Exploiting Social Links for Event Identification in Social Media

Hila Becker*, Bai Xiao
Columbia University

Mor Naaman
Rutgers University

Luis Gravano
Columbia University

ABSTRACT

We explore the use of social links (e.g., comment and authorship connections) for identifying events and their associated documents (e.g., photos, videos) in social media sites. To understand the potential benefits of using social links for this task, we analyze a network of author comments associated with photographs in a large-scale Flickr data set. Our preliminary experiments, building on baselines that use social media document features without link information, suggest that social links can provide a useful indication of document similarity for event identification.

1. INTRODUCTION

Social media sites such as Facebook, YouTube, and Flickr host an ever-increasing amount of user content associated with real-world events (e.g., a presidential inauguration, a concert, a parade). This social media content has a rich associated "context," including user-provided annotations (e.g., title, tags) and automatically generated information (e.g., content creation time). As a distinctive characteristic, social media documents are interconnected via different types of *social links*. Some links might be explicit (e.g., one document might refer to another document), while some other links are indirect (e.g., the author of one document might have commented on another document). We conjecture that social media links can reveal meaningful document connections and hence be helpful for event identification.

While social media documents generally include information that is useful for identifying events, this information is far from uniform in quality and might often be misleading or ambiguous. Several related efforts have focused on extracting high-quality event information from social media sources by analyzing temporal and spatial tag patterns [2, 3], or examined tags with social links for event detection in social media [5]. We are interested in using different types of social links, in addition to the rich context features associated with social media documents, for event identification. By identifying events, and their associated social media documents, we can enable event browsing and search, to complement the search tools that Web search engines provide.

In previous work [1], we focused on learning social media document similarity metrics, and used them in conjunction with a scalable clustering algorithm; ideally each cluster corresponds to an event and includes the social media documents associated with the event. While this work significantly outperformed the appropriate baselines, it did not exploit the variety of social links available in social media sites. We expect social links to be useful in situations where we cannot determine if documents are similar based on their context features alone (e.g., often documents have missing location information). We therefore explore ways to judiciously leverage social links for event identification, to complement the similarity learning models identified in [1].

2. EXPLOITING SOCIAL LINKS

Links such as social network connections, comments, and shared group memberships provide important cues for document similarity in social media. To understand the potential benefits of using (inherently noisy) social links for our event identification task, we analyzed the network of author comments associated with photographs in (a 90,288-document subset of) a large-scale Flickr data set (see Section 3). Out of the distinct document authors in the data set, 45% commented on some other author's document. Interestingly, 44% of authors who made such comments did so purely within one event (i.e., these authors created documents for an event and only commented on documents for that event, not others). Furthermore, 80% of authors made more comments on documents inside events on which they have published content than on documents for other events in the data set. These exploratory statistics hint that social links (e.g., based on author-comment relationships) might help in identifying event content.

We consider different ways to incorporate social links into a document similarity metric. One way is to use different types of link-based similarities on context features (e.g., author) as features for a similarity metric learning model. These similarities may be binary indicators of the authors' social network connections, shared group memberships, and so forth. In isolation, these features are not very revealing, but combined with other similarity metrics (e.g., based on the documents' context features) they may prove helpful for capturing document similarity. However, incorporating social links into the similarity models is a challenging task: for ensemble-based models [1], clusterers using just link-based similarities will likely group together many documents relating to different events, and create a large number of singleton clusters where no links exist; for classification-based models, the true contribution of link-based similarity features might be difficult to capture, since social links are often sparse.

*Contact author: hila@cs.columbia.edu

Copyright is held by the author/owner(s).

This material is based upon work supported by a generous research award from Google and by the National Science Foundation under Grant IIS-0811038.

While social links between document pairs may be too weak to capture similarity, links between *clusters* of documents may be more revealing. We observed that when our clustering algorithm [1] incorrectly splits an event across multiple clusters, it is often due to insufficient similarity between the event's documents rather than due to a strong similarity to documents from other events. As a result, many "pure" clusters, where all documents in the cluster belong to the same event, are created. In a preliminary analysis performed over clusters created by applying our algorithm to (a 90,288-document subset of) the Flickr data set (Section 3), we found that 24% of the events were split across multiple clusters, and half of these were split into "pure" clusters exclusively. These "pure" clusters represent a simple case where strong evidence of social links between two clusters could help us detect that they belong to the same event.

3. PRELIMINARY EXPERIMENTS

We describe our preliminary experiments using author-comment links associated with Flickr photographs to improve the clustering results of a classification-based similarity model based on logistic regression (CLASS-LR in [1]). Our data set consists of 270,425 Flickr photographs, taken in 2006-2008. These photographs were manually annotated by users with event ids, corresponding to events from the Upcoming event database[1]. We order the photographs according to their upload time, and then divide them into three equal parts for training, validation, and testing. We learn the logistic regression similarity model on the training set and use this model to cluster the validation set. We develop our social link-based cluster merging strategies on the validation set and then report the results of the entire procedure (clustering and merging) on the test set.

To decide whether to merge any pair of clusters, we train a learning model using the comments associated with each document across "pure" cluster pairs. We consider a variety of link-based features, including the total number of comments between authors in the clusters, the number of mutual comments (i.e., author A_1 from cluster C_1 commented on the document of author A_2 in cluster C_2 and vice versa), and the percentage of shared comments out of all comments associated with each cluster in the pair.

In our learning scenario, we would like to avoid false positives: a false positive corresponds to merging clusters for different events and hence hurts the quality of the initial clustering solution on which we build. Therefore, we use a cost-sensitive classification approach, training a model that assigns the highest cost to false positive errors. We experimented with different classification models and cost values using the Weka toolkit [4], and found a Multilayer Perceptron to be the best performing model, keeping the number of false positives to just 425 for 397,386 cluster pairs.

Using this classifier, we can predict whether any pair of clusters should be merged. However, we have to address the special case where the classifier's predictions disagree. For example, consider clusters C_1, C_2, and C_3 where the classifier predicts that (C_1, C_2) and (C_1, C_3) should be merged, but (C_2, C_3) should not. We can either merge C_1, C_2, C_3 into a single cluster (*Merge-ALL*), or merge (C_1, C_2) and not C_3, where C_1 and C_2 appear earlier in the data set (*Merge-FCFS*). For the latter approach, we add a confidence threshold θ, to ensure that only high-probability merge predictions would be used (*Merge-FCFS-θ*). We experimented with different threshold settings on the validation set and used the conservative setting that yielded the best performance (θ=0.995) for experiments over the test set.

Model	B-Cubed	B-Precision	B-Recall
CLASS-LR	0.8155	**0.9144**	0.7359
Merge-ALL	0.6765	0.5486	**0.8820**
Merge-FCFS	0.7923	0.7951	0.7895
Merge-FCFS-θ	**0.8226**	0.8980	0.7590

Table 1: Clustering results for the baseline and alternative merging methods, over the Flickr test set.

To evaluate our approach, we used the clustering quality metrics discussed in [1], namely, *Normalized Mutual Information* (NMI) and *B-Cubed*. Although the results follow similar trends for both metrics, we report the performance in terms of B-Cubed, as it can be easily analyzed in terms of precision and recall. Table 1 shows the clustering quality over the test set using the original logistic regression similarity model and our alternative merging strategies. Not surprisingly, all merging strategies hurt the B-Cubed precision, which corresponds to the average proportion of items in every document's cluster that belong to the same event. Similarly, all merging strategies improve the B-Cubed recall, which reflects a decrease in the number of clusters that each event is spread across. However, the only strategy that creates a better balance between the two, measured by the combined B-Cubed score, is the least aggressive strategy, *Merge-FCFS-θ*. While the overall performance improvement offered by this merging strategy is modest, it serves as an initial indication that social links can be useful for event identification in social media. Further improvements may be obtained by considering additional types of social links, which we intend to explore in future work.

4. CONCLUSIONS AND FUTURE WORK

We described a preliminary, exploratory direction for leveraging information from social links to improve similarity learning models [1], used to identify events and their associated social media documents. Our initial statistics and experiments suggest that these links may be useful similarity cues, especially when context features of social media documents (e.g., title, tags, time, location) are not sufficient for inferring similarity. We intend to experiment with techniques that incorporate social links as features for the similarity learning models. In addition, we plan to experiment with other intra-site social links (e.g., explicit social network connections, shared social group memberships) and links across social media sites.

5. REFERENCES

[1] H. Becker, M. Naaman, and L. Gravano. Learning similarity metrics for event identification in social media. In *WSDM*, 2010.

[2] L. Chen and A. Roy. Event detection from Flickr data through wavelet-based spatial analysis. In *CIKM*, 2009.

[3] T. Rattenbury, N. Good, and M. Naaman. Towards automatic extraction of event and place semantics from Flickr tags. In *SIGIR*, 2007.

[4] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2nd edition, 2005.

[5] A. Zunjarwad, H. Sundaram, and L. Xie. Contextual wisdom: social relations and correlations for multimedia event annotation. In *MULTIMEDIA*, 2007.

[1] http://www.upcoming.org