Finding News Curators in Twitter

Janette Lehmann
Universitat Pompeu Fabra
Barcelona, Spain
jnt.lehmann@gmail.com

Carlos Castillo
Qatar Computing
Research Institute
Doha, Qatar
chato@acm.org

Mounia Lalmas
Yahoo! Labs
Barcelona, Spain
mounia@acm.org

Ethan Zuckerman
MIT Center for
Civic Media
Cambridge, MA, USA
ethanz@media.mit.edu

ABSTRACT

Users interact with online news in many ways, one of them being *sharing* content through online social networking sites such as Twitter. There is a small but important group of users that devote a substantial amount of effort and care to this activity. These users monitor a large variety of sources on a topic or around a story, carefully select interesting material on this topic, and disseminate it to an interested audience ranging from thousands to millions. These users are *news curators*, and are the main subject of study of this paper. We adopt the perspective of a journalist or news editor who wants to discover news curators among the audience engaged with a news site.

We look at the users who shared a news story on Twitter and attempt to identify news curators who may provide more information related to that story. In this paper we describe how to find this specific class of curators, which we refer to as *news story curators*. Hence, we proceed to compute a set of features for each user, and demonstrate that they can be used to automatically find relevant curators among the audience of two large news organizations.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*; H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

Social media; journalism; digital curator; news story; automatic learning

1. INTRODUCTION

Online social media platforms have become a powerful tool for the aggregation and consumption of time-sensitive content in general and news in particular. In this study we focus on one of such platforms, Twitter, which as of 2013 counts with more than 200 million active users posting over 400 million tweets per day.[1] Many Twitter users follow news sources and share some of the news articles published by those sources, in some cases commenting or expressing their own opinion about the news [12].

A recent survey on 613 journalists over 16 countries reveals that 54% of them used online social media platforms (Twitter, Facebook and others) and 44% used blogs they already knew to elicit new story angles or verify stories they work on [19]. It has been argued that journalists come together as a community to make sense of events and interpret their importance. The increasing engagement of online users with journalists and news content, might indicate that some individual users of online news have the expectation of being considered as part of that interpretive community [28]. The purpose of our research is to study, from the perspective of journalists, editors or content producers, to what extent the community of engaged readers of a medium – those who share news articles of that medium in social media – could contribute to the journalistic process.

In our previous work we study these communities, referring to them as *transient news crowds* [15], in analogy with the group of passers-by that gathers around unexpected events such as accidents in a busy street. We noticed that news crowds are made of different people playing different roles. For instance, as in many online communities, a minority is actively engaged (posting many articles), while the majority remains largely silent [17].

In addition to differences in the intensity of their participation, there are a set of roles that people play when sharing online news in social media. We focus on the roles of *news aggregator* and *news curator*, and in particular in a sub-type of news curator that we deem as *news story curator*.

Our contributions are the following:

- We define some of the roles people play in news crowds (Section 2).
- We study data from two large international news organizations (Section 3) and characterize these roles (Section 4).
- We show that to some extent we can statistically model these roles (Section 5).

The remainder of this paper is organized as follows. Section 2 introduces a set of user classes, which are studied by hand-coding a sample from the dataset described in Section 3, and obtaining the results shown in Section 4. These labels are then used to build automatic news story curator detectors in Section 5. Section 6 outlines related works. Section 7 presents our conclusions.

[1] http://blog.twitter.com/2013/03/celebrating-twitter7.html

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

2. CONCEPTS

Digital curation is a broad field concerned with the management and preservation of digital data, specially considering future re-use [27]. We focus on the role of *online content curator*, which has been defined as "someone who continually finds, groups, organizes and shares the best and most relevant content on a specific issue online." [4] In this paper, we focus on content curated via Twitter.

News story curators. We are particularly interested in online content curators who follow a developing news story, which we call *news story curators*. A famous example for this type of news curator is Andy Carvin (`@acarvin`), who mostly collects news related to the Arabic world, and became famous for his curatorial work during the Arab Spring [6]. As a news story curator for the Arab Spring, he aggregated reports in real time and tweeted sometimes thousands of tweets per day.

To find candidates for news story curators, we look at the *news crowd of an article* [15] which is the set of users who shared the article's URL in Twitter. We consider that among the users who share an article belonging to a developing story, some of them may follow-up with further tweets. In other words, given a news article and its crowd of users, our goal is to identify which of those users can be suitable curators for the story the article belongs to.

A manual analysis of our data and the characterisitcs of Twitter curators in general revealed different types of curators and we assume that these types are important for our work. We present next the dimensions in which curators can be divided.

Topic-focused/unfocused. We observed two types of users that are intensely engaged with news content in social media. We call them *focused curators* and *unfocused curators*.

A *topic-unfocused curator* is a user that collects contents about diverse topics, becoming a news provider of sorts, disseminating news articles about breaking news and top stories. For instance, `@KeriJSmith`,[2] a self-defined "internet marketer" tweets about various interesting news on broad topics.

A *topic-focused curator* is a more selective user, who collects interesting information with a specific focus. This focus is usually a geographic region or a topic. For instance, `@chanadbh` tweets about news related to Bahrain, whereas `@brainpicker` collects contents about art and design. Topic-focused curators play a pivotal role in the filtering and dissemination of news, and constitute a first line of defense against information overload [22].

With/without user commentary. The way in which different users curate content varies substantially. In most cases, users include links in their tweets to the content they have found. Sometimes, they also provide personal comments and opinions, using Twitter as both a news service and a social network [14]. For instance, `@DruidSmith` is a geolocation/GPS expert who, aside from linking to content from other sources, also shares his own knowledge and experience.

Human/automatic. In Twitter there is a significant amount of *news aggregators*. They collect news articles (e.g. from RSS feeds) and automatically post their corresponding headlines and URLs to Twitter. The majority of them post many tweets per day related to breaking news and top stories, e.g. `@BreakingNews`. A minority are focused on more specific topics, and thus constitute *topic-focused aggregators*. In all cases, they do not provide any personal contents or opinions.

For instance, `@RobinGood` is a widely recognized curator on the topics of media, technology and design. However, `@RobinGood` maintains a personal selection of blog and directories of RSS, which he updates weekly. His account distributes automatically the stories that appear in those sources. Thus, all of his tweets are generated automatically.

Some news aggregators seem to be considered valuable by users, as in the case of `@RobinGood` who has over 17,000 followers as the time of this writing. However, whether all news aggregators provide interesting content to a topic is questionable.

In summary, there are different types of users that aggregate content. They differ in the number of topics they cover (focused/unfocused), in how much commentary they include (only URLs or URLs and comments) and in the way they post information (human/automatic).

These insights allow us to make a first characterization of news story curators. Tweeting about a story but also about many other stories that are not related (e.g. not the same topic or geographic region) indicates that the user is not interested in the story per say, thus, should be considered as a curator for it. Moreover, we are not interested in finding mere news aggregators, who automatically post content from a set of sources, sometimes using automatic filters. Instead, we look for news curators, where there is human intelligence behind the choice of each individual article. As a consequence, we distinguish between human and automatic tweet creation in this paper. The next section describes the datasets used to bring these concepts to practice, in order to find concrete instances of news story curators.

Table 1: General characteristics of our datasets: number of articles, users, and tweets.

		Users		Tweets	
Dataset	Articles	Total	Per crowd	Total	Per crowd
BBC	75	13.3K	177	35.5K	201
AJE	155	8.3K	53	24.0K	154

3. DATASETS

We collected news articles published in early 2013 on two major online news websites, BBC World Service (BBC) and Al Jazeera English (AJE). We downloaded periodically the homepage of each of these websites, from which we sampled at random a subset of news articles. We focused on the headline news, e.g. opinions, magazine and sport news were not included. The sampled articles cover a variety of stories such as Obama's inauguration, the conflict in Mali, and the pollution in Beijing. Table 1 summarizes the main characteristics of our datasets.

For each of the sampled articles, we started a process that used Twitter's API[3] to periodically find tweets containing that article's URL. The earliest tweets followed almost immediately the publication of the article, as each of these

[2]In this paper we follow Twitter's convention of prefixing usernames with an "@" sign.

[3]http://dev.twitter.com/

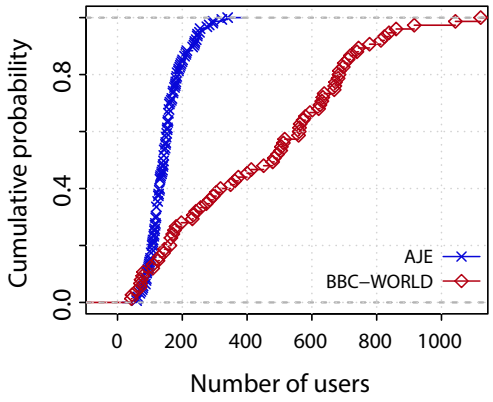

Figure 1: Distributions of crowd sizes, i.e. number of candidate news story curators.

news organizations disseminate their content via their own twitter account(s) (e.g. `@BBCWorld`, `@AJEnglish`).

Spam filtering. In Twitter there is a substantial amount of spam. Therefore, as an heuristic informed by previous works [3, 25], we removed accounts having a clearly anomalous behavior. We tried to keep our intervention to a minimum, and thus discarded only tweets falling in the top 5% of the distribution of a spam-related variable. We discarded accounts having more than 98 tweets per day, more than 90% of retweets, or more than 92% of tweets containing URLs. We also examined manually about 100 of the most prolific accounts and defined a blacklist of high-throughput automatic accounts that do not focus on any particular region, topic, or news provider.

News crowds extraction. The news crowd of an article consists of all the users who tweeted its URL within the first 6 hours after publication [15]. This encompasses about 90% of the tweets an article receives in the datasets we use (87% for BBC and 91% for AJE).

Finally, we excluded articles that did not generate a significant response in Twitter or that generated an extremely high one. Specifically, articles with very small crowds (lower 15% of the distribution) or very large ones (upper 5% of the distribution) were excluded. We kept articles with 50–150 users for BBC news articles and 70–360 users for AJE. We remark that the distribution of users per crowd is very skewed, as shown in Figure 1. Also, the crowds around articles in AJE are smaller than the ones in BBC, following the different sizes of the audiences of these websites.

4. LABELING NEWS STORY CURATORS

We apply the definitions from Section 2 to the dataset described in Section 3. In an automatic system, we would apply supervised learning to detect story curators, and thus a training set (human-provided labels) is required. As we shall see in this section, there are a number of issues that have to be dealt with carefully in order to obtain the labels.

Example. An example helps understand our labeling process. We selected two articles from our data set and looked at their crowds, i.e. the users who posted these articles to their Twitter timelines. Table 2 lists some of these users and their characteristics. We focus on two characteristics

Table 2: Example of users for two news articles. We include the number of followers, tweets per day, fraction of tweets containing URLs and user mentions ("@"), the type of tweet generation and the main topic.

	Foll.	Tweets /day	Fraction URL	@	Type	Topic
16 Jan 2013 – Syria allows UN to step up food aid						
`@RevolutionSyria`	88122	189.13	0.86	0.02	Auto.	Syria
`@KenanFreeSyria`	13388	9.29	0.74	0.28	Human	Syria
`@UP_food`	703	10.22	1.00	0.00	Auto.	Food
18 Jan 2013 – US cyclist Lance Armstrong admits to doping						
`@KevinMcCallum`	15287	60.15	0.18	0.77	Human	Sports
`@huyanxing`	3224	69.19	1.00	0.00	Auto.	Misc.
`@WaseemMansour`	1298	15.33	1.00	0.00	Auto.	Misc.

of them: 1) whether they seem *human or automatic*, which separates news aggregators and curators, and 2) whether they seem to be *interested in the topic of the article or not*, which describes the topical focus of a user.

In Table 2 the first article is about the civil war in Syria. Two of the users who posted this story have several tweets related to Syria: `@RevolutionSyria` provides automatically generated tweets, whereas the content twitted by `@KenanFreeSyria` is collected by hand. We can see this by looking at the number of tweets per day for `@RevolutionSyria` and the fact that it has almost no user mentions (it does not engage in conversations with others). The user `@UP_food` is an news aggregator that apparently tweets anything containing the word "food", but is not relevant for the developing story about Syria. The second article is on the doping scandal of Lance Armstrong. We could detect one curator for this story, `@KevinMcCallum`, who routinely collects sports-related content. The other users in the crowd aggregated breaking and top news in an automatic manner (e.g. `@huyanxing` and `@WaseemMansour`).

Pre-filtering. We define two criteria to reduce the number of users under consideration (i.e. potential story curators). First, we examine only users with at least 1,000 followers, as users with less followers are not influential enough to play a significant role in the Twitter ecosystem [23]. We also apply the method from [15] to detect when an URL posted by a news crowd member is related to the original story that created that crowd. We consider only users whom according to that method posted at least one URL related to the original news article.

Labeling process. We created our training data by selecting a sample of 20 news articles: 10 from AJE, and 10 from BBC. For each news article, we sampled uniformly at random 10 users who posted the article. We then asked three volunteers to provide labels.[4] We provided them examples and typical characteristics of the various types of news aggregator and curator (as discussed in Section 2).

For the labeling task, we showed the title of the news article and a sample of tweets of the user. We showed tweets that were posted directly after the news article, as the lifetime of some stories can be very short. We also presented the profile description and the number of followers of the user.

[4]All three are computer science or engineering graduates with experience in using Twitter.

Then, we asked our annotators to label the user according to the following instructions:

> *You will be presented with the title of a news article, and tweets and profile information of a Twitter user.*
> *Q1) Please indicate whether the user is interested or an expert of the topic of the article story:*
> - *Yes: Most of her/his tweets relate to the topic of the story (e.g. the article is about the conflict in Syria, she/he is often tweeting about the conflict in Syria).*
> - *Maybe: Many of her/his tweets relate to the topic of the story or she/he is interested in a related topic (e.g. the article is about the conflict in Syria, she/he is tweeting about armed conflicts or the Arabic world).*
> - *No: She/he is not tweeting about the topic of the story.*
> - *Unknown: Based on the information of the user it was not possible to label her/him.*

> *Q2) Please indicate whether the user is a human or generates tweets automatically:*
> - *Human: The user has conversations and personal comments in his tweets. The text of tweets that have URLs (e.g. to news articles) can be self-written and contain own opinions.*
> - *Maybe automatic: The Twitter user has characteristics of an automatic profile, but she/he could be human as well.*
> - *Automatic: The tweet stream of the user looks automatically generated. The tweets contain only headlines and URLs of news articles.*
> - *Unknown: Based on the information of the user it was not possible to label her/him as human or automatic.*

The label "unknown" corresponds to the case where the annotators were not able to reach a decision. Possible reasons were the language of the tweets (e.g. the user is tweeting in Chinese). In total, 417 labels were collected. We decided to label whether a user is interested in the topic of the news story or not, instead of asking whether the user is an expert of that topic. We assume that users that are not experts (e.g. eye-witnesses), but interested in the story, could reveal interesting information for journalists.

For the training set, we considered only users for which at least two annotators provided a decisive label (Yes or No, Human or Automatic). We discarded any "maybe", "maybe automatic", and "unknown" labels, as these users could be used neither for training nor evaluation purposes. The distribution of labels is shown in Table 3. While 13% of the AJE users were labeled as both interested in the topic and human, only 1.8% of them had both labels in the case of BBC.

5. AUTOMATICALLY FINDING NEWS STORY CURATORS

In this section we present our approach for automatically finding news curators. We first describe our learning framework and the features we use, then we present the results based on the training set described in Section 4.

Table 3: Distributions of the human-provided labels.

	Interested?			Human or Automatic?		
Dataset	n	`yes`	`not`	n	`human`	`automatic`
AJE	63	21%	79%	71	55%	45%
BBC	58	3%	97%	54	35%	65%

5.1 Learning Framework

We defined two tasks: the first one detects users that are interested in the given story or topics associated with the article (*UserIsInterestedInStory*), and the second one evaluates whether the user is human or generates its tweets automatically (*UserIsHuman*). As discussed in the previous section, we consider users as potential curators only if they have at least 1,000 followers and posted at least one URL related (according to the method in [15]) to the original news article. We use standard evaluation metrics such precision, recall, and AUC, all measured after ten-fold cross validation.

5.2 Features

Previous work including [10, 24, 26] provides us with some useful information about suitable features for the detection of curators. These include network-based features such as the number of followers of a user – shown not to be sufficient on its own as a predictor of expertise by [26] – as well as contextual features including user profile and user lists [24]. Our features try to capture three aspects of users: (1) the *visibility* of a user; (2) characteristics of the user's tweets that might separate human from automatic users; and (3) how focused are the tweets of users with respect to the news media source. We transformed the frequency-based values to provider-specific quantile values in the filtered dataset, as we are merging users coming from two different news providers whose audiences have different sizes, as we showed in Figure 1. These features are denoted by the suffix Q in the feature name.

Visibility. The visibility of a Twitter user is a core characteristic of a curator. There are different features that can be associated with a user visibility. This can be captured by the number of followers (*UserFollowersQ*) or the number of Twitter lists containing that user (*UserListedQ*). We remark that both features are highly correlated in our data ($r^2 = 0.96$), which is consistent with the findings of Sharma et al. [23]. However, we do not know a priori if one of the features is more important than the other in any of the two classification tasks that we attempt.

Tweeting activity. In Section 4 we described the presence of prolific automatic accounts in Twitter. Features that capture the tweeting activity of a user may reflect best the differences between human and automatic accounts. We measure the number of tweets per day (*UserTweetsDailyQ*), the fraction of tweets that contains a re-tweet mark "RT", a URL, a user mention or a hashtag (respectively, *UserFracRetweets*, *UserFracURL*, *UserFracMention*, and *UserFracHashtag*).

Topic focus. A natural measure of topical focus is how many different articles in each dataset has this user tweeted (*UserCrowdsQ*). Additionally, as articles belong to a section in each website (e.g. sports, business, Europe, USA), we also

define the number of distinct sections of the crowds s/he belongs to (*UserSectionsQ*).[5]

5.3 Results

We tried two types of models, one considering only a single input feature, and one considering all the input features. We remark that in all cases there are two implicit rules: the user must have at least 1,000 followers and the user must have posted an article that is likely to be related to the original article.

Simple models. For the task *UserIsHuman*, a basic but effective approach is to apply a simple rule, which yields a precision and recall of 0.85 (for the `human` class), and an AUC of 0.81:

UserFracURL >= 0.85 ⇒ `automatic`, otherwise `human`.

This means that a news aggregator (automatic user) can be readily identified because a large fraction of its tweets contain URLs. This agrees with previous works (e.g. [3]) and our manual analysis of the data in Section 2.

For the task *UserIsInterestedInStory*, the following rule yields a precision of 0.48 (remember the classes are not balanced), a recall of 0.93 (for the `interested` class), and an AUC of 0.83:

UserSectionsQ >= 0.9 ⇒ `not-interested`, otherwise `interested`

This means that if a user does not tweet about many different sections of a news site, one can expect that the sections s/he is tweeting about relate to the topic of the story, thus, s/he is interested in the story of the given article. However, it is not always the case as we can see in Table 2: the curator `@UP_food` collects tweets around the topic "food", which is not relevant for the story about Syria.

Complex models. More complex models, in our case, a random forest after information-gain-based feature selection, as implemented in Weka,[6] perform better for the *UserIsHuman* task (AUC 0.93 vs AUC of 0.85 for the single-feature model). As expected, all features related to the tweeting activity (*UserFracRetweets*, *UserFracURL*, *UserFracMention*, and *UserFracHashtag*) are the most important features for this model.

Adding more features to the model for the *UserIsInterestedInStory* task also yields an improvement in performance when comparing with the single-feature model (AUC 0.90 vs AUC 0.83), and might also improve the robustness of the prediction. Given a large class imbalance for the *UserIsInterestedInStory* task, we applied asymmetric misclassification costs. Specifically, false negatives (classifying an interested user as not interested) were considered 5 times more costly than false positives; values close to this number did not change substantially the obtained results.

All results are summarized in Table 4. Overall, we were able to demonstrate that the considered features can be used to automatically find news (story) curators among the audience of two large news organizations, BBC and AJE. Note that, as shown in Section 3, there is a difference in the sizes of the audiences of these websites (Figure 1); nonetheless, we could identify story news curators for both.

Table 4: Evaluation of models for the *UserIsHuman* and *UserIsInterestedInStory* tasks.

	Precision	Recall	AUC
`automatic`	0.88	0.84	0.93
`human`	0.82	0.86	0.93
`interested`	0.95	0.92	0.90
`not-interested`	0.53	0.67	0.90

5.4 Precision-oriented evaluation

We also compared our method with two baseline approaches: (1) select the users with the largest number of followers among the candidates, and (2) select the users with the largest number of stories detected as related to the original one (using the system in [15]).

Data. We selected a sample of 20 news articles that had at least one curator, detected using the model that uses all the features with a confidence value >= 0.75. For comparison, we extracted for each article the same number of possible curators using the other two approaches. Then, we merged the results together without providing which system identified which curator. We asked the same three assessors to evaluate the results using the question (Q1) of Section 4.

Results. We collected about 210 labels for 70 units. The assessors labeled 71% of the users as *not interested*, 6% as *interested*, and 15% as *maybe interested*. We merged the labels *yes* and *maybe*, and we considered only users for which at least two assessors had the same label. As a consequence an unequal number of labels per approach is given. The worst performance was obtained by the follower-based approach (2/18 = 11%): only two users with a high number of followers were labeled as curators and 18 users with a high number of followers were not curators.

A better performance was obtained by the automatic detection of related stories (5/20 = 25%), but our approach outperformed the other two (6/16 = 37.5%).

6. RELATED WORK

The availability of text collections in which authorship can be identified has generated a significant amount of activity around the topic of expert finding. The first approaches to expert finding where either text-based [7] or network-based [1]. Both paradigms have evolved over the years, and recent approaches combine them e.g. [8].

Expert detection in Twitter has become an active research topic [13], especially with the increased usage of Twitter as a service for news aggregation and consumption. Twitter experts, called curators, collect high valuable and informative news and other content around a topic. They also are known to identify interesting news (that end up becoming popular) earlier on [11].

The detection of curators is a difficult challenge mainly caused by the dynamic nature of Twitter. For instance, Pal et al. [20] argued that network-based features (e.g. follower-network) are not always suitable because the lifetime of a topic can be very short. In addition, users are followed for

[5]The section of an article can be extracted from the prefix of the path of the article. For instance, articles under http://www.bbc.co.uk/news/world-latin-america-21001060 correspond to the section "Latin America" of BBC. In websites organized in a different manner, other ways of defining sections may be necessary.

[6]http://www.cs.waikato.ac.nz/ml/weka/

other reasons than their topical expertise, thus reducing the effectiveness of network-based features to detect experts in Twitter [26]. Network features based on the retweet network in combination with content features were shown to better reflect the dynamic nature of Twitter and as such more suitable for the task of detecting experts [13].

Contextual data (as contained in the user profile including the user name) have to be used carefully, as user studies showed that the name of the user may bias the judgment. Indeed, Pal et al. [20] demonstrated that people rank topic-related tweets from celebrities as more interesting and authoritative than when the same information is tweeted by non-celebrity users. On the other hand, contextual data provides useful information such as the lists a user belongs to [16, 24, 10]. User lists are a widely used Twitter feature that allows users to group other users, for instance, around a same topic. Wagner et al. [24] demonstrated that features based on the user lists perform best, compared to content-based features based on recent tweet/retweets and features based on the profile. Approaches to automatically extend or create these lists exist as well [5]. Nowadays, these lists are also used to filter valuable content for journalists. For instance, *Storyful* is a news agency that provides user lists, developed by journalists for journalists.

Our approach is different because our aim is not to develop a new expert detection approach for Twitter. We took the perspective of journalists and editors who are interested in understanding the users who are tweeting their articles, and want to detect those users that could provide further content to the story of the news article. We referred to these users as news story curators.

Twitter has been used as a source of news recommendations, typically by exploiting Twitter-specific features extracted from post-read social responses [2, 9, 18, 21], tweets content (hashtags, topics, entities), users followees and followers, public timelines and retweeting behavior. However these works aim at building personalized recommender systems; their aim is to suggest news articles based on the inferred topical interests of a user. Our objective is entirely different; we want to provide journalists and editors a tool that recommends them news story curators whom they may wish to follow, as these curators may provide content that complements or extends the one that they have produced.

7. CONCLUSIONS

In this work we have defined and modeled a class of users, news story curators, that has the potential to play an important role in the news ecosystem, particularly for journalists, editors, and readers.

We have found that finding news story curators is a challenging task. First, there is a large amount of automatic activity on Twitter, and some of these news aggregators are actually considered by some users to be good curators. Second, posting a link in Twitter may or may not reflect a long-standing interest on the subject of the link.

In our approach, we have automatically found news story curators. A key aspect of that system is being able to assess how spread are the interests of a user. This matched our intuitions in the sense that the more diverse a user's interests are, the less likely that person is to be a good news curator.

Next, we have tackled this problem by trying to separate automatically-operated accounts from manually-operated ones, showing that while simple rules can be somewhat effective, combining different aspects of the information available about a user can yield better results.

Future work. There are several directions in which the current work can be extended. First, besides topicality other variables can be incorporated, e.g. interestingness and serendipity. Second, a user-centered method could be developed in which the input is a user and the output is a set of news curators. In both cases, incorporating other types of information including e.g. content-based features from lists [10] could improve the performance of these systems.

The results can be extended to reveal a better understanding of the curator ecosystem in Twitter. The application of the framework on other news provider could provide further insights and highlight differences depending on the source of the news article. An open question is whether news aggregators, which are comparable to RSS feeds, possess any kind of functionality in Twitter, especially, if they have a large audience and if their interests are focused around a few topics.

Data availability. The data used in this study is available upon request for research purposes.

8. ACKNOWLEDGMENTS

This work was partially funded by Grant TIN2009-14560-C03-01 of the Ministry of Science and Innovation of Spain. This work was carried out as part of Janette Lehmann internship at QCRI. The authors would like to thank Noora Al Emadi, Diego Sáez Trumper and Hemant Purohit for the labelling tasks and Mohammed El-Haddad and Nasir Khan from Al Jazeera for insightful discussions and comments.

Key references: [15, 22]

9. REFERENCES

[1] M. Abrol, U. Mahadevan, K. McCracken, R. Mukherjee, and P. Raghavan. Social networks for enterprise webs. In *Proc. of WWW*, 2002.

[2] D. Agarwal, B.-C. Chen, and X. Wang. Multi-faceted ranking of news articles using post-read actions. *CoRR*, abs/1205.0591, 2012.

[3] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on twitter. In *Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, July 2010.

[4] R. Bhargava. Manifesto for the content curator: The next big social media job of the future? http://www.rohitbhargava.com/2009/09/manifesto-for-the-content-curator-the-next-big-social-media-job-of-the-future-.html, September 2009.

[5] I. Brigadir, D. Greene, and P. Cunningham. A system for twitter user list curation. In *RecSys*, pages 293–294, 2012.

[6] A. Carvin. *Distant Witness*. CUNY Journalism Press, 2013.

[7] N. Craswell, D. Hawking, A.-M. Vercoustre, and P. Wilkins. P@NOPTIC expert: Searching for experts not just for documents. In *Proc. of AusWeb*, 2001.

[8] H. Deng, I. King, and M. R. Lyu. Formal models for expert finding on DBLP bibliography data. *Data Mining, IEEE International Conference on*, pages 163–172, 2008.

[9] Q. Gao, F. Abel, G.-J. Houben, and K. Tao. Interweaving trend and user modeling for personalized news recommendation. In *Web Intelligence*, pages 100–103, 2011.

[10] S. Ghosh, N. K. Sharma, F. Benevenuto, N. Ganguly, and P. K. Gummadi. Cognos: crowdsourcing search for topic experts in microblogs. In *SIGIR*, pages 575–590, 2012.

[11] C. Hsieh, C. Moghbel, J. Fang, and J. Cho. Experts vs the crowd: Examining popular news prediction perfomance on twitter. In *WSDM, 2013*, 2013.

[12] A. Java, X. Song, T. Finin, and B. L. Tseng. Why we twitter: An analysis of a microblogging community. In *WebKDD/SNA-KDD*, pages 118–138, 2007.

[13] S. Kong and L. Feng. A tweet-centric approach for topic-specific author ranking in micro-blog. In *ADMA*, pages 138–151, 2011.

[14] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *World Wide Web Conference*. ACM Press, 2010.

[15] J. Lehmann, C. Castillo, M. Lalmas, and E. Zuckerman. Transient news crowds in social media. In *ICWSM*, 2013.

[16] Q. V. Liao, C. Wagner, P. Pirolli, and W.-T. Fu. Understanding experts' and novices' expertise judgment of twitter users. In *CHI*, pages 2461–2464, 2012.

[17] B. McConnell and J. Huba. The 1% rule: Charting citizen participation. *Church of the Customer Blog*, 2006.

[18] G. D. F. Morales, A. Gionis, and C. Lucchese. From chatter to headlines: harnessing the real-time web for personalized news recommendation. In *WSDM*, pages 153–162, 2012.

[19] Oriella PR Network. The influence game: How news is sources and managed today. http://www.oriellaprnetwork.com/sites/default/files/research/Oriella%20Digital%20Journalism%20Study%202012%20Final%20US.pdf, 2012.

[20] A. Pal and S. Counts. Identifying topical authorities in microblogs. In *WSDM*, pages 45–54, 2011.

[21] O. Phelan, K. McCarthy, M. Bennett, and B. Smyth. Terms of a feather: Content-based news recommendation and discovery using twitter. In *ECIR*, pages 448–459, 2011.

[22] D. Sasaki. Our friends become curators of Twitter-based news. http://www.pbs.org/idealab/2010/04/our-friends-become-curators-of-twitter-based-news092.html, Apr. 2010.

[23] N. K. Sharma, S. Ghosh, F. Benevenuto, N. Ganguly, and K. Gummadi. Inferring who-is-who in the twitter social network. In *WOSN*, pages 55–60, 2012.

[24] C. Wagner, V. Liao, P. Pirolli, L. Nelson, and M. Strohmaier. It's not in their tweets: Modeling topical expertise of twitter users. In *SocialCom/PASSAT*, pages 91–100, 2012.

[25] A. H. Wang. Don't follow me: Spam detection in twitter. In *Proceedings of the International Conference on Security and Cryptography (SECRYPT)*, July 2010.

[26] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *WSDM*, pages 261–270, 2010.

[27] E. Yakel. Digital curation. *OCLC Systems and Services*, 23(4):335–340, 2007.

[28] B. Zelizer. Journalists as interpretive communities. *Critical Studies in Mass Communication*, 10(3):219–237, 1993.