Signal Briefings: Monitoring News Beyond the Brand

James Brill[1], Dyaa Albakour[1], José Esquivel[1], Udo Kruschwitz[2], Miguel Martinez[1] and Jon Chamberlain[3]

[1] Signal AI, London UK
[2] University of Regensburg, Bavaria, Germany
[3] University of Essex, Colchester, UK

Abstract. Public relations (PR) professionals are responsible for managing an organisation's reputation through monitoring entities of interest and wider industry news. Monitoring and tracking wide news spaces such as industry news can cause a significant work load on PR professionals. We present Signal Briefings, a system which uses a combination of clustering and ranking to produce a small set of impactful articles distributed as a periodic email in a scalable and efficient manner.

1 Introduction

A public relations (PR) team, whether internal to a company or an external agency, is responsible for monitoring news articles that mention entities of interest. Typically this information is summarised in a curated digest of the latest articles identified as having relevance and potential impact on the decision making of the organisation. The PR team routinely monitors news about entities of interest such as their organisation and its competitors; however, short-term monitoring is also used for product releases, promotional campaigns, reputation management, and assessing the impact of disruptive events such as Brexit and Covid-19. Historically, media monitoring tools have used Boolean search [3] as the underpinning search technology. Moving away from Boolean search and to facilitate more accurate information retrieval, companies are using Natural Language Processing (NLP), for example, to identify entities and topics, e.g. *Signal AI news monitoring*[4] which analyses up to 3M news articles a day [6], or to automatically generate query suggestions [9].

Media monitoring products are designed to provide focused search results of the most relevance to the user; however, some articles may be excluded that contain important information due to the high volume of news content being released every day (other articles such as summary articles should in fact be excluded [2]). By increasing the search scope, the task of the PR team becomes exponentially harder and less-efficient. The focus on narrow searches hinder PR teams who try to monitor news more widely (in what we describe as a *broad news space*) to anticipate policy amendments, industrial ecosystem change, and evolving customer habits. This must be achieved whilst only reading a small number of articles that are unique in content and up to date. Additionally, this

[4] https://www.signal-ai.com/

information needs to be delivered inline with organisational procedure, typically being processed by the PR team first thing on a week-day morning and made available the same day within the organisation. In this paper, we describe *Signal Briefings* - a system for monitoring broad news spaces by PR professionals.

2 Architecture and Deployment

System Requirements: In order to identify key requirements for the system, 5 PR professionals were interviewed during a 3-week period in June 2019 in 30-minute informal interviews identifying four key aspects: (a) *Minimal noise*: show no duplicate information; (b) *Novel information*: provide information that would not otherwise be found; (c) *Important information*: identify the most important ("*impactful*") articles within a broad news space; (d) *Authority of article*: report source reputation to reduce the need to check for misinformation.

Architecture: *Signal Briefings* identifies up to ten news articles from a set to represent the most impactful stories in the broad news space.[5] These are sent via email to users at a configured time. *Signal Briefings* was integrated into the news monitoring product of Signal, adding a layer on top of the existing concept-based search by grouping and filtering the search results to provide a more diverse representation of the news being monitored by users. The underlying model for this filtering layer consists of three stages: (1) clustering; (2) ranking the clusters; (3) selecting the best article from each cluster.

We use single-pass clustering [7] with nearest-neighbour classification [1] to assign each new document to the closest cluster or instantiate a new cluster if no clusters exist within a maximum distance. A key element of this stage is that documents are processed in chronological order and represented as a sparse TF-IDF vector (to keep computational costs low) with IDF derived from 3 months of data. Our representation of the cluster is the document that instantiates the cluster, allowing us to use cosine similarity as the distance metric to identify nearest neighbours. By clustering similar documents together and reducing duplication, we address user requirement (1): minimal noise. We define a custom ranking function (based on the user requirements) that outperforms typical baseline ranking – chronological ranking and BM25 [8]:

$$f(C) = |C| \cdot \sum_{a \in C} reach(a) + source\_ranking(a) \tag{1}$$

where $C = \{a_1, a_2, ..., a_n\}$ represents a cluster. At an article level, there are two main components *reach* and *source_ranking*. The *reach* function returns normalised readership figures for an article of a certain news type (print, online) obtained through third-party providers such as SimilarWeb[6]. The *reach* function seeks to model the 'size' of the story. The *source_ranking* is a manual proprietary reputation score of sources (ranging between 0 - 1) based on Signal's expertise of the PR domain, and reflects the value of a source to a PR professional. The

[5] A demonstration video is available on https://tinyurl.com/y5txp6ap showing examples of Signal Briefings

[6] https://www.similarweb.com/

ranking function is designed to address user requirements (2) and (3) (novel and important information) by prioritising *impactful* stories. It makes the assumption that stories will be reported on by reputable news sources and will have high readership figures.

The final stage is to select a single article to represent each cluster thereby increasing the number of news stories that can be covered in a single email (of up to ten stories). The article with the highest *source_ranking* is selected to represent each cluster with second preference being chronological order. This stage addresses user requirement (4): article provenance and source authority. *Reach* is excluded from selection as it does not model article provenance well; for example, a state-owned news company might have a weak reputation for journalistic quality but still maintain very high readership.

Scalable Deployment: The user requirement interviews indicated that organisations need this information first thing in the morning. An assessment of Signal's *Bulletins* product (a more traditional search alert system) showed that 87% of emails were sent out between 6am and 9am. The requirement for user retrieval requests to be completed at the same time places a very high demand on server processing and is not scalable without some form of parallelisation. To solve this problem, we ensure a light processing overhead at serve time. This was achieved by using a streaming approach where, instead of searching the index of the collection of documents for each request, each search request processes a constant stream of documents in small batches. For each new document added to the collection, we find the closest cluster it belongs to, update the cluster score, and store the document with the highest source ranking. By using a streaming approach, the majority of the processing work is not a function of the number of search requests, but the total number of documents flowing through the system.

3 Evaluation

The prototype system was integrated into Signal's email services within the live product in February 2020. The prototype could then be evaluated in a restricted but realistic setting with selected clients briefed on its function. The system was evaluated for user value and technical performance.

User Engagement & Value: The output emails were initially validated with five PR professionals (different from the user requirements interviewees). Subsequently, several restricted tests of the prototype on the live product were performed: (1) All Signal's user base were offered the chance to receive an email on up to 3 predefined search queries via an in-app poll; (2) The evaluation was expanded to let the 44 users who volunteered from the previous stage use the unconstrained system for a limited time. This allowed us to better understand the value of the feature by empowering users to use it on their own searches; (3) The beta system was made available to all clients before being released as an additional premium feature.

The operational metrics [5] we used to monitor user engagement were *open rate* (whether the user opened the email that was sent) and *click rate* (whether the user clicked on an article within the email to read it). Over a 6-month period we assessed the relative differences between *Bulletins* and *Signal Briefings* and found an increase in the open rate of 41.4% and an increased click rate of 115.7% [7], i.e. a substantial boost in terms of both metrics (statistically significant in both cases using Mann-Whitney test at $p < 0.01$) Comparing the daily average number of *Signal Briefings* sent in terms of the first two weeks and the last two weeks showed there was an increase of 950%.

A sample of 4,384 *Signal Briefings* from one month were analysed to determine the number of duplicated articles in each one, and the proportion of unknown words in documents that pass through the pipeline. Identifying duplicate articles is vital to ensure novelty of information and the reduction of noise as identified in the four key requirements. Signal's existing de-duplication service, which uses locality sensitive hashing [4], was used to automatically detect duplicates. 94% of briefs contain no duplicates, and 6% contain 1 duplicate. Quantifying the proportion of unique unknown tokens demonstrates how well our vocabulary for vectorisation understands the documents that users search. A high proportion of unique unknown tokens would impact the performance of the clustering quality. Less than 10% of unique tokens were unknown in 99% of the 22,089,510 documents processed in a month period. The client user base contains PR professionals from every industry, hence the proportion of missing tokens are within acceptable limits and were not expected to impact clustering performance.

Technical Performance: The system uses a streaming approach for processing the document collection to collapse broad news spaces. The technical performance of the system was measured as the time taken to perform the retrieval task to produce each email containing the ten most impactful articles. Analysis during the 2-month test period with more than 10K requests shows that the median average time taken remains reasonably consistent at 0.25 seconds. On further inspection of our system, we find this is mostly due to the amount of data transfer needed at the time of serving a request. At serve time, the *Signal Briefings* database stores 6,333,526 articles in 1,084,580 clusters (for all the alerts setup at the time of writing). This is a compression of 83% in the amount of data transfer out of our systems to cluster, rank, and serve the articles in a briefing.

4 Discussion & Conclusion

We have presented Signal's approach to solving an everyday challenge for PR professionals, *monitoring large news spaces with limited resources*. We found that this new feature has significantly more engagement than Signal AI's existing email alerting product. Additionally, the feature has proven to be valuable to Signal AI who created free *Signal Briefings* on COVID-19 based around different industry verticals. The feature was built around the user needs of Signal AI's clients who did not identify potential bias as a concern, thus addressing bias in news spaces was not a high priority in this context but it may be a concern in other systems. Validating the findings of the evaluation in a business-to-business environment with long sales cycles is difficult due to the many other factors involved and as a result, user satisfaction had to be estimated by measuring feature engagement, and other indicators of quality such as duplicate articles.

[7] Exact numbers are redacted due to company confidentiality

References

1. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Transactions on Information Theory **13**(1), 21–27 (Jan 1967)
2. Fisher, M., Albakour, D., Kruschwitz, U., Martinez, M.: Recognising summary articles. In: Azzopardi, L., Stein, B., Fuhr, N., Mayr, P., Hauff, C., Hiemstra, D. (eds.) Advances in Information Retrieval. pp. 69–85. Springer International Publishing, Cham (2019)
3. Frants, V.I., Shapiro, J., Taksa, I., Voiskunskii, V.G.: Boolean search: Current state and perspectives. Journal of the American Society for Information Science **50**(1), 86–95 (1999)
4. Indyk, P., Motwani, R.: Approximate nearest neighbors: towards removing the curse of dimensionality. In: Proceedings of the thirtieth annual ACM symposium on Theory of computing. pp. 604–613 (1998)
5. Karlgren, J.: Adopting Systematic Evaluation Benchmarks in Operational Settings, pp. 583–590. Springer International Publishing, Cham (2019)
6. Martinez-Alvarez, M., Kruschwitz, U., Hall, W., Poesio, M.: Signal: Advanced Real-Time Information Filtering. In: Advances in Information Retrieval - Proceedings of the 37th European Conference on Information Retrieval (ECIR). pp. 793–796 (2015)
7. van Rijsbergen, C.J.: Information retrieval. Butterworth (1979)
8. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: BM25 and beyond. Now Publishers Inc (2009)
9. Verberne, S., Wabeke, T., Kaptein, R.: Boolean queries for news monitoring: Suggesting new query terms to expert users. In: Martinez-Alvarez, M., Kruschwitz, U., Kazai, G., Hopfgartner, F., Corney, D., Campos, R., Albakour, D. (eds.) Proceedings of the First International Workshop on Recent Trends in News Information Retrieval co-located with 38th European Conference on Information Retrieval (ECIR 2016), Padua, Italy, March 20, 2016. CEUR Workshop Proceedings, vol. 1568, pp. 3–8. CEUR-WS.org (2016), http://ceur-ws.org/Vol-1568/paper1.pdf