TopicFields: Spatiotemporal Visualization of Geo-tagged Social Media With Hybrid Topic Models and Scalar Fields

Ruofei Du, Xiaoxu Meng, Eric Krokos, Shuo Li, Amitabh Varshney
Augmentarium, Department of Computer Science and the Institute for Advanced Computer Studies (UMIACS)
University of Maryland, College Park

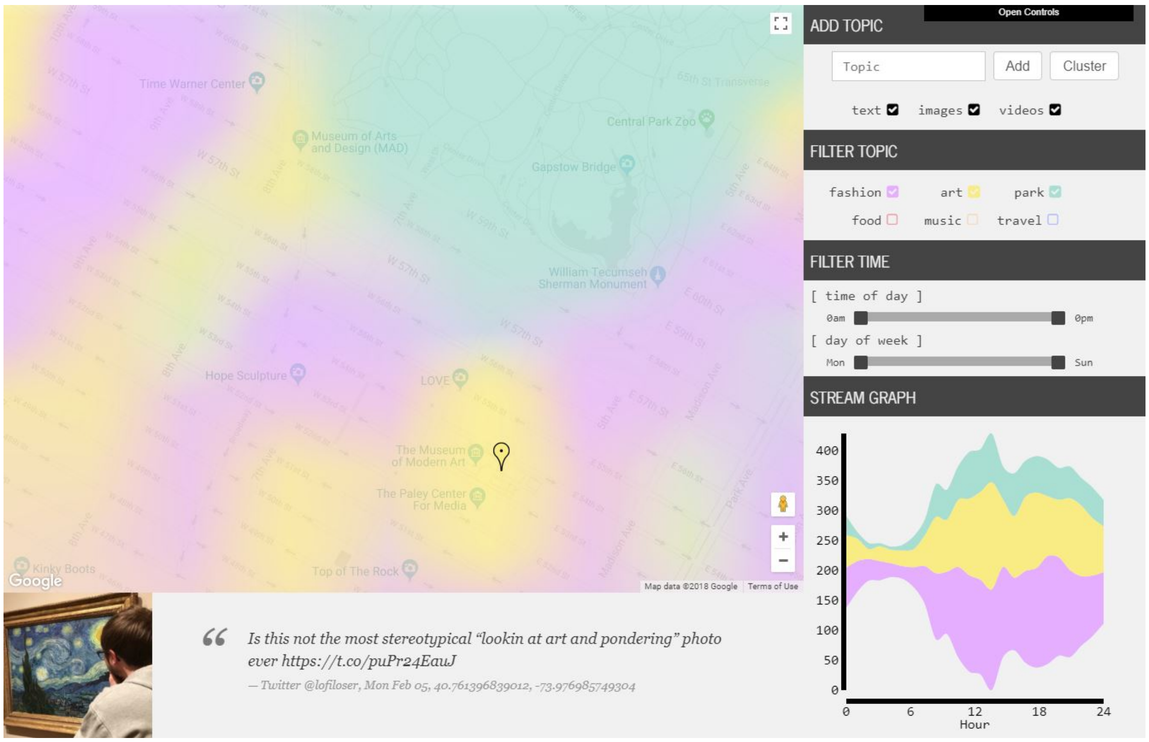


Fig. 1. A screenshot of the *TopicFields* system for visualizing approximately one million geo-tagged social media with hybrid topic models and scalar fields (this figure is best visualized on a computer screen and does not reproduce as well on a printout). The top-left map view is overlaid with an interactive scalar map, with each color indicating a different cluster of topics: fashion, art, and park. We show the ground truth labels such as *Central Park* and *The Museum of Modern Art* from *Google Maps* for reference. The detail view at the bottom shows the corresponding social media text, image, or video, as the user explores and clicks on the map. The control panel on the right allows the user to select, add, and modify the topics to explore. The user could use the "Cluster" button to open the topic matrix diagram (Fig. 4) and adjust the clustering results. The stream graph shows the volume of social media of different topics over the queried time. The corresponding topic matrix diagram is shown in Fig. 5.

Abstract—The state-of-the-art machine learning models can extract hundreds of high-level topics from large-scale social media corpus. Nevertheless, it remains a challenge for the users to interpret the comprehensive distribution of multiple topics in a space-time setting. In this paper, we present *TopicFields*, an interactive system to explore, aggregate, and visualize geo-tagged social media using hybrid topic models, scalar fields, and stream graphs. In the data processing stage, we apply two machine learning models *Word2Vec* and *Inception-v3* to the data and address the relationships among the extracted topics by rearranging them via spectral ordering. In the visualization stage, we allow users to interactively select the preferred topics and alter the transfer function for visualizing the social media on a map with levels of detail. Our system, *TopicFields*, can efficiently estimate the kernel density distribution and visualize the scalar fields of the user-selected topics on a map on the GPU. In addition, we use temporal filters and stream graphs to enhance comprehensibility of the data over time. Here we present the system and its architecture that ingests geo-tagged *Instagram* and *Twitter* messages, extracts topics, hierarchically clusters, and facilitates their interactive visualization on a map. We demonstrate the effectiveness of *TopicFields* with several potential use cases. We envision our system will be useful for visual analytics of geo-tagged social media, tourism itinerary planning, and business intelligence, and mixed reality social media platforms [9].

Index Terms—spatio-temporal visualization; topic model; geographic information system; social media; spectral clustering

1 Introduction

Manuscript drafted on 6/6/2018 for Social Street View and Geollery.
Please cite this and/or full dissertation if you find it useful: Du, Ruofei "Fusing Multimedia Data Into Dynamic Virtual Environments". University of Maryland, College Park, Ph.D. Dissertation, pp. 213, 2018.
`https://doi.org/10.13016/2pnc-vfg0`

Social media today plays an increasingly significant role in our daily lives because of its interactivity, popularity, and social relevance. Every day, billions of users create, share, and exchange messages about their life on websites such as *Twitter*, *Facebook*, and *Instagram*. The influence of social media cannot be understated, as it covers a wide range of topics such as assessment of restaurants and parks, fashion, museums and the arts, and sporting events. The primary motivation behind our research presented in this paper is to use the power of exploratory visual analytics to glean valuable insights from the spatio-temporal social media that can help everyday users.

With recent advances in the deep learning models applied to natural language processing and computer vision, machines are now able to extract hundreds of topics or categories out of the social media data. Nevertheless, there are several problems with directly presenting the topics to the end user:

1. **Topic duplication**. Some, but not all, of the extracted topics could be closely related to each other. In addition, neural networks trained for text and images usually use different classes or labels in the results. From visualization's perspective, we would like to aggregate similar topics from hybrid models. For example, *artists*, *painting*, *art* all refer to *art*. Worse still, machine learning algorithms may fail to classify some features to a specific topic, which requires additional input from the users.

2. **Information overload**. Low-level features such as the unigrams or the image classification labels usually results in hundreds of labels. The variety of results are usually too overwhelming for the user. To deal with this problem, we use spectral clustering and provide the cluster with top frequencies for the user to reduce the scope of their visual search.

3. **Diversity**. Previous visual analytic approaches have investigated heat map visualization of a single topic on a map [20, 23], or heat maps of positive or negative emotions on a map [19]. The challenge of effectively visualizing a large number of topics that are implicit in a diverse spatio-temporal social media has thus far not been adequately addressed.

In this paper, we present *TopicFields* (Figure 1), an interactive visualization system of geo-tagged social media to address these three critical problems. Our goal is to understand and correlate the social topics that occur in the real world at various geographical locations over time.

The main contributions of our work are:

- a novel web-based framework for analyzing, aggregating, and visualizing multiple topics from large-scale geo-tagged social media data,

- clustering hybrid machine learning classification results with spectral ordering algorithm, such as an interactive matrix diagram,

- an efficient and interactive GPU-driven visualization algorithm for visualizing multi-variate scalar data with kernel density estimation and non-linear normalization methods.

We organize this paper as follows. First, we examine related work in Section 2. We present an overview of the system architecture and workflow in Section 3. The processes for data mining, spectral clustering, and topic matrix diagram are described in Section 4. The GPU-driven scalar field visualization algorithm is described in Section 5. In Section 6, we present several potential use cases. In Section 7, we discuss the advances and limitations of the current prototype. We conclude the paper and present future directions in Section 8.

2 Related Work

Our work builds upon the rich literature of previous research on geospatial visualization of social media and topic models for text and visual information.

2.1 Geo-spatial Visualization

Visualizing information in a geo-spatial manner has been around for as long as there have been maps. The ability to map, understand, see patterns, and draw conclusions from information presented in a spatially significant way is potent and intuitive. For a single topic, previous research has used heatmaps with temporal filters to visualize the spatial density.

MacEachren *et al.* [21] present one of the earliest systems for visualizing the heat maps of health reports on a map. Their system offers time-series animation and linked geographic brushing to assist domain experts in exploring the data. Their further work, *SensePlace2*, [22] presents a geo-spatial visualization of Twitter messages with user-defined queries, time filters, spatial filters, and heap maps of tweet frequencies.

An early example of visualizing geo-tagged social media can be seen in TwitterStand [30] and NewsStand [37], where Twitter and news information is distributed on a map of the world as icons or images. In this way, viewers can see what information is available, where it originates from, and the density as well as the absence of the information. Chae *et al.* [5] present a social media analysis system with message plots on a map, abnormality estimation charts, and tables for message content and topic exploration with Latent Dirichlet Allocation (LDA) [3] and Seasonal-Trend Decomposition (STL) [7]. Stefanidis *et al.* [35] visualize Twitter traffic as gridded heat maps. Maciejewki *et al.* [23] employ a multivariate kernel method to create heat maps for visualizing the density of geo-referenced data with a gradient between red and blue. Our heat map visualization is inspired by their formulation but advances the previous research by rendering the multiple scalar fields representing various classes.

Lucasczyk *et al.* [20] apply the topological notion of Reeb graphs to identify hotspots as areas of relatively high event density using kernel density estimates. Lu *et al.* [19] present a novel framework for sentiment modeling of geo-tagged social media and heat map visualization with fixed bandwidth kernel density estimation (KDE) [34]. Hao *et al.* [12] visualize sentiments with color-coded text labels on a map. Scharl *et al.* [31] present an interactive system that allows users to analyze the extracted topics and sentiments with trend charts, temporal controls, and heat maps of the sentiments. Chen *et al.* [6] devise an interactive visual analytics system to investigate the movement patterns and their semantic implication for social media users. Kim *et al.* [16] further visualize the spatio-temporal patterns in the data by employing flow visualization techniques and a 3D gravity model. Using domain-specific knowledge, previous research has analyzed geo-tagged social media to improve emergency responses [40, 45], assist disease control [15], understand the dynamics of neighborhoods [8] and cities [39, 44], and travel route planning [17].

To the best of our knowledge, our work is the first to offer social-media-topic extraction, spectral ordering of related topics, and exploratory visualization of the scalar fields of multiple topics.

2.2 Topic Models

For text information, Latent Dirichlet Allocation (LDA) [3] is one of the most popular and successful topic models. Further research involves numerous variants of LDA algorithms, such as labeled LDA [29], TM-LDA [42], spatial LDA [41], online LDA with variational Bayes [14], and correlated LDA [2]. Recently, neural networks such as Word2Vec with the Continuous Bag-of-Words model and the Skip-Gram model [24], have been widely adopted to extract vectorized features from texts. Tremendous progress also adopted Naive Bayes [18], support vector machines [28], and deep neural networks [36] for sentiment classification in texts. However, even though the accuracy of topic models keep improving, applying topic models on social media data still requires significant efforts in preparing the training data, or accurate filtering of topic features.

Instead of applying topic models to directly extract topic words which may be not suitable to conclude user interests, we turn to another effective solution which first leverages NLP model to extract a relatively large size of candidate topics, then analyzes the correlation of the candidate topics with the sophistical neural network word2vec model

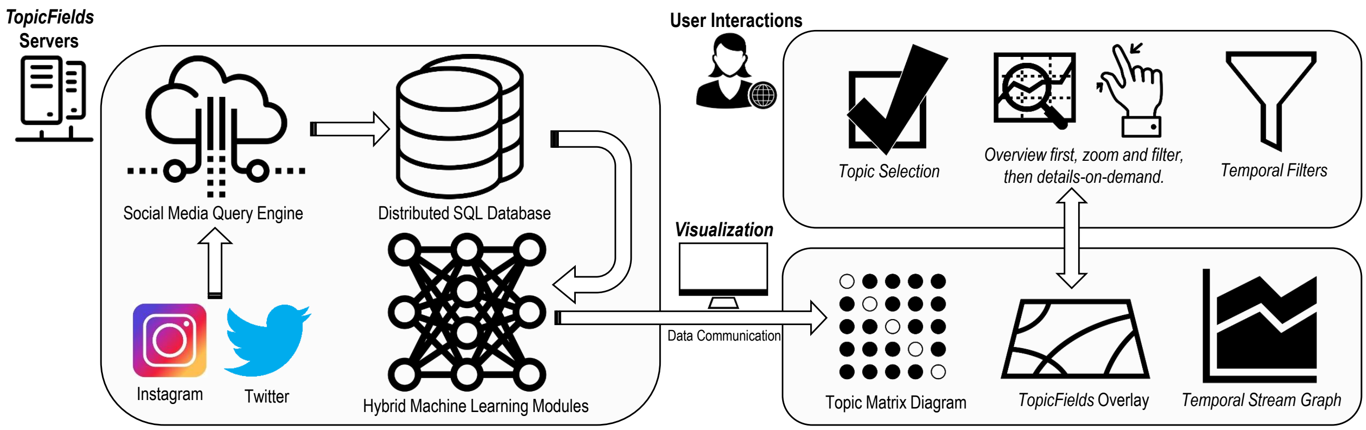


Fig. 2. The architecture and workflow of the *TopicFields* system. The servers consist of social media query engine, distributed SQL databases, and hybrid machine learning modules for topic classification and clustering from text and images. After the aggregation of topics, the server transfers the data to the client visualization system to present the topic matrix diagram, topic fields overlay, and temporal stream graphs. The users can select the desired topics for visualization, have an overview of the distribution, zoom in and filter by keywords and time, and then visualize the details on demand.

[25], and lastly leverages spectral clustering algorithm to obtain the topics which are more consistent to human desire. Moreover, We incorporate the spectral ordering into word2vec model, which facilities users to select the topics of particular interests.

3 System Overview

As shown in Fig. 2, our system consists of a social-media query engine, distributed SQL databases, a machine-learning server that runs hybrid modules, and a client-side visualization system.

First, the user is provided with an interactive map. The user can pan around, zoom in or out to select the region of interest. The boundary of the map is then sent to the social media query engine. In the current prototype, the query engine is able to scrape a few hundred geo-tagged social media from external sources such as *Twitter* and *Instagram* in around a second, but will mostly query from the offline databases. The social media are stored and organized in distributed SQL databases. Next, the machine learning module extracts topic features from the text and images, as described in Section. 4.2. It uses the spectral ordering algorithm on the groups of topics and sends the matrix to the user for filtering. On the client side, the majority groups of the topics after spectral ordering are visualized as a matrix diagram. The user is able to select, add, and remove topic features from the groups and interactively visualize the topic fields on the map.

For exploratory visualization, we have followed the Shneiderman design principle [33] *overview first, zoom and filter, then details-on-demand*. First, we present the map view with topic fields visualization to offer the user an overview of the social media according to the topics. The stream graph offers the user an overview of the volume of the topics over time. Finally, the user is able to zoom in and click on the map view to query proximate social media. We have linked our system with Google Street View to provide an immersive experience to explore the map.

4 Data Processing

In this section, we present the process of data mining, feature extraction, and spectral clustering.

4.1 Data Mining

Our geo-tagged social media scraper is a back-end program written in PHP. Tuchinda *et al.* [38] proposed to model web services as information sources in a mediator-based architecture and have built an exemplary application, *Mashup*. Using a similar architecture, our system is able to integrate information from several web services. In this prototype, we use Twitter and Instagram as the major sources of social media in our proof-of-concept system. We collected the following four types of data:

1. **Geo-spatial and textual location** including latitude and longitude coordinates, street names, and user-tagged location name.
2. **Caption and tags** containing text information of the Tweets or Instagram messages.
3. **Publication time-stamp** containing the exact date and time when published.
4. **User comments and likes** reflecting the popularity level.

We have investigated two major districts on the eastern coast of the United States: *the Manhattan District of New York*, and *the District of Columbia* (Washington D.C.). Over three months from *December* 2017 to *March* in 2018, we have collected $946,856$ *Twitter* and *Instagram* messages with specific geographical labels and publication time-stamps from the public domain, with $589,902$ in the *Manhattan District* and $356,954$ in the *District of Columbia*. The data was scraped using a flood-fill algorithm inspired by Shen *et al.* [32].

4.2 Feature Extraction

Our data consists of two types of social media: text and images. Sometimes, we obtain videos via the social media query engine. Instead of using video data, we use the first frame or the user-defined thumbnail for feature extraction.

As for text messages, we first experimented with the well-known topic model, Latent Dirichlet Allocation (LDA) [4], to extract accurate clusters of topics from our data. However, the state-of-the-art topic models cannot guarantee generation of clean results. For instance, the top three topics we have extracted using LDA [4] are:

enhance, ishootfilm, bend, contemporary, dance, woo, retrospective, ...

sadly, indore, holidayshopping, foodbaby, prk, ...

minidachshund, ana, bulking, busstopdinernyc, fridays, ...

These topics can hardly be used directly for visualization. This led us to the question: can we use machine learning to cluster the most frequent keywords, provide visualization results, and allow the visual analyst to select the desired topics she wants?

Towards this direction, we use the Natural Language Toolkit (NLTK) to extract to top 300 words from the entire social media datasets and

Fig. 3. The top three classification results by applying the *Inception-v3* deep neural network to our dataset. Our topic feature vector of images consists of the entire hierarchical tree of the top three labels from the *WordNet*. For example, pizza belongs to dish, nutriment, and food.

apply the *Word2Vec* neural network to the topics to compute the feature vector for each text data.

Still, 300 features are too much for a visual analyst to interactively explore, so we decided to compute the similarity between each pair of features and use spectral clustering (Section 4.3) with image labels to help further aggregate the topics into topic classes.

For images, we use the *Inception-v3* model on the image dataset to compute the top three classification results. The results above the 80^{th} quantile are used for extracting the feature vectors. We concatenate all labels from the hierarchical tree to find the topics for the image social media. For example, as shown in the first image of Fig. 3, the feature vectors are "pizza, pizza pie / dish / nutriment, nourishment, nutrition, sustenance, aliment, alimentation, victuals / food, nutrient".

4.3 Spectral Clustering

Our topic features consist of unigrams and the image classification labels, which can be 300-dimensional vectors. Some, but not all, of the extracted features could be closely related to each other. When features are in arbitrary order along the x-axis of the design widget (Fig. 4(a)), assigning a meaningful characteristic feature vector may require numerous control points, explicitly defining the value for each dimension. Although the reordering of features does not add to the possible visualizations that can be generated using the machine-learning-assisted approach, the usability issue must be addressed to benefit from the power of high-dimensional representations. We address the relationships among features by rearranging them using spectral ordering, which sorts the features by the eigenvector of the second smallest eigenvalue of a graph Laplacian. First, the normalized Laplacian matrix is generated based on feature-to-feature similarity; then, the eigenvector associated with the second smallest non-negative eigenvalue (the Fiedler vector) is calculated, as shown in Fig. 4(c); finally, the features are sorted based on their values in the Fiedler vector. The result is shown in Fig. 4(b) is an ordering of features where neighboring features are similar. The Fiedler vector and other eigenvectors associated with small eigenvalues also form the basis of spectral clustering. Many pairs of the 300 features are indeed highly correlated as can be seen by several dark pixels. Nevertheless, an arbitrary order of features does not take advantage of such correlations, resulting in a disorganized similarity matrix in Fig. 4(a).

After rearranging the topic features, we cluster the adjacent features using disjoint-set data structures and partition refinement algo-

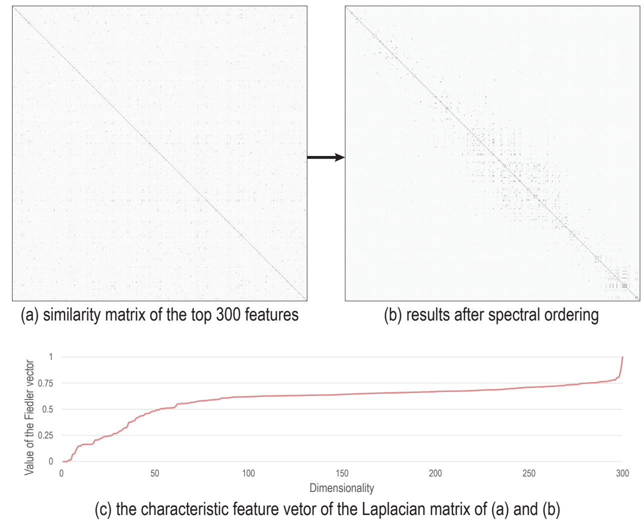


Fig. 4. Spectral ordering of the topic features; (a) shows the similarity matrix of the top 300 features according to the frequencies, while (b) shows the results after spectral ordering (best viewed on a computer screen). Note that spectral ordering eliminates the randomness of the data and clustering similar groups together. (c) shows the Fiedler vector of the normalized Laplacian matrix of (a) and (b).

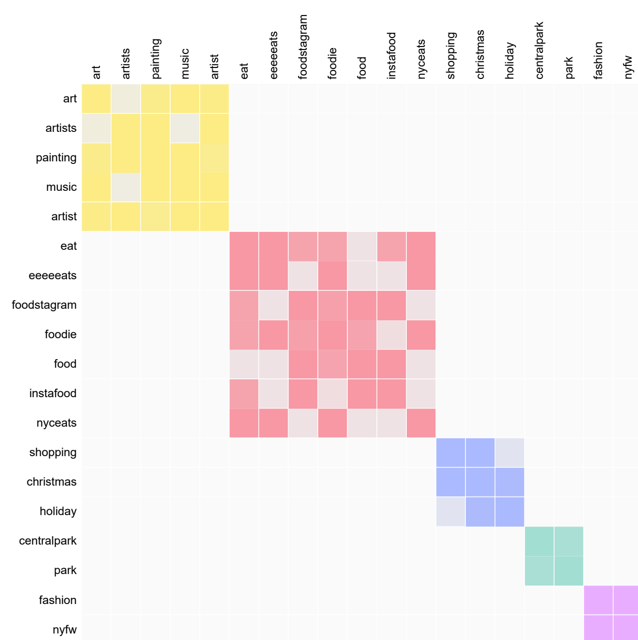
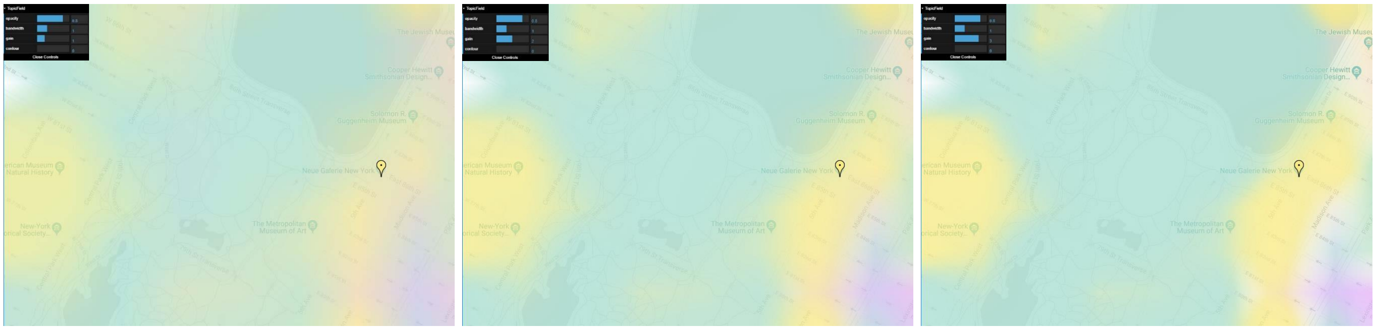


Fig. 5. The resulting matrix diagram after spectral clustering with the following queries: *art*, *food*, *shopping*, *park* and *fashion*.



(a) without nonlinear normalization, k = 1 (b) nonlinear normalization with k = 2 (c) nonlinear normalization with k = 3

Fig. 6. Visualization of the topic fields with different gain factors. (a) shows the baseline visualization using the Gaussian PDF without nonlinear normalization, (b) shows the nonlinear normalization result with a gain factor of 2, and (c) shows the result with a gain factor of 3. The nonlinear normalization significantly increases the contrast between different clusters of topics.

rithms [27]. The pairs of features whose similarity is greater than $\sigma = 0.7$ and a distance of the spectral ordering is smaller than $\delta = 0.5$ are clustered into one disjoint set. We allow the users to change σ and δ in the control panel for the matrix diagram.

The spectrally-ordered similarity matrix places similar features closer together, resulting in large colorful blocks of various sizes along the diagonal. Thus an accessible feature order allows user-directed selection of similar topics using fewer operations in the control panel.

5 Topic Fields Visualization

Our algorithm visualizes the scalar field of user-filtered topics of geo-tagged social media over a map.

Given N geo-tagged social media over the map, with locations $g_i, i = 1,2,\cdots,N; g \in G$, suppose each social media is assigned to a set of M topics $T : \{t_1, t_2, \cdots, t_M\}$. Each topic consists of multiple unigrams. We classify a social media to belong to a topic if and only if the unigram appears in the caption, tag, or the hierarchical tree of the image classification results. We limit $M \leq 6$ in our system, since the capacity of short term memory for processing information is usually seven, plus or minus two [26], so is the number of colors distinguishable in visualization schemes [13].

First, we generate a grid mesh with $W \times H$ vertices and assign a scalar vector f to each vertex. For vertex centered at g_v, we apply the kernel density estimation within its circle of radius R:

$$f_t(g_v) = \frac{1}{NR} \sum_{i=1}^{N} \mathbf{K}\left(\frac{D(g_v, g_i)}{R}\right) \tag{1}$$

, where the kernel function K could be any non-negative function that integrates to one. However, we prefer the kernel functions that smoothly model the falloff of the spatial distribution, such as *Gaussian*, *Quatic*, *Epanechnikov*, or *Triweight* functions. Here we use the *Gaussian Probability Density Function* (PDF) with a bandwidth R:

$$\mathbf{K}(R) = \frac{1}{\sqrt{2\pi}} e^{-R^2/2} \tag{2}$$

Suppose we have a transfer function to colorize each topic t with the color c_t. For each vertex, we can blend the topic fields over the grids by:

$$c = \sum_{t \in T} c_t \cdot \mathscr{N}(f_t) \tag{3}$$

, where $\mathscr{N}(\cdot)$ is a nonlinear normalization method to the scalar fields to emphasize centralized topics:

$$\mathscr{N}(f_t) = \frac{g(f_t)}{\sum_{t \in T} \mathscr{N}(g(f_t, k))} \tag{4}$$

This nonlinear normalization operator partitions the map into different clusters consisting of different topics. In particular, we apply the gain function $g(x,k)$ employed in modern ray tracing frameworks such as *Pixar Renderman* [1]:

$$g(x,k) = \begin{cases} \frac{1}{2} \cdot (2x)^k, & x < 0.5, \\ 1 - \frac{1}{2} \cdot (2-2x)^k, & x \geq 0.5 \end{cases} \tag{5}$$

, where we call k as the gain factor to adjust the contrast of the scalar field. We plot the function in Fig. 7.

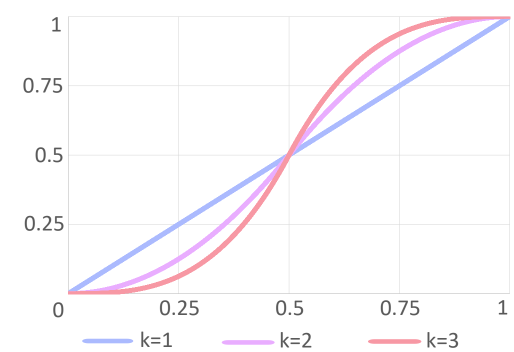


Fig. 7. The gain function remaps the unit interval into the unit interval. It maps 0.5 to 0.5 while expanding expanding the sides and compressing the center.

By default, we take $k = 2.0$. However, we allow the user to change the gain factor using a control panel powered by *dat.gui* for the WebGL rendering. In this way, we assign a scalar vector to every vertex on the planar mesh. Typically, we use 32×32 vertices for the current boundary on the map.

In the fragment shader, we interpolate the normalized scalar field using Lagrange Bicubic sampling [43] and colorize the scalar field using the user-defined colors. Finally, we efficiently render the Topic Fields using WebGL in a modern browser in real time. We show the visualization results with different gain factors in Fig. 6.

6 Use Cases

With *TopicFields* system, we demonstrate two potential use cases for trip planning and searching with temporal filters.

6.1 Trip Planning

First, we demonstrate how *TopicFields* could help a user to plan a short trip near the central park region. Suppose that the user has decided to explore the central park, but has no idea where to go for food and shopping. So the user inputs *park*, *food*, and *shopping* into the topic query box. With the spectral ordering algorithm, the user quickly

(a) topic clustering for food, shopping, and park

(b) seek for food near the central park with *TopicFields*

(c) linking to the street view to identify the spot

(d) seeking for social media in the central park

(e) finally, looking for places for shopping nearby

Fig. 8. These figures show the procedure for planning a trip via *TopicFields* near the central park.

aggregates three clusters including 12 features into the query engine. With the *TopicFields* visualization, the user could quickly identify geo-tagged social media that relates to parks, as shown in Fig. 8(d):

> *#centralparksouth #centralpark #nature #flowers #flowersofinstagram #ferns #flowersofcentralpark #spring #nyc*

From the topic fields, food is distributed everywhere in the map. The user could simply select food, and query close to the central park. One of the results is shown in Fig. 8(b):

> *This is what I am talking about! #food #newyork #follow #Instagram #foodporn #sandwich #love #Angelas #centralpark #Saturday*

The user could further drag the street view pacman on the right-bottom of the map to identify the environment near the spot: it seems like a real sandwich shop.

However, the places for shopping seems to be a little far away from the park, given the small purple distribution on the map. Without the *TopicFields* visualization, one may click anywhere on the map to seek for social media that mentions *shopping*, while the *TopicFields* visualization quickly identifies the *5th Ave* as the concentrated places for shoppers:

> *essentials? #louisvuitton #gucci #guccigang #dolcegabbana #guccicommunity #essentials #shopping #manhattan #nyc #5thavenue #streetstyle*

6.2 Searching with Temporal Filters

We briefly demonstrate how *TopicFields* could help a user to seek for geo-tagged social media with temporal filters. Using the stream graph, we can see that the amount of social media has peak values in around 10am and 5pm. To see the differences regarding the lake in the central park, we adjust the temporal filter to *before 5pm* and *after 5pm*. The results are shown in Fig. 10 and Fig. 9:

Before 5pm, the pond looks clear and beautiful:

Fig. 9. This figure shows the geospatio query with the “park” topic before 5pm.

Fig. 10. This figure shows the geospatio query with the “park” topic after 5pm.

Cooler days as summer turns to fall (and wishing I lived in a four seasons kind of place). #summertofall #centralpark #newyork #nyc #happyfriday

#wagnerscove #wcp #westcentralpark #spring #nature #ponds #nyc beautiful day! #faeries live here ? #nycmydna #timeoutnewyorkcity

After 5pm, the pond has a different atmosphere:

#wagnerscove #wcp #westcentralpark #spring #nature #ponds #nyc beautiful day! #faeries live here ? #nycmydna #timeoutnewyorkcity

7 DISCUSSION

There are a few limitations and improvements that could be made to our algorithm. First, the variety and diversity of Twitter posts were very surprising. The Tweets were written in various manners, with many tweets containing few *real* words, or no real words at all. In addition, the amount and degree of sarcasm and double meaning in tweets also made calculating the actual topic of the tweets very difficult. For example, there are a few social media messages that used the word *park*, to refer to the Park Avenue in New York City.

Second, the deep-learning model on images is not always reliable. Sometimes it provides completely wrong information. For example, in a picture of a dog, the neural network recognized it as a boxer.

Third, the spectral ordering algorithm heavily relies on the neural network that learns the similarity between pairs of word vectors. If the similarity score is not high enough, the algorithm may not place similar words into a cluster.

8 CONCLUSION AND FUTURE WORK

In this paper, we have presented a novel system to explore, summarize, and visualize geo-tagged social media with hybrid topic models and scalar field. Our system, *TopicFields*, can efficiently estimate the kernel density distribution and visualize the scalar fields of the user-selected topics on a map on the GPU. We have presented the system and its architecture that ingests geo-tagged Instagram and Twitter messages, extracts topics, hierarchically clusters, and facilitates their interactive visualization on a map. The advantages of using *TopicFields* are that it allows large volume of spatial and temporal data to be visualized and understood, then correlated with a series of topics. Our system includes an efficient and interactive GPU-driven visualization algorithm for visualizing multi-variate scalar data with kernel density estimation and non-linear normalization methods.

We envision the key components of our system, the spectral ordering for the matrix diagram and the GPU-driven visualization of multi-variate scalar data, could inspire future research and systems to visualize geo-tagged health records, business intelligence, and news articles, *e.g.*, social media platforms for virtual and augmented reality [9–11].

REFERENCES

[1] A. A. Apodaca, L. Gritz, and R. Barzel. *Advanced RenderMan: Creating CGI for motion pictures*. Morgan Kaufmann, 2000.

[2] D. M. Blei and J. D. Lafferty. Correlated topic models. In *Proceedings of the 18th International Conference on Neural Information Processing Systems*, pp. 147–154. MIT Press, 2005.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[5] J. Chae, D. Thom, H. Bosch, Y. Jang, R. Maciejewski, D. S. Ebert, and T. Ertl. Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*, pp. 143–152. IEEE, 2012.

[6] S. Chen, X. Yuan, Z. Wang, C. Guo, J. Liang, Z. Wang, X. L. Zhang, and J. Zhang. Interactive visual discovering of movement patterns from sparsely sampled geo-tagged social media data. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):270–279, 2016.

[7] R. B. Cleveland, W. S. Cleveland, and I. Terpenning. Stl: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6(1):3, 1990.

[8] J. Cranshaw, R. Schwartz, J. I. Hong, and N. M. Sadeh. The livehoods project: Utilizing social media to understand the dynamics of a city. In J. G. Breslin, N. B. Ellison, J. G. Shanahan, and Z. Tufekci, eds., *ICWSM*. The AAAI Press, 2012.

[9] R. Du, D. Li, and A. Varshney. Geollery: A Mixed Reality Social Media Platform. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI, p. 13. ACM, May 2019. doi: 10.1145/3290605.3300915

[10] R. Du, D. Li, and A. Varshney. Project Geollery.com: Reconstructing a Live Mirrored World With Geotagged Social Media. In *Proceedings of the 24th International Conference on Web3D Technology*, Web3D, pp. 1–9. ACM, Jul. 2019. doi: 10.1145/3329714.3338126

[11] R. Du and A. Varshney. Social Street View: Blending Immersive Street Views With Geo-Tagged Social Media. In *Proceedings of the 21st International Conference on Web3D Technology*, Web3D, pp. 77–85. ACM, Jul. 2016. doi: 10.1145/2945292.2945294

[12] M. C. Hao, C. Rohrdantz, H. Janetzko, D. A. Keim, U. Dayal, L. E. Haug, M. Hsu, and F. Stoffel. Visual sentiment analysis of customer feedback streams using geo-temporal term associations. *Information Visualization*, 12(3-4):273–290, 2013.

[13] C. G. Healey. Choosing effective colours for data visualization. In *Proceedings of the 7th Conference on Visualization'96*, pp. 263–ff. IEEE Computer Society Press, 1996.

[14] M. Hoffman, F. R. Bach, and D. M. Blei. Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pp. 856–864, 2010.

[15] M.-H. Hwang, S. Wang, G. Cao, A. Padmanabhan, and Z. Zhang. Spatiotemporal transformation of social media geostreams: a case study of twitter for flu risk analysis. In *Proceedings of the 4th ACM SIGSPATIAL International Workshop on GeoStreaming*, pp. 12–21. ACM, 2013.

[16] S. Kim, S. Jeong, I. Woo, Y. Jang, R. Maciejewski, and D. S. Ebert. Data flow analysis and visualization for spatiotemporal statistical data without trajectory information. *IEEE transactions on visualization and computer graphics*, 24(3):1287–1300, 2018.

[17] T. Kurashima, T. Iwata, G. Irie, and K. Fujimura. Travel route recommendation using geotags in photo sharing sites. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pp. 579–588. ACM, 2010.

[18] B. Liu, E. Blasch, Y. Chen, D. Shen, and G. Chen. Scalable sentiment classification for big data analysis using naive bayes classifier. In *2013 IEEE International Conference on Big Data*, pp. 99–104. IEEE, 2013.

[19] Y. Lu, X. Hu, F. Wang, S. Kumar, H. Liu, and R. Maciejewski. Visualizing social media sentiment in disaster scenarios. In *Proceedings of the 24th International Conference on World Wide Web*, pp. 1211–1215. ACM, 2015.

[20] J. Lukasczyk, R. Maciejewski, C. Garth, and H. Hagen. Understanding hotspots: A topological visual analytics approach. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, p. 36. ACM, 2015.

[21] A. M. MacEachren, F. P. Boscoe, D. Haug, and L. W. Pickle. Geographic visualization: Designing manipulable maps for exploring temporally varying georeferenced statistics. In *Information Visualization, 1998. Proceedings. IEEE Symposium on*, pp. 87–94. IEEE, 1998. doi: 10.1109/INFVIS.1998.729563

[22] A. M. MacEachren, A. Jaiswal, A. C. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, and J. Blanford. Senseplace2: Geotwitter analytics support for situational awareness. In *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 181–190. IEEE, 2011. doi: 10.1109/VAST.2011.6102456

[23] R. Maciejewski, S. Rudolph, R. Hafen, A. Abusalah, M. Yakout, M. Ouzzani, W. S. Cleveland, S. J. Grannis, and D. S. Ebert. A visual analytics approach to understanding spatiotemporal hotspots. *IEEE Transactions on Visualization and Computer Graphics*, 16(2):205–220, 2010.

[24] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[25] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013.

[26] G. A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*,

63(2):81, 1956.
[27] R. Paige and R. E. Tarjan. Three partition refinement algorithms. *SIAM Journal on Computing*, 16(6):973–989, 1987.
[28] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing-Volume 10*, pp. 79–86. Association for Computational Linguistics, 2002.
[29] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pp. 248–256. Association for Computational Linguistics, 2009.
[30] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. Twitterstand: news in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 42–51. ACM, 2009.
[31] A. Scharl, A. Hubmann-Haidvogel, A. Weichselbraun, G. Wohlgenannt, H.-P. Lang, and M. Sabou. Extraction and interactive exploration of knowledge from aggregated news and social media content. In *Proceedings of the 4th ACM SIGCHI symposium on Engineering interactive computing systems*, pp. 163–168. ACM, 2012.
[32] Q. Shen, W. Zeng, Y. Ye, S. M. Arisona, S. Schubiger, R. Burkhard, and H. Qu. Streetvizor: Visual exploration of human-scale urban forms based on street views. *IEEE transactions on visualization and computer graphics*, 24(1):1004–1013, 2018.
[33] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *The Craft of Information Visualization*, pp. 364–371. Elsevier, 2003.
[34] B. W. Silverman. *Density estimation for statistics and data analysis*, vol. 26. CRC press, 1986.
[35] A. Stefanidis, A. Crooks, and J. Radzikowski. Harvesting ambient geospatial information from social media feeds. *GeoJournal*, 78(2):319–338, 2013. doi: 10.1007/s10708-011-9438-2
[36] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, vol. 1, pp. 1555–1565. ACL, 2014.
[37] B. E. Teitler, M. D. Lieberman, D. Panozzo, J. Sankaranarayanan, H. Samet, and J. Sperling. Newsstand: A new view on news. In *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, p. 18. ACM, 2008.
[38] R. Tuchinda, P. Szekely, C. a. Knoblock, and M. Rey. Building Mashups By Example. In *The 13th International Conference on Intelligent User Interfaces*, pp. 139–148. ACM, 2008. doi: 10.1145/1378773.1378792
[39] A. Vaccari, M. Martino, F. Rojas, and C. Ratti. Pulse of the city: Visualizing urban dynamics of special events. 2010.
[40] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 1079–1088. ACM, 2010.
[41] X. Wang and E. Grimson. Spatial latent dirichlet allocation. In *Advances in neural information processing systems*, pp. 1577–1584, 2008.
[42] Y. Wang, E. Agichtein, and M. Benzi. Tm-lda: efficient online modeling of latent topic transitions in social media. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 123–131. ACM, 2012.
[43] D. L. Williamson and P. J. Rasch. Two-dimensional semi-lagrangian transport with shape-preserving interpolation. *Monthly Weather Review*, 117(1):102–129, 1989.
[44] C. Xia, R. Schwartz, K. Xie, A. Krebs, A. Langdon, J. Ting, and M. Naaman. Citybeat: real-time social media visualization of hyper-local city data. In *Proceedings of the 23rd International Conference on World Wide Web*, pp. 167–170. ACM, 2014.
[45] J. Yin, A. Lampert, M. Cameron, B. Robinson, and R. Power. Using social media to enhance emergency situation awareness. *IEEE Intelligent Systems*, 27(6):52–59, 2012.