What to Write? A topic recommender for journalists

Giovanni Stilo and **Paola Velardi**
Sapienza University of Rome, Italy
{stilo, velardi}@di.uniroma1.it

Alessandro Cucchiarelli and **Giacomo Marangoni** and **Christian Morbidoni**
Università Politecnica delle Marche, Italy
{a.cucchiarelli, g.marangoni, c.morbidoni}@univpm.it

Abstract

In this paper we present a recommender system, What To Write and Why (W^3), capable of suggesting to a journalist, for a given event, the aspects still uncovered in news articles on which the readers focus their interest. The basic idea is to characterize an event according to the echo it receives in online news sources and associate it with the corresponding readers' communicative and informative patterns, detected through the analysis of Twitter and Wikipedia, respectively. Our methodology temporally aligns the results of this analysis and recommends the concepts that emerge as topics of interest from Twitter and Wikipedia, either not covered or poorly covered in the published news articles.

1 Introduction

In a recent study on the use of social media sources by journalists (Knight, 2012) the author concludes that "social media are changing the way news are gathered and researched". In fact, a growing number of readers, viewers and listeners access online media for their news (Gloviczki, 2015). When readers feel involved by news stories they may react by trying to deepen their knowledge on the subject, and/or confronting their opinions with peers. Stories may then solicit a reader's *information* and *communication* needs. The intensity and nature of both needs can be measured on the web, by tracking the impact of news on users' search behavior on on-line knowledge bases as well as their discussions on popular social platforms. What is more, on-line public's reaction to news is almost immediate (Leskovec et al., 2009) and even anticipated, as for the case of planned media events and performances, or for disasters (Lehmann et al., 2012). Assessing the focus, duration and outcomes of news stories on public attention is paramount for both public bodies and media in order to determine the issues around which the public opinion forms, and in framing the issues (i.e., how they are being considered) (Brooker and Schaefer, 2005). Futhermore, real-time analysis of public reaction to news items may provide useful feedback to journalists, such as highlighting aspects of a story that needs to be further addressed, issues that appear to be of interest for the public but have been ignored, or even to help local newspapers echo international press releases.

The aim of this paper is to present a news media recommender, What to Write and Why (W^3), for analyzing the impact of news stories on the readers, and finding aspects – still uncovered in news articles – on which the public has focused their interest. The purpose of W^3 is to support journalists in the task of reshaping and extending their coverage of breaking news, by suggesting topics to address when following up on such news items. For example, we have found that a common pattern for news readers is to search events of the same type occurred in the past on Wikipedia, which is not surprising per se: however, among the many possible similar events, our system is able to identify those that the majority of readers consider (sometimes surprisingly) highly associated with breaking news, e.g., searching for the 2013 CeaseFire program in Baltimore during Egypt's ceasefire proposal in Gaza on July 2014.

2 Methodology

Our methodology is in five steps, as shown in the workflow of Figure 1:

Step 1. Event detection: We use SAX*, an unsupervised temporal mining algorithm that we

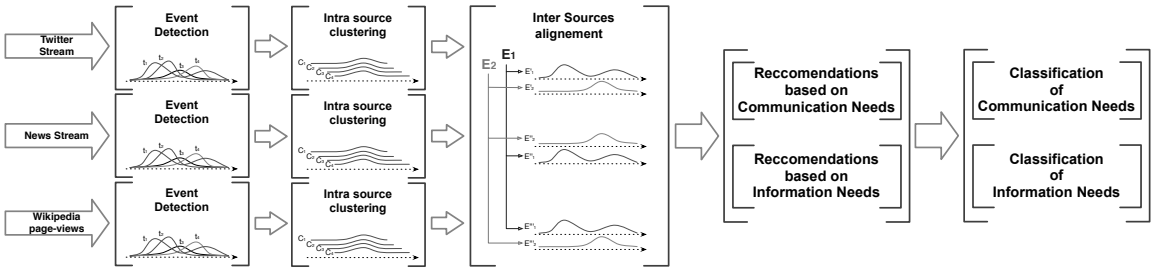

Figure 1: Workflow of W^3

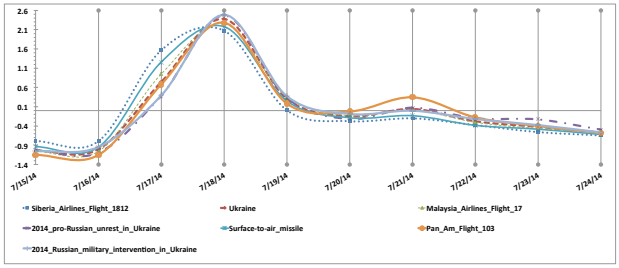

Figure 2: Cluster of normalized time series of Wikipedia page views for Malaysia Airline crash on July 2014.

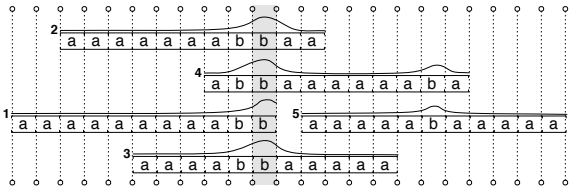

Figure 3: SAX* strings associated to a temporal series $s(t)$ in 5 adjacent or overlapping windows.

introduced in (Stilo and Velardi, 2016), to cluster tokens – words, entities, hashtags, page views – based on the *shape similarity* of their associated signals $s(t)$. In SAX*, signals observed in temporal windows L_k are first transformed into strings of symbols of an alphabet Σ; next, strings associated to *active* tokens (those corresponding to patterns of public attention) are clustered based on their similarity. Each cluster is interpreted as related to an event e_i. Clusters are extracted independently from on-line news (N), Twitter messages (T) and Wikipedia page views (W).

For example, the cluster in Figure 2 shows Wikipedia page views related to the Malaysia Airline crash on July 2014. We remark that SAX* blindly clusters signals without prior knowledge of the event and its occurrence date, and furthermore, it avoids time-consuming processing of text strings, since it only considers active tokens.

Step 2. Intra-source clustering: Since clusters are generated in *sliding windows* L_k of equal length L and temporal increment Δ, clusters referring to the same event but extracted in partly overlapping windows may slightly differ, especially for long-lasting events, when news updates motivate the emergence of new sub-topics and the decay of others. An example is in Figure 3, showing for simplicity a cluster with a unique signal $s(t)$ which we can also interpret as the cluster centroid. The Figure also shows the string of symbols associated with the signal in each window (with $\Sigma = \{a, b\}$).

For a better characterization of an event, we merge clusters referring to the same event and extracted in adjacent windows, based on their similarity. Merged clusters form *meta-clusters*, denoted with m_i^S, where the index i refers to the event and $S \in \{N, T, W\}$ to the data source. With reference to Figure 3, the signals in windows 1, 2, 3 and 4 would be merged, but not the signal in window 5.

An example from the T dataset is shown in Table 1: note that the first two clusters show that initially Twitter users where concerned mainly about the tragedy (clusters C9 and C5), and only later did their interest focus on political aspects (e.g., Barack Obama, Vladimir Putin in C17 and C18).

Step 3. Inter-source alignment: Next, an alignment algorithm explores possible matches across the three data sources N, T and W. For any event e_i, we eventually obtain three "aligned" meta-clusters m_i^N, m_i^T and m_i^W mirroring respectively the media coverage of the considered event and its impact on readers' communication and information needs.

Step 4. Generating a recommendation: The input to our recommender is the news meta-clusters m_i^N related to an event e_i first reported on day d_0 and extracted during an interval I : $d_0 \leq d \leq d_{0+x}$, where d_{0+x} is the day in which the query is performed by the journalist. The system compares the three aligned meta-clusters

Table 1: The Twitter meta-cluster capturing the Malaysia Airlines flight crash event and its composing clusters

Clusters
C9 [tragic, crash, tragedi, Ukraine 1.0, Malaysia_Airlines 0.6, Airline 0.66, Malaysia_Airlines_Flight_17 0.65, Malaysia 0.60, Russia 0.51, Aviation_accidents_and_incidents 0.36, Airliner 0.35, Malaysia_Airlines_Flight_37 0.28, United_States 0.21, Tragedy 0.20, Boeing_777 ...]
C5 [tragic, tragedi, Airline 1.0, Malaysia_Airlines_Flight_17 0.97, Malaysia_Airlines 0.70, Malaysia 0.58, Ukraine 0.40, Twitter 0.39, Gaza_Strip 0.32, CNN 0.26, Tragedy 0.25, God 0.24, Airliner 0.22, Israel 0.22, Malaysia_Airlines_Flight_370 0.22, Netherlands 0.21,..]
C17 [tragedi, tragic, Malaysia_Airlines_Flight_17 1.0, Airline 0.89, Malaysia_Airlines 0.62, Malaysia 0.54, Gaza_Strip 0.42, Twitter 0.38, Ukraine 0.38, Hamas 0.33, Barack_Obama 0.32, Israel 0.29, Vladimir_Putin 0.27, God 0.26, CNN 0.25, Hell 0.25, Airliner 0.23, Malaysia_Airlines_Flight_37 0.20,...]
T_C18 [tragedi, tragic, Malaysia_Airlines_Flight_17 1.0, Airline 0.98, Malaysia_Airlines 0.80, Tragedy , Malaysia 0.54, Gaza_Strip 0.50, Ukraine 0.48, Hamas 0.408, Israel 0.38, Barack_Obama 0.7, Twitter 0.37, Vladimir_Putin 0.36, CNN 0.32, Airliner 0.28, Malaysia_Airlines_Flight_370 0.26, Hell 0.252, ...]
Meta-cluster
[tragedi 0.22, tragic 0.22, airline 0.20, malaysia_airlines_flight_17 0.20, ukraine 0.19, malaysia_airlines 0.19, malaysia 0.17, russia 0.129, tragedy 0.12, vladimir_putin 0.12, airliner 0.12, crash 0.12, gaza_strip 0.11, barack_obama 0.11, aviation_accidents_and_incidents 0.11, cnn 0.106, malaysia_airlines_flight_370 0.10, god 0.10, ...]

to identify in m_i^T and m_i^W the set of most relevant entities [1], respectively E_i^T and E_i^W. A set of entities E_i in either T or W is further partitioned in $R_i^{in\_news}$ and R_i^{novel}, representing, respectively, the event-related topics already discussed, and those not yet considered in news items. The first set is interesting for journalists in order to understand which topics mostly attracted the attention of the public, while the second set includes event-related, but still uncovered, topics that W^3 recommends to discuss. For example, the following is a recommendation generated from the analysis of Wikipedia page views, related to Scottish Independence elections on September 17th, 2014: *[scotland, wales, alex_salmond, united_kingdom, scottish_national_party, flag_of_scotland, william_wallace, countries_of_the_united_kingdom, mary_queen_of_scots, tony_blair, braveheart, flag_of_the_united_kingdom, republic_of_ireland]*. When comparing these entities with the aligned news meta-clusters, the set of *novel* entities R_i^{novel} is: *[flag_of_scotland, william_wallace, countries_of_the_united_kingdom, mary_queen_of_scots, tony_blair, braveheart]* and all the others are also found in news.

Step 5. Classification of information and communication needs: In addition to recommendations, we automatically assign a category both to event clusters m_i^N in news, and to related entities in Twitter and Wikipedia aligned meta-clusters m_i^T and m_i^W, in order to detect recurrent discussion topics and search patterns in relation to specific event *types*. To do so, we exploit both BabelNet (Navigli and Ponzetto, 2010),

[1]We used TextRazor https://www.textrazor.com and DataTXT https://dandelion.eu/semantic-text/entity-extraction-demo/ to extract entities respectively from Twitter and news items

dataset	**# clusters**	**# m.clusters**	**av. size m.clusters**
News	9396	829	122.46
Twitter	4737	413	136.76
Wikipedia	5450	535	6.44

Table 2: Statistics on data and results

a large-scale multilingual semantic network[2], and the Wikipedia Category graph.

3 Discussion

To conduct our study, we created three datasets: Wikipedia PageViews (W), On-line News (N) and Twitter messages (T). Data was collected during 4 months from June 1st, 2014 to September 30th. Table 2 shows some statistics. Note that Wikipedia clusters are smaller, since cluster members are only named entities (page views).

We defined the following evaluation framework: i) Given an event e_i and related news $n_i \in N_i$, we generate recommendations as explained in Step 4, in a selected interval prior to the day of the query. ii) *Automated evaluation*: we we select the top K scored recommendations and measure the *saliency* of $R_i^{in\_news}$ and *serendipity* of R_i^{novel} in an automated fashion, and we compare the performance against a primitive recommender, in analogy with (Murakami et al., 2008) and (Ge et al., 2010); ii) *Manual evaluation*: we select the top K scored recommendations in R_i^{novel} for a restricted number of 21 high-impact world-wide events, and we perform manual evaluation using the *Crowdflower.com* platform, providing detailed evaluation guidelines for human annotators. Using this ground truth, we measure the global *serendipity* of W^3 recommendations.

[2]http://babelnet.org/about

3.1 Automated Evaluation

We first build two *primitive recommenders* (PRs) for Wikipedia and Twitter, which we use as a baseline. The input to a PR is the same as for W^3 (see Step 4).

Wikipedia PR: The Wikipedia PR is based on finding connected components of the Wikipedia hyperlink page graph (like in (Hu et al., 2009)), when considering only the topmost visited pages in a temporal slot. More precisely, for each day d in the interval $I' : d_{0-x} \leq d \leq d_{0+x}$ [3], we select the top $H \geq K$ visited named entities of the day E_d^W. Entities are ranked by frequency of page views[4]. Next, we create clusters c_j^d obtained by extracting the connected components of E_d^W in the Wikipedia hyperlink graph. Let $C^{I'}$ be the set of all clusters $c_j^{I'}$ in I'. From this set, we select the top r clusters based on the Jaccard similarity with news meta-clusters m_i^N. A "primitive" recommendation for event e_i on day d_{0+x} is the set PR_i^W of topmost K ranked entities in the r previously selected clusters. Like in W^3 recommendations, PR_i^W is a ranked list of entities some of which are also found in m_i^N, and some others are novel.

Twitter PR: For each entity $e \in m_i^N$ we retrieve and recommend the top K co-occurring entities in tweets in the considered interval.

Note that both primitive recommenders are far from being naive. A hyperlink graph to characterize users' intent in Wikipedia search is used in (Hu et al., 2009) (although the authors use Random Walks rather than connected components analysis to identify related pages). Co-occurrences with top ranked news terms has been used in (Weiler et al., 2014) to track on Twitter the evolution and the context around events. We generate recommendations using four systems: $W^3(T)$, $W^3(W)$, $PR(T)$ and $PR(W)$. The first two originate from What To Write and Why when applied to Twitter and Wikipedia, respectively. The second two are generated by the two primitive recommenders described above. For all systems, we consider the first K top ranked entities, as we said.

To assess the quality of "not novel" recommended entities in W^3 (and similarly for the other systems), for any $r_j \in R_i^{in\_news}$ we retrieve all the news N_i related to m_i^N meta-clusters, and compute the *saliency* of r_j as follows:

$$saliency\,(r_j, n_i) = \beta \times occ^{title}\,(r_j, n_i) + (1-\beta) \times occ^{snip}\,(r_j, n_i) \quad (1)$$

where $n_i \in N_i$, $occ^{title}(r_j, n_i)$ is the number of occurrences of r_j in the title of n_i, while $occ^{snip}(r_j, n_i)$ is the number of occurrences of r_j in the text snippet of n_i and β has been experimentally set to 0.7. The intuition is that recommended entities in $R_i^{in\_news}$ *are salient if they frequently occur in the title and text of news snippets*, where occurrences in the title have a higher weight. The total saliency of r_j is then:

$$saliency\,(r_j) = \frac{\sum_{n_i \in N_i} saliency(r_j, n_i)}{|N_i|} \times IDF\,(r_j) \quad (2)$$

where $IDF(r_j)$ is the inverse document frequency of r_j in all news of the considered temporal slot, and is used to smooth the relevance of terms with high probability of occurrence in all documents. The average saliency of $R_i^{in\_news}$ is:

$$saliency\left(R_i^{in\_news}\right) = \frac{\sum_{r_j \in R_i^{in\_news}} saliency(r_j)}{\left|R_i^{in\_news}\right|} \quad (3)$$

To provide an estimate of the *serendipity* of novel recommendations, we compute the NASARI similarity (Camacho-Collados et al., 2016) of entities $r_k \in R_i^{novel}$ with in-news entities $r_j \in E_i^N$ and we weight these values with the saliency of r_j. The intuition is that *serendipitous recommendations are those concerning topics which have not been discussed so far in on-line news, but are highly semantically related with highly salient topics in news*:

$$serend.\left(r_k \in R_i^{novel}\right) = \frac{\sum_{r_k \in R_i^{novel}, r_j \in E_i^N} (NASARI(r_k, r_j) \times saliency(r_j))}{|R_i^S|} \quad (4)$$

Note that this formulation is not conceptually different from other measures used in literature (e.g, (Tran et al., 2015),(Murakami et al., 2008)), that commonly assign a value to novel recommendations proportionally to their relevance and informativeness, however given the absence of prior knowledge on users' choices, we assume that semantic similarity with salient entities in news items is a clue for relevance.

In Table 3 we summarize the results of our experiments, that we run over the full dataset (see

[3] Since rumors on an event can be anticipated wrt the day d_0 in which the first news item is published

[4] Note that E_d^W could be straightly used for recommendation, however it would be an excessively rough strategy.

Table 3: Percentage difference in performances between W^3 and PRs on Twitter and Wikipedia

Source	Saliency	Serendipity	F-Value
Twitter d0	-28%	+91%	+15%
Wikipedia d0	+172%	+656%	+371%
Twitter d2	-34%	+81%	+8%
Wikipedia d2	+106%	+547%	+286%

Table 2). We set the maximum number of provided recommendations $K = 10$ for Wikipedia (where clusters are smaller) and $K = 50$ for Twitter. All recommendations are gathered either the same day (d_0) of the first news item on the event e_i, or two days after ($d_2 = d_0 + 2$). In analogy with (Murakami et al., 2008) and (Ge et al., 2010), we show the percentage difference in performance between W^3 and Primitive Recommenders (PRs). Besides *saliencey* and *serendipity*, we also compute the harmonic mean between the two (the F value). The Table shows that for Wikipedia, W^3 outperforms the PR both in saliency and serendipity (it is up to 656% more serendipitous than the baseline) while in Twitter, W^3 shows better serendipity (+91%) but lower salience (-28%). Comparatively, the performance of W^3 is much better on Wikipedia than on Twitter, probably due to the limited evidence provided by the 1% available traffic. We also noted that two days after the main event (x=2), both serendipity and saliency only slightly decrease showing that newswires have covered only a small portion of users' communication and information needs.

3.2 Manual Evaluation

In manual evaluation, in order to start from a clean representation of each event for all systems, we selected 21 relevant (with topmost number of news, tweets and wikipedia views) events in the considered 4-months period, and we manually identified the relevant news items N_i for each event e_i in a $\pm$ 1-day interval around the event peak day d_0. An excerpt of 5 events is shown in Table 4. We then automatically extracted named entities from these news items.

For each of the four systems $W^3(T)$, $W^3(W)$, $PR(T)$ and $PR(W)$ and each event e_i, we generate the first $K = 5$ *novel* recommendations, and we use the *CrowdFlower.com* platform to assess the relevance of these recommendations[5]. For each item of news, annotators are asked to decide

[5]The saliency of $R_i^{in\_news}$ is well assessed by formula (2)

Table 4: Excerpt of selected events

Date	Event
11/06/2014	Al-Qaeda Faction Seizes Key Iraqi City
14/06/2014	England vs. Italy at the 2014 World Cup
30/06/2014	Limiting Rights: Imposing Religion on Workers
05/07/2014	Wimbledon: Novak Djokovic and Roger Federer Reach Men's Final
...	
22/09/2014	Nasa's Newest Mars Mission Spacecraft Enters Orbit Around Mars

if an entity IS or IS NOT relevant with reference to the reported news ("not sure" is also allowed). "Relevant" means that either the entity is semantically related to the domain of the news, or that it is factually related. The task was run on April 23rd, 2017, and we collected 1344 total judgements. To compute the performance of each system, we use the Mean Average Precision (MAP)[6], which takes into account the rank of recommendations. The results are shown in Table 5, which shows, in agreement with the automated evaluation of Table 3, a superiority of W^3 and also confirms that the difference between W^3 and the primitive recommender is much higher in Wikipedia than in Twitter. We also note that the absolute performance of the recommender is higher in Twitter, which is not in contradiction with Table 3, since here we are focusing on world-wide high impact news, those for which our 1% Twitter stream provides sufficient evidence to obtain clean clusters, such as those in Table 1.

3.3 Analysis of Information Needs

To analyze readers' behavior more systematically, we classified events meta-clusters automatically, extending the work in (Košmerlj et al., 2015), were the authors have manually classified 13,883 Wikipedia event-related articles in 9 categories. Furthermore, we classified recommendations, i.e., tokens in m_i^T and m_i^W meta-clusters associated to each event e_i, using BabelNet hypernymy (*IS_A*) relations[7], and their mapping onto Wikipedia Categories. In Figure 4 we plot the category distribution of Wikipedia articles (more specifically, we plot only novel recommendations extracted by W^3) that readers have accessed in correspondence of different event types. The Bubble plot shows several interesting patterns: for example,

[6]https://www.kaggle.com/wiki/MeanAveragePrecision
[7]http://babelnet.org/about

Source	W^3	BR
Twitter	0.934	0.851
Wikipedia	0.789	0.363

Table 5: MAP (mean average precision) of compared systems in Crowdflower.com evaluation (on a sample of 21 breakings news)

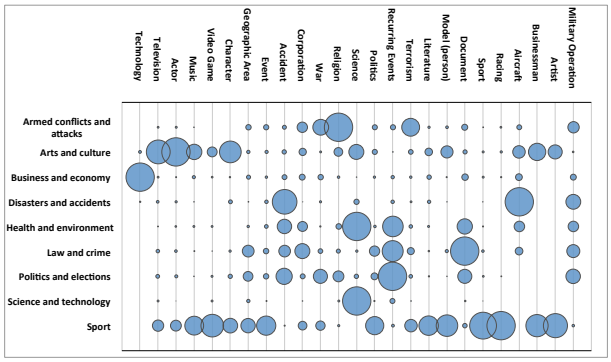

Figure 4: Bubble plot of event categories and associated information needs (during Summer 2014)

Religion is the main searched category for events classified as Armed Conflicts and Attacks, mirroring the fact that religion is perceived as being highly related with latest world-wide conflicts. Accordingly, users try to deepen their knowledge on these aspects. Disasters and accidents mostly include members in the same Wikipedia category (Disasters) and also Aircraft, since the Malaysia crash was the dominating event in the considered period. Business and Economy draw the attention of readers mostly when related to Technology, e.g., new devices being launched. Law and Crime events induce in readers the need to find out more about specific laws and treaties (the category Documents). Finally, we note that Sport is the event category showing the highest dispersion of information needs. While many of the bubbles in Figure 4 indeed show real information needs (e.g, VideoGames refers to the many sport games launched on the market, Model (person) refers to gossip about football players, and in general all people and media related categories refer to the participation of celebrities in sporting events), a number of bubbles can be considered as noise, e.g., Literature, Politics. In fact, Sport was the dominating event type during the considered period (2014 World Football Cup), therefore it is reasonable that sport-related clusters are those cumulating the highest number of system errors.

References

R. Brooker and T. Schaefer. 2005. *Public Opinion in the 21st Century: Let the People Speak?*. New directions in political behavior series. Houghton Mifflin Company.

J. Camacho-Collados, M. Taher Pilehvar, and R. Navigli. 2016. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence* 240:36–64.

M. Ge, C. Delgado-Battenfeld, and D. Jannach. 2010. Beyond accuracy: Evaluating recommender systems by coverage and serendipity. In *Proc. of RecSys'10*. pages 257–260.

P. J. Gloviczki. 2015. *Journalism in the Age of Social Media*, Palgrave Macmillan US, pages 1–23.

J. Hu, G. Wang, F. Lochovsky, J. Sun, and Z. Chen. 2009. Understanding user's query intent with wikipedia. In *Proc. of WWW'09*. pages 471–480.

M. Knight. 2012. Journalism as usual: The use of social media as a newsgathering tool in the coverage of the iranian elections in 2009. *Journal of Media Practice* 13(1):61–74.

A. Košmerlj, E. Belyaeva, G. Leban, M. Grobelnik, and B. Fortuna. 2015. Towards a complete event type taxonomy. In *Proc. of WWW'15*. pages 899–902.

J. Lehmann, B. Gonçalves, J. J. Ramasco, and C. Cattuto. 2012. Dynamical classes of collective attention in twitter. In *Proc. of WWW'12*. pages 251–260.

J. Leskovec, L. Backstrom, and J. Kleinberg. 2009. Meme-tracking and the dynamics of the news cycle. In *Proc. of KDD '09*. pages 497–506.

T. Murakami, K. Mori, and R. Orihara. 2008. Metrics for evaluating the serendipity of recommendation lists. In *Proc. of the 2007 Conf. on New Frontiers in AI*. Springer, pages 40–46.

R. Navigli and S. P. Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proc. of the 48th Annual Meeting of the ACL*. pages 216–225.

G. Stilo and P. Velardi. 2016. Efficient temporal mining of micro-blog texts and its application to event discovery. *Data Min. Knowl. Discov.* 30(2):372–402.

T. Tran, C. Niedere, N. Kanhabua, U. Gadiraju, and A. Anand. 2015. Balancing novelty and salience: Adaptive learning to rank entities for timeline summarization of high-impact events. In *Proc. of CICM'15*. Springer, volume 19, pages 1201–1210.

A. Weiler, M. Grossniklaus, and M.H. Scholl. 2014. Event identification and tracking in social media streaming data. In *Proc. of the Work. of the EDBT/ICDT'14*. CEUR-WS, pages 282–287.