First Story Detection using Multiple Nearest Neighbors

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
Radboud University Nijmegen
Institute for Computing and Information Sciences, Nijmegen, The Netherlands
arjen@acm.org

ABSTRACT

First Story Detection (FSD) systems aim to identify those news articles that discuss an event that was not reported before. Recent work on FSD has focussed almost exclusively on efficiently detecting documents that are dissimilar from their nearest neighbor. We propose a novel FSD approach that is more effective, by adapting a recently proposed method for news summarization based on 3-nearest neighbor clustering. We show that this approach is more effective than a baseline that uses dissimilarity of an individual document from its nearest neighbor.

1. INTRODUCTION

Internet users are turning more frequently to online news as a replacement for traditional media sources such as newspapers or television. For the user, the news stream is a source to both track topics of interest and to become informed about important new events the user was not yet aware of. Automated detection of new events can save to user a great deal of time, for instance by notifying users about new events, which is especially interesting to users and organizations for whom the information is time-critical and who need to act on that information.

FSD systems aim to identify those news articles that discuss an event that was not reported before in earlier stories, without knowledge of what events will happen in the news [2]. Recently, FSD has been suggested as a useful tool to monitor the Twitter feed [7], and while previous work has addressed the efficiency that is required for this purpose, there has been little work on improving the effectiveness in over a decade [7, 8].

In this study, we propose a novel approach that is more effective that the widely used function proposed by Allen et al. that declares a story new if it is dissimilar to its nearest neighbor [1].

2. RELATED WORK

The task of detecting events can be automated using information about the events published online. For this purpose, the Topic Detection and Tracking (TDT) program was initiated to discuss applications and techniques to organize broadcast news stories by the real world events that they discuss in real-time. News stories are gathered from several sources in parallel to create a single stream of constantly arriving news. The problem of first story detection is to identify the stories in a stream of news that contain discussion of a new topic, i.e. whose event has not been previously reported [6].

FSD has been recognized as the most difficult task in the research area of TDT [11]. In early work, Allen et al. detect first stories as news articles whose cosine similarity over tf-idf vectors to its nearest neighbor is less than a threshold, an effective approach that outperforms complex language model approaches in most cases [1]. This baseline is still used for FSD in recent work, in which more focus is put on efficiency than to improve effectiveness [3, 5, 4].

Papka and Allen, argue that a side-effect of the timely nature of broadcast news is that stories closer together on the news stream are more likely to discuss related topics than stories farther apart on the stream. When a significant new event occurs, there are usually several stories per day discussing it; over time, coverage of old events is displaced by more recent events. They use temporal proximity as a distinguishing feature to incorporate the salient properties of broadcast news [2, 6].

In recent work, Vuurens et al. proposed a novel 3-nearest neighbor clustering (3NN) approach to retrieve sentences from news articles that contain novel and useful news facts. In this approach every text is linked to its three nearest neighbors that must be from a different domain [10]. The so-called '2-degenerate cores' constructed by the algorithm correspond to highly similar texts from three different sources. Their existence indicates the importance or salience of the information contained. Temporal proximity is incorporated in the model by weighting the time between news articles in the similarity function used. In [9] normalized information gain is shown to be more effective than cosine similarity for the task of clustering news articles that are topically related.

3. METHOD

In this work, we adapt the 3NN clustering approach to First Story Detection, by clustering news articles rather than sentences, and using a similarity function based on normalized information gain to promote the clustering of news articles that are likely to be topically related.

3.1 Single Linkage

We compare our efforts to the approach described by Allen et al. [1], which is considered a state-of-the-art approach in recent studies on First Story Detection, e.g. [7, 4]. In this approach, documents are represented as tf-idf weighted vectors, and the novelty of a document d is estimated by the cosine similarity to its nearest neighbor n in the collection C [1]:

SIGIR '16, July 17 - 21, 2016, Pisa, Italy

ISBN 978-1-4503-4069-4/16/07...$15.00

DOI: http://dx.doi.org/10.1145/2911451.2914761

$$novelty(d) = 1 - \max_{n \in C} cos(d,n) \quad (1)$$

Then, a news article is marked as a first story when its novelty is below a threshold $\alpha \in [0,1]$.

3.2 3NN First Story Detection

In this study, we propose a novel approach that is based on 3-nearest neighbor clustering (3NN), using the existing open source implementation [10]. In 3NN clustering, every node is assigned to its three nearest neighbors, not allowing links between nodes from the same news domain, and based on temporal proximity between publication dates which allows the clustering to be continuously updated in near real-time. 2-generate cluster cores are formed when three nodes each link to the other two as a one of its 3 nearest neighbors. These clusters contain information that is locally most central and therefore likely to be salient information [10]. The key idea for First Story Detection, is that acting on formed 3NN clusters rather than individual documents is less likely to return false positives. However, instead of truly detecting the first story as was the objective in the TDT program, here we aim to improve detection performance at the expense of slightly delayed detection. It may also be that the story detected as the first of a new event is more central to the information, and therefore more suitable as a seed to start tracking a topic, however, this hypothesis is outside the scope of this study and left for future work.

In [9], news sentences were fitted into a hierarchy that distinguishes between different events and topics by forming clusters of topically related the news articles, for which normalized information gain was shown to be more effective than cosine similarity. Therefore, to promote 3NN clusters to be formed around topically related news articles we use a similarity function based on normalized information gain. In Equation 2, the normalized information gain between two documents d and d' results in a score of 0 between identical documents and a score of 1 between disjoint documents, by dividing the information gain IG between the documents by an upper bound of the information gain IG_{max} that would be obtained if these documents have the same internal distributions over terms but are completely disjoint. For the remainder of this paper we use IG_{sim} as defined in Equation 3 as a similarity function between two documents d, d' based on IG_{norm}.

$$IG_{norm}(d,d') = \frac{IG(d,d')}{IG_{max}(d,d')} \quad (2)$$

$$IG_{sim}(d,d') = 1 - IG_{norm}(d,d') \quad (3)$$

From the obtained 3NN clustering, the newly formed 2-degenerate cores are inspected for first stories. Similar to the Single Linkage baseline, first stories are detected when a newly formed cluster core is dissimilar from news articles seen recently. In 3NN every news article is linked to its three nearest neighbors, therefore the members of a newly formed 2-degenerate core that contains a first story each have two links to the other core members and the third link links to a dissimilar news article. The most similar non-core news article that a core member links to, is then used to estimate the novelty of that cluster core. Formally, in Equation 4 a cluster core A is declared novel when the similarity between a news article $d \in A$ and a news article n in the remainder of the collection C is below a threshold $\phi_{novelty}$.

$$novelty(A) = \max_{d \in A, n \in C-A} IG_{sim}(d,n) < \phi_{novelty} \quad (4)$$

Lastly, we add a threshold to filter out newly formed clusters that are less likely to be topically related to each other. Vuurens et al. show that news articles that have a high normalized information gain are rarely topically related [9]. Following their findings, we filter out clusters that fail the coherence criterium in Equation 5, that enforces that the similarity between all nodes d, d' that are members of the same 2-degenerate core A exceeds a threshold $\phi_{coherence}$, for which different settings are tried to examine the sensitivity and impact on effectiveness.

$$coherence(A) = \min_{d \in A, d' \in A - \{d\}} IGsim(d,d') > \phi coherence \quad (5)$$

3.3 Test set

For the evaluation, we use the TREC Temporal Summarization test sets of 2013 and 2014. The corpus for these test sets is the 2013 TREC KBA Streaming corpus, which contains approx. 150M news articles that are processed in a strict online setting. Table 1 shows the topics from the test sets, which are all types of a crisis that received continuous updates in the media over time. Arguably, the news regarding a single topic could be considered to be all part of the same story, or in some cases be regarded as separate stories within a topic. Here we regard all news articles that are matched to the same topic as part of one news story, for which ideally only the first article should be returned. TREC assessors annotated the sentences that TREC participants retrieved as relevant if they contain a news fact relevant to the topic.

The basis for the evaluation of the FSD systems is a list per topic of all documents that contain relevant news facts according to the TREC ground truth or the online published extended lists that contain duplicate sentences found in the collection. For the combined 23 topics, there are 65,358 documents that were annotated as containing relevant information. For this task, a returned news article is considered as a first for a topic when it is the first relevant article returned by the system, and a false alarm when another relevant article for the same topic was returned earlier. News articles that are not marked as relevant to the topic are ignored in the evaluation.

3.4 Experiment setup and evaluation metrics

The effectiveness of First Story Detection systems is measured by the miss rate, false alarm rate, recall and precision, which we explain using the contingencies in Table 2. For any topic, we only consider articles that are annotated as relevant for the topic, thus if T is the number of documents annotated as relevant for the topic, then $TP + FN + FP + TN = T$. Since there can only be one first story per topic per system, $TP + FN = 1$ and $FP + TN = T - 1$. A miss occurs when the system fails to detect a new event, i.e. $miss\ rate = \frac{FN}{TP+FN}$. A false alarm occurs when the system emits a news article when a first story was already emitted for that topic, i.e. $false\ alarm\ rate = \frac{FP}{FP+TN}$. Recall is the fraction of topics for which a first story was detected $Recall = \frac{TP}{TP+FN}$, and Precision is the fraction of retrieved news articles that is a fist story $Precision = \frac{TP}{TP+FP}$, which here only considers the news articles that are relevant to the topic.

Table 2: Contingency table for evaluation metrics

	Retrieved	Not retrieved
First story	TP	FN
Not first story	FP	TN

Table 1: Topics for the 2013 and 2014 TREC TS track

Topic	Title
1	2012 Buenos Aires Rail Disaster
2	2012 Pakistan garment factory fires
3	2012 Aurora shooting
4	Wisconsin Sikh temple shooting
5	Hurricane Isaac (2012)
6	Hurricane Sandy
8	Typhoon Bopha
9	2012 Guatemala earthquake
10	2012 Tel Aviv bus bombing
12	Early 2012 European cold wave
13	2013 Eastern Australia floods
14	Boston Marathon bombings
15	Port Said Stadium riot
16	2012 Afghanistan Quran burning protests
17	In Amenas hostage crisis
18	2011-13 Russian protests
19	2012 Romanian protests
20	2012-13 Egyptian protests
21	Chelyabinsk meteor
22	2013 Bulgarian protests against the Borisov cabinet
23	2013 Shahbag protests
24	February 2013 nor'easter
25	Christopher Dorner shootings and manhunt

4. RESULTS

In this Section, we compare the effectiveness of first story detection using Single Linkage (SL) to FSD using 3NN.

4.1 Effectiveness

In Figure 1, a DET curve shows the relationship between miss rate and false alarm rates. Overall, the 3NN runs perform better than SL, regardless of the setting used for $\phi_{coherence}$. In Figure 2, we show a tradeoff between recall and precision, which further supports that 3NN is consistently more effective than Single Linkage. Table 3 gives the precision and false alarm rate when the novelty thresholds for both systems are set to the highest precision that can be obtained at recall = 1. When $\alpha = 0.48$ and $\phi_{novelty} = 0.6$ are set to allow for the lowest false alarm rate at a missed rate of 0 (i.e. recall=1), precision is respectively 0.0149 for SL and 0.0618 for 3NN, meaning that SL more redundantly retrieves 4 times more news articles for the same event.

Table 3: Optimal effectiveness at recall=1.

	Precision	false alarm rate
Single Linkage $\alpha = 0.48$	0.0149	0.0195
3NN $\phi_{novelty} = 0.60$	0.0618	0.0053

4.2 Timeliness

In Figure 3, the y-axis shows the aggregated number of relevant news articles per hour over time on the x-axis. In this Figure, we can visually compare the moment a first story was detected against the volume of published news articles. We can see that the systems occasionally missed early detection, e.g. 3NN for topic 3, and Single Linkage for topic 9. On topic 12, detection may be late for 3NN, but there is a difficult tradeoff between early detection and a lower false alarm rate.

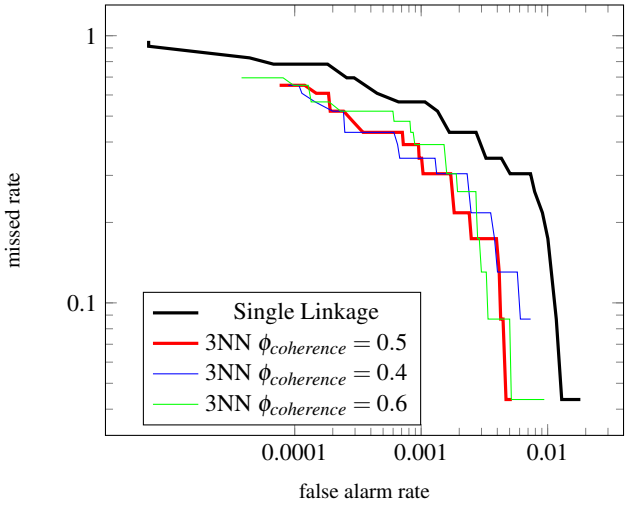

Figure 1: Detection Error Tradeoff curve, closer to the origin is better.

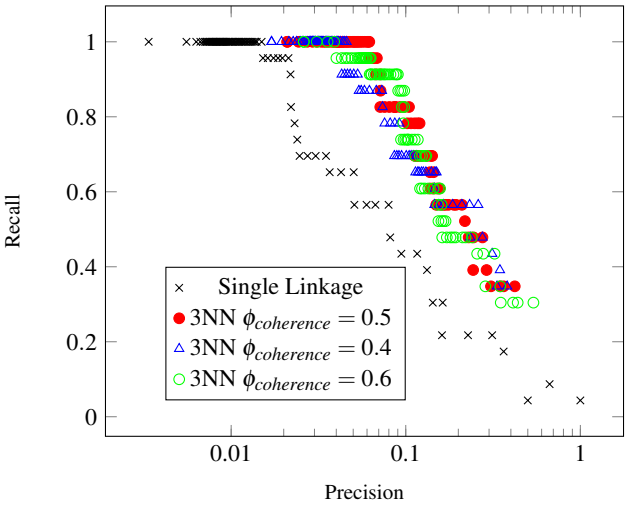

Figure 2: Plotted point show the Recall/Precision that correspond to the systems' effectiveness at the given threshold.

Some topics are related to an incident that is followed by a quick burst (e.g. topic 1), while other topics initially have a phase of little media attention and have intervals of increased interest later in time (e.g. topic 16). An interesting case is topic 18, which concerns the demonstrations that followed the Russian elections. For this topic, the news slowly shifted over the cause of days from a focus on the election itself to the steadily increasing demonstrations. This gradual shift towards a new topic is relatively difficult to detect for the approaches used in this study. The effective detection of these types of event may require a novel FSD approach that is not solely based on dissimilarity.

An inspection on the timeliness of the first stories detected reveals weaknesses in both approaches, and potentially an important aspect that should be taken into consideration in attempts to improve FSD. Timeliness of the detection is currently not addressed by the traditional evaluations that use a DET-curve and the tradeoff between recall and precision. To evaluate future work that ad-

dresses this issue, an additional metric to compare the timeliness of FSD approaches is required.

5. CONCLUSION

In this study, we propose a novel approach for the task of First Story Detection based on clustering news articles that are likely to be topically related, and estimating the novelty of newly formed clusters by comparison to previously seen news articles. We compared this approach to a baseline that estimates the novelty of a single news article by the cosine similarity to its nearest neighbor. The evaluation shows that the proposed model outperforms the existing baseline both in tradeoff between missed first stories and false positives, and in tradeoff between recall and precision. An analysis of the timeliness of the first story detections revealed that both systems missed early detection on some cases, and that there are specific cases such as evolving events that are particularly hard to detect.

Acknowledgment

This work was carried out with the support of SURF Foundation.

References

[1] J. Allan, V. Lavrenko, D. Malin, and R. Swan. Detections, bounds, and timelines: Umass and TDT-3. In *Proceedings of TDT-3 Workshop*, pages 167–174, 2000.

[2] J. Allan, R. Papka, and V. Lavrenko. On-line new event detection and tracking. In *Proceedings of SIGIR 1998*, pages 37–45. ACM, 1998.

[3] M. Karkali, F. Rousseau, A. Ntoulas, and M. Vazirgiannis. Efficient online novelty detection in news streams. In *WISE 2013*, pages 57–71. Springer, 2013.

[4] R. McCreadie, C. Macdonald, I. Ounis, M. Osborne, and S. Petrovic. Scalable distributed event detection for Twitter. In *IEEE Big Data*, pages 543–549. IEEE, 2013.

[5] M. Osborne, S. Petrovic, R. McCreadie, C. Macdonald, and I. Ounis. Bieber no more: First story detection using twitter and wikipedia. In *SIGIR 2012 TAIA Workshop*, 2012.

[6] R. Papka and J. Allan. Topic detection and tracking: Event clustering as a basis for first story detection. In *Advances in Information Retrieval*, pages 97–126. Springer, 2002.

[7] S. Petrović, M. Osborne, and V. Lavrenko. Streaming first story detection with application to twitter. In *Proceedings of NAACL 2010*, pages 181–189. ACL, 2010.

[8] S. Petrović, M. Osborne, and V. Lavrenko. Using paraphrases for improving first story detection in news and twitter. In *Proceedings of NAACL 2012*, pages 338–346, 2012.

[9] J. B. Vuurens, A. P. de Vries, R. Blanco, and P. Mika. Hierarchy construction for news summarizations. In *Proceedings of SIGIR 2015 TAIA Workshop*, 2015.

[10] J. B. Vuurens, A. P. de Vries, R. Blanco, and P. Mika. Online news tracking for ad-hoc information needs. In *Proceedings of ICTIR 2015*, pages 221–230. ACM, 2015.

[11] Y. Yang, J. Zhang, J. Carbonell, and C. Jin. Topic-conditioned novelty detection. In *Proceedings of SIGKDD 2002*, pages 688–693. ACM, 2002.

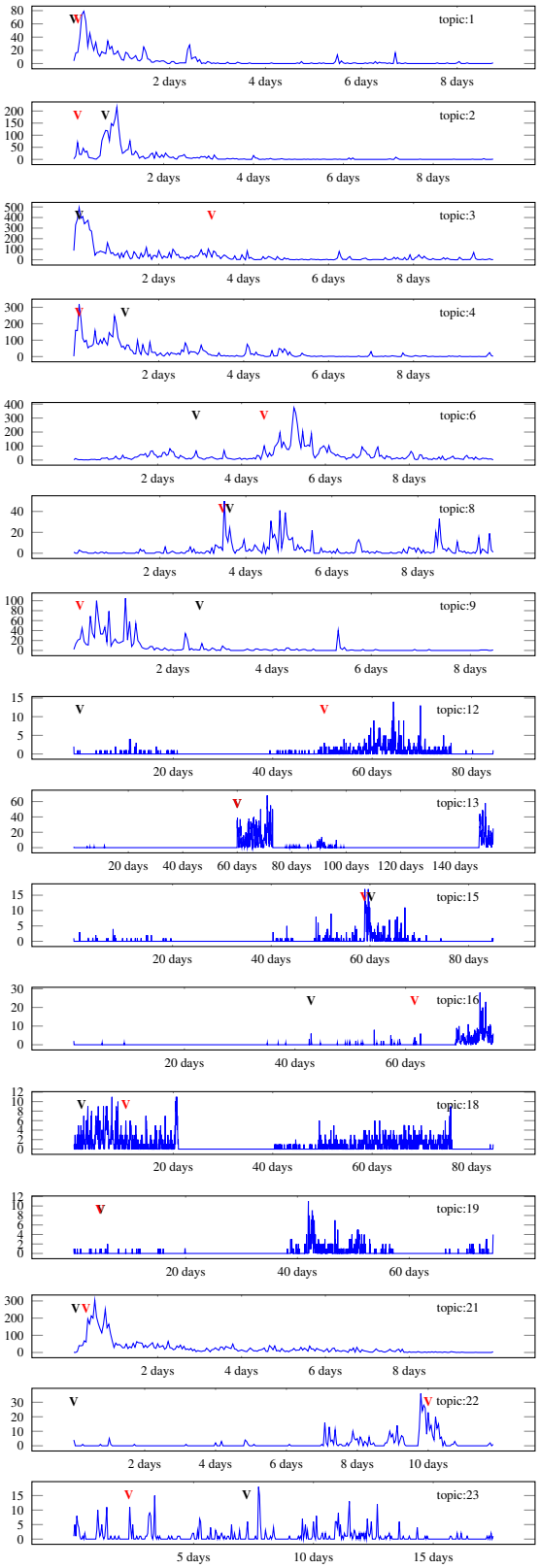

Figure 3: On the y-axis is the number of relevant news articles for the topic per hour, over time on the x-axis. A red V indicates when a fist story is detected by 3NN $\phi_{coherence} = 0.5$, and a black V indicates when a first story is detected by Single Linkage, both at the 'optimal' novelty threshold that obtained recall=1 and the highest precision.