Automatic Generation of Event Timelines from Social Data

Omar Alonso
Microsoft
omalonso@microsoft.com

Serge-Eric Tremblay
Microsoft
sergetr@microsoft.com

Fernando Diaz
Microsoft
fdiaz@microsoft.com

ABSTRACT

Over the past few years, social media has seen phenomenal growth and has become a very important source for getting real time updates from different parts of the world. While the notion of a trend usually reflects current events, the amount of information accumulated over a period of time can be used to provide another perspective for such events in the form of a timeline. In this paper, we present a technique that uses social information as relevance surrogates to generate an informative timeline. A core component is a variation of pseudo relevance feedback that is automatically generated using social data without external evidence. Finally, we describe the implementation of such technique and present evaluation results using a real-world data set.

KEYWORDS

Timelines, social pseudo relevance feedback, social media, Twitter

1 INTRODUCTION

In corpora with temporal information such as news, email, or social media, timelines provide a convenient interface for information needs that can be decomposed into events (e.g., news topics). Usually timeline generation consists of arranging words [8], document titles [6], extracted sentences [2, 4], or social media posts [7] into a sparse linear order.

The problem of timeline generation can be decomposed into *relevance estimation* and *deduplication*. Relevance estimation refers to determining whether a word, title, sentence, or post is relevant to the user's information need. Deduplication refers to collapsing relevant items that refer to the same topic. For example, for the query {`paris attacks`} over, say, a social media corpus, a system would first have to reason about which posts are relevant to the query (e.g., those posts referring to the attack itself, the government response, and so forth) and then about which relevant posts refer to the same event (e.g., group all of the posts related to attack itself, all of those related to the response, and so forth). If a system has good models for each of these, then a set of relevant, deduplicated posts define a timeline. A timeline for an event is, then, a document that contains relevant entries (e.g., sentences, titles, posts) arranged in a temporal order.

In this work, we focus on the relevance estimation problem in the context of social media. Specifically, we hypothesize that social media signals such as retweets (RTs), links, likes, or shares can be effectively used to better estimate post relevance. We take advantage of all these behavioral data in conjunction with relevant tweets that contain links that point to web pages with good titles from trusted domains as the basis for constructing a timeline. To this end, we present an algorithm, based on pseudo relevance feedback demonstrating that this is possible.

2 RELATED WORK

Timeline generation is a subarea of multidocument summarization focused on temporal information. The majority of work in the community has studied news articles. Russell and Jensen present a system for generating a timeline with word clusters as items on the timeline [8]. Jones and Diaz use a language modeling approach to detect relevant document titles from an initial retrieval [6]. Qi *et al.* describe a machine learning approach for selecting sentences from a pool of sentence-segmented documents [4]. This work precipitated the TREC Temporal Summarization track [2]. A review of approaches used there is beyond the scope of this paper. Compared to the tweet timeline generation task, our technique outputs link titles extracted from tweets instead of raw tweets.

Our work focuses on the exploitable information in social media corpora and, as a result, prior work, while related conceptually, lacks precisely the signals we are interested in studying. Our research is most related to the generation of social media timelines [9]. Li and Cardie focus on generating timelines for a specific user's Twitter stream [7]. Our work diverges from this work by generating timelines from popular topics as opposed to a collection of user tweets. The timeline techniques that we propose use modifications to relevance feedback, a well-known technique in information retrieval. A detailed description of different strategies for implementing relevance feedback is presented by Harman [5]. Carpineto and Romano published a comprehensive survey of many automatic query expansion techniques [3].

3 TIMELINE GENERATION

The proposed method works as follows. We assume access to a stream of social media (i.e., the Twitter firehose or similar data feed) and the existence of a component that detects trending hashtags or entities over a period of time such as minutes or hours. Along with the trending data, it is possible to extract a contextual vector, that is, a representative set of n-grams for a hashtag for a given timestamp (i.e., date granule). This contextual vector provides a list of keywords as context for the hashtag.

As a first attempt to derive a timeline, we can select the link from the most voted tweet per granule as each link (and tweet) has an associated *counter* expressed by the number of RTs, shares, likes, or any other behavioral signal. However, this approach suffers from selecting non-relevant links as such counters can be exploited and can produce duplicate entries. We introduce a more promising

WebSci'17, June 25-28, 2017, Troy, NY, USA.
© 2017 Copyright held by the owner/author(s). Publication rights licensed to ACM. 978-1-4503-4896-6/17/06…$15.00
DOI: http://dx.doi.org/10.1145/3091478.3091519

strategy, based on pseudo relevance feedback called *social pseudo relevance feedback* that combines contextual vectors with a voting strategy for identifying relevant link titles extracted from tweets according to the topic. Information extracted from these link titles forms the base for the timeline generation.

3.1 Contextual Vectors

A contextual vector represents a ranked list of n-grams for a set of tweets related to a hashtag or entity. For producing the list of n-grams, we first aggregate all the tweets related to a particular hashtag over the time period of consideration using techniques introduced by Alonso *et al.* [1]. We ingest data from the Twitter firehose and perform the following steps:

- Pre-processing and data cleaning.
- Filtering by tweet quality score, English language, spam, and adult scores.
- Removal of duplicates and near-duplicates tweets using a combination of shingles and Jaccard similarity.
- Clustering short fragments for identifying clusters of tweets.
- Rank the n-grams by relevance.

As an example, based on a single day worth of tweets from January 8, 2015, the Charlie Hebdo terrorist attack was the most important trending topic for that day and `#charliehebdo` one of the most popular hashtags used by people to describe the event on Twitter. By extracting the most relevant n-grams for the hashtag, we construct the associated contextual vectors for that specific granule (8-Jan-2015): `#charliehebdo = (free speech, charlie hebdo, terrorist attack, terror attack, satirical magazine, sad day, paris attack)`.

While not a proper summary, the contextual vector for that specific calendar entry provides enough context for the hashtag. If we compute other contextual vectors per day for the entire duration of the event, we have a data source that contains the many terms and phrases used at specific time granules. This data can be very useful for expanding the original query and re-rank the link titles.

3.2 Social Pseudo Relevance Feedback

Pseudo relevance feedback is an effective mechanism for improving retrieval without any user interaction in contrast to relevance feedback that requires users to mark documents that are relevant or not to a given query.

Our proposed technique, social pseudo relevance feedback (sprf), combines user feedback with query expansion using the contextual vectors presented earlier. The contextual vectors are derived from user generated content (i.e., tweets) and we can think of those n-grams as explicit terms selected by users as votes in aggregate. Similar to pseudo relevance feedback, ranking those document links extracted from tweets by counters (i.e., behavioral data) indicates that the top-k links are relevant to a given hashtag on a specific timestamp. For query expansion, we use the contextual vector to re-rank the links based on how similar the links' titles are to the terms in the contextual vector. Because all these links have been tagged by users with a relevant and frequent hashtag, our hypothesis is that they belong to the same topic.

The technique works as follows. Given a set of tweets that share the same hashtag, we extract a set of document link titles $docs = \{d_1, \ldots, d_n\}$ from those tweets and compute the similarity $sim(q, d)$ between query q (hashtag) and document link title d. For query expansion, we compute the contextual vector $cv = \{t_1, \ldots, t_m\}$ for q (hashtag) and produce a query q' that includes terms t from cv. Given a timestamp, we compute a score that measures the similarity $sim(q', d)$ between the expanded query q' and document link title d multiplied by its counter d_c:

$$\text{score} = sim(q', d) * \log(d_c + 2)$$

where q' contains the original query q and the terms from the contextual vector cv, d is the document link title, d_c is the counter value c for d, and $sim(q', d)$ is implemented as $cos(q', d)$. A counter c here quantifies user engagement via RT, likes, or shares. Each document link title is then ranked by this new score that is computed for a given timestamp as its associated contextual vector and is only valid for that specific timestamp. For each timestamp (i.e., date), which we consider a marker on our timeline, we have a ranked list of link titles to choose from. There are similarities to local context analysis (LCA), a blind relevance feedback technique based on co-occurrence analysis between candidate expansion features and query terms described by Xu and Croft [10]. However, we rely on the contextual vector for candidate expansion terms instead of extracting them from the top-k list.

In summary, sprf takes the following steps to expand a query (hashtag in our case but can be any topic) per time unit on a collection of document links titles and re-ranks them as follows:

- Given a set of tweets, compute contextual vectors for frequent hashtag from tweets per granule. The default granule is date but it can be tuned to any specific time definition. If there is not enough tweet volume, there is no contextual vector for that date.
- Perform an initial retrieval on the set of documents for a given hashtag as query to get a top-ranked set of links extracted from tweets.
- Rewrite the original query with top-t terms from the associated contextual vector.
- Re-rank the retrieval of the link titles using the new score.

3.3 Algorithm

In contrast to other approaches that present a tweet as entry on a timeline, we rely on the link's article title that is included on a tweet as potential candidate for an entry in the timeline. The intuition is that links from trusted news-like domains (e.g., cnn.com, bbc.co.uk, etc.) are likely to be authoritative, contain a title that is well-written in English, and they are easy to read. By relying on a set of worldwide news domains, constructed by mining Twitter accounts and web domains, the timeline has the potential to be diverse.

Both contextual vectors and sprf are local operations because they process the data set given a specific time window (e.g., hours or a single day). The timeline generation is a global operation that needs to ingest data composed of many occurrences of a hashtag and associated contextual vectors over a much large time window (e.g., weeks or months). The main algorithm takes as input a set of timestamped hashtags, each hashtag has a contextual vector and a set of document link titles and outputs a single document that contains a temporal order of link titles that are relevant to a topic

```
Data: Timestamped hashtag $ht$, $docs = \{d_1, \ldots, d_m\}$ and
      contextual vector $cv_t = \{t_1, \ldots, t_m\}$
Result: Timeline in chronological order for $ht$
timeline[] = ∅ ; q = ht ;
foreach timestamp in ht do
    ds[] = ∅ ;
    q' = q + $cv_{timestamp}$ ;
    foreach i in docs do
        score = $\cos(q', d_i) * log(d_i c + 2)$ ;
        add (i, score) to ds ;
    end
    /* every day there is a ranked list of article titles */
    rank(ds) by score ;
    j = 0;
    / * adds top article from the ranked list to timeline. If
      similar article exists, then pick next article */
    if ds[j] not in timeline then
        add $ds[j]$ to timeline;
    else
        repeat
            j++;
            select $ds[j]$ ;
        until $ds[j]$ is not in timeline;
        add $ds[j]$ to timeline;
    end
end
foreach e in timeline do
    print timeline[e] ;
end
```

Algorithm 1: Timeline generation algorithm.

(i.e., hashtag). Pseudo code is presented in the Algorithm 1 listing. The algorithms for computing contextual vectors, sprf, and the final timeline were implemented in the SCOPE language and runs daily over a large distributed computing cluster.

4 RESULTS AND EVALUATION

We now examine the generated timelines for four different events during 2015. All examples show the timelines verbatim and, due to space limitations, a condensed version is presented. The Paris attack example, #parisattacks, presented in Table 1, is very specific as the global event unfolds and condenses the main points. The Republican debate, #gopdebate, presented in Table 2, was part of a long series of debates and shows a preview of one of them in time. Table 3 presents #deflategate, a niche topic but very popular nevertheless and describes correctly the progress of the NFL football tampering scandal and the actors involved. Finally, Table 4 shows the year-long Formula 1 racing calendar, #f1, describing race results as well as intermediate information between grand prixes.

In all cases we can observe gaps in the timelines. In the Paris example, there is no entry for 11/24/2015 which indicates that there was not enough new content discovered by the algorithm, potential removal of duplicate content, or that the associated hashtag did not contain enough data volume to be considered. The NFL example shows more gaps as the topic covers many months and there are days with no activity for such hashtag. Similarly, the F1 example has expected gaps between races. The Republican debate timeline has lesser gaps as many debates were part of the primaries.

As described in Section 2, the TREC TGG data set is a sample (we use the full firehose) and lacks the behavioral signals that we use for our algorithm and therefore is not suitable for comparison. We also note that our timelines are based on link information extracted from tweets and not raw tweets, making it more about informativeness rather than specific data points. In other words, we do not show tweets in the timeline. Instead, our timeline contains the most relevant document link titles extracted from a set of tweets.

We now describe the sprf technique offline evaluation using human judges. We produced timelines for 455 popular hashtags generated over a period of 3 months from October 1 to December 31, 2015. The task consisted on a pair-wise preference selection where each judge has to select which timeline was more informative. We use three timeline generation algorithms: baseline (rank by counter popularity), dedup (rank by counter popularity with no duplicate elimination), and sprf. For the A-B comparison we randomize the order of the algorithms. The task template is presented in Figure 1. We use an internal crowdsourcing platform for gathering labels. Each comparison was assessed by 5 judges and we took majority vote as the final answer.

> *In this task, you will be presented with a Twitter hashtag representing a specific event or topic. There are two timelines summarizing the sequence of important sub-events related to the hashtag. An effective timeline contains only relevant sub-events and avoids duplicate content. The timelines are not complete and can have missing information. Your task is to select which of the two timelines provides a better summary of the important sub-events.*
>
> Hashtag: "#thehashtag"
> [Timeline A] [Timeline B]
>
> *Which of the following timelines is more informative?*
>
> [] A is better [] B is better [] They are the same

Figure 1: A-B relevance comparison task template.

We evaluate our method by examining how much the timeline can be reduced in size with duplicate elimination and how good the output is to a wider audience. We present the results of the timeline size reduction in Table 5. As mentioned early on, the popularity voting scheme (baseline) suffers from including duplicate content on the final result because the same link can be popular over more than one day. By adding exact duplicate removal in dedup (baseline vs. dedup), we can reduce the size of the timeline by 8% on average. If we then compare against sprf (baseline vs. sprf), the final size is cut in half. This is expected as baseline has a problem with potential non relevant links which are removed with sprf.

The relevance label distribution is presented in Table 6. In general, regardless of the type of hashtag, sprf has the advantage over

#parisattacks	
11/13/2015	Paris ’shooting’: Several casualties after gunman opens fire
11/14/2015	Paris terror attacks: eight attackers dead after killing at least 120 people – live updates
11/15/2015	Why Syrian refugee passport found at Paris attack scene must be treated with caution
11/16/2015	Got a French flag on your Facebook profile picture? Congratulations on your corporate white supremacy
11/17/2015	I was held hostage by Isis. They fear our unity more than our airstrikes
11/18/2015	Diesel named as police dog killed in Saint Denis raids hunting for Paris attackers
11/19/2015	Incredible moment woman survives as Paris gunman tries to shoot
11/20/2015	Paris attacks: ’I will not give you the gift of hating you’
11/21/2015	Sir Richard Branson: Blaming all Muslims for Paris attacks like ‘blaming all Americans for past actions of Ku Klux Klan’
11/22/2015	Paris is being used to justify agendas that had nothing to do with the attack
11/23/2015	Frankie Boyle on the fallout from Paris: ‘This is the worst time for society to go on psychopathic autopilot’
11/25/2015	Glenn Greenwald: Why the CIA is smearing Edward Snowden after the Paris attacks
11/26/2015	David Cameron to make case for Syria air strikes
11/27/2015	Paris attacks: France holds service two weeks after massacre

Table 1: Unedited output example for #parisattacks, a major global event in November 2015.

#gopdebate	
8/04/2015	Rick Perry and Rick Santorum Left Off the Republican Debate Stage
8/06/2015	Carly Fiorina Knocks Donald Trump: ’I Didn’t Get a Phone Call from Bill Clinton’
8/07/2015	Rubio, Bush Differ on Common Core Standards
8/08/2015	Donald Trump criticises Fox debate moderator Megyn Kelly
8/10/2015	Backing Christie Is Good Business for Bridge-and-Tunnel Lawyers
8/11/2015	Trump Leads Iowa GOP Field, Rubio Seen as Debate Winner
8/12/2015	Megyn Kelly not apologising for debate comments
8/13/2015	Roger Ailes to Donald Trump: ’We resolve this now...or go to war’
9/17/2015	Donald Trump bankruptcy: Everything you want to know
9/18/2015	No One Performed Better Than Carly Fiorina in Second Republican Debate
9/19/2015	Fiorina defends citing nonexistent abortion video
9/20/2015	Trump fails to say how he’d make America great again
9/21/2015	U.S. Should Use Offensive Cyber Tactics With China: Bush

Table 2: Unedited output example for #gopdebate, a preview of the many Republican debates during the US elections.

#deflategate	
5/06/2015	Patriots ‘probably deflated balls’
5/07/2015	NFL Finds Patriots Employees Likely Deflated Balls
5/08/2015	Cheating Works: NFL tolerates Tom Brady’s ‘Deflategate’
5/11/2015	Tom Brady suspension seems likely
5/12/2015	Tom Brady set to appeal four-game ban as New England Patriots plot strategy
5/15/2015	NFL ban Super Bowl MVP Tom Brady for first four games
7/28/2015	Deflate-gate: NFL upholds 4-game suspension of Tom Brady
7/29/2015	LIVE VIDEO: Coach Bill Belichick speaks out on the NFL’s ‘Deflategate’ decision
8/12/2015	NFL’s Brady and Goodell Squeezed by Judge in Settlement Play
9/03/2015	Patriots’ Tom Brady Faces NFL ‘Deflategate’ Appeal
9/04/2015	Tom Brady case: Judge overturns ‘deflategate’ suspension

Table 3: Unedited output example for #deflategate, an American football related event that spans several months.

dedup. By manually classifying some of those hashtags as “political”, given world events around that time that included unexpected events that have a political connotation, the performance of sprf improves more. This is the kind of scenarios that we believe our timeline method provides the most value: popular events that are heavily discussed in Twitter over longer periods of time. If we look at sports related hashtags, the numbers are very comparable. For games, that usually span a couple of hours, there is not enough differentiation at the day granule. That is, the end result of the game prevails as the most relevant content. A different story is for longer events like a championship where several games are played and a longer timeline is more useful to show the progress of the competition.

For this experiment, we marked a total of 29 unique hashtags under the political category such as #parisattacks, #iran, #isis, #syria, #sanbernardino, #gopdebate, #blacklivesmatter, and

#f1	
7/06/2015	Hamilton & Moss take on Monza
8/19/2015	Nico Hulkenberg partners Sebastian Vettel for 2015 Race of Champions
8/23/2015	Belgian Grand Prix
9/03/2015	Sebastian Vettel and Kimi Raikkonen in harmony over close relationship
9/05/2015	Lewis Hamilton shocked by McLaren's fall from grace
9/06/2015	McLaren reserve Kevin Magnussen says he needs to race in F1 next season
9/07/2015	Red Bull exploring all options for 2016 engine deal
9/16/2015	Renault hint that their engine deal with Red Bull will end this season
9/17/2015	Romain Grosjean reveals he has 'made his decision' on 2016 future
9/18/2015	Volkswagen close to buying Red Bull F1 team
9/19/2015	Red Bull will quit F1 if they don't get a competitive engine in 2016
9/20/2015	Sebastian Vettel starts to dream the 'impossible' after Singapore win
9/21/2015	Daniel Ricciardo believes track invader cost him a shot at winning the Singapore Grand Prix
9/22/2015	Singapore GP track invader charged by authorities
9/23/2015	Force India confirm that Sergio Perez is staying put next season
9/26/2015	Daniil Kvyat crashes in Japanese GP qualifying
9/27/2015	Niki Lauda says Red Bull never came back to confirm engine deal

Table 4: Unedited output example for #f1, a year long racing sports event for 2015.

Timeline comparison	Size reduction %
baseline vs. dedup	8%
baseline vs. sprf	50.5%
dedup vs. sprf	41.5%

Table 5: Timeline size reduction comparison.

Label	All	Political	Sports
baseline	15.5%	10.3%	27%
dedup	35.6%	32.8%	27%
sprf	37.2%	43.1%	27%
same	11.7%	13.8%	19%

Table 6: Relevance evaluation label distribution.

#refugees to name a few. For sports, we marked a total of 22 unique hashtags that describe the name of a club (#barca, #mcfc), the name of a sport (#soccer, #football), or a main competition like the Rugby World Cup (#rwc2015). Finally, timeline evaluation has a number of challenges as well. Judges need to be familiar with the topic and able to recognize a potential sequence of events when assessing. This can be difficult when timelines are longer and somewhat difficult to compare.

5 CONCLUSION AND FUTURE WORK

We presented an algorithm based on social pseudo relevance feedback that constructs a timeline by using only social data as input. The described method has been implemented in a real-world system using the Twitter firehose as input. We provided enough technical information so the technique can be implemented and the experiments and evaluations replicated. We conducted an offline evaluation that shows that our technique produce smaller size timelines with much more relevant content. A data analysis reveals that sprf performs well when the hashtag describes an event that covers multiple days, like political or world events. Sports hashtags have a more specific temporal pattern, mostly due to specific games, so sprf is comparable to other methods.

The overall results are encouraging and we plan to continue improving the techniques. We believe that timeline representations are information rich structures that can be used in many scenarios such as summarization, event descriptions that provide different views or perspectives, event graph annotation, and exploratory search. Future work includes better term ranking for contextual vectors, experimenting with different term weights, and improving the coherence of the timelines.

REFERENCES

[1] Omar Alonso, Sushma Bannur, Kartikay Khandelwal, and Shankar Kalyanaraman. 2015. The World Conversation: Web Page Metadata Generation From Social Sources. In *Proc. of WWW.* 385–395.

[2] Javed Aslam, Fernando Diaz, Matthew Ekstrand-Abueg, Virgi Pavlu, and Tetsuya Sakai. 2013. Overview of the TREC 2013 Temporal Summmarization Track. In *Proc. of TREC.*

[3] Claudio Carpineto and Giovanni Romano. 2012. A Survey of Automatic Query Expansion in Information Retrieval. *ACM Comput. Surv.* 44, 1 (2012), 1:1–1:50.

[4] Qi Guo, Fernando Diaz, and Elad Yom-Tov. 2013. Updating Users about Time Critical Events. In *Proc. of ECIR.* 483–494.

[5] Donna Harman. 1992. Relevance Feedback Revisited. In *Proc. of SIGIR.* 1–10.

[6] Rosie Jones and Fernando Diaz. 2007. Temporal profiles of queries. *ACM Trans. Inf. Syst.* 25, 3 (July 2007), 14.

[7] Jiwei Li and Claire Cardie. 2014. Timeline Generation: Tracking Individuals on Twitter. In *Proc. of WWW.* 643–652.

[8] Russell Swan and David Jensen. 2000. TimeMines: Constructing Timelines with Statistical Models of Word Usage. In *ACM SIGKDD Workshop on Text Mining.*

[9] Yulu Wang, Garrick Sherman, Jimmy Lin, and Miles Efron. 2015. Assessor Differences and User Preferences in Tweet Timeline Generation. In *Proc. of SIGIR.* 615–624.

[10] Jinxi Xu and W. Bruce Croft. 2000. Improving the effectiveness of information retrieval with local context analysis. *ACM Trans. Inf. Syst.* 18, 1 (2000), 79–112.