Online news tracking for ad-hoc queries

Jeroen B. P. Vuurens
The Hague University of Applied Science
Delft University of Technology
The Netherlands
j.b.p.vuurens@tudelft.nl

Arjen P. de Vries
CWI
Delft University of Technology
The Netherlands
arjen@acm.org

Roi Blanco
Yahoo Research Barcelona
Spain
roi@yahoo-inc.com

Peter Mika
Yahoo Research Barcelona
Spain
pmika@yahoo-inc.com

ABSTRACT

Following online news about a specific event can be a difficult task as new information is often scattered across web pages. In such cases, an up-to-date summary of the event would help to inform users and allow them to navigate to articles that are likely to contain relevant and novel details. We demonstrate an approach that is feasible for online tracking of news that is relevant to a user's ad-hoc query.

1. TRACKING EVOLVING NEWS

Internet users are replacing traditional media sources such as newspapers or television shows more frequently for online news. Although the Web offers a seemingly large and diverse set of information sources ranging from highly curated professional content to social media, in practice most sources base their stories on previously published works and add a much more limited set of new information. Therefore, users that seek additional information on a topic, often end up spending significant amount of effort re-reading the same parts of a story before finding relevant and novel information. In such cases, an up-to-date summary of the event would help to inform users and allow them to navigate to articles that are likely to contain relevant and novel details. Online summarization is a crucial aspect of real-world products such as online live streams for natural disasters, product launches, financial or political events, breaking news notifications on mobile devices and topical daily news summaries like the Yahoo! news digest (`https://mobile.yahoo.com/newsdigest`).

In this demonstration, we suggest an alternative to tracking the news by subscribing to hashtags on Twitter, or using Google Alerts. Compared to these existing system the user is presented with a summarization of the most important previously unseen facts along a timeline, which are topically related to a predefined ad-hoc query, and produces results that are less loquatious than Twitter and more insightful than the stream of headlines on Google Alerts. The contribution of this work is to demonstrate an approach that is feasible to tailor continuous news updates to an ad-hoc query.

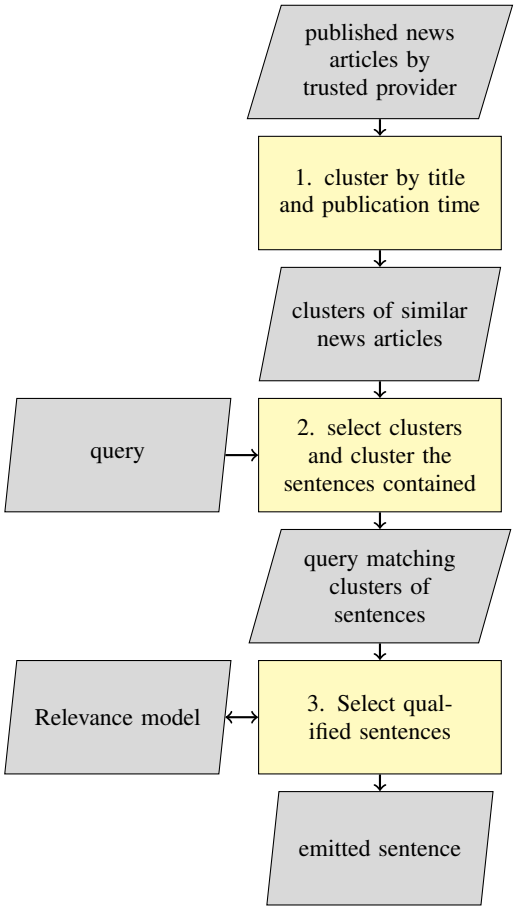

Figure 1: the process that used to extract sentences containing supported news facts from a stream of published news articles that are relevant to a given query.

2. EXISTING WORK

For query based extraction of events along a timeline, Chieu et al. suggested the use of temporal proximity to weigh the importance of terms at a given time [1]. They argue that high similarity between sentences is a possible indication that two sentences report the same event, which can be used in a clustering approach. However, naive clustering approaches require too many computations for large collections to use in an online application. Petrovic et al.

suggested to locate the most similar items with a high probability using Locality Sensitive Hashing (LSH), however, when comparing tweets they observed that more distant documents were not always allocated to same partition as their true nearest neighbor [3].

3. ONLINE NEWS TRACKING FOR AD-HOC REQUESTS

For online news tracking, we propose to use articles that are published on online news sites. The title of news articles can be viewed as a short summarization of its content, and therefore used to determine if articles are likely to describe the same topic. Given the relative low-memory requirements of news headlines, this allows for fast in memory clustering without the need to partition the data. The publication of most news articles can be monitored using RSS feeds, which allows fast access to the articles' title and publication time.

In Figure 1, we describe the process that is used to extract sentences that contain news facts from a stream of published news articles. In step 1, titles of newly published articles are continuously downloaded from RSS feeds. These titles are connected to their nearest neighbors based on similarity measure that considers their title and publication time. Salient sentences are found in sentence clusters of the nearest neighbor graph, when sentences from at least three different sources are most similar, indicating news facts that are more interesting rather than (opinion) information that is not supported. In step 2, per ad-hoc query a graph of sentences is created and maintained. The output of step 1 is monitored, and if a *query matching cluster of titles* is formed or modified, i.e. that contains a title that includes all query terms, then all sentences of the news articles in that cluster are added to the sentence graph of the ad-hoc query. Finally, in step 3, the *qualifying* sentences in the arriving news article are emitted to the user. For qualification, a sentence must (a) be among a cluster of at least three sentences from different news domains, (b) be ranked in the top-K of emitted sentences using a relevance model over the information seen in the last hour, and (c) add information previously not shown to the user.

4. FEASIBILITY

In this demo, we show that online news tracking can be done with reasonable latency in commodity machines.

Typically, in step 1 of Figure 1, the monitoring of RSS feeds and maintaining a nearest neighbor graph over news headlines takes less than 5% of the capacity of a standard computer. Step 2 and 3 of Figure 1 can be processed in parallel for different ad-hoc queries, and therefore scaled up in production systems. Additionally, efficiency can be improved when the news timelines for known entities (e.g. Wikipedia) are cached in advance, allowing steps 2 and 3 for entity related queries to be simplified to a filtering task over a cached timeline.

5. DEMONSTRATION

We provide three ways to participate in the demo. At the stand, we will show a summary of topics that are trending at that time. For these summaries, the information is processed as if online and therefore represent what the user would have received when they subscribed to the query at the first update. Additionally, participants can experience receiving new updates on the topics they wish to track, by subscribing to a live generated RSS feed for current trends. Finally, we will provide limited opportunity to enter ad-hoc queries, depending on the resources needed to downloading the articles for new topics.

In Table 1, we show an example of a time line constructed for the query "Copenhagen", after a terrorist attack on Februari 14th 2015. The first mention on Twitter of the hashtag #CopenhagenShooting was at 17:11 (`http://ctrlq.org/first/`), and the first information was added to Wikipedia at 17:06. A larger static example of the demo results can be viewed online, at `http://media.cwi.nl/newstracker/`.

Table 1: Timeline constructed for the query "Copenhagen" from Feb 14 2015 16:19.

Time	Sentence
2015-02-14 16:19:58	Copenhagen - Shots were fired on Saturday near a meeting in the Danish capital of Copenhagen attended by controversial Swedish artist Lars Vilks, Sweden's TT news agency reported.
2015-02-14 16:49:22	COPENHAGEN, Denmark - At least one gunman opened fire Saturday on a Copenhagen cafe, killing one man in what authorities called a likely terror attack during a free speech event organized by an artist who had caricatured the Prophet Muhammad.
2015-02-14 17:34:26	COPENHAGEN, Denmark (AP) – A gunman fired on a cafe in Copenhagen as it hosted a free speech event Saturday, killing one man, Danish police said.
2015-02-14 18:51:04	After searching for the gunman for hours, police reported another shooting near a synagogue in downtown Copenhagen after midnight.
2015-02-14 19:29:41	One person was shot in the head and two police were wounded in an attack on the synagogue in central Copenhagen, Danish police said, adding that it was too early to say whether the incident was connected to an earlier one at an arts cafe.
2015-02-15 01:56:20	French President Francois Hollande called the Copenhagen shooting "deplorable" and said Thorning-Schmidt would have the "full solidarity of France in this trial."
2015-02-15 03:52:40	Denmark was on high alert and a massive manhunt was under way on Sunday after a man sprayed bullets at a Copenhagen cafe hosting a debate on freedom of speech and blasphemy, killing one person and wounding three police officers.

References

[1] H. L. Chieu and Y. K. Lee. Query based event extraction along a timeline. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 425–432. ACM, 2004.

[2] S. Petrović, M. Osborne, and V. Lavrenko. Streaming first story detection with application to twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 181–189. Association for Computational Linguistics, 2010.