BroadcastSTAND: Clustering Multimedia Sources of News

Jason Zhang, Ai-Te Kuo, Nicole R. Schneider,
Jacob Peters, Hanan Samet

Introduction

The modern media landscape involves diverse channels, including newspapers, social media, radio, and TV

Overview of the NewsStand architecture, which traditionally focuses on online news articles and Twitter posts

Introduction of BroadcastSTAND as an extension to integrate radio and TV broadcasts into the clustering landscape

Objectives and Significance

Objective: Evaluate the viability of incorporating broadcast news into the NewsStand framework

Emphasis on the importance of gaining insights from diverse news sources for a more comprehensive understanding of current events

NewsStand Architecture

Original NewsStand architecture focuses on online news articles and Twitter posts

Utilizes techniques such as content analysis, clustering, and recommendation algorithms for news organization and retrieval

NewsStand demo:

https://player.vimeo.com/video/106352925

NewsStand system:

https://newsstand.umiacs.umd.edu/web/

BroadcastSTAND Framework:

A multimedia approach to news

BroadcastSTAND is an extension to NewsStand

Includes radio and TV broadcast transcripts to broaden the scope of news content

Growing significance of broadcast data, including podcasts and YouTube channels

Enriching the user experience with diverse news perspectives and audiovisual content

Method and Clustering Analysis Results

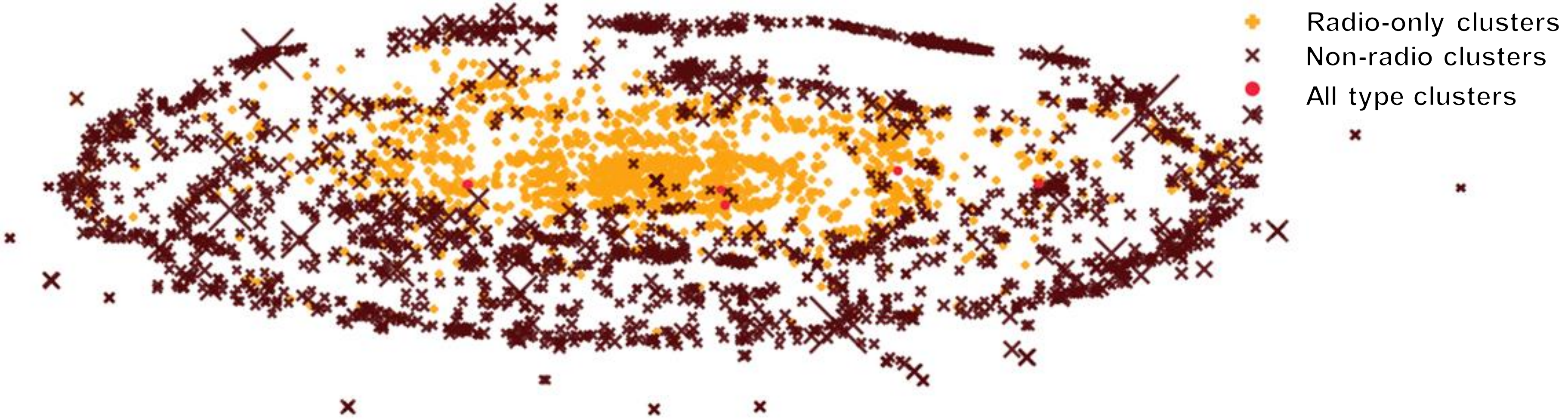
Use of PBS NewsHour transcripts for analysis

Identification of three types of clusters: Broadcast-Only, All type, and Non-Broadcast

Unexpected clustering outcome for broadcast news transcripts

Introduction of BroadcastStand as a response to address clustering challenges associated with broadcast data

Broadcast Media Clusters Mostly Separately from Other News Documents

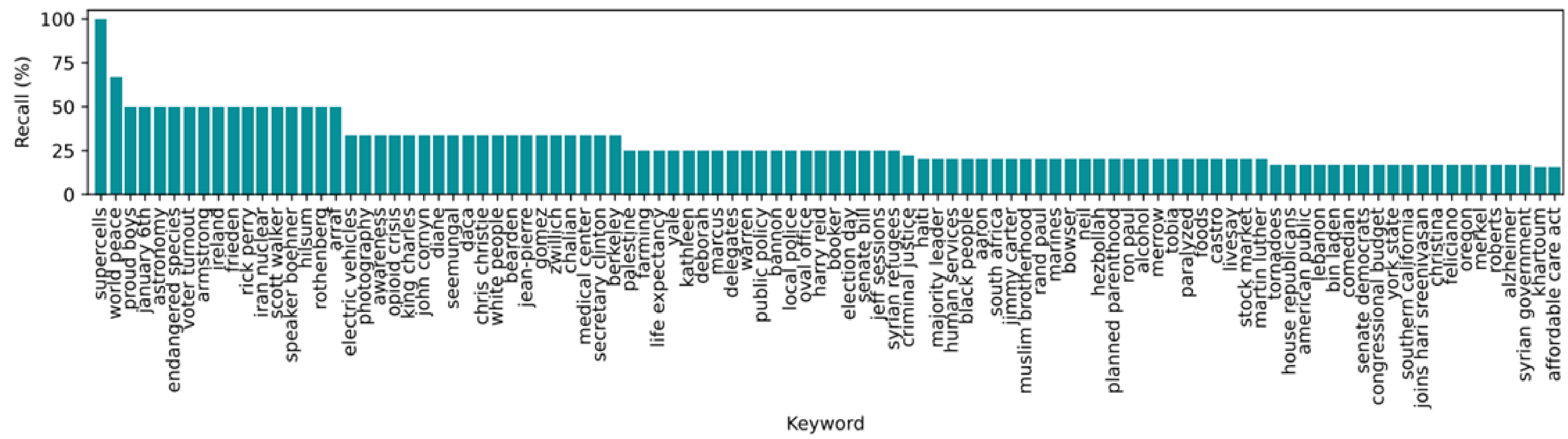


T-SNE visualization of high dimensional cluster space projected into 2 dimensions

Evaluation Metrics

Precision and recall metrics evaluation

The average recall for all the keywords was only 28.46%, while the average precision for all clusters was 99.74%



Conclusion and Future Work

References

[1] C. Bouras and V. Tsogkas. 2012. A Clustering Technique for News Articles Using WordNet. Know.-Based Syst. 36 (dec 2012), 115–128. [2] R. O. Duda and P. E. Hart. 1973. Pattern Classification and Scene Analysis. Wiley Interscience. [3] C. Fu, J. Sankaranarayanan, and H. Samet. 2014. WeiboStand: Capturing Chinese breaking news using Weibo. In LBSN'14. 41–48. [4] I. Gialampoukidis, S. Vrochidis, and I. Kompatsiaris. 2016. A Hybrid Framework for News Clustering Based on the DBSCAN-Martingale and LDA. In Machine Learning and Data Mining in Pattern Recognition. Springer Intl. Publishing, 170– 184. [5] N. Gramsky and H. Samet. 2013. Seeder finder - identifying additional needles in the Twitter haystack. In LBSN'13. 44–53. [6] A. Jackoway, H. Samet, and J. Sankaranarayanan. 2011. Identification of live news events using Twitter. In LBSN'11. 25–32. [7] M. Kretinin and G. Nguyen. 2022. Topic Modeling on News Articles using Latent Dirichlet Allocation. In 2022 IEEE 26th Intl. Conf. on Intelligent Engineering Systems (INES). 249–254. [8] E. Krokos, H. Samet, and J. Sankaranarayanan. 2014. A look into Twitter hashtag discovery and generation. In LBSN'14. [9] A.-T. Kuo and H. Samet. 2021. MusicStand: Listening to song lyrics using a map query interface. In Proceedings of the 29th ACM SIGSPATIAL Intl. Conf. on Advances in Geographic Information Systems. 446–449. [10] M. D. Lieberman and H. Samet. 2012. Supporting rapid processing and interactive map-based exploration of streaming news. In Proceedings of the 20th ACM SIGSPATIAL Intl. Conf. on Advances in Geographic Information Systems. 179–188. [11] M. D. Lieberman, H. Samet, and J. Sankaranarayanan. 2010. Geotagging: Using proximity, sibling, and prominence clues to understand comma groups. In Proceedings of 6th Workshop on Geographic Information Retrieval. [12] G. Quercini and H. Samet. 2014. Uncovering the spatial relatedness in Wikipedia. In Proceedings of the 22nd ACM SIGSPATIAL Intl. Conf. on Advances in Geographic Information Systems. 153–162. [13] G. Quercini, H. Samet, J. Sankaranarayanan, and M. D. Lieberman. 2010. Determining the spatial reader scopes of news sources using local lexicons. Proceedings of the 18th ACM SIGSPATIAL Intl. Conf. on Advances in Geographic Information Systems, 43–52.

[14] G. Salton and C. Buckley. 1988. Term-weighting approaches in automatic text retrieval. Information Processing & Management 24, 5 (1988), 513–523. [15] G. Salton, A. Wong, and C. S. Yang. 1975. A Vector Space Model for Automatic Indexing. Commun. ACM 18, 11 (nov 1975), 613–620. [16] H. Samet. 2014. Using minimaps to enable toponym resolution with an effective 100% rate of recall. In Proceedings of GIR'14. [17] H. Samet, M. D. Adelfio, B. C. Fruin, M. D. Lieberman, and J. Sankaranarayanan. 2013. PhotoStand: a map query interface for a database of news photos. PVLDB 6, 12 (Aug. 2013), 1350–1353. Also Proceedings of the 39th Intl. Conf. on Very Large Data Bases (VLDB). [18] H. Samet, M. D. Adelfio, B. C. Fruin, M. D. Lieberman, and B. E. Teitler. 2011. Porting a web-based mapping application to a smartphone app. In Proceedings of the 19th ACM SIGSPATIAL Intl. Conf. on Advances in Geographic Information Systems. 525–528. [19] H. Samet, J. Sankaranarayanan, M. D. Lieberman, M. D. Adelfio, B. C. Fruin, J. M. Lotkowski, D. Panozzo, J. Sperling, and B. E. Teitler. 2014. Reading news with maps by exploiting spatial synonyms. Commun. ACM 57, 10 (Oct. 2014), 64–77. [20] H. Samet, B. E. Teitler, M. D. Adelfio, and M. D. Lieberman. 2011. Adapting a map query interface for a gesturing touch screen interface. In Proceedings of the Twentieth Intl. Word Wide Web Conf. (Companion Volume). 257–260. [21] J. Sankaranarayanan, H. Samet, B. Teitler, M. D. Lieberman, and J. Sperling. 2009. TwitterStand: News in tweets. In Proceedings of the 17th ACM SIGSPATIAL Intl. Conf. on Advances in Geographic Information Systems. 42–51. [22] N. R. Schneider and H. Samet. 2021. Which Portland is It? A Machine Learning Approach. In Proceedings of the 5th ACM SIGSPATIAL Intl. Workshop on LocationBased Recommendations, Geosocial Networks and Geoadvertising (LocalRec '21). Article 8, 10 pages. [23] M. Steinbach, G. Karypis, and V. Kumar. 2000. A Comparison of Document Clustering Techniques. Proceedings of the Intl. KDD Workshop on Text Mining (06 2000). [24] B. Teitler, M. D. Lieberman, D. Panozzo, J. Sankaranarayanan, H. Samet, and J. Sperling. 2008. NewsStand: A new view on news. In Proceedings of the 16th ACM SIGSPATIAL Intl. Conf. on Advances in Geographic Information Systems. 144–153. [25] G. Zhou and J. Su. 2002. Named Entity Recognition Using an HMM-Based Chunk Tagger. In ACL '02. 473–480. [26] Z. Zhou, J. Qin, X. Xiang, Y. Tan, Q. Liu, and N. N. Xiong. 2020. News Text Topic Clustering Optimized Method Based on TF-IDF Algorithm on Spark. Computers, Materials & Continua 62, 1 (2020).