Modelling Time and Location in Topic Models

Christian Pölitz CHRISTIAN.POELITZ@TU-DORTMUND.DE
TU Dortmund University, Artificial Intelligence Group,
Otto Hahn Str. 12, 44227 Dortmund, Germany

Abstract

Many text collections like news paper or social media blog archives contain texts that often refer to special dates and/or locations. These information can be valuable to investigate topics in certain regions and time spans. We use topic models that integrate time and geographical information extracted from the texts to find such topics. In this extended abstract, we motivate our approach and shortly describe the method. Experimental evaluations and detailed description are in preparation to a full paper.

1. Introduction

Topic models (see for instance (Blei et al., 2003)) have been used extensively to summarize text collections into semantic clusters. Such text collections can contain for instance news paper articles, Blog entries, tweets or any written social media content. The documents in these collections contain often information about locations and dates. These information are valuable for extracting topics for certain regions or time spans. In order to integrate temporal and positional information, we need the corresponding time and location information for our text corpus. Previous approaches assumed that we either directly have information about time and position for each document in the corpus or that a named entity recognition tool finds geographic locations. We propose a hybrid approach that extends standard Latent Dirichlet Allocation (LDA (Blei et al., 2003)) topic models. We assume that the documents can have multiple dates and positions. In order to integrate multiple information for single documents, we propose to extract parts (or chunks) in the documents that contain only information about one location and one date. For example, a document might contain the sentence: "The weather was nice in Berlin last Sunday, but next day at home in Cologne it was cloudy.". In order to reflect all temporal and positional information, NLP tools would divide the sentence in: "The weather was nice in Berlin last Sunday" and "but next day at home in Cologne it was cloudy" and label the words in the first chunk with the time stamp of that corresponding Sunday and the geographical location of Berlin. The words in the second chunk are labeled with the time stamp of that corresponding Monday and the geographical location of Cologne.

Proceedings of the 2^{nd} International Workshop on Mining Urban Data, Lille, France, 2015.

2. Related Work

There are several previous approaches that integrate temporal and positional information into topic models. In (Yin et al., 2011) Yin et al. discuss methods to find and compare topics in documents that have associated GPS coordinate. Speriosu et al. propose in (Speriosu et al., 2010) to use topic models that use non-overlapping regions as latent topics. By this, they model each document as distribution over these regions. Further approaches use topic models with geographic information on social media data to extract activity patterns of users. Hasan and Ukkusuri for instance use in (Hasan & Ukkusuri, 2014) topic models that integrate sequences of activities rather than documents. In (Hong et al., 2012) Hong et al. introduce a sparse generative topic model and in (Kurashima et al., 2013) Kurashima et al. propose a geographic topic model that use Twitter tweets to extract user activities in terms of movement and interests.

3. Method

While the standard topic models group only words and documents in semantically related topics, we are further interested in the distribution of the topics over time and geographic position. In order to extract the distribution of word senses over time and positions, we use topic models that consider temporal information about the documents as well as locations in form of numerical vectors that represent the geographic position. This means, each document has at least one time stamp and one geographic position. The time stamps are assumed to be Beta distributed and the position Normally distributed. The two distributions are simply in-

tegrated in an LDA topic model under the assumption that given the latent topics, the words, the time stamps and geographic positions are independent. We combine the methods by Wang and McCallum (Wang & McCallum, 2006) introduced as topics over time and supervised LDA introduced by Blei et al. (Blei & McAuliffe, 2007).

The generative process of the words, dates and location is:

1. For each topic t:
   (a) Draw $\theta_t \sim Dir(\beta)$
2. For each document d:
   (a) Draw $\phi_d \sim Dir(\alpha)$
   (b) For each chunk $c(d)$:
      i. For each word i in $c(d)$:
         A. Draw $z_i \sim Mult(\phi_d)$
         B. Draw $w_i \sim Mult(\theta_{t_i})$
         C. Draw $t_c \sim Beta(\psi_{z_i})$
         D. Draw $l_c \sim N(\eta' \hat{z_{c(d)}}, \rho^2)$

Assuming a number of topics we draw for each of them a Multinominal distribution over words in this topic from a Dirichlet distribution $Dir(\beta)$ with metaparameter β. For each document we draw a Multinominal distribution of the topics in this document from a Dirichlet distribution $Dir(\alpha)$ with metaparameter α. For each word in the document we draw a topic with respect to the topic distribution in the document and a word based on the word distribution for the drawn topic. Additionally, we draw a time stamp $t_i \sim Beta(\psi_{z_i})$ with $\psi_{z_i} = (a, b)$ the shape parameters of the Beta distribution and the location $l_i \sim N(\eta' \hat{z_{c(d)}}, \rho^2)$ with $\hat{z_{c(d)}}'$ the empirical topic frequencies for document d. The shape parameters ψ are estimated by the method of moments. For each topic z we estimate the mean $\hat{m}$ and sample variance s^2 of all time stamps from the documents that have been assigned this topic. We set $a = \hat{m} \cdot (\frac{\hat{m} \cdot (1-\hat{m})}{s^2} - 1)$ and $b = (1 - \hat{m}) \cdot (\frac{\hat{m} \cdot (1-\hat{m})}{s^2} - 1)$ for each topic. Finally, for the Normal distribution, η is estimated via EM methods, that minimize the likelihood during the estimation of the topic model: $L(\eta) = -\frac{1}{2\rho} \sum_d (y_d - \eta' \hat{z_d})^2 - -\frac{1}{2\sigma} \sum_k \eta_k^2$

Integrating the time stamp as Beta distributed random variable and the geographic location as Normal distributed random variable, we get for the probability of a topic z_i, given a word w in a chunk $c(d)$ with time stamp t and location l and all other topic assignments:

$$p(z_i | w, t, l, z_1, \cdots z_{i-1}, z_{i+1}, \cdots z_T) \propto \frac{N_{w,z_i} - 1 + \beta}{N_{z_i} - 1 + W \cdot \beta} \cdot (N_{d,z_i} + \alpha) \cdot \frac{(1 - t_{c(d)})^{a-1} \cdot t_{c(d)}^{b-1}}{Beta(a, b)} \cdot exp(-\frac{\|l_{c(d)} - \mu_{w,d}\|^2}{2\rho}) \quad (1)$$

Using this, we can estimate the topic model via Gibbs sampling.

References

Blei, David M. and McAuliffe, Jon D. Supervised topic models. In *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, pp. 121–128, 2007.

Blei, David M., Ng, Andrew Y., and Jordan, Michael I. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

Hasan, Samiul and Ukkusuri, Satish V. Urban activity pattern classification using topic models from online geolocation data. *Transportation Research Part C: Emerging Technologies*, 44(0):363 – 381, 2014. ISSN 0968-090X. doi: http://dx.doi.org/10.1016/j.trc.2014.04.003.

Hong, Liangjie, Ahmed, Amr, Gurumurthy, Siva, Smola, Alexander J., and Tsioutsiouliklis, Kostas. Discovering geographical topics in the twitter stream. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pp. 769–778, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1229-5. doi: 10.1145/2187836.2187940.

Kurashima, Takeshi, Iwata, Tomoharu, Hoshide, Takahide, Takaya, Noriko, and Fujimura, Ko. Geo topic model: Joint modeling of user's activity area and interests for location recommendation. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pp. 375–384, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1869-3. doi: 10.1145/2433396.2433444.

Speriosu, M., Brown, T., Moon, T., Baldridge, J., and Erk, K. Connecting Language and Geography with Region-Topic Models. 2010.

Wang, Xuerui and McCallum, Andrew. 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*, KDD '06, pp. 424–433, New York, NY, USA, 2006. ACM. ISBN 1-59593-339-5. doi: 10.1145/1150402.1150450.

Yin, Zhijun, Cao, Liangliang, Han, Jiawei, Zhai, Chengxiang, and Huang, Thomas. Geographical topic discovery and comparison. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pp. 247–256, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0632-4. doi: 10.1145/1963405.1963443.