Viz-TSF: A Visual Twitter Search Tool based on Query-driven Filter Optimization

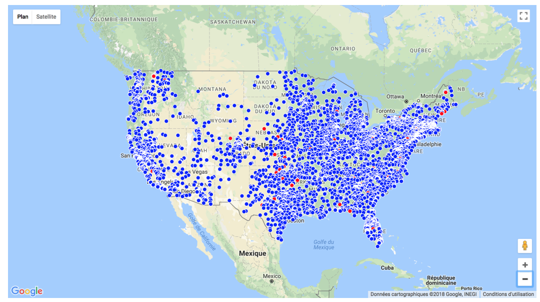
Mohamed Reda Bouadjenek
The University of Toronto
Department of Mechanical and Industrial Engineering
Toronto, Ontario M5S 3G8, Canada
mrb@mie.utoronto.ca

Scott Sanner
The University of Toronto
Department of Mechanical and Industrial Engineering
Toronto, Ontario M5S 3G8, Canada
ssanner@mie.utoronto.ca

ABSTRACT

Geo-temporal visualization of Twitter search results is a challenging task since the simultaneous display of all matching tweets would result in a saturated and unreadable display. Thus, the development of novel intuitive visual interfaces is necessary to help focus a user's attention by filtering information relevant to their query. In this work, we propose a Visual Tweet Search Filter (Viz-TSF) tool for query-driven optimization and ranking of spatial, temporal, and content-based filters to help focus a user's exploration of search results. We leverage a fast greedy optimization algorithm to optimize an approximation of expected F1-Score metric to generate these filters and demonstrate its application to search 2 years of Twitter content, specifically for a user's information need related to natural disasters occurring in the US. Our demo shows that Viz-TSF is capable of extracting geo-temporally coherent filters given search queries, thus aiding the user in visually searching and browsing social network content and enabling new opportunities for the application of Information Retrieval techniques to general visual information exploration interfaces.

Keywords: Adaptive UIs; Visual Search Interfaces; Optimization for IR.



(a) Global display showing search results.

(b) Filtered and focused search interface.

Figure 1: (a) An unfiltered user search interface with geolocated tweets that best match a query related to natural disaster highlighted in red. (b) A filter-optimized version of the interface showing the top three filtered subsets of data identified to have high coverage of relevant content related to the query: a bounding box near New York state limited to data in Feb 2015 and having keyword "Blizzard" and other time and keyword restricted bounding boxes centered on Florida and Washington state.

1 INTRODUCTION

Traditional search engines such as Google or Bing display search results in a vertical list of textual summaries. However, this display mode is certainly not adapted for search results over Twitter content, since related tweets are often geographically and temporally localized. Moreover, given the massive volume of available information in Twitter, displaying all relevant tweets for a given query prevents the visual extraction of relevant information as it would result in a saturated and unreadable display [4]. Thus, adaptive user interfaces (AUIs) are necessary to filter information to provide localized and relevant tweets for each query. Such AUIs would allow the user to efficiently explore the content of the matching tweets and to identify key properties (author, timestamp, geolocation) relevant to the fulfillment of the user's information need. To this end, we propose a Visual Tweet Search Filter (Viz-TSF)[1] tool to optimize and rank spatial, temporal, and content-based filters intended to focus a user's exploration of Twitter search results.

To make the task of AUI filtering more concrete, we introduce an example search filtering use case for AUIs. Consider the case of searching events related to natural disasters that are discussed on Twitter. Typically, as shown in Figure 1(a), there would be some

[1] http://130.220.208.198:8080/AUI4IRSearch/

WOODSTOCK'97, July 1997, El Paso, Texas USA

ACM ISBN 123-4567-24-567/08/06…$15.00
DOI: https://doi.org/10.475/123_4

visual display of all tweets that match the query along with highly relevant Tweets shown as red nodes. Displaying all tweets simultaneously would result in a saturated and an unreadable display for many queries that return a high volume of matching content. To ease the investigation and search task, users could restrict the tweets displayed through a variety of filter settings to get a clear overview of individual events as shown in Figure 1(b) – by panning and zooming in the graph display, by restricting upper and lower bounds on a time filter, and/or by selecting properties (e.g., in a drop-down selection or fielded keyword search in fields such as Tweet content, author, or hashtags).

While the user would typically find it hard to manually set these filters to optimally reveal information about specific relevant events related to his/her query, an AUI that is aware of a probability estimate that each tweet is relevant to the query could automatically optimize and suggest a ranking of filter settings to visually display relevant tweet content with the least amount of irrelevant clutter. The user could then more efficiently browse through the subset of tweets provided by the ranked filter settings to carry out investigation w.r.t. the original information need while covering a large fraction of the relevant content.

Since "retrieved" tweets are not individually selected, but rather chosen through a filter setting (that the user can further modify), the problem is clearly an optimization problem of how to restrict filter settings to show the user the most relevant tweets to his/her query. While this diverges from the standard information retrieval setting where ranked documents are chosen individually [1], these additional constraints do not change the overall information retrieval objective to select relevant information given the user's information need and constraints of the user interface.

To the best of our knowledge, Viz-TSF is the first tool to address a visual information retrieval problem for social media as the explicit optimization of spatial, temporal, and content-based filter selection and ranking w.r.t. surrogates of F1-Score well-suited to this task.

2 Viz-TSF ARCHITECTURE AND FEATURES

The main components of the architecture of Viz-TSF are:

(1) *Crawler and Indexer.* To find and organize tweets for further retrieval.
(2) *Query Analyzer.* To understand the user's information need.
(3) *Scoring Function.* To score the tweets based on their relevance probability w.r.t. the query.
(4) *Filter Optimization.* To optimize and rank filter settings w.r.t. the scoring function and filtering objective.
(5) *Visual Adaptive User Interface.* To present the information to the user in an accessible and interactive form using filters.

While the two first components are common to many IR systems, we briefly describe in the following the remaining components which are at the heart of Viz-TSF.

2.1 Scoring Function

The filter optimization component of Viz-TSF relies on the optimization of an approximation of *expected* F1-Score metric. This expectation is critically based on an estimated conditional probability $p(q|t)$ that a tweet t is relevant to the query q. To obtain this probability estimate, the ranking function component of Viz-TSF uses a language model scoring approach, which assumes that the terms of a query q are "generated" by a probabilistic model based on the content of tweet t, thus directly estimating $p(q|t)$. Specifically for this demonstration, we use a Bayesian smoothing language model using Dirichlet priors as defined in Zhai and Lafferty [6].

2.2 Filter Optimization

The filter optimization component of Viz-TSF relies on the configuration of three sub-filters to select a subset of relevant tweets for the data visualization interface – specifically (i) a keyword content sub-filter, (ii) a temporal sub-filter, and (iii) a spatial (location) sub-filter. A global filter that is generated by *conjoining* these three sub-filters is used to select the overall subset of relevant tweets for a filter. Each sub-filter is constructed following a greedy top-down approach while seeking to optimize an approximation of the expected F1-Score metric. For efficiency, the algorithm implemented in this module uses a greedy binary partitioning search, where the algorithm operates by halving the search space of one sub-filter in each iteration. Details of this algorithm are further provided in [2].

In practice, a single conjoint filter setting chosen by the previous algorithm will narrow the user on a single "event". However, there are likely multiple events and so the user should have a ranked choice of multiple filters to choose from. Hence, after the first filter is produced, all selected tweets by the filter have their probability scores zeroed out (since they are already covered by one filter). The filtering algorithm is then run again, where it will inherently focus on a different content set. This procedure is repeated until the desired number of filters is reached. The user should then be able to choose among a ranking of the filters produced as provided in our demonstration interface.

2.3 Visual Adaptive User Interface

As shown in Figure 2, we designed an interactive visual interface providing a multi-perspective view consisting of: (i) an interactive map, which presents search results for each filter setting that permits further user-driven adaption of the filters (e.g., spatial refinement through panning and zooming) and pop-up information summaries by clicking on filters or tweets, (ii) a query panel allowing the user to specify a query, refine the time window filter, and tune two search parameters (the number of filters produced and the value of β for the F_β-Measure that is optimized), and (iii) a ranked list of the filters produced with respect to a given query, which can be used to select and (un)hide some filters for more clarity.

Viz-TSF is built on the top of the Lucene IR system[2] [5] for the indexing, query processing and retrieval part, and on the Google Maps API for data visualization.

3 WHAT WILL BE DEMONSTRATED?

This demonstration will be illustrated using Twitter data crawled using the Twitter Streaming API for two years spanning 2013 and 2014 [3]. This dataset contains more than 2.5 TB of compressed data, with a total of 829,026,458 English tweets.

In our demonstration, a user can perform interactive searches such as the following example search scenario:

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

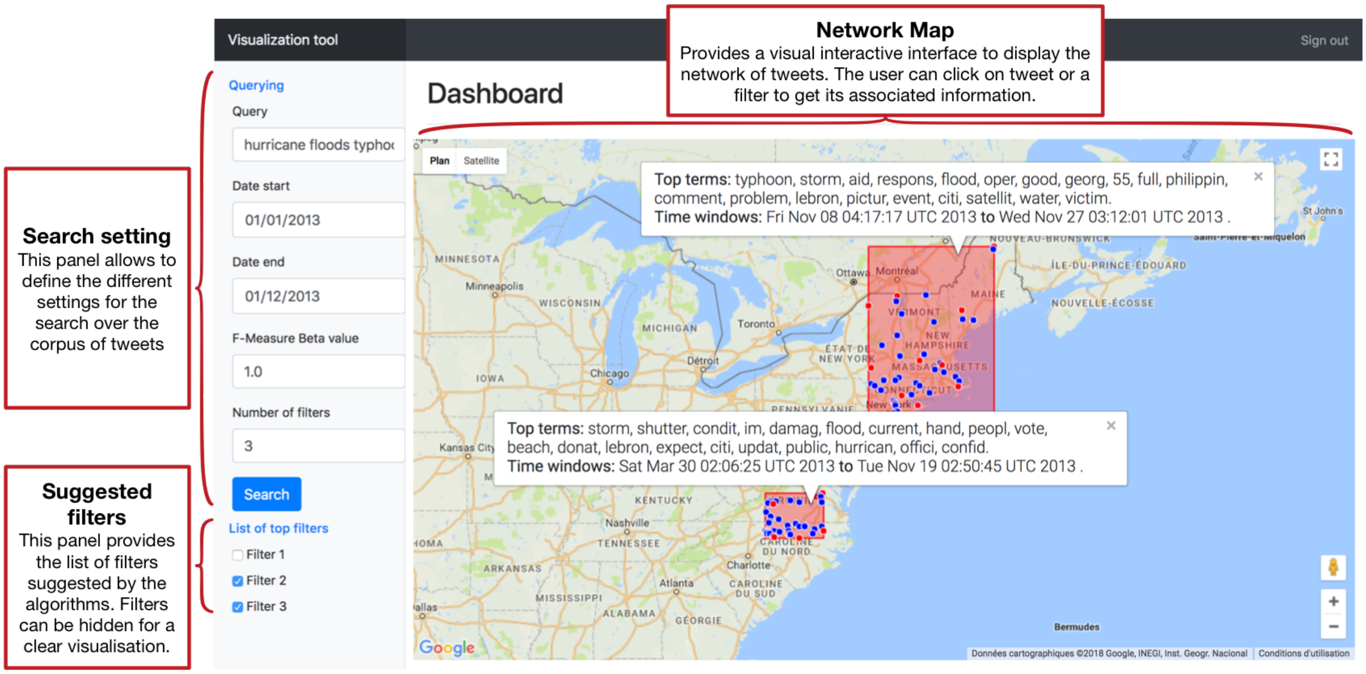


Figure 2: Main interface of Viz-TSF.

- A user begins by searching for information on weather-related natural disaster events. Using the querying panel, the user enters a set of keywords "flood, hurricane, typhoon, drought, storm", selects a search time window (e.g., January 2013 to June 2014), selects the number of filters to be returned, and sets an initial value for β. A low value of β constrains the algorithm to focus on the optimization of the precision, thus obtaining filters with small clusters, whereas a high value of β leads to high recall filter settings with large clusters. The value of β can be interactive and iteratively tuned by the user until the desired granularity or coarseness of filter results are obtained.
- As shown in Figure 3, a list of filters is suggested to the user and displayed on the main map (in this case restricted to the US) along with relevant tweets for the search query, where red nodes indicate highly relevant tweets, and blue nodes indicate lower relevance tweets.

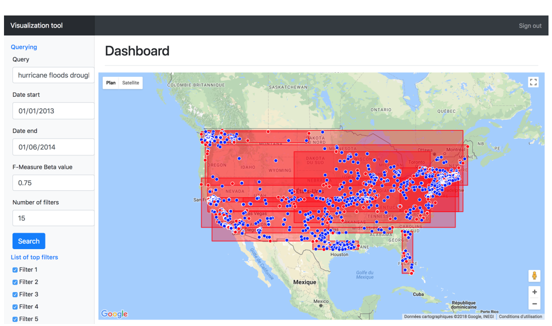


Figure 3: Results showing all the filters that match the query.

- Next, as shown in Figure 4, using the side panel, the user may deselect some filters that he/she thinks are unrelated to their information need in order to focus on a subset of filters that are of interest.

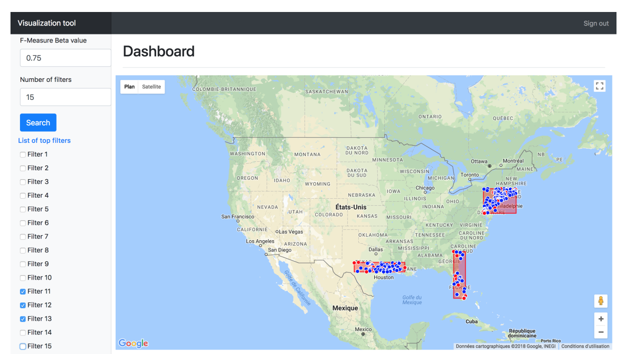


Figure 4: Sub-set of the filters that match the query.

- Then, as shown in Figure 5, the user may zoom in a specific filter for investigation by looking for its top terms, its associated tweets, its time window, etc. In the case of Figure 5, the user investigates a filter that contains tweets related to some flood disasters that happened on the northeast coast of the US during the time period ranging from December 2013 to February 2014.
- Similarly, the user may investigate a second filter in the ranked list by looking at its content. For example, in Figure 6, the filter investigated is related to a series of storms and

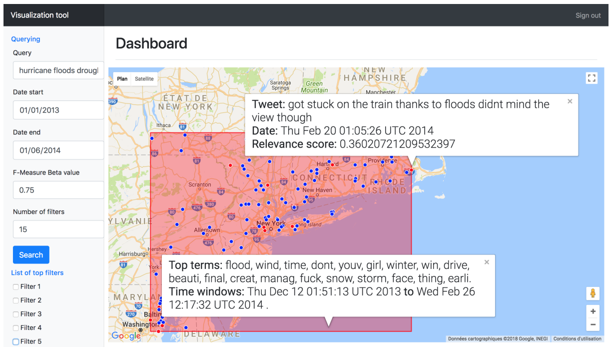


Figure 5: Investigating filter 1.

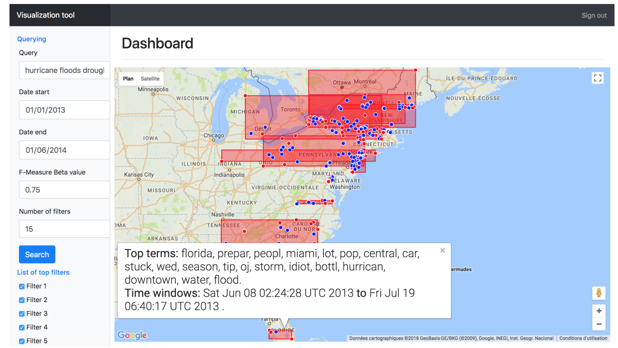


Figure 7: Filtering east US coast.

hurricanes that happened in the southeast coast of the US (mainly around Florida) during the time period ranging from April 2013 to June 2013. Clearly, the two investigated filters are of two different events related to two different natural disasters that happened during two different periods. Thus, this suggests that in this search case, Viz-TSF succeeded in finding two different filters that focus on different (distinct) events related to the initial search query.

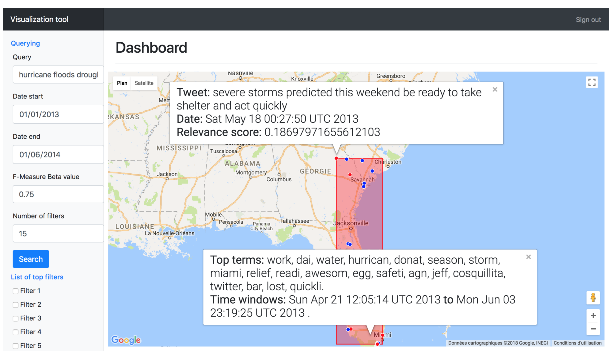


Figure 6: Investigating filter 2.

- Finally, the user may decide the re-run his/her query based on a specific local geographical area. In Figure 7, the query is re-executed on the US east coast, where new local filters are produced, which may be investigated by the user. For example, the bottom filter shown in Figure 7 with a time window from June to July 2013 contains a set of tweets that warn people about the importance of preparing houses for the new upcoming hurricane season.

In summary, the content of this demonstration aims to provide users with a novel way to visually and interactively browse geo-temporally filtered content, specifically in this case for tweets relating to natural disasters in the US over a two year time period.

4 SUMMARY AND FUTURE WORK

In this paper we have introduced Viz-TSF, a tool to search Twitter content through spatial, temporal, and content-based filters to help focus the user's exploration of search results with geo-temporally coherent content. We leveraged a fast greedy optimization algorithm to optimize an approximation of expected F1-Score metric to generate filters and demonstrate its application to search 2 years of Twitter content to better understand tweets that match a given search query. We have described a search scenario of queries related to an information of natural disasters occurring in the US during a two year period and showed the capabilities of using Viz-TSF to search large-scale data in social networks.

Important areas of future work include consideration of the role of (pseudo-)relevance and other explicit or implicit feedback methods to create a tighter and more responsive user interaction loop. Furthermore, in combination with user studies and consideration of human factors, future work should also consider novel application-specific objectives, e.g., in specific visualization frameworks or based on a ranking theory of results presentation (e.g., using size or colour for visual ranking emphasis).

Overall, we aim for this work to bring a formal IR perspective to the important research area of AUIs while opening new research directions for the application of optimization techniques for filter selection in AUIs. As information retrieval has transformed our web experience, we believe there is a similar opportunity to transform the present nascent state of AUIs through optimization and information retrieval principles to make them more ubiquitous in our daily lives. We hope this demonstration serves as one example of such a novel visual information retrieval interface.

REFERENCES

[1] R. A. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., 2 edition, 2010.

[2] M. R. Bouadjenek, Y. Du, and S. Sanner. An Information Retrieval Perspective of Filter Selection for Adaptive User Interfaces. Technical report, University of Toronto, Department of Mechanical and Industrial Engineering, 2018. http://people.eng.unimelb.edu.au/mbouadjenek/papers/bouadjenek2018b.pdf

[3] Z. Iman, S. Sanner, M. R. Bouadjenek, and L. Xie. A longitudinal study of topic classification on twitter. In *ICWSM*, pages 552–555, 2017.

[4] S. Liu, W. Cui, Y. Wu, and M. Liu. A survey on information visualization: recent advances and challenges. *The Visual Computer*, 30(12):1373–1393, Dec 2014.

[5] M. McCandless, E. Hatcher, and O. Gospodnetic. *Lucene in Action, Second Edition: Covers Apache Lucene 3.0*. Manning Publications Co., Greenwich, CT, USA, 2010.

[6] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 334–342, New York, NY, USA, 2001. ACM.