Preface

These proceedings contain the papers of the First International Workshop on Recent Trends in News Information Retrieval (NewsIR'16) held in conjunction with the ECIR 2016 conference in Padua, Italy, on the 20th of March 2016. Nine full papers and three short papers were selected by the programme committee from a total of 19 submissions. Each submitted paper was reviewed by at least three members of an international programme committee. In addition to the selected papers, the workshop features two keynote speeches. Keynote speeches are given by Jochen Leidner "Recent Advances in Information Access at Thomson Reuters R&D: News and Beyond", and Julio Gonzalo "Monitoring Reputation in the Wild Online West". We would like to thank ECIR for hosting us. Thanks also go to the keynote speakers, the program committee, the paper authors, and the participants, for without these people there would be no workshop.

Miguel Martinez, Signal Media Ltd.
Udo Kruschwitz, University of Essex
Gabriella Kazai, Lumi
Frank Hopfgartner, University of Glasgow
David Corney, Signal Media Ltd.
Ricardo Campos, Polytechnic Institute of Tomar / LIAAD-INESC TEC
Dyaa Albakour, Signal Media Ltd.

Programme Committee

Ramkumar Aiyengar, Bloomberg, UK
Omar Alonso, Microsoft, USA
Alejandro Bellogin Kouki, UAM, Spain
Marco Bonzanini, Bonzanini Consulting Ltd
Horatiu-Sorin Bota, University of Glasgow, UK
Igor Brigadir, Insight Centre for Data Analytics, Ireland
Toine Bogers, Aalborg University Copenhagen (AAU-CPH), Denmark
Ivan Cantador, UAM, Spain
Arjen De Vries, Centrum Wiskunde & Informatica (CWI), Netherlands
Ernesto Diaz Aviles, IBM Research, Ireland
Angel Castellanos Gonzalez , UNED, Spain
Julio Gonzalo, UNED, Spain
David Graus, University of Amsterdam, Netherlands
Jon Atle Gulla, NTNU, Norway
Charlie Hull, Flax, UK
Alipio Jorge, University of Porto / LIAAD-INESC TEC, Portugal
Jussi Karlgren, Gavagai, Sweden
Marijn Koolen, University of Amsterdam, Netherlands
David D. Lewis, David D. Lewis Consulting, USA
Stefano Mizzaro, University of Udine, Italy
Elaheh Momeni, University of Vienna, Austria
Miles Osborne, Bloomberg, UK
Filipa Peleja, Yahoo! Research, Spain
Vassilis Plachouras, on Reuters, UK
Barbara Poblete, University of Chile, Chile
Muhammad Atif Qureshi, National University of Ireland, Ireland
Paolo Rosso, Universidad Politecnica de Valencia, Spain
Alan Said, Recorded Future, Sweden
Damiano Spina, RMIT, Australia
Jeroen Vuurens, TU Delft, Netherlands
Colin Wilkie, University of Glasgow, UK
Arjumand Younus, National University of Ireland, Ireland
Arkaitz Zubiaga, University of Warwick, UK

Recent Advances in Information Access at Thomson Reuters R&D: News and Beyond

Jochen L. Leidner
Director of Research
Corporate Research & Development
Thomson Reuters
London, UK

Abstract

In this talk, I report on some recent advances of the Corporate R&D group at Thomson Reuters. Thomson Reuters is divided into the business areas News, Legal, Financial & Risk, Tax & Accounting, IP & Science. In the realm of news, the news recommender system NewsPlus and the real-time Twitter rumor detection tool for journalists, REUTERS Tracer, are discussed. From the area of pharma within IP & Science, I review work on adverse events associated with medical drugs, as mined from Twitter and used for drug repositioning. From the area of law, I report on the advanced search engine technology that powers the Westlaw search engine. From the area of Financial & Risk, I present risk mining, a technique for computer-supported risk identification framed as a relation classification task. Last but not least I conclude with some observed challenges and lessons learned. I conclude with a series of challenges and needs for the news industry.

Biography

Dr. Jochen Leidner is currently Director of Research at Thomson Reuters, where he heads the London (UK) R&D site, which he established. He has worked in many areas including information extraction from legal, news and financial documents, search engine technology and its application to legal information retrieval, automated proofing support for contracts, sentiment analysis, rule based systems, citation analysis and social media.

In: M. Martinez, U. Kruschwitz, G. Kazai, D. Corney, F. Hopfgartner, R. Campos and D. Albakour (eds.): Proceedings of the NewsIR'16 Workshop, Padua, Italy, 20-March-2016, published at http://ceur-ws.org

Monitoring Reputation in the Wild Online West

Julio Gonzalo
UNED
Madrid, Spain

Abstract

Monitoring Online Reputation has already become a key part of Public Relations for organizations and individuals; and current search technologies do not suffice to help reputation experts to cope with the vast stream of online content flooding reputation management experts.

In the talk we will summarize some of the main challenges that Information Access Technologies must face to assist online reputation monitoring tasks, and present some of the results obtained by the UNED research group in the areas of entity name disambiguation, topic tracking for reputation analysis, identification of opinion makers, and reputation-oriented summarization. We will make a special emphasis on the Replab test collections for Online Reputation Monitoring, which provide over half a million manual annotations provided by reputation experts on Twitter data.

Biography

Julio Gonzalo (UNED, Madrid, Spain) is head of the UNED research group in Natural Language Processing and IR (nlp.uned.es). He has recently been co-organizer of the RepLab Evaluation Campaign for Online Reputation Management Systems, co-organizer of the WePS evaluation campaign for Web People Search systems, and co-recipient of a Google Faculty Research Award. His research interests include Entity-Oriented and Semantic Search, Evaluation Methodologies and Metrics in Information Access, and Information Access Technologies for Social Media. A list of his publications can be found at Google Scholar: `https://scholar.google.com/citations?user=opFCmpYAAAAJ`.

In: M. Martinez, U. Kruschwitz, G. Kazai, D. Corney, F. Hopfgartner, R. Campos and D. Albakour (eds.): Proceedings of the NewsIR'16 Workshop, Padua, Italy, 20-March-2016, published at http://ceur-ws.org

Boolean Queries for News Monitoring: Suggesting new query terms to expert users

Suzan Verberne
Radboud University
Nijmegen, the Netherlands
s.verberne@cs.ru.nl

Thymen Wabeke
TNO
The Hague, the Netherlands
thymen.wabeke@tno.nl

Rianne Kaptein
TNO
The Hague, the Netherlands
rianne.kaptein@tno.nl

Abstract

In this paper, we evaluate query suggestion for Boolean queries in a news monitoring system. Users of this system receive news articles that match their running query on a daily basis. Because the news for a topic continuously changes, the queries need regular updating. We first investigated the users' working process through interviews and then evaluated multiple query suggestion methods based on pseudo-relevance feedback. The best performing method generates at least one relevant term among 5 suggestions for 25% of the searches. We found that expert users of news retrieval software are critical in their selection of query terms. Nevertheless, they judged the demo application as clear and potentially useful in their work.

1 Introduction

LexisNexis Publisher[1] is an online tool for news monitoring. Hundreds of organizations in Europe and the US use the tool to collect news articles relevant to their work. An organization typically monitors multiple topics. For monitoring the news for a user-defined topic, LexisNexis Publisher takes a Boolean query as input, together with a selection of news sources and a date range. Two example queries can be found in Table 1.

Table 1: Examples of Boolean queries

Topic	Boolean query
Products in the News	"Output Campaign Manager" or "TransPromo" or "Output Wrap Envelope" or "Adsert" or "OptiMail" or "ePriority" or "output Address Direct" or "PredictionPro" or "offmydesk"
Diversity	Diversity /2 inclusion OR "equal employment" or discrimination or harassment or race or gender or religion or "national origin" or disability

Interviews with users of LexisNexis Publisher indicate that noise in the set of retrieved documents is not very problematic because the user has the option to disregard irrelevant documents in the selection, thereby controlling precision. Recall is more difficult to control because the user does not know the documents that were not found. For the user, it is important that no relevant news stories are missed. Therefore, the query needs to be extended when there are changes to the topic. This can happen when new terminology becomes relevant for the topic (e.g. 'wolf' for the topic 'biodiversity'), when there is a new stakeholder (e.g. the name of the new minister of economic affairs for the topic 'industry and ICT') or when new geographical names are relevant to the topic (e.g. 'Lesbos' for the topic 'refugees'). The goal of the current work is to support users of news monitoring appli-

In: M. Martinez, U. Kruschwitz, G. Kazai, D. Corney, F. Hopfgartner, R. Campos and D. Albakour (eds.): Proceedings of the NewsIR'16 Workshop at ECIR, Padua, Italy, 20-March-2016, published at http://ceur-ws.org

[1] http://www.lexisnexis.com/bis-user-information/publisher/

cations by providing them with suggestions for new query terms in order to retrieve more relevant news articles.

Our intuition is that documents that are relevant but *not* retrieved for the current query have similarities with the documents that *are* retrieved for the current query. Therefore, our approach to query suggestion is to generate candidate query terms from the set of retrieved documents.

In this paper, we present the results of a user study in which we evaluate our methodology for query term suggestion with 9 expert users of LexisNexis Publisher. We first conducted interviews with the users to collect their wishes and needs. Then we developed a demo application for news retrieval with query term suggestion functionality. We used this application to evaluate our approach and compare 12 different methods for query term suggestion.

2 Related work

The task of spotting novel terms in a news stream is related to research on topic detection and tracking (TDT) which has its roots in the 1990s [2, 1]. TDT aims to automatically detect new topics or events in temporally-ordered news streams, and to find new stories on already known topics. The functionality of LexisNexis Publisher is related to news tracking in TDT: the topic is given (in the form of a query) and the tool is expected to find relevant new stories in the news stream [14]. More recent work on TDT is directed at topic tracking in microblog data (Twitter) [10, 5]. Microblog data, like news data, is temporally ordered data that continuously changes.

Our approach to query suggestion – generating candidate query terms from the set of retrieved documents – is related to pseudo-relevance feedback [3], a method for query expansion that assumes that the top-k retrieved documents are relevant, extracting terms from those documents and adding them to the query. Pseudo-relevance feedback has been applied to microblog retrieval, expanding the user query with related terms from retrieved posts to improve recall [6, 8]. It is important to take into account that the language use around a topic continuously evolves when selecting terms from Twitter and news data. One option is to give a higher score to terms that are temporally closer to query time [6]. Our approach to query term suggestion is related to this idea: we aim to find the terms that are prominent in the most recent news articles on a topic.

There are two key differences between pseudo-relevance feedback and our approach: First, instead of adding terms blindly, we provide the user with suggestions for query adaptation. Second, we deal with Boolean queries, which implies that we do not have a relevance ranking of documents to extract terms from. This means that the premise of 'pseudo-relevance' may be weak for the set of retrieved documents.

3 Interviews with expert users

We conducted interviews with three experienced users of LexisNexis Publisher to get to know their way of working, their priorities and their wishes for query assistance. The following paragraphs summarize the insights obtained during these interviews.

Way of working. Queries are not changed frequently; most attention is paid to the initial query. Formulating this query takes several hours up to a whole day. Query constructions with Boolean operators are often re-used, for example to exclude specific sources or newspaper sections. If a query gives too much noise, exclusions are added (using the 'NOT' operator). If a query gives too few results, new terms are added (with the 'OR' operator). Changes that are made a later stage are often changes in person and place names. Some customers have difficulties formulating good Boolean queries. These customers make use of information specialist at LexisNexis to formulate their queries.

Priorities. The experts we interviewed use LexisNexis Publisher to create newsletters for their organization. Typically, they review all the retrieved articles before deciding which are included in the newsletter. This selection is based on redundancy and relevance; in case of overlapping news articles, the longest story from the most reliable source is selected. This is done manually, as it allows users to control the precision of the news articles included in the newsletter. The users indicate that for this reason, it is especially important that no relevant documents are missed by the search. Noise in the result set is not so much an issue; if half of the retrieved articles is relevant, the users are satisfied.

Wishes for query assistance. Users indicate that assistance in query formulation could be helpful, not only when adapting existing queries, but especially when formulating new queries. The users mention assistance in the form of: (a) suggestions of new query terms; (b) suggestions for deleting query terms that give too much noise; (c) suggestions for deleting query terms that give very few results. Of these three tasks, we concentrated on the first: suggesting potential new query terms. One requirement posed by the users is that the user still has full control over the query. Terms should not be added blindly, but be presented as suggestions.

4 Methodology

Our approach to query suggestion is to generate candidate query terms from the set of retrieved documents.[2] The central methodology needed for generating terms from a document collection is term scoring; each candidate term from the document collection is assigned a score that allows for selecting the best – most descriptive – terms. The term scoring methods that we use are defined below.

Problem definition. We have a text collection D (the 'foreground collection') consisting of one or more documents. Our goal is to generate a list of terms T with for each $t \in T$ a score that indicates how *descriptive* t is for D. Each t is a sequence of n non-stopwords; we use $n = \{1, 2, 3\}$ in our experiments.

In most term scoring methods, descriptiveness is determined by comparing the relative frequency of t in the foreground collection D to the relative frequency of t in a background collection. For a given Boolean query, we retrieve the result set R_{recent}, which is the set of articles published in the last 30 days, and the result set R_{older}, which is the set of articles published 60 to 30 days ago.

Methods for generating descriptive terms. We compare three methods for generating the most relevant query terms (see Figure 1 for a schematic overview):

A. Return the top-k terms from T_1, generated using R_{recent} as the foreground collection and a generic news corpus as background collection;[3]

B. Return the top-k terms from T_2, generated using R_{recent} as foreground collection and R_{older} as background collection;

C. First generate T_3, using R_{older} as foreground collection and the generic news corpus as background collection. Then return the top-k terms from the set $\{t : t \in T_1 \wedge t \notin T_3\}$ (all terms from T_1 that are not in T_3).

Term scoring algorithms. We implemented four different term scoring algorithms from the literature that we compare for the task of generating potential query terms from the set of retrieved documents:

- Parsimonious Language Models (PLM) [4], designed for creating document models in Information Retrieval. In PLM, the term frequency for each t in D is weighted with the frequency of t in the background collection using an expectation-maximization algorithm;
- Kullback-Leibler divergence for informativeness and phraseness (KLIP) [12]. Informativeness is determined by comparing the relative frequency of t in D to the relative frequency of t in the background collection. Phraseness is determined by comparing the frequency of t as a whole to the frequencies of the unigram that the n-gram t is composed of; Informativeness of t and Phraseness of t are summed to obtain a relevance score for t.
- Frequency profiling (FP) [11], designed for contrasting two separate corpora. This method uses a log-likelihood function based on expected and observed frequencies of a term in both corpora (the foreground and background collections);
- Co-occurrence Based χ^2 (CB) [7], which determines the relevance of t in the foreground collection by the distribution of co-occurences of t with frequent terms in the collection itself. The rationale of this method is that no background corpus is needed because the set of most frequent terms from the foreground collection serves as background corpus.

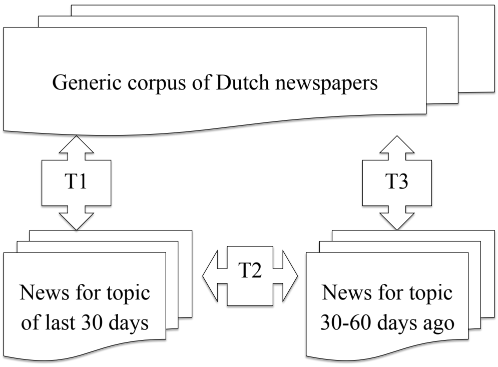

Generic corpus of Dutch newspapers
T1
T3
T2
News for topic of last 30 days
News for topic 30-60 days ago

Figure 1: Schematic view of how the term lists are generated. The query suggester returns one of three term lists to the user: $A = T_1$; $B = T_2$ and $C = \{t : t \in T_1 \wedge t \notin T_3\}$.

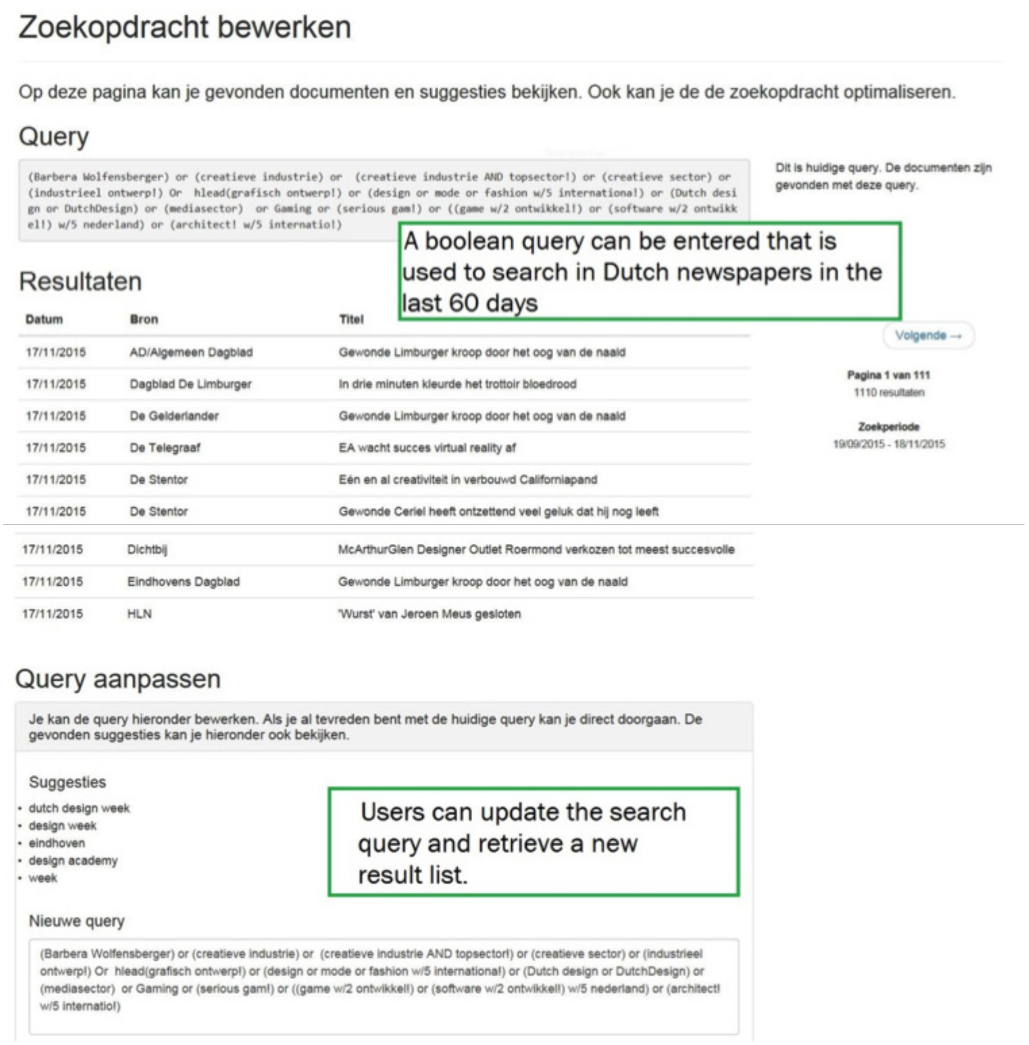
For one query and the corresponding retrieved documents, we generate twelve lists of potential query terms: three different approaches (A–C) with four term scoring algorithms.

5 Experiment and results

We collected feedback from expert users of LexisNexis Publisher to determine the best method for generating term suggestions. For this purpose, we developed an external demo application for news retrieval from the LexisNexis collection that includes query term suggestion functionality. Note that the query term suggestion functionality was not integrated in the existing LexisNexis search interface, but implemented as a standalone web application. Figure 5 shows a screen-

[2] A query term may consist of multiple words.

[3] We used the newspaper section from the Dutch SoNaR-corpus [9], 50 Million words in total. Available at http://tst-centrale.org/producten/corpora/sonar-corpus/6-85



Zoekopdracht bewerken

Op deze pagina kan je gevonden documenten en suggesties bekijken. Ook kan je de de zoekopdracht optimaliseren.

Query

```
(Barbera Wolfensberger) or (creatieve industrie) or  (creatieve industrie AND topsector!) or (creatieve sector) or
(industrieel ontwerp!) Or  hlead(grafisch ontwerp!) or (design or mode or fashion w/5 international!) or (Dutch desi
gn or DutchDesign) or (mediasector)  or Gaming or (serious gam!) or ((game w/2 ontwikkel!) or (software w/2 ontwikk
el!) w/5 nederland) or (architect! w/5 internatio!)
```

Dit is huidige query. De documenten zijn gevonden met deze query.

A boolean query can be entered that is used to search in Dutch newspapers in the last 60 days

Resultaten

Datum	Bron	Titel
17/11/2015	AD/Algemeen Dagblad	Gewonde Limburger kroop door het oog van de naald
17/11/2015	Dagblad De Limburger	In drie minuten kleurde het trottoir bloedrood
17/11/2015	De Gelderlander	Gewonde Limburger kroop door het oog van de naald
17/11/2015	De Telegraaf	EA wacht succes virtual reality af
17/11/2015	De Stentor	Eén en al creativiteit in verbouwd Californiapand
17/11/2015	De Stentor	Gewonde Ceriel heeft ontzettend veel geluk dat hij nog leeft
17/11/2015	Dichtbij	McArthurGlen Designer Outlet Roermond verkozen tot meest succesvolle
17/11/2015	Eindhovens Dagblad	Gewonde Limburger kroop door het oog van de naald
17/11/2015	HLN	'Wurst' van Jeroen Meus gesloten

Pagina 1 van 111
1110 resultaten

Zoekperiode
19/09/2015 - 18/11/2015

Query aanpassen

Je kan de query hieronder bewerken. Als je al tevreden bent met de huidige query kan je direct doorgaan. De gevonden suggesties kan je hieronder ook bekijken.

Suggesties

- dutch design week
- design week
- eindhoven
- design academy
- week

Users can update the search query and retrieve a new result list.

Nieuwe query

(Barbera Wolfensberger) or (creatieve industrie) or (creatieve industrie AND topsector!) or (creatieve sector) or (industrieel ontwerp!) Or hlead(grafisch ontwerp!) or (design or mode or fashion w/5 international!) or (Dutch design or DutchDesign) or (mediasector) or Gaming or (serious gam!) or ((game w/2 ontwikkel!) or (software w/2 ontwikkel!) w/5 nederland) or (architect! w/5 internatio!)

Figure 2: A screen shot illustrating the functionality of the demo application for query term suggestion.

shot of the demo application.[4] The user interface is in Dutch. In the top part of the screen ('Zoekopdracht bewerken' – 'Edit search'), the user sees the current query and the results ('Resultaten') retrieved for that query. In total, 1110 results were retrieved for this query. In the bottom part of the screen ('Query aanpassen' – 'Adapt query'), the user sees a list of term suggestions. This example illustrates the final functionality, in which only the 5 suggestions by the best performing method are shown. In the experimental setting, the user saw a pool of 10–25 terms from different methods.

5.1 Evaluation design

The query suggestion software was evaluated by 9 individual users of LexisNexis Publisher. A 2-hour evaluation session was organized for each participant. The interviews described in Section 3 revealed that queries change more frequently when they are novel. Therefore, each participant was asked to perform two different tasks with the assistance of our demo application during the evaluation session. In the first task, the participant is asked to update a query that is already being used by his company. In the second task, the participant designs a new query for a topic of which they received a short topic description.

The initial (existing or new) Boolean query is issued in LexisNexis Publisher through its API, searching in Dutch newspapers of the last 60 days (the maximum posed by the API). The titles and abstracts of the matching news articles are shown in a result list (in chronological order) and a list of query term suggestions is presented. The participant reviews the set of retrieved documents and improves the query by adding and/or removing terms, optionally using a term from

[4] A video demonstrating the demo application can be viewed here: https://youtu.be/4yIYpvHVugQ

the suggestions. Subsequently, the updated query is issued and the query can be improved again. In both tasks, the participant was asked to review and update the query up to a maximum of five iterations. After the complete evaluation session, the participants filled in a post-experiment questionnaire, in which they could provide additional comments.

5.2 Data

The participants issued 83 searches in total. The Boolean queries are long: 45 terms on average. Terms can be single words or phrases (multi-word terms), and they are combined with Boolean operators. We used the LexisNexis Publisher API to retrieve documents (news articles) published in the last 60 days. On average, $1,031$ documents were retrieved per query (ranked by date), with an average length of 63 words. The short document length is caused by the API allowing us to extract only the summary of the news article, not the full text. This means that the size of the subcollection from which potential new query terms are extracted for a query is on average $1,031*63 = 64,953$ words.

We created a pool of terms from the 12 (3 approaches * 4 term scoring algorithms) term lists per topic. We assume that in a real application, the query suggestion software would show five candidate terms to the user, and we want to be able to evaluate these 5 suggestions for each method. Therefore, the top 5 terms from each term list were added to the pool. The maximum number of terms in a pool is 60 (12*5) but in reality there is quite some overlap: the number of terms per pool is between 10 and 25. For each query, the participants were presented with this pool of 10–25 terms. The terms were ranked by the number of top-5 lists they appear in: the terms that were extracted by most methods were ranked on top of the pool.

5.3 Experimental Results

The selection of query terms and the relevance judgments for the suggested terms in the pool allow us to evaluate and compare the methods. For each method, we have judgments for the 5 highest scoring terms. We count how often one of these terms was selected by a participant, and how often at least one of these terms received a relevance rating of at least 4. The results are in Table 2 and Table 3. The results for the best performing methods (method A with either FP or KLIP as term scoring algorithm, or method C with KLIP) are marked with boldface in the tables. With these methods, participants selected a term from the top-5 suggestions for 13% of the searches, and judged at least one term from the top-5 suggestions as relevant (relevance score >= 4) for 25% of the searches.

Table 2: Results per method in terms of 'selected-success-rate': the percentage of searches for which participants added a term from the top-5 to the query.

	CB	FP	KLIP	PLM
$A = T_1$	10%	**13%**	11%	11%
$B = T_2$	10%	7%	6%	6%
$C = \{t : t \in T_1 \wedge t \notin T_3\}$	10%	0%	11%	11%

Table 3: Results per method in terms of 'relevant-success-rate': the percentage of searches for which participants judged at least a term from the top-5 as relevant (relevance score >= 4).

	CB	FP	KLIP	PLM
$A = T_1$	14%	**24%**	**25%**	20%
$B = T_2$	14%	11%	13%	5%
$C = \{t : t \in T_1 \wedge t \notin T_3\}$	14%	11%	**25%**	20%

The average rating given to the terms in the pool was low: 1.36 on a 5-point scale.

Further analysis of the results showed that the term suggestions were noisy because the sets of retrieved documents are noisy. The Boolean queries return a large set of documents (more than a thousand on average for the last 60 days), without any relevance ranking. The interviews with the users indicated that this is not a problem for the users (because they filter the news items for the newsletter), but it turns out to be a problem for the extraction of relevant terms. In other words, the premise of 'pseudo-relevance' does not hold for Boolean retrieval, and this hurts the quality of query term suggestion based on retrieved documents.

5.4 Qualitative feedback

In the post-experiment questionnaire, participants indicated that the demo application was clear and intuitive (median score of 4 on a 5-point scale for the statement 'the web application is clear'). Half of the participants would be interested in using the tool. However, they felt that the quality of the terms should be improved for the application to be really useful. Suggestions that were provided by the users included:

- Do not to suggest terms that are already covered by wildcards in the query. We improved this in the final version of the demo application.
- Terms that occur in important parts of the text should be more relevant. In fact, this was already taken into account because the API only allowed us to access the abstracts of the documents.
- Multi-word terms should not be suggested. This comment appeared to be in contrast with the users' term selections: of the selected terms by the users (15), the majority (12) are multi-words.

- Add suggestions for the use of Boolean operators. This was beyond the scope of the current project, which focused on term suggestion.

6 Conclusions

The results of our user experiment show that with the best performing method, participants selected a term from the top-5 suggestion list for 13% of the topics, and judged at least one term as relevant for 25% of the topics. Inspection of the results and the post-task questionnaire revealed that the term suggestions are noisy, mainly because the set of retrieved documents for the Boolean query is noisy. We expect that the use of relevance ranking instead of Boolean retrieval, and a post-filtering for noisy terms, will give better user satisfaction.

The relevance judgments for the suggested terms are low compared to another application area for term extraction that we addressed in previous work with the same methodology, namely author profiling [13]. This can partly be explained by the noise in the set of retrieved documents (irrelevant documents lead to irrelevant terms), but may also be caused by expert users of news retrieval software being critical in their selection of query terms. This shows that it is valuable to evaluate query suggestion technology with real users.

References

[1] Allan, J.: Topic detection and tracking: event-based information organization. Volume 12. Springer Science & Business Media (2002)

[2] Allan, J., Papka, R., Lavrenko, V.: On-line new event detection and tracking. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, ACM (1998) 37–45

[3] Cao, G., Nie, J.Y., Gao, J., Robertson, S.: Selecting good expansion terms for pseudo-relevance feedback. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, ACM (2008) 243–250

[4] Hiemstra, D., Robertson, S., Zaragoza, H.: Parsimonious language models for information retrieval. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, ACM (2004) 178–185

[5] Lin, J., Snow, R., Morgan, W.: Smoothing techniques for adaptive online language models: topic tracking in tweet streams. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM (2011) 422–429

[6] Massoudi, K., Tsagkias, M., de Rijke, M., Weerkamp, W.: Incorporating query expansion and quality indicators in searching microblog posts. In: Advances in Information Retrieval. Springer (2011) 362–367

[7] Matsuo, Y., Ishizuka, M.: Keyword extraction from a single document using word co-occurrence statistical information. International Journal on Artificial Intelligence Tools **13**(01) (2004) 157–169

[8] Metzler, D., Cai, C., Hovy, E.: Structured event retrieval over microblog archives. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics (2012) 646–655

[9] Oostdijk, N., Reynaert, M., Monachesi, P., Van Noord, G., Ordelman, R., Schuurman, I., Vandeghinste, V.: From d-coi to sonar: a reference corpus for dutch. In: LREC. (2008)

[10] Phuvipadawat, S., Murata, T.: Breaking news detection and tracking in twitter. In: Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on. Volume 3., IEEE (2010) 120–123

[11] Rayson, P., Garside, R.: Comparing corpora using frequency profiling. In: Proceedings of the workshop on Comparing Corpora, Association for Computational Linguistics (2000) 1–6

[12] Tomokiyo, T., Hurst, M.: A language model approach to keyphrase extraction. In: Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18, Association for Computational Linguistics (2003) 33–40

[13] Verberne, S., Sappelli, M., Kraaij, W.: Term extraction for user profiling: Evaluation by the user. In: UMAP Workshops. (2013)

[14] Yamron, J., Carp, I., Gillick, L., Lowe, S., Van Mulbregt, P.: Topic tracking in a news stream. In: Proceedings of DARPA Broadcast News Workshop. (1999) 133–136

An Analysis of Novelty Dynamics in News Media Coverage

Ronaldo Cristiano Prati
Universidade Federal do ABC
Santo André, São Paulo, Brazil
ronaldo.prati@ufabc.edu.br

Walter Teixeira Lima Júnior
Univerisdade Federal do Amapá
Macapá, Amapá, Brazil
contato@walterlima.net

Abstract

Computer Science has affected almost all fields of human knowledge, contributing to scientific advances in many branches of Natural and Social Sciences. Journalism is one of the fields that is benefiting of the advance of computer science. Among the journalistic concepts that can be analyzed computationally is News Value. Novelty is one of the most important news value. A possible approach to get novelty elements in a story considers word frequency, through of the capacity to collect and analyze massive amounts of data. In this paper, we use the News Coverage Index dataset (NCI), maintained by the Pew Research Center, to analyze the novelty dynamics of news coverage, using the novelty signatures proposed by [12]. As a definition of novelty, we used the first appearance of a new lead newsmaker. Results show a good fit of the model to the dataset. Furthermore, an analysis by media sector and broad topic shows interesting insights for the analysis of media coverage.

1 Introduction

The Computational Science has affected almost all fields of human knowledge, contributing to scientific advances in many branches of Natural and Social Sciences. For instance, the capacity to collect and analyze massive amounts of data has transformed intensely fields such as biology and physics [7].

In Social Science, despite the difficulties to formalize computationally many scientific subjects of the human behavior, "a computational social science is emerging that leverages the capacity to collect and analyze data with an unprecedented breadth and depth and scale" [7]. Unfortunately, most of the advances in this area have been progressing at a much slower pace. However, substantial barriers that might limit progress are being overcome in recent years. The emergence of a powerful new field of data analysis of Social Science has also influenced the research on a branch of it, Journalism. Journalism is an important social practice. Therefore, to find non-trivial information on content produced by journalism, it is necessary to count with the support of the current stage of technologies to advance in analytical techniques "Computation can advance journalism by drawing on innovations in topic detection, video analysis, personalization, aggregation, visualization, and sense making [10].

Among the journalistic concepts that can be analyzed computationally is News Value. News value as a concept was thought by Johan Galtung and Mari Holmboe Ruge's seminal publication in the Journal of Peace Research. In 1965, the paper suggested a range of attributes that establish news values in discursive elements contained in newspapers and broadcast news. Galtung and Ruge established the news values elements as Frequency; Threshold; Unambiguity; Meaningfulness; Consonance; Unexpectedness; Continuity; Composition; Reference to Elite Nations; Reference to Elite People; Reference to Persons; and Reference to Something Negative [3]. These factors have been the base to compose the structure of the theory of newsworthiness. The theory is based on the psychology of individual perception and explain which factors influence newsworthiness of an event [6].

News values are studied considering a range of at-

In: M. Martinez, U. Kruschwitz, G. Kazai, D. Corney, F. Hopfgartner, R. Campos and D. Albakour (eds.): Proceedings of the NewsIR'16 Workshop at ECIR, Padua, Italy, 20-March-2016, published at http://ceur-ws.org

tributes contained in discursive elements. It is also possible to verify the news value through a range of "more specific cognitive constraints that define news values (Novelty, Regency, Presupposition, Consonance, Relevance, Deviance and Negativity, Proximity) [2]. The news value named Novelty can be analyzed by words such as reveal or revelation. These words announce semantically 'unexpected aspects of an event News stories are frequently about happenings that surprise us, that are unusual or rare' [1].

The novelty can be understood by concepts as out of the ordinary, least expected, or not predicted, news values relating to the novelty, newness or unexpectedness of an event/happening [2]. The quality of being interesting enough to the public (newsworthiness) is also based on if a journalistic fact is out of the ordinary, it will have a greater effect than something that is an everyday occurrence (unexpectedness). The unexpectedness power of attraction is in the factor that "there is new information that has been uncovered and evaluations of importance can make the eliteness of a source explicit" [2]. This means that readers or viewers can know facts or different people or unusual to their quotidian, however, "this is the old man-bites-dog syndrome which needs little more explanation" [9, 2]. When a fact or term first come up, the human attention is captured, but "the fact that the novelty of a story tends to fade with time and thus the attention that people pay for it. This can be due to either habituation or competition from other new stories" [13].

As previously observed, novelty is also elaborated "mainly through using evaluative language, references to surprise/expectations and comparisons" [1]. This way of perception of novelty on the construction of journalistic contents is based on analyzes produced by reading the news. However, it is possible to get novelty elements in the story considering word frequency, through the capacity to collect and analyze massive amounts of data. Over the years, there is a massive increase in the availability of journalistic data and creation of new tools to extract the value from data that are helping to understand our lives, organizations, and societies.

In this paper, we used a recent model of novelty dynamics to analyze news coverage. The main idea is to analyze whether different news sources present different novelty dynamics. This paper is organized as follows: Section 2 presents novelty signatures that emerge in some dynamical processes. Section 3 describes the data set used in our study. Section 4 presents the results of applying the novelty signatures to the NCI dataset, and Section 5 concludes the paper.

2 Novelty in dynamical processes

Tria et. al [12] have recently analyzed novelty as new events occurring in a dynamical process evolving over time. Given a sequence of events, a novelty occurs whenever a new element first appears in a sequence. They have analyzed four different data sets: books from Gutenberg project Corpus, annotations in the social bookmarking platform Delicious, songs and singers at Last FM streaming portal and, entries appearance in English Wikipedia. The novelties in these data are, respectively, the occurrence of new words in books, the use of new annotation tags in the bookmarks, the inclusion of a new artist/song in a play list the user had never listen to and the first edition of a page in the collaborative encyclopedia.

They were able to model novelty as a simple mathematical model based on random draws sampling with replacement of an Urn [4] that increases when a novel item is observed. The model predicts statistical laws for the rate at which novelties happen (Heaps' law [5]) and for the probability distribution on the space explored (Zipf's law [14]), as well as signatures of the process by which one novelty sets the stage for another.

The first signature is based on quantifying the rate at which novelties occur in a temporally ordered sequence of elements of length N by analyzing the growth of the number $D(N)$ of distinct elements in this sequence. This relation would imply in a Heap's law, which states that the rate at which novelties occur decreases over time as t^{β}, where β is the coefficient of a power law distribution of $D(N)$ over N fitted over the data.

The second signature is related to the frequency of occurrence of different elements in the data. The frequency-rank distribution would follow an approximate Zipffian distribution (Zipf's law). In this distribution, the frequency of any element is inversely proportional to its rank in the frequency table, *i.e.*, the frequency $F(R)$ of an element at rank R is proportional to $R^{-\alpha}$, where α is the coefficient of a power law distribution of $F(R)$ over R fitted over the data.

It is well known that α and β are inversely correlated [8]. The larger the β coefficient, the higher the frequency of appearance of new elements in the sequence, thus there is a high propensity for novelty. On the other hand, the larger the α coefficient, the higher the occurrence of the most frequent elements in the sequence. The key result reported in [12] is that in the four data sets analyzed, the model was able to capture the novelty behavior in the data. An interesting research question is then whether News delivery also shows these novelty signatures. This paper is an initial attempt towards such analysis.

3 News Coverage Index dataset

In our analysis, we used the data gathered by the Pew Research Center[1]. Every week, this institution produced the News Coverage Index (NCI) by identifying and annotating the main subjects covered by the U.S. mainstream media. The dataset used this research is the most updated dataset (2013), published by Pew Research Center. Until this moment, no other similar dataset that can be used to update the data or serve to comparison.

> The NCI captured and analyzed 52 news outlets in real time to determine what was being covered and what was not in the U.S. news media. The analysis was conducted weekly, Monday - Sunday. The key variables included source, story date, big story, broad story topic, placement, format, geographic focus, story word count, duration of broadcast story and lead newsmaker. The outlets studied came from print, network TV, cable, online, and radio. They included evening and morning network news, several hours of daytime and prime time cable news each day, newspapers from around the country, the top online news sites, and radio, including headlines, the long form programs and talk [11].

By focusing on the topic of the story, the index measures by what percentage of the analyzed news hole is about that topic. Data were collected from January 2007 to May 2012. Table 1 presents the number of news stories collected per year. Note that the year 2012 has a few stories because the collection period ranges from January to May, rather than January to December.

Table 1: Number of news stories per year

Year	2007	2008	2009	2010	2011	2012
CableTV	22823	21892	18856	17087	15324	6472
NetworkTV	21320	19796	19427	13016	11858	5186
Newspaper	6559	7350	7370	5626	5190	1977
Online	6520	6539	7830	7818	7744	3242
Radio	13515	14365	15234	9067	8439	3570
All	70737	69942	68717	52614	48555	20447

The codebook includes variable names, definitions, applicable procedures and changes that were made to certain variables. For each story, it was annotated the date, source, broadcast start time (morning, noon, afternoon, evening and night, or not broadcast), duration in seconds, word counts, placement prominence, story format, big story, geographic focus (local, US national, US international, non-US international), broad story topic, media sector (cable TV, network TV, newspaper, online and radio), and lead newsmaker. The number of outlets and individual programs vary considerably within each media sector, as do the number of stories and size of the audience.

The index is a good source for analyzing, through time, how stories emerge and sink. Other possibilities include how the character or narrative focuses of the story change and how much of the broad topic's categories get more coverage, when compared to the others. However, the index does not provide information for additional possible questions, such as tone, sourcing or other matters.

The key variable chosen in this study was "lead newsmaker", a variable that "determines the person whose actions or statements constitute the main subject matter of the story". In the NCI, the derivation of the "lead newsmaker' variable used a methodology that examined the outlets daily by the coding team. The researchers establish as a definition: variable lead newsmaker determines the person whose actions or statements constitute the main subject matter of the story discussed with at least 50% of the story (in time or space).

Therefore, in our analysis, a news story is flagged as a novelty whenever the first appearance of a new lead newsmaker occurs, considering an ordered sequence of histories by date in the NCI. Obviously, this approach does not completely capture all the aspect of novelty in news coverage. It is perfectly possible (and indeed very common) that some new factor is being published by some lead newsmaker who appeared before. However, this approach does capture some aspect of novelty, in a sense that different subjects are being noticed in the media. Furthermore, the approach sheds some interesting insights, as discussed next.

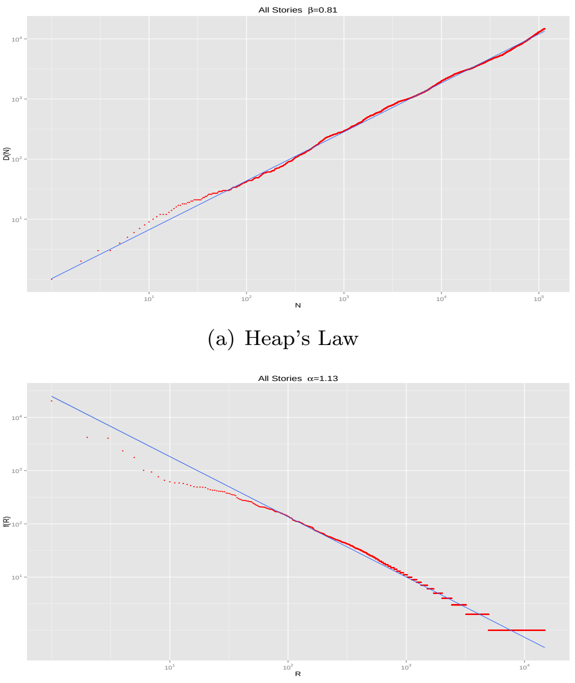
4 Results and Discussion

In this section we present the results of the novelty signatures as proposed by [12] to the NCI dataset. As the main variable used in this study was lead newsmaker, we removed from the dataset all stories where the lead maker was not identified, resulting in a total of 135,205 entries in the dataset.

Figure 1 shows the two novelty signatures for all stories in the NCI dataset, for the Heaps' law and Zipf's law, respectively. The graphs show a very good fit (the blue line in the graphs), indicating that the dynamic of novelties also follows the model proposed in [12] for the NCI dataset.

This is an interesting result per se, but we can move beyond that by conditioning the analysis by some news groups. Figure 2 does this, where we have split the analysis by the media sector (newspaper, online, radio,

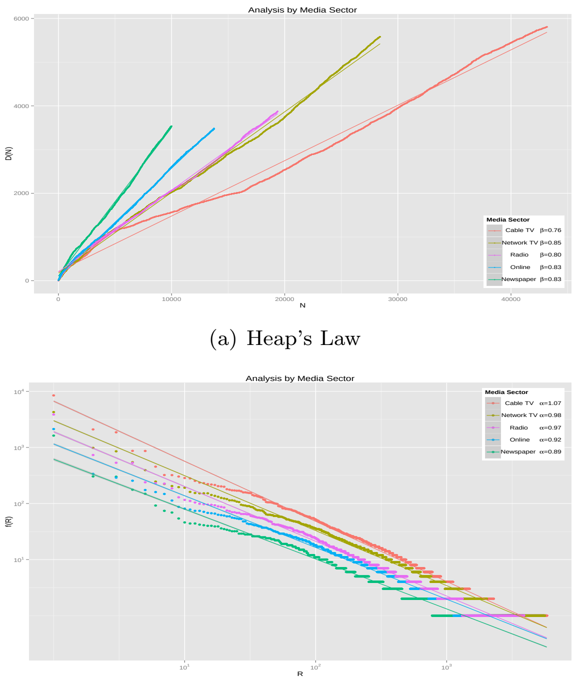
[1] Formerly Project for Excellence in Journalism (PEJ).



(a) Heap's Law

(b) Zipf's Law

Figure 1: Novelty Signatures for lead newsmaker over all stories in NCI dataset. The blue line is the best data fit.



(a) Heap's Law

(b) Zipf's Law

Figure 2: Novelty Signatures for lead newsmaker in NCI dataset grouped by media sector

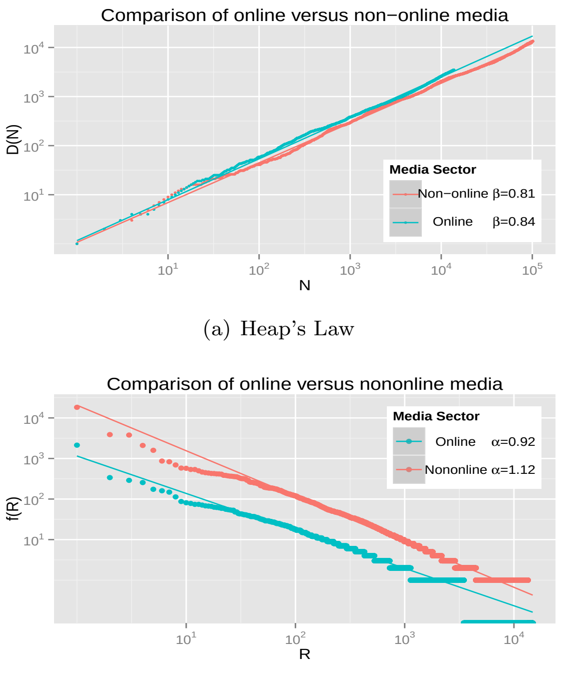
broadcast TV and cable TV). Figure 2(a) shows how novel lead newsmakers appears in the news sequence, for each media sector collected by the NCI. The interpretation of these results is, the steeper the line, the more novelty the media sector has (according to the definition of novelty used in this paper). Surprisingly, newspapers is the media

sector with the larger ratio of lead newsmakers per story, followed by online portals, radio, network TV and cable TV. Figure 2(b) shows an orthogonal insight for this result, which shows the rank distribution of lead newsmakers for each sector. As Heaps' law and Zipfs' law are inverse correlated, the interpretation of these results are, the steeper the line, the more a media sector concentrates the coverage in a few lead makers. Cable TV repeats lead newsmakers more often than other sectors, and (proportionally) uses fewer leads newsmaker than the other media sectors. Newspapers, on the other hand, proportionally use the top ranked lead news makers less often, and have a larger number of histories with different lead makers.

We can speculate that the higher frequency of novel lead newsmakers in newspaper media is due to that this media needs competitiveness in relation to other media (digital and electronic), which are characterized by dissemination of news in real time. As the newspaper is a diary media, it always needs to have something different to present than what was published on the previous day on TV, radio and, Internet. Despite being late in relation to events in one day (it generally publishes stories from the eve), the newspaper still continues to be a source for other rival media because it intends always having something new in their pages. On the other hand, TVs have a rotating audience, and focus on a narrow range of topics. Thus, the presented histories focus in a few lead newsmakers. Radios an online media are somehow in between these two extremes.

To gain some insight in the online versus offline scenario, we break down the analysis in online versus offline media, as shown in Figure 3. The interpretation of the graphs is the same as of 2. Figure 3(a) shows that online sector introduce more lead makers in their stories, and Figure 3(b) indicate that the same lead maker appears less often in online media. As can be seen from the graphs, online media have stronger novelty signatures. Therefore, online media have a bias towards introducing more different lead newsmakers, and a lower tendency to echo the same leading maker in future stories.

A possible reason for this is that online outlets have a high propensity to show new stories due to the difference in media consumption from the target audience. In general, the audience for online news sources is of younger people (as discussed in the previous section). These users have a less tendency to in-depth stories, fo-



(a) Heap's Law

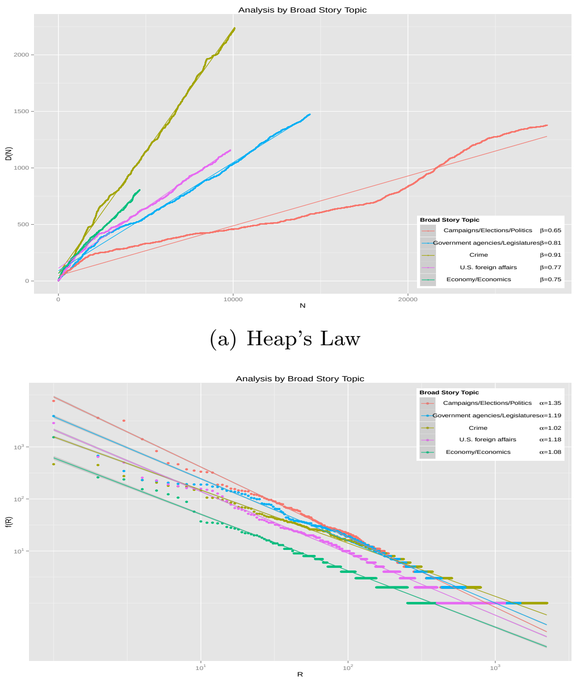
(b) Zipf's Law

Figure 3: Novelty Signatures for lead newsmaker in NCI dataset grouped by online versus offline media

cusing in the headlines. They also are more connected, and access the news more often, thus the necessity of novelty in the news stories.

We did a similar analysis, but conditioning on the five most frequent broad story topics. The broad story topic variable identifies which of the broad topic categories is addressed by a story. NCI has 32 broad story categories, but most of them have low frequencies. These low frequencies difficult an analysis, due to a lack of data. Figure 5 shows these results. The interpretation of the graphs is the same as of 2, except for the fact that instead of media sectors, we have topics in these graphs. Figure 4(a) shows how novel lead newsmakers appear in the news sequence, for each of the five most frequent topics collected by the NCI. In these figures, one traditional attribute of news value, Negativity (any reference that is negative), emerges as through Crime broad story topic. Crime is the topic with the largest rate of novel lead makers, followed by economy/economics, US foreign affairs, government agencies/legislatures and campaigns/elections/politics with the lowest rate. Figure 4(b) indicates that the most frequent lead makers appear proportionally less often in the news than the most frequent lead makers in campaigns/elections/politics. Furthermore, crime is the sector with the largest proportion of lead makers to appear in fewer histories.

A similar analysis was performed by start time of the program, as shown in Figure 6. The interpretation of the graphs is the same as of 2, except for the fact that instead of media sectors, we have the program start time in these graphs. Figure 5(a) shows that, in general, morning programs introduce more often new lead makers, while night programs have few novelty lead makers. On the other hand, Figure 5(b) shows that night program cites more often the more noticed lead makers than morning programs. We believe this also is related to the target audience, which in the evening/night has a higher prevalence of elderly people, which is more interested in-depth coverage.



(a) Heap's Law

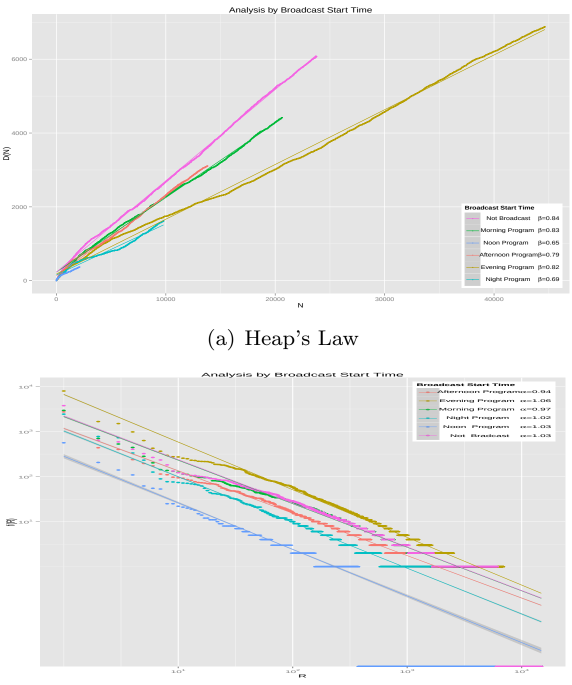
(b) Zipf's Law

Figure 4: Novelty Signatures for lead newsmaker in NCI dataset grouped by media sector (top 5 sectors)

5 Concluding Remarks

In this paper, we examine the dynamic of novelties in the NCI dataset. We used the lead newsmaker as the main variable to define the concept of novelty in our framework. We verified a very good fit of these data to the two novelty signatures discussed in [12].

We obtained interesting and insightful insights when conditioning the analysis do media sector and broad story topic. Regarding media sector, we verified that newspapers is the sector with largest novelty, in terms of the introduction of new lead newsmakers. Furthermore, online media have a largest novelty, when compared to non-line media. In terms of story topic, crime is the sector with more novelty, also in terms of lead newsmakers.



(a) Heap's Law

(b) Zipf's Law

Figure 5: Novelty Signatures for lead newsmaker in NCI dataset grouped by starting time

We believe these patterns somehow tend to follow the interest of the public in order to get her attention. Thus, there is the necessity to provide news on topics to better reach a target audience, tailoring the audience. An interesting future work is to analyze whether these patterns would be similar in the next years, because the online young audience became a generation more mature. Would the behavior be the same, and the sectors have to adapt to the news consumption patterns of this generation or they will change their tastes, showing a similar behavior or their previous generation as accessing the in-depth stories?

This research has two obvious limitations. First, our adopted definition of novelty does not capture all aspects of novelty, as new information can be published about lead Newsmakers which already appeared in the sequence. However, we believe this definition do capture some aspects of novelty, and were able to provide some interesting insights on the topic. Furthermore, the data set has a bias towards the U.S.A. media coverage. An interesting further research direction is to broaden this research to different sources.

References

[1] M. Bednarek and H. Caple. 'value added': Language, image and news values. *Discourse, Context & Media*, 1(2–3):103–113, 2012.

[2] H. Caple and M. Bednarek. Delving into the discourse: Approaches to news values in journalism studies and beyond. Technical report, Reuters Institute for the Study of Journalism, 2013.

[3] J. Galtung and M. H. Ruge. The structure of foreign news the presentation of the congo, cuba and cyprus crises in four norwegian newspapers. *Journal of peace research*, 2(1):64–90, 1965.

[4] J. Haigh. Polya urn models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(4):942–942, 2009.

[5] H. Heaps. *Information Retrieval: Computational and Theoretical Aspects*. Academic Press, New York, 1978.

[6] H. Kwak and J. An. Understanding news geography and major determinants of global news coverage of disasters. In *Computer+Journalism Symposium*, New York, USA, 2014.

[7] D. Lazer, A. Pentland, A. Lada, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne. Life in the network: the coming age of computational social science. *Science*, 323(5915):721–723, 2009.

[8] L. Lü, Z.-K. Zhang, and T. Zhou. Zipf's law leads to heaps' law: Analyzing their relation in finite-size systems. *PLoS ONE*, 5(12):e14139, 12 2010.

[9] M. Masterton. Asian journalists seek values worth preserving. *Asia Pacific Media Educator*, 1(16):41–48, 2005.

[10] B. O'Connor, D. Bamman, and N. A. Smith. Computational text analysis for social science: Model assumptions and complexity. In *Second NIPS Workshop on Comptuational Social Science and the Wisdom of Crowds*, 2011.

[11] Pew Research Center. News coverage index methodology, 2013. `http://www.journalism.org/news_index_methodology/99/`.

[12] F. Tria, V. Loreto, V. D. P. Servedio, and S. H. Strogatz. The dynamics of correlated novelties. *Sci. Rep.*, 4, 2014.

[13] F. Wu and B. A. Huberman. Novelty and collective attention. *Proceedings of the National Academy of Sciences*, 104(45):17599–17601, 2007.

[14] G. K. Zipf. The psycho-biology of language. *Language*, 12(3):196–210, 1935.

Detecting Attention Dominating Moments Across Media Types

Igor Brigadir Derek Greene Pádraig Cunningham

{igor.brigadir, derek.greene, padraig.cunningham}@insight-centre.org
Insight Centre for Data Analytics
University College Dublin, Ireland

Abstract

In this paper we address the problem of identifying attention dominating moments in online media. We are interested in discovering moments when everyone seems to be talking about the same thing. We investigate one particular aspect of breaking news: the tendency of multiple sources to concentrate attention on a single topic, leading to a collapse in diversity of content for a period of time. In this work we show that diversity at a topic level is effective for capturing this effect in blogs, in news articles, and on Twitter. The phenomenon is present in three distinctly different media types, each with their own unique features. We describe the phenomenon using case studies relating to major news stories from September 2015.

1 Introduction

The problem of detecting breaking news events has inspired a host of approaches, extracting useful signals from activity on social networks, newswire, and other types of media. The online communication platforms that have been adopted allow these events to persist in some form. These *digital traces* can never fully capture the original experience, but offer us an opportunity to revisit significant phenomena with different points of view, or help us to characterise and learn something about the processes involved. Many different forms of news media attempt to record and disseminate information deemed important enough to communicate, and as the barriers to broadcasting and sharing information are removed, attention becomes a scarce commodity.

We define the problem of detecting *attention dominating moments* across different media types, as a collapse in diversity in the content generated by a set of online sources in a topic during a given time period. *Media types* here include mainstream news articles, blog posts, and tweets. These media types differ in both the category of topics covered [22], and their use of language [10]. In the context of Twitter, we define *sources* as unique user accounts. For mainstream news and blogs, sources refer to individual publications or outlets. Publications may have different numbers of authors, but as unique author information is not available, we treat each unique blog or news outlet as a single source.

In Section 3, we describe the two stages of our proposed event detection procedure. In the first stage, content generated by the news, blog and tweet sources is grouped into broad topical categories, through the application of matrix factorization to the content generated by these sources. In the second stage, we examine the variation in similarity between content generated by sources within a given topic during a given time period, in order to identify a collapse in diversity within a topic which corresponds to an attention dominating moment. In Section 5, we evaluate this procedure on a collection of one million news articles and blog posts from September 2015, along with a parallel corpus of tweets collected during the same time period.

Rather than formulating the problem as tracking the evolution of topics themselves, we consider the diversity of content within a specific topic over time. The motivation is that, for instance, a collapse in diversity around a major sporting event will be strongly evident in certain news sources, but not evident in others.

In: M. Martinez, U. Kruschwitz, G. Kazai, D. Corney, F. Hopfgartner, R. Campos and D. Albakour (eds.): Proceedings of the NewsIR'16 Workshop at ECIR, Padua, Italy, 20-March-2016, published at http://ceur-ws.org

The distinction is important, as this approach is more suited to retrospective analysis, when the entire collection of documents of interest is available. The topics do not change over time, as opposed to a real-time setting where topics must be updated as new documents arrive [21]. The information need is guided by two major questions. Firstly, when have significant collapses in diversity occurred in a topic of interest? Secondly, are there differences between media types when these events occur?

Our main contributions here are: 1) a diversity-based approach of detecting attention dominating news events; 2) a comparison between traditional news sources, blogs, and Twitter during these events. 3) a parallel corpus of newsworthy tweets for the NewsIR dataset.

2 Related Work

In previous work, attention dominating news stories have been described as *media explosions* [2] or *firestorms* [14]. The idea of combining signals from multiple sources for detecting or tracking evolution of events proved effective in the past. Osborne *et al.* [16] used signals from Wikipedia page views, together with Twitter to improve "first story detection". Concurrent Wikipedia edits were used as a signal for breaking news detection in [19].

Topic modeling applied to parallel corpora of news and tweets has been previously explored by a number of researchers [6, 9, 11]. Extensions to LDA to account for tweet specific features have been proposed [22]. A comparison between Twitter and content from newswires was explored in [18]. A Non-negative Matrix Factorization (NMF) approach is used for topic detection in [20].

How offline phenomena link to bursty behaviour online is discussed in [5] and [12]. In [12] Shannon's Diversity Index was used to detect a "contraction of attention" in a tweet stream by measuring diversity of hashtags. In contrast, we employ a different measure of diversity based on document similarity, applying it to streams from different media types segmented by topic. Methods for automatically detecting anomalies or significant changes in a time series are discussed in [4]. In [15] a change-point detection approach is applied to time series constructed from Tweet keyword frequencies.

As a broad overview, the common components involved in detecting high impact, attention dominating news stories include: selecting relevant subsets of documents; representation and feature extraction; constructing time series from features; event detection and analysis. In this paper we concentrate on a single key feature of breaking news: a collapse in content diversity within a fixed time window.

3 Proposed Method

Our objective is to detect when multiple articles in a topical stream become less diverse, signalling the emergence of an attention dominating news story. We consider attention to a phenomenon as the main driving force behind the decision to produce or broadcast a communication. Using the diversity of content within a time window, we attempt to characterise instances where a particular piece of information becomes dominant. Concretely, for each type of media, NMF is used to assign topics to documents; for documents in a topic, we calculate diversity between documents in a time window. This type of analysis allows us to examine the extent to which the onset of an important breaking news event is accompanied by a collapse in textual content diversity, both within a group of news sources and across different media types.

3.1 Finding Topics

We apply a Non-negative Matrix Factorization (NMF) topic modeling approach to extract potentially interesting topics from a stream of tweets or set of articles. For each *media source*, we build a tf-idf weighted term-document matrix and use this as input to NMF.

We also considered LDA to infer topics in these datasets. The choice of NMF over LDA was primarily due to computation time. LDA was significantly more computationally expensive than NMF with NNDSVD [1] initialisation. NMF also tends to produce more coherent topics [17].

3.2 Measuring Diversity

The same tf-idf representation used for topic modeling is used in diversity calculations. Each article, blog post or tweet is a tf-idf vector. A separate document-term matrix is built for each *media type*. Stopwords and words occurring in fewer than 10 documents are removed.

To measure diversity, we calculate the mean cosine similarity between all unique pairs of articles within a topic for a fixed time window. Given a set of documents D in a time window, the diversity is:

$$diversity(D) = -\frac{\sum_{i,j \in D, i \neq j} cosSim(D_i, D_j)}{\sum_{i=1}^{|D|-1} i}$$

Where $cosSim(D_i, D_j)$ is the cosine similarity of tf-idf vectors of documents i and j in a time window. In practice, calculating similarities between all pairs of documents can be efficiently performed in parallel, and can be calculated in a matter of seconds.

Longer time windows consider more document pairs, which naturally result in smoother trends. In contrast, shorter time windows are more sensitive to brief attention dominating events, but also false positive spikes—where a small number of articles happen to be similar in content, but do not constitute an attention dominating story.

An alternative to content diversity is also considered. Ignoring document content, and just considering the sources of articles, diversity is calculated with Shannon's Diversity Index:

$$H' = -\sum_{i=1}^{R} p_i \ln p_i$$

Where p_i is the proportion of documents produced by the ith source in a time window of interest, R is total number of sources in a given media type.

Both diversity measures produce a single diversity value per time window, generating a univariate time series. Changes in diversity that are 2 standard deviations away from the mean are naively considered to be important enough to warrant attention. Exploring more robust and well established methods for change point detection such as [15, 4] is left for future work.

For the case studies described in Section 5, the window length was set to 8 hours. While the fast-paced "24/7 news cycle" is described as a constant flood of information, we find that all three mediums largely follow a more traditional publishing cycle, with prominent spikes in number of published articles on weekday mornings, and low numbers of articles published outside of normal office hours. A more detailed analysis of publishing times and characteristics will be explored in future work.

4 Datasets

To explore attention dominating news stories, we apply the method described above to three media sources: mainstream news, blogs, and tweets. For the first two sources, the NewsIR dataset[1] is used. For the final source, we use our own parallel corpus collected from Twitter[2]. In contrast to previous work [6, 11] where tweets are retrieved based on keywords extracted from news articles, the parallel corpus was derived from a large set of newsworthy sources, curated by journalists [3]. Journalists on Twitter curate lists[3] of useful sources by location or general topic of interest—for example "US Politics" may contain accounts of US politicians and other journalists who tend to cover US politics related stories.

Gathering all members of such lists covering different countries and topics follows the *expert-digest* strategy from [7]. A tweet dataset collected independently of news and blog articles preserves Twitter-specific features and topics. Source and document counts are summarised in Table 1.

Media Type	*Sources*	*Documents*	*Docs. per 24h*
News	18,948	730,634	8,177
Blogs	73,403	253,488	23,568
Tweets	30,448	3,274,089	125,568

Table 1: Summary of overall source and document counts by media type after filtering, and average number of documents in a 24 hour window.

Of the original 1 million articles provided, 15,878 were filtered as non-English[4] or outside the date range of interest (*i.e.* created between 2015-09-01 and 2015-09-31). Tweet language filtering was performed using meta-data provided in the tweet.

5 Attention Dominating Events

In order to compare the same topics across different media types, we compare the top 10 terms representing the topics from different models. Specifically, when topics from two different models have strongly-overlapping (using Jaccard similarity) top term lists, this indicates that similar events were discussed in both media types.

Topics in a model that do not have any overlapping terms with topics in other models, suggest that content unique to a platform is prominent. For example: the "*live, periscope, follow, stream, updates*" topic in the tweet corpus has no equivalent among the news or blog topics. This reflects the fact that the Periscope app became popular with journalists for broadcasting short live video streams and Twitter is the main platform where these streams are announced. The "*music, album, song, video, band*" topic is prominent in the blogs and Twitter, but is not present in news. This may reflect the fact that most Twitter accounts and blogs are far more personal in nature.

An indicative, but not necessary feature of attention domination news is the presence of a similar topic on multiple platforms. To illustrate the phenomenon of topical diversity collapse, we now describe three case studies.

[1] Available from: `http://research.signalmedia.co/newsir16/signal-dataset.html`

[2] Data: `https://dx.doi.org/10.6084/m9.figshare.2074105`

[3] Examples of such lists are available `https://twitter.com/storyful/lists/` and `https://twitter.com/syflmid/lists`

[4] `https://github.com/optimaize/language-detector` was used for language detection. Interestingly, language detection proved effective for filtering "spammy" articles containing obfuscated text, large numbers of urls, or containing tabular data.

For each case study, we present the following: Top 10 topic terms for a topic in a media type, and a plot of diversity over time, where:

- Solid lines show diversity of documents over time.
- Dashed lines show Shannon Diversity of sources.
- Highlighted time periods are when major developments occurred—based on Wikipedia Current Events Portal[5] for September 2015.
- Dot and Triangle markers indicate periods when diversity drops 2 standard deviations below the mean.

5.1 European Refugee Crisis

The European crisis began in 2015, as increasing numbers of refugees from areas in Syria, Afghanistan, and Western Balkans [8] sought asylum in the EU. Figure 1 shows a plot of diversity for the documents assigned to this topic in each 8 hour time window, for the three media types. To help with visualisation, raw diversity values are standardised with z-scores on the y axis, while the x axis grid separates days.

Media	*Top 10 Topic Terms*
Blogs	refugees, syria, syrian, war, president, government, military, europe, russia, iran
News	refugees, migrants, border, hungary, eu, europe, european, refugee, asylum, germany
Tweets	refugees, syrian, hungary, help, migrants, europe, border, germany, austria, asylum

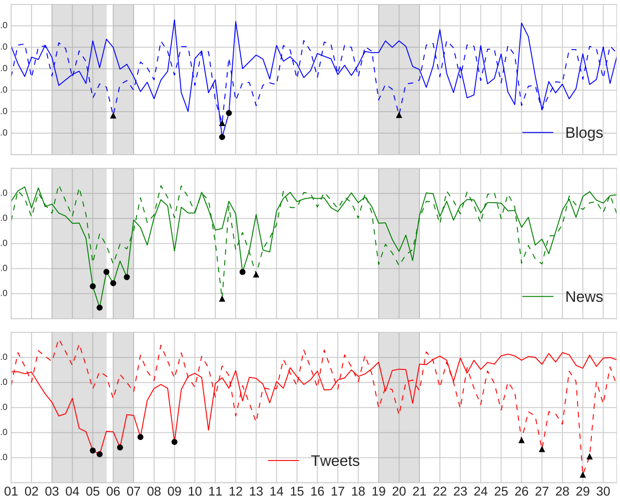


Figure 1: Standardised diversity scores for the European refugee crisis topic during September 2015, across three media types.

The downward trend in diversity between September 3rd and 5th in the refugee crisis topic can be explained by the death of Aylan Kurdi. News of his drowning quickly spread online and made global headlines. This was a particularly far-reaching story, dominating news coverage until an announcement on relaxing controls on the Austro-Hungarian border by Chancellors Faymann of Austria and Merkel of Germany. Both Twitter and mainstream news streams experienced a diversity collapse, while Blogs maintained more diverse set of articles. Between 19th and 21st, smaller drops in diversity coinside with Pope Francis' visit, where the issue of refugees was a prominent topic of discussion.

5.2 Donald Trump Presidential Campaign

Donald Trump's presidential campaign has attracted considerable attention across all types of media[6]. Positions on issues of immigration and religion are particularly polarising, frequently causing controversies in mainstream media.

Media	*Top 10 Topic Terms*
Blogs	trump, donald, republican, presidential, debate, gop, president, candidates, candidate, bush
News	trump, republican, presidential, donald, debate, clinton, bush, fiorina, candidates, campaign
Tweets	trump, im, love, donald, going, debate, happy, gop, president, think

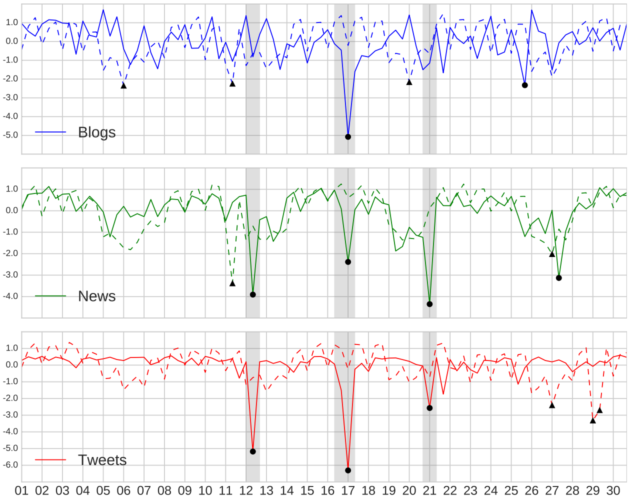


Figure 2: Standardised diversity scores for Donald Trump Presidential Campaign topic

Significant events marked around 12th, 17th, 21st in Figure 2 relate to: Trump's comments on Senator Rand Paul on Twitter which was discussed on mainstream news around 12th, but not as prominently on blogs. On the 16th-17th coverage of a republican presidential debate hosted by CNN; and 21st—mainstream news coverage of reactions to events on 17th: during

[5] https://en.wikipedia.org/wiki/Portal:Current_events/September_2015

[6] https://en.wikipedia.org/wiki/Donald_Trump_presidential_campaign,_2016

a town hall meeting in Rochester, Donald Trump declined to correct a man who said that President Obama is a Muslim.

The statement prompted a significant drop in the diversity of stories across all platforms. On the 25th, during a speech given to conservative voters in Washington, Trump called fellow Republican presidential candidate Marco Rubio “a clown”. Based on the data, it appears that the reaction to the latter on Twitter was not as pronounced as among journalists and bloggers.

5.3 Pope Francis visits North America

The visit of Pope Francis spanned 19 to 27 September 2015, where the itinerary included venues in both Cuba and the United States. This event is a good illustrative example as it was widely documented[7], and highlights a case where a collapse in diversity did not occur at the same time on different media platforms.

Media	*Top 10 Topic Terms*
Blogs	pope, francis, church, catholic, visit, cuba, popes, climate, philadelphia, vatican
News	pope, francis, catholic, church, philadelphia, popes, cuba, united, vatican, visit
Tweets	pope, francis, visit, house, congress, popeindc, cuba, white, popeinphilly, philadelphia

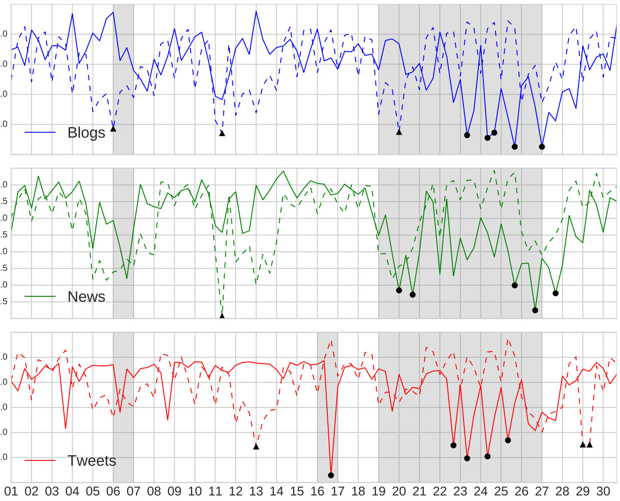


Figure 3: Standardised diversity scores for the Papal visit topic during September 2015.

In the case of news publishers, the largest drop in diversity coincided with the beginning of the Pope’s visit to Havana. Twitter users and bloggers reacted more on September 23rd and 24th, when the Pope met with Barack Obama and became the first Pope to address a joint session of US Congress.

In the Twitter stream, the notable event around 16th-17th is due to large numbers of similar tweets as preparations for the visit were being discussed, and #TellThePope trended briefly.

Earlier in the month, we see evidence of overlapping attention dominating events. Between 6th and 7th September, the Pope announced the Vatican’s churches will welcome families of refugees. This announcement followed a significant development in the ongoing European refugee crisis: around 6,500 refugees arrived in Vienna following Austria’s and Germany’s decision to waive asylum system rules. This suggests that an attention dominating news event in one topic can trigger events in other topics, especially where prominent public figures are involved.

6 Discussion

While the diversity measure we propose is relatively simple, it can be easily augmented to account for other factors. In the simplest form, every similarity value between a unique pair of articles within a time window carries an equal weight in the diversity calculation, implying that a strong similarity between two highly influential publishers is just as important as between two inconsequential publishers with a small audience. However, this weight could be tuned, either manually or automatically using external information (*e.g.* Alexa rankings). Accounting for social context [13] could also be achieved by augmenting the topic modeling stage of the process. Instead of using a classic tf-idf vector space model, alternative representations that capture more semantic similarity between documents can be used. We aim to explore extensions to this measure in future work.

The sequence of events in the European refugee crisis and papal visit case studies suggest that it may be possible to identify and track major developments with global impact by linking attention dominating moments across multiple topics, as well as across sources on different platforms. Social media communities both influence and are influenced by traditional news media [11]. Stories break both on Twitter and through traditional news publishers. Tracking or linking instances of diversity collapse to explain the direction of influence between the different media types is also a potential avenue for future work.

Acknowledgments: This publication has emanated from research conducted with the support of Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289.

[7] `https://en.wikipedia.org/wiki/Pope_Francis'_2015_visit_to_North_America`

References

[1] C. Boutsidis and E. Gallopoulos. Svd based initialization: A head start for nonnegative matrix factorization. *Pattern Recognition*, 41(4), 2008.

[2] A. E. Boydstun. *Making the news: Politics, the media, and agenda setting*. University of Chicago Press, 2013.

[3] I. Brigadir, D. Greene, and P. Cunningham. Adaptive representations for tracking breaking news on twitter. *CoRR*, abs/1403.2923, 2014.

[4] P. Esling and C. Agon. Time-series data mining. *ACM Computing Surveys (CSUR)*, 45(1):12, 2012.

[5] Y. Gandica, J. Carvalho, F. S. D. Aidos, R. Lambiotte, and T. Carletti. On the origin of burstiness in human behavior: The wikipedia edits case, 2016.

[6] W. Gao, P. Li, and K. Darwish. Joint topic modeling for event summarization across news and social media streams. In *Proc. 21st ACM international conference on Information and knowledge management*, pages 1173–1182. ACM, 2012.

[7] S. Ghosh, M. B. Zafar, P. Bhattacharya, N. Sharma, N. Ganguly, and K. Gummadi. On sampling the wisdom of crowds: Random vs. expert sampling of the twitter stream. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1739–1744. ACM, 2013.

[8] E.-M. P. Giulio Sabbati and S. Saliba. Asylum in the eu: Facts and figures. *European Parliamentary Research Service*, (PE 551.332), mar 2015.

[9] Y. Hu, A. John, F. Wang, and S. Kambhampati. Et-lda: Joint topic modeling for aligning events and their twitter feedback. In *AAAI Conference on Artificial Intelligence*, 2012.

[10] Y. Hu, K. Talamadupula, and S. Kambhampati. *Dude, srsly?: The surprisingly formal nature of Twitter's language*, pages 244–253. AAAI press, 2013.

[11] T. Hua, F. Chen, C.-T. Lu, and N. Ramakrishnan. Topical analysis of interactions between news and social media. *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, 2016.

[12] A. Jungherr and J. Pascal. Forecasting the pulse: how deviations from regular patterns in online data can identify offline phenomena. *Internet Research*, 23(5):589–607, 2013.

[13] J. Kalyanam, A. Mantrach, D. Saez-Trumper, H. Vahabi, and G. Lanckriet. Leveraging social context for modeling topic evolution. In *Proc. 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 517–526, 2015.

[14] H. Lamba, M. M. Malik, and J. Pfeffer. A tempest in a teacup? analyzing firestorms on twitter. In *Proc. International Conference on Advances in Social Networks Analysis and Mining*, pages 17–24, 2015.

[15] S. Liu, M. Yamada, N. Collier, and M. Sugiyama. Change-Point Detection in Time-Series Data by Relative Density-Ratio Estimation. *ArXiv e-prints*, Mar. 2012.

[16] M. Osborne, S. Petrovic, R. McCreadie, C. Macdonald, and I. Ounis. Bieber no more: First story detection using twitter and wikipedia. In *SIGIR Workshop on Time-aware Information Access*, 2012.

[17] D. OCallaghan, D. Greene, J. Carthy, and P. Cunningham. An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, 42(13):5645 – 5657, 2015.

[18] S. Petrovic, M. Osborne, R. McCreadie, C. Macdonald, I. Ounis, and L. Shrimpton. Can twitter replace newswire for breaking news? In *Proc. 7th International Conference on Weblogs and Social Media, ICWSM*, 2013.

[19] T. Steiner, S. van Hooland, and E. Summers. Mj no more: Using concurrent wikipedia edit spikes with social network plausibility checks for breaking news detection. In *Proc. 2nnd International Conference on World Wide Web*, pages 791–794, 2013.

[20] C. K. Vaca, A. Mantrach, A. Jaimes, and M. Saerens. A time-based collective factorization for topic discovery and monitoring in news. In *Proceedings of the 23rd international conference on World wide web*, pages 527–538. ACM, 2014.

[21] K. Zhai and J. Boyd-Graber. Online latent dirichlet allocation with infinite vocabulary. In *Proc. 30th International Conference on Machine Learning*, pages 561–569, 2013.

[22] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval*, pages 338–349. Springer, 2011.

Semi-Supervised Events Clustering in News Retrieval

Jack G. Conrad
Thomson Reuters
Corporate Research & Development
Saint Paul, Minnesota 55123 USA
jack.g.conrad@thomsonreuters.com

Michael Bender
Thomson Reuters
Thomson Reuters Global Resources
Baar, Zug 6340 Switzerland
michael.bender@thomsonreuters.com

Abstract

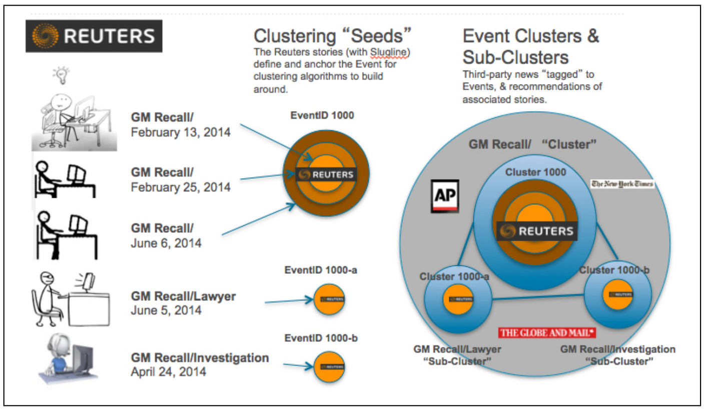
The presentation of news articles to meet research needs has traditionally been a document-centric process. Yet users often want to monitor developing news stories based on an event, rather than by examining an exhaustive list of retrieved documents. In this work, we illustrate a news retrieval system, *eventNews*, and an underlying algorithm which is event-centric. Through this system, news articles are clustered around a single news event or an event and its sub-events. The algorithm presented can leverage the creation of new Reuters stories and their compact labels as seed documents for the clustering process. The system is configured to generate top-level clusters for news events based on an editorially supplied topical label, known as a 'slugline,' and to generate sub-topic-focused clusters based on the algorithm. The system uses an agglomerative clustering algorithm to gather and structure documents into distinct result sets. Decisions on whether to merge related documents or clusters are made according to the similarity of evidence derived from two distinct sources, one, relying on a digital signature based on the unstructured text in the document, the other based on the presence of named entity tags that have been assigned to the document by a named entity tagger, in this case Thomson Reuters' *Calais* engine.

1 Introduction

1.1 Motivations

Thomson Reuters has been exploring alternative models for organizing and rendering articles found in its news repository. Whether the users are editors, financial analysts, lawyers or other professional researchers, a more effective means of examining a set of event-related news articles beyond that of a ranked list of documents was expressly sought. The presentation of news articles based on events aligns well with contemporary research use cases, such as those arising in the finance and risk sectors, where there is a salient need for more effectively organized news content through the lens of events. Other news organizations such as Google have experimented with news clustering, but in the absence of the concrete use cases of Thomson Reuters' professional users.

This project uses semi-supervised clustering capabilities in order to group news documents based upon shared news events. Germinal Reuters stories with editorially assigned labels (a.k.a. 'sluglines') are used as seed documents for event identification and organization. This task addresses the fundamental aim of the project.



Clustering "Seeds"
The Reuters stories (with Slugline) define and anchor the Event for clustering algorithms to build around.

Event Clusters & Sub-Clusters
Third-party news "tagged" to Events, & recommendations of associated stories.

Figure 1: News Events Clustering Process

In: M. Martinez, U. Kruschwitz, G. Kazai, D. Corney, F. Hopfgartner, R. Campos and D. Albakour (eds.): Proceedings of the NewsIR'16 Workshop at ECIR, Padua, Italy, 20-March-2016, published at http://ceur-ws.org

1.2 Objectives

The main objective of this project is to develop an event-centric news paradigm that solves the challenge of event validation and event story clustering at scale. This goal is in response to feedback received from consumers on news in their products. In addition to organizing news results around events rather than documents, another goal of this study is to provide a mechanism for clustering third-party (non-Reuters) news documents together with corresponding Reuters articles around common news events. This is aided by leveraging metadata tags that exist in Reuters news articles about the same topical event. Since these tags distinguish Reuters news documents from third-party content, it is possible to consider using them as the basis for grouping news articles together. The initial plan for this project was developed in conjunction with R&D's partner, the news asset owner and subject matter expert (SME), to use the initial or top-level story labels known as primary sluglines (e.g., VOLKSWAGEN-EMISSION-FRAUD/) as an organizing principle for top-level clusters, and an algorithmic means for creating lower-level clusters which can incorporate second tier story labels known as secondary sluglines (e.g., VOLKSWAGEN-EMISSION-FRAUD/COMPENSATION).

1.3 Workflow Illustration

In Figure 1, we see an example involving the "General Motors Recall" for faulty ignition switches. Through regular editorial practices, journalists write and tag event-related stories. The first story with the first "GM Recall" tag serves as the seed story for initiating the cluster. As Reuters writes and tags more stories about the GM Recall, the set of tags and text defining the GM Recall event expands. As it expands, so too does the algorithm's grasp of the event, helping it to better identify cluster candidates, in particular, within third-party news. Both the editorially generated slugline responsible for the birth of the cluster and the algorithmic identification and population of subsequent sub-clusters are depicted in the figure.

2 Previous Work

Previous work published on the topic of news events structuring has been largely academic in nature, for example, as in Borglund [6]. This thesis includes three contributions: a survey of known clustering methods, an evaluation of human versus human results when grouping news articles in an event-centric manner, and lastly an evaluation of an incremental clustering algorithm to see if it is possible to consider a reduced input size and still get a sufficient result.

In addition, there have been journal articles that have explored the computational complexity of the algorithms necessary to cluster real-time news articles [5]. But they have focused largely on the math behind the clustering rather than the use case and practitioners benefitting from it.

Some of the earliest work in this area was pursued under DARPA and NIST funding and resulted in reports written by various forums created to advance the state of the art in event detection [3, 1].

There have also been research group work and dissertations on the subject of topic detection and tracking resulting from the above research [12, 11]. Subsequent work has attempted to capture some of the structure of events and their dependencies in a news topic by creating a model of events, a.k.a. 'event threading' [10]. Yet more recently there have been actual forums under large umbrella organizations like ACL focusing on automatically computing news stories (and their titles) [2, 14].

There is also another field of research that addresses event extraction in the ACE tradition[1] that is relevant to the context of our current work, e.g., [9]. What is distinct about our present project, however, is the use of SME-defined seed stories and labels in a semi-supervised manner and the subsequent clustering stages at scale for real world news streams.

Worth noting is that one of the building blocks of the current work is represented by an initial form of 'local' clustering that involves the identification and grouping of exact and fuzzy duplicate documents [8]. This takes place in the stage immediately preceding the final, aggregated clustering step.

3 Data Resources

The news repository under examination in this effort is known as NewsRoom. It is a Thomson Reuters news aggregation platform. It consists of approximately 15-30 million documents per year from 12,000 independent news sources which consist of national and local newspapers, periodic journals, radio program transcriptions, etc. From 2012 to 2015, NewsRoom consisted of approximately 80 million news articles. These were the target of our investigation for this project (Table 1).[2]

In order to test our news workflow and the clustering algorithms that support it, we focus on chunks of data representing approximately three months of documents at a time.

Having investigated baseline news clusters in earlier research efforts (i.e., baseline algorithm, its granularity, speed and complexity) we have subsequently pursued improvements and efficiencies to help us approach

[1] http://www.itl.nist.gov/iad/mig/tests/ace/

[2] Thomson Reuters has long made comparably large news collections available for external research: http://trec.nist.gov/data/reuters/reuters.html

Table 1: NewsRoom Integrated Data Sources

Year	Sources	Document Count
2012	Reuters / Diverse	14.6M
2013	"	20.3M
2014	"	27.8M
2015	"	20.0M
Total	"	82.7M

our objectives more effectively.

4 Methods

Given our substantial data resources and our goal to build a flexible experimental retrieval environment, we have established three stages for processing and clustering a large set of news documents around news events (Figure 2). These stages include: (1) document extraction (Reuters and non-Reuters articles) from our news repository; (2) local clustering based on duplicate document detection of identical and fuzzy duplicates [7]; and (3) aggregate clustering performed over the result set from stage 2. We have determined empirically that the local clustering stage works highly effectively [8]. It is the aggregate clustering stage that has spawned ongoing research, evaluation and refinement. This stage consists of the application of hierarchical agglomerative clustering, where different types of cluster centroid representations were examined. Although we provide descriptions of each of the three processing stages below, it is the third of these stages that is the principal focus of our latest efforts and this research report.

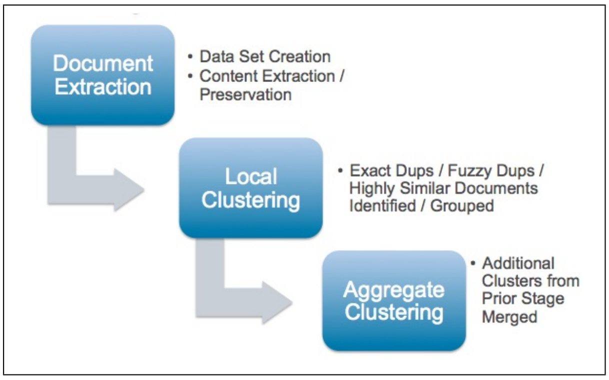

Document Extraction
• Data Set Creation
• Content Extraction / Preservation
Local Clustering
• Exact Dups / Fuzzy Dups / Highly Similar Documents Identified / Grouped
Aggregate Clustering
• Additional Clusters from Prior Stage Merged

Figure 2: News Events Clustering Functional Stages

4.1 Document Extraction Stage

The document extraction process can be customized to facilitate experimentation such as that undertaken for this study. NewsRoom represents a news repository of both Reuters and non-Reuters sources covering roughly 12,000 news sources. Given a date range, e.g., [20141001T0000000Z 20141231T235959Z], one can extract all of the 'recommendable' news documents in the repository, or some user-defined subset of them. Since the repository contains substantial numbers of Reuters and non-Reuters financial documents, for example, some stories are largely non-textual, e.g., containing tabular information only; very short, e.g., stubs for stories in progress stories; or meta-data snippets for topics that were not substantiated. These types of documents would be considered non-recommendable and thus are not retrieved for subsequent processing. In general, over half of the documents in the repository would be classified as recommendable for this use case. The NewsRoom environment comes with a recommendation classifier. Additional details beyond those provided above would be beyond the scope of our current focus.

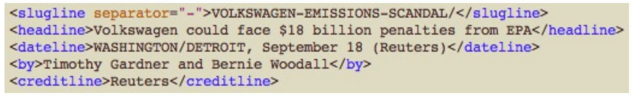
The extraction process results in all recommendable documents being loaded from the repository to an Apache Derby JDBC relational database. The tabular data structures that store the documents and subsequent clusters contain basic information such as doc_id, dataset_name, doc_date, title, article_source, source_url (if applicable), body, body_length, together with tens of additional features that can be used to discriminate and used by various classifiers, e.g., primary news code, short sentence count, ticker count, quantity of numbers, quantity all-caps, quantity of press releases, etc. These additional features are available for subsequent downstream processing such as classification, routing or clustering.

4.2 Local Clustering Stage

The next process, local clustering, is designed to rapidly and efficiently identify initial clusters based on documents that satisfy criteria for identical or fuzzy duplicates. Documents are compared using two types of digital signatures that harness the most discriminating terms, one, smaller and more compact leveraging O(10) terms, is used to identify identical duplicates; another, more expansive, leveraging O(100) terms, is used to identify fuzzy duplicates. The process being executed uses techniques reported on in [8]. For this application, a rolling window of n days is used, where ($n < 10$). Documents falling within this window are compared. Heuristics relying on features such as doc_length, are also invoked to reduce the number of comparisons required. For example, when a document exceeds the length of another by 20% or more, though they may satisfy a containment relationship, according to our definition, they would not be considered 'duplicates.'

4.3 Aggregate Clustering Stage

During the third, aggregate clustering stage, the clusters are initiated via seminal Reuters articles containing slugline tags. These tags are distinct from headlines, as shown in Figure 3. The articles with sluglines may be singletons or they may exist in one of the local clusters formed in preceding stage. Both of these ‘objects’ qualify to serve as a cluster ‘seed.’



```
<slugline separator="-">VOLKSWAGEN-EMISSIONS-SCANDAL/</slugline>
<headline>Volkswagen could face $18 billion penalties from EPA</headline>
<dateline>WASHINGTON/DETROIT, September 18 (Reuters)</dateline>
<by>Timothy Gardner and Bernie Woodall</by>
<creditline>Reuters</creditline>
```

Figure 3: Reuters Article - Slugline Illustration

Two main challenges confronted when implementing this hierarchical, agglomerative clustering stage were, first, finding the best set of features and metrics to decide whether a pair of singletons or local clusters justify merging into larger clusters while still remaining sufficiently cohesive, and, second, identifying the optimal sequence for comparing these clusters when considering merging (Figure 4).

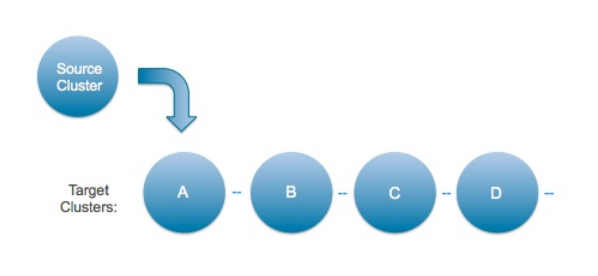


Figure 4: Approaches to Source to Target Merging

Based upon observations made by subject matter experts who created exemplar news clusters to support the project, we determined that there were two, often independent, means by which documents could be identified as belonging to the same news event. One involves the unstructured text of an article; the other involves the structured text, in our case, documents that have been tagged by the Calais named-entity tagging engine [13, 4]. Given that articles involving news events can be found to be similar based on either of these two feature spaces, our approach to aggregate (stage 3) clustering is robust: a decision to merge two of these documents or local clusters can be based on the similarity between the unstructured text of two objects, the tagged named entities that have been identified by Calais (listed below), or both.

- People – person name entities
- Reuters Instrument Codes (RICs) – for companies
- Reuters Classification System (RCS) – for topics & industries
- Topics – domain independent topical phrases
- Smart Terms – topical taxonomy terms

Operationally, the hybrid feature set described above is used to decide whether or not to merge two clusters. It consists of two data structures, both repesented in vector form. The first is a term-based vector. It is used to determine the degree of overlap between two cluster centroids, constituted by two central ‘documents’ (e.g., longest, most recent, true centroid, etc.). The second is a tag-based vector, representing a set of Calais tags present in the cluster’s documents. The similarity measures used in each of these cases is thresholded, with the threshold determined empirically. In the case of the term vectors for the unstructured text, the thresholds are set high, although not as high as those for duplication detection used in stage 2. In the case of the set of Calais tags for the structured text, a weighted sum is used, whereby various combinations of named entities can be assembled to satisfy the threshold for merging.

Table 2: Experimental Processing (4 QTR 2014)

Stage	Name	Type	Count
1.	Document Extract	documents	3.63M
2.	Local Clustering	clusters	2.10M
3.	Agglom. Clustering	clusters	1.67M

5 Evaluation

Given the objectives of this study with respect to retrieval performance and organizational structure, evaluation is an essential piece of the validation process. After having conducted a number of trials to establish various thresholds (document or cluster similarity, named entity similarity, etc.), we conducted a trial which focused on a number of news events chosen by subject matter experts (SMEs) from the final quarter of 2014. We focused on the set of high-level news events shown below.

1. Halliburton Buying Baker Hughes (Nov. 13, 15)
2. Defense Secretary Hagel Resigns (Nov. 24, 25)
3. Air Asia Crash (Dec. 28, 31)
4. Pope Urges Tolerance in Turkey (Nov. 28)
5. Lufthansa Braces for Next Strike (Dec. 3)
6. Iran Rouhani Says Will Try to Clinch Nuclear Deal in Talks (Dec. 15)
7. Alstom Nearing $700M Bribery Settlement (Dec. 16)

For each of the events identified, result sets were created and stored in worksheets (Table 2 presents dataset details). The result sets consisted of numerous clusters on the subject of the event (often involving named entities such as Halliburton, Hagel, the Pope, Rouhani, Alstom, etc.), some of which are on the topic of the news event, some of which address the entity in other contexts. For those that *were* on the subject of the event, the clusters represent sub-topical (second-level) clusters (see VW example in Section 1.2). Regarding the result worksheets, in addition to doc ids, they included local cluster and batch cluster ids, date and time stamp, document title, document length and URL link to the complete news article (if available). The worksheets were presented to two evaluators, both subject matter experts from the news domain.[3]

Two metrics were used to evaluate these experiments. First, the assessors scored each cluster for coherence and accuracy, making sure that all of the documents that belong to a specific cluster were present, and that all of the documents that didn't belong were not present. The cluster database was queried broadly, e.g., 'Defense Secretary Hagel', in order to permit the assessors to have access to clusters both about and not about the event in question, again, in order to inspect those documents that belong in the relevant clusters and those that do not. For this task, they used a five-point Likert scale, A (very good) thru F (very weak), codified as 5-to-1.[4] Secondly, the assessors determined a 'cluster edit distance' for each cluster solution, indicating which sub-clusters they would merge and which they would split, if any, to achieve an optimal solution. Each merge or split step would be the cluster equivalent of an 'edit' in the standard character-based edit distance measure. The results of this assessment task are presented in the Table 3.

In general, we see that with few exceptions, the majority of clusters returned for our queries were about the underlying event(s) (Table 3, column 4). In addition, the coherence/accuracy scores for the clusters reviewed were in the 4.0 or 'B' range, some higher, some lower. When the same entities, but out-of-event clusters are included (column 3), their scores are slighty higher, still in the 4.0 or 'B' range.[5] In terms of the cluster edit distances measured, for the seven news events represented in the table, the mean number of 'splits' required for each cluster set was λ=1.15 (σ=1.2) while the mean number of merges was λ=4.7 (σ=4.3).

Clearly the larger numbers appearing in the context of merges have been influenced significantly by a couple of the outliers found in the list of events, i.e., nos. 2 and 7. In the case of the latter, there was greater variety in the news sources and articles reporting on the statements coming from the Iranian leader, and as a result, the algorithm may not have captured the overarching similarity among the documents. In addition, there was a greater variety of persons mentioned in these articles who were responding to President Rouhani.

Regarding the queuing strategy and its impact on agglomerative clustering and merging (Figure 4), we conducted a series of experiments that involved different strategies, including least-recently-used and most-recently-used. Other strategies tended to have a significant impact on computational complexity insofar as it was necessary to perform real-time tracking of dynamic cluster characteristics. Although the spectrum of considerations involved in those experiments may be beyond the scope of the current reporting space, we found that the most-recently-used was as effective a queuing strategy as the majority of others investigated.

There is clearly room for improved performance and additional evaluation. One way of addressing some of the disparities revealed above is by tuning the joint thresholds for document signature and named entities tagged. Alternatively, one could have the thresholds learned and optimized depending on features associated with the documents (e.g., range of idfs in the signatures, number and type of entities in the document). Moreover, one could use a variable weighted sum of the similarity scores, depending on the contribution of the named entities and distinguishing terms present in the articles being compared.

6 Conclusions

The news events clustering efforts summarized in this report and depicted in Figure 1 represent a combination of semi-supervised clustering techniques and human-generated, labeled data. They aim to deliver an effective solution by leveraging Reuters' labels and validating the scope of events at scale. The ultimate goal of the study is to determine to what extent combined human-computer resources can produce event-based clusters that are considerably more useful – i.e., more effective – than exhaustive lists of unstructured documents. In addition, third-party content can be gathered and organized around existing clustered content based upon Reuters' own editorially labeled and classified news events. The variety of challenges confronted – using Reuters' metadata, getting the granularity right, and scaling the solution – all depend on the right mix within this integration. By tracking the steps outlined above, we anticipate having a more robust working model available for evaluation in the near

[3] The first SME assessed the quality of both types of clusters, those about the event and those not; the second SME assessed the quality of the event clusters only.

[4] The five grades used in the American educational system are A-B-C-D-F, which range from exceptional (A) to failure (F). E is not used.

[5] Although in aggregate, the mean of the grades assigned the clusters by the two SMEs were comparable, when we calculated the weighted Kappa score for inter-reviewer agreement, we found that they were not as uniform, as the scores generally fell into the bottom quartile. The reviewers assigned identical grades in only about a third of the cases. In the majority of the other cases, they were one and sometimes two grades apart.

Table 3: Graded Assessments of News Events Clusters

No.	**Event Title**	**No. Clusters**	**No. Clusters on Event**	**Mean Avg Score for All Clusters**	**Mean Avg Score for Event Clusters**	
					SME #1	SME #2
1.	Halliburton Buying Baker Hughes	6	5	4.33	4.20	4.00
2.	Defense Secretary Hagel Resigns	24	17	4.02	3.94	2.94
3.	Air Asia Crash	14	7	3.93	3.64	3.50
4.	Pope Urges Tolerance in Turkey	7	6	4.29	4.17	4.33
5.	Lufthansa Braces for Next Strike	5	2	4.00	3.00	4.50
6.	Iran Rouhani Tries to Secure Nuclear Deal	59	47	3.99	3.86	3.93
7.	Alstom Nearing $700M Bribery Settlement	5	3	3.80	3.50	4.50
T.	Total			Avg = 4.05	Avg = 3.73	Avg = 3.95

future. Anticipated amendments or extensions of the model are addressed below.

7 Future Work

In future work, we will extend our evaluations by comparing our results with exemplar clusters identified by our SMEs, both in terms of granularity and in terms of completeness, at the top, topical cluster level and lower, sub-topical level of resulting clusters. This form of assessment addresses overall cluster precision. We will also need to conduct tests that approach evaluating recall, i.e., of all the possible news events in the data set or sample, how many do we capture and represent at top and lower levels of the shallow hierarchy?

8 Acknowledgments

The authors thank Sarah Edmonds at TRGR for her diligent work assessing result sets. We are also grateful to Brian Romer with Reuters Data Innovation Lab for his innovative work on the UI and demo (to be shown at the workshop).

References

[1] *Topic Detection and Tracking Workshops*, Washington, D.C., 2004. NIST.

[2] *First Workshop on Computing News Storylines (CNewS 2015)*, Beijing, PRC, July 2015. ACL.

[3] James Allan, Jaime Carbonell, George Doddingtom, Jonathan Yamron, and Yiming Yang. Topic detection and tracking pilot study final report. In *DARPA Broadcast News Transcription & Understanding Workshop*, Feb. 1998.

[4] Samet Atdag and Vincent Labatut. A comparison of named entity recognition tools applied to biographical texts. In *2nd International Conference on Systems and Computer Science (ICSCS13)*, pages 228–233. IEEE, Aug. 2013.

[5] Joel Azzopardi and Christopher Staff. Incremental clustering of news reports. *Algorithms*, 5:364–378, 2012.

[6] Jon Borglund. Event-centric clustering of news articles. Masters thesis, University of Uppsala, Sweden, Oct. 2013.

[7] Jack G. Conrad, Joanne C. Claussen, and Jie Lin. Information retrieval systems with duplicate document detection and presentation functions. U.S. Patent #7,809,695, Oct. 2010.

[8] Jack G. Conrad, Xi S. Guo, and Cindy P. Schriber. Online duplicate document detection: Signature reliability in a dynamic retrieval environment. In *Proceedings of the 12th Conference on Information and Knowledge Management (CIKM03)*, pages 243–252. ACM Press, Nov. 2003.

[9] Qi Li, Heng Ji, and Liang Huang. Joint event extraction via structured prediction with global features. In *Proceedings of the 51st Annual Meeting of the ACL*, pages 73–82. Association for Computational Linguistics, Aug. 2013.

[10] Ramesh Nallipati, Ao Feng, Fuchun Peng, and James Allan. Event threading within news topics. In *Proceedings of the 13th Conference on Information and Knowledge Management (CIKM04)*, pages 446–453. ACM Press, Nov. 2004.

[11] Ron Papka. *On-Line New Event Detection, Clustering, and Tracking.* Ph.d. thesis, University of Massachusetts - Amherst, Sept. 1999.

[12] Jakub Piskorski, Hristo Tanev, Martin Atkinson, and Erik van der Gout. Cluster-centric approach to news event extraction. In *2008 Conference on New Trends in Multimedia and Network Information Systems*, pages 276–290, 2008.

[13] Thomson Reuters. Open Calais Named^{TM} Entity Tagging Engine. `http://www.opencalais.com`, 2016.

[14] Piek Vossen, Tommaso Caselli, and Yiota Kontzopoulou. Storylines for structuring massive streams of news. In *Proceedings of the First Workshop on Comparing News Storylines*, pages 40–49. ACL and Asian Federation of NLP, July 2015.

Cross-lingual Trends Detection for Named Entities in News Texts with Dynamic Neural Embedding Models

Andrey Kutuzov
University of Oslo
Postboks 1080 Blindern 0316, Oslo, Norway
andreku@ifi.uio.no
Elizaveta Kuzmenko
National Research University Higher School of Economics
Moscow, Russia
eakuzmenko_2@edu.hse.ru

Abstract

This paper presents an approach to detect real-world events as manifested in news texts. We use vector space models, particularly neural embeddings (prediction-based distributional models). The models are trained on a large 'reference' corpus and then successively updated with new textual data from daily news. For given words or multi-word entities, calculating difference between their vector representations in two or more models allows to find out association shifts that happen to these words over time. The hypothesis is tested on country names, using news corpora for English and Russian language. We show that this approach successfully extracts meaningful temporal trends for named entities regardless of a language.

1 Introduction

We propose an approach to track changes happening to real-world entities (in our case, countries) with the help of constantly updated distributional semantic models. We show how one can train such models on new textual data arriving daily and draw conclusions about events based on changes in word vectors induced by new contexts. In other words, subtle *semantic shifts* which the words undergo over time, influenced by real-world events, are detected by the presented method.

Detecting semantic shifts can be of use in a variety of linguistic applications. First, this method can be of help in the problem of automatically monitoring events through the stream of texts [AGK01]. Detected semantic shifts can potentially be used as additional features in the algorithms aimed at extracting the course of events. Without unsupervised approaches, it is impossible to process all the continuously generated data. This is the primary motivation factor for our research. Second, the developed approach can be used to study language shift and compare temporal corpora slices. This language area is traditionally studied by linguists, who put a lot of efforts into describing semantic shifts with the help of dictionaries, corpora and sociolinguistic research. At the same time, it is impossible to grasp all the language vocabulary and describe every lexical shift manually. Distributional semantic models facilitate this task.

The approaches to events detection and modeling of language shifts have a lot in common. First techniques employed various frequency metrics [JS09] and shallow semantic modeling [KNR15], [HBB10]. With the emergence of distributive semantic models detection of semantic shifts acquired new potential, as it was shown that word embeddings significantly improve the performance of algorithms [KARPS15].

The rest of the paper is organized as follows. In Section 2 we introduce the basics of prediction-based vector models of semantics. Section 3 describes the

In: M. Martinez, U. Kruschwitz, G. Kazai, D. Corney, F. Hopfgartner, R. Campos and D. Albakour (eds.): Proceedings of the NewsIR'16 Workshop at ECIR, Padua, Italy, 20-March-2016, published at http://ceur-ws.org

principles of comparing such models, trained on pieces of text which follow each other in time. Specifics of our datasets are covered in Section 4, followed by the description of experimental setting in Section 5. Section 6 evaluates the results and in Section 7 we conclude.

2 Distributed Semantic Models

Vector space models (VSMs) are well established in the field of computational linguistics and have been studied for decades (see [TP^{+}10], [Reh11]). Essentially, a model is a set of words and corresponding vectors, which are produced from typical contexts for a given word. The most widespread type of contexts is other words co-occurring with a given one, which means that the set of all possible contexts generally equals the size of the vocabulary of the corpus. The dimensionality of the resulting *count model* can be reduced with well-known techniques like Principal Components Analysis (PCA) or Singular Value Decomposition (SVD). But in turn, this effectively forbids online training (continuously updating the model with new data), because after each update one has to perform computationally expensive dimensionality reduction over the whole co-occurrence matrix.

To overcome this, we employ a type of VSMs called *prediction-based models*: particularly, Continuous Bag-of-Words (CBOW) algorithm ([BDV03], [MSC^{+}13])[1]. Predictive models rather approximate co-occurrence data, instead of counting it directly, and show a promising set of properties. Using them, one directly learns dense lexical vectors (*embeddings*). Vectors are initialized randomly and then, as we move through the training corpus with a sliding window of a pre-defined width, gradually converge to values maximizing the likelihood of correctly predicting lexical neighbors. Such models as a rule use artificial neural networks to train; this is why they are sometimes called *neural models*.

For our task, it is important that predictive models can be updated with new co-occurrence data in a quite straightforward way. As already said, this is usually not the case with count models which demand computationally expensive calculations each time a new text is added.

3 Introducing Temporal Dimension to Vector Models

Detecting semantic shifts which words undergo over time demands the ability to somehow compare reference ('baseline') and updated models, representing later periods of time.

The idea of employing changes in distributional semantic models to track semantic shifts is not in itself new. [KCH^{+}14] proposed to detect language change with chronologically trained models. However, they used rather simplified measure of 'distance' between word vectors at different time slices, namely, raw cosine distance. We employ more sophisticated methods as described further. [POL10] developed an approach to the First Story Detection in Twitter posts. Their research is similar to ours in that it deals with streaming data. The authors explore the space of documents and compare new tweets to the existing ones. However, the algorithm is developed specifically for short texts like tweets, which differ radically from news pieces analyzed in the presented paper.

Updating a neural model with new texts (in addition to the base training corpus used for initial training) is technically straightforward. After that, we have two models M_1 and M_n, where the former is the 'baseline' reference model, and the latter is the updated one (or a sequence of n updated models, each corresponding to the next time period), probably bringing new semantic shifts. This dynamic model in a way tries to imitate human brain learning new things, gradually 'updating' its state with new input data every day.

What are the possible ways to extract these changes? Suppose there is a set S of named entities (organizations, locations or persons we are interested in). Initially in the model M_1, each element of S can be thought of as possessing a number of topical '*associates*' or '*nearest neighbors*': words with their respective vectors closest to this element vector, ranked by their closeness or similarity. The exact number of nearest neighbors we consider in the simplest case is defined arbitrarily (for example, 10 nearest words). As we update the model with new data, co-occurrence counts for the elements of S are gradually growing (the model sees them in new contexts). It means than in each successive model M_n learned vectors for elements of S can be different.

If contexts for these words remain pretty much the same throughout the training data, the list of associates (nearest neighbors) in M_n will also remain intact. However, if a word acquires new typical contexts or loses some previous ones, its neural embedding will change: a *semantic shift* happens. Accordingly, we will see a new list of associates. For example, the vector representation for the word *president* may change so that its nearest neighbor is the vector for the name of the actual president of a country, instead of the previous one.

In this way, lists of nearest neighbors can be compared across models trained on different corpora or across one and the same model after an incremental update (as in the presented research). Substantial

[1] The well-known *word2vec* tool also implements SkipGram, which is another predictive algorithm. However, it is more computationally expensive, and we leave its usage for future work.

changes or *bursts* in such lists for the named entities we are interested in may signal that these entities have undergone or are undergoing semantic shifts, which in turn reflects real-world events. We dub this approach '*dynamic neural embedding models*'.

Sets of neighbors in different models can be compared in many ways. Approaches to this range from simple Jaccard index [Jac01] to complex graph-based algorithms. We test two methods:

1. *Kendall's τ coefficient* [Ken48], which measures similarity of item rankings in two sets. Intuitively, it is important to pay attention not only to raw appearance of some words in the nearest neighbors set, but also to their rankings in it.

2. *Relative Neighborhood Tree* (RNT), introduced by [CGS15]. It essentially produces a tree graph with the target word as its root, nearest neighbors as vertexes and similarities between them as weighted edges. We then select the immediate neighbors of the target word in this tree and rank them according to their cosine similarity to the target word. These rankings are then compared across models using the same *Kendall's τ*.

The reason behind the second method is that it theoretically allows a deeper analysis of nearest neighbors' sets structure. Obviously, the neighbors participate in similarity relations not only with the target word but also between themselves. These relations convey meaning as well, making it possible to find the most 'important' neighbors. Graph-based methods to analyze relations between words in distributional models were also used in [KWHdR15]; note, however, that the problem they deal with is inverse to ours – they attempt to trace changes in surface words for a stable set of concepts, while we attempt to trace semantic shifts (changes in underlying concepts for a stable set of words).

We hoped that this graph-supported 'pre-selection' would allow Kendall's τ to improve the performance of the model. However, these expectations failed and simple ranking turned out to be more efficient than graph-based methods; see Section 6.

4 Data Description

We test our approach on lemmatized corpora of English and Russian news texts. The English corpus consists of *The Signal Media Dataset*[2], which contains 265,512 blog articles and 734,488 news articles from September 2015. The size of the corpus (after lemmatizing and removing stop words) is 222,928,287 words. We employ Stanford POS tagger [TKMS03] to extract lemmas and to assign each lemma a part-of-speech tag.

In order to test whether extracted semantic shifts are consistent across languages, we use a corpus of news articles in Russian published in September 2015 (unfortunately, not available publicly due to copyright restrictions). It contains about 500,000 texts extracted from about 1000 Russian-language news sites. The size of the corpus (after lemmatizing and removing stop-words) is 59,167,835 words. We employ Mystem [Seg03], a state-of-the art tagger for Russian to produce lemmas and part-of-speech tags.

5 Experimental setting

News texts from September 2015 do not seem to be a good training set alone. This is because such a corpus is inevitably limited in language coverage, lacking relations to events that happened earlier. Therefore, we first train a 'reference' or 'baseline' model which aims to mimic some background knowledge, which is then exposed to daily updates. For English, we used British National Corpus[3] (about 50 million words) to train this reference model, while for Russian it was the corpus of news articles published in the months preceding September 2015, precisely June, July and August (taken from the same source as the September articles). This corpus contains about 250 million words.

We acknowledge it is not quite correct to employ different types of corpora for 'reference' models in English and Russian. However, in a way, we compensate the quality and balance of BNC with the larger size of the reference corpus in Russian. In the future we plan to eliminate this inconsistency by using an analogous set of English news published in summer months or by employing Wikipedia dumps as reference corpora for both languages.

Both corpora were merged with same-language texts released in the first half of September 2015 (before 14th of September), in order to seed baseline models with some initial 'knowledge' of events and entities belonging to this month. Then, Continuous Bag-of-Words models were trained for both corpora, using negative sampling with 10 samples, vector size 300, symmetric window size 5 and 5 iterations. Words with frequency less than 10 were ignored during training.

After that, we successively updated these models with texts released in the following September time periods: 14th–15th, 16th–17th, 18th–20th, 21th–22th, 23th–24th, 25th–27th, and 28th–30th. Granularity of 2 or 3 days was chosen in order to enlarge the amount of data fed to models: for example, some one-day Russian corpora corresponding to weekends contained only

[2] `http://research.signalmedia.co/newsir16/signal-dataset.html`

[3] `http://www.natcorp.ox.ac.uk/`

several thousand words. For this reason, we additionally tried to include week-ends in 3-days periods, to make news stream more evenly distributed. As a result, average time period size in tokens was 18,774,000 for English data and 5,332,000 for Russian data.

We once again emphasize that our baseline models were not re-trained from scratch with new texts added from new corpora. Instead, we continued training the same model, gradually updating word vectors with new contexts. All interim states were saved as separate models, and in the end we had 8 successive models for each language.

We extracted English and Russian countries names from Wikipedia list of all world countries[4] and manually checked and normalized it, bringing all name variants to one lexeme. Then we filtered out the entities with frequency less than 30 per million words in either of our two reference corpora (English and Russian), producing a set CS of 36 frequent country names[5].

Finally, for each of the successive models, we found nearest neighbor sets for each entity in CS and compared them to the sets from the model state at the previous time period. Kendall's τ and Relative Neighborhood Tree (RNT) were used to compute similarity coefficients for each country within the given pair of models. This provided us with two lists of countries (for each language) ranked by their similarity to the same country in the 'previous' model. Supposedly, countries in which some major events happened during the last days have to position low in these lists, because their associations in news texts drifted towards the recent event or an opinion burst.

Let's illustrate how news texts and changes in the models reflect the real-life events by comparing 10 nearest associates for *Chile* in the English and Russian corpora. On the 16th of September 2015 there was an earthquake in Chile, and we can detect its 'echo' in the changes between our models for 14th–15th and 16th–17th of September (see Table 1).

Before the 16th of September, associates for *Chile* in both models were mostly the neighboring countries. However, after the earthquake things have completely changed: there was a strong bias towards this topic in news and blogs, and this is reflected in vectors for the word. 60% of English and 20% of Russian associates are now related to the event.

Kendall's τ coefficient between these two neighbors lists is as low as 0 (neighbors are completely replaced) for English and 0.56 for Russian. Average Kendall's τ for CS is 0.56 in the English models for the two days in question, with standard deviation 0.12. Thus, in the case of English, the change to the neighbors' set can be considered a significant burst, well above simple chance. In the case of Russian, Kendall's τ lies only 1 point below the average value of 0.57. It is obvious that Russian mass media paid less attention to the earthquake (they are more concerned with Michelle Bachelet, Chile's president), but the event is still reflected in the nearest neighbors set.

Table 1: Change in *Chile*'s neighbor set

14th–15th September		16th–17th September	
English	Russian	English	Russian
peru	бачелет	*quake*	аргентина
bolivia	аргентина	*earthquake*	бачелет (bachelet)
colombia	коста-рика	santiago	никарагуа
argentina	перчик	chilean	мексика
honduras	никарагуа	*tremor*	бельгия
brazil	швейцария	*tsunami*	исландия
ecuador	бельгия	*aftershock*	тунис
nicaragua	исландия	chileans	*магнитуда* (magnitude)
paraguay	аргентин	*temblor*	*землетрясение* (earthquake)
enchiladas	гватемала	kyushu	коста-рика

The next section describes how we employed cross-linguality of the data to evaluate the presented approach.

6 Cross-Lingual Evaluation of Events Detection

There is no 'golden standard' or ground truth which would allow to evaluate precision and recall of our events and associations extraction, and to tune hyperparameters of the algorithms. However, there is a way to indirectly estimate their performance in a kind of intrinsic evaluation.

We hypothesize that the better is an algorithm of detecting semantic shifts, the closer should be its results on model sequences trained on different language corpora. Obviously, national media focus on different topics, but this mostly concerns the domestic news. As for the world news, the worst scenario could be that a news story is not covered in national media of a particular country. However, such scenarios should be rare. In other cases, the perspective on a story can differ, but the 'burst' should remain the same[6].

Thus, English and Russian countries lists ranked by their 'burstiness' can be compared using Spearman's ρ

[4] `https://en.wikipedia.org/wiki/List_of_sovereign_states`

[5] Low-frequency country names bring in noise, because their vectors are susceptible to wild fluctuations when exposed to even a small amount of new contexts.

[6] Analyzing the degree to which the vision of events is different in national media is beyond the scope of the present research.

Table 2: 5 countries with most changed neighbors' sets (of total 36) between September 18–20 and 21–22

Rank	English	Russian (translated)
1	Italy	**Japan**
2	**Georgia**	Brazil
3	Malaysia	**China**
4	**Japan**	Spain
5	**China**	**Georgia**

[Spe04] for each time period. As there are 7 shifts from one time period to another, we use median of ρ values for these 7 cases as a tentative measure of algorithm's performance. The Table 2 gives an example of such country rankings for the changes between 18–20 and 21–22 of September. One can see that the top lists are highly similar, with 3 of 5 countries appearing in both (actual Sperman's ρ for the total lists of 36 countries between these periods is 0.5).

Overall results of applying this approach to the whole dataset using two our algorithms (with different sizes of nearest neighbors' sets to consider) are presented in the Table 3. We also applied it to a simple baseline method, where nearest neighbors are words which most frequently occurred in the window of 5 tokens to the right and to the left of the target entity in the given corpus.

Table 3: Cross-lingual evaluation

Algorithm	Neighbors' set size	Median Spearman's ρ
Raw co-occurrences baseline	5 10 100	0.26 ($p = 0.12$) 0.15 0.06
CBOW and Kendall's τ	5 10 100	0.25 0.25 **0.28** ($p = 0.09$)
CBOW and Relative Neighborhood Tree	5 10 100	0.20 0.16 0.14

Kendall's τ consistently renders better results without additional selection of 'important' associates by a relative neighborhood tree (additionally, it is much faster). This once again raises questions about whether vector models can be efficiently processed with graph representations. Kendall's τ also outperforms the baseline approach: the margin is as small as two points, but it is supported by higher significance ($p < 0.1$).

Note that qualitative analysis of the baseline results shows that they are mostly inappropriate for any practical task. For the time period which is described in the Table 1, the baseline approach almost does not reveal any differences between neighbors sets: average Kendall's τ is 0.92 for English and 0.99 for Russian. Thus, if in the case of English the earthquake event is at least detected (we observe the emergence of 4 new related neighbors), in the case of Russian the neighbor set remained strictly the same. It seems that the raw co-occurrences approach suffers from overestimating the influence of the reference corpora, which are much larger than the daily updates. Dynamic neural embedding models overcome this problem.

Interestingly, the wider sets of neighbors taken into account results in better performance only for CBOW with Kendall's τ. For the baseline and for CBOW with RNT, increasing the size of processed neighbor sets actually results in poorer performance. The reason for this behavior in RNT can be that the algorithm begins to 'roam' in the graph attracting more far-away associates as immediate tree neighbors to the target word. In the baseline method it simply leads to much language-dependent noise, which semantically aware models filter out at the training stage.

7 Conclusions

We presented a method of detecting semantic shifts for countries in news texts with the help of dynamic neural embedding models. We explored the difference between entities' vector representations in the models from different temporal stages and discovered association shifts that happen to these words over time. This can be employed to trace trends and events in streaming news texts using a completely unsupervised approach.

We showed that distributional semantic models are rather efficient when detecting associations shifts and are in most cases language-independent. In our test sets, there is a statistically significant correlation between lists of 'semantically shifted' countries in English and Russian sequences of models for the same time period.

However, there is still room for improvement. First of all, some ways to evaluate semantic shifts extraction have to be developed (including creation of ground truth datasets). Additionally, we plan to test other ways of comparing neighbor sets and tune algorithms' hyperparameters. It would be also useful to improve the quality of corpora (e.g. eliminate more noise and stop words). Finally, we plan to experiment with using different algorithms or parameter sets for different languages: preliminary tests show promising results.

References

[AGK01] James Allan, Rahul Gupta, and Vikas Khandelwal. Temporal summaries of new topics. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 10–18, New York, USA, 2001.

[BDV03] Yoshua Bengio, Rejean Ducharme, and Pascal Vincent. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, 2003.

[CGS15] Amaru Cuba Gyllensten and Magnus Sahlgren. Navigating the semantic horizon using relative neighborhood graphs. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2451–2460, Lisbon, Portugal, September 2015.

[HBB10] Matthew Hoffman, Francis R. Bach, and David M. Blei. Online learning for latent dirichlet allocation. In *Neural Information Processing Systems 23*, pages 856–864, Vancouver, Canada, 2010.

[Jac01] Paul Jaccard. *Distribution de la Flore Alpine: dans le Bassin des dranses et dans quelques régions voisines.* Rouge, 1901.

[JS09] David Jurgens and Keith Stevens. Event detection in blogs using temporal random indexing. In *Proceedings of the Workshop on Events in Emerging Text Types*, pages 9–16, Borovets, Bulgaria, 2009.

[KARPS15] Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635, Florence, Italy, 2015.

[KCH+14] Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. Temporal analysis of language through neural language models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, page 61, Baltimore, USA, 2014.

[Ken48] Maurice George Kendall. *Rank correlation methods.* Griffin, 1948.

[KNR15] Manika Kar, Sérgio Nunes, and Cristina Ribeiro. Summarization of changes in dynamic text collections using Latent Dirichlet Allocation model. *Information Processing & Management*, 51(6):809–833, 2015.

[KWHdR15] Tom Kenter, Melvin Wevers, Pim Huijnen, and Maarten de Rijke. Ad hoc monitoring of vocabulary shifts over time. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 1191–1200, New York, NY, USA, 2015. ACM.

[MSC+13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems 26*, pages 3111–3119, 2013.

[POL10] Saša Petrović, Miles Osborne, and Victor Lavrenko. Streaming first story detection with application to twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 181–189. Association for Computational Linguistics, 2010.

[Reh11] Radim Rehurek. *Scalability of semantic analysis in natural language processing.* PhD thesis, Masaryk University, 2011.

[Seg03] Ilya Segalovich. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In *MLMTA*, pages 273–280. Citeseer, 2003.

[Spe04] Charles Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101, 1904.

[TKMS03] Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 NAACL-HLT Conference-Volume 1*, pages 173–180. Association for Computational Linguistics, 2003.

[TP+10] Peter Turney, Patrick Pantel, et al. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188, 2010.

Using news articles for real-time cross-lingual event detection and filtering

Gregor Leban
Jožef Stefan Institute
Ljubljana, Slovenia
gregor.leban@ijs.si

Blaž Fortuna
Jožef Stefan Institute
Ljubljana, Slovenia
blaz.fortuna@ijs.si

Marko Grobelnik
Jožef Stefan Institute
Ljubljana, Slovenia
marko.grobelnik@ijs.si

Abstract

The written medium through which we commonly learn about relevant news are news articles. Since there is an abundance of news articles that are written daily, the readers have a common problem of discovering the content of interest and still not be overwhelmed with the amount of it. In this paper we present a system called Event Registry which is able to group articles about an event across languages and extract from the articles core event information in a structured form. In this way, the amount of content that the reader has to check is significantly reduced while additionally providing the reader with a global coverage of each event. Since all event information is structured this also provides extensive and fine-grained options for information searching and filtering that are not available with current news aggregators.

1 Introduction

News publishers daily produce large numbers of news articles. Most of these articles describe happenings that are currently occurring in the world, such as natural disasters, meetings of important politicians, crime, business and sport events. Not all reported information is equally important – some events get higher media coverage, while other events get reported only by a small set of publishers.

In order to learn about current events, people nowadays usually either go to their favorite news publisher's web site and browse through the frontpage articles or they use of some type of aggregator, such as Flipboard or Bloomberg Terminal. Neither of the two approaches are optimal. By browsing a publisher's web site you typically learn about a small subset of current events (usually constrained to the geographic location of the news source) that are not necessarily unbiased and objective but instead implicitly promote political, social and religious views of the publisher/author. Using a news aggregator on the other hand can provide the readers with a coverage of the same events from multiple news sources, but unfortunately also overwhelms the reader with huge amounts of news articles (Bloomberg Terminal daily provides over 1 million articles). Using a news aggregator is also helpful since it usually allows one to specify a particular topic to follow, such as Business, Technology, Apple or Android. The list of topics is however quite narrow and does not allow one to specify long-tail interests.

In this paper we will describe a system called Event Registry [4] that tries to alleviate the aforementioned issues with news consumption and is freely available at [1]. Just as news aggregators it collects news articles published globally from more than 100,000 news sources in over 10 different languages. However, unlike the aggregators, Event Registry identifies from the articles the actual events that are being described in the articles. For Event Registry, an event is defined as any significant happening in the world that was reported in at least a few articles. Two examples of events are the death of David Bowie on Jan 11, 2016 that was reported in over 4,000 news articles as well as the news reported in 13 articles on Jan 23, 2016, that in Smithsonian's National Zoo, the Giant Panda was really enjoying the snow.

Grouping of news articles into events has several ad-

In: M. Martinez, U. Kruschwitz, G. Kazai, D. Corney, F. Hopfgartner, R. Campos and D. Albakour (eds.): Proceedings of the NewsIR'16 Workshop at ECIR, Padua, Italy, 20-March-2016, published at http://ceur-ws.org

[1] http://eventregistry.org/

vantages. First, given an event, the reader can choose to read articles from various news sources that reported about the event. Providing the complete and global coverage of the event allows the reader to construct an unbiased view of the event and all related details. Secondly, when browsing through the current events, the reader does not have to go through hundreds of news articles, where several articles report about the same event. Instead, all articles about the same event are grouped together and shown only once, which easily reduces the amount of content for one or two orders of magnitude. Lastly, for each event in Event Registry there is also abundant semantic information that is extracted from the articles, such as the location of the event, date, who and what the event is about, etc. This semantic information allows the reader to determine very specifically what his interests are and get a custom-tailored feed of events and news.

The rest of the paper is organized as follows. We will first describe the process in which Event Registry identifies events from news articles. We will also describe in more details the process in which the articles about the same event can even be linked although they are written in different languages. Additionally we will also describe the concept of a topic page which can be used by readers to very specifically determine the news articles and events of interest. We end the paper with a conclusion and some ideas for future work.

2 Event Registry

Event Registry consists of a pipeline of services that collect, process and analyze news articles collected globally in different languages. We will now briefly describe the major components in the pipeline.

2.1 Collecting news

In order to collect the news we developed a service called Newsfeed [5] that monitors RSS feeds of over 100,000 news publishers. Whenever a new article is detected in a feed, we crawl the web page and extract from it the news article and the available metadata information. In this way we collect daily between 200,000 and 300,000 news articles in various languages.

2.2 Semantic enrichment

The collected news articles provide information in unstructured form which requires a human to interpret it.

One way in which we extract structured/semantic information from the articles is by identifying and disambiguating relevant entities (people, locations and organizations) and non-entities mentioned in the articles. Examples of relevant non-entities would be things, such as Zika virus, murder, movie, automobile, etc. Identification of concepts (entities + non-entities) is done by wikification, which is a process of entity linking that uses Wikipedia as the knowledge base. As a result, each mentioned concept is annotated with a URI that is the link to the corresponding Wikipedia page. Since Wikipedia provides pages for the same concept in several languages (Barack Obama has a Wikipedia page in 225 languages), the question is which URL to take as the concept URI. We use the link to the English Wikipedia, when it is available, and the link to original (article) language otherwise. "Normalizing" the concepts to the same URI is very important since it allows the readers to find content regardless of the language in which it is written. The URI for the concept of the Sun, for example, would be the same, regardless if it is found in an English, Slovene (as 'Sonce'), Italian (as 'Sole') or any other language. Along with the URI, we also compute the relevance of the concept for the article. The relevance is computed depending on the number of times the concept is mentioned as well as it's locations in text and can be in the range between 1 and 5.

Another type of semantic enrichment we perform is categorization of the news articles based on the article's content. Currently we categorize news articles into a DMOZ [1] taxonomy. This taxonomy contains over a million categories, but we only consider top 3 levels, which amounts to 5,000 categories. The taxonomy was built for organizing web pages so it is not the optimal fit for categorizing news content. A more appropriate categorization would be to the IPTC's Media Topics taxonomy [2], which contains about 1.400 topics structured into 3 levels. Unfortunately we have not yet been able to obtain an annotated corpus of articles that we could use to train the models for this taxonomy.

Additionally we also extract from news articles all mentions of dates. Extracting dates is relevant for the following steps when we want to determine when the event described in the text occurred. In order to extract the dates we created an extensive set of regular expressions for individual languages that can detect date mentions in various forms.

2.3 Clustering of news articles

In order to group all articles that describe the same event we use an online clustering algorithm. The clustering is applied on each language separately and in short works as follows. Each collected article is first represented as bag-of-words – a representation in which we only keep an unordered list of words from the article and the number of times they occurred in the article. After applying TF-IDF weighting we compute the similarity of the article with centroids of existing clusters. The criteria that is used when computing similarity between the article and the cluster centroid are the cosine similarity of the text, similarity of the mentioned concepts and the date difference. If computed similarity of the most similar cluster is above the threshold, the article is put into the cluster, otherwise a new (micro) cluster is created, containing only the single article. Micro clusters are not considered to be events until they reach a certain number of articles. The threshold value for becoming an event depends on the language and was empirically determined to be between 3 – 6 articles.

News about an event are typically reported only for a limited amount of time. For this reason we also want to remove clusters after they reach a certain age. Currently, when a cluster becomes 5 days old we remove it, which means that new articles can not be assigned to it anymore. In this way we can maintain high performance of the system as well as prevent incorrect assignments of new events to old clusters.

2.4 Construction of events

Each time a micro-cluster of articles reaches a certain size, we form in Event Registry an event and associate it with the cluster of articles. Clustering has to be done for each language separately so each event is initially mono-lingual. Most relevant world events are however covered by various publishers globally that report in various languages. To represent such clusters as a single event we use a machine learning approach that will be described in more details in the next section.

Each created event is represented in Event Registry with a unique identifier that can be used to reference it. For each event we also want to extract it's core information – what occurred, where, who as involved, etc. To determine these details we use the available semantic and meta information provided by the articles assigned to the event.

To determine the date of the event, we can analyze the publishing date of the articles in the clusters. The naive approach would be to use the date of the first article as the date of the event. In practice this approach generates erroneous results for events that are reported in advance (such as various meetings of politicians, product announcements, etc.) as well as when the collected publishing dates of the articles are inaccurate. A more error prone approach that we use is to analyze the density of reporting and use the time point where the reporting intensified as the date of the event. Additional input can be provided by the mentioned date references – a particular date that is consistently mentioned across the articles most likely the correct date of the event.

In order to determine who is involved in the event we can analyze and aggregate the entities mentioned in the articles. A list of entities and their associated relevance can be obtained by analyzing the frequency of their occurrence in the articles as well as their assigned scores. Entities can be scored and ranked according to this criterion which provides an accurate aggregated view on what and who is the event about.

Location of the event is another important property. Since the event location is commonly mentioned in the articles, we can identify it by analyzing the frequently mentioned entities that are of type location. Additional signal for determining the event location can be obtained by inspecting the datelines of the articles. A dateline is a brief piece of text at the beginning of the news article that describes where and when the described story happened. The datelines are unfortunately not present in all news articles and even when they are, they sometimes represent the location where the story was written and not the actual location of the event. To determine which location, if any, is the event location, we apply an SVM classifier. Each mentioned city is considered to be a candidate for the event location and we generate for it a set of learning features. The features we use are based on the number times the city is mentioned in the articles and the number of times it is mentioned in the dateline. The SVM model that we use was trained on 200 events for which location was manually determined. Using 5-fold cross validation on this training data we found that the achieved classification accuracy of the model is 98%.

3 Cross-lingual linking of clusters

Since same events can be reported in multiple languages we need a way for identifying clusters in different languages that are discussing the same event so that they can be merged and represented as a single event. In short, we need an approach that given two clusters of articles determines if they describe the same event or not.

To perform the task we again represent it as a learning problem. From the two tested clusters we extract a set of learning features that can be used for training a classification model. There are three groups of learning features that we use:

Cross-lingual article similarity. Using an approach based on CCA [3] we can compute an estimated similarity between articles in different languages. Given this measure we can compute how similar individual articles in one cluster are to the individual articles in the other. From these results we can generate a number of learning features such as the maximum similarity, the average similarity, standard deviation, etc.

Concept-related features. Articles in Event Registry are annotated with concepts that have language independent URIs. For each cluster, we can analyze the associated articles and determine the top concepts based on how frequently they appear in these articles and what are their assigned scores. Using two such weighted vectors, one for each cluster, we can compute a list of informative features. Examples of these features include cosine and Jaccard similarities of the two vectors. Additional features can also be computed separately for the entities and non-entities in the vectors.

Miscellaneous features. Additional set of features can be computed reporting (a) whether the event locations found for the two clusters are the same or not, (b) the absolute difference in hours between the events in the two clusters and (c) the similarity of the dates that are being mentioned in the articles in the two clusters.

To evaluate how accurately we can, given these features, predict whether two clusters are about the same event or not, we performed the following experiment. Using two human experts we have manually annotated 808 pairs of clusters in English, Spanish and German language. The dataset contained 402 examples of cluster pairs that report about the same event and 406 examples where they do not. By training a linear SVM model and by using 10-fold cross validation schema we were able to achieve 89.2% classification accuracy.

4 Topic pages

Whenever an event is identified or updated, the information is stored in the Event Registry. Currently, Event Registry holds information about 3.6 million events that it identified from 88 million news articles, which were collected since January 2014. The users can use the web interface to search for events based on various criteria, such as relevant concepts, news sources that reported about it, location of the event, category, date, size and others. The users can also simply observe the stream of new/updated events as they are shown on the Event Registry home page.

An even more useful functionality than observing the whole feed of events, is the option for the users to create their own feed of articles and events based on their own interests. We call this functionality a topic page, where a topic can be defined using a set of relevant concepts, keywords, news sources and/or categories. The user can define the topic page using an interface shown in the top part of Figure 1. To each specified concept, keyword, news source and category, the user also assigns a weight of relevance for the topic. Each article and event that is processed by Event Registry is then scored according to the specified criteria and only those that achieve high enough score (a parameter specified by the user) are then shown to the user in the feed of the topic page.

More specifically, the scoring is done as follows. Let's assume that the user defines a topic T using a set of conditions $c_i, i = 1..n$ and their associated weights w_i, where conditions consist of one or more concepts, keywords, news sources and/or categories. For each new event e, a score $S_T(e)$ is computed as

$$S_T(e) = \sum_{i=1}^{n} w_i \cdot in(c_i, e) \cdot val(c_i, e)$$

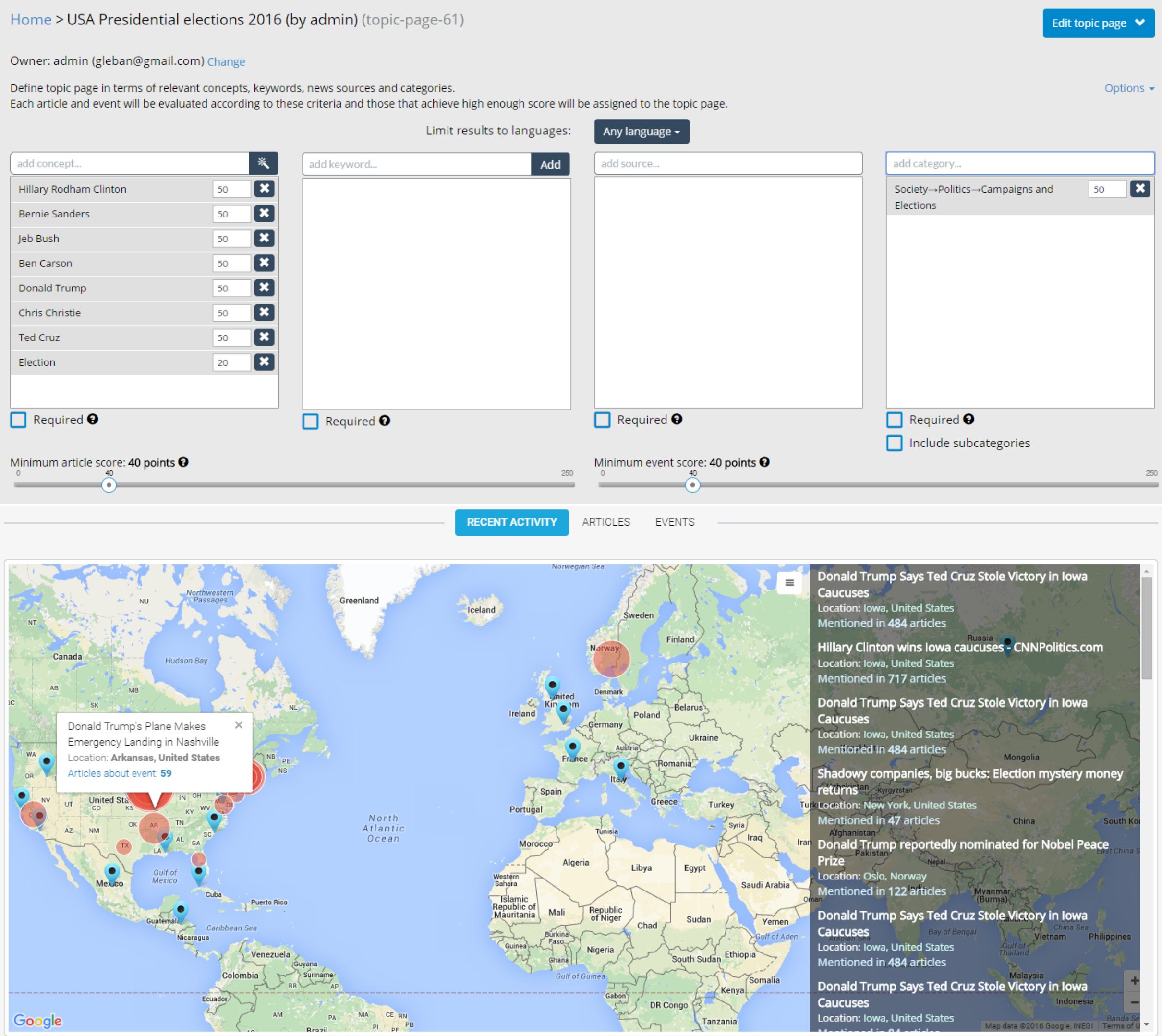
$$in(c_i, e) = \begin{cases} 1 & c_i \in e \\ 0 & otherwise \end{cases}$$

$$val(c_i, e) = \begin{cases} e_{c_i}/100 & c_i \text{ is a concept} \\ 1 & otherwise \end{cases}$$

The score $S_T(e)$ is therefore a simple sum over all conditions, where for each condition c_i we multiply the associated weight w_i with a Boolean function $in(c_i, e)$ and a scoring function $val(c_i, e)$. Function $in(e, c_i)$ simply determines if the condition c_i matches the event e or not. In case the condition is a concept or a category, the function is true when the event is annotated with it. In case the condition is a news source, the function is true if the event contains an article written by the news source. Lastly, in case the condition is a keyword, the function is true if the keyword appears in any of the articles assigned to the event. The scoring function $val(c_i, e)$ is trivial, except in the cases when c_i is a concept. When concepts c_j are associated with an event e, they are assigned a score e_{c_j} that is in range between 1 and 100, which represents how important the concept is to the event. The function $val(c_i, e)$ therefore simply ensures that for all conditions, the returned value is in range between 0 and 1. The scoring function for scoring articles is almost the same, except that the normalization constant in function $val()$ is 5, each concept in an article is assigned a score between 1 and 5. The events and articles that match the topic page can be then visualized on a map or displayed in a feed. An example topic page for USA presidential elections is available at Figure 1.

5 Conclusion

In this paper we have presented a system called Event Registry with fixes several shortcomings in the ways how news content is currently being consumed. Firstly, it is able to aggregate large amounts of news articles into actual events. Instead of flipping through tens or hundreds of articles about the same event in your


Home > USA Presidential elections 2016 (by admin) (topic-page-61)

Owner: admin (gleban@gmail.com) Change

Define topic page in terms of relevant concepts, keywords, news sources and categories.
Each article and event will be evaluated according to these criteria and those that achieve high enough score will be assigned to the topic page.

Limit results to languages: Any language

add concept...	
Hillary Rodham Clinton	50
Bernie Sanders	50
Jeb Bush	50
Ben Carson	50
Donald Trump	50
Chris Christie	50
Ted Cruz	50
Election	20

□ Required

add keyword... Add

□ Required

add source...

□ Required

add category...	
Society→Politics→Campaigns and Elections	50

□ Required
□ Include subcategories

Minimum article score: 40 points

Minimum event score: 40 points

Donald Trump's Plane Makes Emergency Landing in Nashville
Location: Arkansas, United States
Articles about event: 59

Donald Trump Says Ted Cruz Stole Victory in Iowa Caucuses
Location: Iowa, United States
Mentioned in 484 articles

Hillary Clinton wins Iowa caucuses - CNNPolitics.com
Location: Iowa, United States
Mentioned in 717 articles

Donald Trump Says Ted Cruz Stole Victory in Iowa Caucuses
Location: Iowa, United States
Mentioned in 484 articles

Shadowy companies, big bucks: Election mystery money returns
Location: New York, United States
Mentioned in 47 articles

Donald Trump reportedly nominated for Nobel Peace Prize
Location: Oslo, Norway
Mentioned in 122 articles

Donald Trump Says Ted Cruz Stole Victory in Iowa Caucuses
Location: Iowa, United States
Mentioned in 484 articles

Donald Trump Says Ted Cruz Stole Victory in Iowa Caucuses
Location: Iowa, United States

Figure 1: The interface for defining the topic page (top) and the feed of current events that match the criteria (bottom). The feed can be displayed on a map or as a list of matching articles and events.

news aggregator, a single item can be shown, together with the structured information about the event (who, what, when, where,...). If interested in the event, the user can then open the details of it and read individual articles (even in different languages) about it. By reading multiple articles, the user can form a more complete and unbiased view of the event as if he would be able to by just reading about it from a single news publisher. Having extensive structured information about the events allows the users of Event Registry to also create custom feeds based on a combination of general or long-tail topics of interest.

6 Acknowledgments

This work was supported by the Slovenian Research Agency as well as X-Like (ICT-288342-STREP) and xLime (ICT-611346-STREP) projects.

References

[1] DMoz, open directory project, *http://www.dmoz.org/*.

[2] Media topics, *https://iptc.org/standards/media-topics/*.

[3] S. T. Dumais, T. A. Letsche, M. L. Littman, and T. K. Landauer. Automatic cross-language retrieval using latent semantic indexing. In *AAAI spring symposium on cross-language text and speech retrieval*, volume 15, page 21, 1997.

[4] G. Leban and et. al. Event registry – learning about world events from news. In *Proceedings of*

23rd International World Wide Web Conference, 2014.

[5] M. Trampus and B. Novak. Internals of an aggregated web news feed. In *Proceedings of 15th Multiconference on Information Society 2012 (IS-2012)*, 2012.

Temporal Random Indexing: a Tool for Analysing Word Meaning Variations in News

Pierpaolo Basile
Dept. of Computer Science
Univ. of Bari Aldo Moro
name.surname@uniba.it

Annalina Caputo
Dept. of Computer Science
Univ. of Bari Aldo Moro
name.surname@uniba.it

Giovanni Semeraro
Dept. of Computer Science
Univ. of Bari Aldo Moro
name.surname@uniba.it

Abstract

The availability of data spanning different epochs has inspired a new analysis of cultural, social, and linguistic phenomena from a temporal perspective. This paper describes the application of Temporal Random Indexing (TRI) to the news context. TRI is able to build geometrical spaces of word meanings that consider several periods of time. Hence, TRI enables the analysis of the evolution in time of the meaning of a word. We propose some examples of application of TRI to the analysis of word meanings in the news scenario; this analysis enables the detection of linguistic variations that emerge in specific time intervals and that can be related to particular events.

1 Introduction

The analysis of word meaning variations over time periods is a crucial task for identifying changes in social and cultural phenomena. The diachronic analysis of a language allows to discover linguistic variations over time. Generally, a diachronic analysis is performed on a large time interval since linguistic variations happen quite slowly. However, this is not the case for fast data-streaming scenarios like the Web, and in particular social media such as Twitter or Facebook, where socio-cultural and linguistic phenomena quickly rise and fall. Although the news scenario is generally characterized by the use of a regular language, the large number of events that occur along the time line causes sudden topic shifts, making the analysis of this data similar to the data-streaming scenario.

In this paper we describe a technique called Temporal Random Indexing (TRI) that we have successfully applied to several diachronic analyses of the language [BCS15]. TRI is able to build several geometrical spaces of word meanings, called Distributional Semantic Models (DSM), one for each time interval, by skimming through huge corpora of text in order to learn the context of usage of words over time. In the resulting spaces, semantic similarity between words is expressed by the closeness between word-points. Thus, the semantic similarity can be computed as the cosine of the angle between the two vectors that represent the words. We show how to adopt TRI as a tool to discover particular phenomena in news data-streaming and how to link these linguistic changes to interesting events reported in the news content.

2 Methodology

TRI is based on Random Indexing (RI) [Sah05], a dimensionality reduction methodology and computational framework for distributional semantics. Given a term-term co-occurrence matrix A, RI builds a new matrix B where the Euclidean distance between points is preserved. Formally, given a corpus D of n documents, and a vocabulary V of m words extracted from D, we perform two steps: 1) assign a random vector r_i to each word $w_i \in V$; 2) compute a semantic vector sv_i for each word w_i as the sum of all random vectors assigned to words co-occurring with w_i in a given context. The context is the set of c words that precede and follow w_i. The second step is defined by the following equation:

In: M. Martinez, U. Kruschwitz, G. Kazai, D. Corney, F. Hopfgartner, R. Campos and D. Albakour (eds.): Proceedings of the NewsIR'16 Workshop at ECIR, Padua, Italy, 20-March-2016, published at http://ceur-ws.org

$$sv_i = \sum_{d \in D} \sum_{\substack{-c<j<+c \\ j \neq i}} r_j \qquad (1)$$

The set of semantic vectors assigned to words in V represents the *WordSpace*.

The classical RI does not take into account temporal information, but it can be easily adapted to our purposes by applying the methodology proposed in [JS09]. Specifically, if the corpus of n documents D is annotated with metadata containing information about the publication date, we can split the collection in p subsets $D_1, D_2, \ldots, D_p$, where p is the number of different time periods we want to analyse. The first step in the classical RI is unchanged in TRI: a random vector is assigned to each word in the whole vocabulary V. This represents the strength of this approach: the use of the same random vectors for all the spaces makes them comparable. The second step is similar to the one proposed for RI but it takes into account the temporal information: a different *WordSpace* T_k is built for each time period D_k. Hence, the semantic vector for a word in a given time interval is the result of its co-occurrences with other words in the same time interval, but the use of the same random vectors for building the word representations over different time spans guarantees their comparability along the timeline. This means that a vector in the *WordSpace* T_1 can be compared with vectors in the space T_2.

Let T_k be a period that ranges from $t_{k_{start}}$ to $t_{k_{end}}$, where $t_{k_{start}} < t_{k_{end}}$. In order to build the *WordSpace* T_k we consider only the documents d_k whose publication date falls within the time interval T_k as follows:

$$sv_{i_{T_k}} = \sum_{d_k \in D_k} \sum_{\substack{-c<j<+c \\ j \neq i}} r_j \qquad (2)$$

Using this approach we can build a *WordSpace* for each time period T_k over a corpus D tagged with information about the publication year. The word w_i has a distinct semantic vector $sv_{i_{T_k}}$ for each time period T_k built by accumulating random vectors according to the co-occurring words in that period. The great potentiality of TRI lies on the use of the same random vectors to build different *WordSpace*s: semantic vectors in different time periods remain comparable because they are the linear combination of the same random vectors.

3 Case study

The main goal of this case study is to show how to adopt TRI[1] to discover interesting phenomena in the news scenario. Specifically, we can analyse the similarity between the vector representations of a term across different time periods in order to detect changes in the usage of the term. Then, we can scrutinise both the neighbour terms and the news related to such a term during the period of time when the similarity has changed in order to understand if a specific event occurred.

Table 1: Neighbour terms of the word "scandal" in the two time periods 14-20 and 21-27 September 2015.

14-20 September 2015		21-27 September 2015	
allegations	0.60	cheating	0.86
called	0.60	volkswagen	0.83
corruption	0.59	rigging	0.80
made	0.59	automaker	0.79
apology	0.59	tests	0.79
met	0.58	carmaker	0.77
became	0.58	deception	0.77
case	0.58	german	0.76
initially	0.58	diesel	0.76
forced	0.58	emissions	0.76

We adopt the Signal Media One-Million News Articles dataset that consists of 1 million articles scraped during the time interval 1-30 September 2015. News are extracted from Reuters, in addition to local news sources and blogs. We split the dataset in five time periods of about one week: 1-6, 7-13, 14-20, 21-27, and 28-30. The split reflects the start and end of weeks in the month of September 2015. Then, for each period we build a *WordSpace* exploiting TRI. In particular, we analyse the 150,000 most frequent words in the whole corpus and we set the vector dimension to 500 using two non-zero elements in the random vector.

In each time interval, we try to discover terms that change their semantics with respect to the previous periods. Formally, given two time periods T_h and T_k, where T_h precedes T_k, and a term t_i, we can simple compute the cosine similarity between the semantic vector of t_i in T_h and the semantic vector of t_i in T_k $(sim(sv_{i_{T_h}}, sv_{i_{T_k}}))$. The similarity is a good indicator of the variation of semantics of the term t_i: a low similarity suggests a meaning shift. Using this approach we can rank all terms according to their similarity in ascending order. Top terms in the rank are good candidates for further analysis. However, in order to limit our analysis to those terms that frequently occur in the whole collection, the similarity scores have been multiplied by the term document frequency. By looking to such ranks, we discover that the word "scandal" had a semantic shift between the 3rd and the 4th week as showed in Table 1.

Another interesting analysis is the variation in similarity values between pairs of words over time: an

[1] TRI is available as an open-source project at: https://github.com/pippokill/tri

upsurge in similarity reflects the increment of co-occurrences between the two words in similar contexts. Figure 1 reports the similarity between "scandal" and "Volkswagen" over time. The plot shows a spike in the similarity value starting from the fourth time interval (21-27 September), which corresponds to the scandal about the Volkswagen diesel emission.

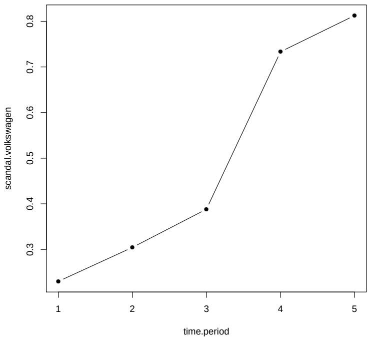


Figure 1: Word-word similarity between the terms: "scandal" and "Volkswagen".

Semantic vectors can be exploited to implement a semantic information retrieval system [BCS11]. The idea is to provide a vector representation for both documents and queries. In particular given a text W (e.g. a document or a query) composed of k terms we can build the vector representation of W as the vector sum of the k semantic vectors occurring in W. Formally, given $W = t_1 t_2 \dots t_k$ the sequence of k terms in W, its vector representation is $\mathbf{w} = sv_{t_1} + sv_{t_2} + \cdots + sv_{t_k}$. Using the same approach we can build the vector representation $\mathbf{q}$ for a query Q. Then the similarity between a query Q and a document D is given by the cosine similarity between $\mathbf{q}$ and $\mathbf{d}$. TRI provides different *WordSpaces* for each time period T_k. Then, the vector representation of a document published during the period T_k is built by exploiting only the semantic vectors of the corresponding *WordSpace*. At query time, the query representation is built by taking into account the semantic vectors of the time period we want to search.

As showed in Figure 1, in the third time period the similarity between "scandal" and "Volkswagen" starts to increase. We try to investigate this phenomenon from the information retrieval point of view. Table 2 reports the first three snippets retrieved by the query "scandal" and "Volkswagen" in the third time period. The column VSM reports results obtained with a classical vector space model implemented by Lucene[2], while the column TRI reports results obtained by TRI.

The VSM model gives more importance to documents that contain both terms, this is the case of the first two documents, while the third document is not relevant at all. The first three documents retrieved by TRI are all relevant for the given query since they all talk about events related to the Volkswagen diesel emission scandal. However, it is interesting to notice that no document contains explicitly the word "scandal". These results can be explained by the nature of the semantic search, which does not rely on string matching, but rather assigns a rank to documents on the basis of their proximity to the semantic vector $scandal + Volkswagen$ taken from the third time period. Semantic search based on TRI opens new opportunities for implementing effective semantic search engines that take into account word meaning variation over time. We plan to deeply investigate this aspect in future research.

Table 2: Search results for query "scandal Volkswagen" in the third time interval: 14th Sept.-20th Sept. 2016

VSM	**TRI**
Volkswagen multi billion pollution coverup *scandal*...	*Volkswagen* to recall 500,000... a device that disguises pollution levels...
Volkswagen emissions cheating... investigations over an emissions *scandal*...	EPA, California investigate *Volkswagen* for clean air violations...
The reinvention of *Volkswagen*. In the *Volkswagen* Group, there is a sense...	*Volkswagen* Ordered to Recall Half a Million Cars After It Cheated on Smog Checks...

References

[BCS11] Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. Integrating sense discrimination in a semantic information retrieval system. In *Information Retrieval and Mining in Distributed Environments*, volume 324, pages 249–256. Springer, 2011.

[BCS15] Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. Temporal random indexing: A system for analysing word meaning over time. *IJCL*, 1(1):55–68, 12 2015.

[JS09] David Jurgens and Keith Stevens. Event Detection in Blogs using Temporal Random Indexing. In *Proc. of the Workshop on Events in Emerging Text Types*, pages 9–16, 2009.

[Sah05] Magnus Sahlgren. An Introduction to Random Indexing. In *Methods and Applications of Semantic Indexing Workshop at TKE 2005*, volume 5, 2005.

[2] `https://lucene.apache.org/core/`

What do a Million News Articles Look like?

David Corney, Dyaa Albakour,
Miguel Martinez and Samir Moussa
Signal Media
16-24 Underwood Street, London N1 7JQ
{first.last}@signalmedia.co

Abstract

We present a detailed description and analysis of the Signal Media One-Million Articles dataset. We have released this dataset to facilitate research on emerging news-related information retrieval (IR) challenges. In particular, we have observed over the past decade emerging novel paradigms for publishing and consuming news, where users can get updated on the go with news from multiple sources, and at the same time, news providers are increasingly using social media and citizen journalism as powerful news sources. As a result, a number of news-related IR tasks have emerged and attracted attention in industry and academia. These include news verification, temporal summarization of multiple news sources, and news recommendation among others. A number of news datasets were created and shared for news IR in the past. However, such datasets are often drawn from a single outlet, and heavily preprocessed and cleaned. Also, they have become outdated and are not suitable any more for the emerging news IR challenges described above. Our dataset aims to address this because it is a recent collection from a wide range of sources reflecting many real-world issues in news collection and analysis.

We present insights obtained from an analysis of certain characteristics of the dataset, such as article lengths; similarity between articles; and the temporal characteristics of news publishing. We also discuss the opportunities and the limitations of our dataset.

1 Introduction

We present an analysis of The Signal Media One-Million News Articles Dataset[1]. We have created and shared this collection to stimulate research into new and improved methods for large-scale text analysis related to a wide range of applications. The articles were collected from a variety of news sources in September 2015. The sources include major national and international outlets, such as Reuters, the BBC and the New York Times, along with many sources that have fewer readers and less impact, including news magazines, blogs, local outlets and specialist publications. The collection is shared under a Creative Commons licence[2] while the copyright of the articles remains with the original publishers.

There are several existing collections of news-related texts that have been widely used by the information retrieval (IR) and natural language processing (NLP) communities, such as the Twenty Newsgroups collection [twe99]; the Reuters-21578 test collection[Lew96]; several Reuters Corpora, such as RCV1 [LYRL04], RCV2 and TRC2; and more recently Yahoo's user-news feed interaction data set [Yah16]. Common Crawl[3] provides a collection of 1.8 billion pages, but this represents a snapshot sample of the entire web, rather than a focussed news collection.

While such datasets continue to be useful for evaluating and guiding research, most are limited to a single source or a website. Furthermore, such datasets

In: M. Martinez, U. Kruschwitz, G. Kazai, D. Corney, F. Hopfgartner, R. Campos and D. Albakour (eds.): Proceedings of the NewsIR'16 Workshop, Padua, Italy, 20-March-2016, published at http://ceur-ws.org

[1] Available for research purposes from `http://research.signalmedia.co/newsir16/signal-dataset.html`

[2] Attribution, non-commercial `https://creativecommons.org/licenses/by-nc/3.0/`

[3] `http://commoncrawl.org/`

are highly curated and refined, which may result in an over-estimation of performance: if an algorithm performs well on such a clean set, will it still perform well when presented with more noisy data, such as content obtained from web-scraping or other less-controlled sources? In the fast changing era with news publishing and consumption, such datasets have become outdated and are not suitable for emerging IR tasks on news. Our dataset addresses this by providing a recent large sample of news articles from multiple sources over a one month period.

Like our collection, the British National Corpus [Bur07] is monolingual (English-only), synchronic (sampled from a single time period), general (not limited to any genre or topic), and sampled from a wide range of sources. Our dataset shares many of these features, except that while the BNC is definitively from a single nation, our collection spans the globe. While focussed on news, our collection also contains imaginative works of blogs and transcriptions of broadcast speech.

The articles of the dataset were originally collected for Signal Media by Moreover Technologies from 1st–30th September 2015. This included repeated sampling from over 93,000 different news sources ranging from large-scale mainstream media outlets to single-author blogs. Recent decades have seen a blurring of the distinction between mainstream and citizen journalism[Gla08], hence our inclusion of multiple sources in the collection. As well as being an attractively round number, one million articles is a large enough collection to develop and evaluate a wide range of tools and models, while still being manageable enough to not require specialist or sophisticated infrastructure.

Due to the scale of the collection, and the importance of speed, the collection process is largely automated. This, along with the variation in quality of source websites, inevitably leads to imperfections in the collection. For example, the goal was to collect English-language content only, but manual analysis shows that a small proportion of articles are in other languages or a mixture of languages. Similarly, some articles contain fragments of HTML, PHP or JavaScript code, due to problems with encoding, rendering or scraping, and some are duplicated (see Section 2.2). We consider such issues as the inevitable presence of noise in any real-world collection. One of our aims in sharing this dataset is to encourage the development of tools and methods that are robust enough to cope with data of this nature.

In the remainder of the paper, we provide insights from an analysis conducted on a number of the dataset characteristics (Section 2). We share some tools for accessing the data (Section 3), and discuss the opportunities and the limitations of the dataset in Section 4.

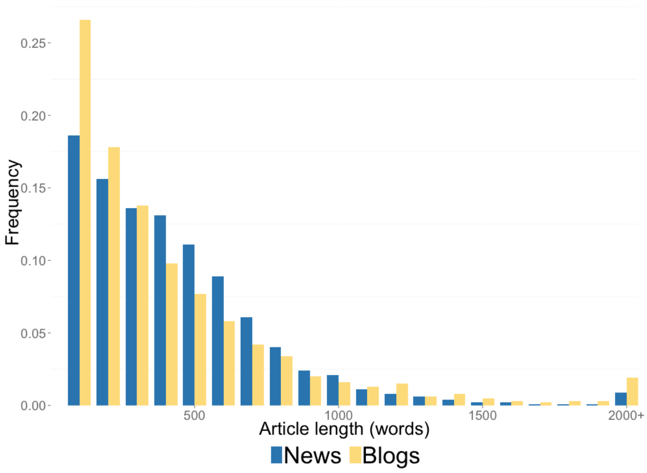


Figure 1: Distribution of article length for each media type (blogs and news)

2 Characteristics of dataset

2.1 Article length

Some news articles are little more than a single statement, especially if a news story is still breaking. Other articles may contain a lot of details and background information or discussion. To investigate this, we use a simple tokenizer[4] to count the number of words in each article. Figure 1 shows the distribution of article lengths, comparing the length of 'news' articles to the lengths of 'blog' articles. This shows that typical articles are a few hundred words long, but with a long tail reaching up past 2000 words. Note also that articles between 400 and 1000 words long are more likely to be news articles than blogs, while very short articles (200 words or less) and long articles (1200 words or more) are more likely to be blogs than news.

The tokenization also allows us to calculate that there are 407,754,159 words in the dataset; that there are 2,003,254 *distinct* words in the dataset; and that the average number of words per article is 407.75.

Which articles are unusually long? And which articles are unusually short? Just 144 articles in the collection have more than 10,000 words. The longest article has 12,450 words and is a transcript of a US college football match. Other long articles include an installment of serialized novel; an updated about a fantasy football competition; detailed personal memoirs; and a detailed list of fishing reports from Florida. While these may never attain very large readerships, they

[4] We used the Standard Tokenizer in ElasticSearch v1.7, which splits on white space and punctuation symbols, while allowing for abbreviations.

Title	First Week of ICE November 20th Options Trading
Content	Investors in Intercontinental Exchange Inc. (ICE) saw new options become available this week, for the November 20th expiration.
Media-type	News
Source	Town Hall
Published	2015-09-22T16:31:56Z

Figure 2: The shortest article in the collection, with 18 words of content.

will no doubt be of interest to certain audiences. At the other extreme, the shortest article in collection is nothing more than a brief announcement about options trading with no background discussion or detail. The article is shown in its entirety in Figure 2, including available metadata.

2.2 Duplicated articles

Identical, or near-identical articles may appear in the collection for several reasons, including:

Syndication One publisher may publish the same article through several outlets, such as regional newspapers.

Updates One source may publish multiple versions of the same article over time, especially in the case of updating breaking news stories.

Aggregation Some news aggregation sites (such as `wn.com` and `www.newslookup.com/`) display copies of articles originally published elsewhere. These articles may have already been collected from the primary sources.

Access issues Some sites give the same content when an automated tool attempts to access multiple articles on the site (e.g. a copyright or login notice), with the full text being behind a firewall.

To investigate this, we measured the cosine similarity between pairs of articles. This compares the frequency of terms found in each article, ignoring the word order. A score approaching one means that the same words appear at the same frequencies; a score close to zero means that entirely different words appear. If two articles differ by only a few words, we still wish to treat them as near-duplicates. In this way, we can ignore differences caused by different page layouts, changes in the byline or other minor edits.

In Figure 3, we show the probability distribution of cosine similarity scores between pairs of articles. To generate this, we randomly sampled 250 articles and measured the cosine similarity between each of these and each of the million articles. We removed stopwords and calculated the term frequencies of the remaining words (though similar results were obtained without removing stopwords). The cosine similarity scores were grouped into 100 equal-sized bins to generate the graph. Note that we are sampling from the space of similarity scores and not the space of articles: full pairwise analysis of a smaller corpus would lead to a very different, sparser distribution.

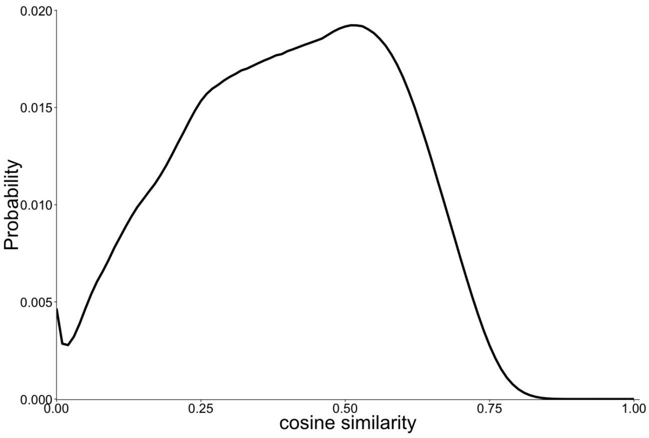


Figure 3: Distribution of pairwise cosine similarities. This shows the probability of generating different scores when comparing 250 × 1m articles.

The broad peak in central portion shows that when choosing two articles at random, their cosine similarity is likely to be around 0.5. 90% of article pairs have a cosine similarity between 0.13 and 0.70. This shows a substantial degree of overlap between articles in terms of word frequencies, despite the wide range of sources and topics covered. The slight up-tick at the extreme left shows that around 3% of articles share no terms (except stopwords) with any other article. These are typically very short articles containing rare terms (such as rare proper nouns). This includes some articles that have been corrupted during collection, and contain only fragments of code. At the other extreme, 0.0561% of pairs have a similarity score of greater than 0.95. This suggests that articles have on average around 2.2 duplicates or near-duplicates. This result is consistent with the duplicate recognition component of Signal Media's news monitoring platform[5].

2.3 Typical articles

Beyond considering duplicates, we investigated which articles are 'typical' and 'atypical' with regard to the rest of the collection. To do this, we remove stopwords and generate term frequency vectors for each article, as before. We then combined these into a single term

[5] `signalmedia.co`

Term	TF	DF
2015	930524	352629
time	642021	331982
people	531947	235954
1	518949	208164
september	471723	204583
market	423744	129407
company	416838	161796
2	414198	185599
day	387289	206976
world	365728	188220
news	364983	216282
information	351809	177431
business	327894	144896
3	313890	154149
10	312377	176963
home	293098	156997
million	278202	120327
including	275302	183904
week	271468	153210
team	263352	128957

Figure 4: The 20 most common terms, excluding stopwords, showing the total number of term occurrences (TF) and the document frequency (DF).

frequency vector representing the centroid of the entire collection. We then measured the cosine similarity between each article and this centroid. Articles with a high score can be seen as typical of the collection, in that the distribution of terms is similar. We plan to repeat this analysis using TF·IDF, but we expect to find a similar pattern of results overall.

Figure 4 lists the twenty most common terms in the collection (excluding stopwords). These include the month and year of the collection, along with words such as 'news', 'market', 'business' and company, indicating the typical focus of stories. .

One of the most typical stories by this definition, with a centroid cosine similarity of 0.9620, is about a new portable charge storage device[6]. It contains words such as 'power', 'time', 'heat' and other terms commonly found in news articles as shown in this extract: "*Harnessing the thermal energy from the included heating pot, it generates 5-watt power that can charge a device via the attached USB cable. You provide the heat source and liquid, and the PowerPot charges any compatible electronic with a USB connection. Charging time is actually comparable to a standard outlet, at about 1 to 2 hours to fully charge a phone.*"

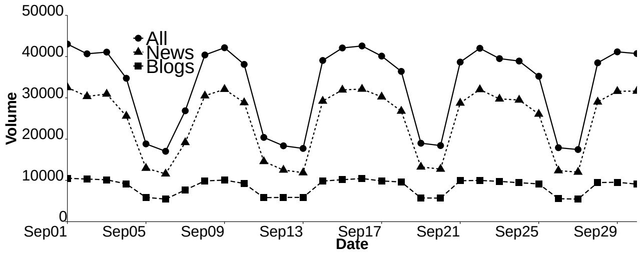


Figure 5: Daily volume over time for the different media types (Blogs and News).

One of the least typical stories is an article about tomato soup, in a small-town American newspaper[7] : "*That slight chill in the evening air? That's fall. And what better way to chase it than with a bowl of steamy, homemade soup. Chef Shereen Pavlides' recipe for Tomato Basil Soup is just what the season ordered, made with chicken stock and hearty San Marzano tomatoes.*" This has a cosine similarity with the centroid vector of just 0.00160.

While this centroid approach gives some indication of typical and atypical articles, it ignores the fact that the collection is far from homogeneous. A more sophisticated analysis could define a number of representative centroids, each representing a typical article belonging to a particular category. Such a segmentation of the collection could be achieved using any of a wide variety of document clustering algorithms.

2.4 Volume over time

In Figure 5, we plot the daily volume of articles published across the full duration of the dataset (September 2015). We can see the expected weekly pattern with more activity (published articles) during weekdays and less activity during the weekends. This is true for both media types (blogs and news). However, a couple of exceptions can be observed in these weekly cycles. First, there are fewer articles published on Monday, September 7^{th} compared to other Mondays and other weekdays in the month. This was a public holiday in the United States (Labor Day), which resulted in less media activity. Note that the dataset is sampled only from English-language sources and therefore the majority of articles originate from English speaking countries (with the USA the largest of those). The second exception is Friday, Sept. 11^{th} where the volume has dropped due to a downtime in the collection process that occurred on that day.

We now consider the hourly distribution of article publication in Figure 6. We calculate the average volume of articles in each hour of the day using the GMT timezone over the 30-day period. We observe a sharp

[6] Article ID 05bab596-55a3-40aa-adde-1a2cb48ebc41

[7] Article ID 497a3712-dd6d-4b0a-9365-5a7ade79d905

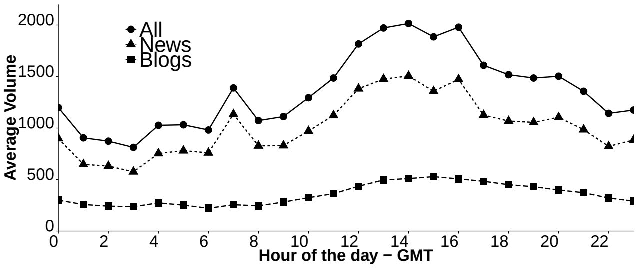


Figure 6: Average hourly volume for the different media types (Blogs and News).

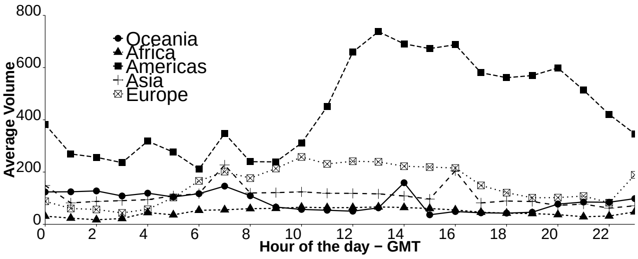


Figure 7: Average hourly volume for News sources across different geographical regions.

rise in the "News" volume at 7:00 GMT. The volume of "News" increases throughout the day and it reaches a peak between 12:00 GMT and 16:00 GMT. Afterwards the volume starts to decrease through the late hours of the day. To explain this behaviour, we further break down the "News" sources based on the continent of their sources' origin in Figure 7. We can observe that the peak at 7:00 GMT is mainly due to increased activity in Europe with the early hours in the working day there. Hours later, when it is morning in America (12:00 to 16:00 GMT), we observe an increase of activity from American sources which in fact dominate the stream.

For "Blogs" (Figure 6), we observe a different picture where the volume is more stable throughout the day with slight increases during the morning and afternoon hours.

3 Open Source Tools

We have created an open source repository on GitHub [8] to host useful tools and programming scripts for processing the dataset. For example, we have added scripts to index the data with ElasticSearch and to convert it into the TREC format for easier compatibility to some other IR tools. We will continue to maintain and promote this repository encouraging the community to provide similar tools for processing the dataset.

4 Discussion

The Signal Media One-Million News Articles Dataset provides a distinctive research asset. The articles come from multiple sources and reflect many of the realities of practical large-scale text analytics. However, there are inevitably limitations to the data, some of which we now consider.

Language The dataset is almost monolingual, being dominated by English-language articles and with the remaining articles mostly being a mixture of English and non-English text.

Date range The articles were all collected during September 2015; a small number were published in the preceding days and weeks but only detected and downloaded in that month. By limiting the sample to a single calendar month, we achieved a relatively high density of related articles, such as multiple articles written about the same events. One month is also long enough for some of the news stories to change over time, making the dataset suitable for topic detection and tracking studies. Of course, by restricting the collection to this period makes it less useful for longer-term topic tracking.

Missing links and multimedia resource The dataset is text-only, and does not contain links to the original articles. This is partly due to issues around image licensing, but also to avoid problems with link rot: the ever-changing nature of the internet means that any published URLs would not reliably point to the same text as in the dataset. A dataset containing archived copies of multimedia content related to news would be very useful, but is beyond the scope of this work.

Labelling Many tasks require labelled documents, such as to assign documents to topic categories or to disambiguate entities. Although the collection currently has no such labels, our goal is to encourage the wider community of IR researchers to use this collection for a variety of tasks and share label sets for parts (or all) of the articles.

In conclusion, we believe that this dataset will be useful for developing, evaluating and comparing a wide range of news-related information retrieval tools and algorithms. The analysis provided here provides an initial description of the collection and forms the basis for future work.

[8] https://github.com/SignalMedia/Signal-1M-Tools

References

[Bur07] Lou Burnard. Reference guide for the British National Corpus (XML edition). `http://www.natcorp.ox.ac.uk/XMLedition/URG`, February 2007.

[Gla08] Mark Glaser. Distinction between bloggers, journalists blurring more than ever. `http://mediashift.org/2008/02/distinction-between-bloggers-journalists-blurring-more-than-ever059`, 2008.

[Lew96] David Lewis. Reuters-21578 text categorization test collection. `http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html`, 1996.

[LYRL04] David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *The Journal of Machine Learning Research*, 5:361–397, 2004.

[twe99] 20 Newsgroups dataset. `http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html`, 1999.

[Yah16] Yahoo News Feed dataset. `http://webscope.sandbox.yahoo.com/catalog.php?datatype=r&did=75`, 2016.

Exploring a Large News Collection Using Visualization Tools

Tiago Devezas[1, 2]
tdevezas@fe.up.pt

José Devezas[2]
jld@fe.up.pt

Sérgio Nunes[1, 2]
ssn@fe.up.pt

INESC TEC[1] and DEI[2], FEUP, University of Porto
Rua Dr. Roberto Frias, s/n
4200-465 Porto, Portugal

Abstract

The overwhelming amount of news content published online every day has made it increasingly difficult to perform macro-level analysis of the news landscape. Visual exploration tools harness both computing power and human perception to assist in making sense of large data collections. In this paper, we employed three visualization tools to explore a dataset comprising one million articles published by news organizations and blogs. The visual analysis of the dataset revealed that 1) news and blog sources evaluate very differently the importance of similar events, granting them distinct amounts of coverage, 2) there are both dissimilarities and overlaps in the publication patterns of the two source types, and 3) the content's direction and diversity behave differently over time.

1 Introduction

Finding valuable information in large collections of data can resemble looking for a needle in a haystack. An effective way to address this problem is the use of data visualization tools to explore datasets [Kei01]. The presentation of abstract data through interactive visual tools leverages human perceptual abilities and enhances cognitive performance, thus promoting discovery and sensemaking. In this paper, we present three distinct visualization tools for exploring large news collections, and apply them to the Signal Media One-Million News Articles Dataset[1], a collection of one million news and blog articles.

We show three use cases that highlight how these tools allow the investigation of distinct dimensions of the data. The first case evaluates how the hierarchy of importance given to a set of select global events, manifested through the amount of coverage, varies between news and blog sources. The second investigates the publication patterns of both source types during 24-hour and seven-day weekly cycles. The third use case studies the variation of topical diversity for news and blogs over time and employs a visualization tool developed specifically for this work. To develop this tool, an analysis was conducted to identify the topic vectors representing the directions followed daily by the articles' contents, compute a diversity score, and measure the topic diversity over time for news and blogs.

2 Corpus Characterization

The Signal 1M Dataset is comprised of one million articles published by 93,345 distinct media sources of two types: news and blogs. An analysis of the articles' media type reveals that 18,533 sources published exclusively news articles, 74,333 sources published only blog stories, and 479 had documents of both types. As for the article count by media type, nearly three-fourths were news (734,488 or 73.4%) and one-fourth blog items (265,512 or 26.6%). Thus, despite its lower amount, news sources were responsible for the publication of the majority of articles.

Even though the publication period extends from Jul 2nd 2015 to Sep 30th 2015, the majority of the articles were published between Sep 1st 2015 and Sep 30th 2015 (987,248 or 98.7%). Of these, 734,488 (74.4%)

In: M. Martinez, U. Kruschwitz, G. Kazai, D. Corney, F. Hopfgartner, R. Campos and D. Albakour (eds.): Proceedings of the NewsIR'16 Workshop at ECIR, Padua, Italy, 20-March-2016, published at http://ceur-ws.org

[1] http://research.signalmedia.co/newsir16/signal-dataset.html

were news articles and 265,512 (26.9%) blog articles. The highest number of articles published by a single source was 192,228 and the lowest amount, a single article. Regarding the overall distribution of articles, the majority of the sources (91,693 or 98.2%) published 100 articles or less, 1,565 sources (1.7%) published between 101 and 1000 articles, 85 (0.09%) between 1,001 and 5,000, one (0.001%) between 5001 and 10000, and one between 10,000 and 20,000 articles.

The topic analysis conducted for each media type stream (see Section 5.3.2) found that the top five n-grams, based on the TF-IDF score of the topic vectors, were 'south africa', 'pope francis', 'total volume table', 'high school football', and 'college football' for news articles, and 'star wars', 'school district', 'syrian refugees', 'executive director', and 'kansas city' for the blog document set.

3 Visualization of Large News Archives

The visualization and analysis of large volumes of news content is an emerging field of research [KBMK10]. The ThemeRiver application [HHN02] was one of the first efforts in this domain. It provides an interactive visualization of thematic changes across a large set of news documents over time. It uses a metaphor of a river to assist in the recognition of relationships, trends and patterns in the data. Themes are displayed as colored streams whose width — the measure of its strength — varies as it flows across time from left to right. A similar river-like visual metaphor is employed by the NewsLab system [GLYR07], which allows exploratory analysis of the temporal variation of themes, and their hierarchical structure, from a large collection of news videos.

Krstajić et al. [KBK11] present CloudLines, a visualization technique to display a compact view of multiple time series, each showing a sequence of related events and event episodes (high density sequences of events). The relative importance of events is conveyed through variations in the clusters' opacity and size. The system also permits fine-detailed analysis of individual event data points.

The complexities of visualizing the dynamics of news data streams are addressed by Krstajić et al. [KBMK10]. The system displays the evolution of news in real-time by converting the stream into threads comprised of similar articles. In addition to showing recent threads, the system computes the threads' relevance on the fly — based on the items' age and their relationships — to determine which threads to keep on screen and which ones to remove.

The development of news stories and their relationships through time is also explored by Story Tracker [KNAMK13]. The application represents the evolution of stories over time, and how they merge and split. Story clusters are displayed as rectangles whose size corresponds to the number of articles and include labels for the story title and the most important keywords. Related clusters have the same color, are edge-connected, and can be zoomed to the level of the individual articles that compose them.

The NewsStream service [NGSM15] provides several interactive tools to visually explore a continuously updated collection of financial articles, published via the RSS feeds of multiple news and blog sources. The system displays occurrences and co-occurrences of financial and geographic entities in the news, the related sentiment, a summary of the linked content through tag clouds, and temporal country co-occurrence networks displayed on a world map.

4 The MediaViz Platform

The MediaViz platform [DNR15] aims to assist in gaining insight from a large archive of news through interactive visualization tools. It comprises two components. The first is a back-end application that fetches and stores articles published via the RSS feeds of multiple online news sources and provides access to the data through an API. The second is a client application which retrieves the data provided by the API and allows its exploration through interactive visualization tools. Our approach is based on open technologies and was built with extensibility in mind: the client application is decoupled from the back-end so it can be configured to work with different datasets with minimal effort. For this paper, we stored the Signal 1M Dataset in a relational database and built a simple API. No major modifications were required for the existing visualization tools to work with the new API. However, a new tool was developed to explore topic diversity over time for news and blog articles. A fully functional demo is available online[2].

5 MediaViz Visualization Tools

Rather than focusing on individual sources, we opted to explore the two types of media sources that comprise the corpus — news and blogs —, as they allow a macro-level analysis and comparison of the dataset.

5.1 Variations in Coverage

The dynamics of the coverage that each source type granted to different themes over time are displayed by the Keywords tool. Users can insert multiple search terms and see how many articles (in absolute terms or

[2] http://irlab.fe.up.pt/p/mediaviz/newsir/

as a percentage of all articles published on the respective day) with those keywords were published daily during the selected period. Additional context can be obtained by clicking the data points, which displays a list of all related articles. Each list item includes the title, summary, publication date and the source's name, and can be clicked to display the full text.

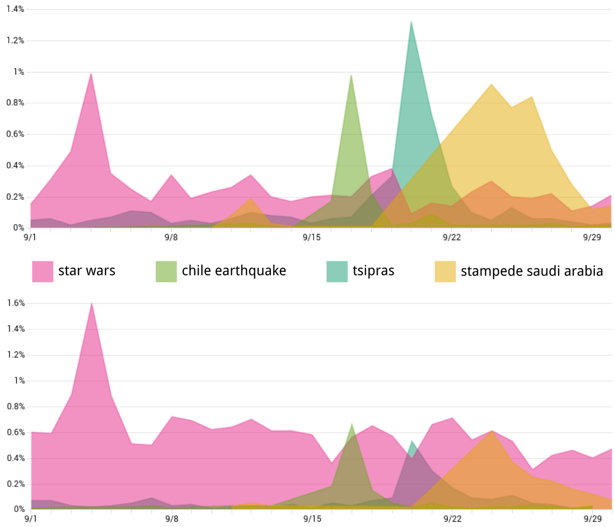


Figure 1: MediaViz Keywords tool. Top: Daily percentage of articles published by all news sources containing the given terms. Bottom: Daily percentage of articles published by all blog sources containing the same terms.

Figure 1 displays the daily percentage of articles published between Sep 1st 2015 and Sep 30th 2015 by each source type with the terms 'star wars', 'chile earthquake', 'tsipras', and 'stampede saudi arabia'. These particular terms were chosen because they are related with some relevant global events — identified after consulting several online resources — that took place on September 2015. The visualization's peaks highlight the selected events: the merchandise for the latest Star Wars movie was released on Sep 4th; an earthquake in Chile which led to the evacuation of millions of people took place on Sep 16th; on Sep 20th Alexis Tsipras was reelected as Prime Minister of Greece after resigning and calling for a snap election; and, on Sep 24th, hundreds of people died after a stampede during the annual pilgrimage to Mecca, in Saudi Arabia. As shown in Figure 1, the attention given to these events differed greatly between the two source types. News sources (top), gave similar attention to each event, while in blogs (bottom), the primacy belongs to articles mentioning Star Wars.

5.2 Publication Patterns

The Sources tool allows the comparison of publication patterns (count and percentage of articles) for multiple sources according to distinct temporal granularities: weekly, monthly and 24-hour cycles. To have comparable results, publication times are converted to the UTC time standard. The ability to compare several sources in the same screen can thus provide meaningful perspectives regarding their production cycles. This can be seen in Figure 2. News sources published a higher percentage of articles than blogs during business days, a behavior that is reversed during the weekend. While this pattern might be expected, given the particularities of each media type, the Sources tool quantitatively shows that such assertion is indeed true.

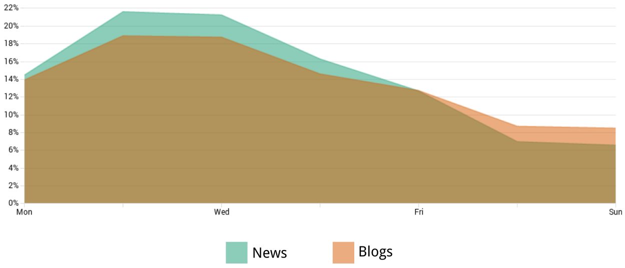


Figure 2: MediaViz Sources tool. Percentage of articles published by both source types for each day of the week.

When looking at a 24-hour cycle, news and blog sources exhibit similar patterns. As Figure 3 displays, publications follow a typical working schedule: the most active publication period occurs between 08:00 and 16:00 UTC and then gradually decreases. One possible explanation for this overlap is the growing professionalization and influence of blogs, which often compete with traditional news sources for online eyeballs. The most significant difference between the two patterns, the news sources' peak at 07:00, can be potentially explained by the publication of early morning news.

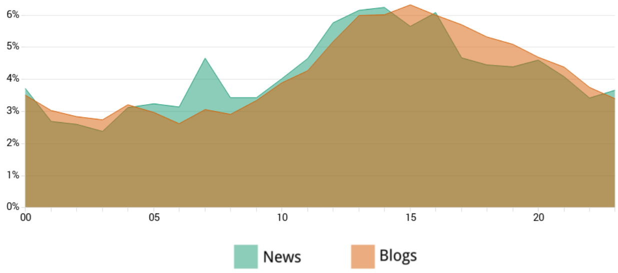


Figure 3: MediaViz Sources tool. Percentage of articles published by both source types during a 24-hour cycle.

5.3 Diversity Explorer

The Diversity Explorer tool was developed specifically for this work. Below we describe our strategy for detecting topics and measuring topical diversity between the news and blog streams.

5.3.1 Topic Detection

Our topic detection strategy was based on the clustering of text documents using n-grams of size $n = 2$ (bigrams) and $n = 3$ (trigrams) as features. The base strategy consisted of, for a given day, transforming each document into a bag of n-grams and then running k-means [HW79] using the n-gram frequencies as features. The value of k was selected based on the Silhouette method [Rou87], by testing successive values of $k \in [2, 15]$ for a random sample of 100 or less documents — in case less than 100 documents were available. Constraining the value of k, indirectly enforced the number of topics to range between 2 and 15. The result of this process was a set of k topics, represented by the centroid of each cluster and associated with the documents for each day.

Prior to the clustering phase, and in order to ensure performance, we reduced the number of features by removing n-grams that were over 99.6% sparse, i.e., features with more than 99.6% zeros, that were less useful in distinguishing documents, were simply discarded. The sparsity threshold of 99.6% was determined empirically, by experimenting with the largest daily document set and ensuring that the number of features would not explode (99% decrease from 1,834,310 to 350 features for the largest daily document set), but also with smaller daily document sets to ensure that the number of features would not be too small (nearly 0% decrease for daily document sets with less than 100 documents). After completing the feature reduction process, we repeated the previously described clustering process for the smaller matrix, obtaining k topic vectors that illustrated the different directions of followed contents in daily news.

5.3.2 Measuring Topic Diversity

In order to measure topic diversity within a corpus, we took the topic vectors for a given day and did an element-wise aggregation based on the maximum weight of each n-gram. This resulted in a set of daily vectors, describing the overall topical direction of news and blog articles per day.

Our approach to measuring topic diversity was based on a combined distance metric between all n-gram daily vectors, for a given corpus — the more distant the topics are from every other topic, the higher the diversity. We computed the normalized cosine distances X for each pair of n-gram daily vectors, separately for the news and blog corpora. Next, we calculated the mean and standard deviation for the obtained values, and combined the mean $E[X]$ and standard deviation $\sigma(X)$ into a diversity score, as described in Equation 1.

$$\begin{aligned} score(X) = E[X] - 2 \times (F(E[X]; 0.5, 1/50) \times 0.5) \\ \times (1 - E[X]) \times (E[X] \times \sigma(X)) \end{aligned} \tag{1}$$

$$F(x; \mu, s) = \frac{1}{1 + e^{-\frac{x-\mu}{s}}} \tag{2}$$

The idea was for the variance to affect the mean cosine distance in the following way: for a low mean, a low variance would result in a small increase, while a high variance would result in a large increase; for a high mean, a low variance would result in a small decrease, while a high variance would result in a large decrease. For example, given a mean cosine distance of 0.9, with a 0.9 standard deviation, we know that there are several values below the mean and that, since we are using a normalized cosine distance, its maximum is one. Thus, it makes sense that we would decrease (negative sign) the diversity score with the intuition that a subset of documents would be less diverse among themselves than average. On the other hand, for a mean cosine distance of 0.1, it would only make sense to increase (positive sign) the value based on the standard deviation. To determine sign, we took advantage of a logistic distribution (Equation 2), centered on $\mu = 0.5$ and scaled to $s = 1/50$. We used this as a sign function by shifting the result by -0.5 and multiplying by 2, which gave us a value in the interval $[-1, 1]$ with a sigmoidal behavior. We then combined the mean and standard deviation to obtain the absolute value of increase or decrease, and multiplied it by the sign function.

We repeated this process for news, blogs, and the concatenated n-gram daily vectors of both corpora, for an overall topic diversity measurement. This resulted in a diversity score between zero and one, where zero meant that all the topics were exactly the same, while one meant that all the topics were completely distinct. Based on our results, topics have, overall for the combined samples, a diversity score of 0.970, a value that is as high as 0.986 for blogs, and as low as 0.976 for news. Topic diversity is similarly high in either case, despite blogs having a slightly higher diversity score.

5.3.3 Exploring Diversity Over Time

We also measured topic diversity over time, for small temporal windows, comparing news and blogs. Figure 4 shows the resulting diversity score for a sequence of 5-day windows starting at the given date (x-axis), from Sep 1st to Sep 30th 2015, with news in green and blogs in red. As we can see, both corpora have a diversity behavior that is similar over time, with the exception of the temporal windows from Sep 15th to Sep 19th 2015. Correlation between the two diversity score distributions is 28.9% for the whole month of September, but raises to 69.3% when ignoring the period of 15–19 Sep. We calculated the differences between diversity scores over time and found that the temporal window starting at Sep 19th 2015 represented the largest break in consistency between news and blogs, with a difference in diversity of 0.205.

We analyzed the n-grams of the topics, for each corpus, within this temporal window. For the news corpus, we found 111 unique n-grams out of 175 total n-grams, meaning that 63.43% of the n-grams are unique, which indicates a high diversity. On the other hand, for the blog corpus, we found 64 unique n-grams out of 164 total n-grams, meaning that 39.02% of the n-grams are unique, which indicates a low diversity. This is consistent with our diversity score. We also calculated the Jaccard index for the set of n-grams of each corpora, for the Sep 19th 2015 temporal window, finding that 15.89% of the total number of unique n-grams appears in both news and blogs.

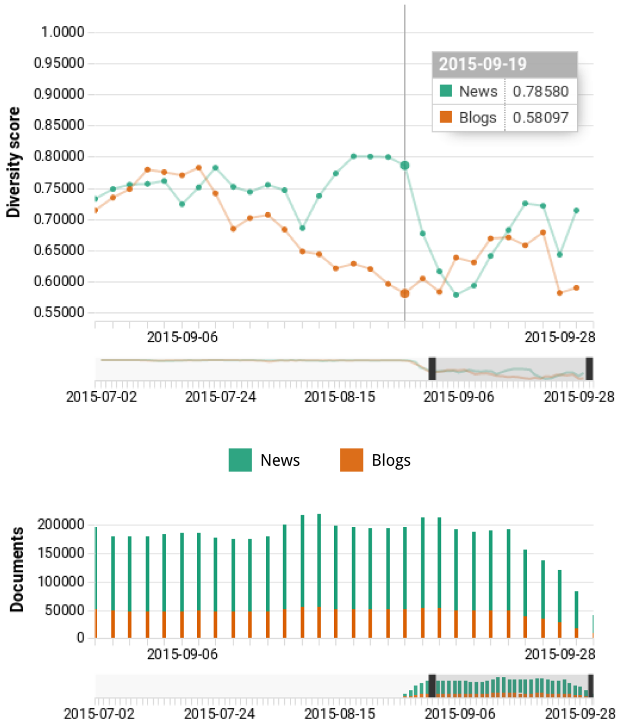


Figure 4: MediaViz diversity explorer. Top: diversity over time for windows of 5 days, starting at the given date. Bottom: number of documents for windows of 5 days, starting at the given date.

6 Conclusion

In this paper we presented the exploration of the Signal 1M Dataset, which comprises a large collection of news and blog articles, using distinct visualization tools. The visual analysis of the corpus provided interesting perspectives that would be much more difficult to obtain without the assistance of such tools. The Keywords tool allowed us to see that news and blog sources granted different levels of importance to a given set of keywords related with major global events that took place on September 2015. It was also evident, using the Sources tool, that the temporal publication patterns of these two media behaved differently — blogs published a higher percentage of content during the weekend than news sources —, but also in a similar fashion — both sources followed an identical curve during a 24-hour cycle. Finally, through the Diversity Explorer tool, we were able to visualize variations in the dynamics of topical diversity over time for each media type's content stream.

Acknowledgements

Project 'NORTE-01-0145-FEDER-000020' is financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF).

References

[DNR15] Tiago Devezas, Sérgio Nunes, and María Teresa Rodríguez. MediaViz: An interactive visualization platform for online media studies. In *Proceedings of the 2015 International Workshop on Human-centric Independent Computing*, pages 7–11. ACM, 2015.

[GLYR07] Mohammad Ghoniem, Dongning Luo, Jing Yang, and William Ribarsky. Newslab: Exploratory broadcast news video analysis. In *Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on*, pages 123–130. IEEE, 2007.

[HHN02] Susan Havre, Beth Hetzler, and Lucy Nowell. Themerivertm: In search of trends, patterns, and relationships. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):9–20, 2002.

[HW79] J A Hartigan and M A Wong. A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society*, 28(1):100–108, 1979.

[KBK11] Miloš Krstajić, Enrico Bertini, and Daniel A Keim. Cloudlines: Compact display of event episodes in multiple time-series. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2432–2439, 2011.

[KBMK10] Miloš Krstajić, Enrico Bertini, Florian Mansmann, and Daniel A Keim. Visual analysis of news streams with article threads. In *Proceedings of the First International Workshop on Novel Data Stream Pattern Mining Techniques*, pages 39–46. ACM, 2010.

[Kei01] Daniel A Keim. Visual exploration of large data sets. *Communications of the ACM*, 44(8):38–44, 2001.

[KNAMK13] Miloš Krstajić, Mohammad Najm-Araghi, Florian Mansmann, and Daniel A Keim. Story tracker: Incremental visual text analytics of news story development. *Information Visualization*, 12(3-4):308–323, 2013.

[NGSM15] Petra Kralj Novak, Miha Grcar, Borut Sluban, and Igor Mozetic. Analysis of financial news with newsstream, technical report IJS-DP-11965. *CoRR*, abs/1508.00027, 2015.

[Rou87] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

Exploiting News to Categorize Tweets: Quantifying The Impact of Different News Collections

Marco Pavan
University of Udine, Udine, Italy
marco.pavan@uniud.it

Stefano Mizzaro
University of Udine, Udine, Italy
mizzaro@uniud.it

Matteo Bernardon
University of Udine, Udine, Italy
matteo.bernardon@gmail.com

Ivan Scagnetto
University of Udine, Udine, Italy
ivan.scagnetto@uniud.it

Abstract

Short texts, due to their nature which makes them full of abbreviations and new coined acronyms, are not easy to classify. Text enrichment is emerging in the literature as a potentially useful tool. This paper is a part of a longer term research that aims at understanding the effectiveness of tweet enrichment by means of news, instead of the whole web as a knowledge source. Since the choice of a news collection may contribute to produce very different outcomes in the enrichment process, we compare the impact of three features of such collections: *volume*, *variety*, and *freshness*. We show that all three features have a significant impact on categorization accuracy.

1 Introduction

Social Network contents are analyzed for several purposes: identifying trends [MK10], categorizing and filtering news [JG13, SSTW14], measuring their importance, spread etc. [NGKA11]. Other researchers try to categorize short texts posted on social networks (e.g., tweets), using contents taken from the WWW, to understand user interests, to build user models etc. However, platforms like Twitter limit the text length, and users tend to use abbreviations and acronyms to write even faster. In a lot of cases the posted texts have a very low number of characters[1]; therefore, an automatic categorization process with topic extraction methodologies could be not enough reliable. In these cases, exploiting an additional source of information could help, providing additional text to analyze. Since short texts posted by users are often related to recent events (sharing their opinions and thoughts with friends), our approach is to use news collections instead of generic web contents in the categorization process.

On this basis, we study how the choice of the news collection affects the results: in particular, how different news collections with different properties impact the categorization effectiveness. More specifically, we analyze, by means of three experiments, three features of news collections: (i) *Volume*, to see how different numbers of news provide different sets of terms for the enrichment phase and, consequently, affect the categorizations; (ii) *Variety*, to see how news of different nature impact the enrichment process; and (iii) *Freshness*, to highlight the different effectiveness by using news from different time windows (i.e., same temporal context, 1 year old, 2 years old etc.). We exploit the methodology proposed in [MPSV14], based on a text enrichment with new set of words, extracted from news on webpages of the same temporal context,[2] and a categorization by querying the Wikipedia category tree as external knowledge base.

2 Related work

All the works in the literature addressing the problem of classifying tweets recognize that "data sparseness" and ambiguity represent a serious issue. For instance,

In: M. Martinez, U. Kruschwitz, G. Kazai, D. Corney, F. Hopfgartner, R. Campos and D. Albakour (eds.): Proceedings of the NewsIR'16 Workshop at ECIR, Padua, Italy, 20-March-2016, published at http://ceur-ws.org

[1] Several surveys show that the mode of characters is 28 [twi16a].

[2] A set of news published in the same period of the short text.

in [HH15] the authors use the “bag-of-words” approach, adopting dimensionality reduction techniques, to reduce accuracy and performance problems.

In [AGHT11] the authors introduce several enrichment strategies (i.e., entity-based, topic-based, tweet-based and news-based) to relate tweets and news articles belonging to the same temporal context, in order to assign a semantic meaning to short messages. In [YPF10] another enrichment-based approach is proposed to classify generic online text documents, by adding a semantic context and structure, using Wikipedia as a knowledge source. In [GLJD13] the authors define a framework to enrich and relate Twitter feeds to other tweets and news speaking about the same topics. Hashtags (for tweets) and *named entities* (for news) are used to achieve such goal. A cluster-based representation enrichment method (CREST) is introduced in [DSL13]: such system enriches short texts by incorporating a vector of topical relevances (besides the commonly adopted tf-idf representation). Finally, topics are extracted using a hierarchical clustering algorithm with purity control. Enrichment techniques can also be quite sophisticated like, e.g., in [WZX+14] where a short texts are classified exploiting link analysis on topic-keyword graphs. In particular, after the initial topic modeling phase, each topic is associated to a set of related keywords. Afterwards, link analysis on a subsequent topic-keyword bipartite graph is carried out, to select the keywords most related to the analyzed short text.

Machine learning can play a fundamental role in classifying short texts: for instance, in [DDZC13] supervised SVM (Support Vector Machine) techniques are used to classify tweets into 12 predefined groups tailored for the online community of Sri Lanka. In [ZCH15] a completely automatized unsupervised bayesian model is used. In particular only tweets related to events are selected, exploiting a lexicon built from news articles published in the same period.

So far, it is clear that the problem of classifying short texts (whatever the related semantic domain) must rely on some forms of background knowledge, to fill the gaps and lack of information of the original messages. Such knowledge base can be found in external semantic platforms like, e.g., Wikipedia (as in some of the above mentioned works, and in the INEX Tweet Contextualization Track [ine13]), the WWW or other, possibly more focused, archives/structures. Hence, it is of utmost importance to study how the choice of the external collection influences the accuracy of the short text categorization process.

3 Features of News Collections

To run a set of experiments to analyze the collections features, we use two different open source document collections, which differ in number and kind of documents included, have different sizes, span from 2011 to 2013, and also have some temporal overlaps to allow several comparisons. They are shown in Table 1 and allow us to analyze the following three key features:

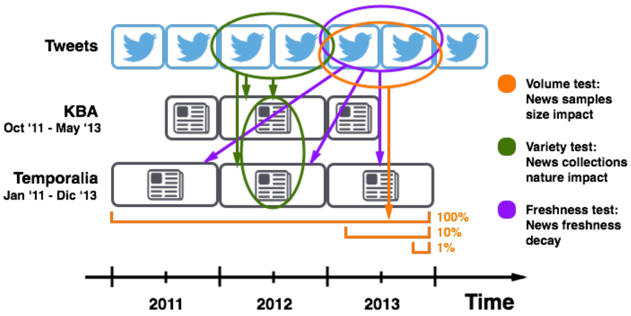


Figure 1: News collections distribution with features based tests

- *Volume:* we want to see the impact of news samples with different cardinality, extracted from the same collection in different percentages. With this test we aim to measure how the amount increment correlates to the final enrichment effectiveness.
- *Variety:* news are often different in nature, such as texts from blogs, forums, online newspapers etc., and different variety of texts could have different impact on the text enrichment. We want to measure how the news variety affects the results.
- *Freshness:* short texts are often related to recent events, therefore, it is interesting to study how important is to have the publishing time of the news close to the publishing time of the short text being enriched, and how the enrichment effectiveness changes using increasingly older news.

Figure 1 shows a representation of the two collections distributed over time and tweets as short texts to analyze. The *Volume test*, highlighted in orange, aims to compare the categorization results with samples of news from the same collection but with different sizes; the *Variety test*, in green, compares results among news samples with same cardinality but with different kinds of news; and the *Freshness test*, in purple, exploits news from the same collection but in different years. The figure shows only some examples; the details of all the experiments are described in the next section.

4 Experimental evaluation

4.1 Experimental design

To evaluate the impact of each news collection on the categorization process we selected a set of 5 popular Twitter account famous in different fields. In particular, David Cameron (@*David_Cameron*)

Table 1: The two news collections used in the experiments

Acronym	Name	# of docs/ size	kind of docs	Timespan
Temporalia	NTCIR Temporal Information Access 2012[a]	~2M / ~20GB	blogs news	Jan2011 – Dec2013
KBA	Knowledge Base Acceleration 2012[b]	~20M / ~930GB[c]	blogs, news, forums, social	Oct2011 – May2013

[a] http://ntcirtemporalia.github.io/NTCIR-12/collection.html

[b] http://trec-kba.org/

[c] Data extracted from the 3rd stream corpora http://s3.amazonaws.com/aws-publicdatasets/trec/kba/index.html

for Politics, Harry Kane (@*HKane*) for Sport, Bill Gates (@*BillGates*) for Technology, Neil Patrick Harris (@*ActuallyNPH*) for Cinema and Rihanna (@*rihanna*) for Music. We extracted a set of tweets from each account in a specific time window, according to the test we planned to run, in order to have a sufficient amount of short texts to enrich and categorize. We used a Python wrapper [pyt16] around the official Twitter API [twi16b] to retrieve tweets. We repeated this process to have a sample of 1000 tweets for each test which involves a large temporal window (e.g., six months or one year). Instead, for tests focused on one month, we built samples of 250 tweets. We then defined the benchmarks as follows in the next sections.

4.1.1 Volume test

To measure the impact of collections volume we defined 2 tests, "Test 1a" based on Temporalia and "Test 1b" on KBA. We analyzed samples using news subsets with different cardinality. With these tests we can see how changing the amount of news affects the results, and also if the results will generalize across different collections. The 2 tests are defined as follows:

Test 1a: Tweets posted in whole 2013, categorized with Temporalia 1%, Temporalia 10% and Temporalia 100%.

Test 1b: Tweets posted in whole 2013, categorized with KBA 1%, KBA 10% and KBA 100%.

4.1.2 Variety test

We defined "Test 2a" and "Test 2b" to measure how the variety of news inside a collection could impact the enrichment phase and consequently the categorization process. We selected news samples with the same cardinality from different collections and from different time windows, in order to see the effects of changing news varieties, and also if on a wider time window of 6 months we have the same effects we get on only 1 month. The 2 tests are defined as follows:

Test 2a: Tweets posted in January 2013, categorized with Temporalia Jan 2013 (60K news sample), KBA Jan 2013 (60K news sample) and Temporalia+KBA Jan 2013 (30K+30K news sample).

Test 2b: Tweets posted in the second half of 2012, categorized with Temporalia Jul-Dec 2012 (400K news sample), KBA Jul-Dec 2012 (400K news sample) and Temporalia+KBA Jul-Dec 2012 (200K+ 200K news sample).

4.1.3 Freshness test

To benchmark how the news freshness is important we defined 3 tests, "Test 3a", "Test 3b", based on different news "aging", and "Test 3c", based on a different collection. For the first test we want to see the difference between enriching the tweets with news extracted from the same temporal context (i.e., at most 1 month before the publishing date) and news in the same year of publishing (i.e., at most 1 year before the publishing date). In the second test we want to extend this analysis to more than 1 year before the publishing date, in particular we benchmark the results using news related to event of the same year of the tweets, 1 year old and 2 years old. The third test aims to compare the same "aging effect" with a different collection. The 3 tests are defined as follows:

Test 3a: Tweets posted in whole 2013, categorized with Temporalia 2013 - *contextualized*[3] and Temporalia Jan 2013 (both samples are composed of 60K news).

Test 3b: Tweets posted in whole 2013, categorized with Temporalia 2013, Temporalia 2012 and Temporalia 2011 (both samples are composed of 90K news).

Test 3c: Tweets posted in whole 2012, categorized with KBA 2012 - *contextualized*, KBA Jan 2012 and KBA 2012 (both samples are composed of 100K news).

4.2 Measures

To evaluate the experiments and to benchmark the collections effectiveness we carried out an expert evaluation to assess each analyzed feature over short texts samples composed of either all tweets for one month based tests (250) or a set of 250 randomly extracted tweets for tests based on larger temporal windows.

We used a categorization prototype system [MPSV14] for the categorization of short texts which

[3] Only news from the same month when the tweet has been posted.

provides, as final outcome, a list of labels extracted from Wikipedia category tree. The system includes a module which analyzes text, searches related documents into a news collection, and extracts a set of words used to enrich the original short text.

The texts have been submitted to the categorization system with different news collections according to the three tests described in Section 4.1. For each test, in order to assess the news impact over the enrichment process, the set of categories yielded by the system has been evaluated by expert users. The latter assigned a rating, i.e., a number between 1 and 5 (1=lowest value, 5=highest value) indicating how the categories properly represent the topic discussed in the tweet.

In particular for the Volume test, we run the evaluation several times, with news samples randomly rebuilt each time, where we used only a portion of the entire collection. We kept the average ratings obtained with different sub-collections, avoiding bias due to the random set of news. Specifically for samples with 10% or 1% of news we run respectively the evaluation 3 or 5 times, approximating the average ratings to the nearest integer value.

4.3 Results

Results are reported in the following charts, which show distribution functions of ratings obtained by each test with the different experiment settings. In particular, we display the cumulative distribution function (CDF), the inverted complementary cumulative distribution function (I-CCDF), and a table reporting the mean ratings. The I-CCDF is provided for an easier reading, showing the data in ascending order and thus highlighting the news collection performing better as the line at the top of the chart.

4.3.1 Volume Test

Figure 2 shows the results related to Test 1a and 1b, highlighting how for both collections the number of news is an important feature to consider. We can observe a noticeable improvement with Temporalia 100% compared to smaller samples. Increasing the volume allows us to include a large number of both relevant and not relevant news: the first ones yield a global improvement, while the second ones have a low overall impact. The general improvement is also confirmed by the Wilcoxon test. Then, we notice only a slight difference between Temporalia 1% and 10%, where the news increase in number from an order of magnitude 10K to 100K. The Wilcoxon test, over the latter couple of rating distributions, confirmed a non statistically significant difference between those samples, with a p-value>0.05. On the other hand, with KBA we already have a noticeable difference between KBA 1% and KBA 10%, due to order of magnitude from 100K to 1M, and even better using KBA 100% (10M). This fact emphasizes how increasing the sample sizes has considerable effects on the results only when a certain amount of news is reached. The diverse impact of Temporalia and KBA is probably also due to other factors than the only difference in size. Of course the same percentage, applied to collections with very different sizes, yields sets of extracted documents whose cardinality is very different; whence we can also expect a different variety of such sets. Moreover, for instance, KBA does not fully cover year 2013, whence the effectiveness could be affected by the publishing date of the analyzed short texts. Such aspects are taken into consideration in the remaining experiments.

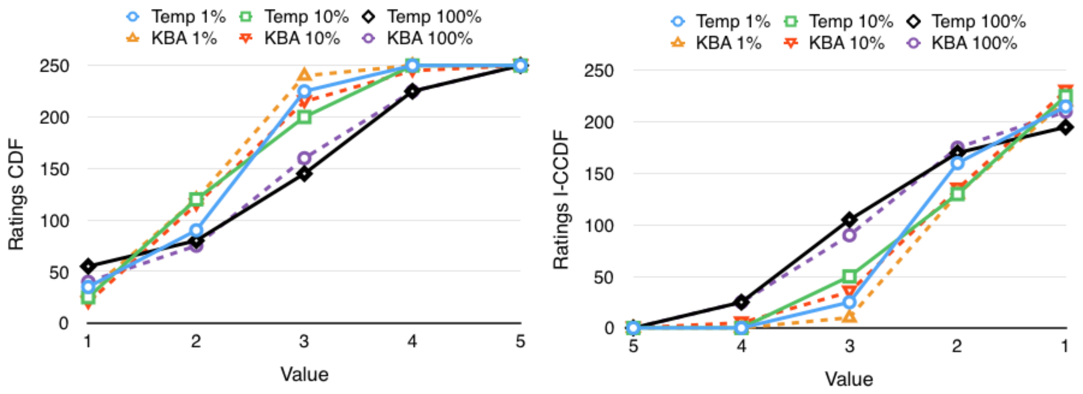
4.3.2 Variety Test

Figure 3 shows how the variety of news inside the analyzed samples affects the enrichment effectiveness. Continuous lines represent the results over 1 month of news (Test 2a), and dotted lines over 6 months (Test 2b). For both experiments there is a noticeable difference among the samples which highlights how increasing the variety of news allows to improve the final categorization also on different time windows. The Wilcoxon test over the sample pairs of each test confirms the statistically significant difference between all the rating distributions. This fact highlights how important is to increase the variety of news in order to improve the set of words to use as text enrichment.

4.3.3 Freshness Test

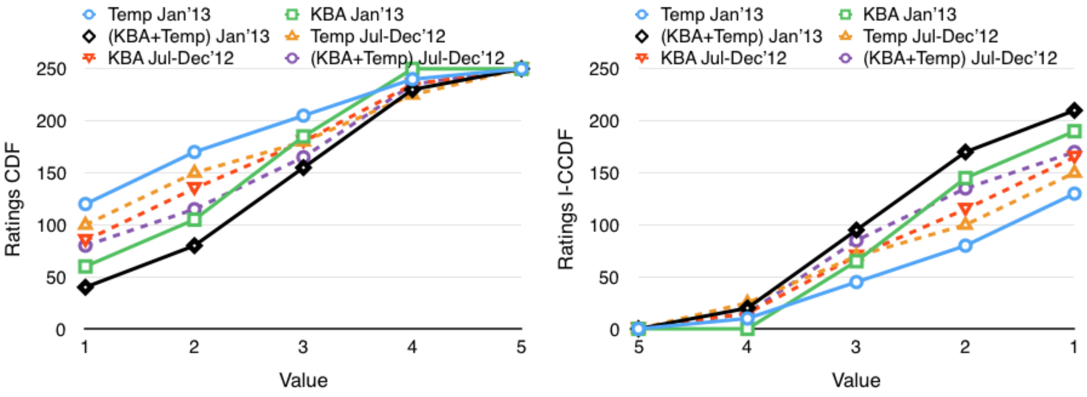
The chart in Figure 4 shows the results related to Test 3a, 3b and 3c, and it is possible to notice how the news freshness affected the results especially when the news get older. Collections with contextualized news got the best effectiveness due to the news publishing time close to the tweets (same month), therefore they allow to have more relevant additional text to exploit. The system has worsened the categorization process with tweets randomly selected from whole 2013, and using collections of news extracted from the same year, either equally distributed over all months or only in January. The effectiveness decreases drastically when the news get older in previous years. In particular we can notice how we got the same lowest effectiveness with Temporalia 2012 and Temporalia 2011, highlighting how 1 (or more) year old news are poor of information for these purposes.

Test 3a results, related to Temporalia 2013, show how large is the difference between news distant only some months in time, and Test 3b results, where we analyzed three years of Temporalia news, highlight how going back to 1 year is crucial for the categorization



Mean Ratings	
Temp 1%	2.60
Temp 10%	2.62
Temp 100%	2.98
KBA 1%	2.44
KBA 10%	2.62
KBA 100%	3.00

Figure 2: Volume impact CDF, I-CCDF, and mean ratings



Mean Ratings	
Temp Jan'13	2.06
KBA Jan'13	2.60
(KBA+Temp) Jan'13	2.98
Temp Jul-Dec'12	2.38
KBA Jul-Dec'12	2.46
(KBA+Temp) Jul-Dec'12	2.62

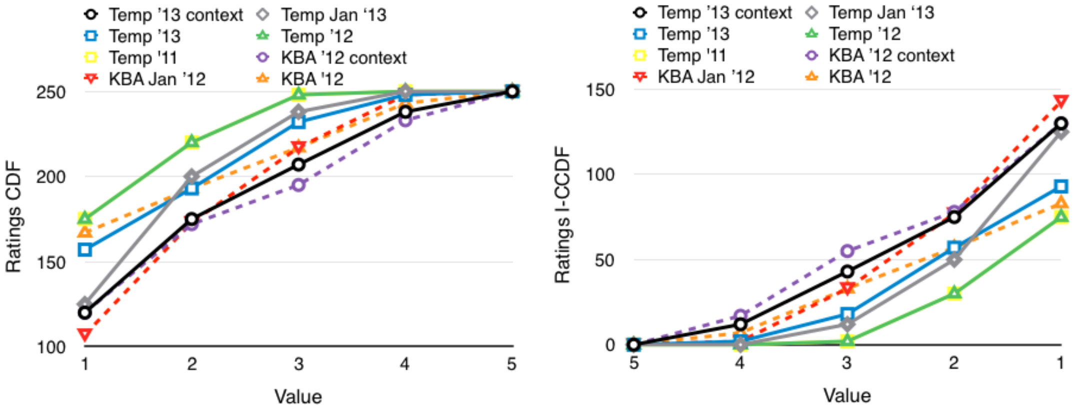
Figure 3: Variety impact CDF, I-CCDF, and mean ratings

process. With KBA collections we can notice how the results are similar and the rating distributions, represented by dotted lines, highlight better effectiveness with higher news freshness. Wilcoxon tests confirm that there is statistical significant difference among the rating distributions in both Temporalia and KBA, except for Temporalia '11 and '12 which obviously have equal values. This is a further confirmation that few months old news have a strong impact as those from previous years.

5 Discussion and Conclusions

The experiments performed in this work have demonstrated that text enrichment is sensibly affected by the features of the news collections that we have analyzed. More precisely, there is a critical threshold for what concerns the collection Volume, that allows to have a sufficient amount of news to reach a good level of effectiveness. Moreover, such threshold seems to be dependent on the whole size of the collection taken into consideration. Our benchmarks confirm the importance of news variety, highlighting how increasing the number of available kinds yields a better enrichment both for texts selected in one month and in the larger time window. The news Freshness appears to be a sensible feature since news published close to the same period of the short text provide a better set of terms to use in the enrichment phase. Indeed, as soon as the news begin to age (even of just a few months) the effectiveness of the categorization drastically decreases.

For future work, we plan to refine and complete the experiments on the three focused features. For instance, it could be interesting to look at the impact of the number of documents extracted from the news collection and used to categorize short texts. As we pointed out in Section 4.3, a larger database will produce a higher number of elements (with the same percentage), and this fact can have subtle implications on the final outcomes. We also plan to carry on further experiments about the variety, investigating which kinds of news it is important to include in the collection, and which ones are marginal. As the freshness is concerned, we could investigate more precisely, varying the granularity of the time windows, which is the temporal threshold causing a quick decrease of the effectiveness of the enrichment process. Moreover, we plan to carry on further experiments on



Mean Ratings	
Temp '13 context	2.04
Temp Jan '13	1.75
Temp '13	1.68
Temp '12	1.43
Temp '11	1.43
KBA '12 context	2.12
KBA Jan '12	2.02
KBA '12	1.72

Figure 4: Freshness impact CDF, I-CCDF, and mean ratings

different news collections and new kinds of short texts (e.g., instant chat messages, online comments). Unfortunately we could not use the Signal Media collection available at `http://research.signalmedia.co/newsir16/signal-dataset.html`; indeed, a collection covering a one-month period is not sufficient for the kind of experiments we described in this paper (think, e.g., of the freshness test).

References

[AGHT11] Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. Semantic enrichment of twitter posts for user profile construction on the social web. In *The Semanic Web: Research and Applications*, pages 375–389. Springer, 2011.

[DDZC13] Inoshika Dilrukshi, Kasun De Zoysa, and Amitha Caldera. Twitter news classification using SVM. In *Proc. of ICCSE'13*, pages 287–291. IEEE, 2013.

[DSL13] Zichao Dai, Aixin Sun, and Xu-Ying Liu. Crest: Cluster-based representation enrichment for short text classification. In *Advances in Knowledge Discovery and Data Mining*, pages 256–267. Springer, 2013.

[GLJD13] Weiwei Guo, Hao Li, Heng Ji, and Mona T Diab. Linking tweets to news: A framework to enrich short text data in social media. In *ACL (1)*, pages 239–249, 2013.

[HH15] Yin-Fu Huang and Chen-Ting Huang. Mining domain information from social contents based on news categories. In *Proc. of IDEAS'15*, pages 186–191. ACM, 2015.

[ine13] INEX 2013 Tweet Contextualization Track. `http://inex.mmci.uni-saarland.de/tracks/qa/`, 2013.

[JG13] Nirmal Jonnalagedda and Susan Gauch. Personalized News Recommendation Using Twitter. In *Proc. of WI-IAT'13*, pages 21–25. IEEE Computer Society, 2013.

[MK10] Michael Mathioudakis and Nick Koudas. Twittermonitor: trend detection over the Twitter stream. In *Proc. of ACM SIGMOD'10*, pages 1155–1158. ACM, 2010.

[MPSV14] S. Mizzaro, M. Pavan, I. Scagnetto, and M. Valenti. Short text categorization exploiting contextual enrichment and external knowledge. In *SIGIR '14 Proceedings*. SoMeRA, SIGIR, July 2014.

[NGKA11] Nasir Naveed, Thomas Gottron, Jérôme Kunegis, and Arifah Che Alhadi. Bad news travel fast: A content-based analysis of interestingness on Twitter. In *Proc. of WebSci'11*, page 8. ACM, 2011.

[pyt16] Python wrapper around the Twitter API. `https://dev.twitter.com/rest/public`, 2016.

[SSTW14] Timm O Sprenger, Philipp G Sandner, Andranik Tumasjan, and Isabell M Welpe. News or noise? using twitter to identify and understand company-specific news flow. *Journal of Business Finance & Accounting*, 41(7-8):791–830, 2014.

[twi16a] The Next Web. `http://thenextweb.com/twitter/2012/01/07/interesting-fact-most-tweets-posted-are-approximately-30-characters-long/#gref`, 2016. [Online, visited Feb-2016].

[twi16b] Twitter REST APIs. `https://dev.twitter.com/rest/public`, 2016.

[WZX+14] Peng Wang, Heng Zhang, Bo Xu, Chenglin Liu, and Hongwei Hao. Short text feature enrichment using link analysis on topic-keyword graph. In *Natural Language Processing and Chinese Computing*, pages 79–90. Springer, 2014.

[YPF10] Hiroki Yamakawa, Jing Peng, and Anna Feldman. Semantic enrichment of text representation with wikipedia for text classification. In *Proc. of SMC'10*, pages 4333–4340. IEEE, 2010.

[ZCH15] Deyu Zhou, Liangyu Chen, and Yulan He. An unsupervised framework of exploring events on twitter: Filtering, extraction and categorization. In *Proc. of AAAI'15*, 2015.

Visualising the Propagation of News on the Web

Svitlana Vakulenko*, Max Göbel†, Arno Scharl* and Lyndon Nixon*

* MODUL University Vienna
† Vienna University of Economics and Business
Vienna, Austria
{svitlana.vakulenko,arno.scharl,lyndon.nixon}@modul.ac.at
max.goebel@wu.ac.at

Abstract

When newsworthy events occur, information quickly spreads across the Web, along official news outlets as well as across social media platforms. Information diffusion models can help to uncover the path of an emerging news story across these channels, and thereby shed light on how these channels interact. The presented work enables journalists and other stakeholders to trace back the distribution process of news stories, and to identify their origin as well as central information hubs who have amplified their dissemination.

1 Introduction

Newsworthy events are communicated via traditional news media sources such as CNN and the New York Times, as well as social media platforms. However, the specific path a story takes via various news distributors and the interplay with the social network discussion is not well studied yet. This limits further research on rumour detection and news content verification. This paper presents an approach developed in the EU-funded Pheme project (www.pheme.eu), tracking information contagions across various media sources including major online news publishers as well as single Twitter users.

2 Related Work

Information diffusion is an established research field traditionally applied to explicit networks such as social media, but less studied in communication scenarios where information sources tend to be implicit.

One research area that links news articles to trace the origin of an information piece is text reuse (plagiarism) detection. This approach has been recently applied to analyse information exchange networks based on historical newspaper texts [CIK14] and to study the evolution of memes [SHE$^+$13]. In contrast to this work, our approach does not track stable phrases, but uses information pieces directly as relations.

Yang and Leskovec [YL10] model the total number of infected nodes over time determined by the influence function of nodes infected in the past. They formulate this problem as an instance of *Non-Negative Least Squares* and use it to predict the volume of information diffusion in the future. Their approach differs from ours since it does not model implicit network to surface implicit links between the information sources.

3 Information Diffusion Model

3.1 Modeling Information Contagions

We propose a 'bag-of-relations' document representation model to capture the essential information contained in textual documents, such as news articles. The main idea behind our approach is to represent each document as a set of relations, represented as n-grams-like similarity strings. Unlike n-grams, these strings are constructed from grammatical dependency relations instead of the sequential order of words in a sentence. We employ a dependency parser to obtain parse trees for each of the sentences and extract the relations by traversing these trees. The relations are then modeled as triples of the form:

$$\mathbf{s}\text{ (subject)} - \mathbf{p}\text{ (predicate)} - \mathbf{o}\text{ (object)}$$

In: M. Martinez, U. Kruschwitz, G. Kazai, D. Corney, F., R.and D. Albakour (eds.): Proceedings of the NewsIR'16 Workshop at ECIR, Padua, Italy, 20-March-2016, published at http://ceur-ws.org

We start off with the task of finding all the predicates in the sentence, which play the role of triggers to finding the corresponding relations. We normalize the predicates to the form: '{*synsets (or lemmas)*} + {*flags*}', by detecting for each verb the corresponding WordNet synset (or taking the verb's lemma otherwise), tense, voice, negation and auxiliary verbs (e.g. 'did not say' is transformed to 'state D N').

We define a set of words to be excluded from the predicate phrase to improve the results. For example, there are trivial relations, which are common among all news articles and which we would like to eliminate, e.g. the ones triggered by the predicates: 'print', 'post', 'update'. Words that do not carry any semantic information of the predicate, but are used solely for grammatical purposes (e.g. 'will', 'do'), are also excluded.

We introduce special symbols to preserve the grammatical information removed at the previous step. As such, D indicates the past tense, F – future tense, N – negation, A - auxiliary verb ('would'). Since there are multiple ways to express negation or past tense, this approach allows to disambiguate and group together semantically-equivalent relations.

Then, for each predicate we pick the adjacent branches with clauses that correspond to the subject and objects of the relation. We designed a simple heuristic for English language texts: assign the node to the subject-element if it precedes the predicate in the sentence, and to the object otherwise (i.e. when it follows the predicate).

We construct separate relations for each object-element related to the predicate and one relation with an empty object, if the subject is not empty. This simple heuristic allows us to create several fine-grained relations with different levels of detail. For example, a sentence "The plane landed in Panama on Tuesday" will be decomposed into: 'plane - land D', 'plane - land D - in Panama', 'plane - land D - on Tuesday'. This approach enables us to spot those articles that report on the same event but provide complementing or contradicting details.

3.2 Modeling Diffusion Cascades

We assume that all articles sharing the same information contagion are related to each other, i.e there is a path for every pair of articles within the diffusion graph. We included this assumption into our model by enforcing the connectivity requirement over our diffusion graph: for each node (except the root node), we generate an incoming edge that links the node to its source. Here, we also use the single source assumption: for all nodes (except the root node), there is exactly one incoming edge linking the node to its source (the closest neighbour). This assumption allows us to simplify the model and avoid making assumptions about the similarity threshold value, i.e. how similar the articles should be to be linked in the diffusion model.

The diffusion process is modeled as a graph with two types of edges: (1) explicit links referencing the source URL - edge direction: from the source to the post with the URL; (2) implicit links to connect similar posts that share the same information contagions - edge direction: from the older to the more recent post.

We link news articles to social media posts by querying the Twitter API with the URL of a news article to obtain all the tweets which reference it explicitly. News media often do not cite their information sources apart from the references to the major news agencies, e.g. Reuters. Therefore, we focus on uncovering the latent relations between the news articles, which we construct based on content similarity. We construct the diffusion graph with edges generated using the pairwise similarity values computed over the relation bags of the articles.

There are two methods to compute similarity between a pair of news articles: (1) considering the intersection of the relation bags, (2) hashing the relation bags and computing the similarity between the relation hashes. While the first method, returning an integer for the number of shared relations, is simple and intuitive, it is limited to considering only exact matches between relations. The second method is more powerful by allowing for approximately similar relations.

We test both methods to compute similarity for any two relation bags complementary to each other to evaluate which of them performs better in practice. We use Nilsimsa hashing and Hamming distance to generate and compare the relation hashes. Nilsimsa is one of the most popular algorithms for locality-sensitive hashing and is traditionally employed for spam detection in emails. Hamming distance measures the proportion of positions in the hash at which the corresponding symbols are different.

4 Experiment

4.1 Dataset and Configuration

The dataset is based on a recent news media snapshot exported from the PHEME dashboard [SWG+16], which contains 71,000 articles published between 27 November and 3 December 2015. We ran the relation extraction procedure on this corpus and picked one of the frequent information contagions to illustrate how it can be backtracked across the online media:

s: president barack obama – **p**: state D – **o**:

This relation provided us with a cluster of 12 news articles. It is able to capture all the expressions with

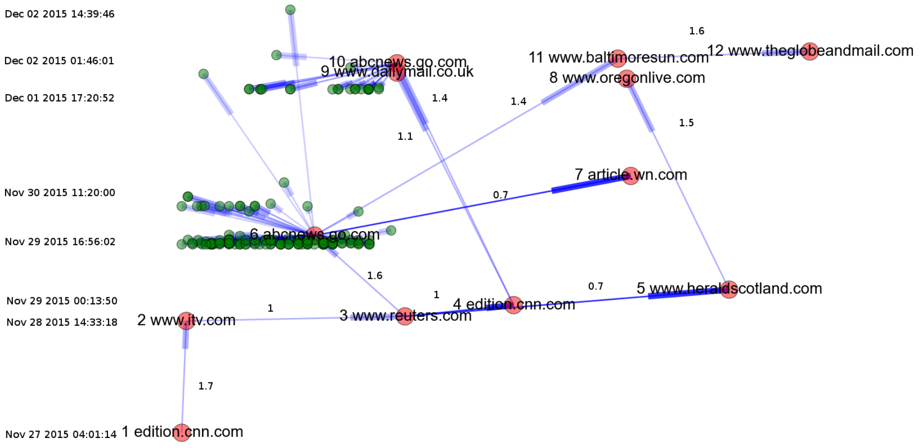


Figure 1: Sample information diffusion model: ‘president barack obama state D’

a predicate that belongs to the WordNet synset ‘state’ and is used in the past tense (‘D’), such as “president barack obama said”, thereby indicating statements made by President Obama.

For each article we retrieved the tweets via its URL using Twitter Search API, which resulted in 150 tweets (127 and 23 for two of the articles). We used the networkx[1] and matplotlib[2] Python libraries to visualize the resulting diffusion graph (see Figure 1).

4.2 Results

The nodes of the graph in Figure 1 represent the individual posts published at discrete time intervals (red: news articles; green: tweets). The sources get infected in the sequence order aligned along the vertical time axis as indicated on the node labels. The same source may appear more than once within the same network if it has published multiple articles containing the same information contagion within the given time interval.

The edges of the graph represent *direct links* in case of tweets, or *content similarity* in case of articles. Content similarity values are indicated as weights over the corresponding edges. Values closer to 0 indicate more similar articles. *Light edges* indicate that the adjacent articles share a single information contagion, while *solid edges* indicate that the articles have more than one information contagion in common.

5 Conclusion and Future Work

We showed how to uncover the latent relations between news articles and used them to infer a model of the implicit diffusion network, which constitute an important step towards rumour detection research. The results of our initial experiment indicate that our relation-based modeling approach is promising and merits further research. In future work we will further evaluate our approach against baseline methods.

6 Acknowledgments

The presented work was conducted within the PHEME Project (www.pheme.eu), which has received funding from the European Union’s Seventh Framework Programme for Research, Technological Development and Demonstration under Grant Agreement No. 611233.

References

[CIK14] Giovanni Colavizza, Mario Infelise, and Frederic Kaplan. Mapping the Early Modern News Flow: An Enquiry by Robust Text Reuse Detection. In *Social Informatics*, pages 244–253, 2014.

[SHE+13] Caroline Suen, Sandy Huang, Chantat Eksombatchai, Rok Sosic, and Jure Leskovec. NIFTY: A System for Large Scale Information Flow Tracking and Clustering. In *22nd International Conference on World Wide Web*, pages 1237–1248, 2013.

[SWG+16] Arno Scharl, Albert Weichselbraun, Max Göbel, Walter Rafelsberger, and Ruslan Kamolov. Scalable knowledge extraction and visualization for web intelligence. In *49th Hawaii International Conference on System Sciences*, pages 3749–3757, 2016.

[YL10] Jaewon Yang and Jure Leskovec. Modeling information diffusion in implicit networks. In *10th International Conference on Data Mining*, pages 599–608, 2010.

[1] networkx.github.io

[2] matplotlib.org

Comparative Analysis of GDELT Data Using the News Site Contrast System

Masaharu Yoshioka
Hokkaido University
N14 W9, Kita-ku, Sapporo-shi,
Hokkaido, 060-0814, Japan
yoshioka@ist.hokudai.ac.jp

Noriko Kando
National Institute of Informatics
2-1-2, Hitotsubashi, Chiyoda-ku,
Tokyo, 101-8430, Japan
kando@nii.ac.jp

Abstract

The News Site Contrast (NSContrast) system analyzes news articles retrieved from multiple news sites based on the concept of contrast set mining. It can extract terms that characterize different topics of interest across news sites, countries, and regions. In this study, we used NSContrast to analyze Global Database of Events, Language, and Tone (GDELT) data by comparing news articles from different regions (e.g., USA, Asia, and the Middle East). We also present examples of analyses performed using this system.

1 Introduction

It has become possible to access a wide variety of news sites from across the world via the Internet. Each news site has its own culture and interpretation of events, so we can obtain a greater diversity of information than ever before by using multiple news sites. Opinions and interests expressed in news articles vary across countries, and we can obtain different points of view regarding a topic if we access news sites from different countries. For example, Asian, European, and American news sites share some common views on diplomatic issues related to North Korea, as well as having their own characteristic opinions. Therefore, it is important to clarify the characteristics of each specific news site when analyzing events reported by multiple sites.

The News Site Contrast (NSContrast) system was developed to analyze the characteristics of news sites [YK12]. However, since it is not easy to construct news databases from different countries, NSContrast only uses small numbers of news sites from East Asian countries (Japan, China, Korea) and the USA to characterize the differences between them.

Recently, a Global Database of Events, Language, and Tone (GDELT) [LS13] [1] was released. This database is based on larger numbers of news sites from all over the world and it contains extracted metadata information from news articles. In this paper, we propose a method to utilize GDELT to analyze the characteristics of news article from different countries and regions by adding country and region information for the news sites in the database. By using these data, we can compare news articles from various countries and regions (e.g., USA, Asia, South America, and Africa) worldwide instead of our original small database of news articles. We also present examples of analyses performed using the NSContrast system.

2 NSContrast

2.1 System description

NSContrast employs the following four methods to analyze news articles.

- **Burst analysis** [Kle02] identifies the daily burst terms and the regional distribution of a specific bursty term. (Figure 1)
- A **term collocation analysis graph** shows relationships among collocated terms and the given query. NSContrast uses highly collocated terms from all regions based on contrast set mining and ordinal collocation analysis. These collocation terms are visualized with a spring model using fdp in Graphviz.[2].
- A **news article retrieval system** is used to understand the meanings of the terms in the collocation analysis and the burst analysis.
- A **multifaceted interface for analyzing news articles**.

The system uses multiple facets (e.g., keyword,

In: M. Martinez, U. Kruschwitz, G. Kazai, D. Corney, F. Hopfgartner, R. Campos, and D. Albakour (eds.): Proceedings of the NewsIR'16 Workshop at ECIR, Padua, Italy, 20-March-2016, published at http://ceur-ws.org

[1] http://www.gdeltproject.org/

[2] http://www.graphviz.org/

named entity, polarity, news site, and country) to analyze news articles. The interface supports the construction of structured queries that use one or more facets, where the facet information can be represented using various styles (e.g., time sequence graph, table, or bar chart). (Figure 2)

2.2 Data conversion

To apply NSContrast to the analysis of GDELT data, it was necessary to convert the GDELT data into news article data. There are two databases in GDELT: GDELT Event and GDELT Global Knowledge Graph (GKG). GDELT GKG is a database based on a raw output format of the original news articles for constructing the GDELT Event database. Because the GDELT Event database does not have detailed original news article sources, GDELT GKG was used for NSContrast.

GDELT GKG was constructed by extracting the following metadata information from the original news articles: DATE, THEMES, LOCATIONS, PERSONS, ORGANIZATIONS, TONE (as a real value; 0 means neutral), CAMEOEVENTIDS (references to the GDELT Event database), SOURCES, and SOURCEURLS. When there are two or more articles that share all name sets (THEMES, LOCATIONS, PERSONS, and ORGANIZATIONS), those news articles are aggregated as one datum and SOURCES and SOURCEURLS have multiple entries. Example of SOURCES and SOURCEURLS information for one datum in January 19, 2016 are shown below.

SOURCES punchng.com; punchng.com; onlinenigeria.com; onlinenigeria.com

SOURCEURLS http://www.punchng.com/25909-2/, http://www.punchng.com/i-am-resolved-to-better-lagos-ambode/, http://news2.onlinenigeria.com/news/general/453949-i-am-resolved-to-better-lagos-%E2%80%93w-ambode.html, http://news2.onlinenigeria.com/news/general/453949-i-am-resolved-to-better-lagos-ambode.html

Two types of multiple SOURCEURLS are shown above. In one, almost the same content has a different URL for the same news site (the first two URLs and the last two URLs above) and the other is a different URL with different news sites (the first and third URLs).

Most of the former cases are simply URL variations of the same content; e.g., the first URL is redirected to the second URL and the third URL is a variation of the fourth URL (the URL encoding of "%E2%80%93w" is "-" for UTF-8). It is better to select one of them for deduplication. The latter cases are meaningful for representing the importance of the contents, because different news sites have selected the same content for their sites.

By using these metadata, the following information was constructed for NSContrast.

Date Date of the article.

Person, Organization, Location Lists of people, organizations, and locations extracted from the article using the GDELT GKG.

Polarity We classified articles into three types (positive, negative, and neutral) to simplify the analysis of the polarity information. The tone extracted by the GDELT GKG was used for classification (tone > 1: positive; tone < -1: negative; other: neutral).

Site Site information extracted from GDELT GKG. To count the number of articles from different news sites, we duplicate one datum for each site. However, if there are two or more entries for the same news site information, one of these entries is used for deduplication. In the above example, one datum is duplicated for "punchng.com" and "onlinenigeria.com."

URL URL for the original news article. When there is one URL for a site, the corresponding URL is used for each site. However, when there are two or more URLs for a given news site, the shortest URL is selected for each news site (e.g., http://www.punchng.com/25909-2/ for punchng.com).

SiteCountry We constructed a database of news sites to identify their countries of origin. We used `http://www.world-newspapers.com/` to extract these relationships. For "BBC monitoring," we used "United Kingdom" as the site country for the news site. In addition, if news sites used country code top-level domains (e.g., .jp for Japan), we used this domain information to estimate the site country. Finally we used a geolocation service [3] to estimate the site country by using the IP address of the top domain. However, the country was left blank if we could not obtain appropriate location information from the geolocation service.

SiteRegion Countries were grouped into the following eight regions: USA, Asia, Europe, Middle East, Africa, Oceania, North America (excluding USA), and South America. News articles that lacked site country information were categorized as Unclassified.

We could use all of these information types other than the URL to perform multifaceted analyses.

3 NSContrast with GDELT

We set up our system based on the GDELT GKG from July 20, 2015 to January 19, 2016. Using the data conversion process described above, we extracted 31,584,327 articles from 70,781 news sites.

[3] https://freegeoip.net/, http://ip-api.com/, and http://ipinfo.io

First, we present information related to the country and region estimation. Because our manually constructed news site list is small, only 2201 sites (8,555,263 articles) were identified by using this information. Table 1 shows the number of articles (sites) by the top-level domain of URLs (Top 6). Because 81.2% (47,259/70,781) of news sites and (71.9% (22,716,591/31,584,327) of articles have .com as their top-level domain, only 10,139 sites (5,671,259 articles) were identified by their top-level domain.

Table 1: Number of articles (sites) for top-level domains (Top 6)

.com	22,716,591 (47,259)	.au	2,623,813 (1048)
.uk	1,682,960 (2705)	.org	1,029,232 (8049)
.net	645,996 (3015)	.ca	326,184 (1008)

Finally, by using the geolocation service 57,459 sites (16,816,980 articles) were identified. As a result, most of the sites (98.6%: 69,799/70,781) and articles (98.3%: 31,043,442/31,584,327) were classified into countries and regions.

Table 2 shows the number of articles for each region. From this table, news articles from the USA were dominant in the database (61.6%: 19,443,005/31,581,063). In contrast, there were only 903,811 articles from North America excluding the USA. With such unbalanced numbers of articles, making a category North America including the USA is almost equivalent to USA alone. Therefore, we divided North America into the USA and North America (excluding USA).

Table 2: Number of articles for each region

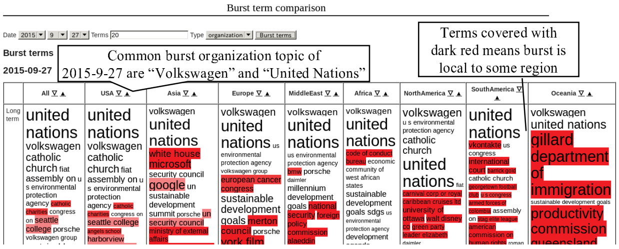
USA	19,443,005	Europe	3,696,359
Oceania	2,962,792	Asia	2,891,865
North America (excluding USA)			903,811
Africa	726,626	Middle East	373,411
South America	45,573	Unclassified	537,621

Our multifaceted analysis interface was used to compare the results with different query conditions. Figure 2 shows a time-sequence graph of polarity in different countries: all countries (upper left), China (upper right), the USA (lower left), and Europe (lower right). These graphs were constructed by adding new query conditions when selecting the data. For example, the graph for China uses news articles that included "Asian Infrastructure Investment Bank" (AIIB) as the organization, an article date ≥ July 20, 2015, and the SiteCountry = "China."

This figure shows that there were many positive articles about AIIB in China. Europe was slightly positive than the USA. This information reflects the attitudes to AIIB in these countries (or regions).

4 Conclusion

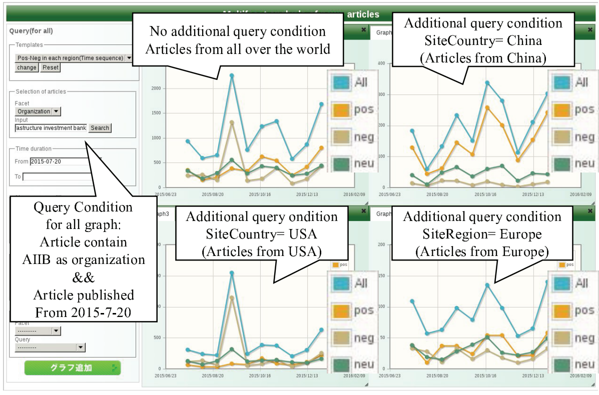
In this study, we have analyzed the characteristics of GDELT data and propose a data conversion process to utilize this information for NSContrast. In this conversion process, we conducted deduplication of news article URLs and added source country and region information to analyze the characteristic differences between them. Because of the large coverage of news sites, the system can conduct comparative analyses of various countries and regions by using large numbers of news articles from different news sites. However, for future work, it may be better to check the appropriateness of the estimated country by using a geolocation service.



Common burst organization topic of 2015-9-27 are "Volkswagen" and "United Nations"

Terms covered with dark red means burst is local to some region

Figure 1: Burst analysis results on 2015-9-27



No additional query condition
Articles from all over the world

Additional query condition
SiteCountry= China
(Articles from China)

Query Condition
for all graph:
Article contain
AIIB as organization
&&
Article published
From 2015-7-20

Additional query ondition
SiteCountry= USA
(Articles from USA)

Additional query condition
SiteRegion= Europe
(Articles from Europe)

Figure 2: Multifaceted analysis for the query "AIIB"

Acknowledgement

This work was partially supported by JSPS KAKENHI Grant Number 25280035.

References

[Kle02] Jon Kleinberg. Bursty and hierarchical structure in streams. In *Proceedings of the 8th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, pages 91–101, New York, NY, USA, 2002. ACM Press.

[LS13] Kalev Leetaru and Philip A. Schrodt. Gdelt:global data on events, location, and tone, 1979-2012. In *ISA Annual Convention 2013*, volume 2, page 4, 2013.

[YK12] Masaharu Yoshioka and Noriko Kando. Multifaceted analysis of news articles by using semantic annotated information. In *Proceedings of the fifth workshop on Exploiting semantic annotations in information retrieval*, ESAIR '12, pages 19–20, New York, NY, USA, 2012. ACM.