Cosmos An International Journal of Art & Higher Education
A Refereed Research Journal
Vol 1/No 2 / Jul-Dec 2012 ISSN: 2319-8966

A STUDY ON LOCATION DRIVER MODELS

[1]Yadavaraj Gajiyavar

Introduction

In the location-driven model, we simply cluster the documents based on their locations. Each document cluster corresponds to one topic. $p(z|d)$ is the probability of topic z given document d from the location clustering result. We then estimate the word distribution μ^z for topic z by $p(w|z)\alpha \; {}_{d \in CD} \; p(w|d)p(d|z)$, where $p(d|z)$ is obtained from $p(z|d)$ by Bayes' theorem. In Festival dataset in Example 1, after we cluster the photos according to their locations, those photos close to each other are merged into the same cluster; And then we can generate the geographical topics (i.e., festival descriptions for each region) based on tags in each cluster. To cluster objects in 2-D space, we can use partition-based clustering like KMeans, density-based clustering like Mean-shift and DBScan, and mixture model based clustering. After we get the word distribution μz for topic $z \in Z$ based on the clustering result, we would like to know the topic distribution in geographical location $p(z|l)$ for topic comparison. Therefore, we prefer a generative model for location clustering because we can get the estimation of $p(l|z)$. $p(z|l)$ can be obtained by Bayes' theorem from $p(l|z)$. A popular generative model is Gaussian Mixture Model (GMM). In GMM, we assume that each cluster is mathematically represented by a Gaussian distribution and the entire data set is modeled by a mixture of Gaussian distributions. Although the location-driven model is straightforward, it is likely to fail if the document locations do not have good cluster patterns. A geographical topic may be from several different areas and these areas may not be close to each other. For example, in Landscape dataset in Example 2, there are no apparent location clusters; mountains exist in different areas and some are distant from each other.

Location-Text Joint Model

In this section, we propose a novel location-text joint model called LGTA (Latent Geographical Topic Analysis), which combines geographical clustering and topic modeling into one framework.

a. General Idea

To discover geographical topics, we need a model to en-code the spatial structure of words. The words that are close in space are likely to be clustered into the same geographical topic. In order to capture this property, we assume there are a set of regions. The topics are generated from regions instead of documents. If two words are close to each other in space, they are more likely to belong to the same region. If two words are from the same region, they are more likely to be clustered into the same topic.

1. The regions can be the areas in different cities, so the discovered geographical topics are different festivals.
2. The regions can be different areas such as the long strips along the coast and the areas in the mountains, so the discovered geographical topics are different landscapes. In Food data set in Example, the regions can be different areas that people live together, so the discovered geographical topics are different food preferences. We would like to design a model that can identify these regions as well as discover the geographical topics.

b. Latent Geographical Topic Analysis

In this, we introduce our LGTA framework for geographical topic discovery and comparison. The notations used in the framework are listed in Table.

Table: Notations used in LGTA framework

R	The region set, r is a region in R
	The topic distribution set for R, i.e., $\{\ {}_r\}_{r \in R}$
μ	The mean vector set for R, i.e., $\{\mu_r\}_{r \in R}$
	The covariance matrix set for R, i.e., $\{\ {}_r\}_{r \in R}$
α	The region importance weights

i) Discovering geographical topics

We would like to discover K geographical topics. The word distribution set of all the topics is denoted as θ, i.e., $\{\theta_z\}_{z \in}$z. Let us assume there are N regions and denote the region set as R. We assume that the geographical distribution of each region is Gaussian, parameterized as $(\mu;\ ) = f\{(\mu_r;\ {}_r)\}_{r \in R}$ where μ_r and ${}_r$ are the mean vector and covariance matrix of region r. α is a weight distribution over all the regions. $p(r|\alpha)$ indicates the weight of region r and ${}_{r \in R}\ p(r|\alpha) = 1$. Since topics are generated from regions, we use $\ = \{\ {}_r\}_{r \in R}$ to indicate topic distributions for all the regions. ${}_r = \{p(z|r)\}_{z \in Z}$ where $p(z|r)$ is the probability of topic z

1 *Research Scholar, Manav Bharti University, Solan, Shimla, HP*

given region r. $_{z\in Z}$p(z|r) = 1 for each r.

In our model, topics are generated from regions instead of documents and the geographical distribution of each region follows a Gaussian distribution. The words that are close in space are more likely to belong to the same region, so they are more likely to be clustered into the same topic. The generative procedure of the model is described as follows. To generate a geographical document d in collection D:

1. Sample a region *r* from the discrete distribution of region importance α, r $Discrete(\alpha)$.
2. Sample location ld from Gaussian distribution of μr and Σr.

$$p(l_d|\mu_r,\Sigma_r) = \frac{1}{2\pi\sqrt{|\Sigma_r|}}\exp(\frac{-(l_d-\mu_r)^T\Sigma_r^{-1}(l_d-\mu_r)}{2})$$

3. To generate each word in document d:
   (a) Sample a topic z from multinomial $_r$.
   (b) Sample a word w from multinomial θ_z.

Instead of aligning each topic with a single region, each topic in our model can be related to several regions. There-fore, our model can handle topics with complex shapes. Our model identifies the regions considering both location and text information. Meanwhile, it discovers the geographical topics according to the identified geographical regions. Let us denote all parameters by ψ . Given the data collection $\{(\mathbf{w_d}; l_d)_{d\in D}$ where $\mathbf{w}_d$ is the text of document d and l_d is the location of document d, the log-likelihood of the collection given [a] is as follows.

$$L(\psi; D) = \log p(D|\psi)$$

$$= \quad p(\mathbf{w}_d; l_d|\psi)$$

In this, we show how to estimate all the parameters using an EM algorithm.

ii) Comparing geographical topics

To compare the topics in different geographical locations, we need to get $p(z|l)$ in Definition 3 for all topics z Z given location $l = (x, y)$ where x is longitude and y is latitude. Given the estimated ψ, we estimate the density of location l given topic z.

$$\begin{aligned} p(l|z,\Psi) &= \sum_{r\in R} p(l|r,\Psi)p(r|z,\Psi) \\ &= \sum_{r\in R} p(l|\mu_r,\Sigma_r)\frac{p(z|r)p(r|\alpha)}{p(z|\Psi)} \end{aligned}$$

where $p(z|\psi) = {}_{r\ R}\, p(z|r)p(r|\alpha)$ and $p(l|\mu_r; {}_r)$ is based on

Equation

After we get $p(l|z; \psi)$, we can get $p(z|l; \psi)$ according to Bayes' theorem.

$$\begin{aligned} p(z|l,\Psi) &\propto p(l|z,\Psi)p(z|\Psi) \\ &\propto \sum_{r\in R} p(l|\mu_r,\Sigma_r)p(z|r)p(r|\alpha) \end{aligned}$$

References

1. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. Journal of Machine Learning Research, 3:993-1022, 2003.
2. L. Cao and L. Fei-Fei. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In ICCV, pages 1-8, 2007.
3. M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In KDD, pages 226-231, 1996.
4. T. Hofmann. Probabilistic latent semantic indexing. In SIGIR, pages 50-57, 1999.
5. L. S. Kennedy and M. Naaman. Generating diverse and representative image search results for landmarks. In WWW, pages 297-306, 2008.
6. Q. Mei, D. Cai, D. Zhang, and C. Zhai. Topic modeling with network regularization. In WWW, pages 101-110, 2008.