Capturing users' information and communication needs for the press officers

Giovanni Stilo[1], Christian Morbidoni[2],
Alessandro Cucchiarelli[2], and Paola Velardi[1]

[1] Sapienza University of Rome, Italy,
stilo@di.uniroma1.it, velardi@di.uniroma1.it,
[2] Università Politecnica delle Marche, Ancona, Italy,
c.morbidoni@univpm.it, a.cucchiarelli@univpm.it

Abstract. The way in which people acquire information on events and form their own opinion on them has changed dramatically with the advent of social media. For many readers, the news gathered from online sources becomes an opportunity to share points of view and information within the micro-blogging platforms such as Twitter, mainly aimed at satisfying their communication needs. Furthermore, the need to deepen the aspects related to the news stimulates a demand for information that is often met through online information sources, such as Wikipedia. This behaviour has also influenced the way in which journalists write their articles, requiring a careful assessment of what actually interests readers. The goal of this article is to define a methodology for a recommender system able to suggest to the journalist, for a given event, the aspects still uncovered in news articles in which the readers' interest focuses. The basic idea is to characterize an event according to the echo it had 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 identifies as recommendations the concepts that emerge as topic of interest from Twitter and Wikipedia, not covered in the published news articles.

Keywords: recommender system, wikipedia, twitter, social networks, media, press agents, events detection

1 Introduction

In a recent study on the use of social media sources by journalists [12] 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 [6]. When readers fell 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 behaviour on online knowledge bases, and their discussions on popular social platforms. What is more, on-line public's reaction to news is almost immediate [17] and even anticipated, as for the case, e.g., of planned media events and performances, or for disasters [16]. A related issue is the so called *fact-checking problem*, i.e. the need to validate the veracity of news. The growing relevance of this type of information need is also demonstrated by the recent announcement by Facebook of the Journalism Project to help fight fake news [31].

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) [1]. Futhermore, real-time analysis of public reaction to news may provide a useful feedback to journalists, such as highlighting aspects of a story that need to be further addressed, issues that appear to be of interest for the public but have been ignored, or even to help local newspapers echoing international press releases.

The aim of this paper is to present a methodology to effectively exploit social data sources for the purpose of news media recommendation and for summarizing the *outcome* of news stories on the public, defined in terms of the shorter-term effects that news can have, such as informing, engaging, and mobilizing audiences [21]. The purpose of the the recommender 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.

The paper is organized as follows: in section 2 we review related works, in section 3 we describe our dataset and additional resources used in our methodology, which is presented in section 4. Finally, section 5 is dedicated to the discussion of the preliminary results and section 6 contais the concluding remarks and the future work directions.

2 Related Works

To the best of our knowledge, this is the first system for recommending journalists what to write, focusing on presenting users' needs that come from different sources while keeping their original motivation (*information* and *communication*).

Many available studies are concerned with the task of predicting the response of social media to news articles [13] [27] rather than extracting users' interest related to news articles to help journalists focalize on additional, yet uncovered, aspects of a reported event. Other works analyze the symmetric problem of recommending news to social media users. Among these, the authors in [19] are concerned with the task of recommending articles to readers in a stream-based scenario, when large user-item matrixes are not available and time constraints are strict. In their work, they derive a number of statistics extracted from the PLISTA [11] dataset used during the ACM Recsys News Challenge 2013. They also compare performances of several existing recommending algorithms show-

ing that the precision of algorithms depends upon the particular news articles domain. The study in [22] also deal with real time recommenders. As before, the considered task is to recommend topical news to users. The authors present Buzzer, a system able to mine real time information from Twitter and RSS feeds and use overlapping keywords in most recent tweets and feeds as a basis for recommendation. Evaluation is performed on a small group of 10 participants over a period of 5 days.

A line of study closer to our work is concerned with the task of identifying social content related to a given event. In [28] the following task is considered: given a news article, find the Twitter messages that "implicitly" refer to the same topic, i.e. messages not including an explicit link to the considered article. They are interested in discovering utterances that link to a specific news article rather than the news event(s) that the article is about. First, the authors analyze the KL-divergence between the vocabulary of news articles (using the NT Times as a primary source) and various social media, such as Twitter, Wikipedia, Delicious, etc. They find that, unless part of the original article is copied in the message, which subsumes explicit reference, the vocabularies might be quite different. The method used by the authors is in three steps: they derive multiple query models from a given article, which are then used to retrieve utterances from a target social media index, resulting in multiple ranked lists that are finally merged using data fusion techniques. Evaluation is performed, in line with other scholars, using messages with explicit mention to an article, and then removing the mention. However, as observed by the same authors, evidence suggests that these messages often copy part of the article, an eventuality that could boost performances. In [15] the objective is to combine news articles and tweets to identify not only relevant events but also the opinions expressed by social media users on the very same event. Like us, they use the news article as the query, and tweets as the document collection. They use a latent topic model to find the most relevant tweets wrt a given news topic. Besides topic similarity, they use additional features such as recency, follower count etc, which are then combined using logistic regression or Adaboost. Relevance judgement for evaluating the system have been collected from 11 computer science students.

Only two papers aim to help journalists find relevant content in social media, as we do. In [3] the authors present a tool to help journalists at identifying eyewitnesses in the context of an event. In [30] a system is described to assist journalists in the use of social media. The authors use SVM to identify newsworthy messages on Twitter based on a manually annotated dataset. Their work however is concerned more with the design of a user interface to help journalists in digging into trending topics than on algorithms to extract such content automatically.

3 Datasets and Resources

To conduct our study, we have created three datasets: Wikipedia PageViews (W), On-line News (N) and Twitter messages (T). Data has been collected during 4 months from June 1st, 2014 to September 30th, 2014 in the following way:

1. **Wikipedia PageViews**: we downloaded Wikipedia page views statistics from the data dumps provided by the WikiMedia foundation[3]. We considered only English queries and we retained only those matching a Wikipedia document, removing redirected requests. Overall, we obtained 27.708.310.008 clicks on about 388 million pages during the considered period. An example is: *en Ravindra Jadeja 12 345* where the number of requests (12 in the example) refers to a time span of 1 hour.
2. **On-line News**: We collected news from GoogleNews (GN) [4] and HighBeam (HB) [5]. Due to existing limitations, we extracted at most 100 news per day from GN, while for HB we downloaded all available news. Each news item has a title, source, day of publication and an associated snippet, e.g., *GN - "8 1 2014","Bleacher Report",[6],"India Will Be Left Furious by the Ravindra Jadeja and James","On Friday, following a six-hour hearing in Southampton, judicial commissioner Gordon Lewis found both Anderson and Jadeja not guilty of breaching the ICC "*. Overall, we extracted 351,922 news from 88 sources in GN and 1,181,166 from 325 sources in HB during the considered period. Snipptes were about 25 words long in average.
3. **Twitter messages**: we collected 1% of Twitter traffic, the maximum freely allowed traffic stream using the standard Twitter API [7]. Overall, we collected 235 million tweets, e.g., *"James Anderson and Ravindra Jadeja have both been found not guilty by judicial commissioner Gordon Lewis. #Cricket #ENGvIND"*.

Furthermore, in this research we used the following resources:

1. NASARI embedded semantic vectors for Wikipedia pages, generated as described in [2]. We used the second release[8] covering 4.40 million Wikipedia pages.
2. Dadelion Entity Extraction API (DataTXT)[9] and TextRazor[10]. Both are commercial tools providing entity recognition REST APIs that, given a text snippet, identify, disambiguate and link named entities to Wikipedia. DataTXT is based on previous research [4], and has been recently further developed and engineered [24].

[3] `https://dumps.wikimedia.org/other/pagecounts-raw/`
[4] `https://news.google.com/`
[5] `https://www.highbeam.com`
[6] `http://bleacherreport.com/articles/2148916`
[7] `https://dev.twitter.com/docs/streaming-apis`
[8] `http://lcl.uniroma1.it/nasari/\#two`
[9] `https://dandelion.eu/semantic-text/entity-extraction-demo/`
[10] `https://www.textrazor.com`

4 Proposed Method

4.1 Overview

Our methodology is based on the following steps:

1. *Events identification*: first, we identify breaking news using SAX^{++}, an enhanced version of the temporal mining algorithm presented in [25] and [26]. Terms related to breaking news are extracted from on-line news (N), Twitter (T) and Wikipedia page-views (W), respectively in representation of media news providers and of public's communication and information needs. Terms are grouped into clusters represented as ranked lists of words and named entities;
2. *Intra-source clustering*: within each data source (N, T and W) we attempt to group terms clusters with the same peak day and related to the same breaking news, creating *meta-clusters*.
3. *Inter-source alignment*: an alignment algorithm explores possible matches across the three data sources N, T and W. For breaking news n_i, we thus obtain three meta-clusters mirroring respectively the media coverage of the considered event, and its impact on readers' communication and information needs.
4. *Identification of missing information*: the final step is comparing the three meta-clusters to identify in T and W meta-clusters the most relevant words and named entities, considering both their *impact* on users and *novelty* wrt to what has already been published in N. These terms can then be used to recommend journalists additional aspects to cover or deepen when following up on a news item. At the current stage of our research, we are in the phase of defining an effective method to automatically derive, rank and evaluate such recommendations. However, in section 5 we discuss preliminary results in this direction.

4.2 Events identification

Algorithm In this section we shortly summarize the SAX^{++} algorithm, a multi-thresholds version of the SAX^{*} algorithm, presented in [25] and [26], with pluggable domain dependent variants. The phases of the new algorithm, named SAX^{++}, are the followings:

1. The temporal series associated to terms/wikipages and hashtag are sliced into sliding windows of length W, normalized and converted in symbolic strings using Symbolic Aggregate ApproXimation [18]. The parameters of this step are the dimension of the alphabet $|\Sigma|$ and the number $\frac{W}{\Delta}$ of partitions of equal length Δ.
2. Using a set of seed keywords related to known events, we convert their temporal series into symbolic strings and automatically learn regular expressions representing common usage patterns. For example, with an alphabet of 3

symbols, we learn the following expression:

$$(a + [bc]?[bc][bc]?a+)?(a + [bc]?[bc][bc]a*)?$$

which captures all the temporal series with one or two peaks and/or plateaus in the analyzed window. These are common temporal patterns of breaking news. Only terms/wikipages with frequency higher than a threshold f' and hashtags with frequency higher than a threshold f'' and matching the learned regular expressions are considered in the subsequent steps. These are hereafter denoted as *active tokens*.

3. Tokens are analyzed in sliding windows W_i and the detected active tokens are clustered in each W_i using a bottom-up hierarchical clustering algorithm with *complete linkage* [8] and similarity threshold δ.
4. In order to cope with *temporal collision* (i.e. co-occurring but unrelated events) an additional cluster splitting step is performed, which considerably improves clustering results. First, we build a graph $G = (V, E)$ for each cluster c previously detected by SAX*[11] in a window W. A graph G is built associating each vertex $v \in V$ with a token t_i and adding an edge (t_i,t_j) if token t_i and t_j:
   - co-occurs in a number of documents greater than a threshold τ (for social networks or for news domain);
   - or show "sufficient" semantic similarity, as derived by an external resource (for Wikipedia domain); specifically, we use NASARI vectors to compute similarity between two Wikipedia Pages that must be higher than a threshold *nas* [12];

   Next, we detect connected components in G. Each connected component is a split of the original cluster. Extracting connected components from a graph is a well-established problem [7] and does not have a heavy impact on the computational cost of the entire algorithm [23], mainly due to the fact that the size of the graphs, when pruning non-active tokens, is relatively small.

Entities Enrichment Clusters extracted from News and Twitter are made out of single words, however, including named entities would be preferable both for relevance (named entities are pervasive in texts and especially in news) and to compare results among different data sources (N, W and T). Named Entity expansion is applied only as post-processing phase over all the produced clusters, because extracting names on the full Twitter and News stream is computationally very demanding. Starting from all the tokens included in a cluster, we retrieved the d best matching documents querying the original data-set (tweets or news). We performed entity recognition for those documents only. Next, we consider all the named entities recognized in these d documents, we score them by occurrence, from 0 to 1, and add the top e entities to the related clusters.

[11] for efficiency we maintain just one graph at a time in memory

[12] Experiments shows that reeling on the Wikipedia graph doesn't split clusters effectively.

Note that for every source (Twitter, News) we adopted the entities tagger that best performs on the input text (see subsection 4.2).

Parameter Setting In our experiments we ran SAX^{++} using different parametrizations for each of the three sources, then we manually evaluated the resulting clusters considering 10 known events to select the best parameters configuratios, shown in table 1. The value of the Δ parameter (the time granularity) was set to 24 hours in all the datasets as this is the mininum granularity in news, where the exact time of publication is not present.

As tweets and news are very different in nature (length and style), before performing the Named Entity Enrichment phase we processed some sample documents with two available systems, DataTXT and Textrazor, and selected the tool which provided more accurate results: DataTXT for news articles an TextRazor for tweets. We then experimentally set optimal values for d and e (see table 1).

Table 1. $SAX+$ parameters settings used for the different sources

Source	$\lvert\Sigma\rvert$	Δ	f'	f''	δ	τ	nas	d	e
Twitter	2	24	250	25	0.25	6	-	200	100
News	2	24	1000	-	0.12	20	-	50	50
Wikipedia	2	24	50000	-	0.12	-	0.01	-	-

4.3 Intra-source clustering

Since SAX^{++} works on sliding windows, the same event is usually captured by different clusters extracted from adjacent windows. In order to obtain a better characterization of an event, we aggregate similar clusters with the same peak day, forming meta-clusters which contain the most relevant terms for the event. When considering clusters of the same events we note that the *pivot cluster*, i.e. the cluster whose peak day is closer to the centre of the considered window, shows a higher precision as compared to those clusters twith a peak day closer to the extremes of the window.

First, we select all the *pivot clusters* P^d from the set of clusters C^d with the same peak day. We then build a similarity graph $G^J = (C^d, E^\gamma)$ (where an edge is built if two clusters have a Jaccard similarity coefficient higher than a threshold γ), comparing every *pivot cluster* $p \in P^d$ with all the clusters of the original set, $c \in C^d \setminus P^d$. Each extracted connected components from the graph G^J, represents a meta-clusters composed by the identified clusters.

To represent each meta-cluster in a compact and readable way, we create a scored list of all the terms (words, hash-tags ad named entities) contained in the meta-cluster. The score of a term is calculated as the normalized ratio between

the sum of the *terms scores* in all the clusters and the number of clusters. We refer to such a scored set of terms as the *meta-cluster signature*.

4.4 Inter-source Alignment

The subsequent phase aligns meta-clusters from the three sources (T, N and W) corresponding to the same popular event. We use as "seeds" the News meta-clusters, and find the most similar meta-clusters from Twitter and Wikipedia. As there might be a slight difference in peak days in different data sources for the same event, we use a similarity measure $TempSym$ with two components: a content based component and a time based one. The content based component is the Jaccard similarity between terms of the meta-cluster's signature, while the time based component takes into account the distance between the two peak days: the closer the two, the higher the similarity. Considering two meta-clusters m^a and m^b we use the following formula:

$$TempSym(m^a, m^b) = Jaccard(m^a, m^b) \times \alpha^{(|peak(m^a) - peak(m^b)|)}$$

Where alpha is a decay coefficient. The smallest is alpha, the less past clusters are considered similar.

Table 2. Results statistic for the three data sources

dataset	**# clusters**	**# meta-clusters**	**average size of meta-clusters**
News	9396	829	122.46
Twitter	4737	413	136.76
Wikipedia	5450	535	6.44

In table 2 we show some statistics of the obtained results wrt the three data sources: the total number of clusters extracted by SAX^{++}, the total number of meta-cluster obtained running the intra-source clustering and the size of the meta-clusters expressed as the average number of terms in such meta-cluster's signatures.

5 Discussion of preliminary results

We present here an example of the result produced by the intra-source clustering phase and three examples produced by the inter-source alignment algorithm; these examples allow us to present a qualitative discussion of our results and to show the difficulty of a systematic evaluation.

In table 3 we show a sample meta-cluster obtained as output of the intra-source clustering step, along with its composing clusters and the scored set of

terms derived as discussed above. The meta-cluster clearly refers to a popular event: the crash of the Malaysia Airlines flight 17 on 17 July 2014[13].

Table 3. The Twitter meta-cluster capturing the Malaysia Airlines flight crash event and its composing clusters

Clusters
T_1405036800000_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, Vladimir_Putin 0.27, United_States 0.21, Tragedy 0.20, Boeing_777 0.19924139305113062 ...]
T_1405296000000_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, Barack_Obama 0.32, Vladimir_Putin 0.27, CNN 0.26, Tragedy 0.25, God 0.24, Airliner 0.22, Israel 0.22, Malaysia_Airlines_Flight_370 0.22, Netherlands 0.21, ...]
T_1405123200000_C17 [tragedi, tragic, Malaysia_Airlines_Flight_17 1.0, Airline 0.89, Malaysia_Airlines 0.62, Malaysia 0.54, Tragedy 0.47, 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_1404950400000_C36 [crash, russian, tragedi, tragic, ukrain, Ukraine 1.0, Malaysia_Airlines 0.42, Malaysia 0.400, Russia 0.40, Airline 0.36, Airliner 0.18, Malaysia_Airlines_Flight_17 0.18, Aviation_accidents_and_incidents 0.13, Vladimir_Putin 0.114, Kuala_Lumpur 0.09, Eastern_Ukraine 0.09, Boeing_777 0.075, Jet_aircraft 0.068, ...]
T_1405209600000_C5 [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, God 0.25, ...]
Meta-cluster signature
17 Jul 2014 [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, ...]

In table 4 we show some examples of the inter-source alignment algorithm. Three popular events from different domains are considered: the celebration of the USA Independence Day, the FIFA 2014 World Cup final match and the Malaysia Airlines crash. For each event we show the news meta-cluster signature, used as seed in the alignment algorithm, and the most similar meta-cluster's signatures emerged in Twitter and Wikipedia. In addition, we mark in bold the novel terms in T and W, which could be the candidate recommended terms. Due to lack of space, we do not show all the terms, except for the Wikipedia where the meta-clusters are smaller on average.

Some emerging terms, especially in Wikipedia cluster, clearly highlight information needs related to the corresponding events. For example, the emerging of terms like *american_revolutionary_war* and *the_start-spangled_banner* suggests a keen interest to deepen the knowledge of the historical events that led to the

[13] BBC page of the event: http://www.bbc.com/news/world-europe-28357880

Table 4. Examples of aligned meta-clusters for popular events

Independence Day (04 Jul 2014)
News
03 Jul 2014 [united_states_declaration_of_independence 0.25, independence_day 0.24, life,_liberty_and_the_pursuit_of_happiness 0.24, natural_and_legal_rights 0.14, continental_congress 0.13, all_men_are_created_equal 0.12, thomas_jefferson 0.12, self-evidence 0.12, washington,_d.c. 0.12, human_events 0.12, fireworks 0.11, united_states_house_of_representatives 0.10 ...]
Twitter
03 Jul 2014 [independence_day 0.29, textbffourth 0.28, **safe** 0.22, **bbq** 0.16, **grill** 0.16, **sparkler** 0.15, fireworks 0.14, united_states 0.14, **barbecue** 0.13, **parad** 0.12, **coffee** 0.10, god 0.10, **pittsburgh_steelers** 0.10, **heinz_field** 0.09, **canada** 0.09, ...]
Wikipedia
Jul 04 2014 [**the_star-spangled_banner** 0.16, independence_day 0.16, **american_revolutionary_war** 0.12]

FIFA World Cup 2014 final match (13 Jul 2014)
News
13 Jul 2014 [germany_national_football_team 0.30, fifa_world_cup_0.27, overtime_(sports)_0.27, argentina 0.26, germany 0.26, argentina_national_football_team 0.24, brazil_0.24, mario_goetze 0.24, rio_de_janeiro 0.23, maracana_stadium 0.22, 2014_fifa_world_cup 0.22, brazil_national_football_team 0.21, lionel_messi 0.21 ...]
Twitter
13 Jul 2014 [**shakira** 0.33, gervsarg 0.29, **kramer** 0.29, **gerarg** 0.29, **argvsger** 0.29, germany_national_football_team 0.29, lionel_messi 0.28, argentina_national_football_team 0.26, argentina 0.24, champion 0.22, germany 0.21, fifa_world_cup 0.21, **neuer** 0.19, **ceremoni** 0.19 ...]
Wikipedia
13 Jul 2014 [**2018_fifa_world_cup** 0.19, **2026_fifa_world_cup** 0.19, **2022_fifa_world_cup** 0.12]

Malaysia Airlines fligth 17 crash (17 Jul 2014)
News
17 Jul 2014 [malaysia 0.38, ukraine 0.33, malaysia_airlines 0.33, surface-to-air_missile 0.30, kuala_lumpur 0.28, eastern_ukraine 0.27, malaysia_airlines_flight_17 0.27, boeing_78 0.272, amsterdam 0.27, ...]
Twitter
17 Jul 2014 [malaysia 0.37, aircraft 0.31, plane 0.29, **condol** 0.29, malaysian 0.29, airlin 0.29, missil 0.29, passeng 0.29, ukraine 0.29, malaysia_airlines 0.28, ukrain 0.28, ... , **tragedi** 0.12, eastern_ukraine 0.11, **aviation_accidents_and_incidents** 0.11, boeing_777 0.11, missile 0.11, passenger 0.11, kuala_lumpur_international_airport 0.10, **interfax** 0.09, **jet_aircraft** 0.09, , ... , **russian_language** 0.08, ..., **tragic** 0.06, **expens** 0.06, ...]
Wikipedia
17 Jul 2014 [**siberia_airlines_flight** 0.93, boeing 0.85, malaysia_airlines_flight 0.73, **iran_air_flight** 0.71, **korean_airlines_flight** 0.44, **pan_am_flight** 0.41, kuala_lumpur 0.28, surface-to-air_missile 0.26, malaysia_airlines 0.26, malaysia 0.24, buk_missile_system 0.24, ukraine 0.12, **bermuda_triangle** 0.06, **2014_ crimean_crisis** 0.06]

US independence and of the US national anthem, respectively. These could be topics that are worth deepening, e.g., in editorials. Looking at T meta-clusters, the emerging of popular terms like *bbq*, *grill* or *parad* (stem of *parade*) in tweets immediately before Independence Day may simply suggest that most people are preparing to celebrate, while other terms like *pittsburg_steelers* and *heinz_field* refer to co-occurring related sports events and could be reasonably labelled as noise.

Looking at the second event, terms like *(2018—2022—2026)_fifa_world_cup* in the W meta-cluster expose a widespread interest in future editions of the football World Cup, which, again, could suggest related topics to be deepened. In the T meta-cluster, the appeareance of the term *shakira*, referring to the popular singer, in association with the FIFA football match seems apparently unrelated. However "googling" the term highlights a strong connection, as the singer sang the theme song of the 2014 World Cup during the FIFA world cup closing ceremony; *ceremoni* is another term in the same cluster, confirming this interpretation. The terms *gerarg* and *argvsger* are popular hashtags used to comment the match on Twitter; while not novel per-se, finding relevant hashtags for an event may prove useful in some contexts.

Around the time of the Malaysia Airlines crash it is not surprising that most people are encouraged to check Wikipedia about similar incidents in the past, e.g., Siberia Airlines flight 1812, shot down by the Ukrainian Air Force over the Black Sea in 2001, and about somehow related topics, e.g. *bermuda_triangle*. Finding past similar events is a common information need, frequently highlighted in our data. Terms like *condol* and *tragic* mirror a popular mood emerging among Twitter users, while for other terms it is hard to say if they are noisy or not, e.g. *expens*. Finally, the term *interfax*, apparently unrelated, turned out to be related to the event, since Interfax is a Moscow-based wire agency which reported that Ukrainian rebel forces had the airplane black boxes and they had agreed to hand them over to the Russian-run regional air safety authority; this news sub-topic captured the attention on Twitter.

6 Conclusions and future work

In this paper we presented a methodology to derive recommendations for journalists based on the detection and analysis of readers' information needs on Wikipedia and communication needs on Twitter. Preliminary experiments suggest that our methodology succeeds in aligning manifestations of interest wrt to highly popular events in all three different information sources: online news, twitter and wikipedia. No strong conclusions can be drawn from our preliminary experiments, even if results suggest that many relevant and recurrent behaviours can be extracted by systematically comparing the three sources.

Future works are focused on defining an effective method to extract missing information that might provide useful insight for press agents, from T and W aligned meta-cluster of a given News $n \in N$. Measuring the quality of the results in a meaningfull way is still an open issue. Specifically, we want to identify

terms that expose *relevant* and *novel* topics wrt to what have already been published. Several works [5, 9, 10, 14, 20, 29] try to formally define how to evaluate relevance and novelty, but this task still remain difficult even for humans, since connections among topics might not be evident and what may seem noise at first glance, turn out to be relevant and interesting after an in depth (and time-consuming) analysis. On the other hand, trying to reduce noise, e.g. discarding loosely connected topics, could dramatically affect novelty.

7 Acknowledgments

We thank SpazioDati and TextRazor for granting us the use of their entity recognition API beyond the non commercial use limit. We would also like to thank Giacomo Marangoni for his kind support in developing the workflow related to the Wikipedia source.

References

1. Brooker, R., Schaefer, T.: Public Opinion in the 21st Century: Let the People Speak? New directions in political behavior series, Houghton Mifflin Company (2005)
2. Camacho-Collados, J., Pilehvar, M.T., Navigli, R.: Nasari: a novel approach to a semantically-aware representation of items. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 567–577. Association for Computational Linguistics, Denver, Colorado (May–June 2015), http://www.aclweb.org/anthology/N15-1059
3. Diakopoulos, N., De Choudhury, M., Naaman, M.: Finding and assessing social media information sources in the context of journalis. In: Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems. pp. 24151–2460. CHI 2012, ACM, New York, NY, USA (2012)
4. Ferragina, P., Scaiella, U.: Tagme: On-the-fly annotation of short text fragments (by wikipedia entities). In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management. pp. 1625–1628. CIKM '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1871437.1871689
5. Ge, M., Delgado-Battenfeld, C., Jannach, D.: Beyond accuracy: Evaluating recommender systems by coverage and serendipity. In: Proceedings of the Fourth ACM Conference on Recommender Systems. pp. 257–260. RecSys '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1864708.1864761
6. Gloviczki, P.J.: Journalism in the Age of Social Media, pp. 1–23. Palgrave Macmillan US, New York (2015)
7. Hopcroft, J., Tarjan, R.: Efficient algorithms for graph manipulation. Commun. ACM 16(6), 372–378 (Jun 1973), http://doi.acm.org/10.1145/362248.362272
8. Jain, A.K.: Data clustering: 50 years beyond k-means. Pattern Recogn. Lett. 31(8), 651–666 (Jun 2010), http://dx.doi.org/10.1016/j.patrec.2009.09.011
9. Jenders, M., Lindhauer, T., Kasneci, G., Krestel, R., Naumann, F.: A Serendipity Model for News Recommendation, pp. 111–123. Springer International Publishing, Cham (2015), http://dx.doi.org/10.1007/978-3-319-24489-1_9

10. Kaminskas, M., Bridge, D.: Measuring surprise in recommender systems. In: Adamopoulos, P., Bellogn, A., Castells, P., Cremonesi, P., Steck, H. (eds.) Procs. of the Workshop on Recommender Systems Evaluation: Dimensions and Design (Workshop Programme of the Eighth ACM Conference on Recommender Systems) (2014)
11. Kille, B., Hopfgartner, F., Brodt, T., Heintz, T.: The plista dataset. In: Proceedings of the 2013 International News Recommender Systems Workshop and Challenge. pp. 16–23. NRS '13, ACM, New York, NY, USA (2013), http://doi.acm.org/10.1145/2516641.2516643
12. Knight, M.: 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 (2012)
13. König, A.C., Gamon, M., Wu, Q.: Click-through prediction for news queries. In: Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 347–354. SIGIR '09, ACM, New York, NY, USA (2009), http://doi.acm.org/10.1145/1571941.1572002
14. Kotkov, D., Wang, S., Veijalainen, J.: A survey of serendipity in recommender systems. Know.-Based Syst. 111(C), 180–192 (Nov 2016), http://dx.doi.org/10.1016/j.knosys.2016.08.014
15. Krestel, R., Werkmeister, T., Wiradarma, T.P., Kasneci, G.: Tweet-recommender: Finding relevant tweets for news articles. In: Proceedings of the 24th International Conference on World Wide Web. pp. 53–54. WWW '15 Companion, ACM, New York, NY, USA (2015)
16. Lehmann, J., Gonçalves, B., Ramasco, J.J., Cattuto, C.: Dynamical classes of collective attention in twitter pp. 251–260 (2012)
17. Leskovec, J., Backstrom, L., Kleinberg, J.: Meme-tracking and the dynamics of the news cycle. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 497–506. KDD '09, ACM, New York, NY, USA (2009)
18. Lin, J., Keogh, E., Lonardi, S., Chiu, B.: A symbolic representation of time series, with implications for streaming algorithms. In: Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. pp. 2–11. DMKD '03, ACM, New York, NY, USA (2003), http://doi.acm.org/10.1145/882082.882086
19. Lommatzsch, A., Albayrak, S.: Real-time recommendations for user-item streams. In: Proc. of the 30th Symposium On Applied Computing, SAC 2015. pp. 1039–1046. SAC '15, ACM, New York, NY, USA (2015)
20. Murakami, T., Mori, K., Orihara, R.: Metrics for evaluating the serendipity of recommendation lists. In: Proceedings of the 2007 Conference on New Frontiers in Artificial Intelligence. pp. 40–46. JSAI'07, Springer-Verlag, Berlin, Heidelberg (2008), http://dl.acm.org/citation.cfm?id=1788314.1788320
21. Napoli, P.M.: Measuring media impact an overview of the field. Rutgers University (2014)
22. Phelan, O., McCarthy, K., Smyth, B.: Using twitter to recommend real-time topical news. In: Proceedings of the Third ACM Conference on Recommender Systems. pp. 385–388. RecSys '09, ACM, New York, NY, USA (2009)
23. Reingold, O.: Undirected connectivity in log-space. J. ACM 55(4), 17:1–17:24 (Sep 2008), http://doi.acm.org/10.1145/1391289.1391291
24. Scaiella, U., Prestia, G., Del Tessandoro, E., Ver, M., Barbera, M., Parmesan, S.: Datatxt at microposts2014 challenge. vol. 1141, pp. 66–67 (2014)

25. Stilo, G., Velardi, P.: Efficient temporal mining of micro-blog texts and its application to event discovery. Data Min. Knowl. Discov. 30(2), 372–402 (Mar 2016)
26. Stilo, G., Velardi, P.: Hashtag sense clustering based on temporal similarity. Computational Linguistics pp. 1–32 (2017/01/17 2016), http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00277
27. Tsagkias, E., de Rijke, M., Weerkamp, W.: Predicting the volume of comments on online news stories. In: Proceedings of CIKM 09. ACM, New York, NY, USA (2009)
28. Tsagkias, M., de Rijke, M., Weerkamp, W.: Linking online news and social media. In: King, I., Nejdl, W., Li, H. (eds.) WSDM. pp. 565–574. ACM (2011)
29. Vargas, S., Castells, P.: Rank and relevance in novelty and diversity metrics for recommender systems. In: Proceedings of the Fifth ACM Conference on Recommender Systems. pp. 109–116. RecSys '11, ACM, New York, NY, USA (2011), http://doi.acm.org/10.1145/2043932.2043955
30. Zubiaga, A., Ji, H., Knight, K.: Curating and contextualizing twitter stories to assist with social newsgathering. In: 18th International Conference on Intelligent User Interfaces, IUI '13, Santa Monica, CA, USA, March 19-22, 2013. pp. 213–224 (2013)
31. Zuckerberg, M.: New facebook project to fight 'fake news' (Jan 2017), https://www.dawn.com/news/1307895/new-facebook-project-to-fight-fake-news