INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE

ISSN: 2057-5688

Hashtagger+: Efficient High-Coverage Social Tagging of Streaming News

BOKKA SRINU
PG Scholar, Department of M.C.A,
S.K.B.R P.G College,
Amalapuram, E.G.Dt., A.P, India.
srinuinnocent9100@gmail.com

Mr. NAGA. SRINIVASA RAO*
Asst. Professor, Dept of M.C.A,
S.K.B.R P.G College,
Amalapuram, E.G.Dt., A.P, India.
naagaasrinu@gmail.com

Abstract—
News and social media are now joined at the hip. There are many publications that come from social media, and there are many institutions and factors that rely on social media information for content. Twitter has taken a central role in the dissemination and the consumption of news. Twitter hashtags serve as a natural way of connecting users from Twitter with news organizations, allowing them to explode momentum for stories that are popular, hashtagged, and trending. Hashtaggers provide news organizations with an efficient way to find out what people care about. On the social media side, news editors rely on social media information for following their audience's attention. For example, they tweet in an attempt to reach their followers. We therefore propose, an efficient learning-to-rank framework. Penalized pointwise algorithms empirically outperform multi-class classification and other learning-to-rank approaches. To this end, and motivated by significant accounting issues, and requirement of extensive dataset collection as well as feature engineering, we propose new data collection heuristics and reweighting of given features. We show that all implemented approaches achieve unattainable (with prior art methods) maximal possible coverage (the golden ratio in "n" for general n-class classification.

Index Terms: Learning-to-rank, dynamic topics, social tags, news, real-time hashtag recommendation

I. Introduction

Hashtags tend to appear spontaneously around breaking news or developing news stories, and are a way for news readers to connect to a particular story and community, to get focused updates in real-time. News organizations use hashtags to target Twitter communities in order to promote original content and engage readers. Journalists sometimes introduce new hashtags, but the Twitter crowd is the one that most often creates and drives the usage of a few of many competing hashtags, thus echoing the current social discourse (e.g., #Brexit, and the opinion camps of #VoteLeave and #Remain for the EU referendum story).

A news story can have multiple hashtags, and is likely to have different hashtags at different stages of the story. For example, in the Umbrella Revolution story (a series of street protests in Hong Kong in 2014), Twitter played a huge role: thousands of people were protesting and reporting on

ISSN: 2057-5688

ongoing events by tweeting with their phones. Three main hashtags are used during the event: #HongKong, #OccupyCentral and #UmbrellaRevolution. Each hashtag dominates the discussion at different time points: #HongKong, the location of the events, is popular at the beginning of the story. #OccupyCentral becomes popular when sit-in protests begin to attract wide attention, particularly on Twitter. Finally, #UmbrellaRevolution dominates the topic as it refers to the protesters using umbrellas to protect themselves from teargas. The relationship between the news story and the hashtags is very dynamic, with new hashtags being created and adopted by Twitter users at a rapid pace. It may be seen from this example that for applications aiming to exploit hashtagging, it is critical to capture the dynamic co-evolution of news and hashtags, as the news story evolution influences the Twitter discussions, which in turn may affect the news. We note that the content of some articles may not be obviously related to a story, but a hashtag recommender can use the social discourse to create a bridge between news articles. This work proposes a real-time hashtag recommendation approach that is able to efficiently and effectively capture the dynamic evolution of news and hashtags. Most prior approaches for hashtag recommendation work on static datasets and do not account for the emergence and disappearance of hashtags. Many approaches use topic/class modeling, by considering hashtags as topics, and mapping news articles to topics using content similarity. As the relevant hashtags change quickly and the news and Twitter environments are highly dynamic, approaches that use multi-class classification need continuous retraining to adapt to new content. Additionally, to train models, these methods rely on tweets that contain both hashtags and URLs. Such tweets are very few and tend to be noisy, which may explain the low accuracy of prior methods (e.g., 50 percent precision reported in recent work).

II. RELATED WORK

Existing work on hashtag recommendation tackles the problem from either a class/topic modeling point of view, or from a learning-to-rank perspective. We discuss recent literature from both categories of approaches, as applied to tweets or news articles. Hashtag Recommendation for Tweets. Prior work focusing on hashtag recommendation for tweets relies on MCC modeling on static datasets. The work of [1] builds Na€ıve Bayes or SVM classifiers for hashtags, where (i) a hashtag is seen as a class and (ii) the tweets tagged with that hashtag are assumed to be labeled data for that class. Hashtag recommendation for tweets can be adapted to recommendation for news, by treating the news headline as a rich tweet. As we show in our experiments, MCC approaches are overwhelmed by the data scale, sparsity and noise characteristics of tweets.

Many other approaches employ topic modeling with PLSA [2], DPMM [3] and LDA [4]. For example [5] fits an LDA model to a set of tweets in order to recommend hashtags. They combine the LDA model with a translation model, to address the vocabulary gap between tweets and hashtags. LDA-type approaches face drastic challenges regarding both scalability and accuracy of recommendation, since either hashtags that are too general are recommended, e.g., #news,

ISSN: 2057-5688

#life, or ones that are not actively used by Twitter users. This happens because the focus is on recommending hashtags solely driven by the content of tweets. These models are also not efficient as they need to be constantly retrained to adapt to newly emerging hashtags.

Some recent methods formulate hashtag recommendation based on multi-class modeling with deep neural nets. The work in [6] proposed an attention-based Convolutional Neural Network model for hashtag recommendation to tweets. This approach works on a static dataset and improves the state-of-the-art results, but the recommendation precision is still around 50 percent. The work in [7] uses pairwise L2R for hashtag recommendation for tweets. This work is tailored for tweets with at least one URL and one hashtag in their body, a very small subset of the overall tweet pool discussing news. Training on this small and noisy tweet set can pose serious problems for the recommendation, resulting in low Precision and low coverage, i.e., few tweets receiving any recommendation at all (reported coverage of 50 percent). The data collection is seeded by an external set of 135 trending hashtags collected from hashtags.org each day. This means that many of the hashtags used as seed do not relate to news at all, but just happen to be trending on hashtags.org at the time of collection. Furthermore, there is no focus on news nor on efficient recommendation which is critical for our setting. In contrast to the approach, we use the actual news articles to drive the selection of tweets and candidate hashtags. In our experiments we compare to the method and show that our model achieves much better coverage and Precision@1. Hashtag Recommendation for News: There is little prior work focusing specifically on hashtag recommendation for news. The approach in [8] relies on a manual user query to retrieve related articles, which are then clustered to create a topic profile. A hashtag profile is also created from tweets collected from a set of manually selected accounts. Hashtags with a similar profile to a cluster, are recommended to that cluster. Since the experiments are done on a static collection, the user engagement with the hashtag is not considered. In [9] we proposed a high-precision pointwise L2R framework for hashtag recommendation for news. In this paper, we improve the efficiency and coverage of that method, while preserving high-precision. We explore different methods for retrieving relevant tweets for news articles and evaluate the end-to-end effect on recommendation. There are several published methods for retrieving tweets for news articles.

III. EXISTINGSYSTEM

- Gong et al proposed an attention-based Convolutional Neural Network model for hashtag recommendation to tweets. This approach works on a static dataset and improves the state-of-the-art results, but the recommendation precision is still around 50 percent.
- Sedhai et al used pairwise L2R for hashtag recommendation for tweets. This work is tailored for tweets with at least one URL and one hashtag in their body, a very small subset of the overall tweet pool discussing news.

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

- Shi et al proposed a high-precision point wise L2R framework for hashtag recommendation for news.
- Gruetzed et al focused on temporal aspects of hashtag recommendation and proposes two content-based models implemented in a distributed manner
- **Disadvantages**: Low coverage. Recommendation slowly.

IV. PROPOSED SYSTEM

- This work proposed Hashtagger+, an efficient learning-to-rank framework for merging news and social streams in real-time, by recommending Twitter hashtags to news articles.
- This work provides an extensive study of different approaches for streaming hashtag recommendation, and show that point wise learning-to-rank is more effective than multi-class classification as well as more complex learning-to-rank approaches.
- This work improves the efficiency and coverage of a state-of-the-art hashtag recommendation model by proposing new techniques for data collection and feature computation.

- **Advantages**: High coverage. Recommendation quickly.

V. MODULES

1. Load News Articles & Extract Keywords
2. Relevant Tweets Extraction
3. Candidate Hashtags Extraction & calculate Article-Hashtag Feature Vector
4. Recommended Hashtags

Load News Articles & Extract Keywords:

- First, this module loads lot of news articles.
- Followed by this module extract keywords using PoS tagger.

Relevant Tweets Extraction:

- Furthermore, this module extracts relevant tweets based on keywords.
- It takes each keyword as a query and extracts tweets based on this query.

Candidate Hashtags Extraction & calculate Article-Hashtag Feature Vector:

- This module extracted candidate hashtags from relevant tweets.
- Followed by, this module calculates feature vector for each candidate hashtages on all available articles.
- **Recommended Hashtags:**
- This module first applies Learning-to-Rank model.
- This module recommends best hashtags.

VI. CONCLUSION

In this work we have presented Hashtagger+, an approach for efficient, high-coverage real-time hashtag recommendation for streaming news. Our work has advanced the state-ofthe-art by proposing an L2R model together with a set of efficient algorithms for data collection and feature computation. We have presented a detailed breakdown and analysis of our model, and provided an extensive empirical study of each building block. We showed that pointwise L2R approaches vastly outperform content-based and pairwise/listwise L2R approaches for real-time hashtag recommendation. Finally, we

ISSN: 2057-5688

showed that L2R approaches behave better for recommending hashtags to niche news articles, a setting where most other approaches do not perform well due to lack of data for robust feature computation.

VII. REFERENCES

1. Improving product marketing by predicting early reviewers on E-Commerce websites
S. Kodati, M. Dhasaratham, V. V. S. S. Srikanth, and K. M. Reddy, "Improving product marketing by predicting early reviewers on E-Commerce websites," Deleted Journal, no. 43, pp. 17–25, Apr. 2024, doi: 10.55529/ijrise.43.17.25.

2. Kodati, Dr Sarangam, et al. "Classification of SARS Cov-2 and Non-SARS Cov-2 Pneumonia Using CNN." Journal of Prevention, Diagnosis and Management of Human Diseases (JPDMHD) 2799-1202, vol. 3, no. 06, 23 Nov. 2023, pp. 32–40, journal.hmjournals.com/index.php/JPDMHD/article/view/3406/2798, https://doi.org/10.55529/jpdmhd.36.32.40. Accessed 2 May 2024.

3. V. Srikanth, "CHRONIC KIDNEY DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS," IJTE, pp. 106–109, Jan. 2023, [Online]. Available: http://ijte.uk/archive/2023/CHRONIC-KIDNEY-DISEASE-PREDICTION-USING-MACHINE-LEARNING-ALGORITHMS.pdf

4. V. SRIKANTH, "DETECTION OF PLAGIARISM USING ARTIFICIAL NEURAL NETWORKS," International Journal of Technology and Engineering, vol. XV, no. I, pp. 201–204, Feb. 2023, [Online]. Available: http://ijte.uk/archive/2023/DETECTION-OF-PLAGIARISM-USING-ARTIFICIAL-NEURAL-NETWORKS.pdf

5. V. SRIKANTH, "A REVIEW ON MODELING AND PREDICTING OF CYBER HACKING BREACHES," IJTE, vol. XV, no. I, pp. 300–302, Mar. 2023, [Online]. Available: http://ijte.uk/archive/2023/A-REVIEW-ON-MODELING-AND-PREDICTING-OF-CYBER-HACKING-BREACHES.pdf

6. S. Kodati, M. Dhasaratham, V. V. S. S. Srikanth, and K. M. Reddy, "Detection of fake currency using machine learning models," Deleted Journal, no. 41, pp. 31–38, Dec. 2023, doi: 10.55529/ijrise.41.31.38.

7. "Cyberspace and the Law: Cyber Security." IOK STORE, iokstore.inkofknowledge.com/product-page/cyberspace-and-the-law. Accessed 2 May 2024.

8. "Data Structures Laboratory Manual." IOK STORE, www.iokstore.inkofknowledge.com/product-page/data-structures-laboratory-manual. Accessed 2 May 2024.

9. Data Analytics Using R Programming Lab." IOK STORE, www.iokstore.inkofknowledge.com/product-

ISSN: 2057-5688

page/data-analytics-using-r-programming-lab. Accessed 2 May 2024.

10. V. Srikanth, Dr. I. Reddy, and Department of Information Technology, Sreenidhi Institute of Science and Technology, Hyderabad, 501301, India, "WIRELESS SECURITY PROTOCOLS (WEP,WPA,WPA2 & WPA3)," journal-article, 2019. [Online]. Available: https://www.jetir.org/papers/JETIRDA06001.pdf

10. V. SRIKANTH, "Secured ranked keyword search over encrypted data on cloud," IJIEMR Transactions, vol. 07, no. 02, pp. 111–119, Feb. 2018, [Online]. Available: https://www.ijiemr.org/public/uploads/paper/1121_approvedpaper.pdf

11. V. SRIKANTH, "A NOVEL METHOD FOR BUG DETECTION TECHNIQUES USING INSTANCE SELECTION AND FEATURE SELECTION," IJIEMR Transactions, vol. 06, no. 12, pp. 337–344, Dec. 2017, [Online]. Available: https://www.ijiemr.org/public/uploads/paper/976_approvedpaper.pdf

12 . SRIKANTH MCA, MTECH, MBA, "ANALYZING THE TWEETS AND DETECT TRAFFIC FROM TWITTER ANALYSIS," Feb. 2017. [Online]. Available: http://ijmtarc.in/Papers/Current%20Papers/IJMTARC-170309.pdf

14 Srikanth, V. 2018. "Secret Sharing Algorithm Implementation on Single to Multi Cloud." International Journal of Research 5 (01): 1036–41. https://journals.pen2print.org/index.php/ijr/article/view/11641/11021.

5. K. Meenendranath Reddy, et al. Design and Implementation of Robotic Arm for Pick and Place by Using Bluetooth Technology. No. 34, 16 June 2023, pp. 16–21, https://doi.org/10.55529/jeet.34.16.21. Accessed 20 Aug. 2023.

16. Babu, Dr P. Sankar, et al. "Intelligents Traffic Light Controller for Ambulance." Journal of Image Processing and Intelligent Remote Sensing(JIPIRS) ISSN 2815-0953, vol. 3, no. 04, 19 July 2023, pp. 19–26, journal.hmjournals.com/index.php/JIPIRS/article/view/2425/2316, https://doi.org/10.55529/jipirs.34.19.26. Accessed 24 Aug. 2023.

17. S. Maddilety, et al. "Grid Synchronization Failure Detection on Sensing the Frequency and Voltage beyond the Ranges." Journal of Energy Engineering and Thermodynamics, no. 35, 4 Aug. 2023, pp. 1–7, https://doi.org/10.55529/jeet.35.1.7. Accessed 2 May 2024.

18. K. Meenendranath Reddy, et al. Design and Implementation of Robotic Arm for Pick and Place by Using Bluetooth Technology. No. 34, 16 June 2023, pp. 16–21, https://doi.org/10.55529/jeet.34.16.21. Accessed 20 Aug. 2023