Towards Real-time Lifetime Prediction of Information Diffusion

Io Taxidou, Anas Alzoghbi, Peter M. Fischer, Christoph Schöller
University of Freiburg, Germany
{taxidou,alzoghbi,peter.fischer}@cs.uni-freiburg.de, chschoeller@web.de

1. MOTIVATION AND CHALLENGES

In this paper, we provide the first steps towards real-time, large-scale prediction of the lifetime of information diffusion processes.

Analyzing information diffusion at large scale and in real-time speed provides a broad range of benefits for many use cases such as online journalists: Since social media exerts a considerable influence on how people are being informed, journalists need to incorporate it into their investigations, as well as into their own publishing process, which both corresponds to diffusion analysis.

In this context, predicting the lifetime in a real-time manner has several benefits: Understanding how much longer a piece of information will last indicates how many resources to allocate to the observation. Furthermore, in real-time analysis of cascades [3], predicting the lifetime gives an indication on the completeness and further spread. Yet, to the best of our knowledge, there is little work on lifetime prediction, and none that targets real-time prediction.

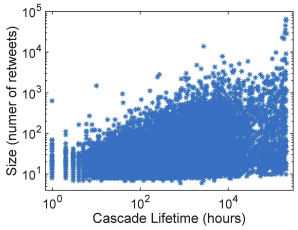
There is, however, a lot of research concerning predictions on other properties of information diffusion, like size [2], scale and speed [4]. The closest work is [1], which implements incremental predictions of cascade size over time, while authors support that the final state of each cascade is inherently unpredictable. As a result, we can conclude that cascade lifetime is a hard to predict property; it is even more challenging than predicting the cascade size (in terms of messages), since a single message can alter the lifetime dramatically.

To support these insights, we present the correlation of lifetimes and sizes of cascades in our test data set (described in Section 3); as Figure 1a shows, there is no correlation, meaning that existing solutions to predict the size [1] are of limited use. The pronounced skew in the cascade lifetime distribution imposes another challenge in solving such a problem, as shown in Figure 1b, again derived from the same dataset: Large cascades that reach virality (that are the most relevant to predict for our use cases) are extremely rare, while small cascades are over-represented. In case we treat all cascades in the same way, we will bias the results in favor of short lifetime predictions.

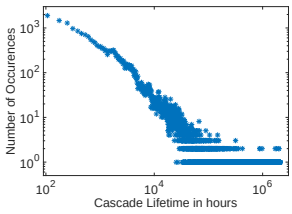
While not (yet) providing a complete solution to the problem of real-time lifetime prediction, this paper contributes the following:

WebSci '15 June 28 - July 01, 2015, Oxford, United Kingdom

ACM 978-1-4503-3672-7/15/06 ...$15.00.
http://dx.doi.org/10.1145/2786451.2786926



(a) (Missing) Correlation of Lifetime and Size



(b) Distribution of Lifetimes

- It determines the feasibility of cascade lifetime prediction.
- It defines an approach for incremental predictions.
- It investigates relevant algorithms and features.

2. METHODS AND FEATURES

The challenges outlined in Section 1 lead to the conclusion that predicting the very end of the cascade is not feasible. Instead, we model a binary classification problem that determines if a cascade will survive for the next period of time. This leads to incremental prediction problem which will be updated with new observations gathered from the evolving diffusion process. As shown in Figure 2, we define an *observation window* of the current lifetime and we set a *prediction window* relative to the current lifetime. This problem definition matches our use case: Online journalists are interested in whether a piece of information will last for the next hours or days in order to keep observing it.

The first challenge to tackle is how to *model* the properties of cascades to allow effective prediction. We investigated a *stage-based* approach, in which we split each cascade into the same number of ranges of its lifetime. The ranges are defined as fractions of the full cascade lifetime, thus eliminating the different overall lifetimes. For example, in our study, we sliced each cascade in 10 stages starting from the beginning of each cascade: 5, 10, 20, 30, 40, 50, 60, 70, 80, 90 % of its lifetime. Every one of these stages splits the cascade into an observation window and a prediction window, as shown in Figure 2.

Furthermore, we investigated which parts of the cascade so far (within the observation window) are more predictive: the full observation window or the most recent part of it? The latter may capture the recent behavior better, while the former may provide a broader picture. As Figure 2 shows we considered 10, 30, 60% of the most recent parts of it and we compare with the full (100%) observation window.

The next question we want to answer is what constitutes a meaningful prediction goal and how far can we go with the predictions from a given stage. For that, we varied the predictions windows relative to the observation windows: 0.25, 0.5, 1, 2 times the full

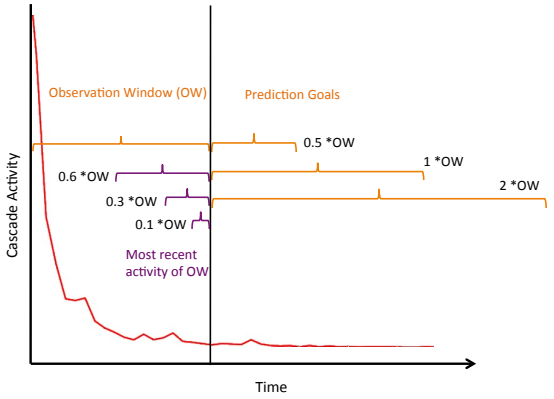


Figure 2: Varying Observation and Prediction Windows

observation window and we observe the prediction accuracy.

We considered two *learning algorithms*: *ZeroR* as baseline and *Random Forest* as more elaborate algorithm. ZeroR is a naive algorithm that assigns the most frequent class found in the data. In this context, we expect high performance if the the prediction window and the observation window reflect each other by having high self-similarity. The choice of a random forest algorithm is connected with its speed and efficiency for training and evaluation procedures, appropriate for large scale, real-time analysis.

We used a limited amount of simple but dynamic *features* to match the scaling of computations in a large scale, real-time scenario and show the feasibility of these methods. We selected six features that can be easily computed or extracted from the messages. We group them into two main categories: *temporal* and *user* related features, computed in every observation window. For temporal features we considered: time elapsed from the last tweet to the current tweet, cascade duration so far and observation window duration to the number of retweets. These features show how viral a piece of information is and how much activity we observe per time unit.

For user-related features we considered: number of retweeters (=number of unique retweets), average number of retweeters' followers, audience size of the last hour (= users exposed to the message). Such features reveal the power of intermediate recipients and forwarders in the diffusion process.

3. EVALUATION

We collected datasets from Twitter covering the period from July to September 2012 with the hashtags "Olympics" and "London2012", retaining cascades with at least 5 retweets. We ended up with with 32.584 cascades with message completeness of 97% and lifetimes between few hours and several months.

Since a fully real-time evaluation for training and predicting is not our goal, instead we are focused on proving the feasibility of such an approach, we performed a standard 10 fold cross validation and reported the average accuracy.

With the results gathered from our experiments, we can provide initial answers to several hypotheses:

H1: Incremental predictions is a useful approach to tackle the cascade lifetime prediction problem and yield useful results

According to Table 1, we identify effective predictions: For 100% observation windows, the prediction rates are between 75 and almost 90 %. The results become worse overall the longer the prediction window is: for short prediction windows, the observation windows tends to be quite similar to the prediction windows, yielding also very high scores. For longer predictions windows, these baselines drop quickly, while Random Forest exploits the features better and provides higher gains.

Prediction Window	100% obs. window	60% obs. window	30% obs. window	10% obs. window
0.25*obs. window	**ZR: 0.89** RF: 0.88	**ZR: 0.92** RF: 0.92	**ZR: 0.92** RF: 0.92	**ZR: 0.91** RF: 0.91
0.5*obs. window	ZR: 0.69 **RF: 0.80**	**ZR: 0.77** RF: 0.78	**ZR: 0.79** RF: 0.76	**ZR: 0.76** RF: 0.74
1*obs. window	ZR: 0.59 **RF: 0.77**	**ZR: 0.68** RF: 0.71	**ZR: 0.71** RF: 0.70	**ZR: 0.69** RF: 0.70
2*obs. window	ZR: 0.61 **RF: 0.74**	ZR: 0.51 RF: 0.63	ZR: 0.53 RF: 0.62	ZR: 0.52 RF: 0.62

Table 1: Prediction Results

H2: Choosing a limited observation window will affect the prediction quality

H3: An elaborate learning algorithm will provide better results than a naive, statistical one

As the results in Table 1 show, these two hypotheses need to be evaluated together. ZeroR, as the naive based algorithm, yields similar or better results when targeting short prediction goals (0.25&0.5% $\times$ obs.window), meaning that there is high similarity between the observation window and limited prediction window.

Random Forest nearly always benefits from more available training data, showing its best performance on the 100% observation window. At larger prediction windows (1&2 $\times$ obs.window), Random Forest establishes a clear lead over ZeroR, providing gains up to 18%. For smaller prediction windows, the odds are changing. At 0.25%, the effect of self-similarity is so strong that ZeroR beats Random Forest in most cases. 0.5 % is somewhat of the middle ground, where Random Forest has a very small lead over ZeroR.

Our conclusions from such analysis are that in a real time system, we can make cheap predictions (up to 0.5 $\times$ obs.window) by monitoring the most recent part of the observation window and discard the rest of the content. For longer prediction windows we need information from the beginning of each cascade, which is more costly to compute in a real-time set up but still feasible.

4. CONCLUSION AND FUTURE WORK

This work sets the baselines for incremental predictions of cascade lifetime. We have proposed a method to overcome the difficulties of skewedness and variability of cascade lifetimes and predict how much longer they will last. Also, we investigate which parts of the cascades are more predictive and proposed simple and "cheap" to compute features that give reasonably good results.

Future work includes the implementation of a real-time predictor of lifetimes: as well as the investigation of more complex features.

5. REFERENCES

[1] J. Cheng et al. Can cascades be predicted? In *WWW*, 2014.

[2] A. Kupavskii et al. Prediction of retweet cascade size over time. In *CIKM*, 2012.

[3] I. Taxidou and P. M. Fischer. Online analysis of information diffusion in twitter. In *WWW (Companion Volume)*, 2014.

[4] J. Yang and S. Counts. Predicting the speed, scale, and range of information diffusion in twitter. *ICWSM*, 10, 2010.