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 news about a specific event can be a difficult task as new information is often scattered across web pages. 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.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Filtering

General Terms

Clustering, Multi-document summarization

1. TRACKING EVOLVING NEWS

Internet users are replacing traditional media sources such as newspapers or television shows more frequently by 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 for tracking the news via Twitter hashtag subscriptions or Google Alerts. Compared to these existing approaches, the user is presented with a summary that contains the most important previously unseen facts along a timeline, topically related to a predefined ad-hoc query. The result is a timeline that is less redundant than Twitter and more insightful than the stream of headlines on Google Alerts. Instead of reporting social media utterances, we propose to give updates directly from the journalists' writings, by selecting relevant sentences with novel information from the news articles themselves.

The main contribution of this work is to demonstrate an approach that is feasible to tailor continuous news updates to ad-hoc queries. The technical contribution is to apply a multi-step three-nearest-neighbors clustering approach that can keep up with all the news arriving from hundreds of RSS feeds. For each of those feeds, we fetch the full articles from the source, and process their contents to identify the most novel and relevant sentences. The resulting system participates in the TREC temporal summarization task, but the demonstration gives more interpretable and comparable results than the TREC evaluation measures indicate. Also, users of the demo can track their own information needs.

2. EXISTING WORK

The system used can be viewed as a hybrid combination of techniques for query based online news tracking and summarization, adapted from [2, 1]. A side effect of broadcast news is that stories with near identical publication times are more likely to discuss related events [2], which is what we use in this study to cluster news. Our approach differs by a stronger emphasis on novelty of information emitted (like [3]). Hereto, we estimate the amount of previously unseen information to use only sentences that are likely to contain novel information.

3. TRACKING 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 this demo, we summarize a stream of online news articles in a three-step process: *cluster titles*, *cluster sentences* and *qualify sentences*. In the first step, 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

SIGIR'15, August 09-13, 2015, Santiago, Chile.
ACM 978-1-4503-3621-5/15/08.
DOI: http://dx.doi.org/10.1145/2766462.2767872 .

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, the monitoring of RSS feeds and maintaining a nearest neighbor graph over news headlines takes less than 10% of the capacity of a standard computer. The clustering and qualifying of sentences is processed independently for different ad-hoc queries and therefore can be processed in parallel and 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 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://newstracker.github.io/`.

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.

Acknowledgment

This demonstration has been developed with support of the Yahoo Faculty Research and Engagement Program and COMMIT/ project Infiniti, and was carried out on the Dutch national e-infrastructure with the support of SURF Foundation.

References

[1] J. Allan, R. Gupta, and V. Khandelwal. Temporal summaries of new topics. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 10–18. ACM, 2001.

[2] J. Allan, R. Papka, and V. Lavrenko. On-line new event detection and tracking. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 37–45. ACM, 1998.

[3] E. Gabrilovich, S. Dumais, and E. Horvitz. Newsjunkie: providing personalized newsfeeds via analysis of information novelty. In *Proceedings of the 13th international conference on World Wide Web*, pages 482–490. ACM, 2004.