INTERNATIONAL JOURNAL
OF TECHO-ENGINEERING

ISSN : 0975-4520

HASHTAGGER+ EFFICIENT HIGH-COVERAGE SOCIAL TAGGING OF STREAMING NEWS

KONA SALINI MISRA
PG Scholar, Department of Computer Science,
SVKP & Dr K S Raju Arts & Science College,
Penugonda, W.G.Dt., A.P, India.
konashalini1999@gmail.com

A.N.RAMAMANI*
Associate Professor in Computer Science,
SVKP & Dr K S Raju Arts & Science College,
Penugonda, W.G.Dt., A.P, India.
manivasu6@gmail.com

Abstract

News and social media now play a synergistic role and neither domain can be grasped in isolation. On one hand, platforms such as Twitter have taken a central role in the dissemination and consumption of news. On the other hand, news editors rely on social media for following their audience's attention and for crowd-sourcing news stories. Twitter hashtags function as a key connection between Twitter crowds and the news media, by naturally naming and contextualizing stories, grouping the discussion of news and marking topic trends. In this work, we propose Hashtagger+, an efficient learning-to-rank framework for merging news and social streams in real-time, by recommending Twitter hashtags to news articles. We provide an extensive study of different approaches for streaming hashtag recommendation, and show that pointwise learning-to-rank is more effective than multi-class classification as well as more complex learning-to-rank approaches. We improve the efficiency and coverage of a state-of-the-art hashtag recommendation model by proposing new techniques for data collection and feature computation. In our comprehensive evaluation on real-data, we show that we drastically outperform the accuracy and efficiency of prior methods. Our prototype system delivers recommendations in less than 1 minute, with a Precision@1 of 94 percent and article coverage of 80 percent. This is an order of magnitude faster than prior approaches, and brings improvements of 5 percent in precision and 20 percent in coverage. By effectively linking the news stream to the social stream via the recommended hashtags, we open the door to solving many challenging problems related to story detection and tracking. To showcase this potential, we present an application of our recommendations to automated news story tracking via social tags. Our recommendation framework is implemented in a real-time Web system available from insight4news.ucd.ie.

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

1. INTRODUCTION

1.1 Introduction:

In a recent study, nearly 9 in 10 Twitter users say they use Twitter for news, and the vast majority of those (74 percent) do so daily. Most of the news stories spreading on Twitter have names, in the form of hashtags that contextualize the stories. These keyword based tags, describing the content of a tweet, are a natural way to label tweets for news stories. For example, #Brexit, #Remain, #VoteLeave was heavily used for Twitter discussions on the EU membership referendum held in UK in June 2016.

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 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).

1.2 Purpose:

We aim to map a stream of news articles to a stream of Twitter hashtags, in real-time, with high-precision and high-coverage. Real-time refers to the efficiency of a solution: given a new article, how quickly can we recommend hashtags ? For example, we prefer a solution that can deliver recommendations in under 5mins, while a solution that takes hours is not acceptable. High-precision refers to the quality of recommendations. For example, for a news headline such as, Deadly car bomb targets Afghan bank, we prefer focused recommendations, e.g., #afghanistan #helmand, that refer to specific entities involved in this story, instead of generic hashtags such as #news. High-coverage refers to how many articles get recommended hashtags within 5mins. For example, 8 out of 10 articles with at least one recommendation is an example of good coverage, while 1 out of 10 is not.

1.3 Scope:

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 hash tags change quickly and the news and Twitter environments are highly dynamic, approaches that use multi-class classification need continuousre training 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.

1.4 Motivation:

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 hash tags change quickly and the news and Twitter environments are highly dynamic, approaches that use multi-class classification need continuousre training 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.

1.5 Overview:

Hashtag Recommendation for Tweets. Prior work focusing on hashtag recommendation for tweets relies on MCC modeling on static datasets. The work of builds Na€ıve Bayes or SVM classifiers for hashtags, where (i) a hashtag is seen as a classand (ii)the tweets tagged withth at hash tag 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.

2. LITERATURE SURVEY

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, #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.

The method in [10] gathers news articles and tweets independently of each other, then uses co-occurence terms for the two datasets (i.e., articles, tweets) to connect the news and tweets. This approach to collect tweets is not appropriate, as it may result in little or no overlap with the news dataset. The method in [11] explores different ways of generating queries from news articles, but is limited to working only with tweets with URLs. We also generate queries from news articles, but do not restrict ourselves to tweets with URLs. We further advance the work with an extensive study of MCC and L2R approaches evaluated on the task of streaming hashtag recommendation for news.

3. PROBLEM STATEMENT

- 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.
- 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.

3.1 Disadvantages:

➢ These existing works provides poor caching ratio and less hit ratio

4. 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.

4.1 Advantages:

High coverage. Recommendation quickly.

5. Modules Description:

1. Load News Articles & Extract Keywords
2. Relevant Tweets Extraction

3. Candidate Hashtags Extraction & calculate Article-Hashtag Feature Vector
4. Recommended Hashtags

5.1 Load News Articles & Extract Keywords:

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

5.2 Relevant Tweets Extraction:

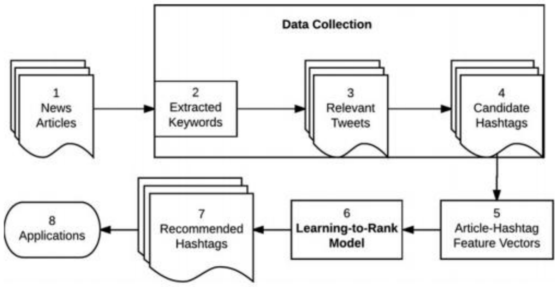
- Furthermore, this module extracts relevant tweets based on keywords.
- It takes each keyword as a query and extracts tweets based on this query.

5.3 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.

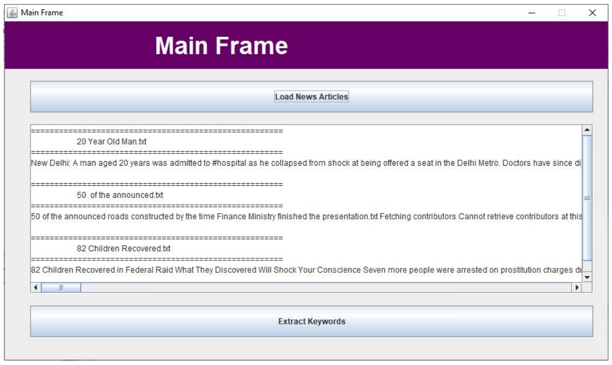
Recommended Hashtags:

- This module first applies Learning-to-Rank model.
- This module recommends best hashtags.

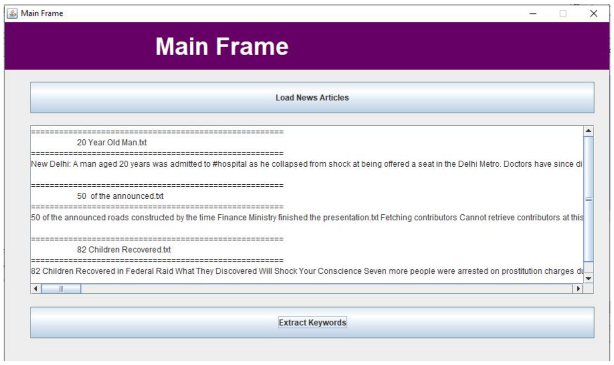


System Architecture

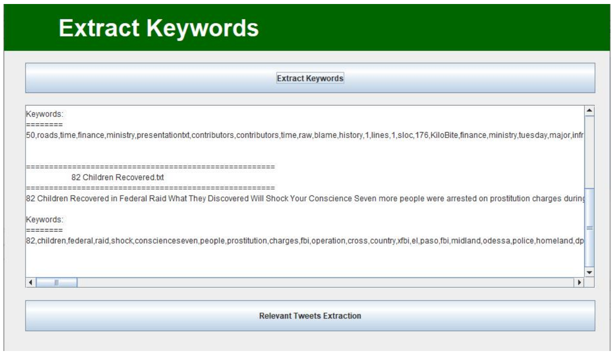
6. OUTPUT RESULT



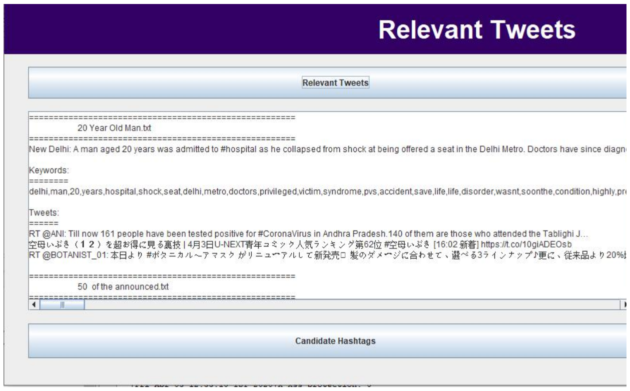
```
=========================================================
              20 Year Old Man.txt
=========================================================
New Delhi: A man aged 20 years was admitted to #hospital as he collapsed from shock at being offered a seat in the Delhi Metro. Doctors have since di

=========================================================
              50  of the announced.txt
=========================================================
50 of the announced roads constructed by the time Finance Ministry finished the presentation.txt Fetching contributors Cannot retrieve contributors at this

=========================================================
              82 Children Recovered.txt
=========================================================
82 Children Recovered in Federal Raid What They Discovered Will Shock Your Conscience Seven more people were arrested on prostitution charges du
```



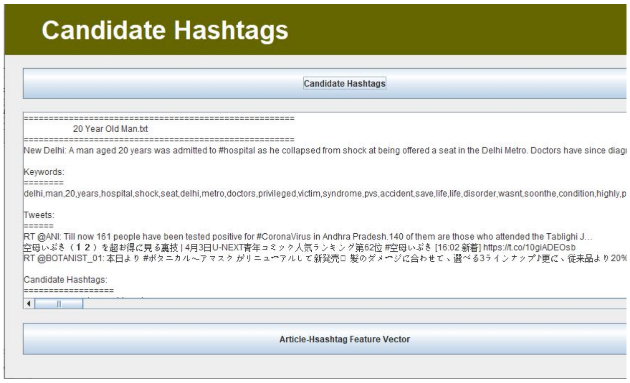
```
=========================================================
              20 Year Old Man.txt
=========================================================
New Delhi: A man aged 20 years was admitted to #hospital as he collapsed from shock at being offered a seat in the Delhi Metro. Doctors have since di

=========================================================
              50  of the announced.txt
=========================================================
50 of the announced roads constructed by the time Finance Ministry finished the presentation.txt Fetching contributors Cannot retrieve contributors at this

=========================================================
              82 Children Recovered.txt
=========================================================
82 Children Recovered in Federal Raid What They Discovered Will Shock Your Conscience Seven more people were arrested on prostitution charges du
```



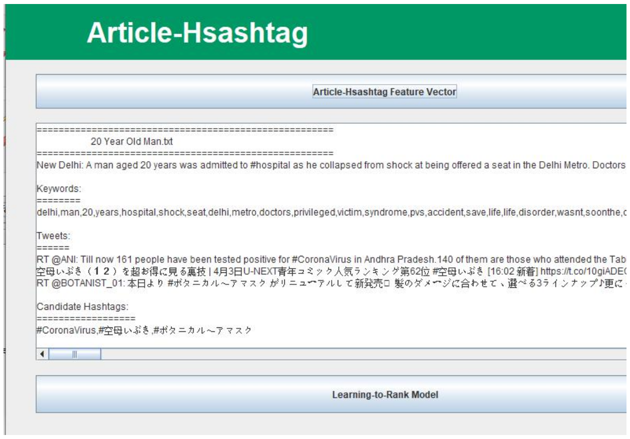
```
Keywords:
========
50,roads,time,finance,ministry,presentationtxt,contributors,contributors,time,raw,blame,history,1,lines,1,sloc,176,KiloBite,finance,ministry,tuesday,major,infr


=========================================================
          82 Children Recovered.txt
=========================================================
82 Children Recovered in Federal Raid What They Discovered Will Shock Your Conscience Seven more people were arrested on prostitution charges durin

Keywords:
========
82,children,federal,raid,shock,conscienceseven,people,prostitution,charges,fbi,operation,cross,country,xfbi,el,paso,fbi,midland,odessa,police,homeland,dp
```



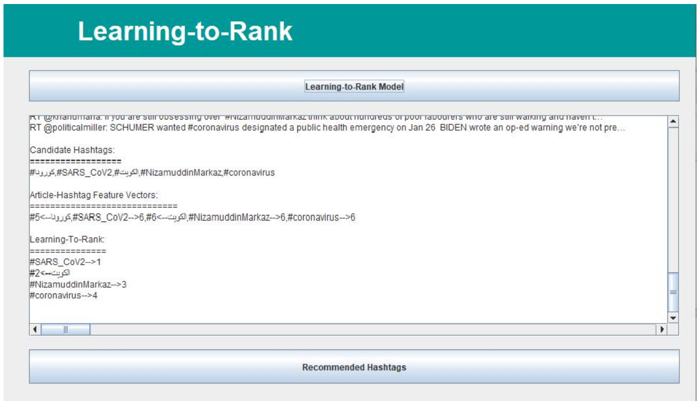
```
=========================================================
          20 Year Old Man.txt
=========================================================
New Delhi: A man aged 20 years was admitted to #hospital as he collapsed from shock at being offered a seat in the Delhi Metro. Doctors have since diagn

Keywords:
========
delhi,man,20,years,hospital,shock,seat,delhi,metro,doctors,privileged,victim,syndrome,pvs,accident,save,life,life,disorder,wasnt,soonthe,condition,highly,pr

Tweets:
======
RT @ANI: Till now 161 people have been tested positive for #CoronaVirus in Andhra Pradesh.140 of them are those who attended the Tablighi J…
空母いぶき（１２）を超お得に見る裏技 | 4月3日U-NEXT青年コミック人気ランキング第62位 #空母いぶき [16:02 新着] https://t.co/10giADEOsb
RT @BOTANIST_01: 本日より #ボタニカル〜アマスク がリニューアルして新発売□ 髪のダメージに合わせて、選べる3ラインナップ♪更に、従来品より20%

=========================================================
          50  of the announced.txt
=========================================================
```



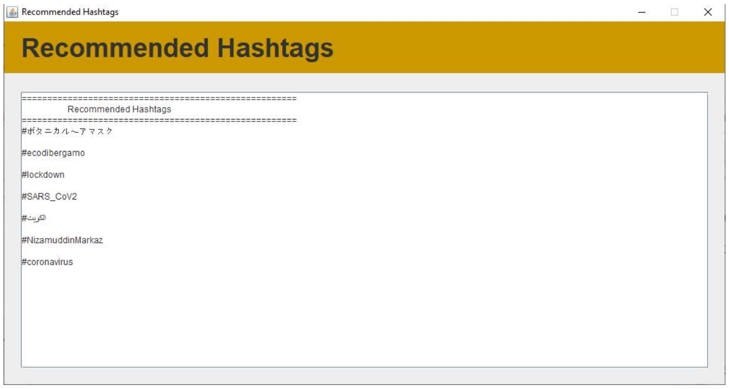
```
=========================================================
          20 Year Old Man.txt
=========================================================
New Delhi: A man aged 20 years was admitted to #hospital as he collapsed from shock at being offered a seat in the Delhi Metro. Doctors have since diag

Keywords:
========
delhi,man,20,years,hospital,shock,seat,delhi,metro,doctors,privileged,victim,syndrome,pvs,accident,save,life,life,disorder,wasnt,soonthe,condition,highly,p

Tweets:
======
RT @ANI: Till now 161 people have been tested positive for #CoronaVirus in Andhra Pradesh.140 of them are those who attended the Tablighi J…
空母いぶき（１２）を超お得に見る裏技 | 4月3日U-NEXT青年コミック人気ランキング第62位 #空母いぶき [16:02 新着] https://t.co/10giADEOsb
RT @BOTANIST_01: 本日より #ボタニカル〜アマスク がリニューアルして新発売□ 髪のダメージに合わせて、選べる3ラインナップ♪更に、従来品より20%

Candidate Hashtags:
==================
```



```
=========================================================
          20 Year Old Man.txt
=========================================================
New Delhi: A man aged 20 years was admitted to #hospital as he collapsed from shock at being offered a seat in the Delhi Metro. Doctors

Keywords:
========
delhi,man,20,years,hospital,shock,seat,delhi,metro,doctors,privileged,victim,syndrome,pvs,accident,save,life,life,disorder,wasnt,soonthe,c

Tweets:
======
RT @ANI: Till now 161 people have been tested positive for #CoronaVirus in Andhra Pradesh.140 of them are those who attended the Tab
空母いぶき（１２）を超お得に見る裏技 | 4月3日U-NEXT青年コミック人気ランキング第62位 #空母いぶき [16:02 新着] https://t.co/10giADE
RT @BOTANIST_01: 本日より #ボタニカル〜アマスク がリニューアルして新発売□ 髪のダメージに合わせて、選べる3ラインナップ♪更に

Candidate Hashtags:
==================
#CoronaVirus,#空母いぶき,#ボタニカル〜アマスク
```



```
RT @khanumarfa: If you are still obsessing over #NizamuddinMarkaz think about hundreds of poor labourers who are still walking and haven't…
RT @politicalmiller: SCHUMER wanted #coronavirus designated a public health emergency on Jan 26  BIDEN wrote an op-ed warning we're not pre…

Candidate Hashtags:
==================
#كورونا,#SARS_CoV2,#الكويت,#NizamuddinMarkaz,#coronavirus

Article-Hashtag Feature Vectors:
=============================
#كورونا-->5,#SARS_CoV2-->6,#الكويت-->6,#NizamuddinMarkaz-->6,#coronavirus-->6

Learning-To-Rank:
===============
#SARS_CoV2-->1
#الكويت-->2
#NizamuddinMarkaz-->3
#coronavirus-->4
```



```
=========================================================
          Recommended Hashtags
=========================================================
#ボタニカル〜アマスク

#ecodibergamo

#lockdown

#SARS_CoV2

#الكويت

#NizamuddinMarkaz

#coronavirus
```

7. CONCLUSION AND FUTURE ENHANCEMENTS

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 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.

8. BIBLIOGRAPHY

[1] R. Dovgopol and M. Nohelty, “Twitter hash tag recommendation,” CoRR, vol. abs/1502.00094, 2015, http://arxiv.org/abs/1502.00094

[2] Z. Ma, A. Sun, Q. Yuan, and G. Cong, “Tagging your tweets: A probabilistic modeling of hashtag annotation in twitter,” in Proc. 23rd ACM Int. Conf. Conf. Inform. Knowl. Manage., 2014, pp. 999–1008.

[3] Y. Gong, Q. Zhang, and X. Huang, “Hashtag recommendation using dirichlet process mixture models incorporating types of hashtags,” in Proc. Empirical Methods Natural Language Process., 2015, pp. 401–410.

[4] F. Godin, V. Slavkovikj, W. De Neve, B. Schrauwen, and R. Van de Walle, “Using topic models for twitter hashtag recommendation,” in Proc. 22nd Int. Conf. World Wide Web, 2013, pp. 593–596.

[5] Z. Ding, X. Qiu, Q. Zhang, and X. Huang, “Learning topical translation model for microblog hashtag suggestion,” in Proc. 23rd Int. Joint Conf. Artif. Intell., 2013, pp. 2078–2084.

[6] Y. Gong and Q. Zhang, “Hashtag recommendation using attention-based convolutional neural network,” in Proc. 25th Int. Joint Conf. Artif. Intell., 2016, pp. 2782–2788.

[7] S. Sedhai and A. Sun, “Hashtag recommendation for hyperlinked tweets,” in Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inform. Retrieval, 2014, pp. 831–834.

[8] F. Xiao, T. Noro, and T. Tokuda, “News-topic oriented hashtag recommendation in twitter based on characteristic co-occurrence word detection,” in 12th Int. Conf. Web Eng., 2012.

[9] B. Shi, G. Ifrim, and N. Hurley, “Learning-to-rank for real-time high-precision hashtag recommendation for streaming news,” in Proc. 25th Int. Conf. World Wide Web, 2016, pp. 1191–1202.

[10] O. Phelan, K. McCarthy, and B. Smyth, “Using twitter to recommend real-time topical news,” in Proc. 3rd ACM Conf. Recommender Syst., 2009, pp. 385–388.

[11] M. Tsagkias, M. de Rijke, and W. Weerkamp, “Linking online news and social media,” in Proc. 4th ACM Int. Conf. Web Search Data Mining, 2011, pp. 565–574.

ABOUT AUTHORS:

KONA SALINI MISRA is currently pursuing MCA in SVKP & Dr K S Raju Arts & Science College, affiliated to Adikavi Nannaya University, Rajamahendravaram. Her research interests include WebTechnology and Internet of Things.

A.N.Ramamani is working as Associate Professor in SVKP & Dr K S Raju Arts &Science College, Penugonda, A.P. She received Masters Degree in Computer Applications from Andhra University and Computer Science & Engineering from Jawaharlal Nehru Technological University Kakinada, Kakinada, India. Her research interests include Software Engineering, Web Technology, Internet of Things.