maiti_2019_spatial_aggregation_facilitates_discovery_of_spatial_topics

Year

2019

Author(s)

Maiti, Aniruddha and Vucetic, Slobodan

Title

Spatial Aggregation Facilitates Discovery of Spatial Topics

Venue

ACL

Topic labeling

Manual

Focus

Secondary

Type of contribution

Established approach

Underlying technique

(Presumed) manual labeling

Topic labeling parameters

\

Label generation

General topic themes are assumed to be manually provided by the authors.

Additionally, since data (each tweet) was aggregated by:

- k-means clustering on the latitude and longitude information
- time intervals (into 12 days)

The extracted topics can also be flagged as spatial (NYC) or temporal (Halloween) depending on the kind of tweets associated with them:

"We selected the 10 most frequent words in each discovered topic and labeled each tweet from the corpus based on the presence of these words. If a tweet contains any of the 10 words it is assigned to the corresponding topic. We call all tweets assigned to the given topic the positive tweets. If the topic is strongly spatial, we would expect the assigned tweets to be strongly spatially clustered. If the topic is strongly spatio-temporal, we would expect the assigned tweets to cluster within a particular spatio-temporal area"

Table 2: Evaluation of the topic quality using SaTScan

Topic General Theme	Deviation (Δ)	Topic Type
Power	26504.53	Temporal
NYC	25282.17	Spatial
NFL	12275.18	Temporal
Presidential Debate*	11089.34	Temporal
Snow	8624.95	Temporal
New Jersey*	8355.10	Spatial
Halloween*	7679.58	Temporal
Pennsylvania*	6728.94	Spatial
NYC Airport*	6424.54	Spatial
Weather	2220.64	Temporal

Table 3: General theme of topics and related words

Topics	Words
Power	power sandy generator trees electricity tree open lights safe hurricane
NYC	york brooklyn nyc park manhattan city square mta island halloween
NFL	cowboys steelers romo giants harden church redskins touchdown eagles party
Snow	snow snowing cold weather delay boone wind blizzard snowed outside
Weather	barometer humidity temperature mph wind rain blacksburg steady wnw rising

† Offensive words are removed

For example: “We also found that large metropolitan areas such as New York City, Philadelphia, and Pittsburgh are represented as separate spatially distinct topics”

Figure 9: NYC airport-specific distinct topic

And: “We identified several purely temporal topics in this way, including the Halloween topic. Figure 10 also contains another temporally distinct topic associated with the 2012 Country Music Association (CMA) Award event that happened on the same day”

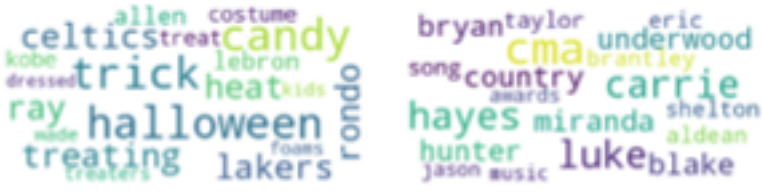

Figure 10: Halloween and CMA temporal topics

Motivation

\

Topic modeling

NMF, (LSA and LDA for comparison)

Topic modeling parameters

Nr. of topics (k): 500

α: 0.1

ρ: 0.5

Nr. of topics

500

Label

Single and multi-word theme + Temporal/Spatial binary label

Label selection

\

Label quality evaluation

\

Assessors

\

Domain

Domain (paper): Spatial aggregation (for topic modeling)
Domain (corpus): Social media (Twitter)

Problem statement

Spatial aggregation refers to merging of documents created at the same spatial location. We show that by spatial aggregation (or pooling) of a large collection of documents and applying a traditional topic discovery algorithm on the aggregated data we can efficiently discover spatially distinct topics.

By looking at topic discovery through matrix factorization lenses we show that spatial aggregation allows low rank approximation of the original document-word matrix, in which spatially distinct topics are preserved and non-spatial topics are aggregated into a single topic.

Our work indicates that different forms of document aggregation might be effective in rapid discovery of various types of distinct topics from large collections of documents.

Corpus

Origin: Twitter
Nr. of documents: 4.7 million tweets (Organised into 2400 patio-temporally aggregated documents)
Details: Hurricane Sandy (2012) Twitter corpus spanning 12 days

Document

A single tweet geotagged to one of 13 states along the East Coast of the U.S.

Pre-processing

- Transformed all characters to lowercase
- Removed stopwords and special characters.
- Excluded repetitive letters that convey enthusiasm (e.g., birthdayy, birthdayyy, birthdayyyy).
- TF-IDF document-word matrix is constructed using the 20, 000 most frequent words in the corpus.

Spatio-temporal clusters

- k-means clustering on the latitude and longitude information for each tweet is used to identify 200 cluster centers in space. Each tweet is assigned to its nearest cluster center for spatial aggregation.
- In addition to the k = 200 spatial clusters we divided the time interval into 12 days, resulting in a total of 2, 400 spatio-temporally aggregated documents.

```
@inproceedings{maiti_2019_spatial_aggregation_facilitates_discovery_of_spatial_
topics,
    title = "Spatial Aggregation Facilitates Discovery of Spatial Topics",
    author = "Maiti, Aniruddha  and
      Vucetic, Slobodan",
    booktitle = "Proceedings of the 57th Annual Meeting of the Association for
Computational Linguistics",
    month = jul,
    year = "2019",
```

```
    address = "Florence, Italy",
    publisher = "Association for Computational Linguistics",
    url = "https://aclanthology.org/P19-1025",
    doi = "10.18653/v1/P19-1025",
    pages = "252--262",
    abstract = "Spatial aggregation refers to merging of documents created at the same spatial location. We show that by spatial aggregation of a large collection of documents and applying a traditional topic discovery algorithm on the aggregated data we can efficiently discover spatially distinct topics. By looking at topic discovery through matrix factorization lenses we show that spatial aggregation allows low rank approximation of the original document-word matrix, in which spatially distinct topics are preserved and non-spatial topics are aggregated into a single topic. Our experiments on synthetic data confirm this observation. Our experiments on 4.7 million tweets collected during the Sandy Hurricane in 2012 show that spatial and temporal aggregation allows rapid discovery of relevant spatial and temporal topics during that period. Our work indicates that different forms of document aggregation might be effective in rapid discovery of various types of distinct topics from large collections of documents.",
}
```

#Thesis/Papers/Initial