Real-time News Story Detection and Tracking with Hashtags

Gevorg Poghosyan and **Georgiana Ifrim**
Insight Centre for Data Analytics
University College Dublin
Dublin, Ireland
{gevorg.poghosyan, georgiana.ifrim}@insight-centre.org

Abstract

Topic Detection and Tracking (TDT) is an important research topic in data mining and information retrieval and has been explored for many years. Most of the studies have approached the problem from the event tracking point of view. We argue that the definition of stories as events is not reflecting the full picture. In this work we propose a story tracking method built on crowd-tagging in social media, where news articles are labeled with hashtags in real-time. The social tags act as rich meta-data for news articles, with the advantage that, if carefully employed, they can capture emerging concepts and address concept drift in a story. We present an approach for employing social tags for the purpose of story detection and tracking and show initial empirical results. We compare our method to classic keyword query retrieval and discuss an example of story tracking over time.

1 Introduction

We study the problem of automatically extracting and tracking[1] the storyline of news (i.e., the news articles covering the story events) for the purpose of improving the news presentation, both for consumption and research purposes (as targeted also in (Ahmed et al., 2011; Conrad and Bender, 2016; Leban et al., 2016)). Although this problem is widely addressed in the research literature from machine learning, data mining and information retrieval communities, the issue of efficiently and effectively mapping large volumes of news articles to story timelines in real-time, remains challenging.

A news story often discusses multiple related events, which happen in different time periods and may as well involve different entities (people, countries, organisations). Some stories are relatively short in time, such as the 2016 Champions League final, and some others span many years and discuss multiple events, such as the Ebola outbreak or the migrant crisis. The story of the Syrian war, for example, has evolved in time, shifting the discussion **topic** (*Middle East*, *migration*, *human rights*, *politics*), the discussed **entities** (*Assad*, *ISIS*, *Putin*, *USA*, *Islamic State*, *Turkey*, *Hungary*, *Belgium*) and **events** (*rebel uprising*, *destruction of Syria's chemical weapons*, *Yazidi massacres*, *camerawoman kicking a migrant*, *liberation of Palmyra*). Figure 1 illustrates this drift in the projected topic-event-entity combined dimensions over time, in the news article space. The figure also shows that stories may share articles. For example, the article "Turkey carries out air strikes" may appear in several stories: *Syrian war*, *PKK in Syria*, *Turkey elections 2015*.

We propose to model story tracking as a real-time information retrieval problem. We assume to have access to a collection of news articles annotated with social tags extracted in real-time from social media platforms such as Twitter. This approach takes advantage of crowdsourced content as a form of real-time, continuous tagging of news. Additionally, social tags (aka hashtags)[2] are not necessarily topical:

[1] Corresponding to Topic Detection and Topic Tracking research applications defined by TDT community at http://www.itl.nist.gov/iad/mig/tests/tdt/

[2] The terms *social tag* and *hashtag* are used interchangeably in the

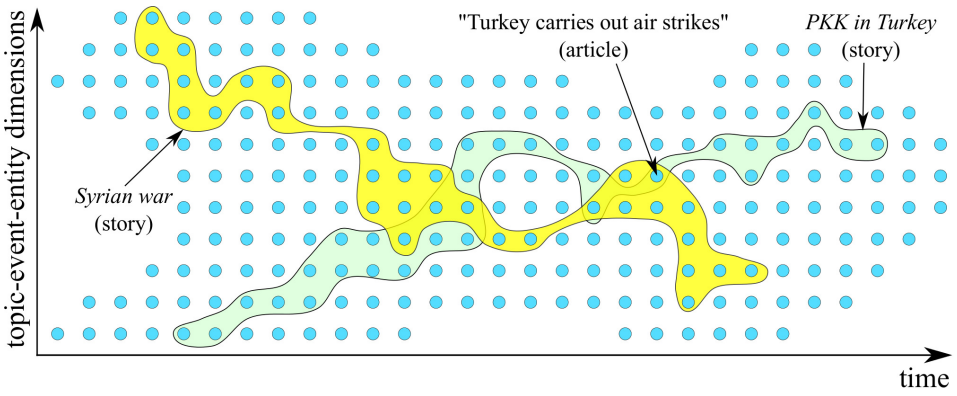

Figure 1: Stories' drift in topic-event-entity-time space.

they have the advantage of grouping together articles belonging to the same story (e.g., racial conflicts in US, #ericgarner, #blacklivesmatter, #icantbreathe) and allow the user to focus on diverse aspects of a story (e.g., Greek economic crisis, #grexit, #tsipras, #merkel, #ecb, #imf, #finland). We model a story as a query in an information retrieval framework where the query can mix keywords and hashtags.

From an application point of view, our work aims to automate the creation of story-focused pages and corresponding timelines, to enrich the storyline with context from social media (videos, tweets, posts, etc.) such as on `www.NewsDeeply.org` pages, and to provide story detection and tracking capabilities.

The choice of social tags is motivated by the following factors: (i) hashtags are inherently suitable for story tracking, as they are used on social platforms such as Twitter for tagging topics of interest[3], (ii) creation, popularity and abandonment of hashtags implicitly encode the concept drift in the story, (iii) hashtags allow cross-platform multi-modal content linking (text, image, video), (iv) tagging articles imposes no structural limitations for organising news as in single-linkage clustering (Ahmed et al., 2011; Conrad and Bender, 2016; Hou et al., 2015; Leban et al., 2016; Pouliquen et al., 2008). This approach is also consistent with trends in media: (a) social media oriented news providers like *AJ+* have embraced the usage of manually assigned hashtags to organise their content, (b) both *The Guardian*[4] and *The Huffington Post*[5] are giving importance to hashtags by writing articles about the popular social tags and informing the public on discussion trends, (c) *The Sun* had published a newspaper with a hashtag[6] alongside an article[7].

By including hashtags in the query we can facilitate better query formulation, and therefore a better story tracking process. For example, searching for *#rmucl* (the hashtag for Real Madrid's UEFA Champions League (UCL) season) can help to narrow the search down to Real Madrid's solely UCL games (i) even if the articles do not contain any of the keywords (e.g., an article titled *"Zidane's squad beats Manchester City 1:0 (video)"* may have a short body not containing any of the search keywords), (ii) avoiding the noise from the Club's activities in the Spanish League (with dedicated hashtag *#rmliga*). The method implicitly allows the choice of story granularity and navigation to substories. Contrary to the *#rmucl* example, some hashtags, e.g., *#rip*, can also group articles about unrelated events and entities in a non-topical fashion.

The remainder of this paper is structured as follows. Section 2 describes the related work. In Section 3 we briefly introduce the notation used in the paper and the problem setup. Section 4 describes our approach for story detection and organizing the news

paper.

[3] http://www.nytimes.com/2011/06/12/fashion/hashtags-a-new-way-for-tweets-cultural-studies.html

[4] www.theguardian.com/technology/hashtags

[5] www.huffingtonpost.com/news/hashtags/

[6] www.huffingtonpost.com/2014/03/26/sun-hashtag-newspaper-murdoch-british_n_5034639.html

[7] "By printing hashtags alongside our news we are making it easy for readers to share their opinions and continue the story online," Sun editor David Dinsmore said in a statement

articles into stories. Section 5 explains our proposed method for story tracking. In Section 6 we present the evaluation of the proposed method, and, in Section 7 we conclude the paper and discuss future research.

2 Related Work

In this section we present related research spanning different research communities.

Event detection and tracking: Work presented by (Allan et al., 1998) tracks 25 predefined events, processing the articles in chronological order and making a binary decision on event relatedness for each article, before processing any subsequent articles. An assumption is made that each article discusses a single event. (Brown, 2001) provides an extension to (Allan et al., 1998) for real-time event detection. (Kuzey and Weikum, 2014) describe an offline system for populating event classes of knowledge bases by first mapping news to *Wikipedia* categories, then mapping the latter to WordNet event classes. (Leban et al., 2016) have designed a real-time system which groups articles about an event across languages. Articles are clustered based on their cosine similarity. An event is registered after a cluster reaches a certain size, whereas the cluster will be removed when it becomes older than 5 days.

Tracking stories: (Navrat et al., 2009) propose a system with a focused crawler which works on the user side and tracks a story by smartly selecting the links on the article page. The performance on the experimental set was noisy and the system purely relies on the existing links on the page. In the *European Media Monitor* from (Pouliquen et al., 2008) news articles are clustered using an agglomerative clustering algorithm. Stories are formed from clusters linked based on their cosine similarity. The system introduced in (Hou et al., 2015) represents each news article in the dimensions of entities, topics and events. A knowledge base is used for linking topics and linking entities and thus creating links between the articles. For a given query, a list of articles is returned ranked by the weighted sum of relevance and topic scores. (Ahmed et al., 2011) model news storyline clustering by applying a topic model to the clusters, while simultaneously generating single-linkage clusters using the Recurrent Chinese Restaurant Process. This approach allows the number of stories to be determined by the data. The system accuracy is evaluated on 2,525 manually judged "must-link" and "cannot-link" article pairs. (Conrad and Bender, 2016) have designed an event-centric hierarchical agglomerative clustering algorithm operating in real-time for providing a news browsing experience in a structured way, given the editorially supplied top-level story labels. *MediaMeter* introduced in (Nomoto, 2015) uses a tagger called *WikiLabel* for assigning Wikipedia labels to news articles and detects trending topics based on labels with high burstiness scores. (Vossen et al., 2015) discuss a framework for structuring massive news streams into storylines. The authors discuss a computational model of storylines and guidelines for storyline evaluation, but no comprehensive empirical study is presented.

Query expansion: (Verberne et al., 2016) studied query term suggestions for Boolean queries in a news monitoring system. They found that the premise of 'pseudo-relevance' does not hold for Boolean retrieval when the set of retrieved documents is noisy. (Anagnostopoulos et al., 2012) introduced a query expansion algorithm based on a semantic network of Twitter hashtags. They have shown that the social intelligence can be used to describe information and successfully applied it in query expansion.

Contribution: State-of-the-art systems rely on keyword/semantic matching and require often slow-to-change offline snapshots of knowledge bases (Kuzey and Weikum, 2014) or need computationally expensive, complex clustering or semantic models, where parameters such as the number of topics (Hou et al., 2015), timespan of stories (Conrad and Bender, 2016; Leban et al., 2016) and cluster sizes (Pouliquen et al., 2008), significantly affect the system performance.

Unlike the methods described above, we model storyline extraction as a pattern mining and real-time retrieval problem based on social annotations of news articles. The proposed approach has the following advantages which are important for our problem: (i) it is non-parametric over stories, allowing any size, duration, number of events, number of named entities, etc., (ii) stories are not bounded to predefined topics or taxonomies and the choice

of query hashtags allows "zooming in" to substories, (iii) articles can be shared between storylines, so articles relevant to multiple stories can appear in each, not penalizing the recall of either story, (iv) no reliance on information from external knowledge bases, which may lag behind the relevant events, (v) the story will track the emerging as well as deprecating concepts (in the form of hashtags or keywords) relevant to the story, (vi) real-time performance is achieved without limiting the story size in articles or span in time, and without a need for recomputing any clusters or semantic models when new data arrives.

3 Preliminaries and Basic Notation

We assume to have a dataset of articles with recommended hashtags. The social tags can be manually assigned to an article by a journalist or by an automated hashtag recommender. *Hashtagger* presented in (Shi et al., 2016) and the method proposed in (Efron, 2010) recommend hashtags to news articles. We build on top of Hashtagger, which recommends up to 10 hashtags to an article, which are updated every 15 minutes over a period of 24 hours from the article publication time.

The notations used in this paper are summarized in Table 1. An article $Article_j$ may get up to $10 \times 24\,hours \times 4\,per\,hour = 960$ unique hashtags denoted as $\#tag_1^j$ ... $\#tag_{960}^j$, each recommended with a confidence[8] $conf_t^j \in [0.5,\ 1]$, where $0 < t \leqslant 960$. A single hashtag can be recommended to the same article with different confidences at different points in time. A query for story retrieval and tracking is composed of keywords $w_1, ..., w_n$ and hashtags $\#tag_1^q, ..., \#tag_m^q$, where $n + m > 0$. Each query also includes a time period from which the articles will be retrieved. We denote articles retrieved by query expansion by the superscript ^{ex}. The retrieval score of $Article_j^{ex}$ is denoted as $score_j$.

We use the terms *substory* and *superstory* to refer to stories correspondingly narrower or wider in scope, than the reference story.

[8]Manually assigned hashtags can get confidence set to 1 if no value is given.

n	number of terms in query
m	number of hashtags in query
q	query, where $n + m > 0$
p	number of full months in the query time period
w_i	i^{th} keyword in the query q, where $0 < i \leqslant n$
$\#tag_i^q$	i^{th} hashtag in the query q, where $0 < i \leqslant m$
$Article_j$	the j^{th} article in the database
$\#tag_t^j$	t^{th} hashtag recommended to $Article_j$, where $0 < t \leqslant 960$
$conf_t^j$	recommendation confidence of $\#tag_t^j$, where $0 < t \leqslant 960$
$Article_j^{ex}$	the j^{th} article retrieved for query expansion, where $0 < j \leqslant 10+p$
$score_j$	retrieval score of the j^{th} article retrieved for query expansion
b	number of hashtag confidence bins
k	number of highest score confidence bins, where $k \leqslant b$
$\#tag_i$	i^{th} hashtag in the query expansion hashtag set
$score_{\#tag_i}$	score assigned to $\#tag_i$
l	number of story expansion hashtags, where $l \leqslant 10+p$
M	number of articles retrieved with the expanded query

Table 1: Notation used in the paper.

4 Story Detection via Frequent Hashtag Set Mining

We propose a method that maps news articles to stories in real-time, by grouping the articles with connected events, entities and topics that are discussed together on social platforms like Twitter. We use Twitter hashtags to group the news into stories. Each hashtag represents a story (e.g., *#turkey* or *#syria*) and a combination of hashtags represents a substory for each of the stories it is part of, e.g., {*#turkey*, *#syria*} represents the story of Turkish involvement in Syrian war.

4.1 Story Detection

This section describes a method for story detection using frequent pattern mining over hashtags.

A single news article can be covering multiple stories (Vossen et al., 2015) and some other articles can cover either a substory or a superstory of a story covered in the first article. We have observed that multiple articles covering the same story get assigned the same set of hashtags. We exploit this phenomenon to detect popular news stories by mining frequent hashtag sets which are being assigned to the same set of articles. Each frequent hashtag set (e.g., {*#turkey*, *#syria*, *#kurdistan*}), which is a superset of another hashtag set (e.g., {*#turkey*, *#syria*}), is a popular story too and is a substory of the story represented by the hashtag superset. This representation enables us to use the hierarchical structure of the story coverage for better navigation in the huge sea of stories.

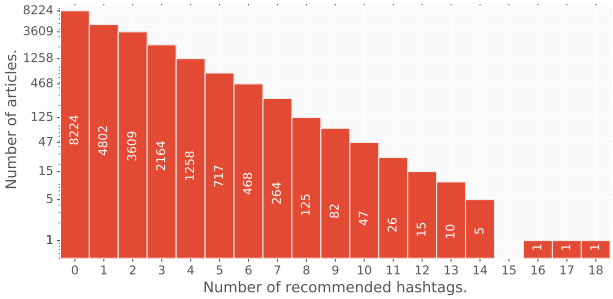

Figure 2: Distribution of number of recommended hashtags per article in the full set of all 21,819 articles from May 2016. 13,595 articles have at least one recommended hashtag with an average of 2.55 hashtags per article. 13,270 articles have at least one non-spammy recommended hashtag.

To study the possible advantages of the chosen representation, we run an experiment on a subset of 13,270 articles[9] from May 2016 (which represent 60.8% of all articles in this period) that have been linked to at least one hashtag. Overall 5,107 unique hashtags were recommended to the articles in May 2016. The histogram of number of recommended hashtags per article is shown in Figure 2. We have defined the articles as a support domain and have extracted frequent hashtag patterns co-occurring for a large number of articles.

We use an implementation of ECLAT[10] (Zaki, 2000) for mining the frequent hashtag sets. Running the ECLAT algorithm with a minimum support requirement[11] of 5 articles resulted in 6,839 frequent hashtag patterns of a form shown in Table 2. For example, the third line in Table 2 shows that there are 35 articles which got both {*#farewellboleyn, #whufc*} hashtags recommended to them.

The extracted patterns give an overview of all the topics covered in the news article set. In the following section we discuss how a user can navigate the big set of hashtag patterns which define detected stories.

[9]We filter out the recommendations of spammy hashtags which don't contribute to any certain story: #news, #business, #breaking, #politics, #jobs, #world, #rt, #sport, #breakingnews #follow.

[10]http://www.borgelt.net/eclat.html

[11]The support threshold can be varied to change the extracted set of patterns.

Pattern	Support
$\#mufc$	758
$\#trumptrain$ $\#makeamericagreatagain$ $\#trump$ $\#trump2016$	59
$\#farewellboleyn$ $\#whufc$	35

Table 2: Example frequent hashtag patterns mined from news articles in May 2016.

4.2 Hierarchical Story Representation

Visualization of mined stories can be done in many ways allowing navigation through the stories. The essential factors that have influenced our choice of visualization are: (i) the hierarchical structure of substory-superstory relationship must be shown, (ii) the user must be able to zoom in to substories, (iii) a substory can be navigated to from either of its superstories[12].

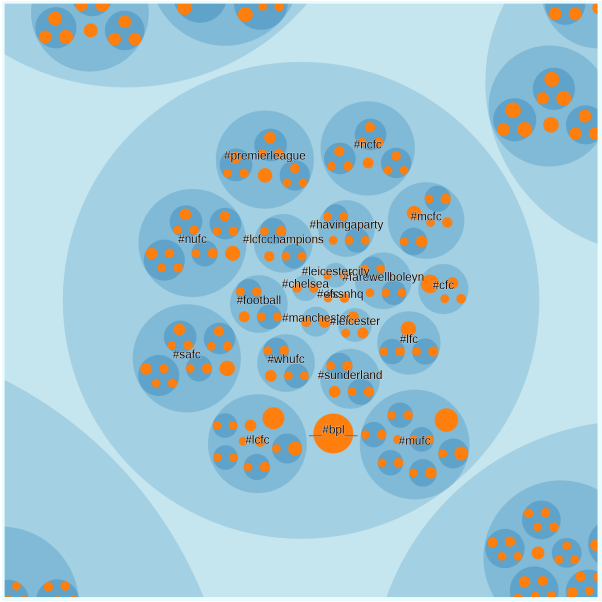

Figure 3: Screenshot of Barclays Premier League, aka #bpl story visualized.

We have used Zoomable Circle Packing[13] interactive visualization from D3 library[14] to visualize the stories. Figure 3 shows the *#bpl* story of Barclays Premier League of soccer. The inner blue bubbles represent the substories of *#bpl* and each includes a specific football club hashtag (e.g., *#mufc* for

[12]This results in duplication of substories under each existing superstory.

[13]http://bl.ocks.org/mbostock/7607535

[14]http://github.com/d3/d3/wiki/Gallery

Manchester United football club). In this form of visualization, the {*#bpl, #mufc*} substory can be navigated from either *#bpl* or *#mufc* story. This choice results in duplicated data but keeps the strictly hierarchical structure of news stories.

The interactive visualization of May 2016 stories and the corresponding data are available online[15].

In contrast to related work, this pattern set structure allows us to represent the news in a hierarchical, multiple-linkage browsable structure, where one can "zoom in" into a multi-hashtag substory while, at the same time, allowing a hashtag (and articles linked to it) to be a part of another story. The frequent sets can be maintained and updated upon the fresh data arrival or alternatively mined again in a periodic fashion (e.g., once an hour).

5 Story Tracking via Retrieval with Social Tags

In the previous section we discussed how to detect stories from article-hashtag sets. We now formulate *story tracking* as a retrieval task with queries that allow mixing of keywords and hashtags. This allows tracking stories on-the-fly rather than being restricted to a pre-determined set of stories. We represent an article by its headline, subheadline, body, a set of summary keywords and a set of hashtags recommended to the article. The recommended hashtags are binned into $b = 20$ confidence bins with ranges from $(0.975, 1.0]$, down to $(0.5, 0.525]$ and indexed to enable an efficient search on article fields with different weighting using the BM25 algorithm (Robertson et al., 1994). The retrieval is done with the following settings:

- Keywords $w_1, ..., w_n$ are matched on article keywords, headline, subheadline and content with score boost of correspondingly $\times 4$, $\times 3$, $\times 2$ and $\times 1$. The idea behind this weighting is that an article matching a query in its headline, is more likely to belong to the requested story, rather than in the case when the matching keywords appear somewhere in the article body.

- Hashtags $\#tag_1^q, ..., \#tag_m^q$ are matched on the top-k hashtag confidence bins with score boosting of $6 - \frac{(i+1)\times 2}{b}$ for a match on bin $0 < i \leqslant k$. The idea behind the decaying per bin boost is that more confidently recommended hashtags are more likely to be relevant to the story.

Figure 4 gives an overview of the proposed method. To retrieve the articles covering a certain story, we do a two-step retrieval in the given time period by expanding the original query in the "hashtag space", then retrieving the story articles with the expanded query. The step-by-step process is the following:

1. To get a set of potentially relevant to the story hashtags, we use the recommended hashtags of the top-$(10+p)$ articles (shown as $Article_j^{ex}$ on Figure 4) from the initial search results, where p is the number of full months in the queried time period. The intuition behind the parameter p is that longer stories would possibly include more events, entities and topics and presumably these may be covered in more articles. On the other hand, using too many articles for the query expansion is bound to introduce more noise. Among the top-$(10 + p)$ articles, we only use for query expansion those whose $score_j \geqslant 0.5 \times score_1$, i.e., the matching score is not lower than 50% of the top match article score. This approach allows to narrow down to the more focused story when an overlap of coverage exists between the query terms.

2. The query expansion hashtags are chosen from the set of hashtags recommended to any of the $10+p$ articles. Because the same hashtag may be recommended to several of chosen $10+p$ articles, we first weight hashtags by the product of their recommendation confidence to an article and the article matching score on the query, and then we take the highest of these scores for a given hashtag. Hashtags of the filtered set of articles are weighted by the following formula:

$$score_{\#tag_i} = max_{1<j\leqslant 10+p,\ \#tag_i=\#tag_r^j}(score_j \times conf_r^j)$$

where $score_j$ is the matching score of the j^{th} article from $(10 + p)$ retrieved articles, tag_r^j is the r^{th} recommended hashtag with confidence $conf_r^j$ for the article $Article_j^{ex}$. The resulting set is limited to $(10 + p)$ unique hashtags with the highest scores $score_{\#tag_i}$, where $0 < i \leqslant (10+p)$, that will serve as a query expansion set of hashtags, denoted as $\#tag_1, ..., \#tag_l$ on

[15] http://github.com/gevra/may2016-stories

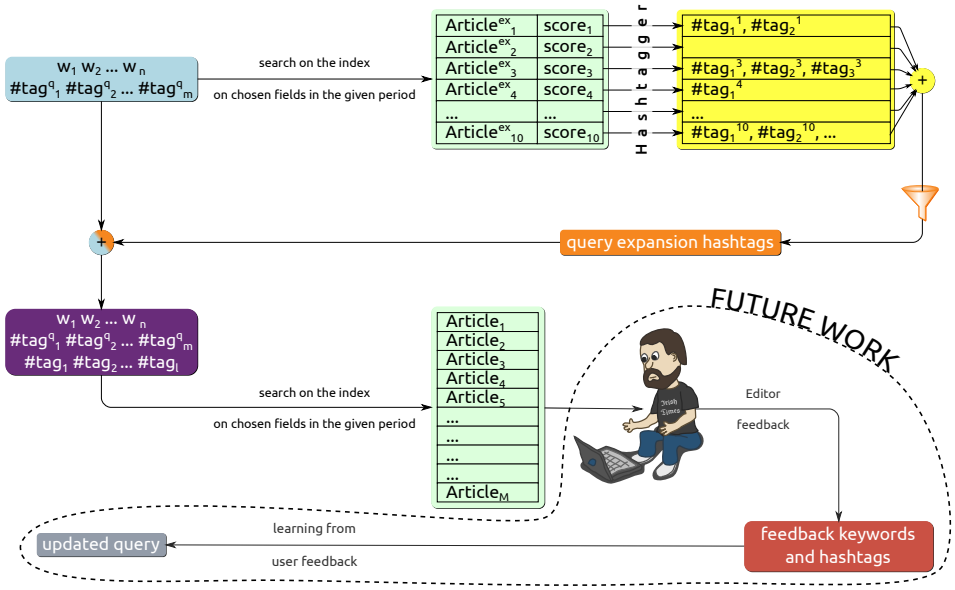

Figure 4: Story retrieval process diagram.

Figure 4. The initial query together with the query expansion hashtags form what we call the story tracking query.

3. A second retrieval using the expanded query, returns the final set of ranked articles $Article_1, ..., Article_M$. The method works even for cases where there are no hashtags recommended to an article, but the presence of hashtags allows for more refined queries.

The hyper-parameters of the system, like the query boosting parameters, b, k, as well as the heuristic methods like the choice of $10+p$ articles are yet not fully evaluated, and the used values were chosen empirically.

The stories can be tracked by simply re-initiating the whole process described in Figure 4 at anytime. A fresh retrieval, i.e., re-issuing the same query, allows to (i) capture the previously unseen events and entities covered in newly arrived articles, (ii) capture emerging relevant story hashtags in the query expansion, (iii) retrieve currently relevant articles, which might not be matching the previously used query. Hashtags once present in the query expansion, will remain in the story defining query to ensure the completeness of story coverage.

6 Evaluation

We have been tracking 27 RSS news feeds from 8 (mostly Irish[16]) news organisations, starting from August 2015. This allows us to track stories that have started capturing the public attention almost a year ago. A side effect of our news selection is that the social content, and therefore also the hashtags that are linked to articles, are biased towards Ireland-related issues.

For a preliminary evaluation of our method we compare the performance of article retrieval with the expanded query to a retrieval with the initial input query in identical setup. To show the effect of query expansion hashtags, we also include in the evaluation a retrieval with expansion hashtags only. The evaluation is performed on the following 5 queries: *migrant crisis*, *refugee crisis*, *US election*, *EURO 2016* and *#euro2016*. Many news providers offer story-focused pages with curated collections of news articles. We select news articles from The Irish Times story-pages as a ground truth for each story[17].

[16] *The Irish Times*, *The Irish Independent*, *RTÉ*, *TheJournal.ie*, *Irish Examiner*, *BBC*, *Reuters* and *Al Jazeera*

[17] The URLs of the curated story pages corresponding to our selected 5 queries are:
http://www.irishtimes.com/news/world/europe/migrant-crisis
http://www.irishtimes.com/news/world/us/us-election
http://www.irishtimes.com/sport/euro-2016

The curated story-pages serve as a natural ground truth for evaluating our story extraction, because our method aims to automate the creation of this kind of pages. There are only few active curated story-pages on news platforms, which limits our ground truth collection and is the main reason behind our choice of queries. For the purpose of evaluation, we correspondingly limit the retrieval from the database to The Irish Times articles from the time period covered by the curated story page. Our goal is not necessarily to evaluate the retrieval quality, but to show that the social tags can improve the story extraction. Table 3 presents the evaluation metrics including the Recall and NDCG@k as defined in (Manning et al., 2008). NDCG@k reflects the quality of ranking, and in our case shows how similar our stories are to those on curated story pages.

Query	Query expansion hashtags	Time period	Number of articles on the curated page and in our database	Match	Articles	Recall	NDCG @10	NDCG @25
migrant crisis	#lybia #pope #popefrancis #eu #health #migrants #greece #turkey #utah #lesbos #italy #francis	07 Apr - 14 Jun	45	initial query	782	88.9%	0.2083	0.2549
				expansion hashtags only	294	46.7%	0.2903	0.2449
				expanded query	984	91.1%	0.3398	0.3565
refugee crisis	#refugees #bono #eu #turkey #un #refugee #china	07 Apr - 14 Jun	45	initial query	783	84.4%	0.6204	0.3896
				expansion hashtags only	244	28.9%	0.2201	0.1741
				expanded query	964	88.9%	0.5321	0.3819
US election	#modiinus #hillaryclinton #deleteyouraccount #peru #freedom #primaryday #eu #imwithher #modifiedforeignpolicy #kuczynski #hillary	06 Jun - 15 Jun	19	initial query	516	94.7%	0.2248	0.2424
				expansion hashtags only	31	36.8%	0.3909	0.2971
				expanded query	524	94.7%	0.4748	0.3420
EURO 2016	#euro2016	30 May - 15 Jun	12	initial query	572	83.3%	0	0
				expansion hashtags only	155	41.7%	0	0.0607
				expanded query	604	100%	0	0.0293
#euro2016	#romania #england #wal #eng #russia #marseille	30 May - 15 Jun	12	initial query	155	41.7%	0	0.0607
				expansion hashtags only	165	41.7%	0	0.0252
				expanded query	165	41.7%	0	0.0252

Table 3: Evaluation results.

Table 3 shows that the query expansion does not contribute dramatically to Recall. However for *migrant crisis* and *US election* stories, the improvement in NDCG is significant. The *refugee crisis* story is a good example for demonstrating the sensitivity of the query expansion as the evaluation for it was done on the same ground truth curated page as for the *migrant crisis*. Also one can notice the presence of #bono[18] in the query expansion set. Bono has several times visited refugee camps and spoken for the rights of migrants. Nevertheless the importance and relatedness of #bono to the story is slightly overshadowed by a significant coverage about Bono in our news corpus of mostly Irish sources. Querying *EURO 2016*, although successfully expands the query with the relevant hashtag, retrieves noisy results. The performance figures can be explained by the high ambiguity of the query, as *EURO* may refer to the currency or politics. Regardless of the unimpressive performance metrics for the latter query, the takeaway point is that the ambiguity problem can be solved by issuing a query *#euro2016* instead, and this is one of the key features of our method, offered as a solution to the ambiguity problem.

The live system used in the experiments for evaluation is described in our previous work (Poghosyan et al., 2016) and is available online[19].

Story Tracking and Concept Drift: In Table 4 we show the query expansion hashtags for the *migrant crisis* story for each two weeks from January 1st to May 31st 2016, to illustrate the potential of hashtags to track a story with the proposed method. It can be noticed that the query expansion has successfully captured the newly emerged entities in the story. Methods relying on offline knowledge bases may not be responsive enough to capture the new relationships of entities in stories.

Period (2016)	Query expansion hashtags for the query *migrant crisis*
Jan 1 - Jan 15	#migrantcrisis #crisis #migrant
Jan 1 - Jan 31	#migrant #germany #calais #corbyn
Jan 1 - Feb 15	#calais #corbyn #germany #migrant
Jan 1 - Feb 29	#calais #greece #migrant #corbyn #germany #refugees
Jan 1 - Mar 15	#calais #migrants #macedonia #greece #turkey #migrant #corbyn #germany #refugees #eu
Jan 1 - Mar 31	#calais #migrants #macedonia #greece #turkey #migrant #corbyn #germany #refugees #eu
Jan 1 - Apr 15	#calais #migrants #macedonia #greece #turkey #migrant #corbyn #germany #refugees #eu
Jan 1 - Apr 30	#calais #migrants #macedonia #greece #turkey #migrant #corbyn #germany #refugees #lesbos #eu
Jan 1 - May 15	#calais #migrants #macedonia #greece #turkey #migrant #corbyn #germany #refugees #lesbos #eu
Jan 1 - May 31	#calais #migrants #macedonia #greece #turkey #migrant #corbyn #germany #refugees #lesbos #eu

Table 4: Example of the expanded query evolution in time for the query *migrant crisis*.

[18] Bono is an Irish musician best known as the lead vocalist of rock band U2

[19] http://lovelace.ucd.ie/tutorial_video

7 Conclusions and Future Work

In summary, we propose a new angle on story detection and tracking based on frequent pattern mining and real-time retrieval of tagged news articles. To the best of our knowledge there is no other method which exploits real-time hashtag recommendations for this purpose. We present a frequent pattern-based story detection which allows "zooming in/out" into substories and superstories. The advantage of our proposed story tracking solution is that it quickly adapts to emerging entities or events and their relatedness, because it does not require a slow-to-change knowledge base. Our solution is real-time and does a retrieval on-demand without the need of recomputing any clusters or semantic models when new data arrives. The weaknesses of our story tracking approach include the strong reliance on the hashtag recommender (although Hashtagger has 85% Precision@1) and the potential lack of story discussions on social platforms, e.g., Hashtagger recommends at least 1 hashtag to about 65% of all articles. This can be mitigated to some extent possibly by expanding our scope to other social platforms that increasingly adopt social tags. Yet another workaround for compensating for the partial hashtag coverage is discussed below in the future work.

Future Work. We intend to have a deeper evaluation of the story detection and tracking by expanding the experiments to multiple news sources and a larger set of stories. We also consider necessary to perform an evaluation involving manual annotation of the retrieved articles by a domain expert. The heuristic elements of the method have an intuition behind and are set only empirically. These elements require an evaluation of their contribution.

We believe a more accurate query expansion with weighted hashtags will allow to distinguish the stories which have the same set of linked hashtags, but different dominant hashtags. Weighted queries will also enable the automation of query updates for story tracking and incorporation of human feedback for refining the query over time.

The query formulation largely affects both the query expansion and subsequently also the final article retrieval (as we have shown in Section 6). The task of composing good queries is not trivial and an exploration of a substory may not be achieved only by modifying the query. For this reason a user feedback loop may be added to allow the query issuer to steer the story in the desired direction (see Figure 4).

The issue of diversity of news articles within a retrieved (and possibly curated) story is also interesting. We plan on studying the literature on aspect-based information retrieval (Santos et al., 2010), where the hashtags would serve as natural aspects in our framework.

To compensate for the partial hashtag coverage of articles (60.8% for May 2016), the keywords extracted from the articles, along with the assigned hashtags, can be included in frequent pattern mining for story detection. This may significantly change the mined stories, as the tag space density and subsequently the mined patterns' cardinality may change.

Along with simple market basket type analysis to discover frequent subsets of hashtags linked to sets of articles, we plan to extend the storyline organization with including new dimensions like the source and the time. n-ary frequent pattern mining techniques like the one described in (Cerf, 2010) can extract patterns of form $\{source_1, ..., source_i \times month_1, ..., month_j \times \#tag_1, ..., \#tag_k\}$ which will help to analyze the temporal-topical patterns of sources and how these patterns are similar or different between the sources. We plan to explore the frequent patterns of hashtags linked to the articles of a source, to possibly discover the response of the audience using a similar vocabulary to the one of the source. We are also interested in news coverage comparison between the sources for different given stories with patterns of form $\{source_1, ..., source_i \times \#tag_1, ..., \#tag_k\}$.

Finally we plan to build on the story tracking method to automate story timeline and summary generation, similar to the ones found on www.NewsDeeply.org or The Irish Times curated pages.

Acknowledgments

This work was funded by Science Foundation Ireland (SFI) under grant number 12/RC/2289.

References

Amr Ahmed, Qirong Ho, Choon H Teo, Jacob Eisenstein, Eric P Xing, and Alex J Smola. 2011. Online inference for the infinite topic-cluster model: Storylines from streaming text. In *International Conference on Artificial Intelligence and Statistics*, pages 101–109.

James Allan, Jaime G Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. 1998. Topic detection and tracking pilot study final report.

Ioannis Anagnostopoulos, Vassilis Kolias, and Phivos Mylonas. 2012. Socio-semantic query expansion using twitter hashtags. In *SMAP*.

Ralf D Brown. 2001. A server for real-time event tracking in news. In *Proceedings of the first international conference on Human language technology research*, pages 1–3.

Loïc Cerf. 2010. *Constraint-based mining of closed patterns in noisy n-ary relations*. Ph.D. thesis, INSA de Lyon.

Jack G. Conrad and Michael Bender. 2016. Semi-supervised events clustering in news retrieval. In *ECIR*.

Miles Efron. 2010. Hashtag retrieval in a microblogging environment. In *SIGIR*.

Lei Hou, Juanzi Li, Zhichun Wang, Jie Tang, Peng Zhang, Ruibing Yang, and Qian Zheng. 2015. Newsminer: Multifaceted news analysis for event search. *Knowledge-Based Systems*, 76:17–29.

Erdal Kuzey and Gerhard Weikum. 2014. Evin: Building a knowledge base of events. In *WWW*, pages 103–106.

Gregor Leban, Blaz Fortuna, and Marko Grobelnik. 2016. Using news articles for real-time cross-lingual event detection and filtering. In *ECIR*.

Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al. 2008. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge.

Pavol Navrat, Lucia Jastrzembska, and Tomas Jelinek. 2009. Bee hive at work: Story tracking case study. In *WI-IAT'09. IEEE/WIC/ACM*, volume 3. IET.

Tadashi Nomoto. 2015. Mediameter: A global monitor for online news coverage. *ACL-IJCNLP 2015*, page 30.

Gevorg Poghosyan, M. Atif Qureshi, and Georgiana Ifrim. 2016. Topy: Real-time story tracking via social tags. In *ECMLPKDD*.

Bruno Pouliquen, Ralf Steinberger, and Olivier Deguernel. 2008. Story tracking: linking similar news over time and across languages. In *MMIES*.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at trec-3. In *TREC*.

Rodrygo LT Santos, Jie Peng, Craig Macdonald, and Iadh Ounis. 2010. Explicit search result diversification through sub-queries. In *ECIR*.

Bichen Shi, Georgiana Ifrim, and Neil Hurley. 2016. Learning-to-rank for real-time high-precision hashtag recommendation for streaming news. In *WWW*.

Suzan Verberne, Thymen Wabeke, and Rianne Kaptein. 2016. Boolean queries for news monitoring: Suggesting new query terms to expert users. In *ECIR*.

Piek Vossen, Tommaso Caselli, and Yiota Kontzopoulou. 2015. Storylines for structuring massive streams of news. *ACL-IJCNLP 2015*, page 40.

Mohammed Javeed Zaki. 2000. Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering*, 12(3):372–390.