Exploring Significant Interactions in Live News

Erich Schubert Andreas Spitz Michael Gertz

Heidelberg University, Germany
{schubert,spitz,gertz}@informatik.uni-heidelberg.de

Abstract

News monitoring is of interest to detect current news and track developing stories, but also to explore what is being talked about. In this article, we present an approach to monitoring live feeds of news articles and detecting significant (co-)occurrences of terms compared to a learning background corpus. We visualize the result as a graph-structured semantic word cloud that uses a stochastic neighbor embedding (SNE) based layout and visualizes edges between related terms. We give visual examples of our prototype that processes news as they are crawled from dozens of news sites.

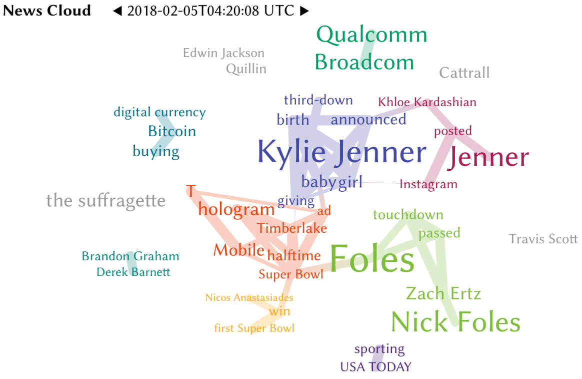


Figure 1: News coverage of the Super Bowl LII

1 Introduction

The prospect of obtaining an overview of recently published news at a glance is becoming increasingly difficult, simply due to the sheer number of available articles that are published daily by a multitude of news outlets. Ultimately, any system that is designed to provide this functionality to a user has to present selected content from such articles to the user, who then selects interesting items and investigates them further. Thus, any system that is designed to give an overview of news has to present the entirety of current news in a way that provides both a summary of contents and that enables the user to select items for subsequent reading. To address this task, numerous approaches have been proposed that aim to aggregate and visualize current news in a format that is interpretable by the user. For example, the articles can be embedded spatially and displayed in a geographic representation on a map, based on the location of the described news events [TLP$^+$08]. Similarly, an exploration along a temporal axis is possible, which leads to the generation of timelines of news [SA00]. However, such a temporal exploration is best employed for archives of news and suffers when processing live streams with very short intervals of interest. Another natural direction for news aggregation and exploration are the events or incidents that are contained in the articles [FA07]. However, the detection of events in unstructured text is a challenging task with relatively low recall, and is thus not necessarily ideal for obtaining an encompassing view on current news with many emerging events. A frequently used alternative for the exploration of (arbitrary) texts are word clouds (for an overview, see [BKP14]), which provide intuitive representations of document contents and do not rely on an extended stack of natural language processing tools. However, while word clouds provide the user with an overview of words in the documents, they do not account for the *co*-occurrences of words or their significance.

To address these shortcomings, we investigate the application of semantic word clouds to the task of representing and exploring news streams in real time. To obtain an informative layout, instead of relying on just the placement of words, we utilize the connections between individual words that we derive from a normalization in relation to a background corpus and stochastic neighbor embedding [SSW$^+$17]. The result-

In: D. Albakour, D. Corney, J. Gonzalo, M. Martinez, B. Poblete, A. Vlachos (eds.): Proceedings of the NewsIR'18 Workshop at ECIR, Grenoble, France, 26-March-2018, published at http://ceur-ws.org

ing representation then provides a quick overview of recently published news, while emphasizing significant word cooccurrences in the selected timeframe (e.g., see Figure 1). Furthermore, our proposed system can easily serve as an index of related news articles for further exploration by the user.

Related work. The European Media Monitor [AV09] monitors 5000 sources in 70 languages (with a focus on Europe), and provides the top 10 clusters for each language in their NewsBrief. Our approach is more dynamic, and emphasizes the interactions of terms and entities. In [KMS+10], the authors explore interaction networks of persons in news, but aggregated over a static four months snapshot, while we explore increased interactions within the latest news only. SigniTrend [SWK14, SWK16] was primarily built for analyzing a stream of Tweets, but has also been applied to a stream of Reuters news. It is conceptually similar to our approach, but we focus on the requirements for news exploration rather than just event detection. Implicit networks of terms or entities support the representation of document collections or streams as graph structures [SG16], and can be used for searching entity relations in the documents or visualizing them as explorable subgraphs [SAG17]. However, they do not include a normalization of relations against a background corpus and are thus limited in assessing the significance of emerging news.

2 Methodology

While our approach is inspired by SigniTrend [SWK14], we divert from it in several important ways. Much of our text processing is derived from [SSW+17], and this paper can be seen as a continuation of this work.

We process documents in the input stream in microbatches of 25 articles, because we need enough data to avoid false significance detections. The new data contributes 10% to the current counts, while 90% is based on the previous counts, which yields an exponentially weighted moving window. After about 200 documents, the weight of a document is reduced by half. This is necessary to have stable enough frequency estimates, while keeping the computational effort sufficiently low.

We use a distributed microservice architecture around an Apache Kafka message broker to enable live processing of streaming data. One service receives push notifications for new news results and adds them to the processing queue. The crawler service then crawls the articles and stores them in a database. The third service extracts the text content from these articles, and removes the boilerplate (parts of the web site such as navigation, menus, and advertisements that are not part of the actual news article). The resulting extracted text is then processed using Stanford CoreNLP [MSB+14] and the entity extraction discussed before, and fed into our model. An initial semantic layout is computed on the server side and pushed to the clients via a Web socket. A force-graph in D3.js is used for the final layout fine-tuning and the result is rendered in the browser.

2.1 Text Processing

We extract news articles from the news websites (English or German, for now) and split them into paragraphs based on their HTML structure. We then use Stanford CoreNLP [MSB+14] to further split the paragraphs into sentences and annotate parts-of-speech (POS). To lemmatize words (and to split compound words in German news), we employ an approach based on the Hunspell spell checking system that is used, for example, by Firefox, Google Chrome, and LibreOffice.

For our analysis, we only use entities, verbs, and nouns (the latter two based on their POS tag). Other words are not considered when counting cooccurrences. Since we only consider verbs and nouns, we only need a small set of stop words with common verbs such as "be", "have" and "give". Entites are merged into a single token with the canonical name, which accounts for synonyms and abbreviations, such as "EU" and "European Union", as long as they are included in the entity linker's dictionary.

We use a Gaussian weighting scheme for cooccurrences with weight $w_d = \exp(-d^2/2\sigma^2)$ for words at a token distance of d, and use $\sigma = 4$ with a window width of 12. We currently only consider cooccurrences within the same sentence for this paper (this differs, e.g., from the LOAD model for implicit entity networks [SG16]).

In contrast to SigniTrend [SWK14], we do not consider a standard deviation, which we leave as future work. Such an estimation of the standard deviation as done in their model may be beneficial to filtering trivial recurring patterns such as weekday names.

2.2 Learning the Background Model

Our initial word cooccurrence frequency model is trained by counting word occurrences and cooccurrences on Wikipedia. We used the January 1st 2018 dump of the English Wikipedia for the experiments presented here. Our entity detection system is also trained on this data, based on the probability that a page containing the entity string contains a link to the expected target page. To reduce the memory requirements (counting all word cooccurences on Wikipedia requires an enormous amount of memory), we use a hashing-based approach as previously used by SigniTrend [SWK14, SWK16]. The accuracy of this hashing technique is studied in detail in [SSW+17].

Because the word distribution in Wikipedia is different from news sources, and since we also want to emphasize the new developments, we continually incorporate the documents we have already seen into our background model. To this end, we update our background model at the end of each micro-batch such that it combines a fraction $1-\eta$ of the previous model, and a fraction η based on the documents we just processed. We used $\eta = 1\%$ in our model, so that the half-life time of data is about 70 batch iterations. Using the hash table, this approach can be implemented efficiently and does not require storing or revisiting documents outside the current micro-batch.

2.3 Judging Significance

In order to evaluate the significance of a term or a pair of terms t, we have to compare the observed frequency $c(t)$ with an estimated frequency $E[t]$. However, there are several biases for which we need to adjust. The bias term $\beta_D = \frac{1}{2}/|D|$ accounts for the document sizes of the batch, while $\beta_C = \frac{1}{2}/|C|$ accounts for a corpus size bias, where $|D|$ is the number of documents in the batch, and $|C|$ is the number of documents in the normalization corpus. The prior probability $p = k/|W|$ is the number of words k we want to choose from the vocabulary of size $|W|$ (details can be found in [SSW+17]). We obtain the overall relevance of a term as

$$r(t) = \max\left\{0, \frac{c(t)-\beta_D}{E[t]+\beta_C}\right\} \cdot p \tag{1}$$

The weighted (co-)occurrence $c(t)$ is obtained by counting term frequencies in the current micro-batch, whereas the value $E[t]$ is estimated from the background model's hash table as the minimum of all buckets, similar to count-min sketches [CM05].

The motivation of this measure is to capture the *interestingness* of a term (or term cooccurrence), and yields a ratio coefficient where values of 1 and below correspond to a non-interesting frequency. This ratio then is converted into a probability using the common transformation $p(t) = \frac{r(t)}{r(t)+1}$ to transform the ratio to a probability.

2.4 Word Cloud Visualization

We use semantic word clouds [SSW+17] for visualization. However, rather than using t-stochastic neighbor embedding (tSNE), we found the results to be better with the predecessor SNE [HR02]. We attribute this to the sparsity of the similarity matrix (where most entries will have 0 significance of cooccurrence, which is different from the distance-based matrixes commonly used), and the heavier tails of tSNE, that cause it to overemphasize separation.[1] The Student t-distribution used by tSNE causes many non-zero pairwise repulsive forces, while in regular SNE, these forces become 0 once words are sufficiently separated.

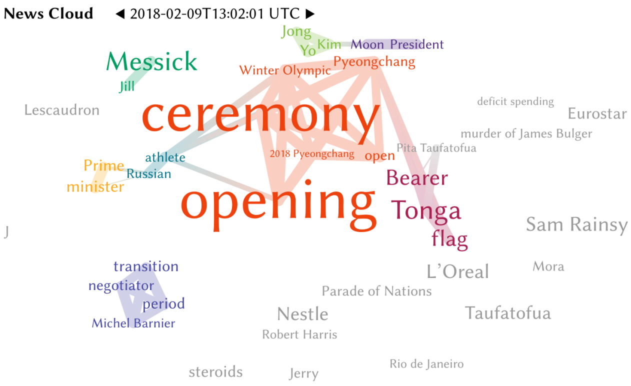


Figure 2: Word cloud for the opening of the Olympic Winter Games in South Korea, with North Korean Kim Yo-jong meeting South Korean president Moon Jae-in. The oiled Tonga flag bearer, Pita Taufatofua, got a lot of attention.

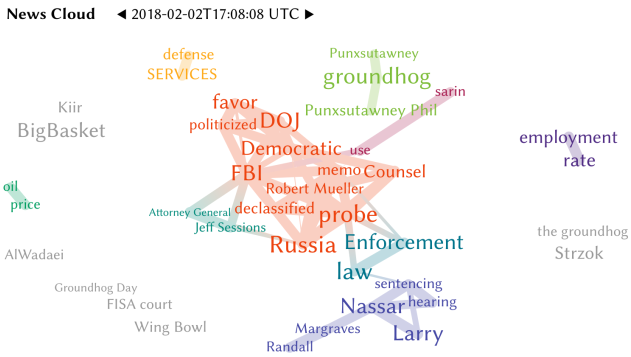


Figure 3: Declassification of the Nunes memo, Larry Nassar sentencing, and Groundhog Day.

The initial word layout is generated in the backend, while the web browser frontend uses a force directed graph to optimize the layout for the user's screen size. We employ forces that draw words to the desired location, forces that try to keep linked words close to each other, and additional forces that try to avoid word overlap based on a bounding box collision algorithm.

A key contribution of the visualization is the inclusion of links between cooccurring words and entities. The link strength visualizes the value of the probability p described above, i.e., it measures how much *more* common this connection is than expected. It is not just displaying the count, which would produce too many uninteresting links.

[1] A demonstration of this effect can be seen at `https://stats.stackexchange.com/a/264647`

2.5 Word Clustering

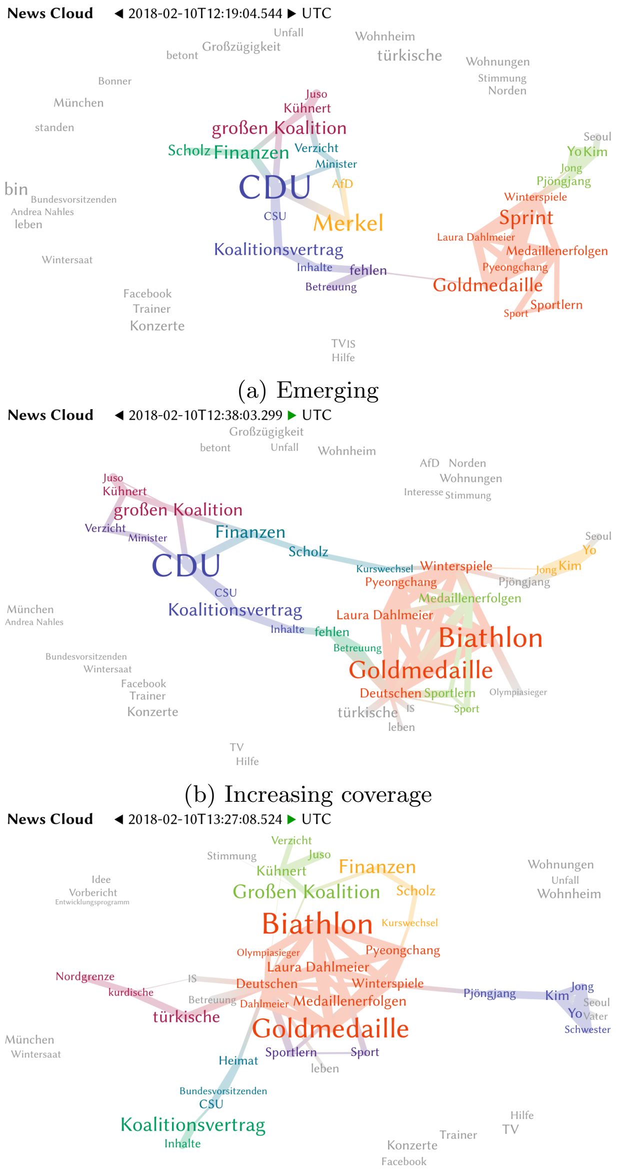
We cluster words based on their link strength, using the very classic AGNES hierarchical clustering algorithm [KR90] with group-average linkage, and extract clusters using the method described in [SSW+17]. This approach tries to find 8 clusters with a minimum size of 2 words, while it puts unclustered words into a separate "noise" set, inspired by DBSCAN clustering [EKSX96, SSE+17]. The main benefit over DBSCAN is that we do not need to set appropriate thresholds, but can extract clusters based on a desired number and minimum cluster size. Furthermore, the group-average linkage strategy takes the pairwise relatedness of all cluster members into account, and has a lower tendency to transitively merge topics. Experimentally, it produced more intuitive results compared to other linking methods. Our clustering implementations are derived from the open-source implementations available in the ELKI data mining framework [SKE+15].

In future work, we intend to cluster links instead of words (or a hybrid of these two approaches) because we repeatedly observe words that are used in two separate topics. For example, in Figure 3, the word "use" causes an undesired connection between the Nunes memo event, and the use of Sarin gas in Syria.

3 Experiments

The objective evaluation of explorative data mining methods such as clustering is inherently difficult [FGK+10]. For example, a user study that involves asking users to summarize recent events when presented with either a word cloud or just a ranked list of articles, takes considerable effort. In the case of our model in particular, such a study is unlikely to contribute to the deeper understanding of the approach. Since we add additional information to the visualization by clustering words and adding relations between them, it is difficult to imagine a setting in which a user will find them less informative than a traditional word cloud. In the following, we therefore present some anecdotal evidence and leave it to the reader to decide on whether this approach is worth exploring.

Figure 1 on the first page shows the Super Bowl LII news coverage, including the touchdown by Zach Ertz after a pass by Nick Foles, but also the announcement of Kylie Jenners baby (and pregnancy), which broke Instagram records. Figure 2 is a snapshot of the Olympics opening ceremony, and Figure 3 of the Nunes memo. The latter also shows the limitations of our approach: the declassification of this memo has been discussed in the media since January 18, and did not come at much surprise. Much of the new coverage discusses the concern that it might be used to try to fire Jeff Sessions, and thus Robert Mueller.



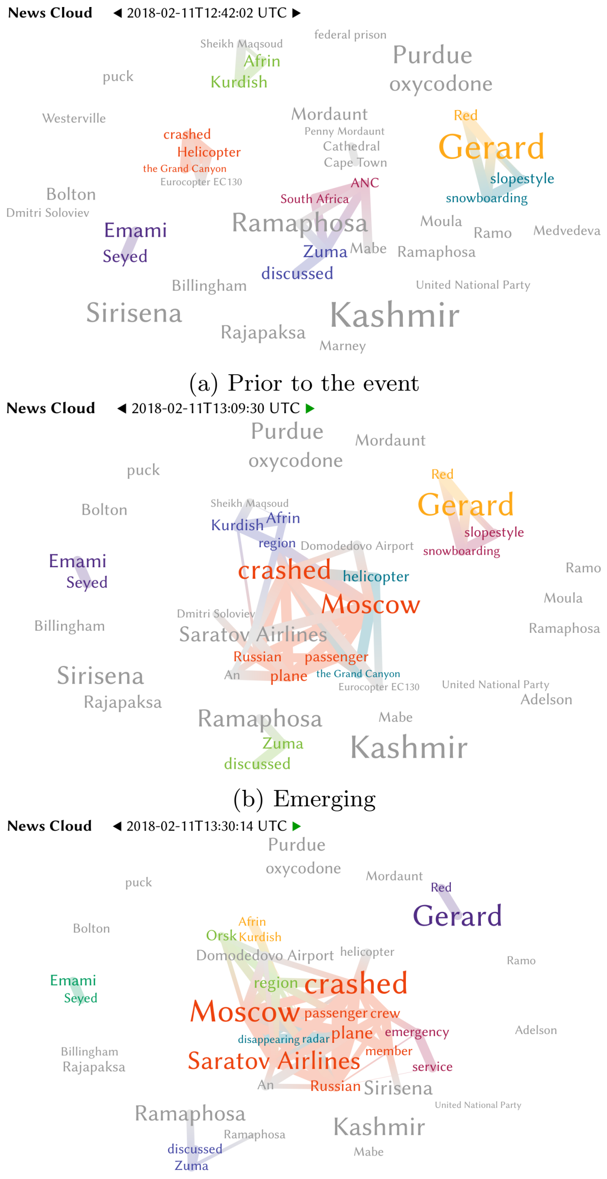
(a) Emerging

(b) Increasing coverage

(c) Olympics taking center stage

Figure 4: Temporal development of the first Olympic gold medal for a German athlete in Biathlon (in German news) taking center stage, while recent political news on government coalition troubles slowly fade.

In Figure 4, using German news, we show the emergence of a cluster around the first gold medal for a German athlete in the 2018 Winter Olympics. Prior to this event, the center stage was occupied by politics, and the ongoing struggle of German politicians to form a coalition government. Within an hour, the main news events are centered around the Olympic games (unsurprisingly, as there has been no major new developments on the political side; so this is the desired behavior). Figure 5 showcases the development of news coverage of a passenger airplane crash in Russia, exhibiting a similar development in English news.



(a) Prior to the event

(b) Emerging

(c) Dominant topic

Figure 5: Temporal development of news covering the plane crash of the Saratov Airlines passenger flight.

4 Conclusions and Future Directions

This paper calls for future research in many directions.

The current visualization is limited by readability aspects because we cannot display much more than 50 words. To improve the user experience when exploring the data, it will be interesting to instead cluster many more words and allow a drill-down approach in which the user can zoom into groups of words that are then displayed in more detail with additional words.

Our prototype provides a single link for each word. A Google News style user interface that allows the user to browse related results from different news sources based on short snippets would be beneficial.

Words occurring in different contexts can cause undesired links between clusters. For example, "police car" and "train car" share the common term "car", but will often belong to different events. Thus, a more complex clustering based on links instead of words – or a hybrid that uses both – could improve the clustering quality. Instead of clustering, the use of topic models for colorizing the plots is also worth exploring, but it is not obvious how to adapt topic models to use (co-)ocurrence *significance* instead of mere frequency.

Recurring patterns are problematic, such as month names, weekdays, or the word "weekend". While calendar words are easily added to a stopword list, a learning approach is preferable. Entity linking with Wikipedia performs very well with frequent entities, but cannot adapt to entities not yet present in Wikipedia [GTBd14, EAM^{+}17], as needed for emerging events. For example, an article on Larry Nassar was not created until January 18, 2018. The system will benefit from an improved entity disambiguation and the ability to learn new entities (such as names) on the fly, even when they cannot yet be linked.

Temporal aspects are of high interest: when do clusters and links emerge, and when do they disappear? In our prototype, the user can navigate between microbatches (each covering about 15-30 minutes of news data) with the arrow buttons, but we do not overlay frequency histograms onto the words as done in some of the related work [LHKC10, LBSW12]. Different events may require different timeframes for analysis. For example, the Super Bowl itself was in the media days before the actual event, and we observe a gradual ramp-up. In-game events will occur at a much shorter time frame. This may require using multiple estimators, learning at different rates.

The quality of the results depends strongly on the crawling lag, i.e., the delay between when the publication time of the article and the time when it is added to the feed. With regular crawling, it is common to see lags of much over 30 minutes. Best results are obtained if push notifications by the publishers notify the system of new contents, which will usually reduce the lag to less than 5 minutes (compared to the reported publishing time; we cannot avoid lag on the publisher side between assigning the publishing date, the actual availability on the web site, and the notification being delivered). Republished articles and duplicates add further data quality challenges. This requires complex systems to gather the data, until news outlets provide reliable and low-latency push notification APIs.

All-in-all, the significance of cooccurrences and the visualization thereof, is an interesting step towards exploring the interactions of multiple entities in current news articles, and allows both for interesting applications as well as further research.

References

[AV09] M. Atkinson and E. Van der Goot. Near real time information mining in multilingual news. In *WWW*, 2009.

[BKP14] L. Barth, S. G. Kobourov, and S. Pupyrev. Experimental comparison of semantic word clouds. In *SEA*, 2014.

[CM05] G. Cormode and S. Muthukrishnan. An improved data stream summary: the count-min sketch and its applications. *J. Algorithms*, 55(1), 2005.

[EAM+17] J. Esquivel, D. Albakour, M. Martinez-Alvarez, D. Corney, and S. Moussa. On the long-tail entities in news. In *ECIR*, 2017.

[EKSX96] M. Ester, H.-P. Kriegel, Jörg Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *ACM KDD*, 1996.

[FA07] A. Feng and J. Allan. Finding and linking incidents in news. In *ACM CIKM*, 2007.

[FGK+10] I. Färber, S. Günnemann, H.-P. Kriegel, P. Kröger, E. Müller, E. Schubert, T. Seidl, and A. Zimek. On using class-labels in evaluation of clusterings. In *MultiClust Workshop*, 2010.

[GTBd14] D. Graus, M. Tsagkias, L. Buitinck, and M. de Rijke. Generating pseudo-ground truth for predicting new concepts in social streams. In *ECIR*, 2014.

[HR02] G. E. Hinton and S. T. Roweis. Stochastic neighbor embedding. In *NIPS 15*, 2002.

[KMS+10] M. Krstajic, F. Mansmann, A. Stoffel, M. Atkinson, and D. A. Keim. Processing online news streams for large-scale semantic analysis. In *ICDE Workshops*, 2010.

[KR90] L. Kaufman and P. J. Rousseeuw. *Agglomerative Nesting (Program AGNES)*. John Wiley & Sons, Inc., 1990.

[LBSW12] S. Lohmann, M. Burch, H. Schmauder, and D. Weiskopf. Visual analysis of microblog content using time-varying co-occurrence highlighting in tag clouds. In *Advanced Visual Interfaces, AVI*, 2012.

[LHKC10] B. Lee, N. Henry Riche, A. K. Karlson, and S. Carpendale. Sparkclouds: Visualizing trends in tag clouds. *IEEE Trans. Vis. Comput. Graph.*, 16(6), 2010.

[MSB+14] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In *ACL System Demonstrations*, 2014.

[SA00] R. C. Swan and J. Allan. Automatic generation of overview timelines. In *ACM SIGIR Conf. on Research and Development in Information Retrieval*, 2000.

[SAG17] A. Spitz, S. Almasian, and M. Gertz. EVELIN: exploration of event and entity links in implicit networks. In *Int. Conf. on World Wide Web, WWW Companion*, 2017.

[SG16] A. Spitz and M. Gertz. Terms over LOAD: Leveraging named entities for cross-document extraction and summarization of events. In *ACM SIGIR Conf. on Research and Development in Information Retrieval*, 2016.

[SKE+15] E. Schubert, A. Koos, T. Emrich, A. Züfle, K. A. Schmid, and A. Zimek. A framework for clustering uncertain data. *P. VLDB Endowment*, 8(12), 2015.

[SSE+17] E. Schubert, Jörg Sander, M. Ester, H.-P. Kriegel, and X. Xu. DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS)*, 42(3), 2017.

[SSW+17] E. Schubert, A. Spitz, M. Weiler, J. Geiß, and M. Gertz. Semantic word clouds with background corpus normalization and t-distributed stochastic neighbor embedding. *CoRR*, abs/1708.03569, 2017.

[SWK14] E. Schubert, M. Weiler, and H.-P. Kriegel. SigniTrend: Scalable detection of emerging topics in textual streams by hashed significance thresholds. In *ACM KDD*, 2014.

[SWK16] E. Schubert, M. Weiler, and H.-P. Kriegel. SPOTHOT: Scalable detection of geo-spatial events in large textual streams. In *SSDBM*, 2016.

[TLP+08] B. E. Teitler, M. D. Lieberman, D. Panozzo, J. Sankaranarayanan, H. Samet, and J. Sperling. Newsstand: a new view on news. In *ACM SIGSPATIAL*, 2008.