FAST: Forecast and Analytics of Social Media and Traffic

Venkata Rama Kiran Garimella
Qatar Computing Research Institute
Doha, Qatar
gvrkirann@gmail.com

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

CSCW'14 Companion, Feb 15–19 2014, Baltimore, MD, USA
ACM 978-1-4503-2541-7/14/02.
http://dx.doi.org/10.1145/2556420.2556784

Abstract
We present FAST (http://fast.qcri.org/), a platform for real-time traffic predictions in online news sources. FAST accurately forecasts the future number of page views of an article based on user traffic and social media engagement signals. To our knowledge, this is the first industrial scale, real-time system for predictive web analytics.

Author Keywords
news media, social media, web analytics

ACM Classification Keywords
H.3.5 [Online information systems]: Web-based services; H.3.4 [Systems and Software]: User profiles and alert services

Introduction
Digital news sources have grown in audience while traditional ways of being informed lose ground [1]. Basically all major news websites capture detailed data about their online audience. This data is aggregated and visualized through web analytics products such as Google Analytics,[1] Visual Revenue[2] and Chartbeat[3], that can

[1] http://analytics.google.com/
[2] http://www.visualrevenue.com/
[3] http://www.chartbeat.com/

generate real-time reports on website traffic, referrals, and social media reactions. However, it is not easy to convert such data into actionable insights. In particular, helping digital newsrooms anticipate changes in the interests of their viewers remains elusive. Producing new content of high quality may take hours or days, and by the time the content has been produced the traffic volumes may have changed substantially. In a typical news website, the article that is receiving the most visits at this moment is unlikely to be the most visited article tomorrow.

In this context, we introduce FAST (Forecast and Analytics of Social Media and Traffic), a system that produces traffic predictions for online news articles in real-time.

Web predictions is a well-established research topic with many results including predicting popularity of content on Youtube and Digg [4], predicting activity on Slashdot.com [3], and predicting the number of comments on news stories [5], etc. The interested reader is referred to [2] for an overview.

We exploit the dynamic relation between social media reactions and visits over time, and show that both are useful to predict future visit patterns. Our predictions are made using data collected in real-time with a very short lag between the time when the article is first seen and a prediction is made (1 hour), making them useful for taking actionable decisions.

Our solution

We embed an Open Web Analytics[4] Javascript code in all the web pages we want to track. This tracking code sends information to our platform which is aggregated in real-time using a high-performance Apache S4 deployment.[5]

When an article from the site makes it into the 30 most visited pages on a 5-minute window, we launch a separate process that periodically asks Twitter and Facebook for information about its URL. This information contains the number of postings as well as the entropy of the vocabulary of messages posted in Twitter. All this data is stored using an efficient Cassandra NoSQL database, to enable fast computations. Information on previously-published articles is used to learn a statistical model of website traffic, which is then used to generate traffic predictions for new articles.

Previous studies have shown that there are different classes of news articles with typical attention curves [2]. Based on this idea, we classify articles (based on URL patterns) into two broad types, (i) "news" containing daily news updates, breaking news, etc. and (ii) "other" containing articles such as documentaries, opinions, editorials, etc. We build separate models for $news$ and $other$ articles at different time resolutions (1–6, 12, 24, and 48 hours) and recompute these models periodically on new articles in our database that have passed the 3-day threshold. The features we use for our models are based on the research results presented in [2], which include information about page views at the different time resolutions and various social media signals like entropy of the vocabulary of tweets, Facebook likes/shares, and average number of followers/friends/statuses of users on Twitter.

[4] http://www.openwebanalytics.com/

[5] http://incubator.apache.org/s4/

Accuracy of the predictions
We test the accuracy of our models on the news portal Al Jazeera English (http://www.aljazeera.com/). Most news articles exhibit fairly predictable trajectories, almost like a ballistic trajectory, with visits per minute going up and then down following a smooth curve. However, not all articles start with the same speed or generate the same reaction in the audience. The accuracy of the prediction improves as time passes. This is because most articles saturate after around 6 hours after publication.

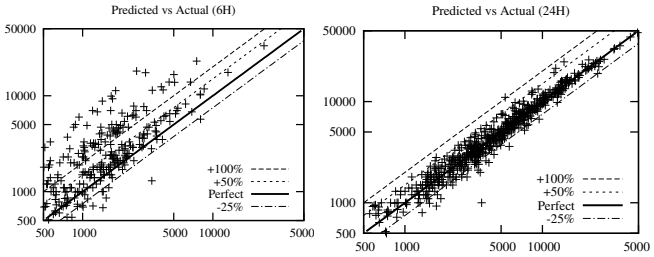


Figure 1: Analysis of the performance of our models. Each point is an article, with the x-axis being the prediction and the y-axis the actual number of visits after 3 days.

An analysis of the quality of our predictions using two different models is provided in Figure 1. We can see from the figure that as time passes, the prediction accuracy increases, with the 24 hour model getting most of the predictions in the 25% range. The 6 hour model mostly over predicts because of unpredictable increases in traffic from social media immediately after publication, which subside after a few hours.

Features of the demonstration

FAST consists of three main dashboards, (i) the 5-min traffic dashboard consisting of the top 30 articles in terms of page views sorted by the number of page views in the last 5 minute window. This list is refreshed every 5 seconds as new page views to articles arrive and the articles are re-sorted accordingly. (ii) The 3-day predictions dashboard, consisting of articles sorted by the *expected* number of visits in the next 3 days. This dashboard is refreshed every 10 minutes and new predictions are computed. A screenshot of the 3-day prediction dashboard is depicted in Figure 2.

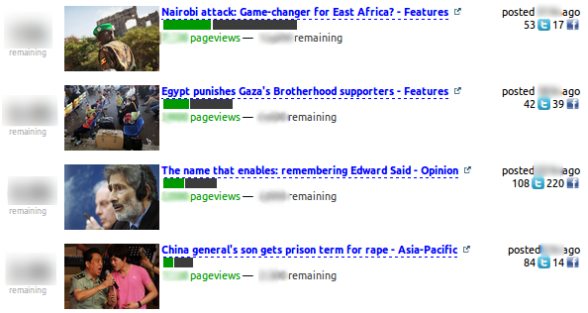


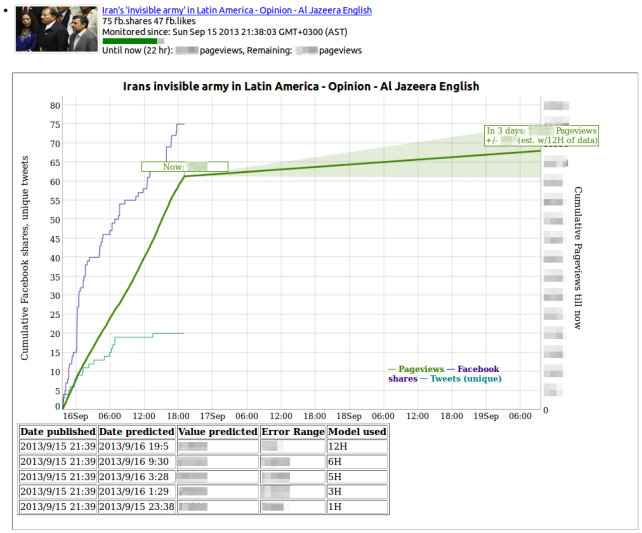
Figure 2: A screenshot of the 3-day predictions dashboard using data from Al Jazeera English. The green bars represent visits received so far by each article, the gray bars visits still to be received according to the predictions.

In both dashboards, the article title links to the article dashboard (Figure 3) where a user can see the time series of the visits, Facebook shares and unique tweets until now, as well as the predicted number of page views. We also include the error bars with 90% confidence intervals on the plot and predictions made using previous models. Actual traffic numbers are shown only to authorized users due to the business-sensitive character of the data.

Our platform performs real time predictions on online news articles and can be integrated into any online news

portal. Since the predictions are made in real time, they can be very useful for news editors to make informed decisions based on the results, such as creating a follow-up article, interactive, or photo gallery for news stories that have high traffic predictions.

FAST is currently being used in the Al Jazeera English news room. Anecdotally, we have observed that content producers try to "beat" the predictions by promoting more aggressively in social media the articles that FAST predicts will under-perform.



Date published	Date predicted	Value predicted	Error Range	Model used
2013/9/15 21:39	2013/9/16 19:5			12H
2013/9/15 21:39	2013/9/16 9:30			6H
2013/9/15 21:39	2013/9/16 3:28			5H
2013/9/15 21:39	2013/9/16 1:29			3H
2013/9/15 21:39	2013/9/15 23:38			1H

Figure 3: A screenshot of the article dashboard, indicating the profile of visits to an article and the history of predictions done by the different models.

Relevance to CSCW

This is a novel social computing application: by incorporating social media signals, it is the first to demonstrate predictive web analytics in an industrial scale. It serves the needs of online media, particularly in the domain of news. It complements the research article [2] by extending its analysis into a real-world application with which conference attendees can interact.

Acknowledgements

The authors wish to thank Al Jazeera English for the data used for this study, and Mohammed El-Haddad, Jürgen Pfeffer and Matt Stempeck for feedback received while developing this system.

References

[1] D. Agarwal, B.-C. Chen, and X. Wang. Multi-faceted ranking of news articles using post-read actions. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 694–703. ACM, 2012.

[2] C. Castillo, M. El-Haddad, J. Pfeffer, and M. Stempeck. Characterizing the life cycle of online news stories using social media reactions. In *To appear in Proc. of CSCW*, 2014.

[3] A. Kaltenbrunner, V. Gomez, and V. Lopez. Description and prediction of slashdot activity. In *Web Conference, 2007. LA-WEB 2007. Latin American*, pages 57–66. IEEE, 2007.

[4] G. Szabo and B. A. Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88, 2010.

[5] M. Tsagkias, W. Weerkamp, and M. De Rijke. Predicting the volume of comments on online news stories. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1765–1768. ACM, 2009.