Place as topics: analysis of spatial and temporal evolution of topics from social networks data

Giovanni Siragusa
Department of Computer Science - University of Turin
Via Pessinetto, 12, Italy
giovanni.siragusa@edu.unito.it

Abstract

Geography in a commonsense way is about *place*. *Place* is a term used to describe the meaning that humans give to a location. Characterising a location as a *place* requires a huge amount of time to collect and analyse data. Furthermore, a *place* definition associated to a location can become rapidly obsolete. Nowadays, social networks and social media became very popular. People on social networks act like social sensors, reporting information about society, politics, economics, etc. Thus, many researchers have focused on the analysis of *posts*, combining them together with algorithms or extracting their meaning, keywords or users' interests. In this paper, I will describe my research project, a visual framework that aims to simplify the process of *place* definition using topics generated from the application of Blei et al.'s *Latent Dirichlet Allocation* (Blei et al., 2003) on geo-referenced social networks data. My main assumption is that topics allow to capture the *sense of place* shared by social sensors. The framework will allow users to be not overwhelmed by the large amount of time and data required to understand and define *places*.

Keywords: NLU, LDA, topics, places, social networks

1. Introduction

In Geography, *place* is an important concept: it is used to describe the meaning that humans give to a location when it is used and lived. Cresswell, in his articles (Cresswell, 2009; Cresswell, 2011), describes a *place* as a melting-pot of 4 elements: *location*, *locale*, *sense of place* and *practice*. *Location* is a physical point in space with a specific set of coordinates, e.g., latitude and longitude. It refers to the "where" of *place*. A *location* can be a city, a city district, a street, a build or even a ship. *Locale* refers to the way a *place* "looks": the material setting for social relation, such as streets, shops, buildings and so forth. *Sense of place* is a nebulous meaning. It includes feelings, emotions and meanings that a *place* evokes to people. *Sense of place* can be individual and based on biography (e.g., the place where I spent my childhood) or it can be shared. *Practice* represents what people do in place. It can contain historical practice (e.g., a battlefield), mundane practice (e.g., going to work) or a mixture of both. *Sense of place* heavily influences *practice*, and *practice* is leaded by the *sense of place*.

Geographers that are intended to define *places* need to perform a sequence of steps. First, they must define the location to study, then they must collect a set of observations regarding the *place*: feelings, emotions, meanings, practice and so forth. Finally, they must analyse all the data to define the *sense of place*. Unfortunately, this process requires a huge amount of time and the definition produced can be obsolete due to the dynamic nature of the *place* itself.

In the last decades, social network services became very popular not only for people, but also for scientific communities and practitioners. People on social networks act like social sensors, reporting information about society, politics, economics, etc. Thus, many researchers have focused on the analysis of posts, combining them together with algorithms or extracting their meaning, keywords or users' interests. The work proposed in (Rizzo et al., 2016) uses posts to define cities thematic maps through a subspatial cluster algorithm called *GeoSubClu*. In (Sakaki et al., 2010), the authors consider Twitter as social sensor for detecting large events such as earthquakes or typhoons. Cataldi et al., in their work (Cataldi et al., 2013), proposed an approach to provide to users the most emerging topics expressed by the community. Cataldi et al. see users as real-time news sensors. The work proposed in (Allisio et al., 2013) uses tweets in conjunction with Sentiment Analysis to capture how much citizen of Italian cities are happy. In their work, they also proposed a graphic framework that allows the user to apply Sentiment Analysis and to infer why people are happy (or unhappy) in a city. Furthermore, Allisio et al.'s work can be viewed as an analysis of the feeling content of *places*.

In this paper, I will describe my research project, a visual framework that aims to simplify the process of *place* definition using topics generated from the application of Blei et al.'s *Latent Dirichlet Allocation* (Blei et al., 2003) on geo-referenced social networks data. My main assumption is that topics allow to capture the *sense of place* shared by social sensors. Moreover, word co-occurrences in topics can be used to infer further information regarding *places*. In details, the framework has a threefold impact:

1. it will allow to capture the *sense of place* of a designed location and how it is geographical distributed over the *place*;

2. it will allow to infer why a *place* has a specific *meaning*, how it is shared and how it evolves over time;

3. it will allow practitioners (e.g., sociologists or psychologists) to apply the LDA model and to generate plots with a single click.

2. Related Works

As previously mentioned in Section 1., social network services such as *Facebook* (`www.facebook.com`) and *Twitter* (`www.twitter.com`) became very popular. Users use social networks to share their thoughts, pictures or videos. On social networks, a user indicates that he/she wants to get notified ("follow") or becomes a friend of another user.

Nowadays, new platform services have emerged. These services have gone beyond information, enabling people to have a direct link with their neighbours and discover local businesses or associations. Examples of such platforms are *MyNeighbourhood* (`www.my-n.eu`) and *Polly&Bob* (`www.pollyandbob.com`). Furthermore, platforms started to use map-based services to push the attention at problems that have to change in cities. *FixMyStreet* (`www.fixmystreet.com`) allows people to report, discuss or view local problems. *Ushahidi* (`www.ushahidi.com`), instead, allows users to report or get notified about what's happening, where and when.

In last years there exists an increasing trend to georeferencing information. Facebook added the possibility to geotag posts, while platforms that analyze users' location, geotag and hashtag arise. For example, *Trendsmap* (`www.trendsmap.com`) aims to show latest trends from Twitter on a map. Differently from these platforms and common social networks, *First Life* (`www.firstlife.org`) (Antonini et al., 2015) is a social network oriented to the person as the citizen, where information are not rooted on the personal life of users, but on their collective way of living a place. First Life combines different sources of information (posts, blogs, open data, etc), that are geo-referenced, and POIs (Point of Interests), and shows them on an interactive map. Furthermore, data can be associated with a temporal dimension which is used to filter or to order the data.

In the context of social networks, Blei et al.'s *Latent Dirichlet Allocation* (LDA) (Blei et al., 2003) was successfully applied. LDA is a generative model that treats a document as a finite mixture of topics, where a topic is a distribution over words. In details, each topic captures word co-occurrences inside documents. In the work proposed in (Pennacchiotti and Gurumurthy, 2011), authors used LDA to automatically discover users' interests. Users can be represented as a mixture of topics, the parameter θ, and these mixtures can be used to suggest friends or people to follow through the computation of dissimilarity functions (e.g., Kullback-Leibler divergence) or cosine similarity. In (Cha and Cho, 2012), Cha and Cho used LDA to analyze the relationship graph of popular social networks. The author's goal was to cluster a set of nodes using topics and to label each edge with a topic group number, obtaining a model that has a twofold impact: it can be used to suggest users and infer why a new user chose to initially follow certain users. Zhang et al., in their work (Zhang et al., 2007), proposed a model called *SSN-LDA* (Simple Social Network LDA) to discover communities from social networks. In their model, communities are represented by latent variables. Eisenstein et al. suppose, in their work (Eisenstein et al., 2010), that pure topics' word co-occurrences are corrupted by geographical information. The model assigns words to a topic according to a geographic region, which is modelled by a latent variable labelled with r.

Recent works have focused on tracking the evolution of topics over time. The framework proposed in (Cui et al., 2011) allows to capture both topics distribution over time and critical events, such as birth, split, merge or death of topics. Furthermore, the model captures and represents word co-occurrences and co-occurrences frequency using threads. First the model defines main words computing a set of weights, then it represents co-occurrences through the wave bundle of the thread. The amplitude of the wave represents the number of co-occurrences between the main word and the other words inside a topic: high amplitudes represent elevate co-occurrences frequency. In (Wang and McCallum, 2006) Wang and McCallum proposed a modified LDA model, called *Topics Over Time*, where topic discovery is influenced both by word co-occurrences and temporal information. In their work, the authors model the time as a continue distribution, defined by a Beta distribution over a parameter Ψ, associated with each topic which is responsible to generate both patterns and topics distribution. Lau et al. in (Lau et al., 2012) proposed a novel method to track emerging events in microblogs (e.g., Twitter). Their method defines a window of time slices, where each time slice contains several documents, and updates parameters α and β for each old word and document. Novel words and documents are initialised using two parameters, α_0 and β_0, that are defined a priori.

Another related work, not linked to the LDA model, is (Di Caro et al., 2011). Di Caro et al. proposed a framework called *TMine* which defines a navigable *tag-flag*. A *tag-flag* can be thought as a topic because it contains a set of related words.

3. Research Questions and Objectives

In this section I describe my research questions and research objectives that will lead to the construction of the framework. My objective is to apply the LDA model to geo-referenced social network data to capture the *sense of place*[1]. My assumption is that topics represent how people live a *place* (Cresswell, 2009; Cresswell, 2011): activities, emotional attachment to *place* and so forth. For example, parks can have a *sport* topic during the afternoon and a *concert* topic during the night. Thus, I am interested in spatial and temporal location of topics. In detail, I will respond to three research questions (labelled with *RQ*):

RQ1 Where is a topic spatially located over time? In RQ1 I am interested to understand when and where a topic emerges and if it can spread in the neighbouring areas due to *social influence*: the change in behaviour that one person causes in another. To answer RQ1, I will study how a topic evolves both temporally and spatially. In details, I will develop a module that associates topics to a location and tracks the spatial and temporal evolution of topics using dissimilarity functions (e.g., Kullback-Leibler divergence) or cosine similarity. I will study how to track a topic over

[1] *Location* and *Locale* are implicitly defined in the selection of the geographical area to study.

time because it can change its structure from a time slice to another. Furthermore, I will try and compare different LDA models, such as Topic Over Time described in (Wang and McCallum, 2006), to find the best model (or models) to extract the *sense of place*.

RQ2 Which topics are presented in the same space over time? In RQ2 I am interested to discover how people live a *place* and infer how their way to live a *place* can change over time. To answer RQ2 I will use the framework developed in RQ1 to analyze the correlation between a *place* and its topics in order to validate my assumption. Furthermore, the analysis of topics in a place in conjunction with their spatial and temporal location analysis will allow to infer further information regarding *places*.

RQ3 Where are users with same interests geographically located? In my project I aim to represent how people live a *place* using topics, but topics depend also on people interests (both subjective and emotional). Thus, I assume that users with same specific interests would refer to similar *places*. First I will study how to cluster users according to specific topics and how to find group of users that have same specific interests and use a specific language. Then, I will cluster users and I will analyze where the members of a cluster are located. Clusters and their shape can be used to improve topic representation inside *places*: we can use the clusters to associate topics to specific areas of the *place*, finding which topic is dominant, how topics overlap and how they are distributed. Clusters distribution inside the *place* can bring more clues about its *meaning*.

In Section 2. I described two works that use topics to implement a user recommendation system: the work described in (Pennacchiotti and Gurumurthy, 2011) and the work described in (Cha and Cho, 2012). Pennacchiotti and Gurumurthy define a model that capture users' interest through topics and compare topics distribution to suggest users; Cha and Cho, instead, define a model that captures social interaction between users through a latent variable which defines a *community*. *Communities* can be used to suggest users to follow to a new user. In my research project, topics capture users' interests. Thus, I can suggest to a user *places* that have most of the topics in common (sufficient and necessary condition is that the *place* suggested cannot be the *place* where the user dwells).

4. Data

In this section I am going to describe the data I will use to validate my main assumption, that topics capture the *sense of place* expressed by users on social networks. To validate my assumption, I will define two datasets: a dataset of tweets taken form Twitter using *Twitter API*, which allows to specify latitude, longitude and a radius in the query, and a dataset of posts taken from *First Life* (Antonini et al., 2015).

Twitter is a social media where users post their through and get in touch (follow) with other users. It is vastly used by researchers, practitioners (e.g., sociologists or psychologists), data journalists and computational linguists due to the huge amount of real human data that it contains. Thus, Twitter was and it is still used to extract information (see (Allisio et al., 2013)) or to test developed application, such as the applications described in (Pennacchiotti and Gurumurthy, 2011; Cha and Cho, 2012; Eisenstein et al., 2010; Lau et al., 2012). Unfortunately, not all information produced by users are useful. For example, popular users (users followed by a large number of users), such as artists, actor and so forth, can produce noisy information. In my research project the first step will be to divide popular users from unpopular ones and analyse topics produced by each group in order to find which ones express the *sense of place*. Defined the user group (popular on unpopular), the second step will be to analyze tweets associated with topics that produced (in the first step) the *sense of place*. This second step will allow to filter the noise in the data, obtaining high quality topics. For example, tweets that express the *sense of place* can contain high frequency of certain words or can be highly re-tweeted. Thus I will analyze words frequency, number of re-tweets and so forth. These filters and their plots will be integrated in the framework (see Section 5. for details).
Differently from Twitter, First Life is a social network focused on the space where a user lives. For this reason, First Life is the perfect candidate to validate my main assumption. However, First Life has two cons: it is not as popular as Twitter and it contains only Italian users. For this social network I will apply the same above-mentioned analysis for Twitter.

5. Framework Architecture

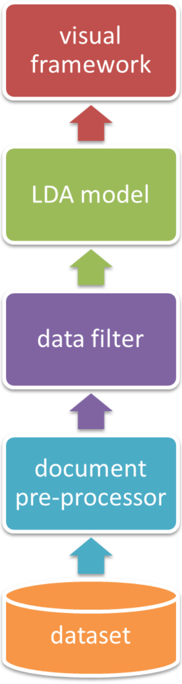

Figure 1: The figure shows the framework architecture.

In this section, I present the framework architecture which is composed by four layers as showed in Figure 1: a blue layer which pre-processes documents in input; a violet layer which filters the data; a green layer that applies the LDA model on cleaned data and a red layer that visualises topics. I will use json for input documents, allowing users to use their datasets. Moreover, the json input format will respect a grammar in order to standardise the input.

The first layer (blue), called *document pre-processor*, deals with the cleaning of documents. First, it will parse the text to extract *Part-Of-Speech* (POS) tags; then it will tokenize documents and it will filter stopwords and all words having POS tags different from *ADJ* (Adjective), *VERB*, *Noun* and *X* (foreign word). The output of this layer is passed in input to the second layer (violet), called *data filter*, which is composed by a set of filters that implements operations described in Section 4. For example, I can filter all tweets that have a number of re-tweets lower than a threshold and, then, filter popular users or viceversa. Users can freely combine filters in sequence and study how topics change according to applied filters. The third layer (green), called *LDA model*, will apply the LDA model on filtered data. This layer only needs the number of topics. To simplify the choice of the number of topics, I will implement a perplexity method (associated with a perplexity plot) that will require a minimum number of topics, a maximum number of topics and a step. Finally, LDA output will be passed in input to the last layer (red) called *visual framework*. This layer will implement all the features described in Section 3. Moreover, the *visual framework* layer will implement a set of plots that will allow users to infer further information.

6. Conclusion

In this paper I presented my research project: a visual framework that allows to extract the *sense of place* from social networks data using topics generated by *Latent Dirichlet Allocation*. The main advantage of my framework does not only regard the extraction of the *sense of place*, but also infer why the *place* has a specific meaning. Furthermore, topics can be used to implement a *geographic recommendation* system, suggesting *places* to users.

7. Bibliographical References

Allisio, L., Mussa, V., Bosco, C., Patti, V., and Ruffo, G. (2013). Felicità: Visualizing and estimating happiness in italian cities from geotagged tweets. In *ESSEM@ AI* IA*, pages 95–106. Citeseer.

Antonini, A., Boella, G., Buccoliero, S., Calafiore, A., Di Caro, L., Giorgino, V., Ruggeri, A., Salaroglio, C., Sanasi, L., and Schifanella, C. (2015). First life, from the global village to local communities. In *1st IASC Thematic Conference on Urban Commons*.

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

Cataldi, M., Caro, L. D., and Schifanella, C. (2013). Personalized emerging topic detection based on a term aging model. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(1):7.

Cha, Y. and Cho, J. (2012). Social-network analysis using topic models. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 565–574. ACM.

Cresswell, T. (2009). Place. In *International Encyclopedia of Human Geography*, volume 8, pages 169–177. Elsevier.

Cresswell, T. (2011). Place–part i. *The Wiley-Blackwell companion to human geography*, pages 235–244.

Cui, W., Liu, S., Tan, L., Shi, C., Song, Y., Gao, Z. J., Qu, H., and Tong, X. (2011). Textflow: Towards better understanding of evolving topics in text. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2412–2421.

Di Caro, L., Candan, K. S., and Sapino, M. L. (2011). Navigating within news collections using tag-flakes. *Journal of Visual Languages & Computing*, 22(2):120–139.

Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. (2010). A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287. Association for Computational Linguistics.

Lau, J. H., Collier, N., and Baldwin, T. (2012). On-line trend analysis with topic models: #twitter trends detection topic model online. In *COLING*, pages 1519–1534.

Pennacchiotti, M. and Gurumurthy, S. (2011). Investigating topic models for social media user recommendation. In *Proceedings of the 20th international conference companion on World wide web*, pages 101–102. ACM.

Rizzo, G., Meo, R., Pensa, R. G., Falcone, G., and Troncy, R. (2016). Shaping city neighborhoods leveraging crowd sensors. *Information Systems*.

Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM.

Wang, X. and McCallum, A. (2006). Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433. ACM.

Zhang, H., Qiu, B., Giles, C. L., Foley, H. C., and Yen, J. (2007). An lda-based community structure discovery approach for large-scale social networks. *ISI*, 200.