GeoFolk: Latent Spatial Semantics in Web 2.0 Social Media

Sergej Sizov
University of Koblenz, Germany
Institute for Web Science and Technologies
sizov@uni-koblenz.de

ABSTRACT

We describe an approach for multi-modal characterization of social media by combining text features (e.g. tags as a prominent example of short, unstructured text labels) with spatial knowledge (e.g. geotags and coordinates of images and videos). Our model-based framework GeoFolk combines these two aspects in order to construct better algorithms for content management, retrieval, and sharing. The approach is based on multi-modal Bayesian models which allow us to integrate spatial semantics of social media in a well-formed, probabilistic manner. We systematically evaluate the solution on a subset of Flickr data, in characteristic scenarios of tag recommendation, content classification, and clustering. Experimental results show that our method outperforms baseline techniques that are based on one of the aspects alone. The approach described in this contribution can also be used in other domains such as Geoweb retrieval.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models

General Terms

Algorithms, Theory

Keywords

Web2.0, Tagging, Geodata, Bayesian learning, MCMC

1. INTRODUCTION

The rapid growth of modern Web 2.0 folksonomies originates in the ease of collaborative content generation and content sharing for non-expert mass users. In the recent years, substantial research progress was achieved regarding folksonomy mining and designing suitable recommendation, search, and retrieval solutions. However, some fundamental questions remain still open. One of the key challenges for Web 2.0 content management and sharing is a deeper understanding of the complex, multi-modal content nature. In general, a Web 2.0 folksonomy provides several facets of media knowledge, including content annotations (e.g. image tags in Flickr), spatial knowledge (e.g. GPS coordinates, names of locations), user-specific aspects, and media properties. However, the construction of appropriate rich folksonomy models and their formal analysis remain a difficult enterprise. For this reason, existing solutions and models are often restricted to particular, most 'promising' folksonomy aspects (e.g. analysis of social networks or content annotations) which are then considered in an isolated manner.

Despite the huge amount of shared content in popular folksonomies (such as Flickr[1] or YouTube[2]), the available data for particular modeling aspects is often sparse and imprecise. The image sharing scenario with Flickr is characteristic for this kind of problems. On one hand, the concentration of media related to popular locations, events and objects is awesome. For instance, the query '*london*' returns on Flickr over 7.5 Mio matches (observed in 2009). However, the prevalent majority of these resources is annotated by few, quite heterogeneous, tags. This information is insufficient for fine-grained similarity estimation and relevance ranking (e.g. for content categorization, personalized filtering, or tag recommendation scenarios).

The support for content sharing tasks can be improved with respect to novel properties of mass social media. In fact, photos and videos produced on state of the art multimedia devices are frequently accompanied with spatial information (e.g. coordinates from integrated GPS receivers). For instance, in the CoPhIR dataset[3], around 4 Million out of 54 Million Flickr images are associated with geographical coordinates. The number of geographically annotated images is supposed to increase in future as more devices will be able to capture the spatial information. In general, we may assume that corresponding resources are produced outdoors, due to physical limitations of GPS localization. For outdoor scenes, location-oriented characterization of content appears reasonable.

Basically, relatedness/similarity of resources or tags can be directly estimated by calculating distances between resource locations (using spatial coordinates) or by measuring overlaps between annotations of resources (using tags, or advanced tag based models like ESA [19]). However, tags or geodata taken alone are often insufficient for reliable con-

WSDM'10, February 4–6, 2010, New York City, New York, USA.

[1] http://www.flickr.com
[2] http://www.youtube.com
[3] http://cophir.isti.cnr.it

tent disambiguation. For instance, in our sample scenario both photos of the London Tower and of the Tower Bridge can originate from the same location (e.g. trip boat on the Thames river) but show two different attractions. On the other hand, views of large buildings and monuments (e.g. the London Eye) can be captured from different locations but show the same object and also have similar annotations. Consequently, we may expect that a combination of both tags and spatial information will allow for better content characterization.

This paper presents our framework GeoFolk (an acronym for **Geo**data in **Folk**sonomies). GeoFolk is designed for Web2.0 content characterization with spatial awareness. Our approach is based on Bayesian statistical models that explicitly describe geodata (spatial coordinates) jointly with tag co-occurrence patterns. From a general perspective, our model can be seen as an extension of Latent Dirichlet Allocation (LDA) introduced in [4]. Beyond LDA-like tag generation process, our model also integrates topic-specific normal distributions which describe location (latitude and longitude). Sampling-based parameter estimation, based on the Monte Carlo Markov Chain method (MCMC), is driven to discover latent topics that simultaneously capture word co-occurrences and spatial locality of characteristic patterns.

The rest of the contribution is organized as follows. Section 2 gives a detailed explanation of the GeoFolk model. Subsequently, we instantiate our model for several application scenarios (content categorization and clustering, tag recommendation) and show results of systematic evaluations with real-life Flickr data in Section 3. In Section 4 we discuss related work. Section 5 summarizes lessons learned and shows directions of our future work.

2. THE GEOFOLK MODEL

For constructing the GeoFolk model, we exploit the idea of Bayesian latent topic models, i.e. mechanisms for discovering low-dimensional, multi-faceted summaries of documents in a probabilistic manner. This section includes the problem formalization for Web 2.0 setting, explains our design choices, and shows possible model instantiations.

2.1 Problem statement

In our model we are given an arbitrary collection $\mathcal{D} = \{d_1..d_D\}$ of D folksonomy resources (e.g. shared photos, videos, etc.). Each resource d in this collection is annotated by some tags $1..N_d$ (we assume $\forall d, N_d \geq 1$). These tags are taken from the vocabulary $\mathcal{V} = \{w_1..w_V\}$ that consists of V different words w_i. Additionally, each resource d is annotated by numeric attributes $lat_d \in \mathbb{R}$ and $lon_d \in \mathbb{R}$ which represent spatial coordinates of the position of its creation. The approach aims to explain observed properties of $\mathcal{D}$ (i.e. tags and coordinates of all $d \in \mathcal{D}$) by the means of a Bayesian model with T latent topics. For the sake of clarity, Table 2.1 summarizes the notation used in following sections.

2.2 Model for tag assignments

The idea of explaining resources by the means of an individual mixture of latent topics (i.e. topics that contributed to their generation) originates in well known Bayesian model coined Latent Dirichlet Allocation (LDA) [4]. In our application scenario, LDA can be directly exploited for explaining assignments of tags taken from $\mathcal{V}$ to resources contained in $\mathcal{D}$. This generative process is summarized as follows. For each resource d, a multinomial distribution θ_d over topics is randomly sampled from a prior Dirichlet distribution with parameter vector $\vec{\alpha}$. To generate each tag w_{d_i}, $(i = 1..N_d)$, a topic z_{d_i} is chosen from this topic distribution, and then a tag w_{d_i} is drawn by a topic-specific multinomial distribution, characterized by $\phi_{z_{d_i}}$. The number N_d of tags attached to the resource d is assumed to be drawn by a Poisson distribution. We also assume that for each topic T, the parameters ϕ_T of the corresponding multinomial distribution over tags are drawn from a prior Dirichlet distribution with parameter vector β.

Symbol	Description
$\mathcal{D}$	Collection of resources (documents)
$\mathcal{V}$	Collection of tags (vocabulary)
D	collection size, i.e. $\lvert\mathcal{D}\rvert$
V	vocabulary size, i.e. $\lvert\mathcal{V}\rvert$
T	number of latent topics in the model
N_d	number of tags assigned to $d \in \mathcal{D}$
lat_d	latitude of d
lon_d	longitude of d
θ_d	distribution of topics specific to resource d
ϕ_z	distribution of tags specific to topic z (multinomial)
ψ_z	joint distribution of coordinates specific to topic z (bivariate normal)
ψ_z^{lat}	separate distributions of latitude/longitude
ψ_z^{lon}	specific to topic z (univariate normal)
z_{d_i}	topic associated with tag i of resource d
w_{d_i}	ith tag in annotation of resource d

Table 1: Summary of GeoFolk notation

1. for $i = 1..T$ do
   - $\phi_T \sim Dirichlet(\beta)$
2. For each $d \in \mathcal{D}$ do
   - $N_d \sim Poisson(\xi)$
   - $\theta_d \sim Dirichlet(\alpha)$
   - for $i = 1..N_d$ do
     - $z_{d,i} \sim Multinomial(\theta_d)$
     - $w_{d,i} \sim Multinomial(\phi_{z_{d,i}})$

Figure 1: Generative process for tag assignments.

The corresponding graphical representation for this tag assignment model is shown in Figure 2, and the generative algorithm (using the common Bayesian pseudocode notion) in Figure 1.

Given a collection of annotated resources $\mathcal{D}$, the learning algorithm estimates parameters of distributions θ_d and ϕ_z that provide a good fit with observed data (i.e. explain the generation of $\mathcal{D}$ with high probability). In contrast, the number of latent topics T and parametrization of the Dirichlet distributions α and β are meta parameters, i.e. they must be specified by model designer and may require empirical tuning.

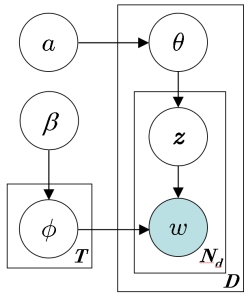

Figure 2: Graphical model for tag assignments.

2.3 GeoFolk Model for spatial information

In our GeoFolk approach, topic discovery is additionally affected by spatial information, which is considered jointly with tag co-occurrences. The key objective is to provide a joint explanation for resource annotation and its coordinates by a mixture of latent topics. However, spatial coordinates are usually assigned to resources only once. Therefore, they cannot be naturally explained by a mixture of topics like tag assignments. To achieve the desired functionality, the model demands certain conceptual adaptation.

The suggested generative process for annotated resources with assigned coordinates can be summarized as follows. For each resource d, a multinomial distribution θ_d over topics is randomly sampled from a prior Dirichlet distribution with parameter vector $\vec{\alpha}$. To generate each tag $w_{d_i}, i = 1..N_d$, a topic z_{d_i} is chosen from this topic distribution, and then a tag w_{d_i} is drawn by a topic-specific multinomial distribution with parameters $\phi_{z_{d_i}}$. In parallel, the topic generates two coordinates (i.e. geostamps), lat_{d_i} and lon_{d_i} from two topic-specific Gaussian distributions ψ_z^{lat}, ψ_z^{lon}. The number N_d of tags attached to the resource d is assumed to be drawn by a Poisson distribution. Gaussian parameters μ_z^{lat} and μ_z^{lon} (i.e. means for topic-specific latitude and longitude) are assumed to be drawn from a certain interval (which represents coordinate ranges in observed data) using independent uniform distributions γ_{lat} and γ_{lon}. Variances of all Gaussian distributions are assumed to be drawn from a certain interval using an independent uniform distribution.

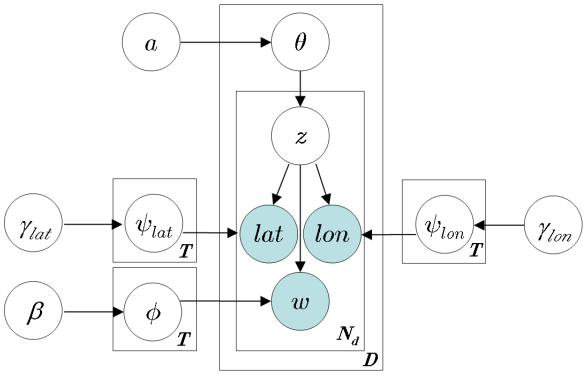

Figure 3: Graphical model for GeoFolk.

The generative process of GeoFolk consists of steps shown in Figure 4, the corresponding graphical representation is shown in Figure 3. The presented solution can be seen as a modeling compromise, in order to capture correlations between coordinates and tags by a mixture of latent topics. In doing so, the GeoFolk model actually describes data in which coordinates are associated with each tag. When fitting our model from real data, coordinates of each training resource are attached to all tags of this resource. However, after fitting, GeoFolk runs as a common generative model. In other words, this process would generate different coordinates for the tags attached to the same resource. This kind of model deficiency is common for other similar LDA extensions, such as topics over time [17]. The actual modeling problem arises with ambiguous prediction of coordinates for new, previously unseen, resources. However, this aspect is of less practical importance for GeoFolk. In our Web 2.0 application scenario, we usually may assume that spatial coordinates of the resources are known (and there is no need to predict them). Practically relevant fields of model application, such as predictions of tags and similarity estimation between resources, are not affected by this issue at all.

We note that the behavior of latent topics in GeoFolk is different from LDA. Since topic-specific location is described by normal distributions, the particular topic is 'active' in a certain spatial area. This (intended) behavior establishes relationships between tag distributions and locations of interest. If the same pattern of frequent tags appears at different places, the resulting model will contain multiple latent topics with similar multinomial distributions over tags but different normal distributions for locations.

1. for $i = 1..T$ do
   - $\phi_T \sim Dirichlet(\beta)$
   - $\psi_T^{lat} < -Normal(Uniform(lat_{min}, lat_{max}), Uniform(var_{min}, var_{max}))$
   - $\psi_T^{lon} < -Normal(Uniform(lon_{min}, lon_{max}), Uniform(var_{min}, var_{max}))$
2. For each $d \in \mathcal{D}$ do
   - $N_d \sim Poisson(\xi)$
   - $\theta_d \sim Dirichlet(\alpha)$
   - for $i = 1..N_d$ do
     - $z_{d,i} \sim Multinomial(\theta_d)$
     - $w_{d,i} \sim Multinomial(\phi_{z_{d,i}})$
     - $lat_{d,i} \sim \psi_{z_{d,i}}^{lat}$
     - $lon_{d,i} \sim \psi_{z_{d,i}}^{lon}$

Figure 4: Generative process of GeoFolk.

The GeoFolk model presented so far combines discrete and continuous distributions in a non-trivial manner. The resulting model is complex and does not allow for exact inference and parameter estimation. For this reason, GeoFolk performs approximate inference by Gibbs sampling, using the Monte Carlo Markov Chain method (MCMC) [2].

Spatial information is continuous by its nature. Coordinate assignments produced by topic z can be naturally modeled by a bivariate normal distribution ψ_z. As an alternative, one may also consider two Gaussian distributions ψ_z^{lat} and ψ_z^{lon} (i.e. separately for latitude and longitude). Our earlier experiments were based on a joint distribution ψ_z, which turned out to be a practical bottleneck, due to poor convergency of parameter estimates. For all the results in this paper we employ the (simpler) second choice and consider latitude and longitude as two separate continuous variables.

As a possible design alternative, one may consider discretization of coordinates. However, the choice of appropriate grid sizes is highly application- and data-specific and may cause additional modeling problems. For instance, the constant grid size can appear too small for some regions (for instance, London vs. Madrid, in our sample scenario) and too large for others (say Westminster vs. Houses of Parliament in London). As a result of granularity problems, the model may miss spatially compact but meaningful topics. From the computational perspective, sparsity of discrete spatial data may also cause fitting problems for Gibbs sampling with MCMC. For these pragmatic reasons, GeoFolk avoids discretization by associating with each topic continuous distributions over coordinates.

2.4 GeoFolk applications

From the application point of view, the set of latent topics can be used to answer various questions about the similarity of resources and tags. In our model interpretation, two resources are similar to the extent that their generation is explained by same topics. Analogously, two tags are similar to the extent that they are generated by same topics.

The distance between probability distributions (specially resource-specific multinomial distributions θ_{d_i}, in our case) can be estimated in a variety of ways [11]. A standard measure for estimating divergence between two distributions is the Kullback-Leibler divergence (KL). Given two resources d_x and d_y with corresponding distributions over topics θ_{d_x} and θ_{d_y}, the KL divergence is defined as follows:

$$KL(\theta_{d_x}, \theta_{d_y}) = \sum_{i=1}^{T} \theta_{d_x,i} \cdot \frac{\theta_{d_x,i}}{\theta_{d_y,i}} \quad (1)$$

To obtain a symmetric distance measure, it is convenient to consider the symmetric variant of KL divergence:

$$KL^{sym}(\theta_{d_x}, \theta_{d_y}) = \frac{1}{2} \cdot (KL(\theta_{d_x}, \theta_{d_y}) + KL(\theta_{d_y}, \theta_{d_x})) \quad (2)$$

As an alternative, we may also consider the symmetrized and smoothed Jensen-Shannon (JS) divergence [10, 11], also known as information radius. This measure estimates dissimilarity between θ_{d_x} and θ_{d_y} through the average of both distributions:

$$JS(\theta_{d_x}, \theta_{d_y}) = \quad (3)$$
$$\frac{1}{2} \cdot \left(KL\left(\theta_{d_x}, \frac{\theta_{d_x} + \theta_{d_y}}{2}\right) + KL\left(\theta_{d_y}, \frac{\theta_{d_x} + \theta_{d_y}}{2}\right)\right)$$

Since the square root of (3) is a proper metric [7] (i.e. it satisfies the usual axioms of nonnegativity, identity of indiscernables and triangle inequality), $\sqrt{JS(\theta_{d_x}, \theta_{d_y})}$ can be considered as a distance measure in the common IR sense.

Of course, it is also possible to consider topic distributions as 'just' numerical, L_1-normalized vectors and to apply simple geometric similarity functions like dot product or cosine measure [12]. Our evaluation has shown that

The distance measures introduced so far can be exploited for a variety of realistic scenarios for social media, including content organization, search, and tag/content recommendation.

Content organization.

In particular, the topic based feature space can be exploited in a straightforward manner for content structuring (e.g. clustering of photos). Analogously, supervised learning methods (classification) can be applied for content categorization and filtering. In our sample scenario, one may think of automatic clustering of photos from the last trip to London, combined with automatic annotation of clusters by most characteristic tags from Flickr for cluster coordinates.

Keyword-based search and ranking.

A keyword-based query $q = \{w_{q1}..w_{qp}\}$ can be treated as a new annotated resource d_q with tags $w_{q1}..w_{qp}$. Consequently, d_q can easily be transformed into the topic-based feature space of the tag assignment model. The required parameter estimation for θ_{d_q} is done by Gibbs sampling, with previously learned (and now fixed) tag-topic distributions ϕ_z for $z = 1..T$. As a result, we obtain for the query q its distribution θ_{d_q} over latent topics, which can be used for search and ranked retrieval of resources in the topical feature space. As an example, we consider the query '*london eye*' for our sample scenario. Query transformation would allow for finding resources annotated by semantically coherent tags (say '*londoneye*', '*millennium-wheel*', etc.) and also filtering out thematically irrelevant matches that show significantly different distribution over latent topics (say '*golden eye*').

Keyword-based search with spatial awareness.

Analogously, a keyword-based query $q = \{w_{q1}..w_{qp}\}$ with spatial preferences lat_q and lon_q can be transformed into the topical feature space of the GeoFolk model. The required parameter estimation for θ_{d_q} is done by Gibbs sampling, with previously learned (and then fixed) tag-topic distributions ϕ_z and spatial distributions ψ_z^{lat} and ψ_z^{lon} for all topics $z = 1..T$. In our sample scenario, the search for '*piccadilly*' may be combined with coordinates of the London city centre. This would help to filter out identically annotated but irrelevant photos of the Manchester Piccadilly train station.

Suggesting locations for queries.

An interesting special case of retrieval with GeoFolk is the prediction of coordinates for keyword-based queries. For a keyword-based query $q = \{w_{q1}..w_{qp}\}$, its distribution over topics θ_{d_q} can be estimated together with most likely coordinates through Gibbs sampling, when the GeoFolk model is not conditioned on fixed values for lat_q and lon_q. This feature can be exploited for assisting the user in navigation through locations, e.g. by displaying the map of the predicted location together with relevant matches identified in the nearest vicinity of this position. In our sample scenario, the search for '*tower*' may display the map with close view on the London Tower and the Tower Bridge, with previews of most relevant photos for both attractions.

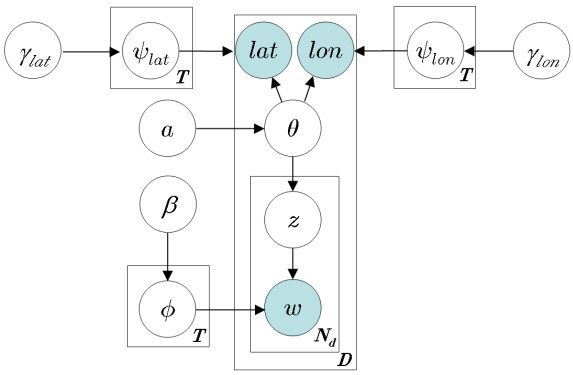

Figure 5: Graphical model for suggesting query locations.

As discussed in Section 2.3, the GeoFolk prediction of locations is ambiguous for queries that consist of more than one keyword. An alternative generative process description of GeoFolk (better suited to explain such queries) is one in which a single pair of values for lat_q and lon_q is generated for q. The graphical model for this alternative is shown in Figure 5. The desired behavior can be achieved with GeoFolk by importance sampling, from a mixture of per-topic Gaussian distributions, with mixtures weight as the per-resource θ_d over topics. In this case, this distribution of coordinates remains parameterized by the set of coordinate-generating Gaussian distributions, but the visible model outcome allows for easier practical interpretation.

Tag recommendation and exploration.

From a different perspective, per-topic tag distributions ϕ_z (i.e. topic-specific multinomial distributions that represent tag generation probabilities) can be used for constructing per-tag feature vectors. In fact, generation probabilities for a particular tag w across topics $\phi_z(w)$ ($z \in 1..T$) can be seen as features of w in a new vector space, which is constructed over latent topics $1..T$ as new dimensions. We note that resulting feature vectors do *not* represent probability distributions and additional L_1-normalization may be necessary. Tag-specific vectors in this space can be exploited in the usual IR-like manner for tag recommendation (based on similarity estimation between tags) and for computation of semantically coherent tag clouds (by classification or clustering of tag vectors).

3. EVALUATION

The key objective of our GeoFolk evaluation is to demonstrate the usefulness of the combination between text and spatial knowledge for understanding of social media. The evaluation strategy includes two directions. On one hand, we analyze statistical properties of the GeoFolk model (such as model fit and model complexity) under different settings. On the other hand, our aim is to demonstrate the practical viability of the introduced modeling approach. Consequently, we instantiate our application oriented evaluation with three realistic scenarios for social media: content classification, content clustering, and tag recommendation. The evaluation is done with real data gathered from Flickr.

In line with our preceding argumentation, we restricted the scope of the evaluation dataset to one particular spatial area (districts of London, analogously to the run-through examples in this paper). From the application point of view, this scenario is especially hard to handle regarding categorization of content and tag recommendation, due to substantial overlap of tag annotations and close proximity of considered locations. From the technical point of view, restriction of the model scope allows for satisfactory scalability, in order to perform systematic series of experiments in reasonable time. From the pragmatic evaluation point of view, the choice of a famous location allows for better content understanding by human experts and, subsequently, high inter-rater agreement.

In our comparative evaluations we consider several modeling approaches for social media. The choice of discussed methods is also motivated by our top-level evaluation objective: analysis of the combined model in comparison with baseline methods that exploit particular aspects of social media (tags, coordinates) in an isolated manner. In particular, the GeoFolk model is primarily compared with LDA-like tag assignment model from Section 2.2. Wherever applicable, we also consider two simple baseline models that represent spatial coordinates and tags without any complex transformations.

3.1 Testbed

The evaluation infrastructure for GeoFolk was implemented as a Java 1.6 framework. For performing approximate inference by Gibbs sampling (using the Monte Carlo Markov Chain simulations), the JAGS framework[4] was used. Results of JAGS simulations (i.e. estimated model parameters) were processed in the statistical computing framework R [5] and finally stored in an Oracle 11g database instance for subsequent application-oriented experiments. The majority of experiments was performed on a lightweight server with 4 Gb main memory and four 3.20 Ghz Xeon cores.

The components of GeoFolk evaluation framework (sample data, BUGS scripts, Java framework implementation) are available as public domain open source implementation[6]. The supplied lists of used Flickr resources can be used for downloading and compiling evaluation datasets.

3.2 Experimental dataset

In our experiments we used parts of the publicly accessible CoPhIR dataset[7] that contains metadata for over 54 Millions of Flickr resources (as observed in 2009). The dataset includes resource descriptors, tag assignments, extracted low-level features, spatial attributes, and other meta knowledge for social media. For evaluating GeoFolk, we were primarily interested in tag assignments and spatial attributes of resources and did not consider further aspects (e.g. low-level features) for modeling. By restricting the spatial scope to locations in London, we obtained a dataset of 28,770 resources with attached spatial metadata (GPS latitude and longitude). The dataset has very heterogeneous tag vocabulary (17,676 distinct tags, 165,000 tag assignments in total). Distributions of tag frequencies and tag counts per resource are shown in Figures 6 and 7.

[4] `http://www-ice.iarc.fr/~martyn/software/jags/`
[5] `http://www.r-project.org/`
[6] `http://isweb.uni-koblenz.de/Research/DataSets/bayes`
[7] `http://cophir.isti.cnr.it`

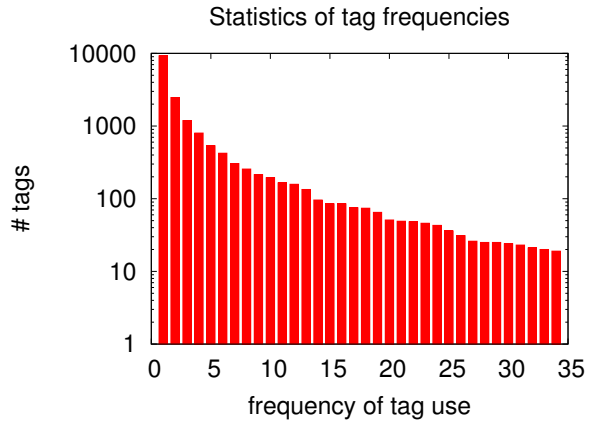

Figure 6: Distribution of tag frequencies in sample Flickr dataset.

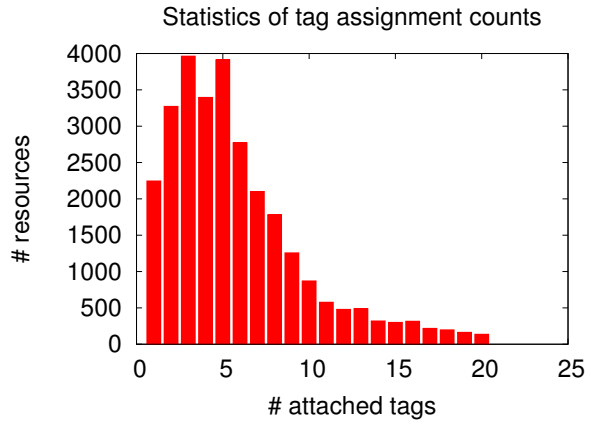

Figure 7: Distribution of tag counts for resources in sample Flickr dataset.

For application-oriented evaluation, 4,100 resources (i.e. Flickr photos) were manually inspected by human experts and associated with one of 28 famous locations in London (such as Westminster, Marble Arch, etc.). This labeling with spatial awareness was then exploited as a gold standard for our resource categorization experiments. Table 2 shows the summary of the resulting evaluation dataset. From the application perspective, categorization of this content is a hard issue. As indicated in Table 2, the evaluation dataset shows a close proximity between particular locations. Furthermore, the tag annotations of particular resources are quite heterogeneous and sparse. In fact, the vocabulary of the evaluation dataset consists of 3,540 distinct tags. However, only 2,200 of them can be identified as relevant for characterizing chosen locations. To make this observation, we simply exploited the fact that almost all locations shown in Table 2 have own articles on Wikipedia, and estimated the string overlap between tags (like *millenniumwheel* or *towercranes*, etc) and words from corresponding Wikipedia sources.

London location	# of Flickr resources	centr. latitude	centr. longitude
Leicester Square	92	51.512	-0.129
Westminster	493	51.500	-0.127
Lambeth North	25	51.496	-0.117
Blackfriars	403	51.510	-0.102
Whitehall	510	51.502	-0.124
Covent Garden	142	51.516	-0.128
Hyde Park Corner	284	51.502	-0.144
The Aldwych	63	51.512	-0.118
South Bank	208	51.506	-0.115
Fitzrovia	66	51.518	-0.138
Soho	40	51.513	-0.134
Piccadilly	21	51.502	-0.140
Farringdon	29	51.518	-0.105
Piccadilly Circus	171	51.510	-0.135
St James	36	51.507	-0.134
Charing Cross	172	51.508	-0.127
Embankment	275	51.504	-0.120
Trafalgar Square	224	51.508	-0.128
Victoria	116	51.497	-0.141
Lincolns Inn	26	51.517	-0.114
The West End	125	51.508	-0.143
London Wall	25	51.518	-0.094
Temple	93	51.512	-0.113
Cheapside	274	51.514	-0.098
Strand	74	51.511	-0.122
Bankside	55	51.508	-0.096
Waterloo	35	51.503	-0.109
Chinatown	29	51.512	-0.132

Table 2: Structure of London evaluation dataset

3.3 Model characteristics

In our experiments, we compared a number of practically relevant model properties and settings. All models discussed in this section were defined by scripts in the JAGS framework, and instantiated with data of the entire reference Flickr dataset described in previous section. The approximate inference was performed by Gibbs sampling (using the Monte Carlo Markov Chain simulations). By convention, the MCMC output is divided into two parts: an initial burn-in period, which is discarded (at least 1000 iterations in each experiment), and the remainder of the run (10,000 iterations in each experiment), in which the output is considered to have converged to the target distribution. Samples from the second part are used to create approximate summary statistics.

In many practical cases, topic models are quite sensitive to the choice of hyper-parameters, i.e. in our case T (number of latent topics) as well as α and β (settings for Dirichlet distributions). In the context of our GeoFolk target application scenario, we observed that the sensitivity to hyper-parameters was not very strong. We used symmetric Dirichlet distributions with $\alpha = 50/T$ and $\beta = 0.1$) in all presented experiments.

The complexity of data (i.e. tag assignments alone vs. tags with coordinates) has direct influence on the model complexity, main memory consumption, and computational overhead for model convergency. The complexity of each model also increases with growing number of latent topics. Table 3 shows practical statistics for models learned in our experiments. Models with larger number of latent topics require rapidly growing processing times but add no principally new value to the observed behavior. The numbers for

model property	LDA 5 topics	LDA 10 topics	LDA 15 topics
stochastic nodes	34,193	34,201	34,213
training time, min	75	192	275
memory, Mb (raw)	126	214	311
model property	GeoFolk 5 topics	GeoFolk 10 topics	GeoFolk 15 topics
stochastic nodes	90,455	90,479	90,515
training time, min	244	372	520
memory, Mb (raw)	397	461	612

Table 3: Properties of evaluation models

observed memory consumption should be seen as raw estimates, since the simulation software incrementally allocates memory in blocks and also applies a number of optimizations.

A natural way to compare Bayesian models is to consider trade-offs between the fit of the data to the model, and the corresponding model complexity. For our models presented so far, we consider the Deviance Information Criterion (DIC) as proposed in [3]. The criterion $DIC = \overline{D} + p_D$ combines the model deviance $D(\vartheta)$ and the model complexity p_D. Models with smaller DIC are better supported by the data. The model deviance is defined as

$$D(\vartheta) = -2logL(data|\vartheta) \quad (4)$$

where ϑ are stochastic parameters of the model. Deviance reflects the model fit for observed data; we consider the total deviance by summing up deviance for all stochastic nodes in the model. The model complexity is measured by estimate of the 'effective number of parameters', given by posterior mean deviance minus deviance evaluated at the posterior mean of the parameters:

$$P_D = \overline{D} - D(\overline{\vartheta}) \quad (5)$$

Analogously, the total complexity is estimated by summing up complexity for all stochastic nodes in the model. In our experiments, complexity was estimated across 5 chains for each learned model. Figure 8 shows the collected DIC values for models in our evaluation. Since LDA and GeoFolk explain different data (without/with consideration of spatial information) their values cannot be directly compared. However, it can be observed that the GeoFolk model with larger number of topics is significantly better supported by the data than the model with few topics. For the LDA model, such significant improvement cannot be observed. We assume that this behavior is caused by the high sparsity of the tag data in our Flickr collection.

3.4 Categorization of resources: classification and clustering

The objective of our classification experiments was to evaluate the ability of presented methods for content categorization in two settings: supervised learning (with explicitly given training data) and unsupervised learning (without apriori available training data). The comparisons include 4 representational models for Flickr data: GeoFolk, LDA like tag assignment model, simple tag based resource representation, and resource coordinates. Conceptually, all mentioned resource representations use the vector space model. However, the way of constructing resource-specific features is

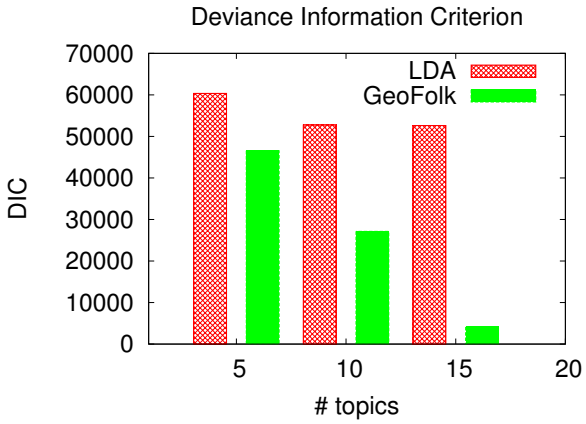

Figure 8: Deviance Information Criterion (DIC) of LDA and GeoFolk models.

quite different. GeoFolk characterizes resources by probabilistic distributions over latent topics, jointly learned for tags and spatial data. Similarly, LDA-like tag assignment method constructs the latent topic model for tags alone. The distance between feature vectors of two resources in these models was estimated by square root of the Jensen-Shannon (JS) divergence (3). As an alternative, simple geometric similarity measures (e.g. cosine measure) can be used as well; notably, at the confidence level 0.95 the variation of results by applying different similarity/distance measures was not statistically significant in our experiments. The tag based resource representation uses the term-based vector space model. Consequently, similarity between two resources is estimated by common cosine similarity between $tf \cdot idf$ weighted feature vectors [12] (in our setting, the tf value is binary - either 1 or 0, and indicates just the presence or absence of the tag in resource annotation). For coordinates, spatial distances between geographic points were considered as distance measure. Spatial distances were approximated by Euclidean distances in two-dimensional coordinate space; due to the relatively small size of the considered area, the three-dimensional nature of the geosphere was neglected. For evaluations, we employed simple baseline categorization methods (kNN for classification, k-means for clustering) which can be directly used with all discussed data models.

Classification.

For the purpose of this evaluation, categories from our Flickr evaluation dataset (Section 3.2) were treated as classes. We kept 50% of resources as training data for modeling the classifier. The remainder was considered as unlabeled test data and passed to the classifier. Our quality measure is the fraction of correctly classified resources (accuracy) among all classes. In all classification experiments, we considered as our baseline method the kNN (k nearest neighbor) classifier with $k = 10$. kNN assigns the majority class of the k nearest neighbors (in the custom sense of distances for each particular method) to a test resource. The method requires no explicit training and can easily be implemented for all considered data models. Subsequently, we computed micro-averaged results for 50 independently performed ex-

periments (with random partitioning of resources into training/test sets). Table 4 summarizes results for baseline methods in comparison with GeoFolk and LDA with $T = 10$ latent topics. All methods provide substantially better accuracy than random assignments (which would lead, for our dataset with 48 topics, to a calculational average accuracy around 0.02). However, it can be observed that neither simple tag representation nor spatial coordinates, taken alone, are sufficient for reliable content categorization. This can be explained by extreme sparsity and diversity of tag annotations and tight proximity of target locations. LDA provides necessary smoothing/stabilization over the tag space and improves the classification accuracy. The joint consideration of text and spatial factors in GeoFolk helps to further increase the robustness and accuracy of classification decisions. At the standard confidence level 0.95, the difference between results for GeoFolk and for baseline methods is statistically significant.

Model	avg(accuracy)
GeoFolk	0.421
LDA	0.374
Tags	0.282
Coordinates	0.187

Table 4: Accuracy of supervised resource categorization (classification)

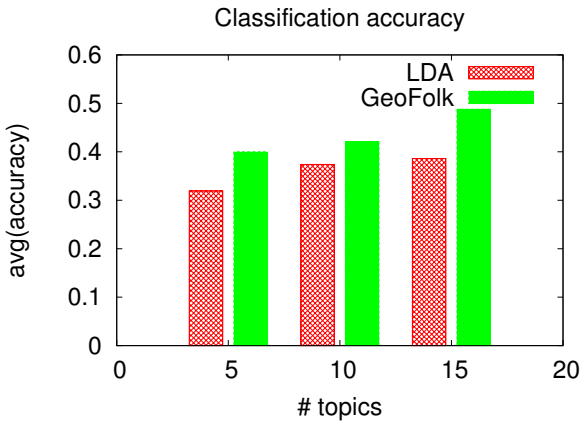

Figure 9: Classification accuracy with LDA and with GeoFolk.

Figure 9 shows evaluation results for GeoFolk and LDA with different numbers of latent topics. As previously observed in Section 3.3, increasing the number of latent topics does not necessarily improves the model quality for LDA, due to the lower information content of tag annotations. In contrast, the richer multimodal GeoFolk data allows for better results with more complex models. However, the price for this is a higher computational overhead, as previously discussed in Section 3.3.

Clustering.

The resource collections and topic labels from Flickr evaluation dataset were also used to evaluate unsupervised clustering. For categorization of resources, we employed the common k-means algorithm [12]. Unlike classification results, the clusters do not have explicit topic labels. Basically, various cluster-class assignments are possible. This makes the accuracy estimation ambiguous and interpretation dependent. For this reason, the notion of clustering accuracy is slightly adjusted. We calculate the clustering accuracy (i.e. the fraction of correctly assigned resources) for the 'optimal' cluster-class interpretation, which maximizes the overall overlap between clusters and classes. Let k be the number of classes and clusters (28 in our case), N_i the total number of clustered resources in $class_i$, N_{ij} the number of resources contained in $class_i$ and having cluster label j. We define the clustering accuracy as follows:

$$accuracy = \max_{(j_1,\ldots,j_k)\in perm((1,\ldots,k))} \frac{\sum_{i=1}^{k} N_{i,j_i}}{\sum_{i=1}^{k} N_i} \qquad (6)$$

Model	avg(accuracy)
GeoFolk	0.328
LDA	0.255
Tags	0.117
Coordinates	0.102

Table 5: Accuracy of unsupervised resource categorization (clustering)

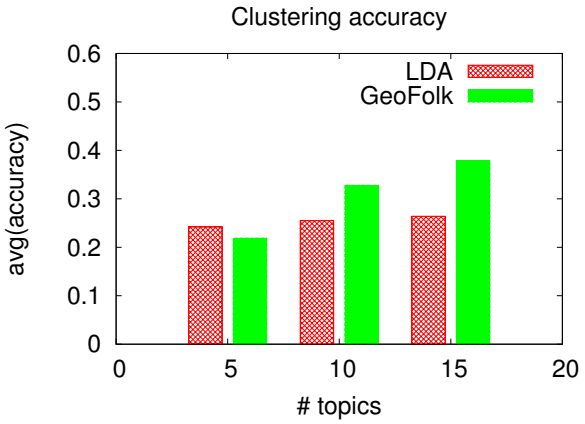

Figure 10: Clustering accuracy with LDA and with GeoFolk.

Table 5 summarizes results for baseline methods in comparison with GeoFolk and LDA with $T = 10$ latent topics. Figure 10 shows evaluation results for GeoFolk and LDA with different numbers of latent topics. The observations are similar to the ones discussed for the supervised case. At the standard confidence level 0.95, the difference between results for GeoFolk and for baseline methods is also statistically significant.

3.5 Tag recommendation

In the tag recommendation scenario, we evaluated the ability of our methods to recognize semantic relationships between tags and to capture semantically coherent tag clouds. As discussed in Section 2.4, tag probabilities in latent topics

can be exploited for constructing tag-characteristic feature vectors. Subsequently, we can estimate similarity between tags. Since tag probabilities across topics do not form probability distributions, we employed cosine similarity measure for this experiment. In our experiments we randomly removed by one some tags from resource annotations and tested the ability of our topical models to reconstruct the missing by producing a similarity-based ranked list of suggestions. Figure 6 shows the mean reciprocal rank for the position of the desired tag in the list of recommended tags. Mean reciprocal rank (MRR) is defined as average of the reciprocal ranks of tag positions for a series of Q tag reconstruction experiments as follows:

$$MRR(Q) = \frac{1}{|Q|} \sum_i \frac{1}{rank_i} \quad (7)$$

Model	MRR
GeoFolk	0.212
LDA	0.119
Tags	0.073
Coordinates	0.027

Table 6: Accuracy of tag recommendation

where $rank_i$ indicates the position of the desired tag in the ranked list of recommendations produced in the ith experiment (lower MRR values indicate lower reconstruction quality). Since the relevance judgement for other tags in the particular list is not known, the result of this experiment indicates the *lower* bound of method accuracy and thus should not be interpreted as a full-fledged evaluation of the tag recommendation scenario. Nevertheless, it can be used as an indication for the ability of methods to identify latent relationships between tags. From this perspective, results from Table 6 indicate the suitable ability of GeoFolk to cope with semantically coherent tags for real tag recommendation scenarios. At the standard confidence level 0.95, the difference between results for GeoFolk and for baseline methods was statistically significant.

4. RELATED WORK

Shared content in Web 2.0 folksonomies is quite different from common collections of text documents or Web pages. In general, the low number of tag assignments to resources makes the available data in folksonomies very sparse. The enrichment of models by simultaneously considering multiple content aspects (tags, spatial knowledge, timestamps, low-level features of shared content, etc.) helps to identify semantically coherent clouds of features.

Recent contributions [5, 9, 15] present different methods for finding semantic relationships between tags. The target application is usually tag recommendation [9, 15] or semantic analysis of tags and tag clouds [5]. Extraction of structured vocabularies from folksonomies is discussed in [13, 14]. Tag recommendation using external data sources (e.g. resource content, anchor text, Wikipedia articles etc.) is discussed in [8, 18]. In contrast to these approaches, GeoFolk aims to explain semantic relatedness of tags by the means of latent topics in a probabilistic Bayesian framework. In our model, tag similarity/relatedness is estimated in a natural manner, by comparing tag distributions over latent topics.

The work of Zhou *et al.* on Explicit Semantic Models (ESA) [19] can be seen as an alternative to latent semantic models discussed in this paper. ESA exploits the semantics of structured large-scale document collections (such as Wikipedia) and constructs the feature space over collection documents, treated as 'explicit' topics. In contrast, our objective with GeoFolk is to exploit the natural multi-modality of folksonomy data, rather than enrichment of the tag-based feature space. Conceptually, GeoFolk and ESA are designed for very different data and thus hard to compare. Nevertheless, it deems reasonable to do in the future comparative evaluations of both methods through shared applications (such as content classification or tag recommendation).

In the field of latent semantic analysis, [1] can be seen as a close competitor for the tag assignment model of GeoFolk. This paper reported on use of SVD-based Latent Semantic Indexing (LSI) [6] for dimensionality reduction in tag-based search (with observed scalability problems). In contrast, our approach exploits the notion of probabilistic Bayesian models which show better scalability and can be easily customized for multi-modal data analysis (as done in GeoFolk for text annotations together with spatial knowledge).

In the area of Bayesian learning, several recent models (at least partly related to and relevant for our Web 2.0 setting) have associated the generation of additional modalities with topics. For example, the Group-Topic (GT) model [16] conditions on topics for both word generation and relational links. Analogously, the Topics Over Time (TOT) model [17] aims to exploit relationships between text and document timestamps. Similarly to GeoFolk, results of GT and TOT evaluations show that jointly modeling an additional modality improves the relevance of the discovered topics.

5. CONCLUSION & FUTURE WORK

This paper has presented GeoFolk, a Bayesian model for Web 2.0 social media that jointly models both tag co-occurrences and spatial attributes of shared resources. Results on real-world Flickr data show the overall viability of the presented approach and improvements in realistic application scenarios, such as content classification, clustering, and tag recommendation. From the practical point of view, the GeoFolk model can be optimally used in quality-oriented applications, as it helps to achieve higher result quality at the price of additional computational overhead for model fitting (parameter estimation through comprehensive sampling).

The relative simplicity of our GeoFolk model allows for several extensions in order to better address the multi-modal nature of social media in Web 2.0 applications. Possible extensions are integration of temporal knowledge (content timestamps), low-level features of shared content, or authorship information. On the other hand, the GeoFolk framework can be easily integrated with models for Web 2.0 social relationships (e.g. for Flickr groups). As a result, we may obtain personalized and thematically focused solutions for content management (recommendation, categorization, filtering) with spatial and social awareness.

Acknowledgements. This work was supported by EU Integrated Project WeKnowIt, 7th Framework Programme: Information and Communication Technologies, grant number ICT-215453.

6. REFERENCES

[1] Rabeeh Abbasi and Steffen Staab. Introducing Triple Play for Improved Resource Retrieval in Collaborative Tagging Systems. In *Proceedings of Exploiting Semantic Annotations in Information Retrieval (ESAIR), workshop at ECIR'08*, March 2008.

[2] C. Andrieu, N. Freitas, A. Doucet, and M. Jordan. An Introduction to MCMC for Machine Learning. *Machine Learning*, 50:5–43, 2003.

[3] Christopher M. Bishop, David J. Spiegelhalter, and John M. Winn. VIBES: A Variational Inference Engine for Bayesian Networks. In *Advances in Neural Information Processing Systems (NIPS), Vancouver, Canada*, pages 777–784, 2002.

[4] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. 3:993–1022, 2003.

[5] Ciro Cattuto, Dominik Benz, Andreas Hotho, and Gerd Stumme. Semantic Grounding of Tag Relatedness in Social Bookmarking Systems. *The Semantic Web - ISWC 2008*, pages 615–631, 2008.

[6] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[7] B. Fuglede and F. Topsoe. Jensen-shannon divergence and hilbert space embedding. *Internationales Symposium on Information Theory*, pages 31–39, 2004.

[8] Paul Heymann, Daniel Ramage, and Hector Garcia-Molina. Social Tag Prediction. In *31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 531–538, New York, NY, USA, 2008. ACM.

[9] R. Jaschke, L. Marinho, A. Hotho, L. Schmidt-Thieme, and G. Stumme. Tag Recommendations in Folksonomies. *Lecture Notes in Computer Science*, 4702:506, 2007.

[10] T. Landauer, D. McNacamara, S. Dennis, and W. Kintsch. *Handbook of Latent semantic Analysis*. Lawrence Erlbaum Associates, 2007.

[11] J. Lin. Divergence Measures based on Shannon Entropy. *IEEE Transactions on information Theory*, 37(14):145–151.

[12] Manning,C.D. and Raghavan,P. and Schütze,H. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[13] Peter Mika. Ontologies are us: A unified model of social networks and semantics. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(1):5–15, 2007.

[14] Patrick Schmitz. Inducing Ontology from Flickr Tags. In *Proceedings of Collaborative Web Tagging, Workshop at WWW'06*, 2006.

[15] Börkur Sigurbjörnsson and Roelof van Zwol. Flickr tag recommendation based on collective knowledge. In *17th international conference on World Wide Web (WWW)*, pages 327–336, New York, NY, USA, 2008. ACM.

[16] X Wang, N. Mohanty, and A. McCallum. Group and topic discovery from relations and text. *11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Workshop on Link Discovery*, pages 28–35, 2005.

[17] Xuerui Wang and Andrew McCallum. Topics Over Time: a Non-Markov Continuous-Time Model of Topical Trends. In *12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), Philadelphia, USA*, pages 424–433, 2006.

[18] Leong Chee Wee and S. Hassan. Exploiting Wikipedia for Directional Inferential Text Similarity. *Information Technology: New Generations, 2008. ITNG 2008. Fifth International Conference on*, pages 686–691, April 2008.

[19] Ding Zhou, Jiang Bian, Shuyi Zheng, Hongyuan Zha, and C. Lee Giles. Exploring Social Annotations for Information Retrieval. In *17th international conference on World Wide Web (WWW)*, pages 715–724, New York, NY, USA, 2008. ACM.