Identifying Top-k Consistent News-Casters on Twitter

Sahisnu Mazumder[†], Sameep Mehta[‡] and Dhaval Patel[†]
[†]Indian Institute of Technology - Roorkee, Uttarakhand, India.
[‡]IBM Research Lab - India , New Delhi, India
sahisnumazumder@gmail.com, sameepmehta@in.ibm.com, patelfec@iitr.ac.in

ABSTRACT

News-casters are Twitter users who periodically pick up interesting news from online news media and spread it to their followers' network. Existing works on Twitter user analysis have only analysed a *pre-defined set of users* for user modeling, influence analysis and news recommendation. The problem of identifying *prominent*, *trustworthy* and *consistent* news-casters is unaddressed so far. In this paper, we present a framework, **NCFinder**, to discover top-k consistent news-casters *directly* from Twitter. NCFinder uses news headlines published in online news sources to periodically collect authentic news-tweets and processes them to discover news-casters, news sources and news concepts. Next, NCFinder builds a tripartite graph among news-casters, news source and news concepts and employs HITS algorithm on it to score the news-casters on daily basis. The daily score profiles of the news-casters collected over a time-period are then used to infer top-k consistent news-casters. We run NCFinder from 11th Nov. to 24th Nov., 2014 and discover top-100 consistent news-casters and their profile information.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Search process, Information filtering.*

Keywords

News, Twitter, News-Tweet, Tweet Authentication, User Profiling.

1. INTRODUCTION

Twitter is playing a key role in disseminating news among large number of people across globe [4]. Broadly, two types of twitter users, *online new sources* and *news-casters* facilitate the news dissemination process. Online news sources like *@timesofindia*, *@BBCWorld* etc. inject fresh news on Twitter from their webpages as early as possible. Whereas, news-casters acquire news from various news websites *based on their interest* and spread them to their followers' network. Thus, news-casters act as *news curators* as well as *opinion leaders* on Twitter and are good candidates for targeted advertisements [10], news recommendation [1], topic-aware influential user analysis [9] etc. applications.

CIKM'15, October 19-23, 2015, Melbourne, VIC, Australia.

DOI: http://dx.doi.org/10.1145/2806416.2806649.

Spotting of news-casters on Twitter can be done by monitoring and profiling individual Twitter user's activity. The profiling task can be performed using user modeling systems, such as TUMS[1]. But, TUMS takes about 1-2 mins to profile a first-time user and often, the system fails to profile certain users (see Table 2). Thus, profiling millions of Twitter users for news-caster discovery is not a feasible task. Even if we consider the follower lists of prominent news sources on Twitter to form a good candidate set, the lists are huge in general and there may be a news-caster who is not follower of any news source on Twitter. Moreover, if we use a set of predefined keywords/hashtags for sampling the Twitter users, it is also not fruitful as the interests of news-casters change over time along with the dynamics of news and it's importance [5]. So, the problem is how can we directly discover the set of Twitter users who have high potential to qualify as news-casters? Further, news-casters may not be consistent in their news advertisement activity and/or have ill-intention of making use of news for advertising non-news sources [2]. Such news-casters are not beneficial for real-world applications and should be separated out from the authentic ones.

In this paper, we present a framework, *NCFinder* that discovers top-k consistent news-casters *directly* from Twitter. NCFinder periodically collects fresh news headlines published in a pre-identified list of popular online news sources, extracts news concepts from the headlines, and uses those concepts to acquire the tweets (news-tweets) that are similar to the headlines. Next, NCFinder evaluates authenticity of the downloaded news-tweets and processes the authentic news-tweets to discover the news-casters and news sources, news concepts referred by them. Finally, NCFinder forms a tripartite graph consisting of news-casters, news sources and news concepts and employs HITS algorithm on it to compute score of each news-caster on daily basis. The daily score profiles of the news-casters collected over a given time-period are then used to output top-k consistent (having k highest average scores) news-casters. Our contributions in developing NCFinder include:

- We present a systematic and effective solution to the problem of *direct discovery* of news-casters on Twitter by leveraging online news source content.
- We design a novel URL validation based news-tweet authentication technique to detect the *trustworthiness* of content published by news-casters on Twitter.
- We extend HITS algorithm to work on a tripartite graph and use it to score the news-casters on daily basis. The scores thus calculated help us to measure *news-awareness* and study *temporal consistency* of the news-casters over a time-span.

We ran NCFinder for two weeks, collected 6,87,740 authenticated news-tweets and obtained top-100 consistent news-casters along with their news interest profile.

[1]`http://wis.ewi.tudelft.nl/tums/`

2. RELATED WORK

A couple of research works have been performed on Twitter user analysis in recent years for influential user discovery [9], news spread analysis [6], user interest modeling and news recommendation [1]. But, these works mainly focus on analysing a *pre-defined set* of Twitter users. Our work leverages news headlines for direct discovery of news casters without using any pre-defined user set. The idea of leveraging online news media is also presented in [8] that analyses implicitly linked social media utterance for a given news article. However, our aim is to perform user centric analysis.

Information credibility is an important problem on Twitter that focuses on identifying the credibility of a tweet content and has mostly been tackled using feature-based learning technique [2]. But, our news-tweet authentication method works in a straight forward way without learning any features. The idea of leveraging news headlines (act as a ground truth) and usage of embedded URL in a news-tweet make it easy to verify the tweet in real-time.

Regarding consistency analysis on Twitter domain, [7] has recently studied the consistency of Twitter user's behavior for measuring topical engagement of the user. However, the work does not concern about top-k consistent user discovery and profiling. Thus considering the existing works, our work can be considered as a new addition that have explored the consistency of the trustworthy news-casters and their news interest profile on Twitter domain.

3. NCFINDER

The working of NCFinder is decomposed into three functional modules: (1) *News-Tweet Corpus Collection*, (2) *News-Tweet Authentication* and (3) *News-Caster Discovery and Ranking*. At present, the first two modules run at an interval of 30 minutes and the third module runs on daily basis.

3.1 News-Tweet Corpus Collection

The News-Tweet Corpus Collection module works by co-operation of two distinct sub-modules: *Headline-Crawler* and *Tweet-Crawler*. The Headline-Crawler module periodically crawls reliable online news sources (selected from various categories like Politics, Sports, Technology, Finance etc.) and captures fresh news headlines. At present, Headline-Crawler crawls 137 news sources and obtains around 200-300 fresh headlines in each execution on an average.

The Tweet-Crawler module periodically downloads the news-tweets using Twitter REST API and the crawled news headlines one by one. As news-tweets are generally posted after the news headlines, Tweet-Crawler selects news headlines that are crawled t hours earlier from the Tweet-Crawler execution time for tweet crawling purpose. In our approach, Tweet-Crawler selects news headlines crawled at 3 hrs. before from its current execution time.

The process of crawling news-tweets using headline H is as follows. First, we process headline H to discover the news concepts. News concepts are basically important terms/keywords in a headline and often used by Twitter users in their tweets to grab the attention of their followers. In our approach, the news concepts are nouns and/or collocation of nouns (noun phrases) and/or collocation of nouns and numbers present in the headline. We use Stanford Part-of-Speech Tagger to tag each word in H with its corresponding parts-of-speech and extract the set of news concepts present in the headline. E.g., considering the headline "*Passenger plane crashes in Taiwan with 58 people aboard 9 killed*", "*Passenger plane crashes*", "*Taiwan*" and "*58 people*" are news concepts.

Let C be the set of news concepts extracted from headline H. Using C, Tweet-Crawler generates a search keyword in the form of a triplet $\langle c_i, c_j$, "http"$\rangle$ such that $c_i, c_j \in C$, $i < j$ *and* $1 \leq i, j \leq |C|$, and uses the triplet for tweet crawling purpose. E.g., If C = {*Passenger plane crashes*, *Taiwan*, *58 people*}, then $\langle$*Passenger plane crashes*, *Taiwan*, *http*$\rangle$ is a search keyword. Note that, using *pair of news concepts* to form the search keyword ensures that the downloaded tweets have high probability to be contextually similar with the headline. Compared to this, *use of single news concept* as a search keyword will download many irrelevant tweets and *use of all news concepts* will discard many contextually similar news-tweets. Besides that, we use "*http*" tag in the search keyword to ensure that each downloaded tweet contains at least one URL. Generally, news-casters embed URL in tweet content to emphasize that the news is not fake/remour and let their followers to click the link for getting the news article details.

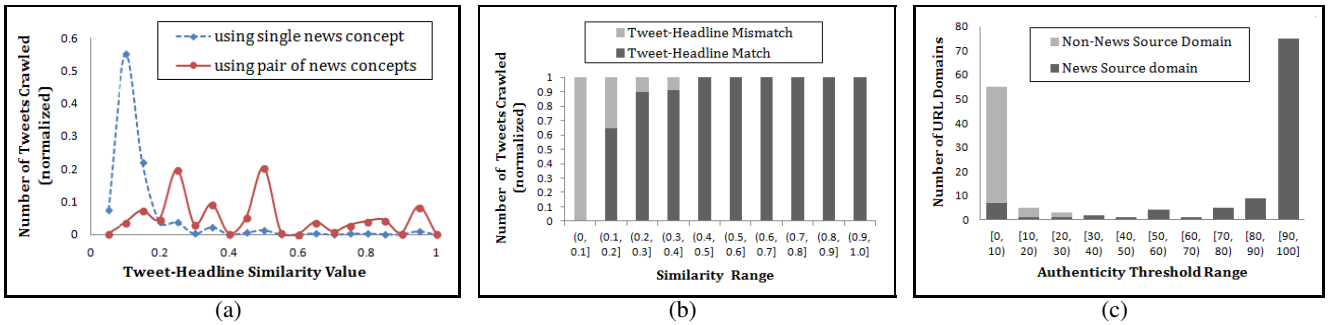
Now, we identify news-tweets from the set of crawled tweets (say, T) that are contextually similar to headline H. In particular, tweet $t \in T$ is a *news-tweet* if $Sim(H, t)$ is greater than or equal to the user-defined similarity threshold. In our approach, $Sim(H, t)$ is an average of unigram similarity and bigram similarity of tweet content with the headline. The unigram similarity, given by $UniSim(H, t) = \frac{|U_H \bigcap U_t|}{|U_H|}$, is the fraction of headline words that appear in the tweet and bigram similarity, given by $BiSim(H, t) = \frac{|B_H \bigcap B_t|}{|B_H|}$, captures the contextual similarity of the tweet with the headline. Here, U_H and U_t (B_H and B_t) are the sets of unigram (bigram) words present in H and t respectively. Before calculating $Sim(H, t)$, we remove single letter words and stopwords from the tweet and the headline. Also, the words starting with '@', "RT", "http://" and '#' are removed from the tweet.

3.2 News-Tweet Authentication

Advertising E-commerce websites using news headline content is a common spamming activity on Twitter. E.g., we discover a news-tweet "#usa » `http://t.co/lNKSIAD8iX` *North Korea to send envoy to Russia ...*" posted by user "**@olga_casey**", where the shown embedded URL points to the fashion website "**`BAKMODA.com`**". As news-tweets are leveraged to discover and rank the news-casters, it is important to remove the news-tweets which intentionally use news media content to advertise non-news sources.

The News-Tweet Authentication module uses the embedded URL of tweet t for its verification purpose. First, we obtain long URL from the shortened URL using URL expansion technique. Next, if the domain of the expanded URL is a news source, we declare the news-tweet as authentic. However, it is not feasible to have the exhaustive list of all news sources for news-tweet authentication purpose. Thus, we devise a technique to incrementally learn a list of news sources as well as a list of non-news sources as follows.

Let $AList$ be the list of news source domains and $BList$ be the list of non-news source domains. Initially, $AList$ and $BList$ are empty. Let, $t.url$ be the expanded URL obtained from the shortened URL in a given news-tweet t. We mark tweet t as authentic if domain of $t.url$ is in $AList$ or as not authentic if the domain is in $BList$. However, if the domain of $t.url$ is not contained by both $AList$ and $BList$, we process tweet t further. We check whether the web page pointed by $t.url$ contains the headline H or not, where H has been used for crawling t. If the webpage does not contain H, the domain of $t.url$ becomes a candidate for non-news source. However, we follow a lazy approach and wait for more evidences in order to declare the domain of $t.url$ as news source; as sometimes news sources remove the news headline after a certain period of time. At present, we make a decision about any URL domain, if it is encountered in 100 different news-tweets. We put the URL domain in $AList$ if out of 100 times, we get evidence for the URL domain of being a news source for at least 30 times. Otherwise, we put it into $BList$. In summary, current module uses $AList$ and $BList$ and discovers the authentic news-tweets.



(a) (b) (c)

Figure 1: (a) Histogram of tweets against various similarity ranges obtained in single concept based and concept pair based tweets crawling. (b) Histogram of tweets (with match and mismatch ratio) against various similarity ranges obtained in concept pair based tweets crawling. (c) Distribution of news-sources and non-news source domains of 160 classified URL domains against various authenticity threshold ranges.

3.3 News-Caster Discovery and Ranking

The Twitter users associated with authentic news-tweets are the candidates for news-casters. However, it is possible that some of the users associated with the authentic news-tweets may be online news sources. In order to remove online news sources from the set of *true* news-casters, we obtain user profile information viz., *original user name* and *profile description* of each Twitter user. If the *original user name* or *profile description* of a given user contains "*news*" word in it, the user is declared as *online news source* and discarded from further analysis.

Now, we proceed to score the discovered news-casters on daily basis. We borrow the idea of HITS algorithm which scores a webpage high if it is associated with high authority webpages and many hubs point to it [3]. In our case, we score a news-caster higher if he/she publishes large number of distinct news-tweets where he/she has referred many popular news sources as embedded links and discussed about many popular news concepts. Here, we consider a news source's popularity higher if many news-casters have referred it in their news-tweets and the news source has published news headlines involving many popular news concepts. Similarly, a news concept's popularity can be explained accordingly.

We first build a weighted tripartite graph $G = (V, E)$ consisting of three types of vertices, namely *news-casters*, *news sources* and *news concepts*. The news-casters are already obtained in the previous paragraph and $AList$ is used as the list of news sources. The news concepts are taken as the concepts that co-occur in both news headline and corresponding news-tweet. Let u, c, s are vertices of graph G, where u is a news-caster, c is a news concept, and s is a news source. The weight of an edge between u and s, denoted as $W(u, s)$, is the number of distinct news-tweets where u has referred s. Similarly, $W(u, c)$ is the number of distinct news-tweets where u has talked about c, and $W(s, c)$ is the number of distinct news-tweets where s and c has co-occurred.

Given the graph G, we compute score for each node in G using HITS algorithm. Compared to the existing version of the HITS algorithm [3], the score of each node in our case is iteratively calculated by enforcing two mutual reinforcement relationships. E.g., the score of news source s, denoted as s_{score} is calculated based on the importance of its relationship with the news concepts and news-casters to which it is connected in G (See Equation 1).

$$s_{score} = \sum_{u \in V_U} W(u, s) * u_{score} + \sum_{c \in V_C} W(s, c) * c_{score} \quad (1)$$

where V_U, V_C are the set of news-caster nodes and news concept nodes and u_{score}, c_{score} are the scores of u and c in G respectively.

We initialize score of each node as 1 and then run the score computation process for N iterations. In each iteration, the score of news source nodes are computed, normalized and used in the score computation of other two types of nodes. In a similar way, the scores of the news concepts and news-casters are subsequently used in next iteration. After N iterations, we consider the normalized u_{score} as the final score of news-caster u on day d.

Once we acquire the daily score profiles of the news-casters over a specified time-period, we compute average score of each news-caster over that time-period and rank them based on their average score values. Finally, we infer news-casters with k highest average scores as the top-k consistent news-casters for that time-period.

Table 1: Examples of new news source domains and non-news source domains discovered by NCFinder.

New News Source Domains	Non-news Source Domains
wsj.com	bakmoda.com
engadget.com	ebay.com
reuters.com	brainyquote.com

4. EXPERIMENTAL ANALYSIS

We evaluate NCFinder empirically at different steps of processing and present sample output inferred by our system. Unless specified explicitly, the result reported in this section is obtained by running NCFinder from 11th Nov. to 24th Nov., 2014.

4.1 Evaluation I: News-Tweet Corpus Collection

We evaluate usefulness of our *news-concept pair based tweet crawling* scheme. We randomly select 10 news headlines, download a tweet dataset using the technique discussed in Section 3.1 and discovered 670 distinct tweets. For comparison purpose, we also download another tweet dataset using *single* news concept as a search keyword for the same set of headlines and obtain 4783 distinct tweets. Next, we plot the distribution of the tweets' similarity values for both the samples [see Figure 1(a)]. We observe that single news-concept based tweet crawling scheme downloads majority of the tweets (88.66%) having low similarity value (range of 0.05 to 0.2). Whereas, only 15.97% of the tweets obtained using our scheme, belong to the said low similarity range. On detailed analysis, we identify that news-tweets with low similarity value are not contextually similar to the selected news headlines. To validate this claim, we have sampled a set of 364 news-tweets from the 670 distinct tweets crawled by our approach and manually annotated each (*tweet*, *headline*) pair in the sample as *match* (contextually similar) or *mismatch* along with corresponding computed similarity value. Figure 1(b) shows the distribution of match and mismatch counts against various similarity ranges. The histogram shows that majority of the tweet-headline mismatches occur in the low similarity value range (0, 0.2] and majority of the tweets downloaded by our method have similarity value higher than 0.2 [see Figure 1(a)] and have high chance of match with headline [see Figure 1(b)]. Thus, our news-tweet corpus collection module is able to

Table 2: Top-5 consistent news-casters and their profile statistics discovered during 11th Nov. to 24th Nov., 2014.

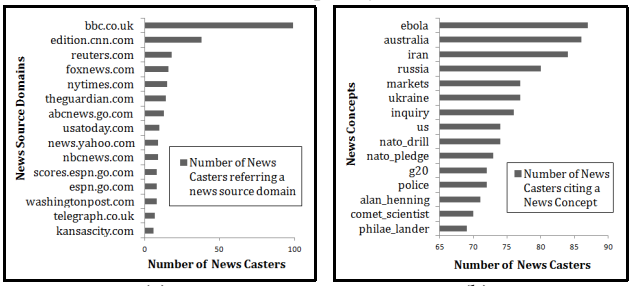
NCFinder				TUMS
Sr. No.	**Twitter User Handle**	**Top News Concepts**	**Top News Sources**	**Topic Profile**
1	TopUp__	australia, uk, france, ebola, roger federer	bbc.co.uk	-
2	krs21da	ukrane, us, putin, obama, g20, ebola, suicide bomber	bbc.co.uk, abcnews.go.com, edition.cnn.com	Entertainment, Culture, Law, Crime, Politics
3	SubrataDhoni	india, australia, england, pakistan, modi	bbc.co.uk, ndtv.com, abplive.in	-
4	guilhermecorde	china, ebola, g20, romania, inquiry	bbc.co.uk	Entertainment, Buisness, Health
5	davidcunha_	china, g20, ebola, romania, growth	bbc.co.uk	-

process many headlines, downloads more relevant news-tweets and meets the Twitter rate limit. E.g., we noted that the said module has crawled 16,212 news-tweets using 211 headlines in 22.41 minutes.

4.2 Evaluation II: News-Tweet Authentication

We evaluate how accurately we have learned $AList$ and $BList$ to infer the authenticity of the crawled news-tweets. We have analysed the URLs accumulated in $AList$ and $BList$ and obtained 160 classified URL domains for one week period, where $AList$ contains 97 URL domains and BList contains 63 URL domains. We have manually verified the webpage of each of the 160 URL domains and observed that out of 97 authenticated URLs, 96 are news sources and 55 URLs in $BList$ are blacklisted (non-news source domains). In summary, F-measure of classifying URL domain as authentic news-source following our technique is 0.956.

Table 1 shows examples of correctly identified 3 new news source domains and 3 non-news source domains discovered by NCFinder. NCFinder has discovered many new news sources that does not belong to our *pre-defined news source set*. In fact, we have discovered 71 new news source domains frequently referred on Twitter.



(a) (b)

Figure 2: Distributions of (a) news source domains and (b) news concepts for Top-100 consistent news-casters.

Note that, the content of $AList$ and $BList$ depend on a threshold value which we use for making decision. Figure 1(c) plots a distribution of news-sources and non-news sources against authenticity threshold ranges. Here, an URL domain l belonging to the range [10, 20) signifies that l has been authenticated at least 10 and at max 20 number of times out of 100 trials. The histogram shows that, majority (86.67%) of the URL domains belonging to the range [0, 30) are non-news source domains. In tweet authentication, precision being more important than recall, we have chosen 30 as the threshold and obtained a precision of 0.99 and recall 0.92. In terms of performance, our multi-threaded implementation of the authentication module has validated 16,212 news-tweets in 5.45 minutes.

4.3 Evaluation III: News-Caster Discovery

We evaluate our news-caster discovery step using user verification process. we have randomly collected sets of 100 users, visited their Twitter page and verified whether the user is news source or not and obtained 0.975 precision and 0.927 recall values.

4.4 Analysis of Output of NCFinder

Table 2 shows top news concepts and news sources cited by top-5 consistent news-casters discovered by NCFinder along their topic profile information obtained from TUMS. The results show that the news interest profile (top news concepts and news sources) discovered by NCFinder are more fine grained compared to the topic profile given by TUMS. In fact, the results of NCFinder for a given news-caster are not only complementing the topic profile, but also specifying detailed behavior of the news-caster over a given timespan. E.g., the news-caster "*guilhermecorde*" has posted tweets about "*ebola*" (Health) by acquiring news from "*bbc.co.uk*".

Figure 2 shows the distribution of news sources and news concepts of Top-100 consistent news-casters. In Figure 2(a) and (b), we see that "*bbc.co.uk*", "*edition.cnn.com*" etc. are popular news sources and "*ebola*", "*australia*" etc. are top news concepts that have been referred by majority of the top-100 consistent news-casters. The skewness of the distributions shows that news-casters mostly browse limited number of popular news sources rather than scrolling through different domains and almost talk about a diverse set of popular news concepts rather than a certain one.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we propose a framework NCFinder to directly discover top-k consistent news-casters on Twitter. We have performed experiments to evaluate the effectiveness and efficacy of NCFinder and analyzed the discovered top-100 news-casters' profile. However, at present, NCFinder does not address the issue of Twitter Bots. As a future work, we aim to incorporate a *Bot filtering step* into NCFinder and also, enrich the ranking process of news-casters by considering their popularity aspects and social influence on Twitter.

6. REFERENCES

[1] F. Abel, Q. Gao, G.-J. Houben, and K. Tao. Analyzing user modeling on twitter for personalized news recommendations. In *User Modeling, Adaption and Personalization*, 2011.

[2] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *WWW*, 2011.

[3] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *JACM*, 1999.

[4] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *WWW*, 2010.

[5] S. Mazumder, B. Bishnoi, and D. Patel. News headlines: What they can tell us? In *IBM I-CARE*, 2014.

[6] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. Twitterstand: news in tweets. In *SIGSPATIAL*, 2009.

[7] A. Schmidt, C. Fink, and N. Bos. Topical engagement on twitter: Using consistency of activity as a means of user segmentation. In *AAAI*, 2014.

[8] M. Tsagkias, M. de Rijke, and W. Weerkamp. Linking online news and social media. In *WSDM*, 2011.

[9] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *WSDM*, 2010.

[10] W.-S. Yang, J.-B. Dia, H.-C. Cheng, and H.-T. Lin. Mining social networks for targeted advertising. In *HICSS*, 2006.