Linking News and Tweets

Xiaojie Lin[1](✉), Ye Gu[1], Rui Zhang[1], and Ju Fan[2]

[1] Department of Computing and Information Systems, The University of Melbourne, Melbourne, Australia
{xiaojiel1,yeg1}@student.unimelb.edu.au, rui.zhang@unimelb.edu.au
[2] Renmin University of China, Beijing, China
fanju1984@gmail.com

Abstract. In recent years, the rise of social media such as Twitter has been changing the way people acquire information. Meanwhile, traditional information sources such as news articles are still irreplaceable. These have led to a new branch of study on understanding the relationship between news articles and social media posts and fusing information from these heterogeneous sources. In this paper, we present a system that is able to effectively and efficiently link news and relevant tweets. Specifically, given a news stream and a tweet stream, the system discovers tweets that are relevant to each news in the news stream.

1 Introduction

Nowadays, a real world event such as a traffic accident or a criminal activity not only is covered by news articles, but also stimulates ordinary people to post their comments on social media such as Twitter[1], Facebook[2] and Weibo[3].

The strong relationship between news and social media interests researchers. Many studies on analyzing these two types of information sources together have been carried out. For example, Yang et al. [4] use relevant social media posts to summarize and extract highlights from news articles; Minkyoung et al. [2] use the relationship to analyze the characteristics of different types of media and the diffusion pattern of news events.

These studies and applications require reliable system to link news articles and relevant social media posts. In this paper, we present such a system which effectively and efficiently links news and relevant tweets.

2 The Linking System

The structure of the system is shown in Fig. 1. The input of the system is a news stream and a tweet stream. When a news article is received, it will be first preprocessed (e.g. tokenization and stemming) and then stored in a buffer for D

[1] https://twitter.com/.
[2] https://www.facebook.com/.
[3] The most popular microblogging platform in China. https://weibo.com.

M.A. Cheema et al. (Eds.): ADC 2016, LNCS 9877, pp. 467–470, 2016.
DOI: 10.1007/978-3-319-46922-5_41

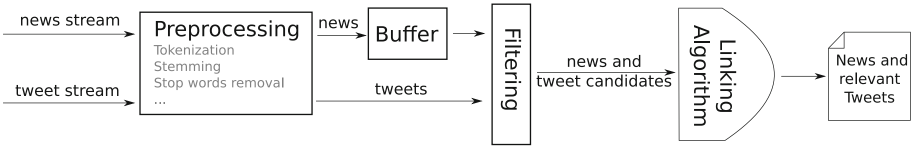


Fig. 1. System structure

days. Also, some indices will be created for the news in the buffer, which facilitate the following filtering and linking processes. When a tweet arrives, it will also be preprocessed, and then the system will use an efficient filtering algorithm (e.g. BM25 with a minimum threshold) to determine if the tweet should be added to the *tweet candidate set* of a news article in the buffer. When the filtering module has processed a certain amount of tweets, it will output a set of news along with their tweet candidates. The more expensive linking algorithm (discussed below) will now do the linking and output the final results — news and their relevant tweets.

We use an SVM classifier for the final linking. For each pair of a news and a tweet, a feature vector is extracted, and SVM will predict if the tweet is relevant to the news. The most important features we used are as follows:

BM25. BM25 computes a relevance score for a document and a query. In our case, we treat the news in the buffer as the document corpus and each tweet as a query.

Time. For a news and tweet published at time t_1 and t_2 respectively, the time feature is computed as $1/(t_2 - t_1 + 1)$. Note that we only consider tweet published after the news, so $t_2 - t_1 > 0$.

Named Entity. We extract named entities and calculate a TF-IDF score for each of them. The named entity feature is computed as:

$$\max_{n \in NE(a) \cap NE(t)} tfidf(n)\,,$$

where $NE(a)$ and $NE(t)$ is the named entities extracted from news a and tweet t respectively.

Event Phrase. We use a dependency parser[4] to extract relations and noun phrases from news and tweets. Collectively, we call them event phrases since they can describe the essence of an event. Examples of extracted event phrases is shown in Fig. 2. We train another SVM classifier to generate a confidence score for each event phrase. The score indicates how well the event phrase describes a news article. For a news and a tweet, the event phrase feature is calculated as: $\max_{e \in EP(t)} confidence(e, a)$, where $EP(t)$ is the set of event phrases extracted from tweet t and a is a news.

[4] http://www.cs.cmu.edu/~ark/TweetNLP/#tweeboparser_tweebank.

- Sigh. What a 9 year old's Uzi accident tells us about gun rights in America
- That time a 12-year-old girl shot dead her assailant. Worth remembering with 9-year-old Uzi accident...
- Child accidentally shoots and kills gun instructor
- A 9 yr old child fatally shoots gun instructor with an UZI..a Submachine Gun!! #NRA whats next?! Tossing live grenades

Fig. 2. Event phrases extracted from tweets relevant to the event of "A 9-year-old girl accidentally shoots and kills her gun instructor with an automatic Uzi".

3 Experiments

We use a dataset derived from Guo's dataset [1], which contains 12,704 news and 34,888 tweets. In the gold standard, a tweet and a news article are considered relevant if the tweet contains a URL pointing to the news article. URLs in the tweets are removed before conducting experiments.

Guo's dataset does not contain the full content of news articles. Also, most of the news articles do not have any relevant tweets. Therefore, we identify the news articles with no less than 20 relevant tweets and download the full contents. A small amount of news are also removed because of download or parsing errors. The final dataset contains 381 news with full contents and all the 34,888 tweets.

Some of the news in the dataset are about the same event, and they are very similar to each other. For example, the news "Scores Dead as Fire Sweeps Through Nightclub in Brazil" and "Hundreds killed in Brazil nightclub fire" are about the same accident. Therefore, we also conduct extra experiments on a clustered version of the dataset, which contains 240 news clusters.

We test a wide range of unsupervised approaches along with ours. The results are shown in Table 1. The unsupervised approaches include the model of Tsagkias et al. [3] which is based on the language model (LM), BM25 using news as document corpus (BM25-news), BM25 using tweets as document corpus (BM25-tweets), cosine similarity of TF-IDF word vectors and the WTMF-G model [1].

For the unsupervised approaches, 5-fold cross-validation is used to determine the cut-off thresholds which maximizes the F_1 score. Precision and recall are reported under the same threshold. For our supervised approaches, the same 5-fold cross-validation is used for training/testing.

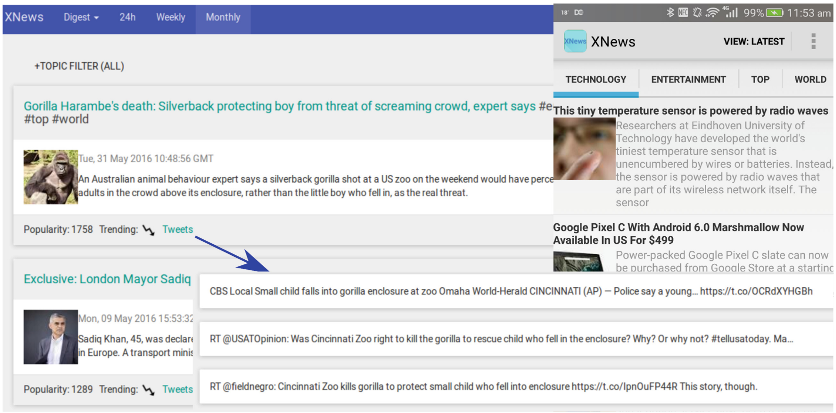
As shown in Table 1, our approach "SVM with event phrase features" performs the best in both the unclustered and clustered versions of the dataset. Note that Tsagkias's model (LM) does not work well in a binary classification setting because the relevance scores generated for different news are very different, so we are not able to find a reasonable cut-off threshold, and the reported metric values are very poor.

4 Demonstration

We build an online news service based on our system. After tweets are linked to news, we also use the relevant tweets to analyze the popularity and trending of each news. Our news service can be accessed via our website, Android client or REST API. Screenshots of the website and Android client are shown in Fig. 3.

Table 1. Performance of different approaches

Approaches	Unclustered			Clustered		
	Precision	Recall	F_1 Score	Precision	Recall	F_1 Score
LM	0.0016	0.0315	0.0030	0.0035	0.6641	0.0069
BM25 (news)	0.4635	0.5379	0.4979	0.6699	0.5658	0.6135
BM25 (tweets)	0.2210	0.3693	0.2765	0.4132	0.3757	0.3936
Cosine similarity	0.4810	0.4830	0.4820	0.6051	0.5539	0.5784
WTMF-G	0.5147	0.4797	0.4966	0.6897	0.5770	0.6284
SVM	0.6474	0.5634	0.6025	0.8183	0.5327	0.6453
SVM (event phrases)	0.6726	0.5544	**0.6078**	0.8145	0.5540	**0.6595**



Gorilla Harambe's death: Silverback protecting boy from threat of screaming crowd, expert says #e #top #world

Tue, 31 May 2016 10:48:56 GMT

An Australian animal behaviour expert says a silverback gorilla shot at a US zoo on the weekend would have perce adults in the crowd above its enclosure, rather than the little boy who fell in, as the real threat.

Popularity: 1758 Trending: Tweets

Exclusive: London Mayor Sadiq

Mon, 09 May 2016 15:53:32

Sadiq Khan, 45, was declare in Europe. A transport minis

Popularity: 1289 Trending: Tweets

CBS Local Small child falls into gorilla enclosure at zoo Omaha World-Herald CINCINNATI (AP) — Police say a young… https://t.co/OCRdXYHGBh

RT @USATOpinion: Was Cincinnati Zoo right to kill the gorilla to rescue child who fell in the enclosure? Why? Or why not? #tellusatoday. Ma…

RT @fieldnegro: Cincinnati Zoo kills gorilla to protect small child who fell into enclosure https://t.co/IpnOuFP44R This story, though.

This tiny temperature sensor is powered by radio waves

Researchers at Eindhoven University of Technology have developed the world's tiniest temperature sensor that is unencumbered by wires or batteries. Instead, the sensor is powered by radio waves that are part of its wireless network itself. The sensor

Google Pixel C With Android 6.0 Marshmallow Now Available In US For $499

Power-packed Google Pixel C slate can now be purchased from Google Store at a startinc

Fig. 3. Website and Android client

References

1. Guo, W., Li, H., Ji, H., Diab, M.T.: Linking tweets to news: a framework to enrich short text data in social media. In: ACL, vol. 1, pp. 239–249. Citeseer (2013)
2. Kim, M., Newth, D., Christen, P.: Trends of news diffusion in social media based on crowd phenomena. In: Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion, pp. 753–758. International World Wide Web Conferences Steering Committee (2014)
3. Tsagkias, M., de Rijke, M., Weerkamp, W.: Linking online news and social media. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, pp. 565–574. ACM (2011)
4. Yang, Z., Cai, K., Tang, J., Zhang, L., Su, Z., Li, J.: Social context summarization. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 255–264. ACM (2011)