Editorial Algorithms: Optimizing Recency, Relevance and Diversity for Automated News Curation

Abhijnan Chakraborty
Indian Institute of Technology Kharagpur, India

Mohammad Luqman
Indian Institute of Technology Kharagpur, India

Sidhartha Satapathy
University of Illinois, Urbana-Champaign, USA

Niloy Ganguly
Indian Institute of Technology Kharagpur, India

ABSTRACT

With a large number of stories emerging from the newsrooms, media websites need to curate interesting news for their readers. Although traditionally news was curated solely by human editors, increasing news volume has led media outlets to adopt *editorial algorithms*. However, such algorithms are often proprietary, and smaller outlets do not have the resources to build them from scratch. In this paper, we present a novel framework 'Samar' to automatically curate news by optimizing recency, relevance and diversity of the selected stories. Evaluations over two real-world news datasets show that Samar outperforms several state-of-the-art baselines in matching the news curation performed by human editors.

ACM Reference Format:
Abhijnan Chakraborty, Mohammad Luqman, Sidhartha Satapathy, and Niloy Ganguly. 2018. Editorial Algorithms: Optimizing Recency, Relevance and Diversity for Automated News Curation. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3184558.3186937

1 INTRODUCTION

Due to the large number of news stories available online, news readers need to rely on news curation or recommendation services to find important news [2]. Traditionally, news curation was the sole domain of expert human editors (e.g., while selecting the news stories for the printed newspaper), who used to select what stories should be consumed by the readers – a process known as *journalistic gatekeeping* [8]. However, the emergence of news aggregators (e.g., Google News), social media newsfeeds (e.g., in Facebook, Twitter) and personalized news recommendations have given rise to *editorial algorithms* [7] which replace the human editorial gatekeeping roles.

Large media organizations are also introducing editorial algorithms in their newsrooms. For example, New York Times has built a slack bot 'Blossom' which recommends stories the editors should promote on social media[1]. Similarly, BBC has developed tools to automate different editorial decisions[2]. However, such tools are often proprietary, and smaller media outlets may not have the technical and human resources to build such tools [9]. In this work, we systemically address different challenges in automatizing editorial decisions and build a framework 'Samar' (named after the Bengali poet and editor Samar Sen (en.wikipedia.org/wiki/Samar_Sen)), which can help the editors to conceptualize innovative offerings.

While curating news, there are three basic metrics of interest – recency, relevance and diversity of stories. Recency captures a story's age, i.e., when the story is published. In *personalized recommendations*, relevance denotes how well a story matches a particular reader's interest. However, editors often curate stories for a broad audience (e.g., the stories in a printed newspaper are same for everyone in a city). In such contexts, relevance refers to the importance of a story judged from the editors' *notions of newsworthiness* [8]. In our earlier work [3], we showed that it is tricky to optimize both recency and relevance of recommended news when relevance is estimated through audience-driven popularity measures. Samar tries to circumvent this difficulty by inferring newsworthiness from observing the editorial decisions on past news data. However, along with newsworthiness, curated stories should avoid covering redundant topics, and instead have diverse topical coverage. Samar efficiently combines all these aspects for automated news curation.

To evaluate the effectivity of Samar in curating news stories, we gather extensive data from two very popular news websites – The Guardian (`theguardian.com`) and NYTimes (`nytimes.com`). We find that Samar outperforms several state-of-the-art baselines in matching the editorial decisions at these websites. We conclude by discussing the potential application of Samar in media newsrooms.

2 METHODOLOGY

Samar selects K news stories from a larger set of candidate stories by first computing recency and relevance scores for the candidates, and then inculcating diversity in the final curated set.

- **Recency:** Recency of a story i is measured as the difference between the curation time and the publish time of the story.

$$recency_i = \frac{1}{time\ since\ i\ is\ published} \qquad (1)$$

where the time difference can be computed in seconds, minutes or hours depending on the context. We then normalize $recency_i$ scores of the candidate stories by the score of the most recent candidate.

$$normalized\_recency_i = \frac{recency_i}{max(\{\forall i\ recency_i\})} \qquad (2)$$

- **Relevance:** To calculate relevance of a story, we develop a supervised binary classifier (two classes denote whether a story will be curated or not), and use the predicted curation probability as the relevance score. To some extent, this score reveals the newsworthiness of the story. We use the following features: (i) story abstract, (ii) author name(s), (iii) list of topics (or keywords) describing the story, (iv) news category (e.g., politics, sports) and (v) no. of stories on

[1] http://www.niemanlab.org/2015/08/the-new-york-times-built-a-slack-bot-to-help-decide-which-stories-to-post-to-social-media
[2] http://www.bbc.co.uk/rd/projects/editorial-algorithms

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3186937

same topic(s) published in last 7 days. As features (i)-(iv) are textual features, we first train four text classifiers with individual features (one Convolutional Neural Networks (CNN) based classifier for feature (i), and three Naive Bayes (NB) classifiers for features (ii)-(iv)), and then use the predicted probalibilities for curated/not-curated classes as features for a SVM classifier (with RBF kernel) at the top level[3]. Thus, the SVM classifier effectively uses nine numeric features – predicted probailities from the textual classifiers and no. of related stories (after appropriate scaling). A story's relevance score is then measured as the curation probability predicted by this SVM classifier (using the method proposed by Lin *et al.* [6]).

The CNN architecture for the textual classifier over the abstract is similar to that used in [5], where every abstract is converted to a $m \times n$ matrix (m is the maximum abstract length, and $n = 50$ is the word vector dimension). A convolution operation is applied to every possible window of h words to produce a feature map. We then apply a max over time pooling operation over the feature map and take the maximum value as a feature. Multiple features are obtained by varying the value of h. These features form the penultimate layer and are passed to a fully connected softmax layer whose output gives the probability distribution over curated/not-curated classes. After computing the recency and relevance scores for a story, we compute a linear combination of these scores as

$$\phi_i = \lambda * normalized\_recency_i + (1 - \lambda) * relevance_i \qquad (3)$$

where λ is a hyper parameter, which can be inferred using maximum likelihood estimates over a given training dataset.

• **Diversity:** Diversity can be measured by how different topics are covered by the curated news stories. Formally, Samar tries to maximize the following function $f(S)$ over the curated set S.

$$maximize \sum_i (\phi_i \cdot \sum_{t \in \tau_i} \frac{1}{freq_t}) \cdot x_i \qquad (4)$$

$$\text{subject to } \sum_i x_i \leq K$$

where x_is are indicator variables ($x_i = 1$ denotes that i is among the curated stories); τ_i is the list of topics covered by i, $freq_t$ is the number of articles in S which cover topic t, and K is the size of S.

It can be proved that $f(S)$ is *non-monotone submodular*, and maximizing such functions w.r.t *cardniality constraints* is NP-Hard [4]. We implement $\frac{1}{3}$-*approximation* algorithm proposed in [4] to solve Eqn 4. Intuitively, we first build S by taking K stories with highest ϕ_i scores. Then, we update S if removing a story from S and adding another story from outside S improves the overall diversity score. This process is repeated until no further change in S is possible.

3 EXPERIMENTAL EVALUATION

Datasets: To evaluate the performance of Samar in curating news stories, we collected all stories appearing on The Guardian and NYTimes, and also the stories selected by editors for the printed newspaper everyday throughout July, 2015 to June, 2016. We gathered 90, 355 Guardian and 242, 125 NYTimes stories; out of which, 13, 580 Guardian stories and 40, 419 NYTimes stories got selected for their print editions during this one year period.

Results: We compare Samar with several baselines: (1) most *recent* stories, (2) most *relevant* stories, (3) most *diverse* stories (proposed in [1]), and (4) stories with highest ϕ_i scores (combination of recency and relevance, which is conceptually similar to the metric *Future Impact* proposed in [3]). To compare the performance of different methods, we consider the selection of stories for the daily newspaper of The Guardian and NYTimes from 1st January to 30th June, 2016. For each day, we consider all stories published in last 3 days as the candidate set, and different methods would predict which stories made it to the print edition. Training data is selected on a sliding basis, i.e., to make a prediction for the newspaper on day m, we consider last six month's data upto day $m - 3$ as training.

Dataset	**The Guardian**			**NYTimes**		
Approach	**Acc**	**P**	**R**	**Acc**	**P**	**R**
Most Recent	0.747	0.180	0.180	0.639	0.066	0.086
Most Diverse	0.688	0.083	0.106	0.648	0.086	0.013
Most Relevant	0.737	0.415	0.651	0.823	0.605	0.614
Most Recent+Relevant	0.815	0.528	0.652	0.866	0.776	0.787
Samar	**0.823**	**0.609**	**0.723**	**0.917**	**0.827**	**0.798**

Table 1: Accuracy (Acc), Precision (P) and Recall (R) in predicting the editorial decision of selecting stories for next day's newspaper.

Table 1 shows the results of each of these approaches. We notice that only considering most recent, most relevant or most diverse articles result in poor precision and recall. Considering recency and relevance together achieves considerable performance gains. However, as we can observe in Table 1, Samar performs best for both datasets by capturing all three salient aspects of editorial curation - recency, relevance and diversity of the stories.

4 CONCLUSION

In this work, we develop a framework Samar to automate editorial decisions. As three major factors – recency, relevance and diversity – guide the editorial decisions in curating important stories for the readers, Samar tries to merge these factors to make an effective algorithmic news curator. However, we do not envision a future without the human editors. We believe that a tool like Samar can complement the editors to offer new innovative schemes to their audience. For example, using Samar, media outlets can generate hourly news digests, weekly/bi-weekly newspapers or even curate news from across the web in their unique editorial styles.

Acknowledgments: A. Chakraborty is a recipient of Google India PhD Fellowship and Prime Minister's Fellowship Scheme for Doctoral Research, a public-private partnership between Science & Engineering Research Board (SERB), Dept. of Science & Technology, Government of India and Confederation of Indian Industry (CII).

REFERENCES

[1] Zeinab Abbassi, Vahab S Mirrokni, and Mayur Thakur. 2013. Diversity maximization under matroid constraints. In ACM KDD.

[2] Abhijnan Chakraborty, Saptarshi Ghosh, Niloy Ganguly, and Krishna P Gummadi. 2013. Can trending news stories create coverage bias? on the impact of high content churn in online news media. In Computation+Journalism Symposium.

[3] Abhijnan Chakraborty, Saptarshi Ghosh, Niloy Ganguly, and Krishna P Gummadi. 2017. Optimizing the Recency-Relevancy Trade-off in Online News Recommendations. In Intl. Conf. on World Wide Web (WWW).

[4] Uriel Feige, Vahab S Mirrokni, and Jan Vondrak. 2011. Maximizing non-monotone submodular functions. *SIAM J. Comput.* 40, 4 (2011).

[5] Yoon Kim. 2014. Convolutional neural networks for sentence classification. In Empirical Methods in Natural Language Processing (EMNLP).

[6] Hsuan-Tien Lin, Chih-Jen Lin, and Ruby C Weng. 2007. A note on Platt's probabilistic outputs for support vector machines. Machine learning 68, 3 (2007).

[7] Raz Schwartz, Mor Naaman, and Rannie Teodoro. 2015. Editorial Algorithms: Using Social Media to Discover and Report Local News. In AAAI ICWSM.

[8] Pamela J Shoemaker, Tim P Vos, and Stephen D Reese. 2009. Journalists as gatekeepers. The handbook of journalism studies 73 (2009).

[9] Mark Stencel, Bill Adair, and Prashanth Kamalakanthan. 2014. The goat must be fed: Why digital tools are missing in most newsrooms. Reporters' Lab (2014).

[3] We experimented with different classifiers and found this combination to work best.