Journal of Intelligent & Fuzzy Systems 36 (2019) 4901–4908
DOI:10.3233/JIFS-179037
IOS Press

Geographical aggregation of microblog posts for LDA topic modeling

Pablo López-Ramírez, Alejandro Molina-Villegas* and Oscar S. Siordia
CONACYT – *Centro De Investigación En Ciencias De Información Geoespacial, Mexico*

Abstract. In this paper we propose an aggregation strategy for geolocated Twitter posts based on a hierarchical definition of the regular activity patterns within a specific region. The aggregation yields a series of documents that are used to train a topic model. The resulting model is tested against the ones produced by two other aggregation strategies proposed in the literature: aggregation by user and by *hashtag*. For comparison, we use quality metrics widely used on the literature. The results show that the Geographical Aggregation performs similarly to *hashtag* aggregation in terms of Jensen-Shannon Divergence and outperforms other aggregation schemes in its ability to reproduce the original cluster labels. One potential application behind this is the discovery of unusual events or as a basis for geolocating messages from text.

Keywords: Probabilistic topic modeling, geolocation, social network

1. Introduction

The increasing amount of information gathered every day from social networks has prove to be a valuable resource for research. More specifically, Spatio-Temporal analysis of Social Media coupled with Natural Language Processing (NLP) can be a valuable avenue of research for exploiting social media information. Combining the approaches of Geospatial Analysis and NLP allows us to examine in a broader sense different aspects of crowd behaviorArab Spring, for example the role of social media in the Arab Spring [1], or the way people react to hazardous events like terrorist attacks [2] or earthquakes [3].

In particular, Probabilistic Topic Modeling has been widely explored to extract insights from the public conversations happening within Social Media. There are several examples in the literature exploring the possibilities offered by techniques such as Latent Dirichlet Allocation (LDA, [4]) to analyze the Twitter stream. However, one of the principal shortcomings of this kind of analysis is the amount of information contained in individual messages. Twitter posts are short and thus contain very little information for the algorithms to learn significant patterns. The general strategy developed to address this issue is to aggregate individual messages into larger *documents*. In this paper we introduce a novel way of aggregating geolocated messages that takes advantage of the spatio-temporal characteristics of the geolocated Twitter stream.

It has been shown in the literature that aggregating individual messages into larger *documents* lead to better quality topics [5] and better performance in terms of training time. So far, several aggregation schemes have been proposed, such as aggregating posts by user [5, 6], *hashtag* [5–7], terms [5] or by *bursts* [7].

However, it is important to notice that the aggregation scheme used has an impact not only on the quality of the topics discovered but also on the actual topics, that is, the probability distribution along the terms that represent each discovered topic. This means that

[*] Corresponding author. Molina-Villegas Alejandro, CONACYT – Centro de Investigación en Ciencias de Información Geoespacial, Mexico. E-mail: amolina@centrogeo.edu.mx.

ISSN 1064-1246/19/$35.00 © 2019 – IOS Press and the authors. All rights reserved

different aggregation schemes will lead to a different set of topics discovered on the same dataset.

This latter observation is important because it means that there is a coupling between the aggregation scheme used to train an LDA model and the possible uses for the resulting model. For example, in [6] a model is trained on documents aggregated by user and *hashtag*, this allows the authors to identify potentially influential users within topical categories. In [5] several aggregations schemes are tested against the tasks of predicting influential messages and classifying users into topical categories. In [7] the authors propose a novel aggregation scheme used to discover emerging topics in the Twitter stream.

Recently, a novel avenue of research is Geographic Topic Modeling and more generally, the relationship between geographic places and message content. In this line, some extensions to the LDA model has been proposed to include the geographic dimension [8, 9] while other authors focus on relating message content with sociodemographic authoritative information [10, 11].

In this paper we aim to bridge research on aggregation schemes for LDA modeling with research on geographic topic discovery. To do so, we propose an aggregation scheme based on a characterization of Twitter activity as a hierarchy of geographical clusters. This characterization aims to capture the regular spatio-temporal activity patterns present in the geolocated Twitter stream. The aggregation proposed is tested against those reported in the literature, the results show that our proposed aggregation performs well in terms of topic quality and separation metrics.

The rest of the paper is organized as follows: in Section 2 we present a brief review of the relevant literature, in Section 3 we describe the geographic aggregation proposed, Section 4 describes the general experimental setting, including the dataset and the metrics used to compare aggregations, in Section 5 we discuss the results and finally, Section 6 presents our conclusions.

2. Previous work

There is strong evidence of a link between the geographic location of social media activity and local characteristics such as land use [12, 13] and local events [14–16]. An important insight that can be drawn from the existing literature is that the place and time where a message is emitted can act as a proxy for the activities the users are undertaking [13, 17, 18]. This is of particular importance for the present research since the aggregation scheme we are proposing aims to capture the differences in the themes treated by social media users at different times, places and scales.

Regarding the application of topic models to short texts, there is a vast body of work proposing and evaluating aggregation schemes to increase the information available to LDA models. Specific to the domain of LDA modeling applied to Twitter posts, in [6] the authors advance the *TwitterRank* algorithm to identify potentially influential users within specific themes. To develop their algorithm the authors aggregate individual messages into larger documents grouping together all posts by the same user. Their results show that such an aggregation scheme is useful in identifying influential users. In [5] the authors undertake an empirical evaluation of different aggregation schemes (aggregation on users and *hashtags* vs. no aggregation) tested against two different real world tasks: predicting popular posts and classifying users into topical categories. The authors find that training LDA models on either posts aggregated by user or by term outperforms training the model on individual posts. More recently [19] tested *hashtag* aggregation using a topic coherence metric that measures the ability of the trained LDA model to reproduce the cluster distribution from which it is drawn (in this case, each *hashtag* acts as a cluster of individual messages). In [7] the authors use author-wise, burst-score (burst is proposed as a score to reflect the saliency of emerging topics), temporal and *hasthag* pooling schemes to aggregate individual messages, they found that the scheme that produces the best topic models (based on topic coherence metrics) is pooling by *hashtag*, this result allows the authors to propose an automatic *hashtag* labeling algorithm that tries to produce the same results that *hashtag* aggregation but without depending on the use of the community defined *hashtags*.

Besides the actual aggregation schemes used and tested on the literature, another important aspect from this review is the different metrics used to test the aggregation schemes. It is possible to identify two broad groups: those who evaluate the trained models intrinsically, that is, based only the characteristics of the discovered topics and those who evaluate the resulting models comparing against additional information such as the cluster from which the messages were drawn or the specific task for which the topics were devised. Examples of the former include the Jensen-Shannon Divergence [5], Point Wise Mutual

Information [20] and coherence metrics [7, 20], while the later includes Normalized Mutual Information [5, 7] and cluster purity [5]. The choice of evaluation metrics largely depends on the task behind the comparison, when developing LDA models fitted to specific tasks, where some sort of ground truth is available (such as the work in [5] or [6]) it makes sense to use metrics that are able to take advantage of external information, but when testing different aggregations in terms of the characteristics of the topic distributions produced, such as separability or homogeneity, the intrinsic metrics are better suited.

The last research topic relevant to our work is the influence of geographic location on the topics discovered by LDA models. In this regard it is possible to identify three broad groups of research: investigations trying to find differences on the topics discovered among geographic regions; research on the relationship between authoritative social and demographical data sources with characteristics mined from social media; and works that propose extensions to the basic LDA framework to accommodate the geographic location of the message. In [11] we can see an example of the first group, where Tweets related to obesity were gathered and geolocated and then a LDA model was fitted to examine the geographical variation of obesity related topics. An example of the second group can be found in [10], where sentiment analysis is carried for a massive dataset of geolocated Tweets, and then the *happiness* of Twitter users is correlated with authoritative data sources describing public wellness. Finally, for the third group, in [21] a model for geographic lexical variation is developed using geolocation as a latent variable in a generative model, or in [8] where an extension of the LDA framework uses the geographic location of Twitter posts as a latent variable, allowing the authors to model the relationships between place and the words used in messages.

In terms of this brief literature review, the work proposed in this paper is a bridge between the research on aggregation schemes and research exploiting the geographic location as a source of information for social media analysis.

3. Geographical aggregation

The aggregation scheme we propose is based on a characterization of the regular activity patterns of the geolocated Twitter stream[1] as a hierarchy of clusters [22]. The main idea of the proposal is that geospatial activity patterns often exhibit a range of scales and that this scales cannot be represented by a flat tessellation (as in [12] or [23]). To overcome this limitation we use a recursive clustering algorithm that is able to extract the structures present at different geographical scales. It is important to notice that the recursive clustering algorithm uses only the spatio-temporal attributes of the messages (that is, the GPS coordinates and the timestamp of the messages), so we are not using the text content to perform geo-referencing and the location used to perform the aggregation is that where the message is emitted.

The rationale behind proposing an aggregation scheme based on geospatial activity patterns is that the documents feeding a LDA model must have some intrinsic semantic significance. In the aggregation schemes examined in the literature, the semantic coherence relates to the topical categories implicit either on users interests or in the way *hashtags* are used to group together conversations about the same theme. In the aggregation proposed in this paper, the semantic coherence is provided by the activities the users are undertaking at the time and place they emit the message. This does not mean we assume that every message will be related to those activities, but that there is a tendency for the users to talk about the activities they are currently undertaking (as discussed in Section 2).

There are two dimensions governing our aggregation scheme: the temporal and the spatial. For the temporal dimension, time is divided into discrete intervals, each interval corresponds to a time window capturing the activities people undertake. Each day is segmented as follows: *Morning*: From 06:00 to 10:00, *Noon*: From 10:01 to 14:00, *Afternoon*: From 14:01 to 18:00, *Evening*: From 18:01 to 22:00 and *Night*: From 22:01 to 06:00. For the spatial dimension, for each time interval, density clusters are calculated recursively using the DBSCAN algorithm [24]. Each recursive application of the clustering algorithm represents a scale level. Each identified cluster is converted to a polygon using the optimal *alpha-shape* [25]. At the end of the process we end up with a collection of polygons that represent the regular activity patterns in the database. The complete work flow for

[1] This is the subset of the Twitter stream that has geographical coordinates attached to its metadata. This coordinates are taken directly from the GPS on the mobile device.

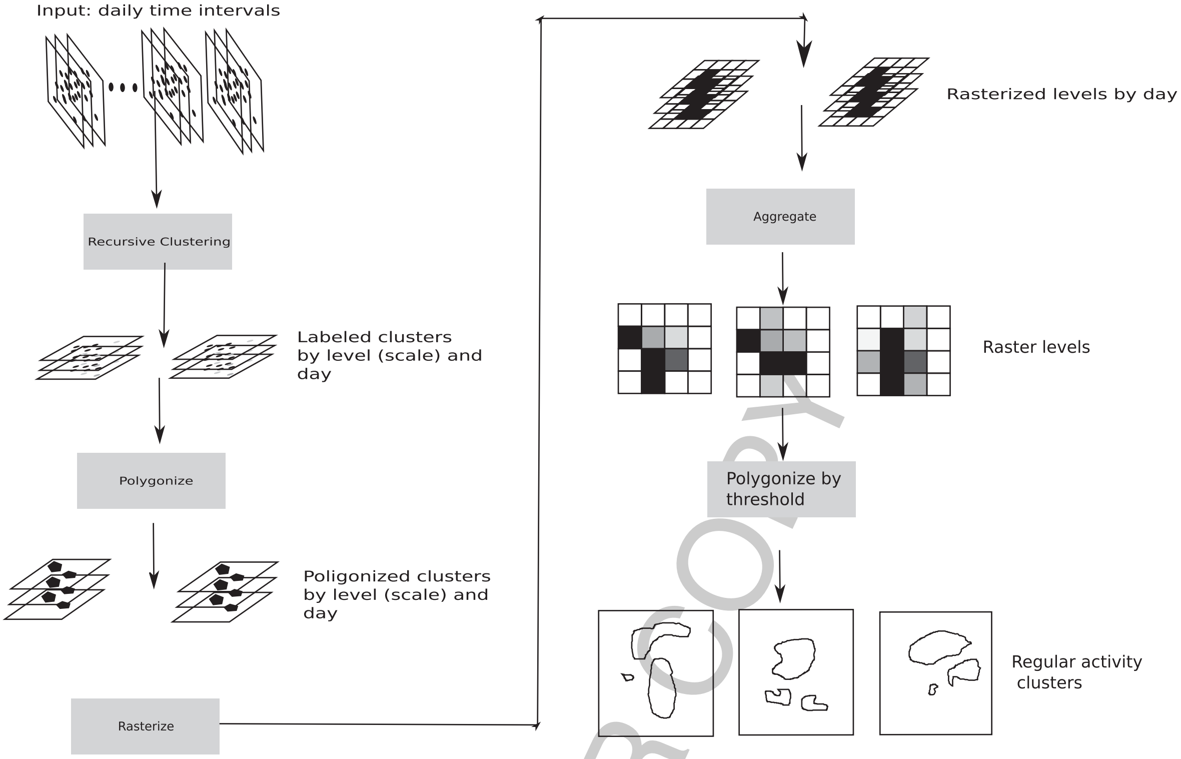

Input: daily time intervals
Recursive Clustering
Labeled clusters by level (scale) and day
Polygonize
Poligonized clusters by level (scale) and day
Rasterize
Rasterized levels by day
Aggregate
Raster levels
Polygonize by threshold
Regular activity clusters

Fig. 1. Diagram showing the workflow for extracting regular activity polygons across multiple scales.

the extraction of regular activity patterns can be seen on Fig. 1.

In Figure 2 the whole hierarchy for the time period corresponding to the afternoon (from 18:00 to 22:00 hours) for the data used in this paper is shown. Each of the dashed lines represents a polygon of regular activity and a document for the LDA model, that is, every tweet inside each polygon is collected and aggregated in a single document.

4. Experiment

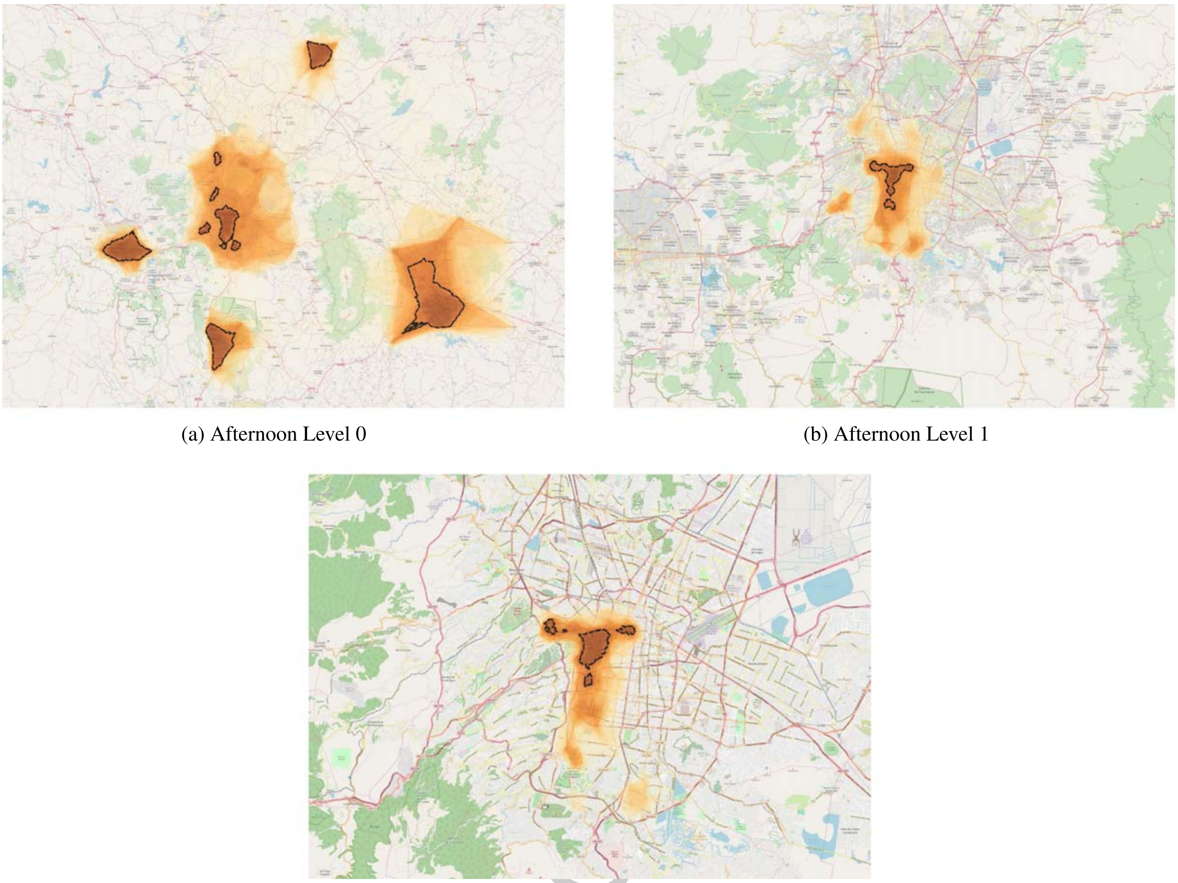
To test out the aggregation proposed we are going to use approximately 2 millions tweets from Central Mexico. We will compare the geographic aggregation described in Section 3 with two aggregations commonly found on the literature: by *user* and by *hashtag*; we will also include unaggregated tweets as a baseline for comparison when pertinent.

4.1. Data

The dataset consists of 2,147,359 geolocated Twitter posts (24,094,521 words) collected from October 2015 to February 2016 in the central region of Mexico using Twitter streaming API and collecting only tweets with explicit geographic coordinates. There are 173,540 unique users and 200,737 distinct *hashtags* in the dataset. The geographically aggregated documents were divided into groups for weekdays and weekends and by time interval (morning, noon, afternoon, evening and night), for each interval and day type three hierarchical average activity scale levels were calculated leading to 24 geographical documents. Table 1 summarizes the number of resulting documents obtained after applying each aggregation method.

4.2. Evaluation

To asses the quality of the proposed geographical aggregation of microblog posts for LDA topic modeling, we trained LDA models for each aggregation method on our study dataset and compare the results using two metrics reported on the literature. In general, topics produced by LDA models can be evaluated in two different ways: (a) in terms of their intrinsic properties or (b) in terms of their ability to classify documents according to *ground*



(a) Afternoon Level 0

(b) Afternoon Level 1

(c) Afternoon Level 2

Fig. 2. Polygons for each scale level for the afternoon period.

Table 1
Number of LDA-documents for each aggregation method

Aggregation method	Resulting documents
Tweet (No aggregation)	2,147,359
Hashtag	200,737
User	173,540
Geographical	24

truth labels. For this work we will use one metric representative of each approach: for (a) we will use the Jensen-Shannon Divergence (JSD) that quantifies how *distinguishable* two or more probability distributions are from each other; thus it measures the *separation* between topics. For (b) we will use the Normalized Mutual Information (NMI) [20, 26] that measures the separation between the ground truth and the predicted labels.

The JSD for two probability distributions takes the value 0 if the distributions are identical and approaches to 1 as they differ more. Thus, small JSD values for two topic distributions X and Y, generated through two different aggregation methods, means that the methods have found similar topic models, that is X and Y Topic Models are almost equivalent. The $JSD(X||Y)$ is calculated as follows:

$$JSD(X||Y) = \frac{1}{2}DKL(X||Z) + \frac{1}{2}DKL(Y||Z) \qquad (1)$$
$$Z = \frac{1}{2}(X + Y)$$

Where DKL, the Kullback-Leibler Divergence between two probability distributions, is calculated as:

$$DKL(X||Y) = \sum_{w=1}^{N} p(w \in X) \log \frac{p(w \in X)}{p(w \in Y)} \qquad (2)$$

Where N is the number of n-grams in the Vector Space Model and $p(w \in X)$ is the probability that the n-gram w belongs to topic X after LDA convergence for a fixed number of topics.

To evaluate the NMI, each aggregation scheme is split into test and training sets and a hard label is assigned to each document by choosing the most likely topic for each document. After this, the correspondence between topics and labels for the training set is taken as the ground truth and the left-out sample is tested against this ground truth. In this setting the NMI is defined as follows:

$$NMI(X, Y) = \frac{2I(X, Y)}{H(X) + H(Y))} \tag{3}$$

where $X = \{x_1, x_2, \ldots, x_K\}$ is the set of documents assigned to each topic, each x_i is a set of documents corresponding to the $i-$th topic; $Y = \{y_1, y_2, \ldots, y_K\}$ is the equivalent set for the ground truth labels; $I(X, Y)$ is the mutual information (Equation 4) between the sets X and Y and H is the Shannon entropy of the corresponding distribution. The mutual information between the two sets is calculated as:

$$I(X, Y) = \sum_X \sum_Y p(x_i, y_j) \cdot \log \left(\frac{p(x_i, y_j)}{p(x_i) \cdot p(y_j)} \right) \tag{4}$$

where $p(x_i)$ and $p(y_j)$ are the probabilities of a document being assigned to topic i and label j respectively; $p(x_i, y_j)$ is is the probability of a document being assigned to topic i and labeled as j. Thus defined, $NMI = 0$ when both distributions are totally different, i.e when there is no agreement between the train and the test classifications and $NMI = 1$ when there is complete agreement.

5. Results

Figure 3 shows the JSD, for topic numbers ranging from $k = 2$ to $k = 100$, between models trained with the proposed geographical aggregation and the other schemes tested. It is interesting to notice that the JSD for the comparison between geographical aggregation and hashtag aggregation is much lower than for the other cases. One way to interpret this result is that geographical aggregation is partially equivalent to the use of hashtags in the sense that is capturing focused *conversation* themes, without users explicitly labeling the messages with hashtags.

In contrast, JSD is close to 1.0 for the comparisons between geographic aggregation vs. Tweet and geographic aggregation vs user. This means that the topics obtained with geographic aggregation differ substantially from those obtained with the user scheme or when no aggregation is performed.

Figure 4 shows the NMI for the geographic and user aggregation. In general the performance of the geographic scheme is much better than the user, meaning that the topics discovered using geographic aggregation are better at reproducing the original cluster labels. This improved classification may be due to two factors, on the one hand the geographic aggregation combines the spatial and temporal dimension and its able to capture the activities the users are engaged at the time an place when they emit the message, on the other hand, geographical documents are much bigger in size so there is more information available to the LDA model.

Tables 2–4 show the top topics and terms for the three different aggregation schemes tested, the number of topics for each aggregation was selected using the metric proposed by [27]. It is interesting to notice that the three aggregations produce very different topics, specifically, user and hashtag aggregation produce few words with geographic references while the geographic aggregation topics contain placenames in

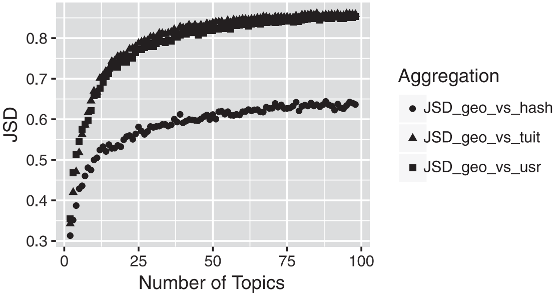


Fig. 3. Jensen-Shannon Divergence for increasing number of topics (k) for different aggregation schemes.

Table 2
Top topics and terms for Geographic aggregation

Topic 1	méxico, mexico, federal, distrito, ciudad, city, hoy, día, gracias
Topic 2	cuauhtémoc, trndnl, cst, tendencia, convertirse, viernes, posición, marthadebayle, vela, torre
Topic 3	votefifthharmony, miguel, tokiohotelconrogergonzalez, hidalgo, buenos, desayuno, días, día, unam, corporativo
Topic 4	buenos, votefifthharmony, días, día, corporativo, universidad, universitario, olímpico, estadio, sábado
Topic 5	cuauhtémoc, benito, miguel, hidalgo, federal, distrito, mexico, juárez, for, city

Table 3
Top topics and terms for hashtag aggregation

Topic 1	casa, noche, buena, tarde, col, viendo, amigo, micanal5, bonita, llego
Topic 2	foro, sol, vive, música, latino, show, concierto, escuchando, disco, vivo
Topic 3	méxico, tendencia, convertirse, ocupando, posición, cst, mención, acaba, rts, cuentas
Topic 4	vamos, estadio, azteca, hoy, cf_america, equipo, partido, final, chivas, américa
Topic 5	justin, bieber, little, lanadelrey, littlemix, mix, follow, justinbieber, one, vote

Table 4
Top topics and terms for user aggregation

Topic 1	smartfit_mex, fitness, fit, smart, gym, sport, club, energy, sportcity_mx, darle
Topic 2	cuernavaca, morelos, studio, mor, dance, tattoo, alicia, laberinto, sábados, domingos
Topic 3	somoscd9, quiero, favor, josdice, amo, tvtelehit, slimecd9kca, cd9, mejor, mas
Topic 4	hotel, ángel, independencia, crossfit, pista, correr, bosque, maria, carrera, nike
Topic 5	gracias, poner, malumafamilydf, malumacolombia, buenas, favor, tardes, pic, podrian, carnaval

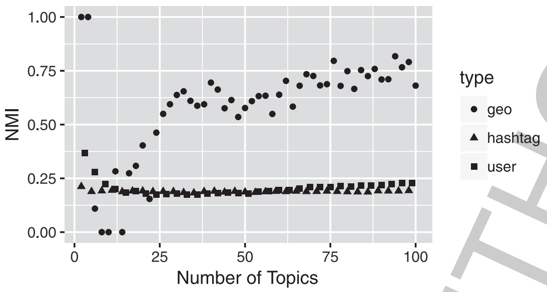


Fig. 4. Normalized Mutual Information for user and geographical aggregations for different topic numbers.

every topic. This tells us that, although the hashtag and geographic aggregation schemes produce similar probability distributions (as shown in Fig. 3), and are thus similar, the focus of geographic aggregation is on the geographic references and is thus capturing a different aspect of the public conversation than hashtag scheme.

6. Conclusions and further work

The Geographical Aggregation presented in this paper represents an interesting alternative to the aggregation schemes reported in the literature. It is interesting that in terms of the *JSD* it performs similarly to the hashtag aggregation which is known to be a useful aggregation scheme [5, 7, 19], while the fact that it performs well at reproducing the cluster labels opens the door for further research in geolocating messages using the text content.

One limitation of the proposed method is that it relies on the geolocated Twitter stream, which comprises only a small fraction of all the messages. Further more, using this method on different data requires that the text content is attached to a specific set of coordinates, thus limiting the possible data sources. On the other hand, for data that has text attached to locations in space, such as Flicker images or 911 reports, the aggregation proposed is an excellent alternative for capturing the spatio-temporal variation of the messages analyzed.

Moving forward on this research, the next step would be to test the geographic aggregation scheme on a larger dataset comprising a whole country and try to capture the regional differences in language use, this would serve as a basis for georreferencing messages by content. Another interesting line of research is testing the aggregation proposed against different geographic aggregation strategies, such as K-means or the use of administrative boundaries.

References

[1] A. Khan, The role of social media and modern technology in arab spring, *Far East Journal of Psychology and Business* **7**(1) (2012), 56–63. Paper 4.

[2] O. Oh, M. Agrawal and H.R. Rao, Information control and terrorism: Tracking the Mumbai terrorist attack through twitter, *Information Systems Frontiers* **13** (2010), 33–43.

[3] A. Crooks, A. Croitoru, A. Stefanidis and J. Radzikowski, #Earthquake: Twitter as a distributed sensor system, *Transactions in GIS* **17** (2013), 124–147.

[4] D.M. Blei, A.Y. Ng and M.I. Jordan, Latent dirichlet allocation, *Journal of Machine Learning Research* **3** (2003), 993–1022.

[5] L. Hong and B.D. Davison, Empirical study of topic modeling in Twitter, *Proceedings of the First Workshop on Social Media Analytics - SOMA '10*, 2010, pp. 80–88.

[6] J. Weng, E.-P. Lim, J. Jiang and Q. He, TwitterRank, *Proceedings of the Third ACM International Conference on Web Search and Data Mining - WSDM '10*, 2010, p. 261.

[7] R. Mehrotra, S. Sanner, W. Buntine and L. Xie, Improving LDA topic models for microblogs via tweet pooling and automatic labeling, *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '13*, 2013, p. 889.

[8] C. Wang, J. Wang, X. Xie and W.-y. Ma, Mining Geographic Knowledge Using Location Aware Topic Model, *Proceedings of the 4th ACM workshop on Geographical information retrieval (GIR)*, 2007, pp. 65–70.

[9] B. Hu and M. Ester, Spatial topic modeling in online social media for location recommendation, *Proceedings of the 7th ACM conference on Recommender systems - RecSys '13*, 2013, pp. 25–32.

[10] L. Mitchell, M.R. Frank, K.D. Harris, P.S. Dodds and C.M. Danforth, The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place, *PLoS ONE* **8** (2013).

[11] D. Ghosh and R. Guha, What are we 'tweeting' about obesity? Mapping tweets with topic modeling and Geographic Information System, *Cartography and Geographic Information Science* **40** (2013), 90–102.

[12] V. Frias-Martinez and E. Frias-Martinez, Spectral clustering for sensing urban land use using Twitter activity, *Engineering Applications of Artificial Intelligence* **35** (2014), 237–245.

[13] R. Lee, S. Wakamiya and K. Sumiya, Urban area characterization based on crowd behavioral lifelogs over Twitter, *Personal and Ubiquitous Computing* **17** (2013), 605–620.

[14] R. Lee, S. Wakamiya and K. Sumiya, Discovery of unusual regional social activities using geo-tagged microblogs, *World Wide Web* **14** (2011), 321–349.

[15] T. Kim, G. Huerta-Canepa, J. Park, S.J. Hyun and D. Lee, What's happening: Finding spontaneous user clusters nearby using twitter, *Proceedings - 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing, PASSAT/SocialCom 2011*, 2011, pp. 806–809.

[16] T. Cheng and T. Wicks, Event detection using twitter: A spatio-temporal approach, *PLOS One* **9** (2014), e97807.

[17] F. Atefeh and W. Khreich, A survey of techniques for event detection in twitter, *Computational Intelligence* **31** (2015), 132–164.

[18] A. Boettcher and D. Lee, EventRadar: A Real-Time Local Event Detection Scheme Using Twitter Stream, *IEEE*, 2012, pp. 358–367.

[19] A.O. Steinskog, J.F. Therkelsen and B. Gambäck, Twitter Topic Modeling by Tweet Aggregation, *Proceedings of the 21st Nordic Conference of Computational Linguistics*, 2017, pp. 77–86.

[20] D. Newman, J. Lau, K. Grieser and T. Baldwin, Automatic evaluation of topic coherence, in: *Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010)*, Los Angeles, pp. 100–108.

[21] J. Eisenstein, B. O'connor, N.A. Smith and E.P. Xing, A Latent Variable Model for Geographic Lexical Variation, 2010, pp. 1277–1287.

[22] P. Lopez-Ramirez, A. Molina-Villegas, O. Sanchez-Siordia, M. Chirinos-Colunga and G. Hernandez-Chan, Multi-scale extraction of regular activity patterns in spatio-temporal events databases: A study using geolocated tweets from central mexico, *Research in Computing Science* **148** (2018).

[23] R. Lee and K. Sumiya, Measuring Geographical Regularities of Crowd Behaviors for Twitter-based Geo-social Event Detection, 2010, pp. 1–10.

[24] M. Ester, H.-P. Kriegel, J. S and X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, AAAI Press, 1996, pp. 226–231.

[25] E. Edelsbrunner and M. Herbert, Three-dimensional alpha shapes, *ACM Trans Graph* (1994), 43–72.

[26] D. Mimno, H.M. Wallach, E. Talley, M. Leenders and A. McCallum, Optimizing semantic coherence in topic models, in: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 2, Association for Computational Linguistics*, 2011, pp. 262–272.

[27] R. Arun, V. Suresh, C.E. Veni Madhavan and M.N. Narasimha Murthy, On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations, Springer, Berlin, Heidelberg, 2010, pp. 391–402.