Detecting Non-Gaussian Geographical Topics in Tagged Photo Collections

Christoph Carl Kling[1], Jérôme Kunegis[2], Sergej Sizov[3], Steffen Staab[4]
[1,2,4] University of Koblenz-Landau, Koblenz, Germany
[3] Heinrich Heine University, Dusseldorf, Germany
{[1]ckling,[2]kunegis,[4]staab}@uni-koblenz.de, [3]sizov@hhu.de

ABSTRACT

Nowadays, large collections of photos are tagged with GPS coordinates. The modelling of such large geo-tagged corpora is an important problem in data mining and information retrieval, and involves the use of geographical information to detect topics with a spatial component. In this paper, we propose a novel geographical topic model which captures dependencies between geographical regions to support the detection of topics with complex, non-Gaussian distributed spatial structures. The model is based on a multi-Dirichlet process (MDP), a novel generalisation of the hierarchical Dirichlet process extended to support multiple base distributions. Our method thus is called the MDP-based geographical topic model (MGTM). We show how to use a MDP to dynamically smooth topic distributions between groups of spatially adjacent documents. In systematic quantitative and qualitative evaluations using independent datasets from prior related work, we show that such a model can exploit the adjacency of regions and leads to a significant improvement in the quality of topics compared to the state of the art in geographical topic modelling.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: General

General Terms

Algorithms, Theory

Keywords

Topic models, Dirichlet process, Graphical model

1. INTRODUCTION

Very large amounts of geographically distributed data are available to social media websites, companies and governments. Common sources of geospatial knowledge include user-generated content with GPS coordinates (e.g. photos and messages sent from smartphones), user data with information about the place of residence, or server logs with IP-based estimations of client locations. Such geographically distributed data are a rich information source, describing the environment in which they were created. Geographical knowledge can be exploited in a broad range of applications that take into account environmental and cultural differences between locations – ranging from marketing campaigns and recommender systems to data mining applications.

Under the bag-of-words assumption, both words and geographic positions can be modelled as being generated by latent topics in a probabilistic model. Existing approaches for geographical topic modelling adopt topic models such as probabilistic latent semantic analysis (PLSA) [10] or latent Dirichlet allocation (LDA) [5] and extend the models by assigning distributions over locations to topics, or by introducing latent geographical regions. In models which extend topics for spatial distributions (such as two-dimensional normal distributions) [15], topics with a complex (i.e. non-Gaussian) spatial distribution cannot be detected. In models with latent, Gaussian distributed regions [18], documents within a complex shaped topic area do not influence the topic distribution of distant documents within the same area. Therefore, topics with a complex spatial distribution such as topics distributed along coastlines, rivers or country borders are harder to detect by such methods. More elaborate models introduce artificial assumptions about the structure of geographical distributions by introducing hierarchical structures [2] or by defining Gaussian process kernels [1] in advance. Additionally, some approaches [12, 18] do not model document-specific topic distributions.

In contrast to existing models, the multi-Dirichlet process (MDP) based geographical topic model (MGTM) presented in this paper uses a MDP mixture model that groups documents by geographical regions. A geographical network between spatially adjacent regions is used to equalise topic distributions within coherent topic areas. Consequently, it allows for constructing generative models which provide a better data fit than existing approaches.

The rest of this paper is organised as follows. Section 2 reviews previous work on geographical topic detection. In Section 3 we describe the theory and the components of MGTM. We establish the baseline reference model, introduce its extension using the MDP, and develop corresponding inference procedures. In Section 4 we evaluate our basic model against state-of-the-art models, using reference datasets that have been used in prior work. Section 5 concludes our work.

WSDM'14, February 24–28, 2014, New York, New York, USA.
Copyright 2014 ACM 978-1-4503-2351-2/14/02 ...$15.00.
http://dx.doi.org/10.1145/2556195.2556218.

2. GEOGRAPHICAL TOPIC MODELS

A geographical topic model is a statistical model of a set of spatially distributed documents that uses word co-occurrences both within texts and within geographical regions. From an application perspective, location-aware topic models should satisfy the following top-level requirements:

(1) Modelling document-specific topic distributions: Documents typically cover a small set of topics, an assumption used for prediction (e.g. tag recommendation)
(2) Recognition of topics with complex (e.g. non-Gaussian) spatial distributions
(3) Detection of coherent topic regions that form complex shaped areas of similar characteristics (e.g. countries, seas, mountain ranges, etc.)
(4) Estimation of parameters from data: Topic models should not require prior knowledge for the parameter setting

In the following, we review existing models for geographical topic modelling, focusing on these requirements.

On a high level, existing approaches to geographical topic detection can be divided in two groups: models using a discrete set of locations and models using a continuous geographical distribution associated with topics. We focus on topic models for modelling text and location only, excluding extensions for other features such as time and authors.

2.1 Methods Using Discrete Locations

The model of Mei et al. [12] extends probabilistic latent semantic analysis (PLSA) [10] to model spatio-temporal information by mixing the topic distribution of documents with location- and time-specific topic distributions. Locations and timestamps are modelled as discrete sets. In practice, the division of data into location and time intervals results in sparse data.

Wang et al. [17] base their model on latent Dirichlet allocation (LDA) [5]. In their model, topics are multinomial distributions both over words and a discrete set of locations. The authors are aware of the fact that some related locations, such as locations within a country, are expected to share a similar topic distribution. They suggest to introduce a hierarchy between locations such as countries or cities to share topic information by merging those locations.

Finally, Yin et al. [18] present a *location driven model* (LDM) based on PLSA, in which geographically distributed data are clustered in a preprocessing step to obtain a discrete set of locations. In the model, all documents within a spatial cluster share a common topic distribution. Clusters are independent and share a global set of topics.

2.2 Methods Using Continuous Locations

A first continuous approach to geographical topic modelling was proposed by Sizov [15]: GeoFolk, a model similar to the model of Wang and based on LDA [5]. Instead of using a multinomial distribution over locations, every topic in GeoFolk has a Gaussian distribution on the coordinates of a document. The drawback of this kind of topic modelling clearly is the limited geographical distribution of topics: Every topic has a normal geographical distribution and topic areas that are not normal distributed are split into independent topics.

Yin et al. [18] therefore introduce *latent geographical topic analysis* (LGTA), an extended version of GeoFolk based on PLSA. Instead of directly assigning normal distributions to topics, in the model of Yin several normal distributions are assigned to *regions* which have a distribution over the set of topics. Clearly, there now can be several Gaussian regions sharing the same topic. Regions now take the role of discrete locations as in the model of Mei. Therefore, the model inherits the problem of merging regions of one kind.

In [2], Ahmed et al. present a hierarchical topic model which models both document and region specific topic distributions and additionally models regional variations of topics. Relations between the Gaussian distributed geographical regions are modelled by assuming a strict hierarchical relation between regions that is learned during inference.

A more general approach for modelling arbitrary, complex features such as geolocations was introduced by Agovic and Banerjee [1]. Given that the similarity between topic distributions of documents directly depends on their respective position in the feature space, topic distributions of documents can be sampled from a Gaussian process (GP) prior which encodes the similarity of documents in the feature space. However, it is unclear how to choose the right GP kernel in the geographical scenario, as the similarity of document-topic distributions across the geographical space typically is hard to predict and involves complex structures such as countries or geographical zones.

2.3 Drawbacks of Existing Methods

The existing geographical topic models described in the previous sections have major drawbacks with respect to the four requirements discussed at the beginning of this section. The models of Yin and Hong [11, 18] do not model document-specific topic distributions; the model of Sizov [15] cannot detect topics with a complex spatial distribution; and the model of Wang [17] supports the merging of semantically related geographical regions but lacks a general merging method. Finally, only the model by Ahmed et al. is parameter-free [2].

Model	**Requirements**			
	(1)	**(2)**	**(3)**	**(4)**
Mei et al. [12]		✓		
Wang et al. [17]	✓	✓	(✓)[1]	
GeoFolk [15]	✓			
LDM [18]		✓		
LGTA [18]		✓		
Agovic et al. [1]	✓	✓	(✓)[1]	(✓)[1]
Ahmed et al. [2]	✓	✓	(✓)[1]	✓
MGTM	✓	✓	✓	✓

Table 1: Requirements met by existing models, and by our model (MGTM). [1]**partial fulfilment**

Models based on a hierarchical relation between regions such as the model by Wang et al. [17] and Ahmed et al. [2] have drawbacks, not only in modelling complexity as mentioned in [17]. Particular hierarchical relations such as *city-state-country* might work for representing geographical topics such as languages or cultural behaviour. However they would be misleading e.g. for topics representing geographical features such as rivers or mountain areas. In most cases, there will be no hierarchy which fits all topics. Additionally, when introducing a hierarchy of Gaussian distributed regions as in [2], geographical topics which fit into such a hierarchy will be preferred over topics with a non-elliptic shape such as, say, coast lines, which would be poorly approximated by a hierarchy of Gaussian regions. Therefore, introducing a hierarchical relation between regions will prevent the model from properly learning topics whose complex geographical distribution does not fit such a simple hierarchical structure. Table 1 summarizes the requirements met by the models presented.

3. MODEL

We consider the general setting of documents consisting of words, and annotated with their geographic location. For topic modelling, words and location of documents are assumed exchangeable. Formally, we have a corpus of documents $D = \{d_1, \ldots, d_M\}$ of size $M = |D|$, and a document d_j consists of a set of N_j words denoted by $\mathbf{w}_j = (w_{j1}, \ldots, w_{jN_j})$ and a geographical location, a latitude and longitude pair $\text{loc}_j = (\text{lat}_j, \text{lon}_j)$. By de Finetti's theorem [5], words and location can be modelled as a mixture of independent and identically distributed random variables generated by latent factors. The document location is generated by L latent factors corresponding to geographical clusters associated with continuous distributions on the geographical space. The K latent factors assigned to words, written as topics $\theta_1, \ldots, \theta_K$, are multinomial distributions over the vocabulary of size V.

In the following section we present three novel geographical topic models: a basic model using a three-level hierarchical Dirichlet process, an extension that considers neighbour relations between regions by model selection and an improved version based on the multi-Dirichlet process introduced in this paper.

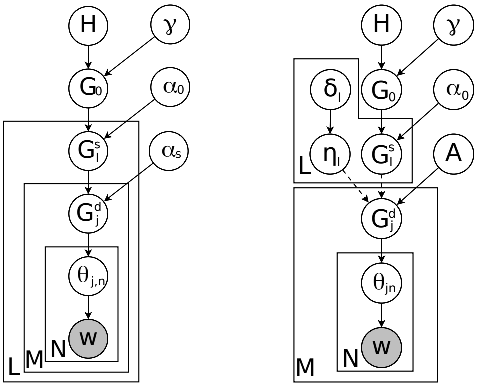
3.1 The Basic Model

For the basic geographical topic model, we model locations and words separately for several reasons:

Detection of coherent topic areas. The separation of spatial clusters and document semantics allows us to define meaningful neighbour relations between spatially adjacent clusters. In fact, as shown later, the use of these spatial adjacency relations allows us to detect coherent topic areas and to significantly improve the topic quality in the final model. Existing models that use continuous document positions do not allow a meaningful definition of spatial adjacency between geographical regions or topics as their position is influenced both by words and document locations.

Computational complexity. Probabilistic clustering methods in two- or three-dimensional space usually converge very fast, while samplers for probabilistic topic models usually take many iterations. Integrating both processes would result in a high computational overhead which is unacceptable for large datasets in real-world applications.

The basic topic model takes a set of geographical clusters as input. In order to get a clustering which also is a generative model of document positions, we fit a mixture of Fisher distributions to the data. The clusters are used to group documents in a three-level hierarchical Dirichlet process in order to ensure that documents within a geographical cluster share similar topics. This basic topic model is identical to the topic model for multiple corpora proposed by Teh et al. [16].



(a) Basic model (b) MGTM

Figure 1: The graphical model for (a) the basic model and (b) MGTM.

3.1.1 Geographical Clustering

Existing approaches for geographical topic modelling rely on a representation of document positions in Euclidean space of latitude and longitude. This causes problems for documents located close to the poles or to the International Date Line. Instead, we use the unit sphere as a model for the shape of the Earth. For geographical clustering, we assume that document locations follow a Fisher distribution. The Fisher distribution is a probability density function on a three-dimensional sphere comparable to an isotropic Gaussian distribution on the plane [9]. The Fisher distribution is defined as

$$f(x \mid \kappa, \mu) = \frac{\kappa}{2\pi(e^{\kappa} - e^{-\kappa})} e^{\kappa \mu^{\mathrm{T}} x}$$

where μ is the mean and κ is the concentration parameter. Given the number of Fisher distributions L and assuming a uniform prior, we use the expectation-maximisation algorithm for parameter estimation. For the concentration parameters we use the approximation given by Banerjee et al. [4]. To construct a non-parametric model where the number of regions is inferred from data, the number of clusters L can be sampled using a Dirichlet process that samples from a space of Fisher distributions. The geographical distribution of topics will depend on the number of regions. In order to ensure comparability, the number of regions is kept as a fixed parameter in the algorithms evaluated in this paper.

3.1.2 Topic Detection

The choice of the underlying topic model is crucial for the task of geographical topic detection. Existing methods typically are based on PLSA [18] or LDA [11, 15]. We decide to base our models on the hierarchical Dirichlet process (HDP) [16] instead as it is non-parametric, yields a sound generative model and supports a grouping of documents by external factors such as geographical clusters. The hierarchical Dirichlet process is a Bayesian approach to the topic detection problem and shares many properties with LDA: There is a Dirichlet-multinomial document-topic distribution assigned to every document and every topic is represented by a Dirichlet-multinomial topic-word distribution. However, in the hierarchical Dirichlet process the number of topics is not fixed. Instead, every topic is sampled from a Dirichlet process with a base distribution over H, the space of possible topic-word distributions. All model parameters can be sampled using hyperparameters, resulting in a fully non-parametric model. In order to share topics between documents, the document-topic distribution is sampled from a higher-level topic-distribution, i.e. a global topic distribution which is itself a draw from a Dirichlet process [16].

It is natural to extend the hierarchical scheme by adding layers for document groups with characteristic topic distributions. We use the three-layer Dirichlet process hierarchy for modelling document corpora proposed in [16] but group the documents by geographical regions instead using the spatial clustering of documents defined before.

The three-level hierarchical topic model using geographical clusters is defined as follows: Given a set of L geographical clusters, each cluster is a subset D_l of the document corpus. First we draw a global probability measure G_0 over the topic space from a Dirichlet process with base distribution H on the continuous topic space:

$$G_0 \sim DP(\gamma, H),$$

where γ is the concentration parameter for the Dirichlet process, influencing the sparsity of the global topic distribution. A symmetric Dirichlet prior is placed over H. The mixture proportions β for the global topic distribution are generated by a stick-breaking process $\beta \sim \text{Stick}(\gamma)$ [16]. For every geographical cluster, we draw a region-specific topic distribution G_l from the global distribution over the topic space G_0:

$$G_l^s \sim DP(\alpha_0, G_0), \qquad l = 1, \dots, L$$

with mixing proportions β_l^s and concentration parameter α_0. The documents from each region-specific document set D_l draw a document-specific topic probability measure from G_l:

$$G_j^d \sim DP(\alpha_s, G_l^s), \qquad d_j \in D_l$$

with mixing proportions π_j. All clusters share the common concentration parameter α_s. The resulting model is given in Figure 1(a). A collapsed Gibbs sampler and strategies for hyperparameter inference are given in [16].

3.2 The Neighbour Aware Model

We modify the basic model to include adjacency relations between geographical clusters. Using geographical neighbour relations has several advantages over the basic model:

Exploiting similarity for smoothing. Geographical clusters adjacent in space often are similar in their topic distribution. Most geographical topics cannot be approximated by a simple spatial probability distribution such as a Gaussian or Fisher distribution and for these complex topic areas, coherent sets of multiple spatial distributions are a reasonable approximation. Therefore adjacent regions may smooth their topic distributions to increase the probability of detecting such coherent topic areas.

Sharing emerging topics. In the basic model, new topics emerge locally, first on the document level, then on cluster level and finally on the global level. Under the assumption that adjacent clusters are likely to be similar, new topics should be actively shared with neighbour clusters. Sharing topics through a network of adjacent clusters is a dynamic, evolutionary process which is similar to the spread of memes in social networks: Strong topics, which describe observed documents well, will survive, while poor topics will perish during the sampling process [8].

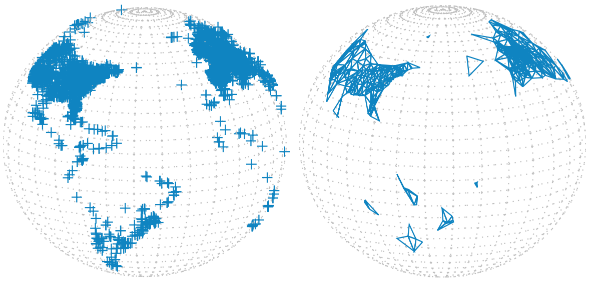


Figure 2: Document positions (left) and geographical network (right) for the car dataset (Section 4.1)

We call this model the *neighbour aware geographical topic model* (NAM). For defining a spatial adjacency relation, we decided to use the Delaunay triangulation, obtained as the dual of the Voronoi tesselation [3]. Triangles with side lengths greater than 1/8th of the earth radius are discarded. An example for the resulting geographical network is shown in Figure 2. Note that other definitions such as k-nearest-neighbour could have been used as well.

The idea behind the extended topic model is to include an uncertainty over the cluster membership of documents in clusters in order to more strongly connect topic distributions of adjacent geographical clusters. Each document topic distribution is assumed to be drawn either from the topic distribution of its geographical cluster or from one of the adjacent cluster distributions. P_l is the union of the cluster index l and the set of neighbour cluster indices, and λ_j indicates from which cluster-specific topic distribution β_r^s the document d_j was sampled. The set of topic distributions $\beta^{\mathbf{s}}$ can be used for Bayesian model selection: Given a uniform prior over the probability for a document d_j to be sampled from G_r^s with $r \in P_l$, the sampling equation for λ_j is

$$p(\lambda_j = r \mid \mathbf{z}, \mathbf{m}, \beta^{\mathbf{s}}) \propto \prod_{k=1}^{K} (\alpha_s \beta_{rk}^s)^{m_{jk}} \qquad (r \in P_l) \qquad (1)$$

which is identical to the sampling equation in [7], except that the weights of the document-specific topic distribution π_j are integrated out. The model structure then is sampled during Gibbs sampling and the rest of the sampler remains the same as for the basic model.

3.3 The Multi-Dirichlet Process Based Geographical Topic Model

The neighbour aware topic model clearly leads to an interaction between adjacent cluster-topic distributions. However, in some cases this interaction does not yield the intended smoothing. Consider the example of two adjacent geographical clusters which both have a high probability for two topics, while other geographical cluster-topic distributions assign very low probabilities to both of the topics. Now, the probability of this model would clearly be maximised if one of the two clusters has a very high probability for the first topic, and the other cluster for the second topic. This (unwanted) effect occurs in cases where there are only few adjacent clusters with high probabilities for a small set of topics. In practice, this is often the case as data are sparse and the number of geographical clusters is small.

To overcome this apparent drawback, we introduce a dynamic smoothing based on the *multi-Dirichlet process* (MDP), a generalisation of the Dirichlet process that combines multiple base measures into a single mixing distribution over the space of the base measures.

3.3.1 The Multi-Dirichlet Process

We define the *multi-Dirichlet process* (MDP) using a notation similar to that used in [16]. Let $G_1, \ldots, G_P$ be probability measures on a standard Borel space $(\Theta, \mathcal{B})$ associated with positive real parameters $\alpha_1, \ldots, \alpha_P$. We define the multi-Dirichlet process $MDP(\alpha_1, \ldots, \alpha_P, G_1, \ldots, G_P)$ as a probability measure G over $(\Theta, \mathcal{B})$, which for every finite measurable partition $(A_1, \ldots, A_r)$ of Θ yields a Dirichlet distributed random vector, denoted $(G(A_1), \ldots, G(A_r))$, with:

$$(G(A_1), \ldots, G(A_r)) \sim Dir(\sum_{p=1}^{P} \alpha_p G_p(A_1), \ldots, \sum_{p=1}^{P} \alpha_p G_p(A_r)) \tag{2}$$

In the following, we will refer to the base measures as parent distributions of the MDP. An alternative notation of the concentration parameters $\alpha_1, \ldots, \alpha_P$ is given by

$$A = \sum_{p=1}^{P} \alpha_p \qquad \eta_p = \frac{\alpha_p}{A}, \quad p \in \{1, \ldots, P\} \tag{3}$$

which gives a convenient parametrisation for the MDP:

$$MDP(A, \eta_1, \ldots, \eta_P, G_1, \ldots, G_P).$$

Using the alternative notation, the MDP can be understood as a Dirichlet process with base distribution $G_0 = \sum_{p=1}^{P} \eta_p G_p$, the weighted sum of parent distributions, and concentration parameter A. Given a set of observed samples from G, $\theta_1, \ldots, \theta_{i-1}$, the probability of a factor $\theta_i \in \Theta$ to be sampled from G can be estimated by integrating out G using the properties of the Dirichlet distributed partitions [13] and replacing the base measure with the weighted sum of parent distributions:

$$\theta_i \mid \theta_1, \ldots \theta_{i-1} \sim \frac{1}{i-1+A} \sum_{j=1}^{i-1} \delta(\theta_j) + A \sum_{p=1}^{P} \frac{\eta_p}{i-1+A} G_p \tag{4}$$

with $\delta(\theta_j)$ being the Dirac delta, giving weight to a single point θ_j. We immediately see that a MDP with a single parent distribution yields a standard Dirichlet process.

3.3.2 Inference

We sample for topic assignments by extending the inference strategies using the "Chinese restaurant franchise" representation given in [16]: For a given two-level hierarchical Dirichlet process, global "dishes" are introduced, corresponding to the Dirichlet distributed random variables on the first level from which the Dirichlet process of the second level samples factors θ_j . For factor sampling, customers corresponding to the factors θ_j form Dirichlet distributed groups sitting at tables in a restaurant and all customers at a table share the same dish. The number of customers at the ith table is given by m_i and the tables are samples from the Dirichlet distributed base distribution of dishes. A detailed explanation of the Chinese restaurant process and its parameters is given in [16]. The Gibbs sampling equation for topic assignment z_{ji} of word w_{ji} in document d_j is:

$$p(z_{ji}{=}k \mid \mathbf{z_{-ji}}, \mathbf{m}, \beta^{\mathbf{s}}) \propto (m_{jk} + \sum_{p \in P_l} \alpha_p \beta^s_{pk}) f_k^{-x_{j}i}(x_{ji}) \tag{5}$$

for topics already sampled, and

$$p(z_{ji}{=}k^{\text{new}} \mid \mathbf{z_{-ji}}, \mathbf{m}, \beta^{\mathbf{s}}) \propto (\sum_{p \in P_l} \alpha_p \beta^s_{pu}) f_{k^{\text{new}}}^{-x_{j}i}(x_{ji}) \tag{6}$$

for new topics where $\beta^{\mathbf{s}}_{\mathbf{p}}$ are mixing proportions of the parent distributions, $f_k^{-x_{j}i}(x_{ji})$ is a topic-specific probability function with parameters from the parent distribution and z_{-ji} denotes the set of all topic assignments except for z_{ji}. The number of customers in document d_j assigned to the kth factor is given by m_{jk}.

For sampling the number of components, the MDP can be interpreted as a multinomial extension of a Dirichlet process. This becomes apparent if we use the alternative representation from Eq. 3. Substituting α_0 by (a sum of) $A\eta_{lp}$ in Equation 40 from [16] gives:

$$p(m_{jk}{=}\mathbf{m} \mid \mathbf{z}, \mathbf{m_{-jk}}) = \frac{\Gamma(\sum_{p \in P_l} A\eta_{lp}\beta^s_{pk})}{\Gamma((\sum_{p \in P_l} A\eta_{lp}\beta^s_{pk}){+}n_{jk})} s(n_{jk}, m_{jk\cdot}) \cdot \\ A^{m_{jk\cdot}} \binom{m_{jk\cdot}}{m_{jk1}, \ldots, m_{jkP}} \prod_{p \in P_l} (\eta_{lp}\beta^s_{pk})^{m_{jkp}} \tag{7}$$

where l is the index of the MDP of document d_j with parent distributions P_l. $s(n, m)$ denotes the unsigned Stirling numbers of the first kind and $\binom{m_{jk\cdot}}{m_{jk1}, \ldots, m_{jkP}}$ the multinomial coefficient. Note that we keep track of the number of tables m_{jkp} for every parent distribution and sample them simultaneously. $m_{jk\cdot}$ denotes the sum of tables over all parents, n_{jk} is the number of customers (topic assignments) for a given document and topic. For sampling the tables, we first drop the Gamma functions which do not depend on $\mathbf{m}$, sample for the sum of tables $m_{jk\cdot}$ and then sample the parent specific table counts m_{jkp} from a multinomial with normalised parameters $\eta_{lp}\beta_{pk}$.

Sampling the weights β_p for each G_p is done using $m_{\cdot kp}$, the sum over all tables of topic k and parent p from documents with parent distribution G_p. If G_p is sampled from a parent Dirichlet process with concentration parameter α_0 and weights β, then

$$\beta_p \sim Dir(m_{\cdot 1p} + \alpha_0\beta_1, \ldots, m_{\cdot Kp} + \alpha_0\beta_K, \alpha_0\beta_u) \tag{8}$$

where β_k denotes the weight of topic k in the parent Dirichlet process and β_u is the weight of the previously unseen topics.

3.3.3 Estimation of Concentration Parameters

Sampling for concentration parameters α_p is similar to the sampling in Dirichlet processes as described in [16]. Instead of directly sampling the concentration parameters α_p, we first sample A and then sample η. We use the probability of the total table counts for all documents in the MDP:

$$p(\mathbf{m_l} \mid \mathbf{n}, \mathbf{m}, \eta, A) =$$

$$\prod_{j \in D_l} \frac{\Gamma(A)}{\Gamma(A + n_{j\cdot})} s(n_{j\cdot}, m_{j\cdot}) A^{m_{j\cdot}} \cdot \binom{m_{j\cdot}}{m_{j1}, \ldots, m_{jP}} \prod_{p \in P_l} \eta_{lp}^{m_{jp}} \tag{9}$$

where D_l is the set of documents which is sampled from the MDP with index l. The left part of the equation is identical to Equation 44 in [16] with parameter A as concentration parameter. Therefore sampling for A is identical as for a normal DP. The document specific table counts $m_{j\cdot}$ are obtained by summing over the sampling results from Equation 7. Obviously, the right side of Equation 9 is a multinomial again. As η governs the influence of parent distributions, we can set a symmetric Dirichlet prior over the sampling parameters for smoothing. For a MDP with index l we then estimate η_l using:

$$\hat{\eta}_{lp} = \frac{m_{\cdot p} + \delta}{m_{\cdot\cdot} + |P_l|\delta} \qquad \eta_l \sim Dir(\delta_l) \tag{10}$$

3.3.4 MDP-based Model

The extension of NAM for the multi-Dirichlet process, the multi-Dirichlet process geographical topic model (MGTM), is obtained by replacing the model selection for uncertain cluster memberships by multi-Dirichlet processes. Instead of sampling for document memberships from the set of potential parent distributions P_l, we use P_l as indices of the parent base distributions of the MDP. Every document has several parental base distributions G^s_r with $r \in P_l$, the indices of the region of the document and the adjacent regions. A schematic representation of the resulting dependencies is shown in Figure 3. The weight of region r in the MDP is given by η_{lr} and is sampled during the topic sampling process. With a concentration parameter $\delta > 1$ we smooth the cluster weights of parent distributions. The resulting model is shown in Figure 1(b). The dashed arrow connecting the cluster specific topic distributions G^s_l and the document specific distributions G^d_j indicates that not every cluster specific distribution is a parent base distribution of each MDP.

In MGTM, all documents of a given region share the same MDP and thus the same weights η_l for the parent topic distributions of the region and its neighbour regions. Each region stores the influence of its adjacent regions on the topic distribution of the contained documents in the parameter η_l which is adjusted during the Gibbs sampling process. Given P_l, the union of the region index l and its neighbour region indices, η_l assigns a probability to every region to be chosen as a base topic distribution by the documents D_l in region l, and thus η_{lr} is an indicator of similarity between the lth region and its rth neighbour.

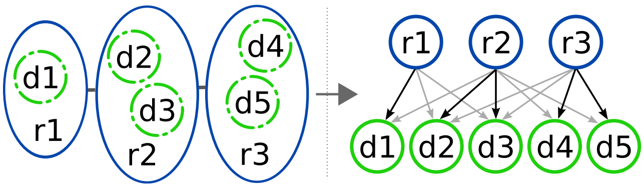


Figure 3: The geographical adjacency of regions (left) is used in the model to derive dependencies of document-specific topic distributions from the topic distributions of regions (right). Dependencies from regions adjacent to the region of a document are shown in grey.

As mentioned, it is possible to smooth the influence of adjacent regions by setting a Dirichlet prior over η. In contrast, a prior for the model selection of NAM (Eq. 1) only could re-weight but not smooth the probability for cluster memberships. Using the MDP, we obtain a more flexible and stable framework for selecting base distributions that correspond – in our case – to cluster-specific topic distributions.

The resulting model has an important advantage: For the neighbour aware model, we made the assumption that adjacent spatial clusters are similar and tried to smooth their topic distributions. The MDP ensures that the model creates homogeneous topic distributions for similar adjacent regions and at the same time prevents a smoothing of dissimilar regions by adjusting the influence parameter η during the sampling process, leading to a *dynamic smoothing* of topic regions.

4. EVALUATION

In this section, we demonstrate the ability of our model to improve the quality of topics by detecting more accurate, coherent topic areas. The evaluation is in four parts: First, we compare the basic model, the neighbour aware model, the multi-Dirichlet process model and a state-of-the-art model for geographical topic detection, LGTA by Yin et al., using the datasets and parameters given in [18]. Second, we analyse the influence of increased region parameters on topic quality for all four methods. We then compare the runtime of the presented methods on the largest dataset for growing region parameters. Finally, we assess the topic quality of LGTA and MGTM on the largest dataset with a user study.

4.1 Datasets

As the evaluation of topic models is heavily dependent on the datasets used for comparison, we use existing datasets in order to guarantee a fair comparison. We use the datasets from [18] created for the evaluation of the LGTA model. The datasets consist of photographs with geographic coordinates and text tags from the photo sharing service Flickr. The landscape dataset contains 5,791 photos, tagged by "landscape" together with the terms *mountains, mountain, beach, ocean, coast, desert* from within the US. Topic models should recognise separated topics for mountain regions, coastal regions and desert areas as they belong to different, almost mutually exclusive geographical landscapes. The activities dataset contains 1,931 images, taken within the US and tagged by "hiking" and "surfing". These two activities should be recognised as own topics of behaviour. The car dataset contains 34,707 globally distributed images annotated with *chevrolet, pontiac, cadillac, gmc, buick, audi, bmw, mercedesbenz, fiat, peugeot, citroen* or *renault*, filtered for event-like images tagged with *autoshow, show, race, racing*. Only the tags from the set of car brands were kept. Concerning the geographical topics, American, German and French car brands are expected to be detected. The manhattan dataset consists of images from New York containing the tag *manhattan*. Different parts of Manhattan should be detected. For the food dataset, Yin et al. filtered geotagged photos containing the tags *cuisine, food, gourmet, restaurant, restaurants, breakfast, lunch, dinner, appetizer, entree,dessert* and kept 278 co-occurring tags. Cultural food patterns such as national cuisines are latent topics hidden in the data [18]. An overview of the data is given in Table 2.

Table 2: Collection period, document count (M) and vocabulary size (V) of the datasets used for comparison [18]

Dataset	Collection period	M	V
Landscape	09/01/2009 – 09/01/2010	5.791	1.143
Activity	09/01/2009 – 09/01/2010	1.931	408
Manhattan	09/01/2009 – 09/01/2010	28.922	868
Car	01/01/2006 – 09/01/2010	34.707	12
Food	01/01/2006 – 09/01/2010	151.747	278

4.2 Experimental Setting

In order to test the generalisation performance of geographical topic models, we calculate the word perplexity. The word perplexity is a widely-used measure in language modelling, corresponding to the inverse of the geometric mean of the per-word likelihood of held-out documents [5] that can be understood as the ability of a topic model to predict words of new documents. Lower perplexity values indicate a better model. Yin et al. used the word perplexity in their evaluation of LGTA which they showed to be superior to GeoFolk [15] and a set of basic geographical topic models. Models that outperform LGTA in perplexity therefore also outperform GeoFolk and the basic methods evaluated by Yin [18].

The comparison between the basic Dirichlet process-based model and its extensions is needed to test the effect of including the additional information of the geographical network in the model and to compare the multi-Dirichlet process with a smoothing mechanism based on model selection.

For each perplexity calculation, a random 80% / 20% split is used to create a training set D_{train} and a test set D_{test}. As explained in Section 3, each document d_j is represented by a word set $\mathbf{w}_j$. We calculate the likelihood of words in held-out documents using the location of documents, the set of topics and other parameters sampled from the training dataset. The word perplexity is defined as [5]

$$\textit{perplexity}(D_{\text{test}}) = \exp\left(\frac{-\log(\prod_{d_j \in D_{\text{test}}} p(\mathbf{w}_j))}{\sum_{d_j \in D_{\text{test}}} |\mathbf{w}_j|}\right).$$

For the hierarchical Dirichlet process-based models, the probability of a document is given by

$$p(\mathbf{w_j}) = \prod_{w_i \in \mathbf{w}_j} \sum_{k=1}^{K+1} \theta_{k,w_i} \pi_{j,k}$$

where θ_{k,w_i} is the probability of word w_i under topic k and $\pi_{j,k}$ is the document-topic distribution for topic k. $K+1$ denotes the index of a previously unseen topic and the topic-word probability θ_{K+1,w_i} is given by $\theta_{K+1,t} = 1/V$, $t \in \{1,\ldots,V\}$ as we use a symmetric Dirichlet prior over the topic space H from which new topics are drawn.

For convenience, the parameters for the (multi-)Dirichlet processes are from the first model in [16]. A $Gamma(1, 0.1)$ prior is assigned to γ and a $Gamma(1, 1)$ prior to α_0, α_s and A. The concentration parameters are initialised to 1. For the multi-Dirichlet process, the weights η of parent distributions are initialised to $1/P$ and the concentration parameter is set to $\delta = 10$. The base measure H is a symmetric Dirichlet distribution with concentration parameter 0.5, except for the car dataset where the parameter is set to 5 for smooth topic-word distributions, as all car brands are expected to appear in all topic areas. We set the number of iterations of the Gibbs sampler to a low value of 200. The source code for MGTM is available from: `http://c-kling.de/mgtm`.

For the evaluation of LGTA, the parameters from the original paper [18] were used. The stopping criterion is set to a change in log-likelihood lower than 0.0001 and the background model weight is set to 0.1. LGTA sets a parameter for the number of normal distributed regions which is analogous to the number of geographical clusters (Fisher distributions) in the models presented in this paper. For comparison, we use identical numbers of regions. The setting depends on the dataset and is taken from the LGTA paper.

As the number of detected topics varies for models based on the hierarchical Dirichlet process, each of the HDP-based methods is run 100 times on each dataset and the resulting perplexity is averaged for topic counts with at least ten samples for all three models. The perplexity of LGTA is calculated for the same number of topics by averaging over ten runs.

4.3 Comparison with LGTA

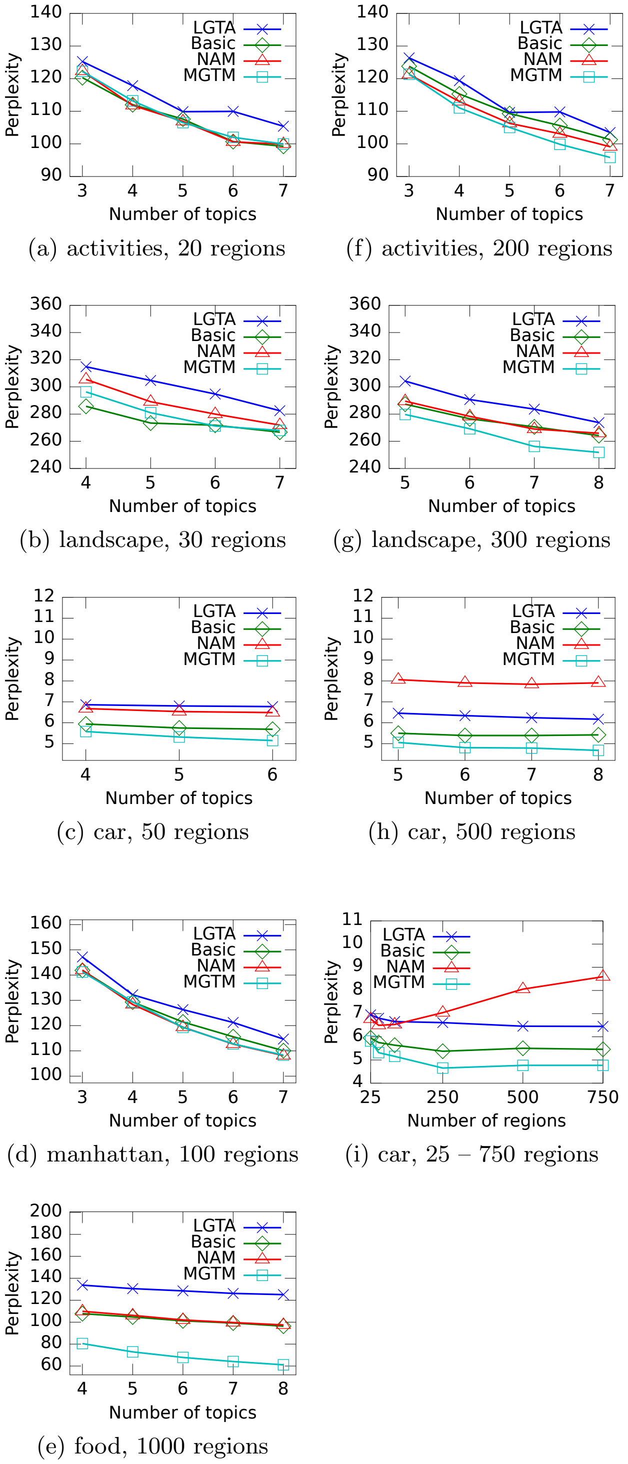
Resulting perplexity scores for each model are given in Figure 4 (a)-(e). The experiments show that the base model, NAM and MGTM are superior to LGTA for all datasets. This finding can be explained by the ability of the models to model document-specific topic distributions that cannot be detected by LGTA. However, the performance of the HDP-based models differs. For the globally distributed datasets (car and food), MGTM performs significantly better than the base model and NAM. In contrast, for local datasets with a small number of regions, all HDP-based methods perform comparably well.

The dynamic smoothing by MGTM helps to detect coherent topic regions and can effectively improve the topic quality for large, complex structured data while for simple datasets the basic model performs similar or even better.

4.4 Effect of the Region Parameter

To further investigate the behaviour of MGTM for complex structured regions, we repeat the experiments for the three datasets with the smallest number of regions but increase the region parameter by a factor of ten. The results are given in Figure 4, (f)-(h). For an increased number of regions, MGTM shows an improved perplexity for all three datasets and outperforms the basic and neighbour-aware model, demonstrating its ability to effectively exploit the adjacency relation between regions for sharing topic information.

The effect of a growing number of regions for the car dataset at a fixed number of five topics is plotted in Figure 4, (i). The car dataset is adequate to demonstrate the usage of geographic information in the topic models as the documents contain only a single word, meaning that intra-document co-occurrences of words do not contribute to the



(a) activities, 20 regions

(f) activities, 200 regions

(b) landscape, 30 regions

(g) landscape, 300 regions

(c) car, 50 regions

(h) car, 500 regions

(d) manhattan, 100 regions

(i) car, 25 – 750 regions

(e) food, 1000 regions

Figure 4: Comparison of average word perplexity for LGTA, the basic model, NAM and MGTM for given topic and region counts. Lower values are better. The HDP-based models sample the number of topics, therefore the average perplexity is shown for topic counts that occur in at least 10 out of 100 runs for each method. For the car dataset, the perplexity at five topics is shown in (i) for growing numbers of regions.

model creation. We observe that LGTA, the basic model and MGTM improve the topic quality for increased region parameters, but the improvement of MGTM is considerably larger than for LGTA and the basic model. The dynamic smoothing based on the MDP helps to improve the topic quality by exchanging topic information between similar adjacent clusters. In contrast, the perplexity of NAM dramatically gets worse for larger region counts due to the instability of the naive smoothing mechanism based on model selection.

4.5 User Study

A lower perplexity does not always indicate an improved topic quality [6]. Therefore we conducted a user study evaluating the semantic coherence of words within the topics detected by LGTA, the basic model, NAM and MGTM for the food dataset with 1000 regions at 4, 6 and 8 topics. Figure 3 shows the words with the highest probability for the topics detected by LGTA and MGTM at eight topics. Participants performed the “word intrusion” task introduced by Chang et al. [6]: For evaluating a topic, users are presented with a set of six words, which consists of the five words with the highest probability under the topic and a word from another topic from the same model. The user’s task is to “find the word which does not fit with the other words”. In case of semantically coherent topic words, the intruder can be easily found. To additionally test the interaction between topics, the intruding word was chosen from a set of words which had a low probability (not in the top 25 words) in the evaluated topic and a high probability (top 5 of the remaining words) in another topic. The study was conducted with 31 users which were presented with word sets in a random order of models and topics. Only one word set per model-topic combination was shown to the user and a total of 1,446 of word sets were rated.

In order to measure the quality of a model, we calculate an overall model precision (the percentage of intruders detected by participants) and per-topic precisions within a given model.

Table 4 shows the average precision and the median of the per-topic precisions for all four models. Clearly, MGTM performs considerably better with both an average model precision and median model precision of around 0.8 compared to about 0.6 for LGTA. Only for the case of 4 topics, the neighbour-aware model shows a comparable precision. However, for 6 topics the precision is worse compared to LGTA and for 8 topics it is only slightly better. Similarly, the basic model is worse than LGTA for 4 topics and only slightly better for 6 and 8.

To analyse the distribution of the per-topic precision, the corresponding box-and-whisker plot for the case of 8 topics is given in Figure 5. Clearly, the quality of the topics de-

	4 topics		**6 topics**		**8 topics**	
	avg /	median	avg /	median	avg /	median
LGTA	0.67	0.64	0.57	0.57	0.60	0.58
Basic	0.45	0.57	0.63	0.61	0.64	0.58
NAM	**0.79**	0.75	0.51	0.48	0.64	0.60
MGTM	**0.79**	**0.80**	**0.82**	**0.81**	**0.78**	**0.75**

Table 4: Model precision and median of per-topic precisions for LGTA, the basic model, NAM and MGTM on the food dataset with 1000 regions.

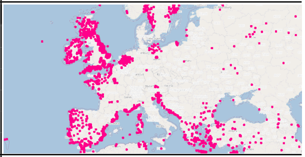
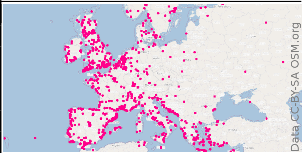
	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Map of Topic 1
LGTA	fish	chocolate	japanese	vegetarian	wine	chinese	mexican	sushi	
	seafood	cheese	sushi	vegan	italian	chicken	bbq	thai	
	rice	bread	ramen	chocolate	coffee	noodles	chicken	korean	
	shrimp	fish	fish	baking	french	soup	burger	japanese	
	crab	wine	noodle	bread	pizza	rice	sandwich	salmon	
	lobster	tapas	sashimi	cheese	chocolate	vietnamese	fries	rice	
	chicken	orange	noodles	bacon	bakery	dimsum	hamburger	tuna	
MGTM	seafood	chocolate	japanese	salad	wine	chinese	mexican	bbq	
	fish	icecream	sushi	cheese	pizza	thai	tacos	burger	
	lobster	strawberry	fish	tomato	coffee	chicken	taco	fries	
	shrimp	baking	ramen	bread	italian	rice	salsa	hamburger	
	crab	cream	sashimi	chicken	pasta	soup	burrito	grill	
	wine	coffee	rice	fish	cheese	noodles	chicken	chicken	
	salmon	pie	salmon	vegetarian	french	korean	chips	sandwich	

Table 3: Topic descriptions for the food dataset detected by LGTA and MGTM. The topics of MGTM were reordered to match the topics of LGTA. The maps show the positions of documents with an above average probability for Topic 1 as detected by LGTA (top) and MGTM (bottom).

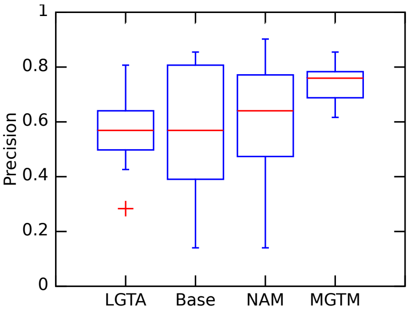


Figure 5: Boxplots of model precision for LGTA, the basic model, NAM and MGTM on the food dataset with 1000 regions, 8 topics. Higher is better.

tected by the basic model and NAM is mixed – the per-topic precision ranges from very low values of about 0.2 to high values greater than 0.8. The precision of LGTA is more consistent, as the topic precision is closer distributed around the median with only one outlier. Finally, the per-topic precision of MGTM is high for all topics and most homogeneous among all models.

The human evaluation supports the findings of the perplexity comparison – indeed, MGTM detects semantically more coherent topic-word distributions by exploiting the spatial structure of topics using dynamic smoothing. The differences between LGTA and MGTM can be explained by taking the topics from Table 3 as an example:

Topic quality. The key difference between LGTA and MGTM is the semantic coherence of the topic-word distributions. Topic 2 of LGTA assigns high probabilities to the terms *chocolate*, *cheese*, *bread* and *fish*. By contrast, the most similar topic of MGTM contains the semantically related words *chocolate*, *icecream*, *strawberry* and *baking* – all related to desserts. The incoherent word-selection of LGTA is due to the fact that these tags often occur within a small region and repetitions of similar word combinations in adjacent regions are not taken into account sufficiently.

Globality. The first topic from LGTA and the corresponding topic from MGTM mostly contain terms related to seafood. Clearly, the words *rice* and *chicken* from the seafood topic of LGTA do not fit – they often occur in Asia, where many photos of seafood are located. The seafood topic from MGTM is more coherent – it assigns a high probability to the word *wine*, as it is often consumed together with fish across Europe. From this example, we can see that the topics of LGTA are heavily influenced by local, region-specific patterns in tag co-occurrences whereas MGTM more is influenced by intra-document co-occurences of tags and the global distribution of topics. The reason is that LGTA does not model document specific topic distributions, instead, all documents within a region share the same topic distribution and therefore individual deviations from the regional topics are not recognised in the model. In contrast, MGTM allows for document-specific topic distributions and permits deviations from the regional topic distribution. By detecting single documents fitting the topic of seafood in coastal regions all over the world, and by exchanging this topic information over the network of adjacent regions, a global topic of seafood is established.

Support for non-compact topics. Some of the topics detected in the food dataset are expected to exhibit a complex spatial distribution. As mentioned before, MGTM is able to detect such complex spatial structures. To give an example, the maps in Table 3 show the geographical distribution of documents with a higher-than-average probability for the seafood topic (Topic 1) as detected by LGTA and MGTM. We expect Topic 1 to have a distribution along coastlines. We see that this is the case for Topic 1 of both LGTA and MGTM, which covers both countries where fish is regularly eaten (such as the UK and the Netherlands) and countries where the seafood topic mainly appears at the coast (e.g. Spain, France). However, the geographical distribution of the seafood topic of LGTA has a large gap on the coast between Spain and France and is not detected in Denmark or on mainland Italy. The reason is that there are not many photos showing seafood in those areas and therefore the evidence is not sufficient for LGTA. Due to the dynamic smoothing of adjacent areas, MGTM still is able to detect such topics and thus correctly detects the seafood topic in documents along the whole coastline as seen on the map.

4.6 Runtime Comparison

Another advantage of the HDP-based models is the separation of geographical clustering and the topic sampling step. By excluding the distance calculation between every document and every region centre from the slowly converg-

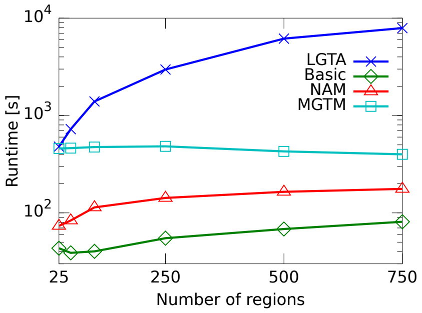


Figure 6: Average runtimes (in seconds) of LGTA, the basic model, NAM and MGTM for the food dataset at growing numbers of regions. Note that the y-axis is a log scale.

ing topic sampling process, we expect to dramatically decrease the runtime of our methods.

For comparing the runtime of the distinct methods, we optimised the implementation of LGTA provided by Yin and measured the runtime on the largest dataset for different settings of the region parameter on a 2.8GHz CPU with 72GB of RAM using a single core. The topic parameter of LGTA is set to 7. The runtime in seconds is given in Figure 6. We see that the runtime of LGTA linearly grows with a higher region count, as in every iteration every document has to be compared with every region for sampling its membership probability. On the other hand, the models presented in this paper use a separate geographical clustering step that can be efficiently implemented and takes only a fraction of the total runtime. The topic sampling is practically not influenced by the number of regions as it only creates additional region-specific topic distributions. MGTM shows a higher runtime compared to the basic model as the sampling of region-topic distributions in the multi-Dirichlet process is more expensive than in a normal hierarchical Dirichlet process. In return, for a larger number of regions, MGTM detects and merges topics with a coherent spatial distribution which results in a lower number of detected topics and a slightly decreased runtime. MGTM thus has a significantly reduced runtime and can be applied to much larger datasets.

Furthermore, it is straightforward to implement a distributed algorithm for MGTM as the distributed Gibbs sampling equations for hierarchical Dirichlet processes from [14] can be directly applied to multi-Dirichlet processes and region-specific topic distributions can be shared across processors with dependent document-topic distributions using the same technique as for sharing topics across processors.

5. CONCLUSION

The results from the user study and extensive quantitative evaluation of our model show a clear improvement in topic quality compared to state-of-the-art methods in topic modelling. MGTM detects more meaningful topics as measured by the perplexity and higher precisions in the user experiments. Additionally, the runtime analysis demonstrates that our method is highly efficient and thus suitable for large-scale applications. The model is the first to make use of adjacency relations between groups of documents for a dynamic smoothing of topic distributions. Our method is based on a multi-Dirichlet process (MDP), a generalisation of the Dirichlet process introduced in this paper. The improved performance of MGTM at higher numbers of regions shows that in real-world datasets, many geographical topics have a complex, non-Gaussian spatial distribution and that their detection can be supported. The presented topic model is just one example of how to use the MDP for dynamic smoothing. We currently are experimenting with integrating the MDP in other existing topic models to account for relations between groups of documents.

6. ACKNOWLEDGEMENTS

This work was supported by the EU 7th FP under grant number IST-FP7-288815 in project Live+Gov (liveandgov.eu)

7. REFERENCES

[1] A. Agovic and A. Banerjee. Gaussian process topic models. *CoRR*, abs/1203.3462, 2012.

[2] A. Ahmed, L. Hong, and A. Smola. Hierarchical geographical modeling of user locations from social media posts. In *WWW*, 2013.

[3] F. Aurenhammer. Voronoi diagrams – a survey of a fundamental geometric data structure. *ACM Comput. Surv.*, 23(3):345–405, 1991.

[4] A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra. Clustering on the unit hypersphere using von Mises–Fisher distributions. *JMLR*, 6:1345–1382, 2005.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *JMLR*, 3:993–1022, Mar. 2003.

[6] J. Boyd-Graber, J. Chang, S. Gerrish, C. Wang, and D. Blei. Reading tea leaves: How humans interpret topic models. In *NIPS*, 2009.

[7] K. R. Canini and T. L. Griffiths. A nonparametric Bayesian model of multi-level category learning. In *AAAI*, 2011.

[8] R. Dawkins. *The Selfish Gene*. OUP, 2006.

[9] R. Fisher. Dispersion on a sphere. *Royal Society*, 217(1130), 1953.

[10] T. Hofmann. Probabilistic latent semantic analysis. In *UAI*, pages 289–296, 1999.

[11] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsiouliklis. Discovering geographical topics in the Twitter stream. In *WWW*, pages 769–778, 2012.

[12] Q. Mei, C. Liu, H. Su, and C. Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *WWW*, pages 533–542, 2006.

[13] R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *J. Comp. Graph. Stat.*, 9(2):249–265, 2000.

[14] D. Newman, A. U. Asuncion, P. Smyth, and M. Welling. Distributed algorithms for topic models. *JMLR*, 10:1801–1828, 2009.

[15] S. Sizov. GeoFolk: latent spatial semantics in Web 2.0 social media. In *WSDM*, pages 281–290, 2010.

[16] Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical Dirichlet processes. *JASA*, 2006.

[17] C. Wang, J. Wang, X. Xie, and W.-Y. Ma. Mining geographic knowledge using location aware topic model. In *GIR*, pages 65–70, 2007.

[18] Z. Yin, L. Cao, J. Han, C. Zhai, and T. S. Huang. Geographical topic discovery and comparison. In *WWW*, pages 247–256, 2011.