Data Corpus

Our analysis uses the message board corpus from the `boards.ie` SIOC Data Competition[5] which was held in 2008 and covers ten years of discussions. Each post belongs to a thread, or conversation, and each thread belongs to a forum. A forum typically covers one particular area of interest, and may contain sub-forums which are more specialised (for example, Music and Hip-Hop). Our analysis uses a subset of the forums and is limited to the posts contained in the final year of the dataset, because these are most likely to have structured data. We examined the posts in the dataset which contained hyperlinks and identified potential sources of structured metadata. These are websites which publish

[4] A post is a message which a user submits as part of a conversation

[5] `http://data.sioc-project.org/` visited January 2011

metadata about objects, such as videos (YouTube) or products (Amazon), and make it available via an API or as Linked Data. In some cases the metadata is published by external sources, e.g. DBpedia [3] provides a structured representation of Wikipedia. The sources on which our study focuses are listed in Table 1, along with the available metadata that we extracted. We consider the first paragraph of a Wikipedia article as a description. For this study, we focus on the most commonly available metadata, but in the future we plan to make use of additional information such as movie directors, book authors and music albums.

Website	Object type	Title	Description	Category	Tags
Amazon	Product	X		X	X
Flickr	Photo	X	X		X
IMDB	Movie	X		X	
MySpace	Music Artist	X		X	
Wikipedia	Article	X	X	X	
YouTube	Video	X	X	X	X

Table 1. External websites and the metadata types used in our experiments.

Amazon, Flickr and YouTube metadata was retrieved from the respective APIs. We obtained MySpace music artist information from DBTune[6] (an RDF wrapper of various musical sources including MySpace), IMDB movie information from LinkedMDB[7] (an export of IMDB data) and Wikipedia article information from DBpedia[8]. The latter three services are part of the Linking Open Data project [5]. The data collection process is described in more detail in [8].

Two groups of forums were chosen for classification experiments - the ten rather general forums shown in Figure 1(a), which we refer to as **General**, and the five more specific and closely related music forums shown in Figure 1(b), **Music**. These forums were chosen since they have good coverage in the dataset and they each have a clear topic (as opposed to general "chat" forums). The percentage of posts in **General** that have hyperlinks varies between forums, from 4% in Poker to 14% in Musicians, with an average of 8% across forums. Of the posts with hyperlinks, 23% link to an object with available structured data. The number of posts containing links to each type of object is shown in Figure 1. For **General**, hyperlinks to Music Artists occur mainly in the Musicians forum, Movies in the Films forum, and Photos in the Photography forum. The other object types are spread more evenly between the remaining seven forums. Note that in rare cases, a post contains links to multiple object types, in which case that post is included twice in a column. Therefore the total counts in Figure 1 are inflated by approximately 1%. In total, **General** contains 6,626 posts and **Music** contains 1,564 posts.

[6] `http://dbtune.org/` visited January 2011
[7] `http://linkedmdb.org/` visited January 2011
[8] `http://dbpedia.org/` visited January 2011

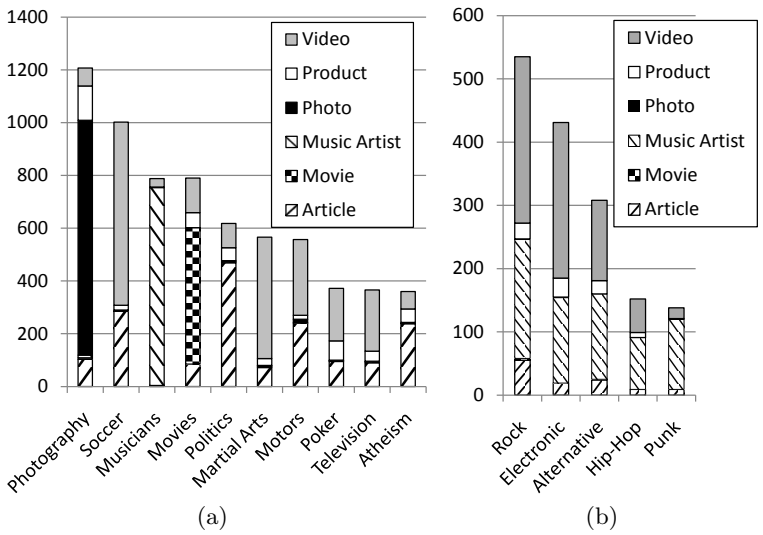

Fig. 1. Number of posts containing each type of hyperlinked object for (a) 10 **General** forums and (b) 5 **Music** forums

3 Experiments

We evaluated the usefulness of various textual representations of message board posts for classification, where the class of a post is derived from the title of the forum in which it was posted. For each post, we derive four different bag-of-words representations, in order to compare their usefulness as sources of features for topic classification. **C-L** denotes the original content of the post, with all hyperlinks removed, while **C** denotes the full original content with hyperlinks intact. **M** denotes the external metadata retrieved from the hyperlinks of the post. **C+M** denotes the combination of **C** and **M** for a given post, i.e. the full original content plus the external metadata retrieved from its hyperlinks. For the 23% of posts which have a title, this is included as part of the content. The average number of unique terms was 38 for post content, and 20 for associated metadata. At present we simply concatenate the text of the metadata values rather than considering the metadata key-value pairs, however it would be interesting to weight the metadata based on which keys provide the most useful descriptors for classification.

Classification of documents was performed using the Multinomial Naïve Bayes classifier implemented in Weka [7]. For each textual representation of each post the following transformations are performed. All text is lower-cased and non-alphabetic characters are replaced with spaces. Stopwords are removed and TF-IDF and document length normalisation are applied.

Ten-fold cross validation was used to assess classifier performance on each type of document representation. To avoid duplication of post content due to

one post quoting another, the data was split by thread so that posts from one thread do not occur in separate folds. Duplication of hyperlinks across splits was also disallowed, so the metadata of an object cannot occur in multiple folds. These restrictions resulted in the omission of approximately 11% of the posts in each forum group. The same ten folds are used to test **C-L**, **C**, **M** and **C+M**.

The results of the classification experiments are shown in Table 2. We performed a paired t-test at the 0.05 level over the results of the cross-validation in order to assess statistical significance. For **General**, all differences in results are significant. Simply using the content without hyperlinks (**C-L**) gives quite good results, but incorporating URL information from hyperlinks (**C**) improves the results. Using only metadata from hyperlinked objects (**M**) gives improved performance, and combining post content with metadata from hyperlinks (**C+M**) gives the best performance. For **Music**, the scores in general are lower, likely due to the higher similarity of the topics. Here, the difference between **M** and **C+M** is not significant, but all other differences are significant. When a post has a hyperlink, object metadata, with or without post content, provides a better description of the topic of a post than the original post content.

Dataset	C-L	C	M	C+M
General	0.783	0.838	0.858	0.902
Music	0.663	0.694	0.780	0.803

Table 2. Micro-averaged F-scores achieved by classifier on each dataset

Detailed results for **General** are shown in Table 3, arranged in order of descending F-score for **C+M**. It can be seen that there is a large variation in classification performance across different forums. For example, classification of a post in the Musicians forum is trivial, since almost all posts feature a link to MySpace. However classification of a Television forum post is more challenging, because this forum can also cover any topic which is televised. Despite the variation between topics, **C+M** always results in the best performance, although not all of these results are statistically significant.

4 Conclusion

We investigated the potential of using metadata from external objects for classifying the topic of posts in online forums. To perform this study, we augmented a message board corpus with metadata associated with hyperlinks from posts. Our experiments reveal that this external metadata has better descriptive power for topic classification than the original posts, and that a combination of both gives best results. We conclude that for those posts that contain hyperlinks for which structured data is available, the external metadata can provide valuable features for topic classification. We plan to continue this work by comparing the usefulness of object titles, descriptions, categories and tags as features for improving topic

Forum	C-L	C	M	C+M
Musicians	0.948	0.976	0.901	0.980
Photography	0.777	0.918	0.895	0.948
Soccer	0.812	0.831	0.909	0.949
Martial Arts	0.775	0.810	0.877	0.914
Motors	0.751	0.785	0.865	0.907
Films	0.744	0.831	0.844	0.880
Politics	0.786	0.801	0.809	0.844
Poker	0.662	0.739	0.794	0.838
Atheism	0.800	0.824	0.771	0.829
Television	0.595	0.634	0.704	0.736
Macro-Averaged	0.765	0.815	0.837	0.883

Table 3. F-score achieved by classifier on each **General** forum

classification. Thus not only textual metadata content will be considered, but also the relationships linking them to the original object. The growing amount of structured data on the Web means that this type of semantically-rich information can be retrieved from a significant and increasing proportion of hyperlinks. Potential applications of the approach include suggesting appropriate forums for new posts, and categorising existing posts for improved browsing and navigation. The enhanced representation of a post as *content plus hyperlink metadata* also has potential for improving search in social media.

References

1. Agichtein, E., Castillo, C., Donato, D., Gionis, A., Mishne, G.: Finding high-quality content in social media. In: Proceedings of WSDM, 2008
2. Angelova, R., Weikum, G.: Graph-based text classification: Learn from your neighbors. In: Proceedings of SIGIR, 2006
3. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. In: Proceedings of ISWC, 2007
4. Berendt, B., Hanser, C.: Tags are not metadata, but "just more content"–to some people. In: Proceedings of ICWSM, 2007
5. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The story so far. International Journal on Semantic Web and Information Systems 5(3) (2009)
6. Figueiredo, F., Belém, F., Pinto, H., Almeida, J., Gonçalves, M., Fernandes, D., Moura, E., Cristo, M.: Evidence of quality of textual features on the Web 2.0. In: Proceedings of CIKM, 2009
7. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The WEKA data mining software: An update. ACM SIGKDD Exp. 11(1) (2009)
8. Kinsella, S., Passant, A., Breslin, J.G.: Using hyperlinks to enrich message board content with Linked Data. In: Proceedings of I-SEMANTICS, 2010
9. Qi, X., Davison, B.: Classifiers without borders: Incorporating fielded text from neighboring web pages. In: Proceedings of SIGIR, 2008
10. Sun, A., Suryanto, M., Liu, Y.: Blog classification using tags: An empirical study. In: Proceedings of ICADL, 2007