Topy: Real-time Story Tracking via Social Tags

Gevorg Poghosyan ✉, M. Atif Qureshi, and Georgiana Ifrim

Insight Centre for Data Analytics, University College Dublin, Ireland
{gevorg.poghosyan,muhammad.atifqureshi,georgiana.ifrim}@
insight-centre.org

Abstract. The *Topy* system automates real-time story tracking by utilizing crowd-sourced tagging on social media platforms. Topy employs a state-of-the-art Twitter hashtag recommender to continuously annotate news articles with hashtags, a rich meta-data source that allows connecting articles under drastically different timelines than typical keyword based story tracking systems. Employing social tags for story tracking has the following advantages: (1) social annotation of news enables the detection of emerging concepts and topic drift in a story; (2) hashtags go beyond topics by grouping articles based on connected themes (e.g., #rip, #blacklivesmatter, #icantbreath); (3) hashtags link articles that focus on subplots of the same story (e.g., #palmyra, #isis, #refugeecrisis).

Keywords: story tracking, news, social media, social tags

1 Introduction

Although keyword and semantic-based matching of news have advanced considerably [1,4], the problem of automatically tracking story timelines and their evolution in real-time remains very challenging. A news story often discusses multiple related events, which take place in different time periods and may involve different entities. Some stories are relatively short-lived, for example, the 2016 Champions League final, and some others span many years and discuss multiple events, for example, the Ebola outbreak. For instance, the story of the Syrian war has evolved in time, by shifting the discussion **topic** (*Middle East*, *migration*, *human rights*, *politics*), the discussed **entities** (*Assad*, *ISIS*, *Putin*, *USA*, *Turkey*, *Belgium*) and the discussed **events** (*rebel uprising*, *destruction of Syria's chemical weapons*, *Yazidi massacres*, *camerawoman kicks a migrant*). Figure 1 illustrates this drift in the news article space projected on the topic-event-entity dimension. Stories may share articles, e.g., the article "Turkey carries out air strikes" may appear in several stories: *Syrian war*, *PKK in Syria*, *Turkey elections 2015*.

Topy takes a different approach to story tracking by building on crowd-sourced social tags and a hashtag recommender, to link news articles in complex story timelines. The choice of Twitter hashtags as rich social annotations is motivated by the following factors: (i) most stories have a lot of quality discussions centered on focused hashtags on Twitter, (ii) creation, popularity and abandonment of hashtags implicitly encode the concept drift in the story, (iii) hashtags allow cross-platform multi-modal content linking (text, image, video). This approach is also consistent with recent trends in news media: (a) *The Guardian*[1], *Huffington Post*[2], *AJ+, BBC*, write articles about popular

[1] www.theguardian.com/technology/hashtags
[2] www.huffingtonpost.com/news/hashtags/

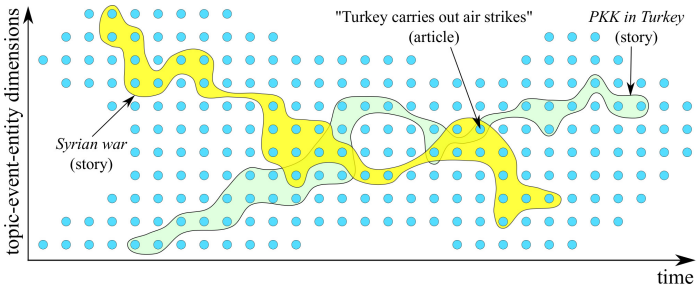

Fig. 1. Stories' drift in topic-event-entity-time space.

hashtags to inform and engage the public on discussion trends, (b) *The Sun* published a newspaper with a hashtag alongside an article to allow readers "to share their opinions and continue the story online"[3]. Most automated story tracking solutions are either limited in the number of events that can be tracked or are not real-time. Some news organizations have story-pages on their websites, i.e., curated collections of news articles that allow the reader to get an overview on particular events, e.g., referendums, elections, budgets. The Irish Times has dedicated story-pages for issues of relevance to the Irish society, e.g., the inquiry into the banking collapse[4] of 2008 (hashtag #bankinginquiry). Preparing these story-pages relies on prior agreement among journalists to manually tag all articles relevant to a set of stories, with the same set of tags. Once a decision is taken to create a story-page, those articles are continuously retrieved from the news archive via the manual tags. The problem with this approach is that it relies on foresight over which stories are worth covering and what is the right tag to use for those story-articles. Additionally, the manual process does not scale well on many stories that require in depth coverage, e.g., tools such as provided by www.newsdeeply.com although useful, update slowly and lack behind the fast pace of the real-world.

To the best of our knowledge there is no similar system that makes use of Twitter hashtags for story tracking. State-of-the-art systems rely on keyword/semantic matching and require often slow-to-change offline snapshots of knowledge bases [3] or need computationally expensive, complex clustering or semantic models, where parameters, such as number of topics [2], timespan of stories [1,4] and cluster sizes [5] significantly affect the system performance.

2 Topy System Overview

Topy maps news articles to stories in real-time by grouping articles with connected events, entities and topics that are discussed together on Twitter. Story tracking is formulated as a retrieval task with queries that allow mixing of keywords and hashtags. This allows tracking stories on-the-fly rather than being restricted to a pre-determined set of stories.

Real-time annotation of news articles with Twitter hashtags. We build on top of the Hashtagger infrastructure [6] for collecting, processing and storing news articles

[3] www.huffingtonpost.com/2014/03/26/sun-hashtag-newspaper-murdoch-british_n_5034639.html

[4] www.irishtimes.com/news/banking-inquiry

and Twitter data. An article is represented by its headline, subheadline, body, a set of summary keywords and a set of hashtags recommended to the article over a period of 24h from the article publication time [6]. Hashtagger is a recent hashtag recommendation method that achieves Precision of more than 85%. Around 70% of processed news articles have at least one recommended hashtag.

Query-based story tracking. The hashtags of each article are binned into 20 confidence bins with ranges from $(0.975, 1.0]$ to $(0.5, 0.525]$ and indexed as child documents for the corresponding article documents. Parent-child relationship and the chosen mapping enable an efficient search on article fields with different weighing using the BM25 algorithm. A query is composed of (i) words $w_1, ..., w_n$, which are matched on article keywords, headline, subheadline and content with score boost of correspondingly $\times 4$, $\times 3$, $\times 2$ and $\times 1$, and (ii) hashtags $\#h_1, ..., \#h_m$, which are searched on $k = 10$ hashtag confidence bins with score boosting of $6 - \frac{(i+1)\times 2}{20}$ for a match on bin $1 < i < k$. To get the articles covering a certain story, we do a two-step retrieval over a time period given by the user. We first expand the query in the hashtag space using the recommended hashtags of the top-10 articles from the initial search. This forms what we call the story tracking query. The second retrieval with the expanded query returns up to 1,000 articles ranked by their relevance, which are presented to the user. Note that the method works even for cases where there are no hashtags recommended to an article.

The Web user interface allows saving and curating stories over time. The *Topy* page provides a search box for queries and a time period menu, from 3 days to a year in the past. Once a query is issued, the user gets a ranked list of relevant articles for that story, together with a list of relevant hashtags. The user can like or remove articles or hashtags from the story. The *MyStories* page shows a dashboard for tracked stories. When loading a saved story, the retrieval is triggered and the updated list of relevant articles is returned. The user can issue the same query over different time periods, each will be saved as a different story in the user dashboard. The system can be seen in action at **http://ada.ucd.ie/tutorial_video/.**

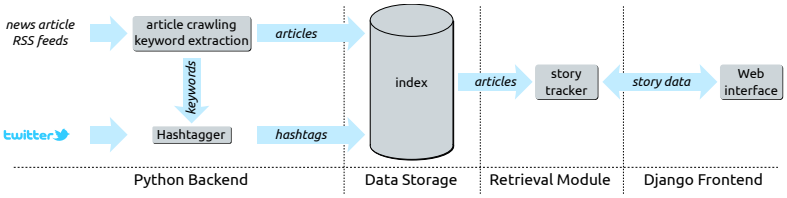

Fig. 2. System Architecture.

3 Topy Use Case

We collect 27 RSS news feeds from 8 news organizations, starting from August 2015. This allows us to track stories that have started capturing the public attention almost a year ago. One such case is the complex refugee crisis that has developed in connection to war conflicts worldwide, Syria in particular. We show here related use cases where the

user is interested in the story of "refugee crisis". The user issues the query "refugee crisis" with a time period of 1 year. This retrieves 19,881 ranked articles, grouped by news source. The retrieval also returns related hashtags: *#eu, #crisis, #refugee, #refugees*. For example, the article "Asylum seekers may receive funding for college" may not match any of the query terms but with Topy it is retrieved by matching the *#refugees* hashtag. The user is interested in searching for subplots of the story, hence issues a new query "#refugeeswelcome". The system retrieves 3,944 ranked articles with related hashtags: *#syria, #turkey, #refugeeswelcome, #refugeecrisis, #aylan, #syrianrefugees*. The user can observe the emphasis on the tragic death of Aylan Kurdi that triggered emphatic reactions from EU citizens towards Syrian refugees. Similarly, the query "#pegida" focuses on an opposite subplot of the refugee story, that emphasizes negative reactions towards refugees, among the related hashtags *#bachmann* is discovered, who is the founder of this movement. Each hashtag can be clicked to get tweets with that hashtag, e.g., *"#German far-right #Pegida founder #Bachmann guilty of race charge https://t.co/YFWYPlhLoP"*. The system can discover hashtags that may cause a topical drift. These can be manually removed to direct the story towards the user's preference[5].

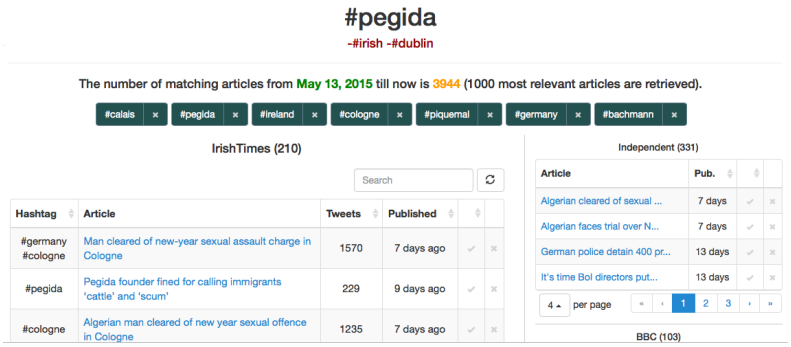

Fig. 3. System Screenshot.

References

1. J. Conrad and M. Bender. Semi-supervised events clustering in news retrieval. NewsIR, 2016.
2. L. Hou, J. Li, Z. Wang, J. Tang, P. Zhang, R. Yang, and Q. Zheng. Newsminer: Multifaceted news analysis for event search. *Knowledge-Based Systems*, 2015.
3. E. Kuzey and G. Weikum. Evin: Building a knowledge base of events. WWW, 2014.
4. G. Leban, B. Fortuna, and M. Grobelnik. Using news articles for real-time cross-lingual event detection and filtering. NewsIR, 2016.
5. B. Pouliquen, R. Steinberger, and O. Deguernel. Story tracking: linking similar news over time and across languages. In *MMIES*, 2008.
6. B. Shi, G. Ifrim, and N. Hurley. Learning-to-rank for real-time high-precision hashtag recommendation for streaming news. WWW, 2016.

[5] This work was funded by Science Foundation Ireland (SFI) under grant number 12/RC/2289.