Computer Communications 000 (2015) 1–7

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

Retweeting beyond expectation: Inferring interestingness in Twitter

William M. Webberley*, Stuart M. Allen, Roger M. Whitaker

School of Computer Science & Informatics, Cardiff University, 5 The Parade, Cardiff, CF24 3AA, UK

ARTICLE INFO

Article history:
Available online xxx

Keywords:
Twitter
Interestingness
Retweet

ABSTRACT

Online social networks such as Twitter have emerged as an important mechanism for individuals to share information and post user generated content. However, filtering interesting content from the large volume of messages received through Twitter places a significant cognitive burden on users. Motivated by this problem, we develop a new automated mechanism to detect personalised interestingness, and investigate this for Twitter. Instead of undertaking semantic content analysis and matching of tweets, our approach considers the human response to content, in terms of whether the content is sufficiently stimulating to get repeatedly chosen by users for forwarding (retweeting). This approach involves machine learning against features that are relevant to a particular user and their network, to obtain an expected level of retweeting for a user and a tweet. Tweets observed to be above this expected level are classified as interesting. We implement the approach in Twitter and evaluate it using comparative human tweet assessment in two forms: through aggregated assessment using Mechanical Turk, and through a web-based experiment for Twitter users. The results provide confidence that the approach is effective in identifying the more interesting tweets from a user's timeline. This has important implications for reduction of cognitive burden: the results show that timelines can be considerably shortened while maintaining a high degree of confidence that more interesting tweets will be retained. In conclusion we discuss how the technique could be applied to mitigate possible filter bubble effects.

1. Introduction

Microblogging services, with Twitter as a prime example, have facilitated a massive interconnection of the world over the past few years [1]. Twitter's support of quick, short, and 'real-time' live content sharing amongst its millions of users allow vast amounts of information to be sent and received very quickly [2]. This has been helped by its growth into the mobile domain, allowing users to share text, photos, or videos directly from a news source or geographic location [3]. It has been especially useful in emergency situations worldwide, such as during the 2010 Haiti earthquake [4] and the 2011 Egyptian protests [5].

Unlike many other media, microblogging services such as Twitter are characterised by convenience and informality - messages are size-limited, making them easy to consume, and may contain pictures and pointers to other web content. The streamed nature of tweets provides channels defined by other Twitter users, where users opt-in to receive content. These subscription relationships provide a social network structure through which content is mediated, with users being able to republish or "retweet" received messages as they wish. However the ease with which content can be published results in a huge volume of potential content, much of this having limited relevance other than to a few users. A user's ability to choose whose content they receive counters to some degree the "long-tail" problem of social media content [6]. However this can introduce noise, where the likelihood of generally uninteresting and mundane content begins to outweigh interesting content [7].

These issues mean that approaches to distinguishing interesting tweets from surrounding noise are valuable in reducing the cognitive burden for users. Identifying interesting tweets represents a form of recommendation system and there are a range of well-known strategies that can be adopted. However, the real-time nature of microblogging combined with limited text from which knowledge can be extracted, means that it is appropriate to look for new and efficient alternative approaches.

In this paper we introduce, formally define and explore a new strategy to quantify the perceived "interestingness" of individual tweets. A brief initial exploration of the underlying approach was presented in [8] as a proof of concept. In comparison, this paper provides a complete specification of the model and necessary implementation details, formally validates the approach against collective and individual assessments of interestingness provided by human participants in a web experiment, and analyses the results to draw conclusions on potential applications.

* Corresponding author. Tel.: +44 29 2087 4812.
E-mail addresses: WebberleyWM@cardiff.ac.uk (W.M. Webberley), AllenSM@cardiff.ac.uk (S.M. Allen), WhitakerRM@cardiff.ac.uk (R.M. Whitaker).

http://dx.doi.org/10.1016/j.comcom.2015.07.016

Our conceptual model interprets the human users of Twitter as collective cognitive agents, who effectively process the semantic content of tweets and respond to cognitive stimulation [9] and cues [10]. When this stimulation is perceived as suitably significant, the agent forwards the message to its neighbours and the process is repeated. The networked nature of Twitter means that message forwarding (or retweeting), when considered in context of the agents and the network structure, holds potentially valuable accumulated perceptions about the quality, relevance and interest of the content. This represents an implicit form of crowdsourcing [11].

A significant benefit of this approach is its efficiency. The human performs sophisticated computation and artificial intelligence can be applied to their subsequent retweeting behaviour, rather than being applied directly to analysis of tweet content. By applying suitable thresholds, a deterministic measure of "interestingness" can be applied both to *filter* content streams from individuals and to *discover* content from outside of the immediate social network.

The remainder of the paper is structured as follows: in Section 2 we identify the key related work; Section 3 introduces the measure of interestingness, as a general metric to capture the notion of interest, beyond expectation, to a significant sub-group of Twitter users; Section 3.2 describes the application of machine learning techniques to characterise retweet behaviour based on selected features; Section 4 involves validating the interestingness metric as a technique and benchmarking it against human selection of interesting content.

2. Related work

Classifying a tweet as "interesting" is challenging since it is a personal and subjective judgement, often relying on the context of surrounding information as to the emotions that might be triggered. In addition, "interesting" content is not simply that which provides positive enjoyment, since content that conveys anger or frustration may also be of interest. Therefore we consider interesting content to be that giving rise to significant affective stimulation [12] for some group of users. An individual who retweets content is signalling that they believe it will have a level of affective stimulation for their followers.

There has been considerable useful research into retweet behaviour, including analyses of retweet propagation [13], retweet volume prediction for a variety of purposes [14–17], and binary retweet decision-making [18–21]. In [22] an approach is discussed for recommending *users* to follow, and the paper provides some useful information on feature selection for training classifiers, identifying mentions, URLs, and hashtags as important. [23] surveys various recommender systems built on Twitter, highlighting only one example [9] which filters a timeline of tweets by predicting their retweet level. More generally, [24] surveys how information is diffused through the Twitter social graph. In the remainder of this section, we highlight a subset of the most relevant works towards identifying interest.

There has been little effort in the literature to explicitly identify "interesting" tweets that may be relevant beyond their immediate audience. [25] attempts to find interesting tweets by analysis the social graph. The HITS algorithm [26] is applied to first filter based on the influence of the author, before a similar process then scores individual tweets. Results are evaluated against a gold standard of tweets identified by two human annotators. [19] uses machine learning techniques to predict the number of retweets using features that range from simple (e.g., includes a hashtag or URL) to more complex including sentiment analysis and term extraction but ignore features relating to the author. It is inferred that tweets that are predicted to be retweeted often, are inherently more interesting. Specific terms play a strong role in their predictions, where tweets containing the term *social* are predicted to be retweeted more often than those containing *sleep* for example.

The concept of "interestingness" is also addressed in research domains outside of social media analysis, particularly with respect to data mining, where [27] surveys several approaches for measuring interestingness in this area. The authors assess metrics such as peculiarity, surprisingness, generality, and diversity for semantically deducing how interesting a piece of data is. [28] uses features, such as validity, novelty, and understandability for a similar aim. *Relative* interest is a term introduced in [29], which uses "common sense knowledge" to mine rules that contradict a user's knowledge, and thus describe relatively interesting information as that which differs from the norm. Affectiveness from stories (including those inducing suspense), news articles, and events was shown to drive interest in information by [30], who also report that increases in interest affect the cognitive application to the information. Recall and learning capacity are also improved as a result. Finally, [31] measures interestingness of mined patterns based on whether or not the data is unexpected or actionable to a user, where information is interesting to a user if it is of use or if it contrasts with belief.

Semantic analysis has been a commonly used approach in many studies. For example, linear regression is used in [32] to score the different components of a tweet's text to produce an average tweet score, allowing users to write tweets that will more likely receive retweets[1]. An estimated retweet count is then obtained through a comparison to a "baseline" score for the tweet's author at that point in time. A feature of this method is that it requires building and continual updating of each user's baseline and links are not make links to information interestingness.

In [7] semantic analysis of tweets is again used, in this case to produce scores to identify *uninteresting* content. A decision tree classifier is used to assign integer scores [1, 5]. However, the categorisation system the authors eventually use is relatively coarse and not able to represent the many types of tweets seen on Twitter. The classification of interestingness involves identifying when a tweet contains a URL, which prohibits a significant amount of potentially interesting content. This means the methods aren't suitable for assessing tweets on a general or user-specific level.

"LiveTweet"[2] is a system introduced in [33,34] for determining interestingness through retweet probability, using a model containing information on features of tweets most popular at a given point in time. The method requires a continual re-building of the semantic model. The authors state that a retweeted tweet is not necessarily an indication of interestingness, due to user influence and temporal factors, but that a single retweet decision does imply that *user's* interest. In [35] information quality is as the driver for the development of a clustering algorithm. However, the scoring method is relatively simplistic and based around identifying the most important tweets surrounding a particular event (such as Michael Jackson's death). This may prohibit other forms of interestingness that don't relate to a specific event.

In summary, from the existing literature there is considerable scope to develop techniques that are: (i) efficient in requiring the use of resources; (ii) generic in capturing interestingness which may arise from diverse sources, for example not necessarily defined by a event or by the inclusion of a web link; (iii) effective in providing some personalisation. These observations have motivated our alternative approach.

3. Inferring interestingness

We assess *interestingness* for a tweet by considering the extent to which it has provided affective stimulation [12] the group of users that have encountered it in their timeline. The signal we use for affective stimulation is a retweet. Although retweeting is a simple cue,

[1] https://sites.google.com/site/learningtweetvalue/home
[2] Available at: http://livetweet.west.uni-koblenz.de

Please cite this article as: W.M. Webberley et al., Retweeting beyond expectation: Inferring interestingness in Twitter, Computer Communications (2015), http://dx.doi.org/10.1016/j.comcom.2015.07.016

Table 1
Tweet (left) and user (right) features used to train and test against the classifier. The nominal feature 'retweet count' is the predictor feature.

Feature	Data type	Feature	Data type
Contains mention	{True, False}	Follower count	Real (numeric)
Tweet length	Real (numeric)	Friend count	Real (numeric)
URL	{True, False}	Verified account	{True, False}
Hashtag	{True, False}	Status count	Real (numeric)
Positive emoticon	{True, False}	Sisted count	Real (numeric)
Negative emoticon	{True, False}	*Max. follower count*	Real (numeric)
Exclamation mark	{True, False}	*Min. follower count*	Real (numeric)
Question mark	{True, False}	*Avg. follower count*	Real (numeric)
Starts with 'RT'	{True, False}	*Max. friend count*	Real (numeric)
Is an @-reply	{True, False}	*Min. friend count*	Real (numeric)
retweet count	{Dynamic}	*Avg. friend count*	Real (numeric)
		Avg. status count	Real (numeric)
		Fraction verified accounts	Real (numeric)

it encapsulates natural human behaviour, indicating the number of users that found a tweet sufficiently interesting to share with their followers. However the retweet metric needs to take into account the relative context of the user and the network. For example, due to the number of their followers, the tweets of a popular user (e.g., celebrity) can generally be expected to be highly retweeted irrespective of the content. Therefore it is important to assess retweet behaviour relative to what can be reasonably expected for a particular author.

Consequently, for a tweet t, we consider the observed retweet count t_O relative to the expected retweet count t_E. The interestingness score for t, denoted $s(t)$ is defined as:

$$s(t) = \frac{t_O}{t_E}$$

where $s(t_1) > s(t_2)$ implies tweet t_1 is *more* interesting than t_2. We note that it is possible to define a set of tweets with at least a particular level of interestingness: i.e., $\{t: s(t) > k\}$. When $k = 1$ this set contains tweets where the observed retweet count is greater than expectation. Applying threshold k in this manner represents a simple application of $s(t)$ to provide a binary classification of tweets based on interestingness.

3.1. Predicting the expected retweet count

Determining the interestingness score for a tweet t requires an estimate of the expected retweet count t_E. We apply machine learning techniques to predict t_E based on easily detectable features exhibited by t. These features extend to the tweet itself, but also capture properties of the author, in terms of their local position in the social graph. A summary of the 31 features adopted for machine learning purposes is presented in Table 1. The lower eight user features (italicised) listed in Table 1 refer to the sampling of each collected user's local network when sampling the data. Note that since feature engineering is implementation- and domain-dependent, we have not focussed on it's investigation here, instead, we utilise features that have proved successful in the literature.

To train a machine learning classifier in prediction of the expected retweet count t_E, a corpus of tweets is required. This has been achieved by randomly walking through Twitter's social graph using the Twitter REST API. For each user visited in the random walk, a set of 1,000 recent tweets (or less if unavailable) from the user's timeline has been collected, alongside the user's features described in Table 1. Subsequent users in the random walk were selected from the user's followers and friends.

It is commonplace that retweeting occurs soon after a tweet being posted. For example [36] identified that approximately 50% of all retweet actions occur within one hour of the tweet being posted, and 75% of retweet actions occur within the first day. As such we use at least day-old tweets in our analyses, which helps to minimise the risk that retweeting behaviour has not yet occurred. Data collection resulted in 240, 717 tweets, denoted T_{full}, authored by 370 Twitter users. For machine learning purposes, the tweets were divided into two sets: 90% formed a training dataset for the classifier, denoted T_{train}. The remaining 10%, denoted T_{test}, were retained for validation of the classifier. The assignment of an authors tweets to T_{train} or T_{test} was made at random and ensured no author had tweets occurring in both sets.

3.2. Categorisation of retweet counts for machine learning

Retweet counts have been shown to follow a long-tailed distribution [17]. Applying a machine learning classifier to a problem with a long-tailed distribution of contiguous data is potentially problematic due to small amounts of training data in the less frequent categories. To counter this, [17] collects training data into intervals in order to predict whether a retweet will fall into one of four broad categories (not retweeted, less than 100 retweets, less than 10,000 retweets or more than 10,000 retweets). Our application requires more granular predictions of retweet behaviour, hence we partition the distribution using variable interval widths such that the total number of instances within each interval is approximately equal. Within training, this increases the opportunity for the retweet counts within the long tail to be identified. This has been achieved by a heuristic displayed in Algorithm 1, which accepts a *requested* number of intervals, R, as an input.

Algorithm 1 Algorithm for dynamically producing containing intervals for retweet counts.

```
procedure GENERATE_INTERVALS(set of tweets T, requested intervals R)
    C ← empty list                          ▷ To hold ordered retweet counts
    I ← empty list                          ▷ To represent container intervals
    for all t ∈ T do
        Add t_O to C
    end for
    Sort C into ascending order
    M ← max(C)                              ▷ Highest instance of t_O
    TSum ← ⌈|C|/R⌉   ▷ Number of tweets to be held in each interval
    H ← empty dictionary                    ▷ To represent the distribution of retweet counts

    for all c ∈ C do
        if c ∈ H then
            Increment H_c
        else
            H_c ← 1
        end if
    end for
    for all i in range 1, ..., M + 1 do
        if i ∈ H then
            s ← s + H_i
        end if
        if s ≥ TSum then
            Add i to I
        end if
    end for
    Return I
end procedure
```

The result of Algorithm 1 on the data set is shown in Fig. 1. To demonstrate the effectiveness of this categorisation, we compare it with an example of linearly defined uniform intervals (Fig. 2), which retains the undesirable long-tail characteristic [13]. Here the lower intervals represent significantly more tweets than the higher ones,

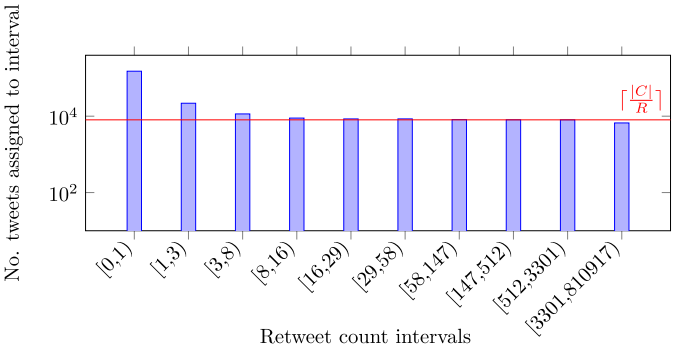


Fig. 1. Cardinalities of dynamically categorised retweet counts of tweets in T_{test} with $R = 15$, yielding an interval count of 10. $\lceil \frac{|C|}{R} \rceil$ represents the target cardinality of each interval (see Algorithm 1).

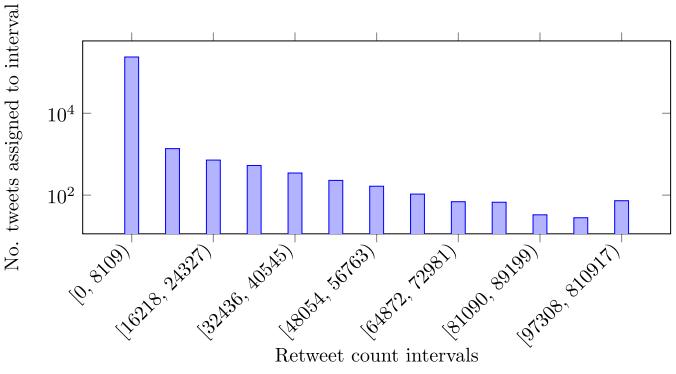


Fig. 2. Cardinalities of linearly categorised retweet counts of tweets in T_{test} with 30 intervals. Note that the final intervals are amalgamated.

Table 2
Cross-validation performance results for the first 10 categorised through the linear and dynamic methods on retweet counts. Note that the remaining categories for Table (a) are excluded as these produce precision and recall values of 0.

Bin interval	Precision	Recall	Bin interval	Precision	Recall
[0,8109)	1.000	0.956	[0,1)	0.935	0.741
[8109,16218)	0.083	0.355	[1,3)	0.218	0.324
[16218,24327)	0.134	0.315	[3,8)	0.190	0.394
[24327,32436)	0.233	0.072	[8,16)	0.240	0.233
[32436,40545)	0.000	0.000	[16,29)	0.291	0.298
[40545,48654)	0.008	0.004	[29,58)	0.265	0.338
[48654,56763)	0.105	0.109	[58,147)	0.232	0.201
[56763,64872)	0.030	0.038	[147,512)	0.256	0.418
[64872,72981)	0.008	0.174	[512,3301)	0.527	0.508
[72981,81090)	0.009	0.343	[3301,810917)	0.519	0.709
(a) Accuracy for linearly-categorised retweet counts with 30 categories.			(b) Accuracy for dynamically categorised retweet counts with $R = 15$.		

resulting in a weaker representation for those tweets receiving a higher retweet counts.

A Bayesian network classifier [37] was used to train against the features defined in Table 1, using both the categorisation schemes in Figs. 1 and 2. Ten-fold cross-validation was carried out with T_{full} for the purposes of comparing the schemes. The results of the comparison are shown in Table 2. These indicate that higher prediction accuracies are obtained with the whole range of variable interval sizes, exhibiting more uniform precision and recall across the intervals, showing that this categorisation method is more effective in classifying the wide range of retweet counts observed in Twitter.

The Bayesian network classifier has been selected via the machine-learning toolkit, Weka[3], on the basis of superior performance against alternative classifiers. Using a randomised subset of T_{full} for training, this technique offers superior performance, in terms of both precision and recall, as compared to the simple logistic, logistic, SMO and Naïve Bayesian alternatives using 10-fold cross validation. Additionally it was the second quickest in terms of training time (of the order of around 1 second on a modern laptop computer). Based on these observations, the Bayesian network classifier and the dynamic retweet interval selection method are adopted in our subsequent experimentation.

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

4. Experimentation and validation

In this section we focus on validation of the interestingness metric from collective (Section 4.1) and individual (Section 4.2) perspectives. In calculating $s(t)$ we adopt the Bayesian network classifier and the dynamic retweet interval selection method as described in the previous section. Given the reported performance of the classifier in predicting t_E as described in Figs. 1, 2 and Table 2, the purpose of our evaluation is to determine the extent to which $s(t)$ identifies tweets that provide affective stimulation, as perceived by users. Note that we are not directly evaluating the predictive retweet capabilities of the classifier. In each of the experiments t_O is the retweet count for the tweet t, as observed in Twitter. This represents the absolute measure of retweet activity for t.

Our evaluation approaches require participants to select from a small set of tweets those that they perceive as most interesting. This is a simple and effective method requiring minimal burden on the participant and reducing the reliance on their interest in the content. From this we can assess the extent to which the most frequently selected tweets are more highly ranked by the measure $s(t)$. This approach allows participants to use their immediate instincts and it removes the need for an individual to calibrate a score of interestingness using an arbitrary scaling. If there is no dominant interest in any tweet the user's choice represents an arbitrary selection of content.

From this approach we measure the extent to which the popularity of the selected tweets is reflected in their interestingness scores, relative to the other tweets displayed. We stress that any ranking of tweets using $s(t)$ is applied here purely for evaluation purposes, and the global ranking of tweets is not the intended primary function of the $s(t)$ metric. It is anticipated that by applying the threshold k to $s(t)$, a first line of content filtering can be provided, distinguishing tweets with a possible higher level of interestingness in a large twitter stream. This functionality allows attention to be managed when following large numbers of users.

4.1. Collective assessment of interestingness

In this test anonymous agents were recruited (using the Amazon Mechanical Turk service) to perform human classification of the interestingness of a sample of tweets from T_{test}. Each Mechanical Turk Worker (MTW) was presented with a series of questions, each containing five tweets from T_{test} authored by the same Twitter user. For each question, MTWs were asked to select the tweets they found the most interesting, and were required to select at least one tweet. The rate of pay was $0.05 per answered question. In total, 150 questions were generated from a set T of 750 tweets randomly chosen from T_{test}, with each question answered by three MTWs. Of these 750 tweets, a set T' of 349 tweets were selected as interesting by at least two out of three workers. In total, 91 distinct workers contributed to the test.

We consider the likelihood of at least two of the three MTWs selecting one of the first i tweets in a given question, ranked by descending $s(t)$ over all questions. This is a useful measure because it gives an insight into the techniques effectiveness when used to shorten (i.e., pre-filter) a user's timeline based on interestingness. Note that this is a relative ranking of the tweets displayed and in particular there is no guarantee that each question will contain a

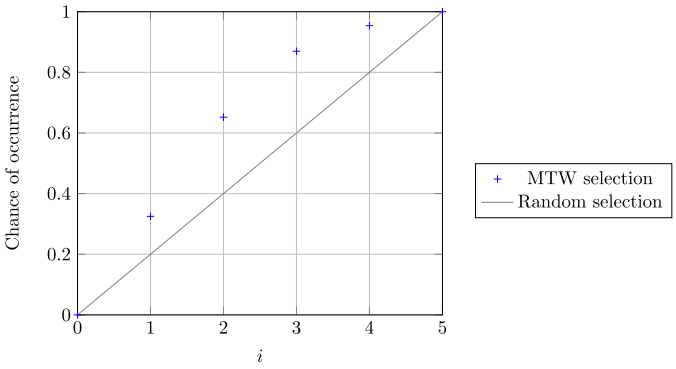


Fig. 3. Likelihood of MTWs selecting one of the first i tweets ranked by descending $s(t)$ over all questions.

tweet t with $s(t) > 1$. Fig. 3 shows that if the question containing five tweets is considered as a timeline, presenting the top three tweets, as ranked by interestingness metric $s(t)$, would capture the majority MTW's timeline choices 88% of the time. However these results represent aggregated views of MTWs, within which consensus is not always possible. As such it is important to assess this from an individual's perspective (Section 4.2).

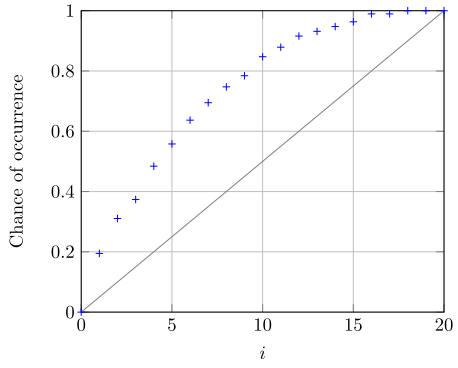
4.2. Individual assessment of interestingness

Individual users of social media readily exhibit their personal differences and dispositions (e.g., [38]). As such, natural variation is likely to affect individual perception of interestingness. In this section we consider people individually, taking into account the user's position within the social network. This involves engaging individual users to identify interesting tweets from a snapshot of their own timelines, and a selection from their neighbours. These selections, and the subsets of tweets from which they were selected, were then evaluated using the interestingness metric.

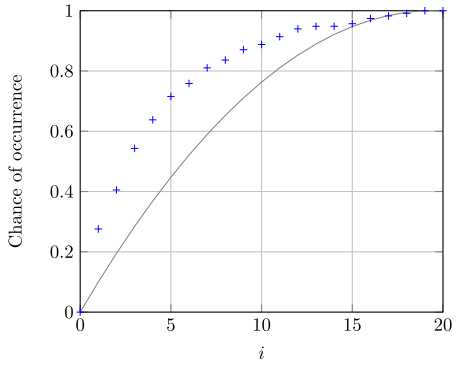
To achieve this a bespoke web application was designed allowing visitors to 'sign-in' using their Twitter credentials. From the user's access keys, as provided through the OAuth mechanism, tweets from the user, their friends and followers were retrieved. Participants were faced with a series of ten tweet timelines; the first representing the participant's current home timeline, and the remaining nine being user timelines from nine of their friends, selected at random with weighting towards more popular users. Each timeline was up to 20 tweets long and participants needed at least 30 friends in order to take part. In each timeline, participants were asked to select the tweet(s) they found the most interesting, and were required to select at least one before moving to the next timeline. Selected tweets were then considered to be 'interesting' and the others uninteresting. On average, participants selected around 1.6 tweets from each timeline.

Users were recruited through voluntary participation (viral web advertising) and through Mechanical Turk. A total of 580 timelines were assessed, consisting of 389 from MTWs and 191 from voluntary participation. In total, the set T^b_{test} of tweets considered by the experiment was authored by 936 unique users and involved 9, 921 tweets. For all these tweets, interestingness scores $s(t)$ were computed by extracting each of their own and their authors' features and classifying the resultant instances against the same classifier model used in Section 4.1. In total, 69.3% of all tweets t selected by the participants had an interestingness score of $s(t) > 1$.

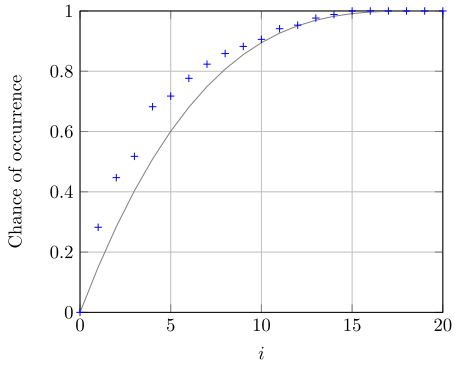
To assess performance we consider ranking the tweets in each timeline considered by the experiment in ascending order of computed interestingness. In Fig. 4 we calculate the chance of a participant selecting one the i highest ranked tweets, as measured by $s(t)$, where $0 \leq i \leq 20$, with 20 being the maximum length of a timeline considered. The random performance in Fig. 4b and c indicates the



(a) In timelines where one tweet was selected



(b) In timelines where two tweets were selected



(c) In timelines where three tweets were selected

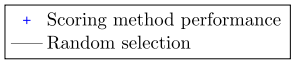


Fig. 4. The chance of a participant selecting one of the *i highest* ranked tweets by $s(t)$ in the timeline.

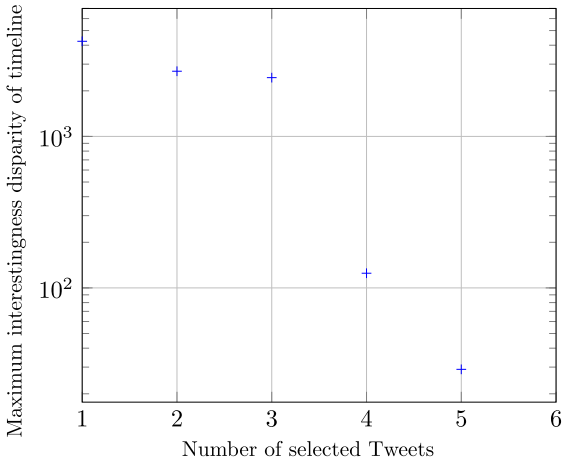


Fig. 5. Relationship between the number of selected tweets in a timeline and the maximum score disparity of the timeline.

likelihood of two and three (respectively) randomly-selected tweets being in the top *i* of those ranked.

The implications of these results are useful, because they demonstrate that the technique can be used to substantially reduce the timeline length, while maintaining a high probability of retaining the tweets that an individual finds interesting. In practical terms this means that cognitive burden can be substantially reduced while preserving the probability of interesting tweets remaining. For example, Fig. 4 shows that the timelines used could be reduced in size by 50% with a 83% chance of retaining the interesting tweets.

The interestingness disparity of a timeline refers to the magnitude of the range of scores observed across its tweets. The maximum disparity of timelines where only one tweet is selected by participants refers to the greatest of such magnitudes. Fig. 5 reports that where this disparity is high, it is more likely that participants select only one tweet from the timeline as interesting, indicating a greater ease of identification of tweets that stand out as being interesting. Timelines with a smaller disparity, and therefore more similarly-interesting tweets, make the selection task more of a cognitive burden to participants, and therefore makes it harder for them to select just one tweet as the most interesting.

5. Conclusions

In this paper we have introduced a method for scoring tweet interestingness using non-semantic methods, and we have demonstrated its ability to infer interesting tweets from the volume and noise within a user's timeline. This has been accomplished by using comparative human tweet assessment in two forms: through aggregated assessment of interestingness using Mechanical Turk, and through a web-based experiment for individual Twitter users. The results provide confidence that the approach is effective in selecting interesting tweets from a user's timeline. This has important implications for reduction of cognitive burden: the results show that timelines can be considerably shortened while maintaining a high degree of confidence that interesting tweets will be retained. The mean length of assessed timelines was 14, yet the vast majority of participants selected only one or two tweets from each as interesting, indicating that a user's own experience of interestingness on Twitter seems less than that that might be achievable through an interestingness-based prioritisation scheme.

The resultant work, as a concept, could be employed in various ways. Implementing such a scheme in Twitter, for example, where "interesting" tweets are given prominence above those less interesting would inevitably lead to the unavailability of non-retweeted tweets which still may be of interest to users. Instead, the research is aimed to help address the "filter bubble" problem [39] by helping to support the notion of the identification of interesting content from beyond the scope of a user's natural social circle. As such, the work would be more useful not for filtering data, but for augmenting the perceived social network structure itself.

For example, TweetBot (and later the official Twitter clients themselves) introduced the "mute" future, in which a user could specify friends to ignore tweets from. However, in terms of usability, this achieves the same effect as temporarily unfollowing the friend, and could therefore be improved by specifying rules that adopt the research we present; 'mute all tweets from user *X* unless a tweet has interestingness score greater than threshold S_{thresh}'. As such, the network has been augmented to simulate a conditional arc through which only a subset of tweets are transmitted.

As a concrete example, consider media sources on Twitter as a potential filter bubble. After classifying news accounts on Twitter as left-wing, centre or right-wing, [40] found that 50% of users followed only sources with a single political leaning. Due to the volume of tweets from each, it could be considered unlikely that these individuals would commit to the cognitive burden of following multiple sources across the political spectrum. However, by filtering these accounts as above, the reader would see the most interesting subset of their content, producing possibly negative, but affective reactions (e.g., anger).

We suggest that our approach is less susceptible to the filter bubble effect than alternative approaches based only on content rather than the author (for example [19]). Firstly, these approaches inherently lead to a focus around certain popular keywords, whereas our approach allows any tweet from any author to emerge. Secondly, our approach is more resilient to gaming and spam. Since interesting tweets are only identified based on their retweet behaviour, an author, for example, cannot simply add popular features (emoticons, URLs and hashtags) in order to raise the profile of their tweet.

The approach is novel in using the implicit intelligence and behaviour of the human, as an agent that responds to interesting received tweets by retweeting them. For an individual user and timeline, the approach determines an expected level of retweeting, using machine learning to take into account tweet and network specific characteristics from a user's perspective. This achieves personalisation, determining whether the actual level of retweeting is significant for an individual, given their network neighbourhood, as well as features concerning the tweet itself. This is a particular strength of the method, making it widely applicable to the diversity of activity found on Twitter, because it addresses the significance of retweeting volume, rather than considering retweeting volume in isolation. Consequently the method is effective in distinguishing between tweet *popularity* and tweet interestingness.

The machine learning approach also offers some interesting characteristics to the overall method. Since the model for expected retweeting is trained on tweets with many features, and across a large variety of retweet counts, there is no need to continually update the model in real time, allowing the method to be used for 'on-demand' inferences with little overhead. Moreover, all tweets from a single user can be evaluated for interestingness using the same predictive scale. These points make the approach flexible and convenient for potential applications. Additionally we observe that the method has generality, with applicability to similar functions found on other social network services, such as 'shares' on Facebook and 'reblogs' on Tumblr. Both of these services provide interesting avenues for further research in this area.

Acknowledgements

This research has been partially supported by the RECOGNITION project grant 257756, an EC-FP7 Future Emerging Technologies project concerning Self-Awareness in Autonomic Systems.

References

[1] A. Java, X. Song, T. Finin, B. Tseng, Why we Twitter: understanding microblogging usage and communities, in: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis (WebKDD/SNA-KDD '07), ACM, New York, NY, USA, 2007, pp. 56–65. http://doi.acm.org/10.1145/1348549.1348556.
[2] D. Zhao, M.B. Rosson, How and why people Twitter: the role that micro-blogging plays in informal communication at work, in: Proceedings of the ACM 2009 International Conference on Supporting Group Work (GROUP '09), ACM, New York, NY, USA, 2009, pp. 243–252. http://doi.acm.org/10.1145/1531674.1531710.
[3] C. Castillo, M. Mendoza, B. Poblete, Information credibility on Twitter, in: Proceedings of the 20th International Conference on World Wide Web (WWW '11), ACM, New York, NY, USA, 2011, pp. 675–684. http://doi.acm.org/10.1145/1963405.1963500.
[4] S. Muralidharan, L. Rasmussen, D. Patterson, J.-H. Shin, Hope for Haiti: an analysis of Facebook and Twitter usage during the earthquake relief efforts, Public Relat. Rev. 37 (2) (2011) 175–177. http://dx.doi.org/10.1016/j.pubrev.2011.01.010.
[5] C. Wilson, A. Dunn, Digital media in the Egyptian revolution: descriptive analysis from the Tahrir Data sets, Int. J. Commun. 5 (2011) 1248–1272.
[6] E. Agichtein, C. Castillo, D. Donato, A. Gionis, G. Mishne, Finding high-quality content in social media, in: Proceedings of the International Conference on Web Search and Web Data Mining, ACM, 2008, pp. 183–194.
[7] O. Alonso, C. Carson, D. Gerster, X. Ji, S.U. Nabar, Detecting uninteresting content in text streams, in: SIGIR Crowdsourcing for Search Evaluation Workshop, 2010.
[8] W. Webberley, S. Allen, R. Whitaker, Inferring the interesting tweets in your network, in: Cloud and Green Computing (CGC), 2013 Third International Conference on, 2013, pp. 575–580, doi:10.1109/CGC.2013.100.
[9] I. Uysal, W.B. Croft, User oriented tweet ranking: a filtering approach to microblogs, in: Proceedings of the 20th ACM International Conference on Information and Knowledge Management(CIKM '11), ACM, New York, NY, USA, 2011, pp. 2261–2264. http://doi.acm.org/10.1145/2063576.2063941.
[10] M.J. Chorley, G.B. Colombo, S.M. Allen, R.M. Whitaker, Human content filtering in Twitter: the influence of metadata, Int. J. Hum.Comput. Stud. (2014). http://dx.doi.org/10.1016/j.ijhcs.2014.10.001.
[11] R.M. Whitaker, M. Chorley, S.M. Allen, New frontiers for crowdsourcing: the extended mind, in: Proceedings of the IEEE 48th Annual Hawaii International Conference on System Sciences, 2015, 2015.
[12] Y. Xu, Relevance judgment in epistemic and hedonic information searches, J. Am. Soc. Inf. Sci. Technol. 58 (2) (2007) 179–189. http://dx.doi.org/10.1002/asi.20461.
[13] W. Webberley, S. Allen, R. Whitaker, Retweeting: a study of message-forwarding in Twitter, in: Workshop on Mobile and Online Social Networks (MOSN), IEEE, 2011, pp. 13–18, doi:10.1109/MOSN.2011.6060787.
[14] Z. Yang, J. Guo, K. Cai, J. Tang, J. Li, L. Zhang, Z. Su, Understanding retweeting behaviors in social networks, in: Proceedings of the 19th ACM International Conference on Information and Knowledge Management(CIKM '10), ACM, New York, NY, USA, 2010, pp. 1633–1636. http://doi.acm.org/10.1145/1871437.1871691.
[15] T.R. Zaman, R. Herbrich, J. Van Gael, D. Stern, Predicting information spreading in Twitter, in: Workshop on Computational Social Science and the Wisdom of Crowds, NIPS, vol. 104, Citeseer, 2010, pp. 17599–17601.
[16] B. Suh, L. Hong, P. Pirolli, E. Chi, Want to be retweeted? large scale analytics on factors impacting retweet in Twitter network, in: Social Computing (SocialCom), 2010 IEEE Second International Conference on, 2010, pp. 177–184, doi:10.1109/SocialCom.2010.33.
[17] L. Hong, O. Dan, B.D. Davison, Predicting popular messages in Twitter, in: Proceedings of the 20th International Conference Companion on World Wide Web (WWW '11), ACM, New York, NY, USA, 2011, pp. 57–58. http://doi.acm.org/10.1145/1963192.1963222.
[18] S. Petrovic, M. Osborne, V. Lavrenko, RT to win! predicting message propagation in Twitter, in: ICWSM, 2011.
[19] N. Naveed, T. Gottron, J. Kunegis, A.C. Alhadi, Bad news travel fast: a content-based analysis of interestingness on Twitter, in: In: Proceedings of the ACM WebSci'11, ACM, 2011, pp. 1–7. http://doi.acm.org/10.1145/2527031.2527052.
[20] J. Zhu, F. Xiong, D. Piao, Y. Liu, Y. Zhang, Statistically modeling the effectiveness of disaster information in social media, in: Global Humanitarian Technology Conference (GHTC), 2011 IEEE, 2011, pp. 431–436, doi:10.1109/GHTC.2011.48.
[21] H.-K. Peng, J. Zhu, D. Piao, R. Yan, Y. Zhang, Retweet modeling using conditional random fields, in: Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on, 2011, pp. 336–343, doi:10.1109/ICDMW.2011.146.
[22] H.B. Celebi, S. Uskudarli, Content Based Microblogger Recommendation, in: Proceedings of the 2012 International Conference on Privacy, Security, Risk and Trust and of the 2012 International Confernece on Social Computing (PASSAT, SocialCom), IEEE, 2012, pp. 605–610, doi:10.1109/SocialCom-PASSAT.2012.124.
[23] S. Kywe, E.-P. Lim, F. Zhu, A survey of recommender systems in twitter, in: K. Aberer, A. Flache, W. Jager, L. Liu, J. Tang, C. Guret (Eds.), Social Informatics, Lecture Notes in Computer Science, 7710, Springer Berlin Heidelberg, 2012, pp. 420–433, doi:10.1007/978-3-642-35386-4_31.
[24] A. Guille, H. Hacid, C. Favre, D.A. Zighed, Information Diffusion in Online Social Networks: A Survey, SIGMOD Rec. 42 (2) (2013) 17–28, doi:10.1145/2503792.2503797.
[25] M.-C. Yang, J.-T. Lee, S.-W. Lee, H.-C. Rim, Finding Interesting Posts in Twitter Based on Retweet Graph Analysis, in: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR '12), ACM, New York, NY, USA, 2012, pp. 1073–1074, doi:10.1145/2348283.2348475.
[26] J.M. Kleinberg, Authoritative sources in a hyperlinked environment, Journal of the ACM 46 (5) (1999) 604–632.
[27] L. Geng, H.J. Hamilton, Interestingness measures for data mining: a survey, ACM Comput. Surv. 38 (3) (2006).
[28] S. Mitra, S.K. Pal, P. Mitra, Data mining in soft computing framework: a survey, IEEE Trans. Neural Netw. 13 (1) (2002) 3–14.
[29] F. Hussain, H. Liu, E. Suzuki, H. Lu, Exception rule mining with a relative interestingness measure, in: T. Terano, H. Liu, A. Chen (Eds.), Knowledge Discovery and Data Mining. Current Issues and New Applications, Lecture Notes in Computer Science, 1805, Springer Berlin Heidelberg, 2000, pp. 86–97.
[30] S. Hidi, W. Baird, Interestingness a neglected variable in discourse processing, Cognit. Sci. 10 (2) (1986) 179–194.
[31] A. Silberschatz, A. Tuzhilin, On subjective measures of interestingness in knowledge discovery, in: Proceedings of the First International Conference on KDD, 95, 1995, pp. 275–281.
[32] S. Gransee, R. McAfee, A. Wilson, Twitter Retweet Prediction, 2012.
[33] A.C. Alhadi, T. Gottron, J. Kunegis, N. Naveed, LiveTweet: Microblog Retrieval Based on Interestingness and an Adaptation of the Vector Space Model, in: Proceedings of the 2011 Text REtrieval Conference (TREC'11), National Institute of Standards and Technology (NIST), 2011.
[34] A.C. Alhadi, T. Gottron, J. Kunegis, N. Naveed, LiveTweet: Monitoring and Predicting Interesting Microblog Posts, in: Advances in Information Retrieval, in: Lecture Notes in Computer Science, 7224, Springer Berlin Heidelberg, 2012, pp. 569–570, doi:10.1007/978-3-642-28997-2_66.
[35] H. Lauw, A. Ntoulas, K. Kenthapadi, Estimating the Quality of Postings in the Real-Time Web, in: Proceedings of SSM Conference, 2010.
[36] H. Kwak, C. Lee, H. Park, S. Moon, What is Twitter, a Social Network or a News Media? in: Proceedings of the 19th International Conference on World Wide Web (WWW '10), ACM, New York, NY, USA, 2010, pp. 591–600. http://doi.acm.org/10.1145/1772690.1772751.
[37] N. Friedman, D. Geiger, M. Goldszmidt, Bayesian network classifiers, Mach. Learn. 29 (2-3) (1997) 131–163, doi:10.1023/A:1007465528199.
[38] M.J. Chorley, R.M. Whitaker, S.M. Allen, Personality and location-based social networks, Comput. Hum. Behav. 46 (2015) 45–56.
[39] E. Pariser, The filter bubble: What the Internet is hiding from you, Penguin UK, 2011.
[40] J. An, M. Cha, P.K. Gummadi, J. Crowcroft, Media landscape in Twitter: a world of new conventions and political diversity., in: International AAAI Conference on Weblogs and Social Media (ICWSM), 2011.

Please cite this article as: W.M. Webberley et al., Retweeting beyond expectation: Inferring interestingness in Twitter, Computer Communications (2015), http://dx.doi.org/10.1016/j.comcom.2015.07.016