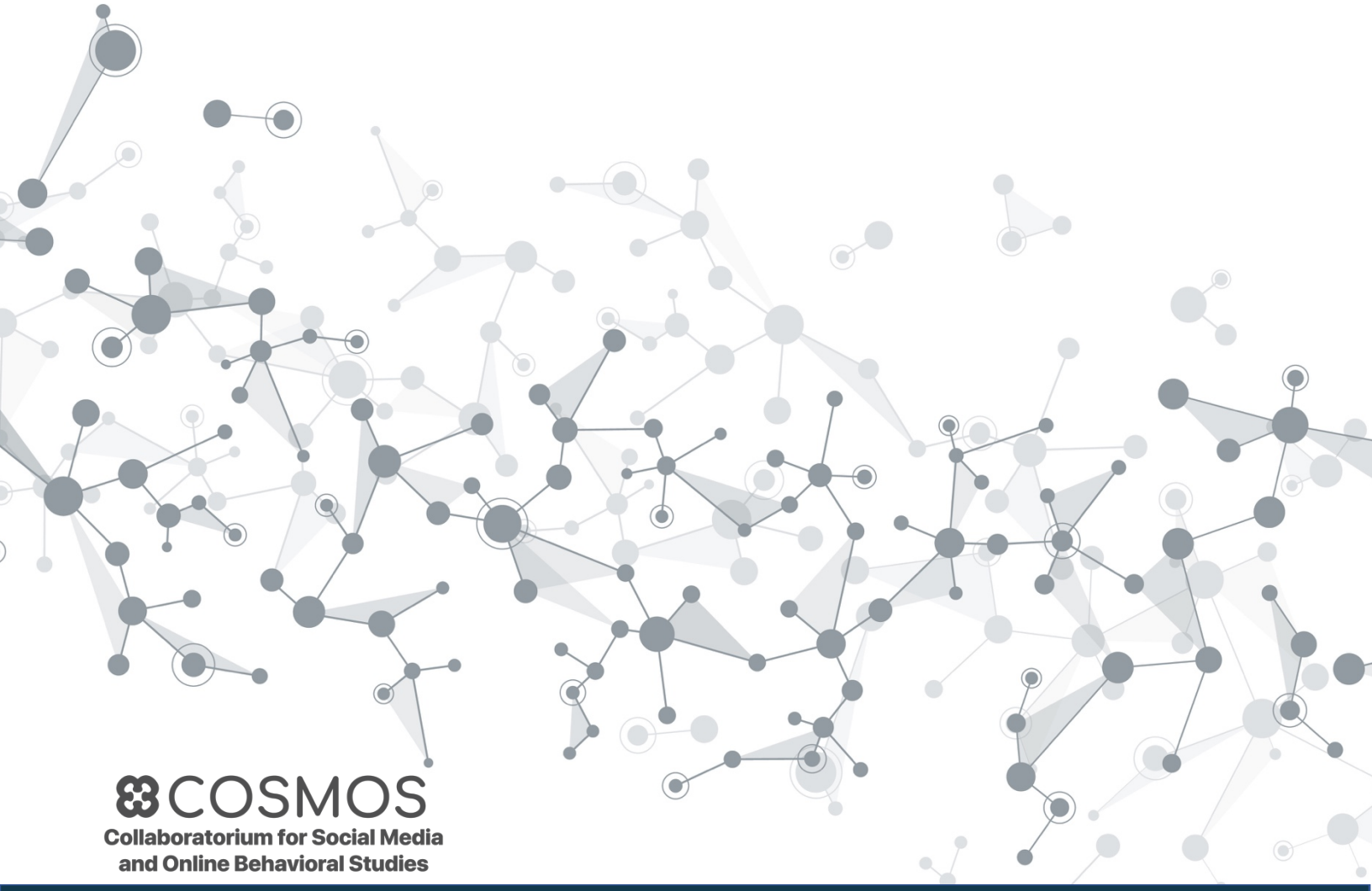
# Blogtrackers

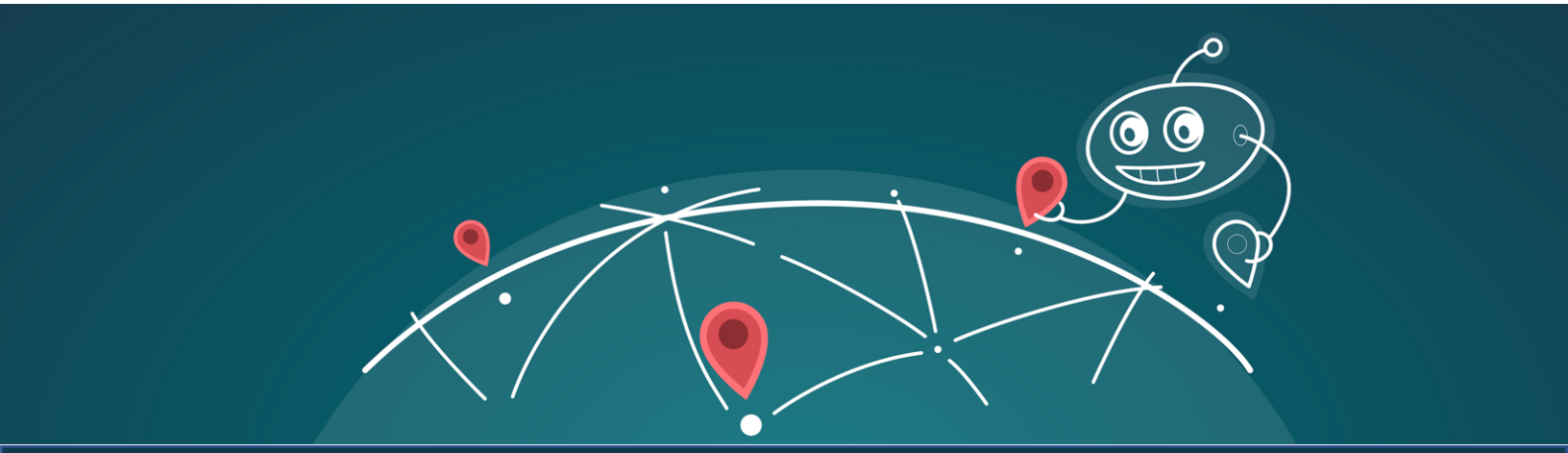
## Technical Resource Blogtrackers Web Application



COSMOS
Collaboratorium for Social Media and Online Behavioral Studies

# Table of Contents





## Executive Summary

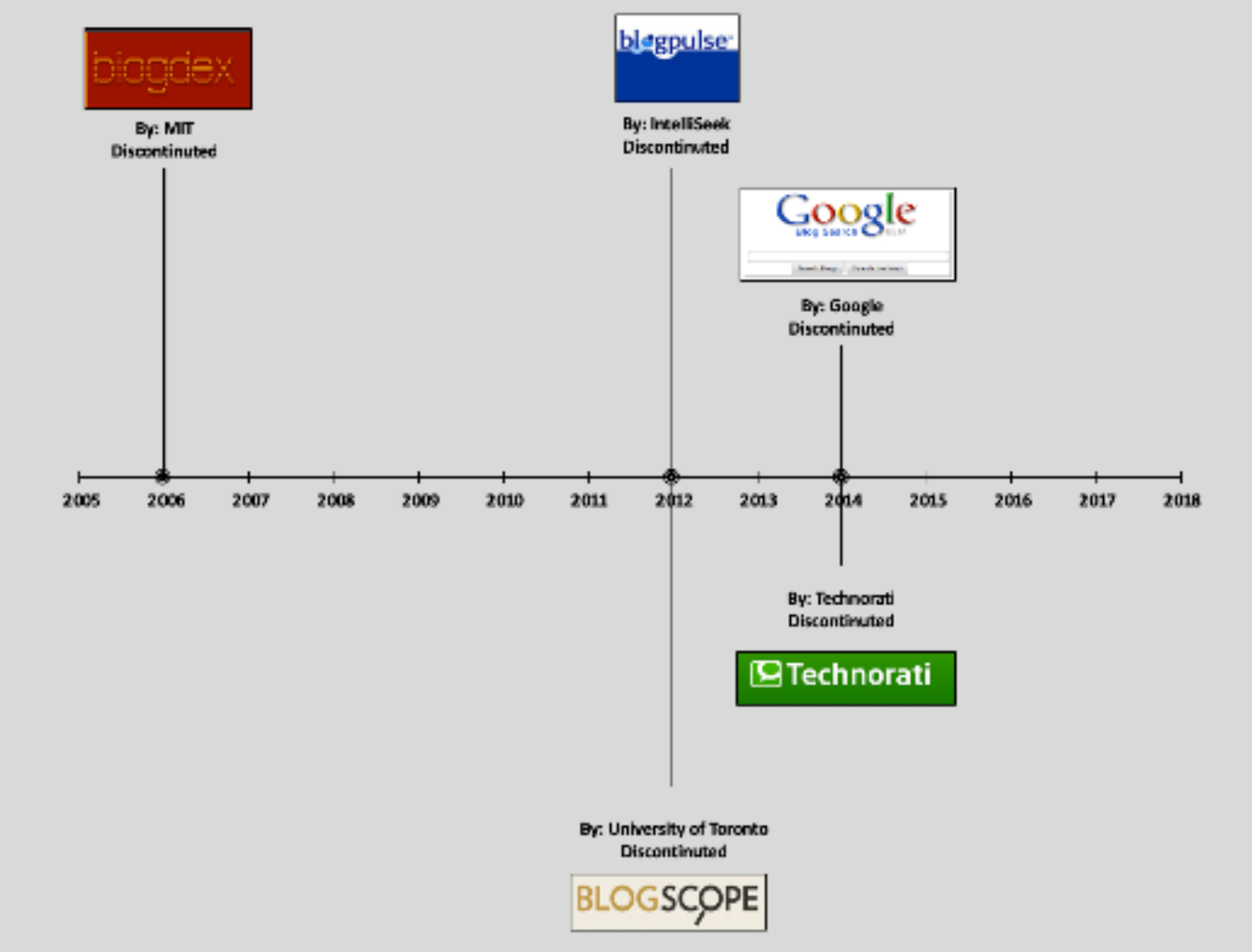
Social media has grown to be the place for voicing one's opinions, share information, and shape discourse. Individuals use social media as a platform to mobilize, coordinate, and conduct cyber campaigns ranging from awareness for diseases or disorders to deviant acts threatening democratic principles and institutions. Blogosphere has continued to rise and afford an effective medium for content framing. With no restriction on the number of characters, many use blogs to set narratives then use other social media channels like Twitter and Facebook to steer their audience to their blogs. Blog content is not structured and hard to collect than other social media channels. Blog monitoring and analysis could be of great use to sociologists, political scientists, communication researchers, journalists and information scientists to examine events. Toward this direction, we present Blogtrackers tool, which is designed to explore the blogosphere and gain insights on various events. Blogtrackers can help in identifying leading information actors, influential bloggers, popular and emerging trends, assess tones, sentiments and opinions, extract entities, and analyze their networks.

## What is Blogtrackers?

Blogging has become a popular channel as users can write without any restriction on topics that interest or affect them. Typically, other networking sites like Twitter and Facebook drive users to the blogs and empower discussions due to character limit. Blogosphere, defined as the network of blogs, is growing at an exponential rate. WordPress.com, which is one of the most popular blogging platforms, recorded activity of more than 87.6 million news posts added each month and over 409 million views per month in 2016 [1]. Analyzing blog data helps in understanding the pulse of a society, knowing what resonates with a community, and recognizing grievances of a group, among other reasons. Since there is no character limit in blogs unlike Twitter, blogs improve quality and inclusiveness of discourse and afford an effective platform to set narratives. Blogs also provide a convenient platform to develop situational awareness during a socio-political crisis or humanitarian crisis in a conflict-torn region or a disaster-struck area. In this demo, we introduce Blogtrackers, a tool that can be used for analyzing blogs to gain situation awareness.

## History of Blogtracking

### Timeline 2005 - 2018



## Challenges of Blogtracking

### Blog Data Collection – Challenges

1. No API support for blog data – Existing services are discontinued.
2. Changing blog structure – Moving target all the time.
3. Noise – Social media plugins such as Facebook share plugins, Twitter share plugins, and advertisements from the blog site could be crawled as JavaScript.
4. No standardization – Conversion required for some attributes - once such attribute is date. For date field extraction from blog posts, we notice it differs in format from blog site to blog site meaning there is no single standard followed. This adds to the workload of converting these kinds of issues to a standard format for further study.
5. No automation – Human intervention is needed.
6. Limitations of tool – Web Content Extractor, does not work with crawling dynamic pages.

## Why Blogtrackers?

Blogtrackers is a tool designed to explore the blogosphere and gain insights about events and how these events are perceived in the blogging community. Blogtrackers provides an analyst with means to develop situation awareness. Following are a few features and analytical
capabilities of Blogtrackers:

1. Setup Tracker: A tracker is a collection of blogs selected by user for analysis. Setup Tracker page allows user to search for a topic of interest and select blogs to create a tracker for analysis. User can create and save any number of trackers.

2. Dashboard gives the overview of the selected tracker. It displays the number of blogs, bloggers, blog posts, total positive and negative sentiments. It also displays blog sites' hosting location and language distribution.

3. Posting Frequency can be utilized to identify any unusual patterns in blog postings. This aids in detecting real-time events that interested the blogging community. This feature also displays a list of active bloggers with number of posts. User can click on any data point on the graph to get a detailed list of the named-entities that were mentioned in blogs during that time-period. Fig 1 shows the posting frequency for NATO blogs during 2016.

4. Keyword Trends provides an overall trend of keywords of interest. It helps track changes in topics of interest in the blogging community. An analyst can correlate keyword trends with events to examine discussion topics and themes relating to that event. The analyst can select any data point on the trend line to view all the blogs. Fig 2 shows the keyword trends related to the ongoing Venezuelan sociopolitical crisis.

5. Sentiments and Tonality: displays the trend of positive and negative sentiments of blogs for any selected time-period (Fig. 3). This helps in understanding the

effect an event has on the blogosphere. Additionally, data analyst can drill down by clicking on any point of interest and view radar charts (Fig. 4) displaying tonality attributes such as personal concerns, time orientation, core drives, cognitive process.

6. Influence: This feature helps identify the influence a blogger or blog post has on theblogosphere. Blogtrackers finds the posts that are authoritative by assigning a score calculated using the iFinder model [7], [8]. This feature lists top 5 influential bloggers and displays a trend line to show the variation in bloggers' influence (Fig. 5). Clicking on a point on the trend line allows a deeper dive into the data. This feature also provides capability to visually distinguish between influential and active bloggers. Further a user can explore the content themes of active-influential, inactive-influential, active-noninfluential, and in-active - influential bloggers.


Fig 1. Traffic pattern graph and list of top bloggers from "Venezuela" tracker.


Fig. 2. Keyword Trends of “Venezuela”, “America”, “Maduro” and “Economy Crisis”


Fig. 3. Positive and negative sentiment analysis on “Venezuela” tracker.

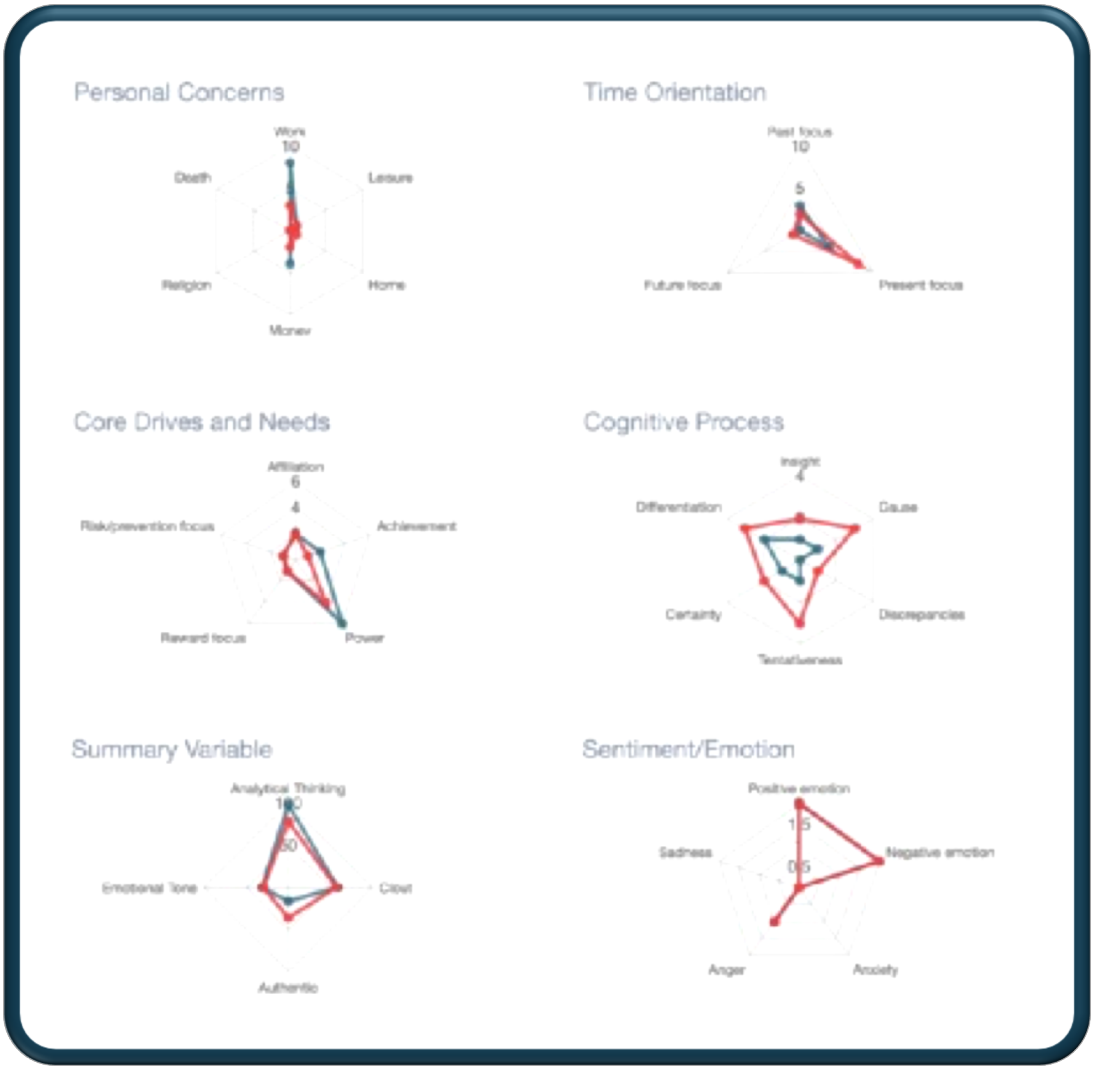



Fig. 4. Radar charts displaying tonality of blogs in “Venezuela” tracker.


Fig. 5. List of Influential bloggers and their trend.

7. Additional Blog Info provides additional information about a blog. It gives a day-of-the week average trend of a blog that helps in determining if the blog is a professional blog or a hobby blog. Also provided are monthly posting trend and sentiments for the past three years to determine the variation in activity and emotions. A list of URLs and domains mentioned in the blog is provided to know the source of information. Fig 6 provides a snapshot of additional blog info page.

# Features

## Topic Modeling

Topic modeling is a type of statistical modeling for discovering the abstract “topics” that occur in a collection of documents. Here, each document is represented as a distribution over topics and each topic is represented as a distribution over words. Latent Dirichlet Allocation (LDA) is a popular example of topic model and is used to classify text in a document to a particular topic. In LDA, each document may be viewed as a mixture of various topics where each document is considered to have a set of topics that are assigned to it via LDA. The topic distribution is assumed to have a sparse Dirichlet prior. The sparse Dirichlet priors encode the intuition that documents cover only a small set of topics and that topics use only a small set of words frequently. There are several scenarios where topic modeling can be very useful, such as text classification and uncovering themes in a text corpus.  We use topic models in blogs to discover latent semantic structures and identify themes from blog posts.

## Clustering

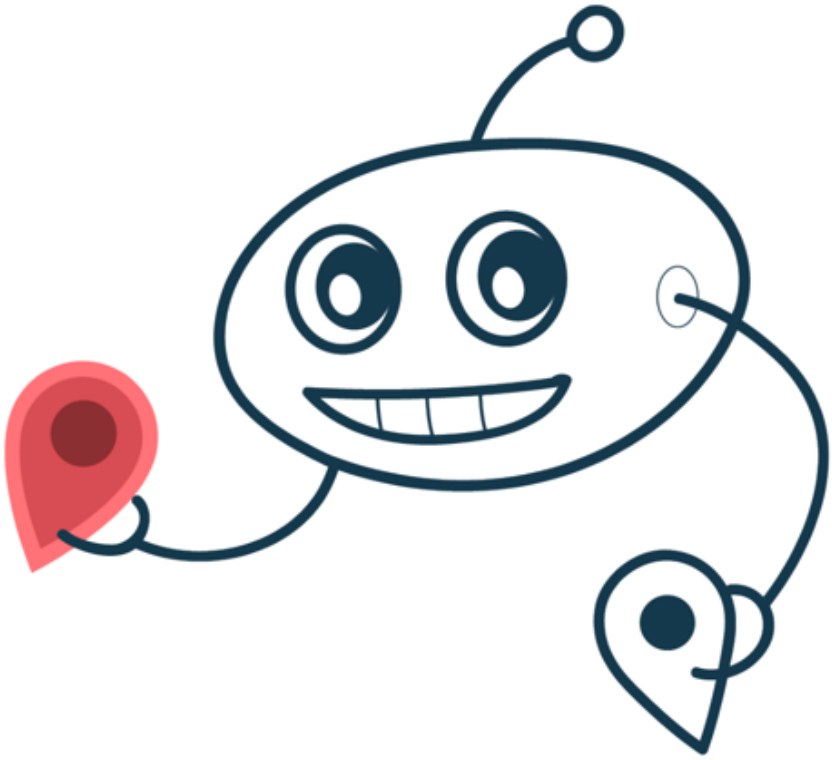
Document clustering is the organization of a large amount of text documents into a small number of meaningful clusters based on content similarity. Each cluster here represents a specific topic. Further, these clusters have similar documents where documents within the same partition exhibit higher degree of similarity among each other than to any other document in any other partition. There are several algorithms in computational text analysis that perform clustering. The algorithm's goal is to create internally coherent clusters that are distinct from one another. Application where clustering can be highly useful are - blogs, news, and tweets, etc. Blogs, for instance, discuss many aspects through posts. Hence, automatic grouping/clustering based on content similarity reveals summarization of prominent topics discussed in posts

## Acknowledgments

This research is funded by the U.S. Office of Naval Research (N000141010091, N000141410489, N0001415P1187, N000141612016, and N000141612412), U.S. Air Force Research Lab, U.S. Army Research Office (W911NF-16-1-0189), U.S. Defense Advanced Research Projects Agency (W31P4Q-17-C-0059) and the Jerry L. Maulden/Entergy Fund at the University of Arkansas at Little Rock. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations. The researchers gratefully acknowledge the support.

## References

[1] “Stats — WordPress.com.” [Online]. Available: https://wordpress.com/activity/. [Accessed: 04-Apr-2017].
[2] T. Furukawa, M. Ishizuka, Y. Matsuo, I. Ohmukai, K. Uchiyama, and others, “Analyzing
reading behavior by blog mining,” in Proceedings of the National Conference on Artificial Intelligence, 2007, vol. 22, p. 1353.
[3] “BlogPulse,” Wikipedia. 08-Mar-2017.
[4] “Blogdex,” Wikipedia. 04-Nov-2016.
[5] N. Bansal and N. Koudas, “Blogscope: a system for online analysis of high volume text streams,” in Proceedings of the 33rd international conference on Very large data bases, 2007, pp. 1410–1413.
[6] “Technorati—the World’s Largest Blog Directory—is Gone,” Business 2 Community.
[Online]. Available: http://www.business2community.com/social-media/technorati-worldslargest-blog-directory-gone-0915716. [Accessed: 04-Apr-2017].
[7] N. Agarwal, H. Liu, L. Tang, and P. S. Yu, “Identifying the influential bloggers in a community,” in Proceedings of the 2008 international conference on web search and data mining, 2008, pp. 207–218.
[8] N. Agarwal, H. Liu, L. Tang, and S. Y. Philip, “Modeling blogger influence in a community,” Soc. Netw. Anal. Min., vol. 2, no. 2, pp. 139–162, 2012.



## Contact Info

nxagarwal@ualr.edu

facebook.com/cosmographers

cosmos.host.ualr.edu

twitter.com/cosmographers

instagram.com/cosmographers

2801 S. University Ave.
Little Rock, AR 72204, USA

NSF · Office of Naval Research ONR Science & Technology · Air Force Research Laboratory · U.S. Department of Homeland Security · DARPA · Science to Shape the Future of the Army, Army Research Office 1951