

# On using the Real-time Web for News Recommendation & Discovery\*

Owen Phelan, Kevin McCarthy, Mike Bennett, and Barry Smyth  
 CLARITY: Centre for Sensor Web Technologies  
 School of Computer Science & Informatics  
 University College Dublin, Ireland  
 firstname.lastname@ucd.ie

## ABSTRACT

In this work we propose that the high volumes of data on real-time networks like Twitter can be harnessed as a useful source of recommendation knowledge. We describe *Buzzer*, a news recommendation system that is capable of adapting to the conversations that are taking place on Twitter. Buzzer uses a content-based approach to ranking RSS news stories by mining trending terms from both the public Twitter timeline and from the timeline of tweets generated by a user's own social graph (friends and followers). We also describe the result of a live-user trial which demonstrates how these ranking strategies can add value to conventional RSS ranking techniques, which are largely recency-based.

## Categories and Subject Descriptors

H.4 [Information Systems Applications]: Miscellaneous

## General Terms

Algorithms, Experimentation, Theory

## Keywords

Content-based recommendation, News recommendation, Real-time recommendation, Social recommendation, Twitter

## 1. INTRODUCTION

We view the emergence of the real-time web as a way of harnessing real-time and time-sensitive data, such as tweets on public and personal Twitter streams, as a basis for recommending news items. Applied research has already been published relating to Twitter data applications [2, 3, 4, 7]. We have previously presented an early basic prototype of the Buzzer system [7] that focussed on real-time ranking.

Here, we widely extend the basic technique with four new overlapping content-based recommendation strategies [1, 5] and describe the result of a large-scale live-user trial with 35 users over a 1 month period, based on more than 30,000 news stories and in excess of 50 million Twitter messages.

Buzzer generates two indexes of content — one from Twitter (including public tweets and Buzzer-user tweets) and one

\*This work is gratefully supported by Science Foundation Ireland under Grant No. 07/CE/11147 CLARITY CSET.

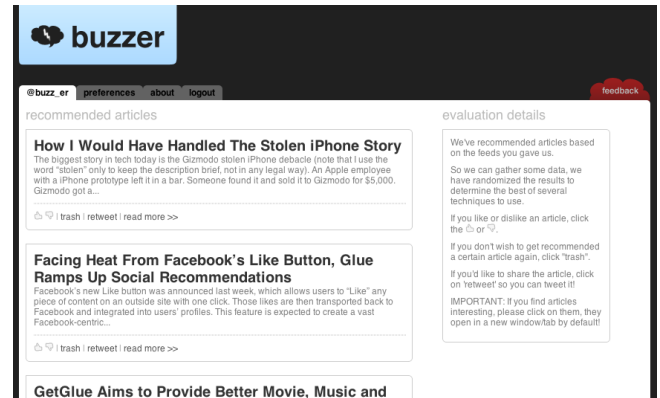


Figure 1: Buzzer's personalized news results

from the RSS feeds of Buzzer users. Buzzer looks for co-occurrences of content between the terms that are present in tweets and RSS articles and ranks articles accordingly. In this way articles with content that appear to match the content of recent Twitter chatter (whether public or user related) will receive high scores during recommendation.

Figure 1 shows a sample list of recommendations for a particular user. Buzzer is a web application and can take the place of a user's normal RSS reader: the user continues to have access to their favourite RSS feeds but in addition, by synching Buzzer with their Twitter account, they may potentially benefit from a more informative ranking of news stories based on their inferred interests.

## 1.1 Algorithms & Strategies

Each Buzzer user brings both their RSS subscriptions and their Twitter social graph to the system. With these sources, we build a number of different ways of combining tweets and RSS content during recommendation, and in this paper we explore 4 different recommendation strategies ( $S1-S4$ ). For example, stories/articles can be mined from a user's personal RSS feeds or those of the wider Buzzer community. Moreover, stories can be ranked based on the tweets of people the user follows, or from the tweets of the public Twitter timeline.

We combine these sources to generate 4 strategies:  $S_1$  — Stories from Personal RSS Articles ranked by the Public Twitter Feed,  $S_2$  — Stories from Personal RSS Articles ranked by your Twitter friends,  $S_3$  — Articles from the Buzzer Community’s RSS Feeds ranked by the Public Twitter Feed,  $S_4$  — Articles from the Buzzer Community’s RSS Feeds ranked by your Twitter friends. Our evaluation uses a 5th strategy of recency of RSS articles as a benchmark.

The system generates results from these recommendation strategies, by mining feeds of tweets and RSS items for co-occurring terms, building a term-frequency vector of these terms and uses this vector as a basis of retrieving relevant RSS items from a given RSS index. Item scoring is a sum of each TF-IDF score of each returned individual item across the vector of terms, so if an item is in the result list of many terms, its score is summed and added to a final result list and ranked. We do this process for each of the four strategies outlined previously, and the system randomly interleaves each of the five ( $S_1$ - $S_5$ ) strategies into one final list.

## 2. USER TRIAL

The trial consisted of 35 active Buzzer users (used the system two or more times) who were emailed a result list daily. The results presented relate to usage data gathered during the 31 days of March 2010. During this timeframe we gathered a total of 56 million public tweets (for use in strategies  $S_1$  and  $S_3$ ) and 537,307 tweets from the social graphs of the 35 registered users (for use in strategies  $S_2$  and  $S_4$ ). In addition, the 35 users registered a total of 281 unique RSS feeds as story sources and during the trial period these feeds generated a total of 31,137 unique stories/articles. During the trial, Buzzer issued 1,085 emails. We considered the participants as active Twitter users, with averages of 145 friends, 196 followers and 1241 tweets sent. Overall our interest is not so much concerned with whether one strategy is superior to others — because in reality we believe that different strategies are likely to have a role to play in news story recommendation — but rather to explore the reaction of users to the combination of strategies.

### 2.1 Results

Figure 2A presents total click-throughs per strategy. Users registered 15 or so RSS feeds as part of their Buzzer sign-up and the stories ranked by  $S_3$  and  $S_4$ , for a given user, came from the 250+ other feeds contributed by the participants. Overall, we can see that strategies  $S_1$  and  $S_2$  tend to outperform the other strategies; for example,  $S_1$  and  $S_2$  received about 110 click-throughs each, just over 35% more than strategies  $S_3$  and  $S_4$ , and about about 20% more than the default recency strategy,  $S_5$ . There is a clear preference for stories from personal feeds, strategies  $S_1$  and  $S_2$  attract more click-throughs than when ranked by recency ( $S_5$ ). But also, participants responded less frequently to stories ranked highly by strategies  $S_3$  and  $S_4$ , although these strategies still attract about 30% of total click-throughs. We feel that this highlights the power of item discovery in our system.

It is also useful to consider whether particular strategies tended to win out over other strategies on a day by day basis. We can judge a strategy  $S_i$  to win on day  $d_j$  if  $S_i$  receives more click-throughs than any other during  $d_j$ . Figure 2B shows the result of this analysis across the 31 trial days, for each of the 5 strategies. Strategy  $S_2$  (user’s personal RSS

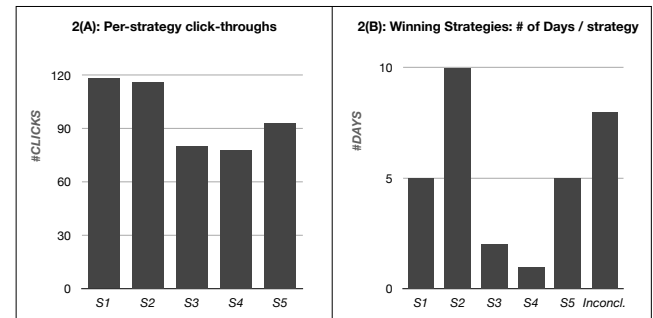


Figure 2: Main Results

feeds ranked by the tweets of their social graph) wins overall, dominating the click-throughs of 10 out of the 31 days. Recency ( $S_5$ ) comes a close second (winning on 8 of the days). Overall strategies  $S_3$  and  $S_4$  do less well here, collectively winning on only 3 of the 31 days.

## 3. CONCLUSION

We have presented a novel content-based real-time news recommendation and discovery system, with a large-scale user trial and interesting results that influence our future directions. We hope to expand the discovery system with a hybrid approach (sharing activities of users as a rating system) and focus on content and metrics from the users personal social graph, as we feel it yields the most interesting results. Further information and in-depth analysis of data from this evaluation is also available [6].

## 4. REFERENCES

- [1] Daniel Billsus and Michael J. Pazzani. A personal news agent that talks, learns and explains. In *AAMAS'99*, pages 268–275. ACM Press, 1999.
- [2] Sandra Garcia Esparza, Michael P. O’Mahony, and Barry Smyth. On the real-time web as a source of recommendation knowledge. In *RecSys 2010*, Barcelona, Spain, September 26-30 2010. ACM.
- [3] John Hannon, Mike Bennett, and Barry Smyth. Recommending twitter users to follow using content and collaborative filtering approaches. In *RecSys 2010*, Barcelona, Spain, September 26-30 2010. ACM.
- [4] Haewoom Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *WWW '10*, pages 591–600, 2010.
- [5] Michael Pazzani and Daniel Billsus. Content-based recommendation systems. *The Adaptive Web*, pages 325–341, 2007.
- [6] Owen Phelan, Kevin McCarthy, Mike Bennett, and Barry Smyth. Terms of a feather: Content-based news discovery and recommendation using twitter. In *ECIR 2011*. Springer-Verlag, 2011.
- [7] Owen Phelan, Kevin McCarthy, and Barry Smyth. Using twitter to recommend real-time topical news. In *RecSys '09*, pages 385–388, New York, NY, USA, 2009. ACM.