Generative Topic Modeling in Image Data Mining and Bioinformatics Studies

A Thesis

Submitted to the Faculty

of

Drexel University

by

Xin Chen

in partial fulfillment of the

requirements for the degree

of

Doctor of Philosophy

December 2012

Table of Contents

List of Tables

List of Figures

Abstract

Generative Topic Modeling in Image Data Mining and Bioinformatics Studies
Xin Chen
Xiaohua Tony Hu, Supervisor, Ph. D.

Probabilistic topic models have been developed for applications in various domains such as text mining, information retrieval and computer vision and bioinformatics domain. In this thesis, we focus on developing novel probabilistic topic models for image mining and bioinformatics studies. Specifically, a probabilistic topic-connection (PTC) model is proposed for co-existing image features and annotations, in which new latent variables are introduced to allow for more flexible sampling of word topics and visual topics. A perspective hierarchical Dirichlet process (pHDP) model is proposed to deal with user-tagged image modeling, associating image features with image tags and incorporating the user's perspectives into the image tag generation process. It's also shown that in mining large scale text corpora of natural language descriptions, the relation between semantic visual attributes and object categories can be encoded as Must-Links and Cannot-Links, which can be represented by Dirichlet-Forest prior. Novel generative topic models are also introduced to meta-genomics studies. The experimental results show that the generative topic model can be used to model the taxon abundance information obtained by the homology-based approach and study the microbial core. It also shows that latent topic modeling can be used to characterize core and distributed genes within a species and to correlate similarities between genes and their functions. A further study on the functional elements derived from the non-redundant CDs catalogue shows that the configuration of functional groups encoded in the gene-expression data of meta-genome samples can be inferred by applying probabilistic topic modeling to functional elements. Furthermore, an extended HDP model is introduced to infer functional basis from detected enterotypes. The latent topics estimated from human gut microbial samples are evidenced by the recent discoveries in fecal microbiota study, which demonstrate the effectiveness of the proposed models.

1. INTRODUCTION

Probabilistic topic models have been developed for applications in various domains such as text mining [95], information retrieval [15]and computer vision [2], [93]. In bioinformatics domain, generative topic model has been previously used to learn protein-protein relations from MEDLINE abstracts of biomedical literatures [9], [110]; it has also been applied to identify gene relations from microarray profiles [36]; the generative topic model is also used to describe the process of constructing mRNA module collections [40].In this thesis, we will focus on developing novel probabilistic topic models for image mining and bioinformatics studies.

The prevalence of digital imaging device, such as digital camera and digital video camera, has generated an increasingly large amount of unlabeled multimedia data, especially unlabeled image data. With nearly a million new images being added in a single day, the Flickr.com, one of the most popular photo sharing websites, now hosts over 3 billion shots of user-uploaded images. Manually annotating such a huge amount of image data is time-consuming, laborious and prohibitively expensive.To face the challenge of enormous explosion of unlabeled online image resources, it is very important to develop context-sensitive robust automatic image annotation system.

In order to develop robust learning algorithm to achieve semantic image annotation, there are four challenging research issues to be addressed: 1) achieve more robust and effective image representations to bridge over the semantic gap; 2)utilize image content and the associated text descriptions; 3) integrate the user contextual information into the image annotation system;4) link image visual appearance to structured human knowledge in scalable image categorization / annotation. With this consideration, a set of novel probabilistic topic models are proposed to leverage image, text and user-created tags to achieve high performance image annotation and retrieval. The techniques and methods developed in this thesis are built on the state-of-the-art methods in statistical learning, image processing, social network analysis, content-based image retrieval and mining. The research will result in improved understanding of the issues involved in designing robust statistical model to integrate user context in image annotation and retrieval.

In the system biology community, there has been a long time focus on studying gene-expression data in isolated organisms and cultures. However, relatively less effort has been made to study the genome-wide gene-expression data from uncultured environment samples (like the ocean, soil and human body) and understand the underlying biological processes. Recently, the development of new sequencing techniques and meta-genomics has dramatically changed the way of genomics data acquiring and analyzing. Next generation sequencing methods (such as Roche/454 Sequencing and Illumina Sequencing) are able to extract very large amount (100 $\sim$ 1000 MB) of DNA fragment sequences from an environmental sample (like the ocean, soil and human body) in only a single run (the acquired data is also known a meta-genomic data). With the fast advancing sequencing techniques, large amount of sequenced genomes and meta-genomes from uncultured microbial samples becomes available. Based on the meta-genome sequences, bioinformatics researchers have done a lot of work to study the underlying biology process such as signal transduction, translation, and molecular functions like the biochemical activity of gene product. However, our knowledge about the biological functions encoded in the meta-genome sequence is still limited. Current functional annotation (genome-level annotation of biological functions) is still far from satisfied. The lack of high quality functional annotation of the major functionality encoded in the gene-expression data of given genome/meta-genome posed a great challenge in the task of interpreting the biological process of meta-genome.

The major objectives of analyzing and interpreting the large amount of meta-genomic data involve answering two questions. The first question is, 'Given a large number of genome fragments from an environmental sample, what genomes are there?' Answering this question requires mapping the meta-genomic reads to taxonomic units (usually a homology-based sequence alignment, this task is also known as taxonomic classification or taxonomic analysis). The second question is, 'What are the major functions of these genomes?' The answers to this question involve annotating the major functional units (such as signal transduction, metabolic capacity and gene regulatory) on the genome-level (a.k.a. functional analysis). Toward these two questions, we present a set probabilistic topic models to identify functional groups from microbial samples. The probabilistic topic models are derived from either taxonomic or functional-element abundance data (such as high abundance of specific functional group, high expression level of specific taxon, gene cluster, or specific metabolic pathway) acquired from either composition-based genome classification or homology-based alignment.

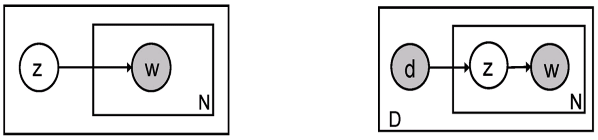
The remainder of this thesis is organized as follows. In Chapter 2, we review related works in generative topic models. In Chapter3, we present a set of novel probabilistic topic models to leverage image, text and user-created tags in semantic image annotation and retrieval. Chapter4 introduces probabilistic topic models for meta-genomic data analysis. We conclude the thesis in Chapter 5.

2. BRIEF REVIEW OF GENERATIVE LATENT SPACE MODELS

In this chapter, I would like to introduce the background of Generative Latent Space Models and review the related works on topic modeling.

2.1 GENERATIVE LATENT SPACE MODELS IN TEXT-MINING

The underlying assumption of generative latent space models in text-mining is that the co-occurrence patterns of words in a document are related to some unseen latent topics, which reflect different semantic context of words. During the last decade, several effective generative modeling approaches such as the Naive Bayesian model, probabilistic LSI (pLSI) [44] and Latent Dirichelet Allocation (LDA) [15] have been proposed. The Naive Bayesian model (Fig. 2.1a) assumes a fixed topic-word distribution over the whole data collection. The topic assignment of words in a document is simply decided by the prior probability of popic z and the Likelihood of word w given topic z. However, it's not the case that all the documents has the same topics, thus the PLSI model [44] is proposed.
The PLSI model (Fig. 2.1b)assumes that each document has a mixture of k topics. Fitting the PLSI model involves estimating the topic specific word distributions $p(w_i|z_k)$ and document specific topic distributions $p(z_k|d_j)$ from the data collection through maximum likelihood estimation (MLE). In PLSI model, the topic mixture probability for documents are fixed the model is estimated.It's not clear how to assign topics to documents outside the training dataset. For new coming document, the model needed to be re-estimated. Therefore, the PLSI model is not scalable.



(a) (b)

Figure 2.1: Basic generative latent space model (a) Naive Bayesian model, (b) the probabilistic latent semantic indexing (PLSI) model

The LDA model [15], initially proposed by Blei et al., has been popular with the text mining community in recent years due to its solid theoretical foundation and promising empirical retrieval performance. Application of LDA model involves text classification [15], social annotation [111], joint modeling of text and citations [76], etc. Compared to the PLSI model, the LDA model treats the probability of latent topics for each document as latent ”random” variables which are subject to change when new document comes.
As illustrated in Fig. 2.2, the LDA model involves two stages, that is, generating the prior probability of latent topics p(z) for each document and generating the conditional probability of words for each latent topic: $p(w|z)$(the sampling process of LDA model will be introduced in the section 4.1.1). The model is estimated via Gibbs Sampling Monte Carlo process [95].

2.1.1 Sampling Process of the LDA Model

In a multinomial distribution, there are n independent events. Each event has a fixed finite number k of possible outcome, with probability: $p_1, \cdots, p_k(p_i \geq 0, \sum_{i=1}^{k} p_i = 1)$ .If we denoted random variables $x_i(i = 1, , k)$ as the times a certain outcome i was observed , then the k-dimension $X = (x_1, , x_k)$ follows a multinomial distribution, in which

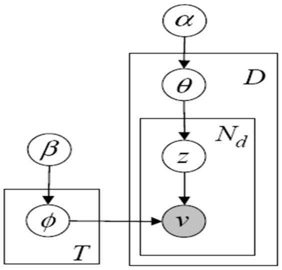


Figure 2.2: The framework of LDA model

$p(X) = \frac{n!}{x_1!\cdots x_k!}p_1^{x_1}\cdots p_k^{x_k}(p_i \geq 0, \sum_{i=1}^{k} p_i = 1)$.

Assuming that each document can be represented as a mixture over latent topics and that each topic is characterized by a distribution over words,given a corpus belong to certain category, if we define the random variables $x_i(i = 1,,k)$ as the times the i-th topic happen in a certain document D in that corpus and the parameters $p_i(i = 1,,k)$ as the prior probability of the topics in the corpus, then we can represent the mixture over topics in a document D by a k-dimension vector $X = (x_1,,x_k)$ which follows a multinomial distribution.

At this stage, only the dimensionality k of the distribution $X \sim Multi(X|k; p_1,,p_k)$ is known and fixed (as the number of topics can be predefined according to the labels in the training dataset). We still need to predict the prior probability of hyper-parameters: $p1,,pk$.

Since the hyper-parameters of the multinomial distribution: $p1,,pk$ are continuous real numbers on the interval [0,1], a Dirichlet priori should be used. The Dirichlet distribution (denoted as $Dir(\alpha)$), is a continuous multivariate probability distributions whose parameter(vector α)is composed with positive real numbers [7]. The probability density function of the Dirichlet distribution is illustrated in eq. (2.1), in which the symbol $\Gamma()$ represent the Gamma function (eq. (2.2)).The Gamma function is an extension of the factorial function to the real numbers. When n is a positive integer number, $\Gamma(n) = (n-1)!$.

$$p(\theta) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_k)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}, \theta_1, \cdots, \theta_k > 0, \sum_{j=1}^{k} \theta_j = 1 \tag{2.1}$$

$$\Gamma(\alpha_i) = \int_0^\infty t^{\alpha_i - 1} e^{-t} \mathrm{d}t \tag{2.2}$$

In Bayesian statistics, a class of prior probability distributions $p(\theta)$ is regarded as conjugate prior to likelihood functions $p(x|\theta)$ when the posterior distributions $p(\theta|x)$ are in the same family as $p(\theta)$. Adopting a conjugate prior is for the algebraic convenience in calculation. As pointed out in [7], the Dirichlet is the conjugate prior distribution for the parameters of the multinomial distribution. For example, if $X \sim Dir(\alpha)$ and $\beta|X \sim Multi(X)$, then $X|\beta \sim Dir(\alpha + \beta)$.

The expectation for Dirichlet distribution $X \sim Dir(\alpha)$ is:

$$E(X_i) = \frac{\alpha_i}{\alpha_0}, \alpha_0 = \sum_{i=1}^{T} \alpha_i \tag{2.3}$$

Given a total of D documents; and assume that there are a total of T latent topics. In the whole collection, supposing that there are a total of N_w text tokens, which belong to W words. The sampling process of LDA model is as follows:

For the d-th document, sample $\theta^d \sim Dir(\alpha)$, in which θ^d is a T-dimensional vector for topics in the document. For the t-th topic, sample: $\varphi_t \sim Dir(\beta)$. In each document, sample word topics $z^j \sim Multi(\theta^d)$ (here z^j means that the topic $z = j$). For each word w_i, sample $p(w_i|z^j) \sim Multi(\varphi^j)$.

The model is then estimated by the Gibbs Sampling Monte Carlo process [95], which involves iteratively estimating the posterior probability for topics from current word-topic assignment, and adopting a Monte Carlo process to determine the assignment of word-topic in the next round. During every iteration of the Gibbs Sampling process, the posterior probability for word-topic is updated as:

$$p(z_{wi} = j|w_i, \mathbf{w}_{-i}, \mathbf{z}_{-wi}) \propto \frac{\beta + n^{w_i}_{-i,j}}{W\beta + n^{w_i}_{-i,j}} \cdot \frac{\alpha + n^{d}_{-i,j}}{T\alpha + n^{d}_{-i,j}} \tag{2.4}$$

In which $n^{w_i}_{-i,j}$ is the total number of words assigned to topic j except for word w_i, and $n^{d}_{-i,j}$ is the total number of words in graph d assigned to topic j except for word w_i.

2.1.2 Dirichlet Process and Hierarchical Dirichlet Process

In nonparametric Bayesian statistics,the Dirichlet Process (DP) are used to learn mixture models whose number of components is automatically inferred from data. It is defined as a distribution of random probability measure $G_0 \sim DP(\gamma, H)$, in which is a concentration parameter and H is a base measure defined on a sample space Θ. By its definition, for any finite measurable partition of $\Theta : \{A_1, \cdots, A_r\}$, $(G_0(A_1), \cdots, G_0(A_r)) \sim Dirichlet(\gamma H(A_1), \cdots, \gamma H(A_r))$. Due to discrete nature of DP [50], it can be constructed by stick-breaking construction as follows (each $\theta_1, \cdots, \theta_k, \cdots, \theta_\infty$ is a distinct value on the space Θ, they are also considered as the parameters of mixture components during modeling).

$G_0 = \sum_{k=1}^{\infty} \beta_k \delta(\theta_k)$, in which $\beta_k = \alpha_k \prod_{i=1}^{k-1}(1-\alpha_i), \alpha_k \sim Beta(1, \gamma)$

The weights of mixture components $\beta = \{\beta_k\}(k = 1, \cdots, \infty)$ are also refer to as $\beta \sim GEM(\gamma)$.

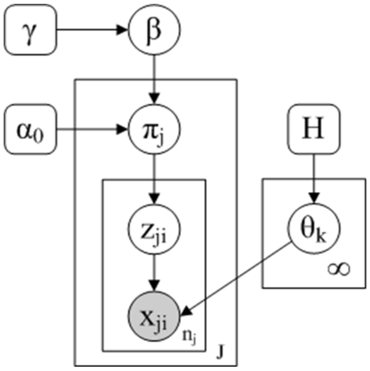


Figure 2.3: Stick-breaking construction of hierarchical Dirichlet process

The Hierarchical Dirichlet Process (HDP) considers $G_0 \sim DP(\gamma, H)$ as a global probability measure across the corpora and defines a set of child random probability measures $G_j \sim DP(\alpha_0, G_0)$ for each document j, which leads to different document-level distribution over semantic mixture components.

$$(G_j(A_1), \cdots, G_j(A_r)) \sim Dirichlet(\alpha_0 G_0(A_1), \cdots, \alpha_0 G_0(A_r)) \tag{2.5}$$

Each G_j can also be constructed by stick-breaking construction as:
$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta(\theta_k)$, in which $\pi_j = \{\pi_{jk}\}(k = 1, \cdots, \infty)$ specifies the weights of integer mixture component indicatork.
Now consider indicator variable set $K_l = \{k : \theta_k \in A_l\}, l = 1, \cdots, r$ for $l = 1,, r;$ then $K_1, \cdots K_r$ become a finite partition of integer indicators.
Substitute the stick-breaking construction of G_0 and G_j to Eq. (2.5), it follows that:

$$(\sum_{k\in K_1} \pi_{jk},\cdots,\sum_{k\in K_r} \pi_{jk}) \sim Dirichlet(\alpha_0 \sum_{k\in K_1} \beta_k,\cdots,\alpha_0 \sum_{k\in K_r} \beta_k) \tag{2.6}$$

Based on the aggregation properties of Dirichlet distribution and its connection with Beta distribution, we can show that:

$$\pi_{jk} = \pi'_{jk} \prod_{l=1}^{k-1}(1-\pi'_{jk}), \pi'_{jk} \sim Beta(\alpha_0\beta_k, \alpha_0(1-\sum_{l=1}^{k} \beta_l)) \tag{2.7}$$

It follows that $\pi_j \sim DP(\alpha_0, \beta)$.

2.2 GENERATIVE LATENT SPACE MODELS FOR IMAGE DOCUMENTS

The Latent Dirichelet Allocation (LDA) model proposed by [15] is originally designed to represent and learn topics from text documents. However, as will be introduced in Section 3.2, with the help of image descriptor quantization technique which maps image descriptors defined in continuous vector space to discretized code-words, we are able to represent an image document by text-like features (such as 'bag of visual words' derived from affine invariant local image descriptors). Once an image document is represented as 'bag-of-visual word', we will be able to achieve topic modeling from image documents in the same way as text documents.

Early approaches of topic modeling in image documents including directly using LDA [2] and using Spatial Latent Dirichlet Allocation (SLDA) [104] , in the following paragraph, we will review both approaches and follow up with a discussion of their contributions and limitations.

Fig. 2.4 demonstrates the LDA model for image documents proposed by [2]. In this

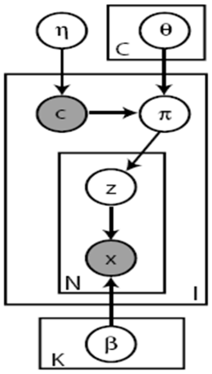


Figure 2.4: LDA model for image documents [2]

model, the salient points are detected by Lowe's difference-of-Gaussian (DoG) detector [63], i.e. the image patches including the salient points are described by 128-dimension SIFT descriptor. After that, the SIFT descriptors are quantized using a fixed 'codebook' of visual words, which was pre-learnt by applying k-means clustering on a large collection of detected patches from different categories of images.

Compared to the original LDA model, this model adds a category variable c for classification. θ is a matrix of size CK, in which C is the number of categories while K is the number of topics. θ_c is a K-dimension Dirichlet parameters conditioned on the category c, π is a K-dimension Dirichlet random variable: $\pi \sim Dir(\pi|\theta_c)$. Given one of the N patches in an image x_n, choose a topic vector $z_n \sim Multi(z_n|\pi)$, in which $z_n^k = 1$ indicate that the k-th topic is selected. β is a matrix which represent the relation between a single code word and a topic, in which $\beta_{kt} = p(x_n^t = 1|z_n^k = 1)$ represent the probability of the t-th codeword is selected when the k-th topic is selected. Finally, the probability of x_n is conditioned on

β and z_n:

$$p(x_n|z_n,\beta) = \prod_{k=1}^{K} p(x_n|\beta_k)^{\delta(z_n^k,1)} \tag{2.8}$$

The learning process of this model is as follows. Given an unknown scene, the likelihood function of x is learnt by margined out the latent variables:

$$p(x|\theta,\beta,c) = \int p(\pi|\theta,c)(\prod_{n=1}^{N}\sum_{z_n} p(z_n|\pi)p(x_n|z_n,\beta))\mathrm{d}\pi \tag{2.9}$$

In variational inference, the goal is to maximize the log likelihood by estimate θ and β. Using the Jensen's inequality, the lower bound of the log likelihood can be obtained. By adopting EM algorithm to maximize the lower bound on the log-likelihood, the approximation of optimal parameter θ and β of the model can be obtained.

In this model, the frequency of visual words in images is equivalent to the term frequency in the text documents. Therefore, this model can be an effective way to learn latent semantic topics from the visual words that extracted from image documents. However, compared to the textual word, the visual word has its unique characteristics. Since an image is a 2-dimension document, if we only focus on the frequency of visual word, we may fail to take into account the spatial correlation among the visual words and lose in touch with the spatial structure of the image.

Recently, Xiaogang et al. proposed the Spatial Latent Dirichlet Allocation (SLDA) model [104] , which was based on the hypothesis that image patches of the same object class should be close in space. Compared to LDA, the major improvement of SLDA is that it is able to model the spatial structure among visual words. The SLDA model first divided an image into local patches by a grid (unlike the SIFT descriptor in LDA model, the local patches in SLDA cover the whole image). Each local patch is then quantized into a visual word. The codebook is created by clustering all the local descriptors in image collection using K-mean.

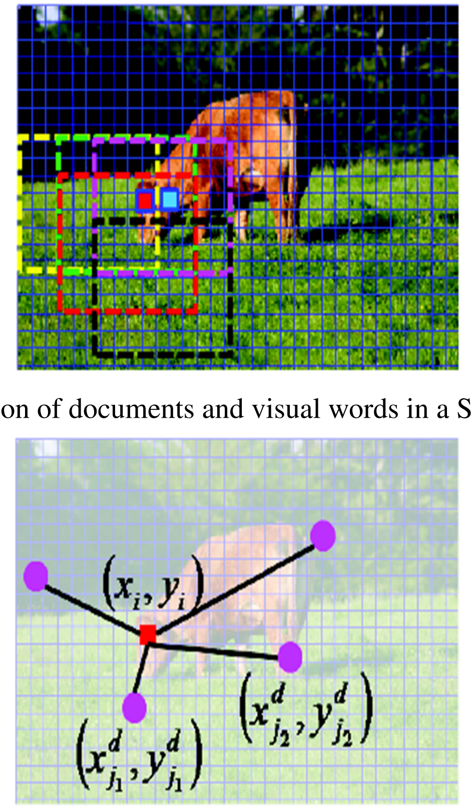
The SLDA model has a unique definition of document. Rather than define the whole image as a document, the SLDA model treats rectangle regions in an image as documents (Fig. 2.5 a), the rectangle regions are densely overlapped in the image and a certain image patch (or visual word, marked by colored spots) can be covered by several rectangle regions (documents).

After that, each document is represented by the location of its centre point. Given the location of visual words and documents, the probability that a visual word belongs to a certain document is:

$$p((g_i,x_i,y_i)|(g^d_{d_i},x^d_{d_i},y^d_{d_i}),\sigma) \propto \delta_{g^d_{d_i}}(g_i)exp\{-\frac{(x^d_{d_i}-x_i)+(y^d_{d_i}-y_i)}{\sigma^2}\} \tag{2.10}$$

By introducing this additional probability to visual words, the SLDA was able to represent the probability that a visual word belongs to a certain spatial group.

Although SLDA was able to represent spatial information of visual words, it still has some problems. Firstly, the notion of document in SLDA seems ill-defined. Intuitively, an image document, rather than a rectangle region, is the visual counterpart of a text document. What's more, since the visual words in a region are spatial closed to each other, according to the author's hypothesis, they are closed in semantic meaning and tend to belong to the same object. Thus, if we treat a region as a document, then the number of topics in a document would be very limited As a result, a rectangle region is not comparable to a text

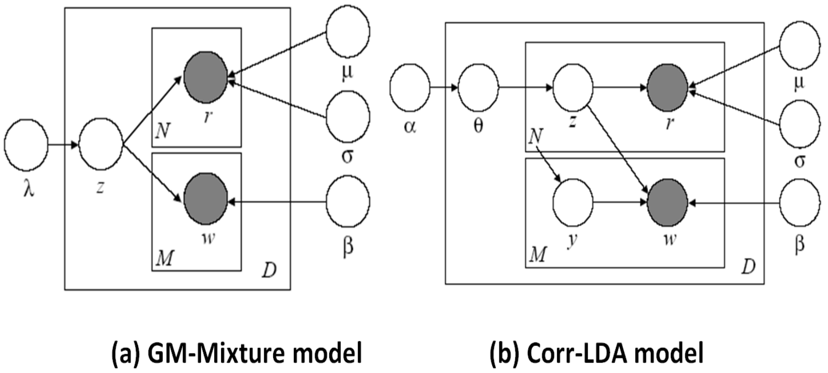


(a) Illustration of documents and visual words in a SLDA model

(b) The spatial relation between documents and visual words

Figure 2.5: Spatial Latent Dirichlet Allocation [104]

document, which is usually a mixture over latent topics. Secondly, in the SLDA approach, the visual words are obtained by directly quantizing the local patches over the whole image (rather than from the salient points). In other words, SLDA considers image patches with salient points equally as the non-salient patches from homogeneous regions (which take up a major part in image). Therefore, the algorithm becomes more computational intensive than LDA with visual words and is less effective in representing the image content (unable to account for the saliency).



(a) GM-Mixture model (b) Corr-LDA model

Figure 2.6: Multinomial mixture model and Correspondence LDA model [14]

2.3 GENERATIVE LATENT SPACE MODEL FOR IMAGE CONTENT AND TEXT ANNOTATIONS

The explosive increase of image data on Internet has made it an important, yet very challenging task to index and automatically annotate image data. One possible solution is developing generative latent space models to learn the correlation between image content and corresponding text annotations.

Toward that end, Blei et al. proposed the Gaussian Multinomial Mixture (GM-Mixture) model and Correspondence LDA (Corr-LDA) model [14]to make image content associated with the latent topics of caption words. As illustrate in Fig. 2.6, both models involve an additional ‘branch’ to generate topics from image feature (which are commonly represented by the mean (μ) and variance (σ) of multidimensional Gaussians in the feature space) besides the branch to generate text topics.

2.3.1 Gaussian Multinomial Mixture (GM-Mixture) Model

For the GM-Mixture model, a single discrete variable z is used to represent a joint clustering of an image and its caption. An image-caption pair is assumed to be generated by first choosing a value of z, and then repeatedly sampling N region description r_n and M caption words w_m conditional on the chosen value of z. The variable z is sampled once for each image-caption pair, and is held fixed when generating other components. The joint distribution of the hidden factor z and the image-caption pair $(\bar{r}, \vec{w})$ is:

$$p(z, \vec{r}, \vec{w}) = p(z|\lambda) \prod_{n=1}^{N} p(r_n|z, \mu, \sigma) \cdot \prod_{m=1}^{M} p(w_m|z, \beta) \tag{2.11}$$

Given a fixed number of factors K and a collection of images-caption pairs, the parameters of a GM-Mixture model can be estimated by using the EM algorithm [5]. This process will end up with K Gaussian distributions over features and K multinomial distributions over words which together describe a clustering of the images-caption pairs. Since each image and corresponding caption are assumed to be generated conditional on the same latent factor z, the resulting multinomial and Gaussian parameters will be corresponded. Image content with high probability under a certain factor will likely have a caption with high probability in the same factor.

Finally, the joint probability of an image-caption pair can be computed by simply marginalizing out the hidden factor z and the conditional distribution of words given the image content can be obtained based on the Bayesian rule (eq. (2.12)).

$$p(w|\vec{r} = \sum_{z} p(z|\vec{r}) p(w|z)) \tag{2.12}$$

When modeling is finished, an image can be annotated by clustering pixel points into regions, extracting feature vector (color and texture) from the regions, and then using the GM-Mixture model to compute the probability of each word being assigned to the image regions.

2.3.2 Correspondence LDA (CorrLDA) Model

The Correspondence LDA (CorrLDA) model [14], initially proposed by Blei et al. in year 2003, provides a natural way to learn the correlation between text words and other entities under an regular LDA model schema. In this model, topics generated from text words are used to generate other entities (such as image features). As mentioned in Chapter 3, the 'bag of visual words' feature is not only able to provide text-like features, which brings computational conveniences, but also more information intensive than the global image features (such as color and texture of image regions). Inspired by the success of directly applying LDA model to latent topics from visual words [2], we may modify the original Corr-LDA model by using the text-like 'bag of visual words' feature (which fits a multinomial distribution) in stead of using the global image features, which are commonly represented as the mean (μ) and variance (σ) of multidimensional Gaussians in the feature space.

The modified Corr-LDA model for visual words and image captions is illustrated in Fig. 2.7, which is followed up with a detailed Gibbs sampling process for the model estimation.

(a) Sampling hyper-parameters of Corr-LDA Model

Assuming that there are a total of D image-caption pairs in the data collection, and that there are a total of T latent topics, N_w text tokens from W words, and a total of $\breve{N}_{\breve{w}}$ visual

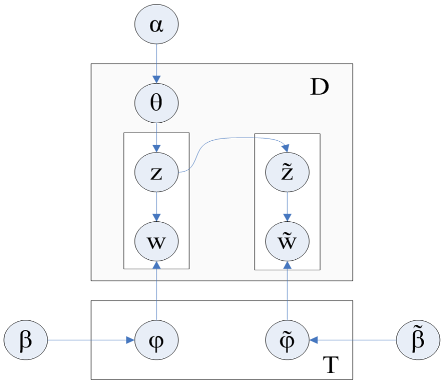


Figure 2.7: Corr-LDA model for visual words and image captions

word entities from $\breve{W}$ visual words. For the d-th image-caption pair, sample $\theta^d \sim Dir(\alpha)$, in which θ^d is a T-dimensional vector for topics in the image-caption pair. For the t-th topic, sample $\varphi_t \sim Dir(\beta)$ and $\breve{\varphi}_t \sim Dir(\breve{\beta})$. In each image-caption pair, sample word topics $z^j \sim Multi(\theta^d)$ and visual word topics $\breve{z}^j \sim Uniform(z^j)$ (here z^j means that the topic $z = j$, and $z = j$ is equivalent to $\breve{z} = j$). For each word w_i, sample $p(w_i|z^j) \sim Multli(\varphi^j)$, for each visual word $\breve{w}_i$, sample $p(\breve{w}_i|\breve{z}^j) \sim Multi(\breve{\varphi}^j)$.

(b) Update of the word-topic probability

Given the settings in Section 3.2, the posterior probability for word-topic is:

$$p(z_{wi} = j|w_i, \mathbf{w}_{-i}, \mathbf{z}_{-wi}) \propto p(w_i|z_{wi} = j, \mathbf{w}_{-i}, \mathbf{z}_{-wi}) \cdot p(z = j|\mathbf{w}_{-j}, \mathbf{z}_{-wi}) \tag{2.13}$$

Recall that $p(w_i|z^j) \sim Multli(\varphi^j)$, by integrating over all the different value of φ^j , we have:

$$p(w_i|z_{wi} = j, \mathbf{w}_{-i}, \mathbf{z}_{-wi}) = \int p(w_i|z = j, \varphi^j, \mathbf{w}_{-j}, \mathbf{z}_{-wi}) p(\varphi^j|\mathbf{w}_{-j}, \mathbf{z}_{-wi}) \mathrm{d}\varphi^j \tag{2.14}$$

In eq. (2.14), we have:

$$p(w_i|z=j,\varphi^j,\mathbf{w}_{-j},\mathbf{z}_{-wi}) = \varphi^j, p(\varphi^j|\mathbf{w}_{-i},\mathbf{z}_{-wi}) \propto p(\mathbf{w}_{-i},\mathbf{z}_{-wi}|\varphi^j)\cdot p(\varphi^j) \tag{2.15}$$

in which $p(\varphi^j) \sim Dir(\beta)$ and $p(\mathbf{w}_{-i},\mathbf{z}_{-wi}|\varphi^j) \sim Multi(\varphi^j)$

Since the Dirichlet is the conjugate prior distribution for the parameters of the multinomial distribution, it follows that: $p(\varphi^j|\mathbf{w}_{-i},\mathbf{z}_{-wi}) \sim Multi(\beta + n^{wi}_{-i,j})$, in which $n^{wi}_{-i,j}$ is the total number of words assigned to topic j except for word w_i.

Therefore, eq. (2.14) is in fact the expectation of φ^j when we already know $\mathbf{w}_{-i},\mathbf{z}_{-wi}$, thus it equals to $\frac{\beta+n^{wi}_{-i,j}}{W\beta+n^{wi}_{-i,j}}$ (recall that the expectation for Dirichlet distribution $X \sim Dir(\alpha)$ is: $E(X_i) = \frac{\alpha_i}{\alpha_0}, \alpha_0 = \sum_{i=1}^{T}\alpha_i$) Similarly, we get:

$$p(z=j|\mathbf{w}_{-i},\mathbf{z}_{-wi}) = \int p(z=j|\theta^d)\cdot p(\theta^d|\mathbf{w}_{-i},\mathbf{z}_{-wi})\mathrm{d}\theta^d \tag{2.16}$$

it follows that: $p(\theta^d|\mathbf{w}_{-i},\mathbf{z}_{-wi}) \propto p(\mathbf{w}_{-i},\mathbf{z}_{-wi}|\theta^d)\cdot p(\theta^d)$.

Since $p(\theta^d) \sim Dir(\alpha)$ and $p(\mathbf{w}_{-i},\mathbf{z}_{-wi}|\theta^d) \sim Multi(\theta^d)$, we have :

$$p(\theta^d|\mathbf{w}_{-i},\mathbf{z}_{-wi}) \sim Dir(\alpha + n^d_{-i,j}) \tag{2.17}$$

in which $n^d_{-i,j}$ is the total number of words in graph d assigned to topic j except for word w_i.

Similar to eq. (2.13), eq. (2.16) is simply calculating the expectation of θ^d when $\mathbf{w}_{-i},\mathbf{z}_{-wi}$ is already known. Therefore, eq. (2.16) should equal to $\frac{\alpha+n^d_{-i,j}}{T\alpha+n^d_{-i,j}}$.

Summing up, we get:

$$p(z_{wi} = j|w_i, \mathbf{w}_{-i}, \mathbf{z}_{-wi}) \propto \frac{\beta + n^{wi}_{-i,j}}{W\beta + n^{wi}_{-i,j}} \cdot \frac{\alpha + n^{d}_{-i,j}}{T\alpha + n^{d}_{-i,j}} \tag{2.18}$$

In which $n^{wi}_{-i,j}$ is the total number of words assigned to topic j except for word w_i , and $n^{d}_{-i,j}$ is the total number of words in graph d assigned to topic j except for word w_i.

Given the word-topic posterior probability, the Monte Carlo process becomes really straightforward, which is similar to throwing dice (based on the posterior probability) to determine the assignment of topics to each words for the next round.

(c)Update of the visual word-topic probability

The posterior probability for visual word-topic is:

$$p(\breve{z}_i = j|\breve{w}_i = v, \breve{\mathbf{z}}_{-\mathbf{i}}, \breve{\mathbf{w}}_{-\mathbf{i}}, \mathbf{z}, \breve{\beta}) \propto p(\breve{w}_i|\breve{z}_i = j, \breve{\mathbf{z}}_{-\mathbf{i}}, \breve{\mathbf{w}}_{-\mathbf{i}}, \mathbf{z}, \breve{\beta}) \cdot p(\breve{z}_i) = j|\breve{\mathbf{z}}_{-\mathbf{i}}, \breve{\mathbf{w}}_{-\mathbf{i}}, \mathbf{z}, \breve{\beta}) \tag{2.19}$$

We have $p(\breve{w}_i|\breve{z}_i = j, \breve{\mathbf{z}}_{-i}, \breve{\mathbf{w}}_{-i}, \mathbf{z}, \breve{\beta}) = \int p(\breve{w}_i|\breve{z}_i = j, \breve{\varphi}^j, \breve{\mathbf{z}}_{-i}, \breve{\mathbf{w}}_{-i}, \mathbf{z}, \breve{\beta}) \cdot p(\breve{\varphi}^j|\breve{\mathbf{z}}_{-i}, \breve{\mathbf{w}}_{-i}, \breve{\beta}) \mathrm{d}\breve{\varphi}^j$ Based on the Bayesian theorem, we have: $p(\breve{\varphi}^j|\breve{\mathbf{z}}_{-i}, \breve{\mathbf{w}}_{-i}, \breve{\beta}) \propto p(\breve{\mathbf{z}}_{-i}, \breve{\mathbf{w}}_{-i}|\varphi^j|) \cdot p(\breve{\varphi}^j)$, in which $p(\breve{\varphi}^j \sim Dir(\breve{\beta})$ and $p(\breve{\mathbf{z}}_{-i}, \breve{\mathbf{w}}_{-i}|\breve{\varphi}^j) \sim Dir(\breve{\beta} + n^{\breve{w}_i}_{-i,j})$. It then follows that: $p(\breve{\varphi}^j|\breve{\mathbf{z}}_{-i}, \breve{\mathbf{w}}_{-i}, \breve{\beta}) \sim Dir(\breve{\beta} + n^{\breve{w}_i}_{-i,j})$.

Therefore, we get:

$$p(\breve{w}_i|\breve{z}_i = j, \breve{\mathbf{z}}_{-\mathbf{i}}, \breve{\mathbf{w}}_{-\mathbf{i}}, \mathbf{z}, \breve{\beta}) \propto \frac{\breve{\beta} + n^{\breve{w}_i}_{-i,j}}{\breve{W}\breve{\beta} + n^{\breve{w}_i}_{-i,j}} \tag{2.20}$$

Recall that $z = j$ is equivalent to $\breve{z} = j$, so the prior probability of $p(\breve{z}_i = j|\breve{\mathbf{z}}_{-i}, \breve{\mathbf{w}}_{-i}, \breve{\beta})$ simply equals to the ration of $n_{\breve{z}}$ over N_w , in which $n_{\breve{z}}$ is the total number of words assigned to topic $\breve{z}$ (because $\breve{z} = j$ is equivalent to $z = j$). So the posterior probability for visual word-topic is:

$$p(\breve{z}_i|\breve{w}_i = v, \breve{\mathbf{z}}_{-\mathbf{i}}, \breve{\mathbf{w}}_{-\mathbf{i}}, \mathbf{z}, \breve{\beta}) \propto \frac{n_{\breve{Z}}}{N_w} \cdot \frac{\breve{\beta} + n^{\breve{w}_i}_{-i,j}}{\breve{W}\breve{\beta} + n^{\breve{w}_i}_{-i,j}} \tag{2.21}$$

In which $n_{\breve{z}}$ is the total number of words assigned to topic $\breve{z}$ ($\breve{z} = j$ is made equivalent to $z = j$).

When the whole model is estimated from the data collection, we will be able to tell the correlation between image content and image captions.

3. TOPIC MODEL FOR IMAGE MINING

3.1 OVERVIEW AND OBJECTIVE

The prevalence of digital imaging device, such as digital camera and digital video camera, has generated an increasingly large amount of unlabeled multimedia data, especially unlabeled image data. With nearly a million new images being added in a single day, the Flickr.com, one of the most popular photo sharing websites, now hosts over 3 billion shots of user-uploaded images. Manually annotating such a huge amount of image data is time-consuming, laborious and prohibitively expensive.To face the challenge of enormous explosion of unlabeled online image resources, it is very important to develop context-sensitive robust automatic image annotation system.

Early text-based image annotation approaches include using lexical chain analysis of the nearby text descriptions on Web pages [89] and using WordNet to disambiguate description words [6]. Even though the text and tags surrounding the image do provide some insight about the semantic meaning of image content, however, they are usually too noisy to be directly used as the image annotation. It's difficult for a purely text-based approach to achieve effective image annotation and searching in a practical application. Therefore, its important to annotate and retrieve images based on their visual content. With this consideration, automatic image annotation approaches have been closely related to computer vision, image processing and content-based image retrieval [6], [59]. During the last decade, we have seen great progress in developing automatic image annotation systems, related works involve considering image annotation as a clustering/categorization problem [60], [35], as an image searching problem [101], and as statistical modeling problem [15], [38], [8], [18]. Despite the success of these works, however, researchers are still facing two major difficulties in providing reliable and accurate annotation for images. One is lacking of benchmark image dataset with clear image hierarchy and comprehensive text descriptions, the other is lacking of effective ways to represent the image content and associate it with the text descriptions.

Most existing algorithms and models for semantic image annotation are either the generative models (including mixture models and topic models) or the discriminative models [41]. The mixture models usually define latent variables to encode the joint distribution of image visual features and annotation words [38], [78], or encode the spatial relations between labeled objects and parts [94] to improve the labeling accuracy. Therefore, this approach can be considered as a non-parametric approach which estimates the density over the concurrence of image and annotations.The topic models, such as latent Dirichlet allocation [15], [52] and hierarchical Dirichlet process [79], represent image as mixture of latent topics, in which each topic is a distribution over image features and annotation words. The parameter estimation involves estimating the image-level topic mixture as well as the topic-specific feature distribution. Topic models have been proposed to build correspondence between text words and image features (both discrete and continuous) [8], [18]. The discriminative models, on the other hand, define a set of classifiers with respect to each individual semantic category (corresponding to a set of annotation words). And use these classifiers to predict whether a testing image belongs to the semantic category associated with particular words. With sufficient positive/negative samples, the state-of-the-art image classification methods (such as SVM classifier with non-linear kernel) has the strength to achieve perfect separation of the hyper-space of image and text features, however, the classifier cannot explicitly tell us which features is more informative and to which extend does a set of visual features correlated with a particular annotation word.

Recently, Web 2.0 tools and environments have made collaborative tagging very popular; in which any user can collaboratively assign open-ended text words, in the form of keywords or category labels, to online shared resources for the purpose of organizing and refinding these resources. Flickr.com is such an example. As one of the most popular photo sharing and online community platforms, Flickr.com allows photo submitters to describe images using tags, which describe different aspect of the picture (such as location, time, author/owner, etc) and allow users to re-find pictures using tags as queries. Even though a number of Flickr images have user-created tags which provide valuable information and metadata to be utilized for context sensitive information retrieval, however, these tags have various functional purposes, some image tags tend tobe subjective in the sense that they might be free-form text, thus not directly relate to the image content; and typically only a few of many possible tags have been added to each image. So it casts a doubt on the importance and usefulness of the image tags. Also, in Web 2.0 environment, the information seeking process starts with the user, the usually missing part of information retrieval (IR) systems is that manyusers not be able to express their information needs well. Thus, in order to develop a context sensitive and user-friendly IR system, it is necessary to explore users' tagging and searching patterns in an online social tagging system.

Furthermore, in our daily life, a large amount of our verbal communication describes the scene/environment around us. Also, recent years have seen increasing amount of online visual resources (such as images and videos) with natural language descriptions. Such information may potentially serve as a rich knowledgebase of how people construct natural language to describe visual content. In order that an image annotation system facilitate extracting and understanding the knowledge encoded in the visual content, it is very important to generate descriptive topic models that combines natural language descriptions with image visual attributes. This work differs from conventional computer vision approaches such as scene recognition and object classification. Instead, it will encode additional semantic information such as the relation between object categories and different visual attributes, which is linked to natural language descriptions of human knowledge (such as Wikipeida) and then used to generate descriptive topic model regarding object with those visual attributes.

In this chapter, four research questions are addressed. After that, robust statistic models are proposed to leverage image, text and user-created tags to enhance the performance of image annotation and retrieval. The four research questions are as follows.

How to achieve more robust and effective image representations to bridge over the 'semantic gap'?

In automatic image annotation and retrieval, how to bridge over the 'semantic gap' [6] between image features and high-level semantic meanings is a major challenge. As introduced in Section 3.2, state-of-the-art image representation approaches either represent image content by its global spatial layout [43], [80], [90], or represent image by saliency model (such as salient part and key-points) [67], [91], [63], [3], [109], [51], [18], yet either approach has its advantages and drawbacks. In our approach, instead of treating these two approaches separately, we utilize the saliency model (salient regions and key-points) as a complement part of spatial layout model. Our motivation comes from the fact that the mechanism of human visual perception allows for very rapid holistic image analysis to provide a coarse context of image scene (spatial layout model), yet it also gives rise to a small set of candidate salient locations in a scene (saliency model) that needs to be intensively studied.

How to utilize image content and the associated text descriptions?

High quality text descriptions of images play a vital role as training and benchmarking data in developing and evaluating an automatic image annotation system. So the first issue of this research question is concerned with building the benchmark dataset for the purpose of training an automatic image annotation and retrieval system. In Section 3.3, we propose to associate image captions in biomedical literatures with semantic concepts from Unified Medical Language System (UMLS) and enrich image in ImageNet dataset by text descriptions from Wikipedia. The second issue is proposing an effective model to study the correlation between image and text descriptions, specifically, a hierarchical probabilistic model with background distribution (HPB) and the probabilistic topic-connection (PTC) model are introduced to enables more effective and robust modeling of the co-existing image features and annotations.

How to integrate the user contextual information into the image annotation system?

After we establish the topic-connection model between image appearance and text descriptions, we plan to further extend this model to the user-created tag information, in which we propose to use a user-perspective model to simulate the generation of tags. In this user-perspective model, one tag may be either derived from the consensus, in other words, the 'standard text description' associated with image appearance (i.e. the associated Wikipedia text descriptions), or from users' specific interests and background. In Section 3.4, the proposed user-perspective model will improve the over-all satisfaction of automatic tag recommendation for specific users and make image retrieval more effective.

How to link image visual appearance to structured human knowledge in scalable image categorization / annotation?

Due to the increasing need of linking visual appearance to structured human knowledge in scalable image categorization/annotation, the extraction of semantic visual attributes has received increasing research focus. By its literal definition, the term ‘attribute’ means ‘a quality or characteristic inherent in or ascribed to an object’. Compared to low-level image features, semantic visual attributes have much stronger relation to both object categories and human knowledge. It should be noted that although various types of attributes can be used to literally describe an object, however, only a small fraction of those attributes may be visible from an object image. Moreover, the usage of textual attributes may differ in different context. In order that the semantic attributes be useful for image annotation, these attributes should be visible and discriminating among different object categories, also, the union of semantic visual attributes should have sufficient coverage, which means that each object category be covered by at least one attribute. In Section 3.5, an effective framework is proposed to reliably extract both categorical attributes and depictive attributes. After that, the obtained semantic associations between visual attributes and object categories are combined in a text-based topic model for descriptive latent topics extraction from external textual knowledge sources.

3.2 IMAGE FEATURE REPRESENTATION TO BRIDGE OVER THE 'SEMANTIC GAP'

In the conventional view, an image scene is usually understood as a spatial configuration of objects, and its semantic recognition needs to initially find the objects and their exact locations. Most conventional image representation approaches are object-based [80], [90]. In a realistic scenario, illuminate changes, dynamic backgrounds, and affine variation usually make object-based image representation approaches less effective. Region-based image representation approach represent image content by segmented image regions as well as their configuration relationships. The blob-world approach [12] is a well-known region-based mage representation approach. The major problem with region-based approach is that it requires reliable region segmentations. The context-based (or holistic) image representation approach, as it named, ignores most details, bypasses the traditional steps of segmentation-recognition and considers the spatial layout of the whole image scene as one individual object.

Recent years have also seen great advances in using saliency model as an intermediate step to interpret the semantic meanings of images. Affine invariant saliency model such as the SIFT descriptor [67], [91], [63] have exhibited very good performance in image categorization and semantic image retrieval across several well-known databases such as the Caltech 101, the TRECVID and the Visual Object Classes (VOC) datasets [35], [60], [1], [39], [66]. Comprehensive study indicates that the SIFT descriptor [63] outperforms other affine invariant local descriptors due to its high robustness to image variations [71], [91], [51], [3], [109], [18].

3.2.1 The Context-based Image Representation

The context-based image representation involves color features, texture features and other statistics that describe the overall distribution of visual contents in images. Typically, color components in RGB, HIS or LUV color space are used as color features; while discrete wavelet transform (DWT) [22] and Gabor filters [23] are used to extract texture features.

More specifically, Belongie and Carson introduced the 'blobworld approach [12], which bring global image features (color, texture) together and represent their spatial distributions as a mixture of Gaussians. However, except for some favorable conditions, global features are always sensitive to the change of light, color and point of view. Therefore, it cannot well represent the highly varied images in the real world. What's more, when the entire image is represented as a whole using global features, we will lose in touch with the specific characters of individual local structures in the image. Therefore, the global image features suffers a significant information loss.

Color Space

In the HSI color space (in which 'H' means 'hue', 'S' stands for 'saturation', while 'I' represents 'intensity'), the color information is represented in polar coordinates.

The LUV color components (eq. (3.1), in which 'L' stands for 'Luminance' while 'U' and 'V' represent 'Chrominance', ''hroma', respectively) have been proven to be more discriminative than traditional RGB color components.

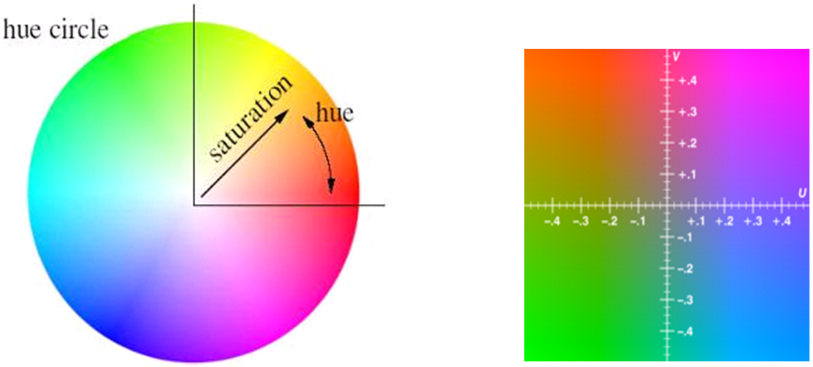


Figure 3.1: Comparison of HSI and LUV color space

$$\begin{pmatrix} L \\ U \\ V \end{pmatrix} = \begin{pmatrix} 0.299 & 0.587 & 0.114 \\ -0.148 & -0.289 & 0.437 \\ 0.615 & -0.515 & -0.100 \end{pmatrix} * \begin{pmatrix} R \\ G \\ B \end{pmatrix}, \begin{pmatrix} R \\ G \\ B \end{pmatrix} = \begin{pmatrix} 1.000 & 0.000 & 1.140 \\ 1.000 & -0.395 & -0.581 \\ 1.000 & 2.032 & 0.000 \end{pmatrix} * \begin{pmatrix} L \\ U \\ V \end{pmatrix} \tag{3.1}$$

Discrete Wavelet Transform

The 2D discrete wavelet transform (DWT) [22] is intensely used in texture features extraction. It separates an image into a lower resolution approximation image (LL) as well as horizontal (HL), vertical (LH) and diagonal (HH) detail components. For an image, the 2-D filter coefficients can be expressed as

$$h_{LL}(m,n) = h(m)h(n), h_{LH}(k,l) = h(k)g(l), h_{HL}(m,n) = g(m)h(n), h_{HH}(k,l) = g(k)g(l) \tag{3.2}$$

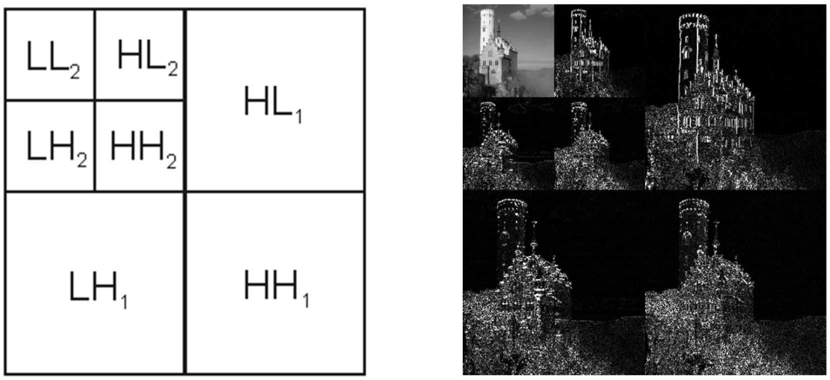


Figure 3.2: Scale 2-Dimensional Discrete Wavelet Transform, a Haar Wavelet transform is applied to the L component (Luminance) of the image [22]

In which the first and second subscripts denote, respectively, the low-pass and high-pass filtering along the row and column directions of the image.The process can then be repeated to computes multiple ‘scale’ wavelet decomposition, as in the 2 scale wavelet transform shown in Fig. 3.2.

Gabor Filters

A typical 2D Gabor filter is formulated in eq. (3.3), which is a product of a frequency shift and a frequency rotation.

$$G(x,y) = \frac{1}{2\pi\sigma\beta} e^{-\pi[\frac{(x-x_0)^2}{\sigma^2}+\frac{(y-y_0)^2}{\beta^2}]} e^{i[\xi_0 x+v_0 y]} \tag{3.3}$$

Gabor filter is able to represent textures in different scale and orientation. Texture components of different scales and orientations in a certain image block can be enhanced by a convolution of image block and corresponding Gabor filter as in eq. (3.4).

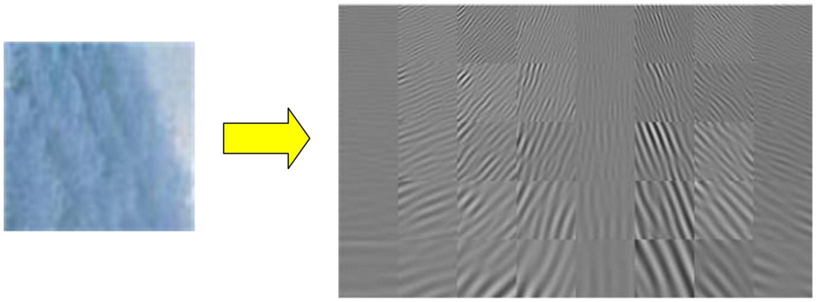


Figure 3.3: Results of convolution, each block represent the convolution of the image (up-left) and a Gabor filter from the filter bank (which has5 scale levels in total and 8 orientations for each scale level)

$$h_{m,n}(x,y) = f(x,y) * G_{m,n}(x,y) \tag{3.4}$$

Where hm,n(x,y) stands for the enhanced texture component; $f(x,y)$ represents the image block; $G_{m,n}(x,y)$ is the corresponding Gabor filter and '* ' denotes the discrete convolution. Fig. 3.3 is an illustration of convolution of image block and corresponding Gabor filter, in which a total of 40 Gabor filters form a so call 'filter bank', with 5 different scale levels and 8 orientations for each scale level. In the convolution results, high amplitude blocks indicated that the image has significant textual components in corresponding scale and orientation.

The Blobworld Approach

The blobworld approach [12]models the feature space and the spatial distribution of an image as a Gaussian mixture model (GMM) [13]. A simplest model which contains only color and spatial information is presented as follows.

Firstly, each pixel is represented by 5-dimensional normalized feature vector, including 3-dimensional LUV color features plus 2-dimensional (x,y) position. In an image with m pixels, a total of m feature vectors: $\mathbf{y}_1, \cdots, \mathbf{y}_m (\mathbf{y}_i \in R^5)$ will be obtained. Then, each image is assumed to be a mixture of n Gaussians in the 5-dimensional feature space and the Expectation-Maximization (EM) algorithm is used to iteratively estimate the parameter set of the Gaussians. The parameter set of Gaussian mixture is: $\theta = \{\mu_i, \sum_i, \alpha_i\}_{i=1}^n$, in which $\mu_i \in R^d$ is the mean of the i-th Gaussian; $\sum_i$ denotes the $d \times d$ covariance matrix; while α_i represents the prior probability of the i-th Gaussian.

At each E-step of the EM algorithm, we estimate the probability of a particular feature vector $\mathbf{y}_i$ belonging to the i-th Gaussian according to the outcomes from the last maximization step (eq. (3.1))

$$p(i|j, \theta_t) = p(z_j = i|y_j, \theta_t) = \frac{p(y_j|z_j = i, \theta_t)p(z_j = i|\theta_t)}{\sum_{k=1}^n p(y_j|z_j = k, \theta_t)p(z_j = k|\theta_t)} \tag{3.5}$$

In which z_i denotes which Gaussian $\mathbf{y}_i$ comes from and θ_t is the parameter set at the t-th iteration. At each M-step, the parameter set of the n Gaussians is updated toward maximizing the log-likelihood, which is:

$$Q(\theta) = \sum_{j=1}^m \sum_{i=1}^n p(z_j = i|y_j, \theta_t) ln(p(y_j|z_j = i|\theta)p(z_j = i|\theta)) \tag{3.6}$$

When the algorithm converges, the parameter sets of n Gaussians as well as the probability are obtained. Based on the estimated GMM model, an image can be decomposed into n regions and will indicate which region a given pixel point most likely belongs to (Fig. 3.4).

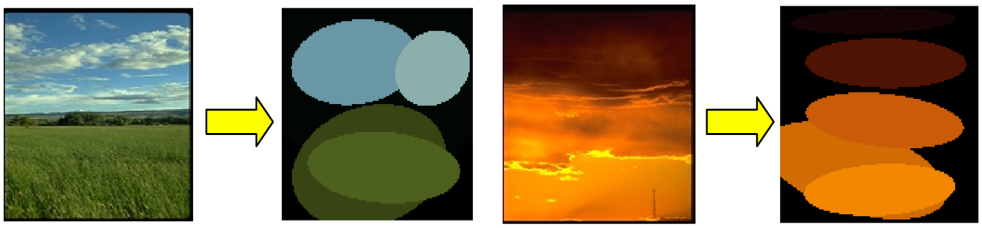


Figure 3.4: Input images and image modeling via Gaussian mixture model (GMM) [13]

3.2.2 Image Representation by Saliency Models

Recently, the 'Bag of local Features (BoF) approach exhibits very good performance in image categorization and semantic image retrieval across several well-known databases such as theLabelMe, the TRECVID and the Visual Object Classes (VOC)datasets [51], [3], [109], [16], [79]. The underlying assumption of this approach is that the visual patterns of different image categories can be represented by different distributions of local structures. Similar to the bag-of-words model in text mining, the BoF method represents images as an unordered collection of salient parts, which help to categorize images.

The 'bag of local features' approach involves two major branches: one represents images by sparse local features, while the other one represents images by dense local features. The underlying assumption for sparse local features is that objects are composed by several unique and salient 'parts', whereas the dense local features approaches assumes that, the patterns of different image categories can be represented by different distributions of key-points/microstructures, thus images should be represented as hundreds of key-points, or 'salient microstructures'. The difference for these two representation approaches are illustrated in Fig. 3.5.

In sparse local features approaches, the Kadir-Brady (KB) saliency detector [54] is usu-

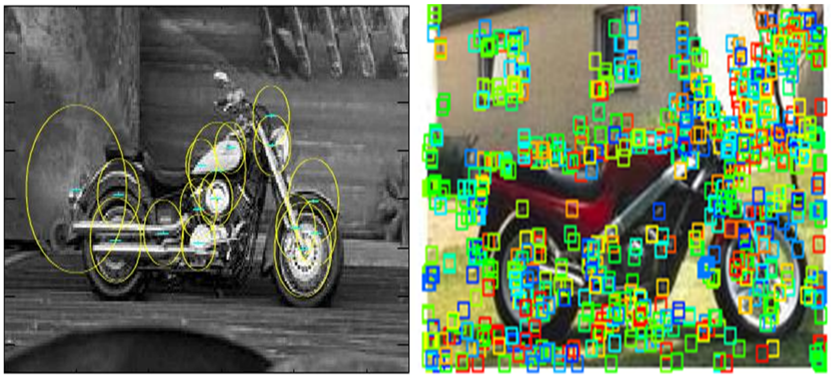


Figure 3.5: Represent images as bag of local features

ally used to detect salient ‘parts’ from images. After that, image patches containing salient ‘parts’ are quantified by performing principal component analysis (PCA) on them. The major problem with sparse local features is that objects may receive insufficient coverage from the feature detector. What’s more, the sparse local features (salient parts) carry little information about the background context of image (such as ‘open county’, ‘inside city’ and ‘beach’). In dense local features approaches, the key-points detection is achieved by either Harris-Laplace detector [70], which estimates the affine neighborhood by the affine adaptation process based on the second moment matrix, or Difference-of-Gaussian (DoG) salient points detector [63], which detect the scale-space extreme points in the difference-of-Gaussian images. The Harris-Laplace detector tend to extract corner points or corner-like regions, which mainly locate around objects, while DoG salient point detector tends to extract salient spots or blob-like regions from images.

Commonly, the Scale Invariant Feature Transform (SIFT) descriptor [63] is used to quantify the detected key-points. The SIFT descriptor is a 128-dimensional feature vector which captures the spatial structure and the local orientation distribution of a patch surrounding key-points. Mikolajczyk et al. evaluated several local image descriptors according to their stability to rotation, scaling, affine transform and illumination change and found the SIFT descriptor performs best [69], [72].

Maximally Stable Extremal Region(MSER)

The Maximally Stable Extremal Region(MSER) is a widely used image feature to represent regions. Unlike the SIFT descriptors, which is derived from key-points, the detected MSER regions are local homogeneous parts in objects (Fig. 3.6). Although the MSER detector output relatively smaller number of MSER features than SIFTdescriptors, their distinctness is higher. Specifically, MSER detection begins with segmenting a set of image regions whose inner intensity value is less than certain thresholds while all intensities around the region boundary is greater than the same threshold. After that, a maximally stable extremal region is obtained when the area of the segments changes the least with respect to the threshold [67]. Extensive study reveal that the set of MSER regions is closed under continuous geometric transforms, thus providing an efficient affine invariant region detector for local image appearance [37].We also extend MSER detector to multiple scales by constructing Gaussian pyramid and applying MSER detection separately in each resolution level.

After MSER detection, each detected elliptical region is normalized to circular patch of constant radius. In order to further improve its scale and affine invariant capability, each patch can be we rectified to canonical orientation following the coordinate transform in [37].In order that image patch representation be compact and highly distinctive, an additional principle component analysis (PCA) can be performed on normalized MSER patches

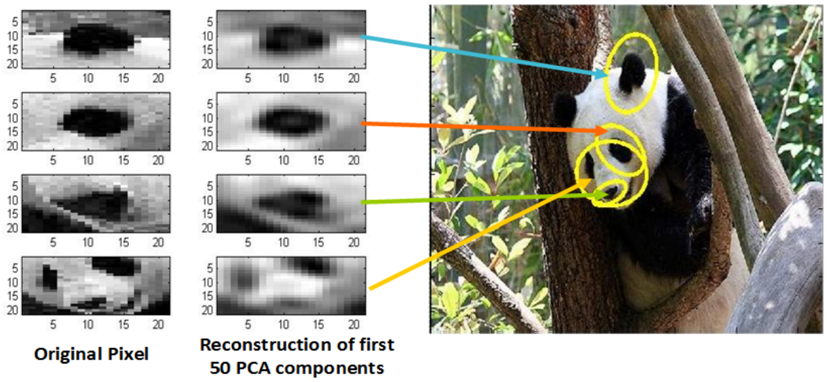


Figure 3.6: Represent images as bag of local features

and provide a robust and compact representation of image regions. More specifically, a covariance matrix can be constructed for all the MSER normalized patches extracted from training dataset. After that, a set of eigenvectors can be obtained by performing eigen-decomposition on the covariance matrix. The first k principal components (i.e. eigenvectors corresponding to k largest eigen-values)then compose the projection matrix, which enables significant dimension reduction of the MSER features while not losing important details (Fig. 3.6).

SIFT Salient Point Descriptor [63]

The Scale Invariant Feature Transform (SIFT) [63] is a widely used methodto detect and describe the salient points in the image.The SIFT approach is composed of four steps: scale-space extreme detection, key-point localization, orientation assignment and key-point descriptor, in which the first and last step is most important.

In the detection part, a difference-of-Gaussian (DoG) function is used to detect the potential salient points. For the first octave in the scale space, images of different scales are produced by repeatedly convolute the 2D Gaussian function to the input image (the value of σ times $2^{1/2}$ each time). The 2D Gaussian function is:

$$G(x,y,\sigma) = \frac{1}{2\pi\sigma^2} e^{-[\frac{(x^2+y^2)}{2\sigma^2}} \tag{3.7}$$

As shown in Fig. 3.7, the sets of scale space images are shown on the left, which are also called the Gaussian images. After that, adjacent Gaussian images are subtracted to produce the difference-of-Gaussian images which are shown on the right. For every next octave, the Gaussian images are down-sampled by a factor of 2, and the process repeated.

In the scale space, by convolute image with a 2D Gaussian function, image structures of spatial size smaller than σ will largely been smoothed away (the value of σ times $2^{1/2}$ each scale). Therefore, in the difference-of-Gaussian images, a local extreme represents that a significant structure with spatial size σ has been smoothed away and eliminates in the next scale level, in another word, the a local extreme in the difference-of-Gaussian images is corresponding to a salient point. For each octave, the local extreme in the scale space can be detected by comparing a pixel to its 26 neighbor pixels in current and adjacent scales (Fig. 3.8). If one pixel is more extreme than its 26 neighbor pixels, it is marked as a potential salient point.

After detecting the local extremes in the scale space, the potential salient points are localized in the original image; their main orientations are also determined according to the image gradients.After that, image patches containing the salient points are rotated to the canonical orientation and divided into 4×4cells. In each cell, the gradient magnitudes at 8 different orientations are calculated, which then form a 128-dimension SIFT descriptor

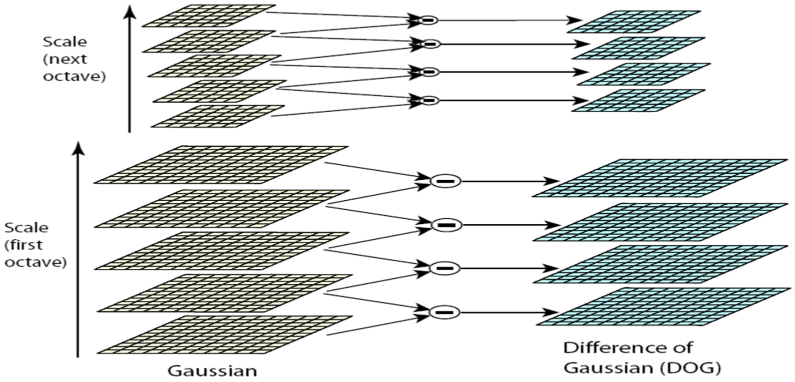


Figure 3.7: The Gaussian images and the difference-of-Gaussian images [63]

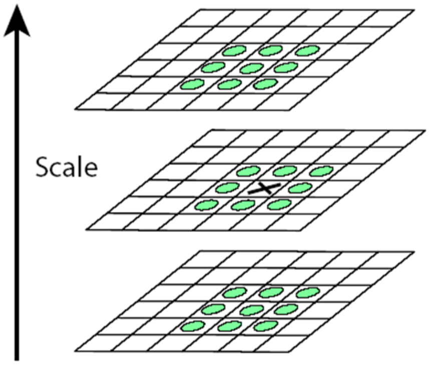


Figure 3.8: Detection of local extreme in difference-of-Gaussian images [63]

(Fig. 3.9).

Compared to other local descriptors, the SIFT descriptor is more robust and invariable to rotation and change in scale and luminance [69], [72]. In my research, each extracted 128-dimension SIFT descriptoris named as a ‘visual token’.

The SIFT approach is in essence an object-based image representation approach as it strongly depends on the existence of reliable distinctive sub-structures, which may to some extend be considered as objects [90]. It should be noted that the SIFT approach may be-

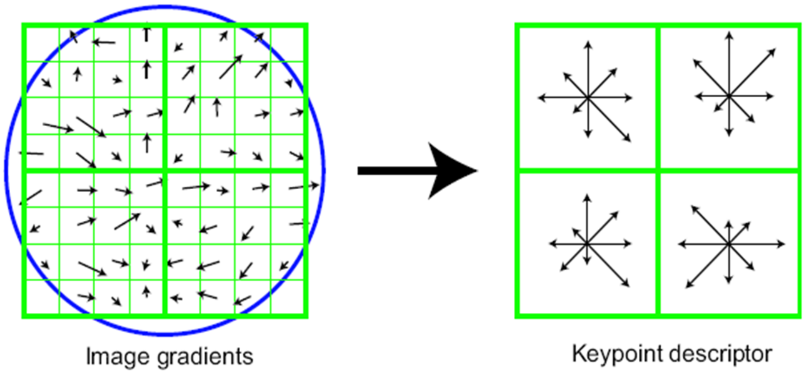


Figure 3.9: The SIFT descriptor of salient points (2×2 cells) [63]

come inefficient when dealing with complex background with too much texture, because the increased number of key-points in a complex background may take longer time to process. Also, since SIFT descriptors are derived from clustering (categorizing) the key-points, it may become less discriminative due to the quantification and increase of cluster number [3]. It has been suggested that researcher may combine both point features (such as SIFT descriptor) and region features (which are derived from local homogeneous parts in objects) to improve the representation of image content [106], [61]. In particular, our study [17] has shown that the Maximally Stable Extremal Region (MSER) [67], a widely used region-based image saliency model, can be used as an effective complementary feature of SIFT descriptors in providing robust representation of image scene.

Bag of Visual Words

Recently, the significance of dense local features has been greatly enriched as the concept 'bag-of-visual-words' being introduced [91].

As mentioned, an image document can be constantly represented by hundreds of key-points

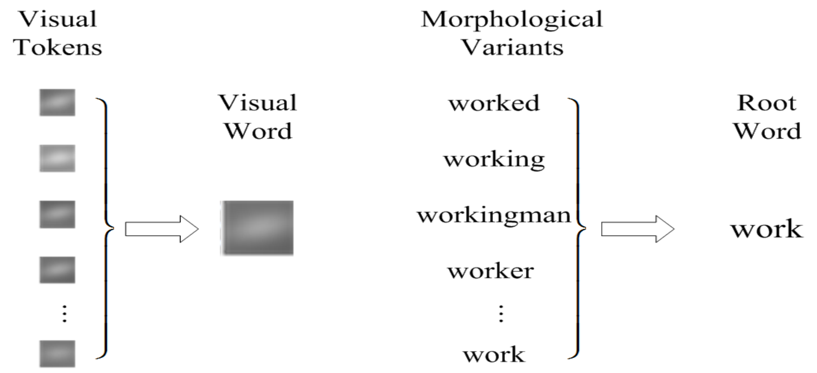


Figure 3.10: An analogy between the visual tokens grouping and the text morphological processing

or unique microstructures. Image patches containing key-points can then be quantified based on affine invariant local descriptors [13], [54], [69], [63]. In an image document, each unique descriptorcan be regarded as a 'visual token', which plays similar roles as its textual counterpart (token) in a text document. Based on the extracted 'visual tokens', researchers further proposed the idea of assigning all the patch descriptors into clusters (Fig. 3.10)and build a vocabulary/codebook of 'visual words' (Fig. 3.11) for a specific image dataset [91], [2].Therefore, 'visual words' can be regarded as a visual analog of text words in animage document.There is no consensus on the actual size of visual word vocabulary. According to the size of image dataset, the visual word size used in existing works varies from several hundred to thousands or tens of thousands [91], [3], [51].

Inspired by the success of vector-space model of the text document representation, the 'bag-of-visual-words' approach usually converts images into vectors of visual words (Fig. 3.12) based on their frequency [91], [3].Many effective text mining and information retrieval algorithms like tf-idf weighting, stop word removal and feature selection have been

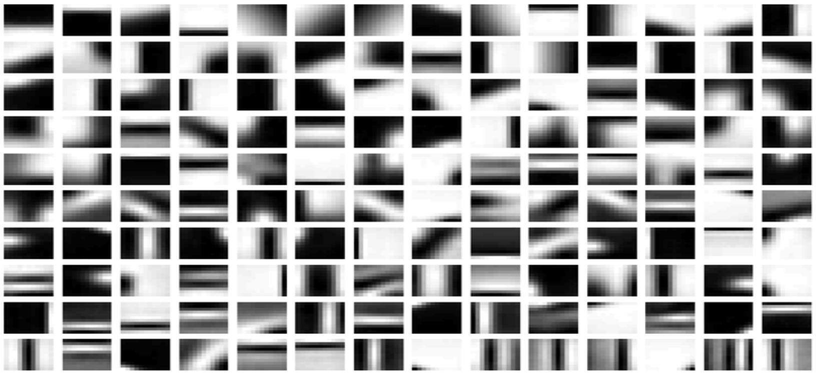


Figure 3.11: Codebook (partial) of visual words [2]

applied to the vector-space model of visual-words. Problems such as how vocabulary size and term weighting schemes affect the performance of ‘bag-of-visual-words’ representation are also studied in recent research works [51], [3].

Despite the success of ‘bag-of-visual-words’ in recent studies, however, there are two problems to be concerned. Firstly, since the ‘bag of visual words’ approach represents an image as an unordered collection of local descriptors, the resulting vector-space model offers us little insight about the spatial constitution of the image. Secondly, as most local descriptors are based on the intensity information of images, no color information is used. There have been some works aiming at incorporating spatial information and color information in the ‘bag-of-visual-words’ model, such as dividing an image into equal-sized rectangular regions and computing visual word frequency from each region [3], using multi-scale spatial grids for locally order-less description of visual words [58] and using Color SIFT descriptors (Fig. 3.13, which has the same framework in the image intensity space as SIFT, while in the color space, the gradient magnitude and orientation are replaced by saturation and hue from HSI color space) [98]. However, to the best of our knowledge,

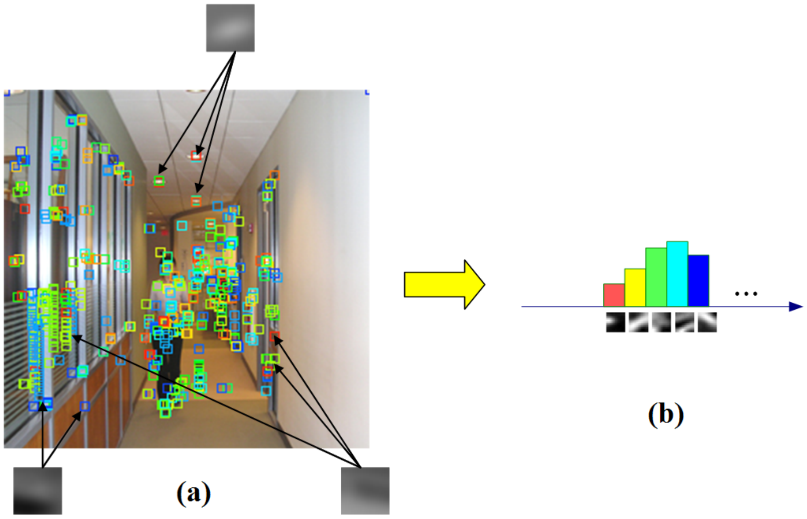


Figure 3.12: (a) illustration of extracted visual words, (b) vector of visual word term frequencies

there hasn't been any study combining visual-word features with the spatial constitution of image content.

3.2.3 Spatial Weighting for the 'Bag-of-Visual-Words'

In summary, the context-based image representations are able to represent the spatial constitution of image content. But they are sensitive to the change of light, color and point of view and unable to represent the specific characters of individual local structures in images. The saliency models, on the other hand, are affine invariant, robust to the change of light, color and point of view. However, they provide little insight about the spatial constitution and color information of the image content. With this consideration, I have developed a spatial weighting scheme for the 'bag-of-visual-words' image feature [107]. The 'bag-of-visual-words' feature represents a set of unique structural elements in images,

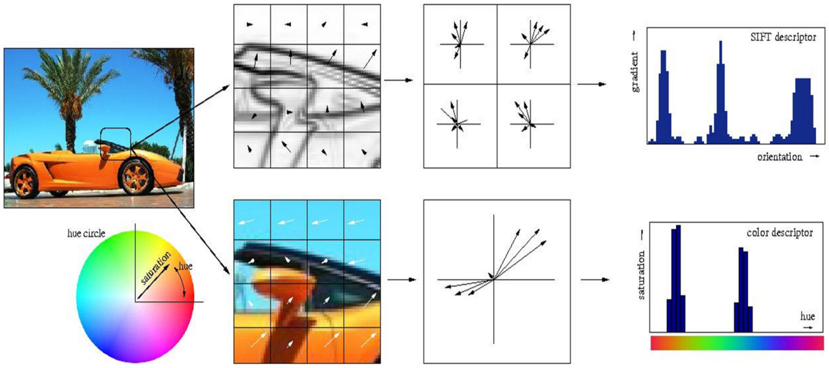


Figure 3.13: The framework of color-SIFT [98]

which are robust to rotation, scaling and luminance changes; it has exhibited very good performance in content-based image retrieval (CBIR). However, since the ‘bag of visual words’ approach representsan image as an unordered collection of local descriptors which only use the intensity information, the resulting model provides little insight about the spatial constitution and color information of the image.

In my approach, I use Gaussian mixture model (GMM) to provide spatial weighting for ‘bag-of-visual-words’: the spatial constitution of image content is represented as a mixture of n Gaussians in the feature space and decomposed the image into n regions (Fig. 3.14c). The spatial weighting scheme is achieved by weighting the ‘bag-of-visual-words’ according to the probability of each visual word belonging to each of the regions in the image. Assuming that local descriptors obtained from the salient point set $\{j_1, j_2, \cdots, j_M\}$ have been assigned to visual word V, then the summation of $p(i|j_k), k = 1, \cdots, M$ will indicate the contribution of visual word V to region i. Therefore, the weighted term frequency of V with regard to region i can be defined as:

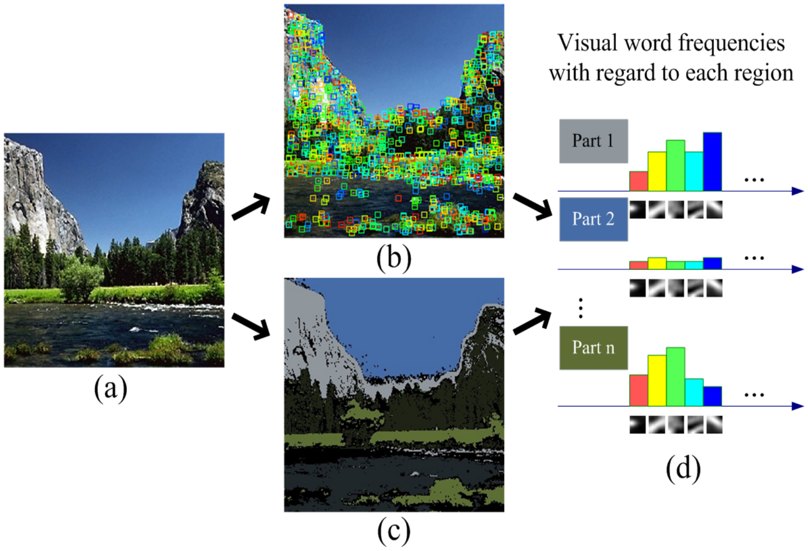


Figure 3.14: The framework for spatial weighting [18] (a) original image, (b) extracted visual words, (c) GMM modeling result, (d) spatial weighted visual word frequencies

$$tf_V = \sum_{k=1}^{M} p(i|j_k) \tag{3.8}$$

Supposing that d_i and d_j are two D-dimensional (D equals the vocabulary size of visual words) vectors of spatial weighted visual word frequencies (Fig. 3.14d), which come from region i and region j, respectively. Then the most natural way to measure the similarity between vectors d_i and d_j is using the cosine similarity (eq. (3.9)).

$$Sim_{cosine}(d_i, d_j) = \frac{\mathbf{d}_i^t \mathbf{d}_j}{||\mathbf{d}_i|| \, ||\mathbf{d}_j||} \tag{3.9}$$

Assuming that I_q is the query image and I_r is an image from the retrieval set. For each region in image I_q, we find its closest region in image I_r based on eq. (3.9). After that, the

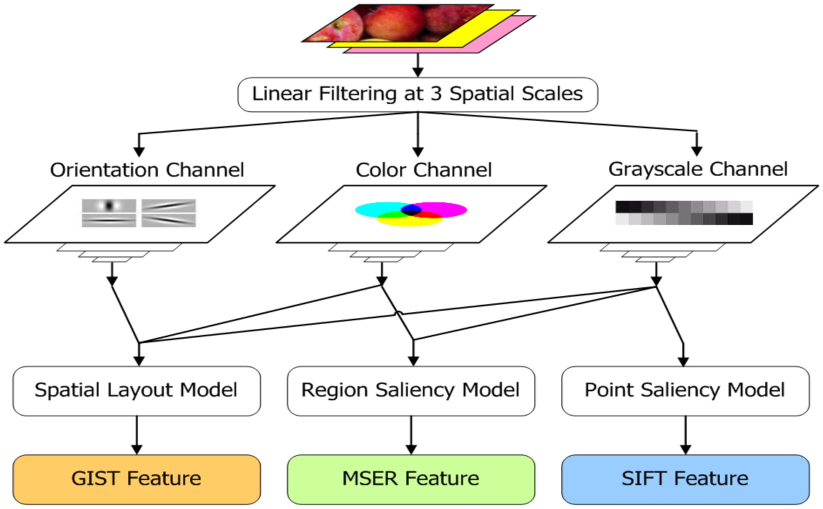


Figure 3.15: Proposed image representation framework

image-level similarity is obtained by taking the average of the pair-wise similarity between regions in I_q and their closest regions in I_r.

3.2.4 An Integrated Image Representation Framework

In our approach, we proposed an integrated image representation framework, in which we utilize the saliency model (including SIFT descriptors and MSER feature) as a complement part of context-based image representations. Our motivation comes from the fact that the mechanism of human visual perception allows for very rapid holistic image analysis to provide a coarse context of image scene (special layout model), yet it also give rise to a small set of candidate salient locations in a scene (saliency model) that needs to be intensively studied.

According to the comprehensive studies of human vision perception [43], the human visual processing system is able to make decision to focus its attention on a small set of interesting (salient) parts in the view field within a very short time (less than 200ms). During the same period, it also extracts enough visual information to accurately recognize functional and categorical properties of the scene. Such a biological mechanism (in which a very rapid holistic image analysis gives rise to a small set of salient parts in a scene) motivates us to develop an integrated image representation framework as Fig. 3.15. The context of image scene and salient parts are two complementary components in image representation. Computing contextual image representation involves accumulating image statistics over the entire scene, while finding salient parts requires finding image regions that stand out significantly different from their neighbors. In our framework, we choose to enrich the contextual image representation by two different saliency models, i.e. the region-based image saliency model the Maximally Stable Extremal Region (MSER) features and point-based image saliency model - SIFT features.

Specifically, an input image is filtered and down-sampled to produce four spatial scales. Then, for each spatial scale, we extract information from a set of low-level visual feature channels (orientation channel, color channel and grayscale channel). Within the low-level feature processing procedure, as suggested in [90], we extract the contrast information from the color channel as it is more robust and show better invariance under different lighting conditions. For the orientation channel, we employ Gabor filters to the image intensity and extract orientation information at four different angles. After that, three low-level image feature channels are merged to produce the image context representations. In our framework, we proposed to use the GIST feature [90] for image context representations. The GIST feature is a low dimensional holistic representation of image content. Specifically, we divide each image into a 4 by 4 grids. For each grid, we incorporate all the image feature channels to generate a vector representation of that location. By merging all localized vector together, we obtain a raw gist feature vector for the whole scene. Finally, we will perform principal component analysis to reduce the dimension of gist feature vectors to a more practical number.

In our framework, we represent the region-based image saliency model by the Maximally Stable Extremal Region (MSER) features. Specifically, we extend MSER detector to multiple scales by constructing Gaussian pyramid and applying MSER detection separately in each resolution level. After MSER detection, each detected elliptical region is normalized to circular patch of constant radius. In order to further improve its scale and affine invariant capability, we rectify each patch to canonical orientation following the coordinate transform in [37]. Specifically, we chose to perform principle component analysis (PCA) on MSER normalized patches toprovide a robust and compact representation of image regions.More specifically, we construct covariance matrix for a total of 140,000 MSER normalized patches extracted from training dataset, each of which is a 21×21 dimensional vector of image intensity. Then, we perform eigen-decomposition on the covariance matrix to obtain the eigenvectors. We then obtain the projection matrix which is composed of first k principal components, i.e. eigenvectors corresponding to k largest eigen-values. In this way, we build the eigenspace for all the MSER normalized patches in our training dataset. To representa new MSER patch, we simply multiply its 21×21 dimensional vector with the projection matrix to obtain its k dimensional projection. After extensive testing, we set k=50 (which means that all the MESR features are represented as 50-dimensional vectors). Reconstruction result in Fig. 3.6 shows that, when using the first 50 principle components, we are able to achieve significant dimension reduction of the MSER features without losing important details.

We represent the point -based image saliency model by SIFT features. Specifically, we extract and index SIFT features as follows. Firstly, we employ the Difference-of-Gaussian (DoG) salient point detector [63] to detect salient points from images. The detection is achieved by locating scale-space extreme points in the difference-of-Gaussian images and the main orientations of salient points are determined by image gradient. Then, image patches containing the salient points are rotated to a canonical orientation and divided into 44cells. In each cell, the gradient magnitudes at 8 different orientations are calculated. Consequently, each salient point is described by a 128-dimensional SIFT descriptor. In this way, each image in the training dataset is represented as a set of SIFT descriptors. After that, the K-mean clustering is performed to quantize all the extracted SIFT descriptors and produce a finite dictionary of appearance patterns called 'code-book of visual words', with each cluster center as a unique 'visual word'. Finally, the indexing of SIFT features is accomplished by computing the term frequency of visual words with respect to each image document.

3.3 PROBABILISTIC TOPIC MODEL FOR CO-EXISTING IMAGE FEATURE AND ANNOTATION

High quality text descriptions of images play a vital role as training and benchmarking data in developing and evaluating an automatic image annotation system. So the first issue of this research problem is to build a benchmark dataset for the purpose of training an automatic image annotation and retrieval system. We propose we propose to associate image captions in biomedical literatures with semantic concepts from Unified Medical Language System (UMLS) and enrich image in ImageNet dataset by text descriptions from Wikipedia. The second issue of this research problem is proposing an effective model to study the correlation between image and text descriptions. In the data mining and information retrieval community, there are many studies focusing on probabilistic topic models to study the correlation between image and text descriptions such as the Correspondence LDA (CorrLDA) model [14], [77]. In my research, a hierarchical probabilistic model with background distribution (HPB) and the probabilistic topic-connection (PTC) model are introduced to enables more effective and robust modeling of the co-existing image features and annotations.

3.3.1 Probabilistic Models for Topic Learning from Images and Captions in Online Biomedical Literatures

Recent researches in biomedical and life sciences produce hundreds of thousands of digital publications each year. Although there are several online digital archives (such as PubMed Central) available for full-text biomedical literatures retrieval, however, it's still very difficult for users to query and retrieve biomedical figures from online publications. Although the CorrLDA model [14] provides a way to learn semantic topics from the co-occurrence patterns of caption words and extracted image features, however, extensive studies of Corr-LDA model show that the discovered topics of CorrLDA model can be overwhelmed by several background words that frequently appear in the database. With this consideration, a hierarchical probabilistic topic model with background distribution is presented, in which there is a switch variable that allows the model to decide whether a word is generated by the background topic of or one of the individual topics.

In this section, I will firstly describe the procedure of preprocessing and indexing of biomedical figures. After that, I will present the extended CorrLDA model and the hierarchical probabilistic topic model with background distribution. Finally, I willprovide the collapsed Gibbs sampling algorithms for inference and learning the proposed probabilistic models.

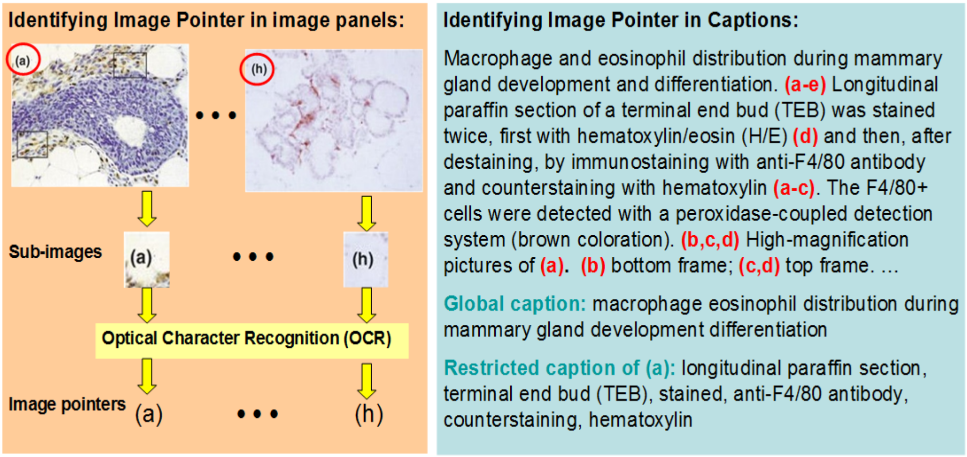
Preprocessing of Biomedical Figures

In our research, we deal with biomedical figures downloaded from the PubMed Central web pages. Generally, a biomedical figure involves two parts, a single image composed with one or multiple image panels (sub-images) and the corresponding captions. Therefore, the preprocessing section of biomedical figures has two parts, the image processing part and the caption processing part.

Within the downloaded biomedical figures, images are segmented into several individual image panels. It should be pointed out that there are image panels which contain flow charts or diagrams. These image panels do not carry substantial visual content. Therefore, they are filtered out using basic region segmentation method.

In caption texts, there are some parenthesized expressions refer to specific image panels. Most of them are simply composed of single letter such as (A), (b) or letters connected by conjunction, such as (a and b), (b,c) and (a-c). We refer to these parenthesized expressions as image pointers (as marked by red color in Fig. 3.16 b). We develop a set of rules to extract these regular image pointers in captions, which is similar to the HANDCODE2 method in [100].

Image pointers are commonly placed in some important positions (such as upper left and


Identifying Image Pointer in image panels:
(a)
(h)
Sub-images
(a)
(h)
Optical Character Recognition (OCR)
Image pointers
(a)
(h)
Identifying Image Pointer in Captions:

Macrophage and eosinophil distribution during mammary gland development and differentiation. (a-e) Longitudinal paraffin section of a terminal end bud (TEB) was stained twice, first with hematoxylin/eosin (H/E) (d) and then, after destaining, by immunostaining with anti-F4/80 antibody and counterstaining with hematoxylin (a-c). The F4/80+ cells were detected with a peroxidase-coupled detection system (brown coloration). (b,c,d) High-magnification pictures of (a). (b) bottom frame; (c,d) top frame. ...

Global caption: macrophage eosinophil distribution during mammary gland development differentiation

Restricted caption of (a): longitudinal paraffin section, terminal end bud (TEB), stained, anti-F4/80 antibody, counterstaining, hematoxylin

(a) Image preprocessing (b) caption preprocessing

Figure 3.16: Biomedical figures preprocessing

lower left corner) of image panels. Therefore, we apply the Asprise OCR Java SDK toolkit for optical character recognition (OCR) in sub-images of image corners (Fig. 3.16 a). The OCR toolkit achieved a moderate precision in our image pointer extraction, which is sufficient for our research. We check the image pointer extraction results and make necessary manual corrections.

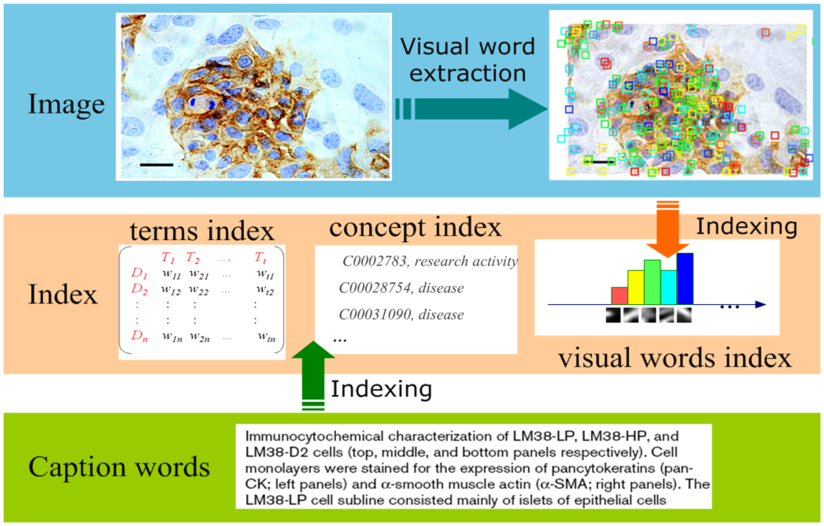
In a figure with multiple image panels, instead of replicating the entire caption to each image panel, we develop a restricted caption scanner to identify restricted captions (Fig. 3.16b) with regard to the image pointer of each image panel. The association of texts and image pointers are determined according to different cases, such as image pointers locate at the beginning of a sentence, preceded by preposition and noun phrases, followed by a clauses, etc. Generally, the undergoing image pointer(s) for captions are disabled when the scanner meets another image pointer or reaches the end of a clause or a sentence. All the texts that don't have any assigned image pointers are regarded as global captions (Fig. 3.16b).

The image panel and captions associated with the same image pointer are named as an image-caption pair. In an image-caption pair, the final caption words are generated via a linear combination of restricted captions and global captions. The combination avoids the over-representation problem and preserves the uniqueness of each individual image panel. Each image-caption pair is assigned a unique ID like 'bcr1011-1a' , in which 'bcr1011' is the PubMed Central article ID, '1' is the number of figure in the article, while 'a' is the name of image pointer of a given image panel.

Image-Caption Pairs Indexing

During the indexing stage, we choose to represent the image content in each image-caption pair as a 'bag-of-visual-word'. First, we adopt the Difference-of-Gaussian (DoG) salient point detector [63] to detect salient points from images. The detection is achieved by locating scale-space extreme points in the difference-of-Gaussian images. The main orientations of salient points are determined by image gradient. Image patches containing the salient points are then rotated to a canonical orientation and divided into 4×4 cells. In each cell, the gradient magnitudes at 8 different orientations are calculated. Consequently, each salient point is described by a 128-dimensional SIFT descriptor. Compared to other local descriptors, the SIFT descriptor is more robust and invariable to rotation and scale/luminance changes [69].

The SIFT descriptors extracted from training images are clustered into 1000 clusters using k-mean clustering to establish a codebook of 'visual words', with each cluster center as a



Immunocytochemical characterization of LM38-LP, LM38-HP, and LM38-D2 cells (top, middle, and bottom panels respectively). Cell monolayers were stained for the expression of pancytokeratins (pan-CK; left panels) and α-smooth muscle actin (α-SMA; right panels). The LM38-LP cell subline consisted mainly of islets of epithelial cells

Figure 3.17: The workflow for image-caption pair indexing

'visual word'. As shown in Fig. 3.17, the image indexing is achieved by computing the term frequency and building index of visual words for each image panel. The indexing of captions results in two parts, the term index and the concept index (Fig. 3.17). The term index is simply obtained by calculating the term frequency of caption words after lemmatizing and stop-word removal. In our approach, the Van Rijsbergen's stop-word lists [99] and the UMLS biomedical stop-word list [10] are used to remove non-content-bearing terms.

The concept index is achieved by calculating the term frequency of concepts according to the results of concept extraction. In biomedical ontology, a concept carries a unique meaning and represents a set of synonymous terms. For example, C0006149 is a concept about the benign or malignant neoplasm of the breast parenchyma in Unified Medical Language System (UMLS) [10]. It represents a set of synonyms including Breast Neoplasm, Breast Tumor, tumor of the Breast and Neoplasm of the Breast. Compared to individual words and multiple word phrases, a concept is more meaningful, therefore, used as indexing terms in large-scale biomedical literatures. In our approach, we adopt MaxMatcher [112], a dictionary-based biological concept extraction tool, to extract UMLS concept from captions.

Topic Modeling from Biomedical Image-caption Pairs

This section focuses on the problem of learning latent topics from biomedical image-caption pairs. The underlying philosophy is that, an image-caption pair may deal with multiple topics; and the co-occurrence patterns of caption words, visual words and biomedical concepts in this image-caption pair are related to some unseen latent semantic variables, which indicate the presence/absence of specific topics.

In this section, we will present two probabilistic models, one is the extended Correspondence LDA (CorrLDA) model and the other is our proposed hierarchical probabilistic topic model with background distribution (HPB). For clarity of the notations, we name each image-caption pair as a document. Some notations to be used in the two probabilistic models are list as follows: D is the number of documents, T is the anticipated number of latent topics, N_d is the total number of text words in document d, N_c^d denotes the total number of extracted biomedical concepts in document d, while M_d represents the total number of extracted visual words in document d. As mentioned, the Corr-LDA model provides a natural way to learn latent topics from text words and other entities. Therefore, the topic learning problem can be addressed by extending the entities in the Corr-LDA model to visual words and ontology-based biomedical concepts (Fig. 3.18). The differences between our

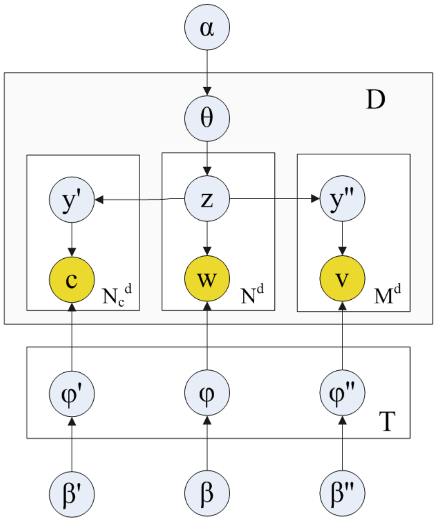


Figure 3.18: The extended Corr-LDA model, yellow cycles represent the observation of words, concepts and visual words

extension and the original Corr-LDA model are two-fold. First, we combine visual words, text captions and ontology-based concepts in one single model. Second, the original model only takes use of global image features such as color and texture, while our extension deals with visual words which are more robust than global image features and have similar statistical properties with text words (which are assumed to fit multinomial distributions). The sampling process for the extended Corr-LDA model is as follows.

```
1.    For the $d^{th}(d = 1 \cdots D)$ document, sample $\theta_d \sim Dir(\alpha)$
2.    For the $t^{th}(t = 1 \cdots T)$ topic, sample $\varphi_t \sim Dir(\beta), \varphi'_t \sim Dir(\beta'), \varphi''_t \sim Dir(\beta'')$
3.    For each of the$N_d$ words $w_i$ in document d:
   a)    Sample a topic $z_i \sim Multi(\theta_d)$
   b)    Sample $w_i|z_i \sim Multi(\varphi_{z_i})$
4.    For each of the $N_c^d$ concepts $c_i$ in document d:
   a)    Sample a topic $y'_i \sim Uniform(z_{w_1}, \cdots, z_{w_{Nd}})$
   b)    Sample $c_i|y'_i \sim Multi(\varphi'_{y'_i})$
5.    For each of the $M_d$ visual words in document d:
   a)    Sample a topic $y''_i \sim Uniform(z_{w_1}, \cdots, z_{w_{Nd}})$
   b)    Sample $v_i|y''_i \sim Multi(\varphi'_{y'_i})$
```

In the first step, a T-dimensional topic-prior vector θ_d is sampled for each document d, with the t-th dimension of the vector represents the prior probability of the t-th topic in d. For each document d, the generative process of the N_d words is achieved by sampling topics from the document-topic multinomial distribution (with Dirichlet prior θ_d) and sampling words from the topic-word multinomial distribution (with Dirichlet prior φ_t). The generative process of the N_c^d concepts and M_d visual words are similar with that of the N_d words; the only difference is that only the topics that associated with the N_d words in document d are used to generate concepts and visual words. Parameters $\alpha, \beta, \beta', \beta''$ are hyper-parameters for the Dirichlet priors. In our approach, we assume symmetric Dirichlet priors, with $\alpha, \beta, \beta', \beta''$ being scalar parameters.

Although the Corr-LDA model is able to learn latent topics from the image-caption pairs and establish direct correlation among words, visual words and concepts, however, after looking into the discovered topics from the data collection, we found several background words appear at the top ranked terms of most discovered topics due to their high frequency. For example, when we use image-caption pairs from online journal: 'Breast Cancer Research' as training data and learn topics using the Corr-LDA model, we found 'breast', 'cancer', 'mammary' are among the top-tanked words of many topics. These words, which we named as 'background words', appear frequently in many topics and take the places of the topic-specific key words. Its necessary to discover these 'background words' from the dataset, otherwise, the topic learning would be less effective.

It should be note that during the caption indexing stage, we have removed the non-content-bearing stopwords according to the Van Rijsbergen's stopword lists [99] and the UMLS stopword list [10]. Obviously, the 'background words' do not belong to regular stopwords. As we have seen, these words carry some contextual information which is shared by most image captions in a biomedical journal. As such 'background words turn to be different from one journal to another, it's better to discover them automatically rather than manually specifying them for each journal.

In [77] the 'SwitchLDA' model is proposed, in which a switch variable is introduced to control the fraction of entities in topics. With similar consideration, we develop a hierarchical probabilistic model with background distribution (HPB model) to capture the background topic z_0. In this model, an additional Binomial distribution λ (with a Beta prior of $\gamma 1$ and $\gamma 2$) was incorporated to control the switch variable x (Fig. 3.19), which decides whether a term should be drawn from a background topic z_0 or a regular latent topic z_i. At this stage, we are not clear whether the background topics (Fig. 28) are related to certain

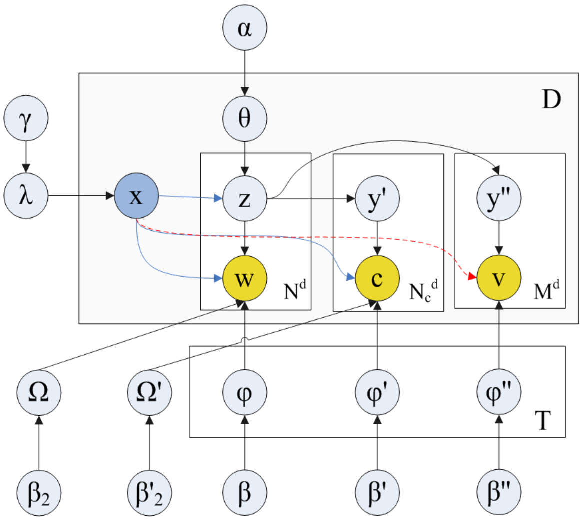


Figure 3.19: The hierarchical probabilistic model with background distribution (HPB), the red dash line denotes a variation of HPB model

image content. Therefore, we also present a variation of the HPB model (HPB2) for testing. The generative process is as follows:

1. For the $d^{th}(d = 1 \cdots D)$ document, sample $\theta_d \sim Dir(\alpha)$ and $\lambda_d \sim Beta(\gamma_1, \gamma_2)$

2. For the $t^{th}(t = 1 \cdots T)$ topic, sample $\varphi_t \sim Dir(\beta), \varphi'_t \sim Dir(\beta'), \varphi''_t \sim Dir(\beta'')$;for background topic, sample $\Omega \sim Dir(\beta_2)$ and $\Omega' \sim Dir(\beta'_2)$

   Variation(for HPB2 model):

   For background topic,sample $\Omega'' \sim Dir(\beta''_2)$

3. For each of the N_d words w_i in document d:

   a) Sample a switch $x_i \sim Bernoulli(\lambda_d)$

   b) If $x_i = 0$, Sample $w_i|z_0 \sim Multi(\Omega)$

   c) if $x_i = 1$, sample a topic $z_i \sim Multi(\theta_d)$ and sample $w_i|z_i \sim Multi(varphi_{z_i})$

4. For each of the N_c concepts c_i in document d:

   a) Sample a topic $y'_i \sim Uniform(z_{w_1}, \cdots, z_{w_{Nd}})$

   b) if $y'_i = z_0$, Sample $c_i|y'_i \sim Multi(\Omega')$

   c) if $y'_i = z_i(i = 1, \cdots, T)$, sample $c_i|y'_i \sim Multi(\varphi'_{y'_i})$

5. For each of theM_d visual words v_i in document d:

   a) Sample a visual topic $y''_i \sim Uniform(z_{w_1}, \cdots, z_{w_{Nd}})$

   b) if $y''_i = z_0$, repeat (a),

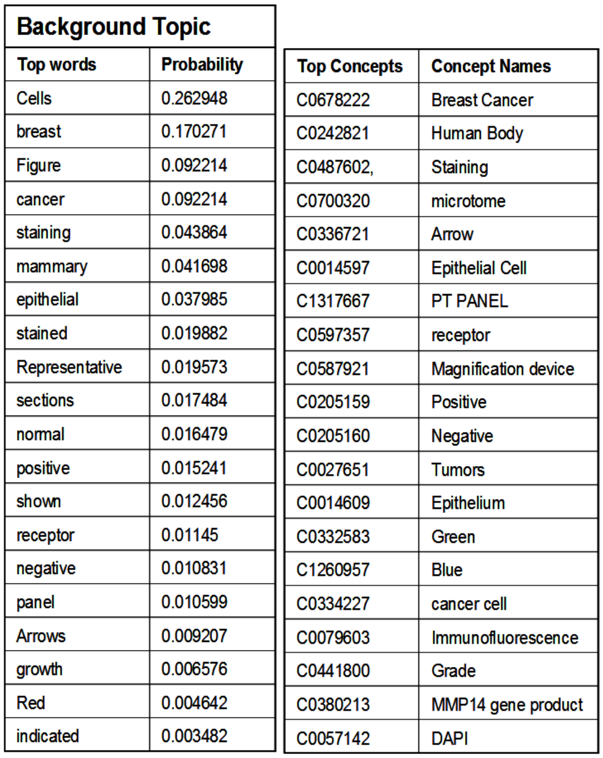
   c) if $y''_i = z_i(i = 1, \cdots, T)$, sample $v_i|y''_i \sim Multi(\varphi''_{y'_i})$

   Variation(for HPB2 model):

   a) Sample a topic $y''_i \sim Multi(\theta_d)$

   b) if $y''_i = z_0$, sample $v_i|y''_i \sim Multi(\Omega'')$

   c) if $y''_i = z_i(i = 1, \cdots, T)$, sample $v_i|y''_i \sim Multi(\varphi''_{y'_i})$



Background Topic

Top words	Probability
Cells	0.262948
breast	0.170271
Figure	0.092214
cancer	0.092214
staining	0.043864
mammary	0.041698
epithelial	0.037985
stained	0.019882
Representative	0.019573
sections	0.017484
normal	0.016479
positive	0.015241
shown	0.012456
receptor	0.01145
negative	0.010831
panel	0.010599
Arrows	0.009207
growth	0.006576
Red	0.004642
indicated	0.003482

Top Concepts	Concept Names
C0678222	Breast Cancer
C0242821	Human Body
C0487602,	Staining
C0700320	microtome
C0336721	Arrow
C0014597	Epithelial Cell
C1317667	PT PANEL
C0597357	receptor
C0587921	Magnification device
C0205159	Positive
C0205160	Negative
C0027651	Tumors
C0014609	Epithelium
C0332583	Green
C1260957	Blue
C0334227	cancer cell
C0079603	Immunofluorescence
C0441800	Grade
C0380213	MMP14 gene product
C0057142	DAPI

Figure 3.20: Top-ranked words and concepts in background topic of journal ‘Breast Cancer Research’

In the proposed model, λ is the Bernoulli parameter for switch variable x. In our experiment, we assume symmetric priors and set $\alpha = 0.1, \beta = \beta' = \beta'' = 0.01, \gamma 1 = \gamma 2 = 0.5$. For clarity, we call the variation of HPB model (in gray color) as HPB2 model. In the HPB model, visual words has nothing to do with the background topic, while in HPB2 model, the presence of background topic z_0 in the caption words of document d is used to generate visual words, which results in direct correlation between visual words and the background topic.

Collapse Gibbs Sampling for Proposed Topic Models

The model estimation is achieved via the Collapse Gibbs Sampling procedure [95], which iteratively estimates the posterior probability conditioned on current entity-topic assignment and adopts a Monte Carlo process to determine the assignment of entity-topic in the next iteration.

Some notations to be used in Collapse Gibbs Sampling are list as following: W accounts for the vocabulary size of indexed words in the testing dataset; N_W denotes the total number of indexed words while W', N'_W and W",N''_W represent the vocabulary size and the total number of concepts and visual words, respectively.

Given the generative process, the next step is to compute the word-topic posterior probability, which is: $p(z_{wi} = j|w_i, \mathbf{w}_{-i}, \mathbf{z}_{-wi}) \propto p(w_i|z_{wi} = j, \mathbf{w}_{-i}, \mathbf{z}_{-wi}) \cdot p(z = j|\mathbf{w}_{-i}, \mathbf{z}_{-wi})$. This probability is intractable, however, it can be approximated by integrating out (collapsing) all the latent variables φ_j and θ_d separately, which is:

$$\begin{aligned} p(w_i|z_{wi} = j, \mathbf{w}_{-i}, \mathbf{z}_{-wi}) &= \int p(w_i|z = j, \varphi_j, \mathbf{w}_{-i}, \mathbf{z}_{-wi}) p(\varphi_j|\mathbf{w}_{-i}, \mathbf{z}_{-wi}) \mathrm{d}\varphi_j \\ &\propto E(p(\varphi_j|\mathbf{w}_{-i}, \mathbf{z}_{-wi}) \sim Dir(\beta + n^{wi}_{-i,j})) = \frac{\beta + n^{wi}_{-i,j}}{W\beta + n^{wi}_{-i,j}} \end{aligned} \tag{3.10}$$

$$\begin{aligned} p(z = j|\mathbf{w}_{-i}, \mathbf{z}_{-wi}) &= \int p(z = j|\theta_d) \cdot p(\theta_d|\mathbf{w}_{-i}, \mathbf{z}_{-wi}) \mathrm{d}\theta_d \\ &\propto \frac{\alpha + n^d_{-i,j}}{T\alpha + n^d_{-i,j}} \end{aligned} \tag{3.11}$$

Therefore, posterior probability for current word w_i is:

$$p(z_{wi} = j|w_i, \mathbf{w}_{-i}, \mathbf{z}_{-wi}) \propto \frac{\beta + n^{wi}_{-i,j}}{W\beta + n^{wi}_{-i,j}} \cdot \frac{\alpha + n^{d}_{-i,j}}{T\alpha + n^{d}_{-i,j}} \tag{3.12}$$

In which $n^{wi}_{-i,i}$ (-i denotes that current word w_i is removed) is the total number of times word w_i being assigned to topic j except for current one,$n_{-i,i}$ is the summation of $n^{wi}_{-i,i}$, and $n^{d}_{-i,i}$ is the total number of words in document d assigned to topic j except for current word. Based on sampled topic variables for each word w_i, the posterior probabilities for visual word-topic and concept-topic can be approximated in similar formations. For simplicity, we give their posterior probabilities in a uniform expression, which is:

$$p(\breve{z}_i = j|\breve{w}_i = v, \breve{\mathbf{z}}_{-\mathbf{i}}, \breve{\mathbf{w}}_{-\mathbf{i}}, \mathbf{z}, \breve{\beta}) \propto \frac{n_j}{N_w} \cdot \frac{\breve{\beta} + n^{\breve{w}_i}_{-i,j}}{\breve{W}\breve{\beta} + n^{\breve{w}_i}_{-i,j}} \tag{3.13}$$

In which n_j is the total number of words in document d assigned to topic j; N_d is the total number of words in document d;$n^{\breve{w}_i}_{-i,i}$ is the total number of entities (concepts /visual words) assigned to topic j except for current entity: $\breve{w}_i$. For concepts, we have: $\breve{W} = W', \breve{\beta} = \beta'$, while for visual words, $\breve{W} = W'', \breve{\beta} = \beta''$.

Similar to the Gibbs sampling procedure in Section 4.1, we derive the sampling equation for proposed HPB model as follows, which allow for joint sampling of the topic variables and the switch variable x for each word w_i:

$$p(x_{w_i} = 0, z_{w_i} = 0|w_i, \mathbf{w}_{-i}, \mathbf{z}_{-wi}, \mathbf{x}_{-wi}) \propto \frac{N^0_{d,-i} + \gamma}{N_{d,-i} + 2\gamma} \cdot \frac{\beta_2 + n^{w_i}_{-i,0}}{W\beta_2 + n^{w_i}_{-i,0}} \tag{3.14}$$

$$p(x_{w_i}=1, z_{w_i}=j|w_i, \mathbf{w}_{-i}, \mathbf{z}_{-wi}, \mathbf{x}_{-wi}) \propto \frac{N^1_{d,-i}+\gamma}{N_{d,-i}+2\gamma} \cdot \frac{\beta+n^{w_i}_{-i,j}}{W\beta+n^{w_i}_{-i,j}} \cdot \frac{\alpha+n^{d}_{-i,j}}{T\alpha+n^{d}_{-i,j}} \tag{3.15}$$

In which $N^0_{d,-i}$ and $N^1_{d,-i}$ are the total number of words (except for current word w_i) assigned to background topic and regular latent topics in document d. In eq. (3.14), $n^{wi}_{-i,0}$ denotes the number of times word w_i being assigned to background topic except for current one, while $n_{-i,0}$ is the summation of $n^{wi}_{-i,0}$. In eq. (3.15), $n^{wi}_{-i,i}$ is the total number of times word w_i being assigned to topic j except for current one, $n_{-i,0}$ is the summation of $n^{wi}_{-i,i}$,and $n^{d}_{-i,i}$ is the total number of words in document d assigned to topic j except for current word. The sampling equations or concept and visual words have two different cases. For the HPB model, we have:

$$p(x_i=0, y'_i=0|c_i, \mathbf{c}_{-i}, \mathbf{y'}_{-i}, \mathbf{w}, \mathbf{z}) \propto \frac{N^0_d}{N_d} \cdot \frac{\beta'_2+n^{c_i}_{-i,0}}{W'\beta'_2+n^{c_i}_{-i,0}} \tag{3.16}$$

$$p(x_i=1, y'_i=j|c_i, \mathbf{c}_{-i}, \mathbf{y'}_{-i}, \mathbf{w}, \mathbf{z}) \propto \frac{N^1_d}{N_d} \cdot \frac{n_j}{N^1_d} \cdot \frac{\beta'+n^{c_i}_{-i,j}}{W'\beta'+n^{c_i}_{-i,j}} \tag{3.17}$$

$$p(y''_i=j|v_i, \mathbf{v}_{-i}, \mathbf{y''}_{-i}, \mathbf{w}, \mathbf{z}) \propto \frac{n_j}{N^1_d} \cdot \frac{\beta''+n^{v_i}_{-i,j}}{W''\beta''+n^{v_i}_{-i,j}} \tag{3.18}$$

In which N^0_d and N^1_d are the total number of words assigned to background topic and regular latent topics in document d. $n^{c_i}_{-i,i}$ is the total number of times concept c_i being assigned to topic j except for current one, while $n^{v_i}_{-i,i}$ is the total number of times visual word v_i being assigned to topic j except for current one. For the variation of HPB model (i.e. the HPB2 model), we have a uniform expression of posterior probabilities for both concept and visual words:

$$p(x_i = 0, \breve{z}_i = 0 | \breve{w}_i, \breve{\mathbf{c}}_{-i}, \breve{\mathbf{w}}'_{-i}, \mathbf{w}, \mathbf{z}, \breve{\beta}) \propto \frac{N_d^0}{N_d} \cdot \frac{\breve{\beta}'_2 + n_{-i,0}^{\breve{w}_i}}{\breve{W}'\breve{\beta}'_2 + n_{-i,0}^{\breve{w}_i}} \tag{3.19}$$

$$p(x_i = 1, \breve{z}_i = j | \breve{w}_i, \breve{\mathbf{c}}_{-i}, \breve{\mathbf{w}}'_{-i}, \mathbf{w}, \mathbf{z}, \breve{\beta}) \propto \frac{N_d^1}{N_d} \cdot \frac{n_j}{N_d^1} \cdot \frac{\breve{\beta}' + n_{-i,j}^{\breve{w}_i}}{\breve{W}'\breve{\beta}' + n_{-i,j}^{\breve{w}_i}} \tag{3.20}$$

3.3.2 Probabilistic Topic-Connection Model for Co-existing Image Features and Annotations

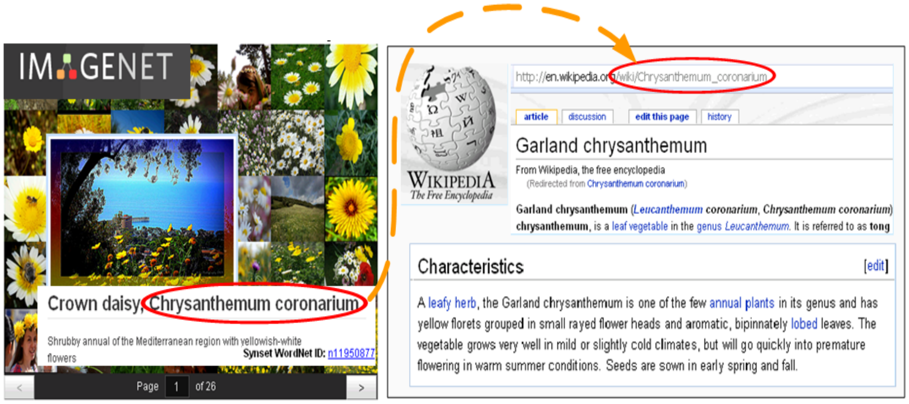
In automatic image annotation, how to bridge over the 'semantic gap' [6]between user and image features is a major challenge. Specifically, its important to identify sets of image features that show strong semantic correlations with textual image descriptions. With this consideration, probabilistic Topic-Connection (PTC) model in this section. We also describe the procedure of enriching the text description by 3rd party knowledge base (Wikipedia). The collapse Gibbs sampling algorithms for inference and learning proposed probabilistic models are presented at the end of this section.

ImageNet Textual Description Enrichment

The Labeled image datasets such as Caltech 101/256 Categories [1], [39], PASCAL [66], LabelMe [87] have been popular with the computer vision community as training datasets. The recently established ImageNet dataset [25] provides large scale ontology of image that is built upon the WordNet Structure [20]. Organized in a hierarchical structure, images in the ImageNet dataset are grouped into sets of cognitive synonyms (synsets), each expressing a distinct semantic concept. Due to its completeness and accuracy, the ImageNet dataset may also serves as a benchmark image dataset. One problem with ImageNet dataset is that it still lacks of comprehensive text descriptions for image data. Therefore, in our research, we utilize Wikipedia as external knowledge source and enrich the ontology structure of ImageNet database with comprehensive and highly-reliable text descriptions from Wikipedia articles.

According to its latest release, ImageNet hosts a total of 15589 synsets(sets of cognitive synonyms, each expressing a distinct semantic concept) of WordNet, with an average of 50-500 images under each synset [25]. The fact that a majority of synsets coincide with Wikipedia entries inspires us to enrich the image hierarchy in ImageNet dataset with high quality text descriptions from Wikipedia articles to provide benchmarking data set for automatic image annotation. Wikipedia is one of the most comprehensive and well-formed electronic knowledge repositories on the web with millions of articles contributed collaboratively by professional subjects. Because of its reliability, accuracy and neutral point of view, Wikipedia has been exploited as external knowledge source in many application of text mining [46], [103], [47]. Although Wikipedia is different from standard WordNet ontology, which is backed up by structured thesaurus, however, each article in Wikipedia only describes one single concept under a hierarchical categorization system. Therefore, the title of each article (which is a succinct phrase) still resembles an ontology term. This feature makes it possible to map an ImageNet synset to a Wikipedia article (Fig. 3.21), which in turn provides text descriptions for images under this synset.

In learning unambiguous semantic topics from text descriptions, polysemies and synonyms are the major barrier. In our previous work [18], we use both ontology-based biomedical concepts and single-word features to overcome the polysemy and synonym problems in biomedical literatures. In public domain, where ontology-based concept is not available



Crown daisy, Chrysanthemum coronarium

Shrubby annual of the Mediterranean region with yellowish-white flowers

Synset WordNet ID: n11950877

http://en.wikipedia.org/wiki/Chrysanthemum_coronarium

Garland chrysanthemum

From Wikipedia, the free encyclopedia
(Redirected from Chrysanthemum coronarium)

Garland chrysanthemum (*Leucanthemum* **coronarium**, ***Chrysanthemum coronarium***) **chrysanthemum**, is a leaf vegetable in the genus *Leucanthemum*. It is referred to as **tong**

Characteristics

A leafy herb, the Garland chrysanthemum is one of the few annual plants in its genus and has yellow florets grouped in small rayed flower heads and aromatic, bipinnately lobed leaves. The vegetable grows very well in mild or slightly cold climates, but will go quickly into premature flowering in warm summer conditions. Seeds are sown in early spring and fall.

(a) ImageNet dataset: images synset "Chrysamthemumcoronarium" (b) Wikipedia article matched the synset

Figure 3.21: Graphical illustration of mapping a WordNet synset to a Wikipedia article

and domain knowledge is rare, we propose to use multiword phrases in conjunction with unigram features. The multiword phrases usually have unchanged meanings, thus reduce the ambiguity in unigram 'bag-of-word' document model. Therefore, the indexing of text descriptions involves two parts, i.e. the term indexing and the phrase indexing. The term indexing is simply achieved by calculating the term frequency of each word after lemmatizing and stop-word removal. In our approach, we propose to use the Van Rijsbergen's stop-word list [99] to remove non-content-bearing terms, and the statistical extraction tool Xtract [92] to identify frequent multiword phrases from the text description.

Probabilistic Topic Connection (PTC) Models for Automatic Image Annotation

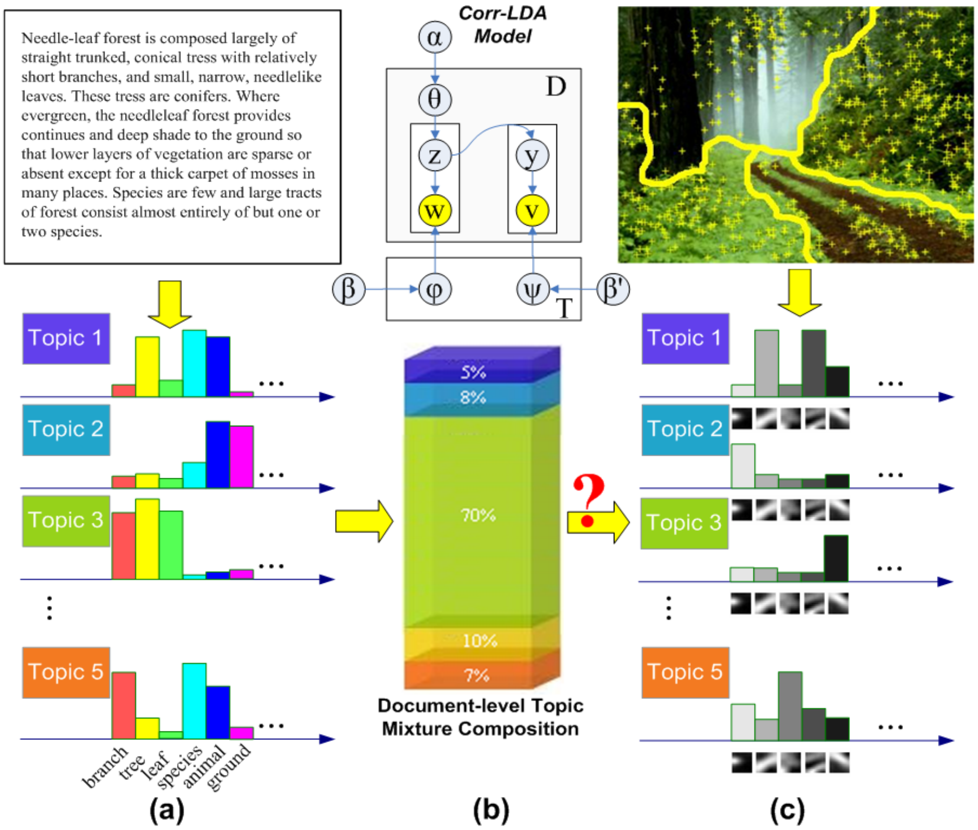
We begin this section with the introduction of extended CorrLDA model, which is the state-of-the-art in modeling image and associated text description [8] [18]. By presenting the generative process of CorrLDA model, we make explicit its problem of topic replicating from text to image. Then, we show how we address the problem in CorrLDA model by introducing new latent semantic variables and relations to thegenerative process.

The CorrLDA model is commonly used by the data mining community to extract latent semantic topics from the co-occurrence patterns of image and text descriptions. A closer look into the generative process of extended CorrLDA model reveals that, the topic composition of image-text pairs is solely decided by the primary entities (such as single-word feature), even though other entities such as the image appearance also serve as a part of description. As a result, in the CorrLDA model, the image topics are made equivalent with the word topics. However, as we know, each word topic may be related to multiple visual topics, enforcing word topics to image features may ignore such a relation and make topic modeling results inconsistent with the underlying image patterns.

To better explain this problem, lets consider a simple image-text modeling problem in Fig. 3.22, for simplicity, we name the entity topic of image feature as 'visual topics'. Assuming that we have a vocabulary of 6 words (branch, tree, leaf, species, animal and ground) and a total of 5 word topics each has a unique distribution of generating words (Fig. 3.22a). Take word topic 3 for example, it has high probability generating branch, tree and leaf while low probability generating 'species', 'anima' and 'ground', so it may be related to the concept of forest. As a comparison, topic 5, which has high probability of generating 'branch', 'species' and 'animal', may represent concept of branch splitting during animal species evolution. Now suppose that we have an image about needle-leaf forest and a piece of text description that explain the needle-leaf forest, and that we choose to represent image content by visual code-words that are derived from SIFT descriptors. As we can see in Fig. 3.22, in the sense of single-word features, this example is almost 'uniform topic, which is mainly composed of topic 3 (Fig. 3.22b). However, in the sense of image feature representation, this example is not really a ‘uniform topic’ case. Although the image purely depicts the scene of needle-leaf forest, however, it still have multiple visual topics corresponding to different image regions such as trunks, leaves, path, grass, etc (Fig. 3.22c). For example, the visual topic ‘trunks’ may favor some visual code-words that occur more frequently in trunks (e.g. vertical lines); similarly, visual topic ‘leaves’ may in turn privilege other visual code-words such as blob-like structures. Since each region takes up similar portion of area in the image, there is no evidence that any of these visual topics be dominant in the entire image. Therefore, the shifting from word topic to visual topics is not as transparent as assumed in CorrLDA model.

With this consideration, in our new model we allow each word topic to connect to multiple visual topics, with different prior probabilities. We call our new model Probabilistic Topic-Connection (PTC) model. This model also allow for different number of word topics and visual topics. For clarity, we name each image-text pair as one document. Some notations to be used in the two topic models are listed as follows: D is the number of documents, T is the anticipated number of latent topics, Nwd is the total number of text words in document d,N_p^d denotes the total number of extracted multiple-word phrases in document d, while N_v^d and N_r^d represents the total number of extracted visual words and MSER regions in document d, respectively. In the model, parameters $\alpha, \beta_w, \beta_p, \beta_v$ are fixed hyper-parameters for the Dirichlet distributions. In our approach, we assume symmetric priors, with $\alpha, \beta_w, \beta_p, \beta_v$ being scalar parameters.

Similar to CorrLDA model, we assume that the observed data is generated by some parameterized random variable known as ‘latent topics’. Specifically, a ‘word topic’ (denotes by



Needle-leaf forest is composed largely of straight trunked, conical tress with relatively short branches, and small, narrow, needlelike leaves. These tress are conifers. Where evergreen, the needleleaf forest provides continues and deep shade to the ground so that lower layers of vegetation are sparse or absent except for a thick carpet of mosses in many places. Species are few and large tracts of forest consist almost entirely of but one or two species.

Figure 3.22: Graphical illustration of mapping a WordNet synset to a Wikipedia article

‘z’) is used to derive the generation of the text words from a topic-specific word distribution (e.g. for a word topic that is related to the concept of forest, the corresponding word distribution will have high probability generating words like ‘branch’, ‘tree’ and ‘leaf’). In a text document, the word topics (which usually relate to some semantic concepts) play an intermediate role between basic elements (words) and high level semantic meanings.The ‘visual topic’ (denotes by ‘y’) is a visual counterpart of the word topic, each had a unique distribution over image features. Specifically, each visual topic is formalized as cluster of features that represent similar image appearance or fit the same distribution in the image feature space. For example, the visual topic 'trunks' may favor some image patterns that occurs more frequently in trunks (e.g. vertical lines), while visual topic 'leaves' may privilege some blob-like image patterns. After identifying latent topics and assigning topic labels to each entity in a document, each document may in turn be represented by a document-level mixture of latent topics. The document-level topic mixture is defined as a probability distribution of latent topics with respect to each document, specifying which word topics are most likely to be generated from observed text description, or which kind of visual topics are most likely to be generated from observed image features.

In Fig. 3.22, we present the graphical representations of the model, in which we highlight the innovation part of proposed model by dashed line. Following the convention in depicting graphical representation of topic models, we use round nodes to represent random variables, in which the white nodes stand for latent random variables, while the gray nodes denote observed ones during the model training. The rounded boxes are used to represent fixed hyper-parameters of the model, while the edges illustrate the conditional dependency underlying the generative process. Detailed explanations of notations used in Fig. 3.23 and following discussions are summarized in Table 3.1.

The generative process for the Probabilistic Topic-Connection (PTC) Model in Fig. 3.23 is:

Table 3.1: Notations in Proposed Topic Model

Symbol	Descriptions
d,w,p,v,r	Instances of variables: d for document, wfor word, pfor phrase, v for visual word, r for MSER region
D,W,P,V	Total number of documents, vocabulary size of words, phrases, visual words
z,z',y,y'	Indicator for word topics (z, z) and visual topics(y, y)
$T1,T2$	The selected number of word topics and visual topics.
$N_w^d,N_p^d.N_y^d,N_r^d$	The number of word tokens, phrases, visual words and MSER regions contained in document d
$C_{kd}^{T_1D},C_{kd,-i}^{T_1D}$	The number of times that word topic k has occurred in document d, with/withoutcounting the current instance
$C_{wk}^{WT},C_{wk,-i}^{WT}$	The number of times that word w is assigned to word topic k, without counting the current instance.
$C_{pk}^{PT},C_{pk,-i}^{PT}$	The number of times thatphrase p is generated from word topic k, with/without counting the current instance.
$C_{vj}^{VT_2},C_{vj,-i}^{VT_2}$	Number of times that visual word v is generated from visual topic j, with/without counting the current instance.
$C_{jk}^{T_2T_1},C_{jk,-i}^{T_2T_1}$	The number of times thatword topic k connects to visual topic j, with/without counting the current instance.
$C_{rj}^{RT_2}$	The number of times that MSERregion r is generated from visual topic j, except current assignment;
θ	A $D \times T$ matrix that indicates the document-topic distribution.
$\alpha,\beta_v,\beta_u,\beta_r,\gamma$	Hyper-parameters of Dirichlet distributions.
λ	A $T1 \times T2$ matrix that indicates theconnection from word topic to visual topic
$\mu_{j,n},\sigma_{j,n}^2$	Parameters of the nth Gaussian distribution with respect to visual topic j
$\bar{u}_{j,n}$	Sample mean ofthe nth Gaussian distribution with respect to visual topic j
$s_{j,n}^2$	Sample variance ofthe nth Gaussian distribution with respect to visual topic j

```
1. For the $d^{th}(d = 1 \cdots D)$ document, sample $\theta_d \sim Dir(\alpha)$
2. For the $k^{th}(k = 1 \cdots T)$ text topic, sample $\varphi_k \sim Dir(\beta_w), \varphi'_k \sim Dir(\beta_p), \lambda_k \sim Dir(\gamma)$
3. For the $j^{th}(j = 1 \cdots T_2)$ visual topic, sample $\psi_j \sim Dir(\beta_v)$
4. For each of the $N_w^d$ words $w_i$ in document d:
   a) Sample a text topic $z_i \sim Multi(\theta_d)$
   b) Sample $w_i|z_i = k \sim Multi(\varphi_k)$
5. For each of the $N_p^d$ phrasespiin document d:
   a) Sample a text topic $z'_i \sim Multi(\theta_d)$
   b) Sample $p_i|z'_i = k \sim Multi(\varphi'_k)$
6. For each of the$N_v^d$ visual words viin document d:
   a) Sample an indicator $s_i \sim Multi(\theta_d)$
   b) Sample a visual topic $y_i|s_i = k \sim Multi(\lambda_k)$
   c) Sample $v_i|y_i = j \sim Multi(\psi_j)$
7. For each of the $N_r^d$ MSER region feature riin document d:
   a) Sample an indicator $s'_i \sim Multi(\theta_d)$
   b) Sample a visual topic $y'_i|s'_i = k \sim Multi(\lambda_k)$
   c) For the n-th dimension of the MSER feature $r_i^{(n)}$
      i. Sample $r_i^{(n)}|y'_i = j \sim N(\mu_{j,n}, \sigma_{j,n}^2)$
```

In the generative process, a T-dimensional topic-prior vector θ_d is sampled for each document d, with the k-th dimension of the vector represents the prior probability of the k-th topic in d. For each document d, the generative process of the N_w^d words is achieved by sampling topics from the document-topic multinomial distribution (with prior θ_d) and sampling

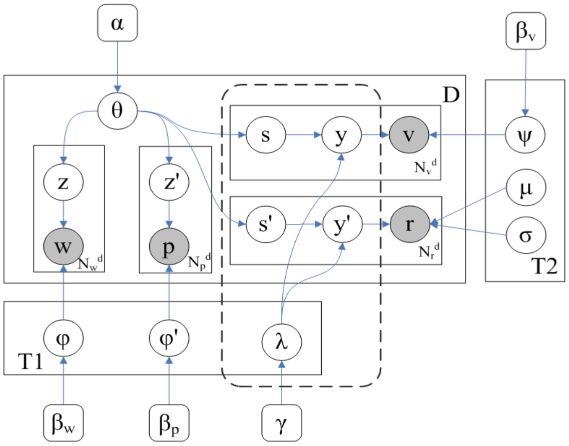


Figure 3.23: Graphical representation of the Probabilistic Topic-Connection (PTC) model

words from the topic-word multinomial distribution (with prior φ_k). After that, instead of being sampled from their own topics, all the other entities (phrases, visual words, etc) are sampled from the same topic as words. Different from Corr-LDA model, new latent variables are introduced to allow for more flexible sampling of word topics and visual topics. Specifically, latent variable s and s' play the role as word topic indicators, while latent variable λ_k serves as the prior probabilities of word topic k connecting to any visual topics. For a given image feature, the model first sample a word topic indicator, then sample the visual topic according to the priori distribution of corresponding word topic connecting to different visual topics.

Collapse Gibbs Sampling For Proposed Topic Model

In recent years, several methods have been developed for estimating the latent variable in topic model, such as the variational expectation maximization, expectation propagation, and Collapse Gibbs sampling [95]. Compared to the other two methods, Gibbs sampling is less computationally intensive, and often yields relatively simple algorithms for approximate inference [95]. With this consideration,we perform the Collapse Gibbs Sampling procedure for model estimation. In the Gibbs Sampling process, a Markov chain is constructed and converges to the posterior distribution on latent topics. The transition between successive states in the Markov chain is modeled by repeatedly drawing a topic for each observed entity from the conditional probability. Due to the space limit, we only introduce our implementation of the Gibbs Sampling for proposed PTC model. For the extendedCorrLDA model, our implementation is similar with that outlined in [8]and [18]. Given the generative process in previous section, our objective is to compute the entity-topic posterior probability and sample topic for each entity from posterior probability. Thus, we derive the posterior sampling equations as follows, in which we follow the standard notations detailed in Table 3.1. Sampling a word topic (z_i) for a given word (w_i)

$$p(z_i = k|w_i = w, \mathbf{z}_{-i}, \mathbf{w}_{-i}, \mathbf{z}', \alpha, \beta_w) \propto \frac{C^{T_1 D}_{kd,-i} + \alpha}{\sum_{k'} C^{T_1 D}_{kd,-i} + T_1 \alpha} \cdot \frac{C^{WT}_{wk,-i} + \beta_w}{\sum_{w'} C^{WT}_{wk,-i} + W\beta_w} \tag{3.21}$$

The above posterior probability is obtained by integrating out (collapsing) all the latent variables φ_k and θ_d separately. Sampling a word topic (z'_i) for a multiple word phrase (p_i)

$$p(z'_i = k|p_i = p, \mathbf{z}'_{-i}, \mathbf{p}_{-i}, \mathbf{z}, \alpha, \beta_p) \propto \frac{C^{T_1 D}_{kd,-i} + \alpha}{\sum_{k'} C^{T_1 D}_{kd,-i} + T_1 \alpha} \cdot \frac{C^{PT}_{pk,-i} + \beta_p}{\sum_{p'} C^{PT}_{p'k,-i} + P\beta_p} \tag{3.22}$$

Sampling a visual topic (y_i) for a visual word feature (v_i)

$$p(y_i = j, s_i = k|\mathbf{v}'_i = v, \mathbf{y}_{-i}, \mathbf{v}_{-i}, \mathbf{y}', \mathbf{z}', \mathbf{z}, \gamma, \beta_v) \propto \frac{C^{T_1D}_{kd}}{N^d_w}\frac{C^{T_2T_1}_{jk,-i}+\gamma}{\sum_{j'} C^{T_2T}_{j'k,-i}+T_2\gamma}\cdot\frac{C^{VT_2}_{vj,-i}+\beta_v}{\sum_{v'} C^{VT_2}_{v'j,-i}+V\beta_v} \tag{3.23}$$

Sampling a visual topic (y'_i) for a MSER region (r_i), in which $r_i = r = (r^{(1)}, \cdots, r^{(n)}, \cdots, r^{(50)})^T$

.

$$\begin{aligned} &p(y'_i = j, s'_i = k|\mathbf{r}_i = r, \mathbf{y}'_{-i}, \mathbf{r}_{-i}, \mathbf{y}', \mathbf{z}', \mathbf{z}, \gamma) \\ \propto &\frac{C^{T_1D}_{kd}}{N^d_w}\frac{C^{T_2T_1}_{jk,-i}+\gamma}{\sum_{j'} C^{T_2T}_{j'k,-i}+T_2\gamma}\cdot\prod_n t_{C^{RT_2}_{rj,-i}-1}(r^{(n)}|\bar{u}_{j,n}, s^2_{j,n}/C^{RT_2}_{rj,-i}) \end{aligned} \tag{3.24}$$

In eq. (3.23), the term in the form of $t_{n-1}(r^{(n)}|\bar{u}, s^2/n)$ is the student-t density with mean $\bar{u}$, variance s^2/n and n-1 degree of freedom. As we place a non-informative prior over the Gaussians, the mean and variance of each Gaussian are purely determined by their sufficient statistics (i.e. the sample mean $\bar{u}$ and sample variance s^2 , respectively). As a result, the student-t density function in eq. (3.23) provides the confidence of drawing the value of $r^{(n)}$ from a topic-specific Gaussian distribution (please refer to section 3.3 for detailed derivation of this conclusion).

During the Gibbs Sampling processes based on above posterior distributions calculations, we may also update single latent variablesin the following manner:

$$\begin{aligned}
E[\theta_{kd}|\mathbf{z},\mathbf{z}',\alpha] &= \frac{C_{kd}^{T_1D}+\alpha}{\sum_{k'} C_{kd}^{T_1D}+T_1\alpha} \\
E[\varphi_{wk}|\mathbf{z},\mathbf{w},\beta_w] &= \frac{C_{wk}^{WT}+\beta_w}{\sum_{w'} C_{w'k}^{WT}+W\beta_w} \\
E[\varphi'_{pk}|\mathbf{z}',\mathbf{p}',\beta_p] &= \frac{C_{pk}^{PT}+\beta_p}{\sum_{p'} C_{pk}^{PT}+P\beta_p} \\
E[\psi_{vj}|\mathbf{y},\mathbf{v},\beta_v] &= \frac{C_{vj}^{VT_2}+\beta_v}{\sum_{v'} C_{v'j}^{VT_2}+V\beta_v} \\
E[\lambda_{jk}|\mathbf{z},\mathbf{z}',\mathbf{y},\mathbf{y}',\gamma] &= \frac{C_{jk}^{T_2T_1}+\gamma}{\sum_{j'} C_{j'k}^{T_2T}+T_2\gamma}
\end{aligned} \tag{3.25}$$

3.4 UTILIZING SOCIAL ANNOTATION FOR USER IMAGE TAGGING

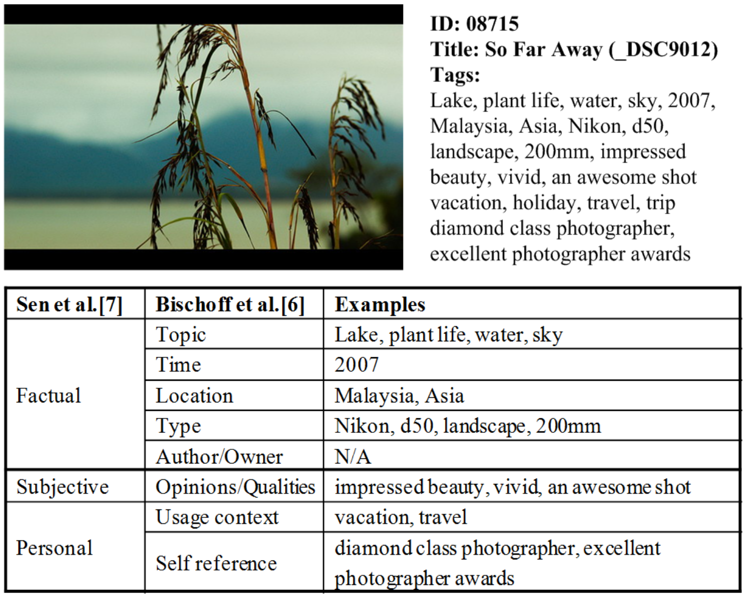
To face the challenge of enormous explosion of unlabeled online image resources, it is important to achieve automatic image tagging for online image resources. The desirable image tagging system should not only be able to interpret the image content but also be able to integrate users' contextual information. Breakthroughs in automatic image tagging algorithms will help with organizing the massive amount of online image resources, promote developing and studying of image storage and retrieval systems, and serve for applications such as interest sharing among online image resource users. The recently established Web 2.0 tools and environments have made collaborative tagging very popular. Take Flickr.com for example, any user can collaboratively assign open-ended text words, in the form of keywords or category labels, to online shared resources for the purpose of organizing and re-finding the images. Moreover, due to its social annotation nature, the tags created by user provide valuable information and metadata which can be utilized to achieve context sensitive information retrieval. With this consideration, it is very important to explore users' tagging patterns in describing the image content. In this section, we propose a robust statistic model to leverage image, text and user-created tags and integrate user context into the image retrieval system.

3.4.1 Background of Social Annotation and Image Tagging Studies

Due to its social annotation nature, Flickr image tags have various functional purposes (Fig. 3.24). For example, the topic tags may refer to any object or person displayed in the picture, such as sky, lake, plant life; the time tags indicate the time when a picture was taken; the location tags provide information about sights, like which country it is from; the type tags include camera settings and photographic styles; the usage context tags suggest the context the picture was collected in, while the self-reference tags contain highly personal information for the tagger himself, such as 'diamond class photographer'. Study on different tag categories suggests that topic and location are two most intensively used tag categories in Flick.com [53].Further study in social tagging behavior suggests that the factual tags [88] (or the first five tag categories identified in [53]) are more closely related to resource content, while the subjective tags and personal tags are more influenced by users' perspectives [64]. Generally speaking, compared to subjective and personal tags,factual tags are more relevant to image content. The subjective and personal tags, on the other hand, are usually free-form texts, but they also provide valuable contextual information of users' tagging preference which can be utilized to customize automatic image tagging for different users.

Study of social tagging in web-based applications has gained increased popularity in the data mining community. Specifically, several probabilistic generative models have been proposed to study users' tagging patterns [82], [111], [64]. In [82], a Conditionally-independent LDA (CI-LDA) model is proposed to infer the generation process of both content word and tags. However, it assumes that tag is purely generated from the textual content of document, while users' impact on the tags generation process is ignored. [111] proposes a social annotation model that considers the impact of both document topic and user interest on tag generation, yet it assumes that words and tags are both generated from the same topics shared by documents and users. In [64], a topic-perspective (TP) model is proposed to infer how both users' perspective and the resource content relate to the generation of social annotations. It improves the generative process of social annotations by separating the tag generation process from the generation process of the resource content. While the resource content (such as text words) is only generated from resource topics, the social tags are generated by both resource topic and user perspective. In this model, the user perspective not only refers to the user's interest, but also covers the users expertise, motivation, language and other personal factors.

On automatic image tagging, another major task is to identify semantic mixture components from the co-existing image content and text descriptions. In the data mining and information retrieval community, there has been a long time focus on using probabilistic topic models to study the correlation between image and text descriptions. Specifically, the Correspondence LDA (CorrLDA) model [14],which imposes correspondence between text word and other semantic entities, provides a natural way to learn latent semantic components(topics) from image features and associate them with text descriptions. Many recent studies, including sophisticated topic models that associate image features with multiple



ID: 08715
Title: So Far Away (_DSC9012)
Tags:
Lake, plant life, water, sky, 2007, Malaysia, Asia, Nikon, d50, landscape, 200mm, impressed beauty, vivid, an awesome shot vacation, holiday, travel, trip diamond class photographer, excellent photographer awards

Sen et al.[7]	Bischoff et al.[6]	Examples
Factual	Topic	Lake, plant life, water, sky
	Time	2007
	Location	Malaysia, Asia
	Type	Nikon, d50, landscape, 200mm
	Author/Owner	N/A
Subjective	Opinions/Qualities	impressed beauty, vivid, an awesome shot
Personal	Usage context	vacation, travel
	Self reference	diamond class photographer, excellent photographer awards

Figure 3.24: Illustration of Flickr image tags and the mapping to different social tagging classification schemas

types of semantic entities (such as protein entities [8], ontology-based biomedical concepts [18]), still follow a similar generative process to the prototype CorrLDA model. In CorrLDA model, each image document has different distribution over semantic mixture components; this feature provides the model a flexibility of adapting to different image contents. However, the CorrLDA model requires specifying the exact number of mixture components, which is fixed for each image document and remains unchanged during the model estimation. In practice, in order to get an optimal number, the researchers have to try out different mixture components numbers and make a choice by comparing the log-likelihood, perplexity and other criteria that indicate how good the model fits the data. The Hieratical Dirichlet Process (HDP) model [108], is a nonparametric extension of the Latent Dirichlet Allocation (LDA)-based topic models, it enables modeling documents with countable infinite mixture components, thus provides the flexibility of modeling images whose actual semantic component numbers are unknown.

3.4.2 Perspective HDP Model for Online User-tagged Image

In this section, we introduce the perspective HDP (pHDP) model for user-tagged images. We present graphical representation of pHDP model in Fig. 3.25. Following the convention in depicting graphical representation of topic models, we use round nodes to represent random variables, in which the white nodes stand for latent random variables, while the gray nodes denote observed ones during the model training. The rounded boxes are used to represent fixed hyper-parameters of the model, while the edges illustrate the conditional dependency in the generative process.

For clarity, we name each tagged image as a document. Some notations to be used in the two models are listed as follows: J is the number of image documents, K and K' (both are countable infinite) indicate the number of semantic mixture components; when K is a finite number, the models become LDA-like models. To represent the image content, we utilize the saliency features (including visual code-words [91] and MSER feature [67]) as a complement part of the holistic GIST features [90]. Our motivation comes from the fact that the mechanism of human visual perception allows for very rapid holistic image analysis to provide a coarse context of image scene (special layout model), yet it also gives rise to a small set of candidate salient locations in a scene (saliency model) that needs to be intensively studied [43]. In Fig. 3.25, N_j^t is the number of tags in document j, while N_j^v and N_j^r represent the total number of extracted visual code-words and MSER regions in document j, respectively. In the model, the holistic representation of an image is replicated 10 times to enable the posterior sampling, so N_j^h denoted the h-th replication of the holistic image representation in document j. In both models, we assume fixed value for Dirichlet process concentration parameters α_0 and γ. We also assume symmetric priors $\alpha_u, \xi_v, \xi_t, \eta and \zeta$ for Dirichlet distributions in the models. Detailed explanations of notations in following discussions are summarized in Table 3.2.

As shown in Fig. 3.25, this model primarily comprises of two parts split by the dash line. The part on the right hand side is essentially the standard HDP model. The generative process of this part begins with drawing a global probability measure $G_0 \sim DP(\gamma, H)$ and for each document j, draw a child Dirichlet process $G_j \sim DP(\alpha_0, G_0)$. Following the stick-breaking construction, it is equivalent to firstly drawing a global weight $\beta \sim GEM(\gamma)$ for semantic component indicators k, then for each document j, draw the document-level weights of semantic component indicators $\pi_j \sim DP(\alpha_0, \beta)$. The data observations in document j are generated by repeatedly drawing semantic component indicator z_{ji} and z_{jl} from π_j and then draw each data observation (i.e. each MSER region and each visual code-word) from the conditional probability of the sampled semantic component.

The left half of the model is for the generation of image tags. As mentioned in previous section, image tags have various functional purposes. For example, some tags (like most factual tags) are closely related to the contents displayed in images, while other tags (including location tags, subjective tags and personal tags) indicate user's contextual information as well as his/her subjective feelings and preferences, which we refer to as 'user's perspectives'. Accordingly, the generative process of user-tagged images should be able to take into account both user's perspectives and semantic components from image contents.

Table 3.2: Notations in Proposed Perspective HDP Model

Symbol	Descriptions
t,p,v,r,h	Instances of variables: t for tags, pfor users perspective, v for visual word, r for MSER feature, h for GIST feature
J,T,U,L	Number of documents, tags, users, users perspectives
z_p,z_{ji},z_{jl},s_j	Indicators for semantic components.
K,K'	The number of components at a certain time point.
N_j^v,N_j^r,N_j^t,N_j^h	Number of visual words, MSER regionsand tags, plus the replication number of GIST features in document j
$C_{vk}^{VZ},C_{vk,-i}^{VZ}$	Number of times visual word vwas assigned to semantic component k, with/withoutcounting the current instance
$C_{rk}^{RZ},C_{rk,-i}^{RZ}$	Number of MSER feature vectors being assigned to component k, with/withoutcounting the current instance
$C_{hk'}^{HS},C_{hk',-i}^{HS}$	Number of GIST feature vectors being assigned to component k', with/without counting current instance.
$C_{pu,-q}^{PU}$	The number of times that perspective p is adopted by user u, except current instance;
$\breve{n}_{j,-q}$	The number of tagsin document d=j generated from users perspectives ($x_{jt}=2$), except current instance;
$n_{j,-q},n'_{j,-q}$	The number of tags in document d=j generated from semantic components ($x_{jt}=0,1$), except current instance
$C_{tp,-q}^{TP}$	The number of times that tag t=q is generated from users perspective p, except current instance;
$C_{tk,-q}^{TZ},C_{tk',-q}^{TS}$	The number of times that tag t=q is generated from semantic component k,k', except current instance;
α_0,γ	Concentration parameters of Dirichlet process.
$\varphi_k^t,\varphi_k'^t$	The tag distribution of semantic component k, k'
φ_k^v	The visual word distribution of semantic component k
π_j	The document-level weights of semantic component indicators for document j
$\alpha_u,\xi_v,\xi_t,\eta,\zeta$	Hyper-parameters of Dirichlet distributions
$\mu_k^{(n)},{\sigma_k^{(n)}}^2$	Parameters of the n-th Gaussian distribution with respect to the k-th semantic component
$\bar{u}_{k,n},s_{k,n}^2$	Sample mean and sample variance ofthe nth Gaussian distribution with respect to the kth semantic component
θ_u,ψ_p	The perspective distribution of user u, and thetag distribution of perspective p.
x,λ_j	Switch variable that decides the source of each tag and the document-level distribution of different x values
β	The global weight of semantic component indicators across the corpora

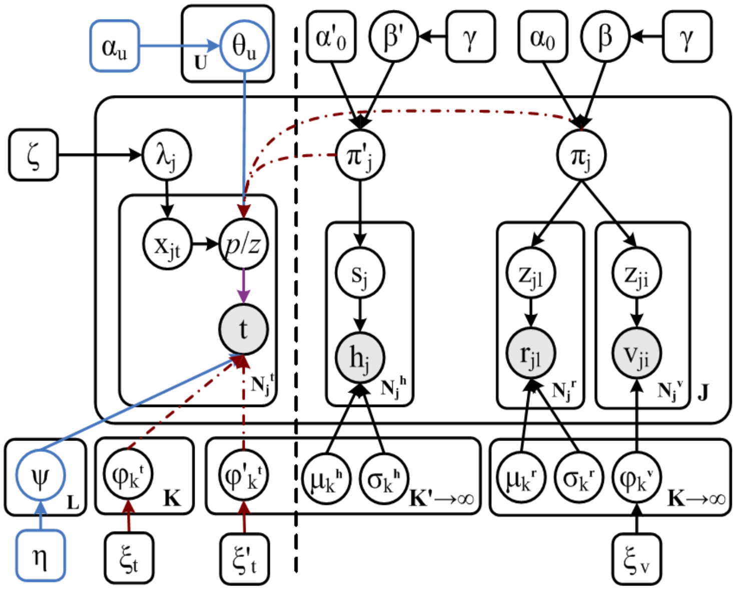


Figure 3.25: Graphical representation of the topic-perspective model for image tagging system (the gray nodes represent the observations from tagged images)

In pHDP model, each tag t created by user u for document j can be either drawn from the semantic components associated with j's image content or from u's perspectives. To decide the source of each tag, a switch variable x is introduced. For each tag t in document j, the value of x_{jt} (which takes values 0-2) is sampled from a multinomial distribution λ_j (with a Dirichlet prior ζ). When the value of x_{jt} equals 0 or 1, the topical indicator of tag t is draw uniformly from the semantic components learned from the image contents(the red dashed arrows in Fig. 3.25 show this process). When x_{jt} equals 2, a user perspective p will be sampled from the perspective distribution θ_u for user u, and tag twill be drawn from the tag distribution ψ_p of perspective p (the blue arrows in Fig. 3.25 illustrate this procedure). The switch variable x plays a critical role in the pHDP model; it is a personalized factor that indicates in which extent the user's perspectives influence the tagging results. It provides the model a flexibility to determine if a specific image tag relates to the semantic components displayed in an image, or it relates to user's context information as well as his/her subjective feeling and preference (users perspective). Overall, the generation process of user-tagged image in the perspective HDP model can be described as in Table 3.3.

3.4.3 Gibbs Sampling for the Model Estimation

In this section, we describe the Gibbs sampling scheme for the proposed pHDP model. The sampling scheme consists of two steps. The first step is sampling for semantic component indicators z as well as the corresponding HDP hyper-parameters β. In order to sample a HDP-like model,one may either follow the Chinese restaurant franchise (CRF) or use direct assignment [108]. In our work, the direct assignment is used(Table 3.4).

3.4.4 Topic Level Image Retrieval Model Using Social Annotation

The discovery of user perspective in image tagging process is one major contribution of our model. In image retrieval, we may expect better performance if we are able to remove those subjective and personal tags in advance based on the discovered user perspective.

Traditional language model (LM) based text document retrieval method considers the generation of a query as a process of independent drawing from a probabilistic distribution associated with the users. Typically, the user first choose a query θ_q, then formulate the query q from the query θ_q with the probability $p(q|\theta q)$. Similarly, the documents are generated word by word from a document θ_d.

In our approach, because of the 'semantic gap' between image documents and text queries,

Table 3.3: The Generative Process of Proposed Models

1.		Draw a global weight $\beta \sim GEM(\gamma)$
2.		For each semantic component k, draw $\lambda_k \sim Beta(1,\zeta), \varphi_k^v \sim Dirichlet(\xi_v), \varphi_k^t \sim Dirichlet(\xi_t)$
3.		For each semantic component k, sample Gaussian-parameters $\mu_{kh}, \sigma_{kh}, \mu_{kr}, \sigma_{kr}$ from sample mean and sample variance;
4.		For each of the U users u, sample perspective distribution $\theta_u^{(u)} \sim Dirichlet(\alpha_u)$, for each of the L user perspectives l, sample $\psi_l \sim Dirichlet(\eta)$
5.		For the jth document, draw $\pi_j \sim DP(\alpha_0, \beta), \pi'_j \sim DP(\alpha_0, \beta')$
6.		For the holistic scene representation of the jth tagged image ,
	a.	sample scene component indicator $s_j = k \sim Discrete(\pi'_j)$
	b.	for the nth dimension of the GIST feature vector hj
		i. sample $h_j^{(n)} \sim N(\mu_{kh}^{(n)}, {\sigma_{kh}^{(n)}}^2)$
7.		For the ith of the Njvvisual code-words in the jth document
	a.	sample object component indicator $z_{ji} = k \sim Discrete(\pi_j)$
	b.	sample its texton id $v_{ji} \sim Multinomial(\varphi_k^v)$
8.		For the lth of the NjrMSER salient regions in the jthdocument
	a.	sample object component indicator $z_{jl} = k \sim Discrete(\pi_j)$
	b.	for the nth dimension of MSER feature vectorrjl
		i. sample $r_{il}^{(n)} \sim N(\mu_{kr}^{(n)}, {\sigma_{kr}^{(n)}}^2)$
9.		For each document j, sample $\lambda_j \sim Dirichlet(\zeta)$
10.		For each tag t in document j created by user u;
	a.	sample a switch variable $x \sim Multinomial(\lambda_i)$
	b.	if $(x = 0)$
		i. Sample semantic component indicator $z_t \sim Discrete(\pi_j)$
		ii. Sample a tag $t \sim Multinomial(\varphi_k^t)$
	c.	if $(x = 1)$
		i. Sample semantic component indicator $z_t \sim Discrete(\pi'{}_j)$
		ii. Sample a tag $t \sim Multinomial(\varphi_k'^t)$
	d.	if $(x = 2)$
		i. Sample a perspective $p_t \sim Multinomial(\theta_u)$
		ii. Sample a tag $t \sim Multinomial(\psi_{pt})$

Table 3.4: The Posterior Sampling of Semantic Components

Preliminaries:
Suppose that at current stage of the sampling, only K of $L \to \infty$ semantic components have been assigned to the observations, define: $\beta_u = \sum_{k=K+1}^{\infty} \beta_k, \gamma_r = \gamma/L, \gamma_u = \gamma(L-K)/L$, then we get:
$\beta = \{\beta_1, \cdots, \beta_k, \beta_u\} \sim Dirichlet(\gamma_r, \cdots, \gamma_r, \gamma_u)$
Repeat for each data observation: Sampling z (may either equals to an existing k or $k_{new} = K+1$)
Firstly, integrate out jto get the marginal probability $p(z|\beta)$:

$$p(\mathbf{z}|\beta) = \int_{\pi_j} p(\pi_j|\alpha_0, \beta) \mathrm{d}\pi_j$$

$$= \prod_{j=1}^{J} \int_{\pi_j} \prod_{k=1}^{K} \pi_{jk}^{n_{jk}+\alpha_0\beta_k - 1} \cdot \frac{\Gamma(\sum_{k=1}^{K})\alpha_0\beta_k}{\prod_{k=1}^{K} \Gamma(\alpha_0\beta_k)} \mathrm{d}\pi_j$$

$$= \prod_{j=1}^{J} [\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + n_j)} \cdot \prod_{k=1}^{K} \frac{\Gamma(\alpha_0 + \beta_k + n_{jk})}{\Gamma(\alpha_0 + \beta_k)}]$$

Secondly, get the posterior probability of zji given the data observations (not counting the current observation v_{ji})

$$p(z_{ji} = k|v_{ji}, \mathbf{z}^{-ji}, \mathbf{v}^{-ji}, \beta) \propto p(z_{ji} = k|\mathbf{z}^{-ji}, \beta) p(v_{ji}|z_{ji}, \mathbf{z}^{-ji}, \mathbf{v}^{-ji}, \beta)$$

$$\propto \{ \begin{matrix} (\alpha_0\beta_k + n_{jk}^{-ji}) f_k^{-v_{ji}}(v_{ji}) \\ (\alpha_0\beta_k) f_k^{-v_{ji}}(v_{ji}) \end{matrix}$$

For visual word v_{ji}, $f_k^{-v_{ji}}(v_{ji}) \propto \frac{C_{vk,-i}^{VZ} + \xi_v}{\sum_{v'=1}^{V} C_{v'k,-i}^{VZ} + V\xi_v} f_{k_{new}}^{-v_{ji}}(v_{ji}) \propto \frac{\xi_v}{V\xi_v}$
For both MESR feature vector and GIST feature vector,
$f_k^{-r_{ji}}(r_{ji}) \propto \prod_n t_{C_{rk,-i}^{RZ}-1}(r^{(n)}|\bar{\mu}_{k,n}, s_{k,n}^2/C_{rk,-i}^{RZ})$
$f_{k_{new}}^{-r_{ji}}(r_{ji}) \propto \prod_n t_{C_{-i}^{RZ}-1}(r^{(n)}|\bar{\mu}_n, s_n^2/C_{-i}^{RZ})$
in which $t_{n-1}(r^{(n)}|\bar{u}, s^2/n)$ denotes the student-t density with mean $\bar{u}$, variance s^2/n and $n-1$ degree of freedom. Sampling m (feasible when $n_{jk} < 200$)
For each j, the auxiliary variable m is sampled as:
$p(m_{jk} = m|\mathbf{m}^{-jk}, \mathbf{z}, \beta) = \frac{\Gamma(\alpha_0\beta_k)}{\Gamma(n_{jk}+\alpha_0\beta_k)} s(n_{jk}, m)(\alpha_0\beta_k)^m$
in which $s(n,m)$ is defined as: $s(0,0) = s(1,1) = 1, s(n,0) = 0, s(n,m) = 0$ for $m > n, s(n+1,m) = s(n,m-1) + ns(n,m)$
Sampling β: accumulate m_{jk} for all document j to get $m_1, m_2, \cdots, m_K$, then draw $\beta \sim Dirichlet(m_1, m_2, \cdots, m_K, \gamma)$

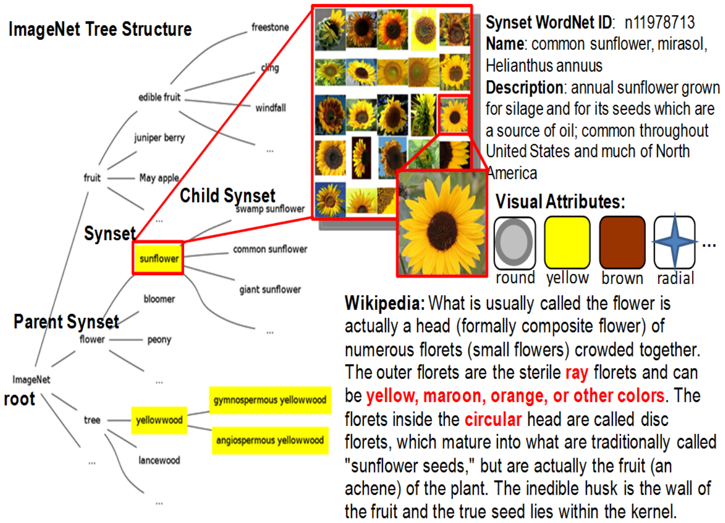
we cannot directly compare them using the word-level language models, instead, we seek to build our image retrieval framework upon the topical information we previously generate from the tagged images. In a conventional query language model, the θ_q is just the empirical word distribution in the query$q = \{w_1, w_n\}$. In our approach, to meet the needs of image retrieval using social tags, we modify the original language model to a higher level, i.e. the topical level. Thus, we consider the semantic categorical c variable uncovered from the tagged image in the previous section, and use such topical information to determine the relevance between tagged images and queries. Recall that in the proposed topic-perspective model, each semantic category/topic c is corresponding to as well as a distribution over image appearance as well as a distribution over tags: $t_j \sim Multinomial((t)_{cj})$. Therefore, we are able to estimate the conditional probability that a tag token in a query correspond to the semantic categorical variable c, say $p(c|t_q)$. Over all topics, we have a vector $v_{c|t} = p(c_1|t_q),, p(c_n|t_q)$. In this manner, we formulate the probability distribution over semantic categories/topics. Finally, we combine all the topical distributions in a query with respect to each tag token to produce a query-level topic distribution (i.e. the topical query LM).

The topical document θ_d can be obtained using maximum log-likelihood estimation from the observed examples from all the tagged images in the dataset, using the framework introduced in the previous section. After that, the closeness between a given query and a given document is simply the Kullback-Leibler (KL) divergence between them:
$D_{KL}(\mathbf{d}, \mathbf{q}) = \sum_c p(c|\theta_q) log \frac{p(c|\theta_q)}{p(c|\theta_d)}$.

3.5 MODELING SEMANTIC RELATIONS BETWEEN VISUAL ATTRIBUTES AND OBJECT CATEGORIES

In our daily life, a large amount of our verbal communication describes the scene / environment around us. Also, recent years have seen increasing amount of online visual resources (such as images and videos) with natural language descriptions. Such information may potentially serve as a rich knowledge base of how people construct natural language to describe visual content. In order that an image annotation system facilitate extracting and understanding the knowledge encoded in the visual content, it is very important to generate descriptive topic models that combines natural languages descriptions with image visual attributes. This work differs from conventional computer vision approaches such as scene recognition and object classification. Instead, it will encode additional semantic information such as the relation between object categories and different visual attributes, which is then linked to natural language descriptions of human knowledge (such as Wikipeida) to generate descriptive topic model regarding object with those visual attributes (Fig. 3.26).

Image annotation was conventionally solved as nearest-neighbor problem [57], [105]. Similar approaches range from studying the relevance between visual similarity and semantic similarity [26], using language entities to construct visual ontologies [102] or jointly modeling images and tags [62]. Most recently, [56] proposed to use conditional random field (CRF) to predict image based potential of how likely the object categories, visual attributes and preposition relationships present in images. However, those approaches are infeasible when labeled reference exemplars are not available. An alternative way is to rely on structured knowledge bases of natural language descriptions (such as Wikipedia).



Synset WordNet ID: n11978713
Name: common sunflower, mirasol, Helianthus annuus
Description: annual sunflower grown for silage and for its seeds which are a source of oil; common throughout United States and much of North America

Wikipedia: What is usually called the flower is actually a head (formally composite flower) of numerous florets (small flowers) crowded together. The outer florets are the sterile **ray** florets and can be **yellow, maroon, orange, or other colors**. The florets inside the **circular** head are called disc florets, which mature into what are traditionally called "sunflower seeds," but are actually the fruit (an achene) of the plant. The inedible husk is the wall of the fruit and the true seed lies within the kernel.

Figure 3.26: Illustration of lexical concept, narrative natural language description and visual attributes

Due to the increasing need of linking visual appearance to structured human knowledge in scalable image categorization/annotation, the extraction of semantic visual attributes has received increasing research focus. By its literal definition, the term ‘attribute’ means ‘a quality or characteristic inherent in or ascribed to an object’. Compared to low-level image features, semantic visual attributes have much stronger relation to both object categories and human knowledge. It should be noted that although various types of attributes can be used to literally describe an object, however, only a small fraction of those attributes may be visible from an object image. Moreover, the usage of textual attributes may differ in different context. For example, in addition to color, texture, shape, body parts, semantic attributes of an animal may also involve its behavior, nutrition, activity, habitat and characters; on the contrary, the attributes about a plant may involve its cultivation and uses, which may be related to botany study. In order that the semantic attributes be useful for image annotation, these attributes should be visible and discriminating among different object categories, also, the union of semantic visual attributes should have sufficient coverage, which means that each object category be covered by at least one attribute.

In our research, we focus on the automation of attribute identification process and semantic relation learning between visual attributes and external textual knowledge sources. The contribution is two-fold, firstly, we provide uniform framework to reliably extract both categorical attributes and depictive attributes. Secondly, we incorporate the obtained semantic associations between visual attributes and object categories into a text-based topic model and extract descriptive latent topics from natural language knowledge base. Specifically, we show that in mining large scale knowledge base of natural language descriptions, the relation between semantic visual attributes and object categories can be encoded as Must-Links and Cannot-Links, which can be represented by Dirichlet-Forest prior. To reduce the amount of manual supervision and labeling in large-scale image categorization, we introduce a semi-supervised training framework using soft-margin semi-supervised SVM classifier (Fig.3.27). We also show that the large-scale image categorization results can be significantly improved by combining automatically acquired visual attributes.

In this section, firstly we will introduce the preliminary task in providing reliable source for attribute learning. We then introduce our approach for image attribute classification. We also present the framework of associate semantic visual attributes with text-based topic models via Dirichlet Forest prior and provide the Gibbs sampler for model estimation.

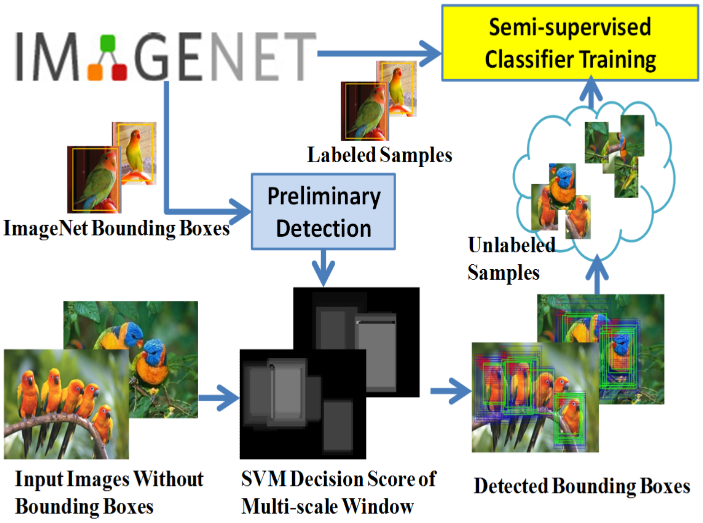

IMAGENET
Semi-supervised Classifier Training
Labeled Samples
ImageNet Bounding Boxes
Preliminary Detection
Unlabeled Samples
Input Images Without Bounding Boxes
SVM Decision Score of Multi-scale Window
Detected Bounding Boxes

Figure 3.27: Bounding boxes as reliable source for attribute learning

3.5.1 Semi-Supervised Large-Scale Multiclass Object Classification

ImageNet dataset [25] is a recently established large scale image ontology (over 15 million images from more than twenty thousand synsets) built upon the WordNet Structure, covering a subset of the nouns of WordNet. In ImageNet dataset, bounding boxes are available for over 3000 popular synsets. For each synset, there areon average 150 images with bounding boxes.The bounding boxes are manually annotated and verified through Amazon Mechanical Turk (AMT) workers. Comparing to attribute learning from full image (FI), the advantage of attribute learning in bounding boxes is obvious, the concept is much cleaner than the full image, no background clutter and other unrelated objects. Related researches have shown that image visual recognition algorithms significantly benefit from explicitly localizing category instance in the image [26]. Moreover, the association between image categories and visual attributes can alsobe significantly strengthen when using bounding box annotation. While high-quality manual labeled bounding boxes have led to impressive object recognition results, however, the main drawback of this approach is that it requires labor-intensive manual labeling and is not scalable to new object categories. In our approach, we proposed to robustly classify object categories and learn visible semantic attributes from automatically detected bounding boxes in ImageNet images (Fig. 3.27).

Bounding Box Detection in ImageNet Images

In our approach, we extract the HOG-LBP feature [105] for bounding box detection. We follow the settings in [57] to train the preliminary non-linear SVM classifiers, in which the kernel of categorical classification is the sum of individual X^2 kernel SVM of each features. We use 80% image data for training and remaining 20% for validation. Achieves average multi-class classification accuracy of 38.5%. Specifically, we densely sample multi-scale detection windows $W_j^T(x,y,s)$ in whole-image range and then perform the 3D mean-shift [19] mode seeking algorithm (in both spatial and scale dimension) on the density map of SVM decision scores across the image to effectively locate the bounding boxes of objects. Given an detected window $x_{(j-1)} = (x,y,s)$, the 3D mean-shift is calculated as:

$$m_j(x) = \frac{\sum_{i=1}^{M} x_i w(x_i) k\left(\|\frac{x_i - x}{h}\|^2\right)}{\sum_{i=1}^{M} w(x_i) k\left(\|\frac{x_i - x}{h}\|^2\right)} - \mathbf{x}_{j-1} \tag{3.26}$$

In which$\{x_i\}_{i=1}^{M}$ are locations corresponding to sliding windows within the neighborhood of $\mathbf{x}_{j-1}, w(x_i)$ is the SVM decision score associated to each location x_i , and k(x) is the profile of kernel K, which satisfies$K(\mathbf{x}) = k(\|\mathbf{x}\|^2)$). We begin with $\mathbf{x}_0 = (x_0, y_0, s = 0)$, iteratively compute j-th mean shift vector $m_j(x)$ and move the estimation window by $m_j(x)$ repeat until convergence. We choose a set of kernel scales around the original image scale as $\{\sigma_s = \sigma_0 * 1.17^s, -n \leq s \leq n\}$, in which n=2 use the Gaussian kernel $K(\mathbf{x}) = exp\{-\|x_i - x\|^2/(2\sigma_s^2)\}$ for the spatial dimension x,y, use flat kernel for scale dimension s, with its shadow kernel $H(s) = 1-(s/n)^2$.

Optimized Kernel Function for Soft-margin Semi-supervised Support Vector Machine

Given the relatively low accuracy (38.5%) in preliminary bounding box detection, directly assigning hard labels to the detected bounding boxes is sub-optimal. Instead, it is reasonable for us to consider the bounding box data as high-quality ‘unlabeled’ data with balanced positive and negative samples (i.e. accurate bounding boxes and inaccurate bounding boxes, respectively). In order to achieve optimal performance in object categorization, we propose to use soft-margin semi-supervised classifier in training (Fig. 3.27). However, one of the major challenges is how to appropriately involve unlabeled examples and efficiently update the discriminative model in an online semi-supervised setting. In our approach, we focus on exploring the intrinsic manifold structure of data marginal distribution and studying its role in kernel function optimization. As shown in [11], general SVM training problem can be extended by considering the ambient space and the marginal distribution of the target function, thus two appropriate penalty terms can be introduced to reflect both the ambient space and the intrinsic structure of the data marginal distribution

P_x. Specifically, the target function could be estimated by:

$$f^* = arg \min_{f \in H_k} \frac{1}{l} \sum_{i=1}^{l} C(\mathbf{x}_i, y_i, f(\mathbf{x}_i)) + \gamma_A \|f\|_{H_k}^2 + \gamma_I \|f\|_I^2 \tag{3.27}$$

The $\|f\|_I^2$ can be estimated as the weighted Laplace-Beltrami operator associated with P_x : $\|f\|_I^2 = \int_{\mathbf{x} \in M} \|\nabla_M f(\mathbf{x})\|^2 \mathrm{d}P_x$. Its shown in [65] that given a set of l labeled examples $\{(x_i, y_i)\}_{i=1}^{l}$ and a set of u unlabeled examples$\{x_j\}_{j=l+1}^{l+u}$, the Laplace-Beltrami operator on the manifold $M \subset R^N$ can be approximated by the graph Laplacian L on the basis of labeled and unlabeled data i.e. $\|\nabla_M f(\mathbf{x})\|^2 :=< \mathbf{f}, L\mathbf{f} >$, in which $\mathbf{f} = [f(\mathbf{x}_1), \cdots, f(\mathbf{x}_{l+u})]^T$. L is the graph Laplacian given by $L = D - W$, in which W_{ij} is similarity between x_i and x_j calculated by kernel function k, $D = diag(D_{1,1}, \cdots, D_{l+u,l+u})$ is a diagonal matrix with the entry$D_{i,i} = \sum_{j=1}^{l+u} W_{ij}$. The optimization problem (3.27) becomes:

$$f^* = arg \min_{f \in H_k} \frac{1}{l} \sum_{i=1}^{l} C(\mathbf{x}_i, y_i, f(\mathbf{x}_i)) + \gamma_A \|f\|_{H_k}^2 + \gamma_I \mathbf{f}^T L\mathbf{f} \tag{3.28}$$

By Representer theorem, the solution is an expansion of kernel functions over both labeled and unlabeled data:

$$f^* = \sum_{i=1}^{l+u} \alpha_i * k(\mathbf{x}_i, \mathbf{x}) \tag{3.29}$$

According to Riesz Representation theorem, define the Gram kernel matrix K with its entries $K_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$,we have:
$\|f^*\|_{H_K}^2 =< f^*, f^* >_{H_K} = \sum_{i=1}^{l+u} \sum_{j=1}^{l+u} \alpha_i^* \alpha_j^* k(\mathbf{x}_i, \mathbf{x}_j) = \alpha^T K \alpha$. Similarly, $\mathbf{f}^{*T} L\mathbf{f}^* =< \mathbf{f}^*, L\mathbf{f}^* >= \alpha^T KLK\alpha$. By substituting into (3.28) the hinge loss function$(1 - yf(\mathbf{x}_i))_+ = max(0, 1 - yf(\mathbf{x}_i))$, the optimization problem can be re-written as:

$$f^* = arg \min_{\alpha \in R^{l+u}, \xi_i \in R} \frac{1}{l} \sum_{i=1}^{l} \xi_i + \gamma_A \alpha^T K \alpha + \gamma_I \alpha^T KLK \alpha \tag{3.30}$$

subject to the relaxed separation constraint:

$y_i(\sum_{j=1}^{l+u} \alpha_j^* k(\mathbf{x}_i, \mathbf{x}_j) + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, \cdots, l$ The above constraint optimization problem (3.30) can be solved by introducing the Lagrangian in which two Lagrange multipliers $\beta_i, \zeta_i \geq 0$ are defined for either constraint:

$$\begin{aligned} L(\alpha, \xi, b, \beta, \zeta) = \frac{1}{l} \sum_{i=1}^{l} \xi_i + \alpha^T (\gamma_A K + \gamma_I KLK) \alpha \\ - \sum_{i=1}^{l} \beta_i \left(y_i \left(\sum_{i=1}^{l+u} \alpha_j * k\left(\mathbf{x}_i, \mathbf{x}_j\right) + b \right) - 1 + \xi_i \right) - \sum_{i=1}^{l} \xi_i \zeta_i \end{aligned} \tag{3.31}$$

Vanishing the derivative of L with respect to b and ξ_i leads to: $\sum_{i=1}^{l} \beta_i y_i = 0, 1/l - \beta_i - \zeta_i = 0, 0 \leq \beta_i \leq 1/l$. Substituting them into (3.31) with b and ξ_i removed, it gives:

$$L^R(\alpha, \beta) = \alpha^T (\gamma_A K + \gamma_I KLK) \alpha - \alpha^T K J^T Y \beta + \sum_{i=1}^{l} \beta_i \tag{3.32}$$

In which $J = [I \quad 0]_{l*(l+u)}, Y = diag(y_1, \cdots, y_l)$.

Taking derivative of (3.33) with respect to leads to: $\frac{\partial L^R}{\partial \alpha} = (\gamma_A K + \gamma_I KLK)\alpha - \alpha^T K J^T Y \beta = 0$, which implies that the l+u expansion coefficients $\alpha_1, \cdots, \alpha_{l+u}$ can be obtained by solving the following quadratic dual program:

$$\left\{ \begin{array}{l} \alpha^* = (\gamma_A I + \gamma_I LK)^{-1} J^T Y \beta^* \\ \beta^* = arg \max_{\beta \in R^l} \sum_{i=1}^{l} \beta_i - \beta^T Q \beta \end{array} \right. \tag{3.33}$$

subject to $\sum_{i=1}^{l}\beta_i y_i = 0$,$0 \leq \beta_i \leq 1/l, i = 1,\cdots,l.$ in which: $Q = YJK(\gamma_A I + \gamma_I LK)^{-1}J^T Y$. (3.33) is a standard restricted quadratic program which can be solved via conjugate gradient descent in Ch. 6 of [11]. During training, the labeled data $\{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$ and unlabeled data $\{(\mathbf{x}_j)\}_{j=l+1}^{l+u}$ are used for solving α^*, β^* by conjugate gradient descent, where $y_i \in \{-1,+1\}$. By substituting the solution α^*, β^* of quadratic program (3.32) to (3.29), we obtain the expansion of kernel function over both labeled data $\{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$ and unlabeled data $\{(\mathbf{x}_j)\}_{j=l+1}^{l+u}$. At the stage of detection, the decision function classified new samples into class +1 or -1 by $y(x) = sign(f^*(x))$.

3.5.2 Descriptive Visual Attribute Extraction and Relation Links

Previous studies on descriptive visual attributes have shown beneficial to improve the performance of object categorization and text description generation [86]. The depictive visual attributes involves color attributes, texture attributes (such as furry, wooden, rough), pattern attributes (spotted, striped) and shape attributes (long, round, rectangle). The attributes may also be associated to the visual similarity with known object classes (for example, giant panda and polar bear both have bear-like attribute). Ideally, these attributes should be able to discriminate between object classes (being associated to some but not all of them), provide sufficient coverage (all classes have at least a single attribute association), and be correlated to visual object class properties that can be observed in images. In our approach, three types of features are used for attribute extraction, i.e. GIST feature [90], densely sampled SIFT [63] and HOG-LBP feature [105]. Each of the three feature types is normalized independently to unit length, then a histogram intersection kernel SVMs [86] is performed to train the attribute classifier. For each classifier, we fit a sigmoid function

[81] to the SVM decision score and convert the output to a probability. The probability $p(Attribute|Category)$ can be aggregated across the whole dataset and eventually build the semantic relations (such as Must-link and Cannot-link) between visual attributes and object categories.

Attributes as Unique Signatures of Object Categories

Given images of an object category, some visual attributes may be presented while some may not, which results in unique attribute signatures associated with each category. Let $a^y = (a_1^y, \cdots, a_M^y)$ be a vector of binary associations $p(a_m|y) \in \{0,1\}$ between attributes a_m and trained object category y. An aggregation of all results (in which $p(a_m|y) = 1$ for positive samples and 0 for negative samples) from binary classifier of attribute a_m can provide an estimation of the conditional probability $p(a_m|y)$ of that attribute being present in category y. By assuming mutual independence among attributes, we have $(a|y) = \prod_{m=1}^{M} p(a_m|y)$.

Following the idea of Direct Attribute Prediction model (DAP) in [57], For an image x with M=K attributes $a_1, \cdots, a_K$ where each attribute corresponds to exactly one conditional probability$p(a_m|y)$, the posterior probability of image x belong to object category y is given as

$$p\left(y|x\right) \propto \prod_{k=1}^{K} \left(\frac{p\left(a_k|y\right)}{p\left(a_k\right)} \right) \tag{3.34}$$

Our experiment results show the performance of object categorization can be significantly improved when the categorization results are smoothed by (3.34), with K=10 most relevant attributes used (Fig. 3.47). By aggregating the results of binary attribute classifiers for all the object categories, we obtain an aggregated attribute-category concurrence map. From which we threshold the aggregation score to produce both Must-Link and Cannot-Link relations for each attribute-category pair (a,y) .

3.5.3 Descriptive Topic Modeling via Dirichlet Forest Prior

In this section, we introduce a novel topic model to infer depictive latent topics from both text corpora and attribute-category relations (Fig. 3.28). Recent studies of large-scale visual classification in ImageNet [24], [26]suggest that visual classification across semantically-defined class boundaries is feasible. In [83], the author proposed to infer object class-attribute association by text-based semantic relatedness on WordNet and Wikipedia.The WordNet is a large scale lexical database of English Language, in which English words are organized into concepts (synonym sets or synsets) according to synonymy and various lexical and semantic relations between lexicalized concepts.

Wikipedia is one of the most comprehensive and well-formed electronic knowledge repositories on the web with millions of articles contributed collaboratively by volunteers. Because of its reliability, accuracy and neutral point of view. Wikipedia has been exploited as external knowledge source in various data mining applications [83], [47]. Although Wikipedia is different from standard WordNet ontology, which is backed up by structured thesaurus, however, each article in Wikipedia only describes one single concept under a hierarchical categorization system. We have found a large amount of Wikipedia articles share the same lexicalized entry as ImageNet synsets, which makes mapping between ImageNet synset to a Wikipedia articles possible.(In our study, about 75% of the ImageNet synsets have corresponding Wikipedia articles). In our approach, the semantic relations between attribute-category pairs (i.e. Must-Links and Cannot-Links) are encoded as Dirichlet Forest prior in the proposed topic model. In order to effectively encode the semantic relations, we explore the WordNet synonym set and extend Must-Links and Cannot-Links between attribute-catry terms (i.e. the lexical word of both attribute and object) to their synonyms.

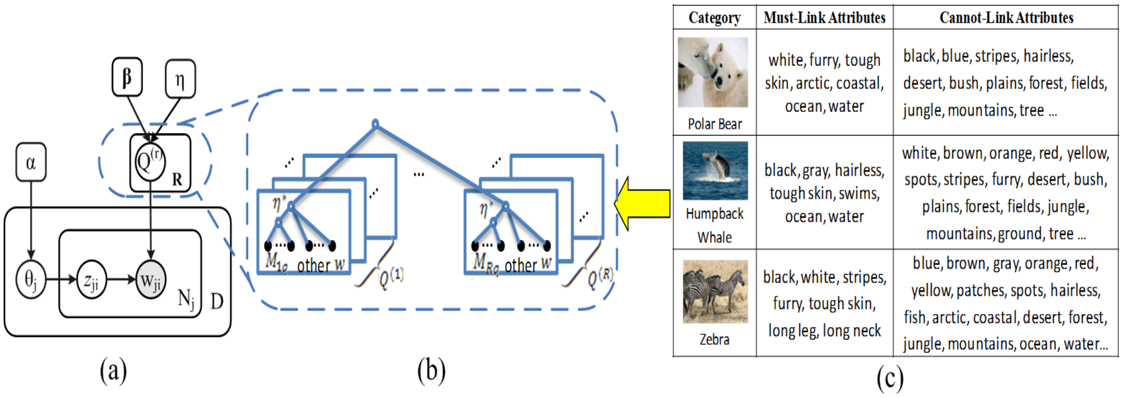
Preliminary of Dirichlet Tree Distribution and the Modeling of Must-Links

The Dirichlet Tree Distribution [73] is a generalization of Dirichlet distribution that allows to break the mutual independence in word generation process, makes the generative process controlled by word-link such as Must-Link (u,v). The Dirichlet-tree distribution is a tree with the words as leaf nodes; let $r^{(k)}$ be the Dirichlet tree edge leading into node k, let c(k) be the immediate children of node k in the tree, L the leaves of the tree, I the internal nodes, and L(k) the leaves in the subtree under k, to generate a sample $\Phi \sim DirichletTree(r)$, one first draws a multinomial at each internal node $s \in I$ from $Dirichlet(r^{c(s)})$, i.e. using the weights from s to its children as the Dirichlet parameters.
The probability Φ_k of a word $k \in L$ is then simply the product of the multinomial parameters on the edges from k to the root.
It can be shown that, the above procedure gives
$DirichletTree(r) \equiv p(\Phi|r) = (\prod_{k \in L} \Phi_k^{r^{(k)}-1})(\prod_{s \in I} \frac{\Gamma(\sum_{k \in c(s)} r^{(k)})}{\prod_{k \in c(s)} \Gamma(r^{(k)})} (\sum_{k \in L(s)} \Phi_k)^{\Delta(s)})$ In which $\Gamma(\cdot)$ is the gamma function, and the notation $\prod_k^L$ means $\prod_{k \in L}$; the function $\Delta(s) \equiv r^{(s)} - \sum_{k \in c(s)} r^{(k)}$ is the difference between the in-degree and the out-degree at internal nodes. (When the difference $\Delta(s) = 0$, for all internal nodes, the Dirichlet tree reduces to a Dirichlet distribution).

Like the Dirichlet distribution, the Dirichlet tree distribution is conjugate to the multinomial. Its possible to integrate out Φ to get a distribution over word counts directly, similar to the multivariate Polya distribution (a.k.a. Dirichlet-Multinomial) in [74]: $p(\mathbf{w}|r) = \prod_{s\in I}(\frac{\Gamma(\sum_k^{c(s)} r_{(k)}}{\Gamma(\sum_k^{c(s)}(r^{(k)}+n^{(k)})} \cdot \prod_{k\in c(s)} \frac{\Gamma(r^{(k)}+n^{(k)})}{\Gamma(r^{(k)})})$ in which $n^{(k)}$ is the number of word tokens in w that appear in L(k), L(k) is the leaves in the subtree under k. c(s) is the immediate children of node s. The definition of Must-Link is transitive. Must-Link (u,v)and Must-Link (u,w) define a transitive closure of Must-Link (u,v,w). In Dirichlet Tree for Must-Links, each transitive closure is subtree, in which words are leaves nodes with symmetric uniform base measure η, β from one internal nodes and each of the internal node s is connected to the root node with weight $|L(s)\beta|$, in which $|L(s)|\beta$ is the size of leavens in sub-tree under s. If $\eta = 1$, then in-degree equals out-degree for any internal nodes(both are $|L(s)\beta$), and the tree reduces to a Dirichlet distribution with symmetric prior β. When we take $\eta = |L(s)|$, it will re-distributethe probability mass at nodes. Which results in increased concentration, and re-distribute the mass evenly in the transitive closure s.The independence (which is enforced in Dirichlet distribution) among Must-Link words is thus eliminated and allows for similar but not identical probabilities for the Must-Link words.

Dirichlet Forest Prior and Cliques of Cannot-Links

From the aggregated concurrence map of attribute-category relations, we are able to assign both Must-Links and Cannot-Links to an object category. It should be noted that, given the presence of an object category in an image, the Must-Links and Cannot-Links corresponding to that category should be simultaneity observed, therefore, such Must-Links and Cannot-Links should be encoded in the same latent topic. With this consideration, we



Category	Must-Link Attributes	Cannot-Link Attributes
Polar Bear	white, furry, tough skin, arctic, coastal, ocean, water	black, blue, stripes, hairless, desert, bush, plains, forest, fields, jungle, mountains, tree ...
Humpback Whale	black, gray, hairless, tough skin, swims, ocean, water	white, brown, orange, red, yellow, spots, stripes, furry, desert, bush, plains, forest, fields, jungle, mountains, ground, tree ...
Zebra	black, white, stripes, furry, tough skin, long leg, long neck	blue, brown, gray, orange, red, yellow, patches, spots, hairless, fish, arctic, coastal, desert, forest, jungle, mountains, ocean, water...

(a) Text topic model (b) Dirichlet Forest prior encoding the visual constraint (c)Must-Link and Cannot-Link Attributes

Figure 3.28: Graphical representation of the proposed method

propose a clique-based topic sampling process as follows .

In our approach, each ‘clique’ is associated with one single object category, it is composed of two parts: the first part is a Dirichlet sub-tree corresponding to Must-Links of that object category, the second parts is all other words (other than words in Must-Links) that are allowed to simultaneously have large probability without violating the Cannot-Links of that object category (Fig. 3.28 b). Each clique is also a Dirichlet tree.

For each object category r, we generate a total of $Q^{(r)} = Q$ cliques $q = 1, \cdots, Q^{(r)}$, in this way, we create a mixture model of $Q^{(r)}$ Dirichlet subtrees, one for each of the $Q^{(r)}$ cliques. In generating the latent topics, the cliques are sampled according to their probability $p(q), q = 1, \cdots, Q^{(r)}$.

The cliques root node connects to an internal nodes(root node of Must-Link sub-tree) with weight $\eta \cdot |L(s)| \cdot \beta$, the node s then connects to words in Must-Links with weight β.The cliques root also directly connects to words that is not in Must-Links (but not violating the

Cannot-Links of that object category) with weight β . This structure will send majority probability mass down to s and then re-distribute it among words in Must-Links. Which results in strong association among Must-Link words.

Let R be the number of object categories. Our Dirichlet Forest prior β, η will consist of $\prod_{r=1}^{R} Q^{r}$ possible Dirichlet trees (cliques), each Dirichlet tree has R branches under the root, one for each connected component, for the r-th branch, there are $Q^{(r)}$ possible Dirichlet subtrees corresponding to $Q^{(r)}$ cliques, which leads to $\prod_{r=1}^{R} Q^{r}$ different Dirichlet trees. Therefore, a Dirichlet tree in the forest is uniquely identified by an index vector $q = (q^{(1)}, \cdots, q^{(R)})$, where $q^{(r)} \in \{1, \cdots, Q^{(r)}\}$.

In generating a Dirichlet Forest model (Fig. 3.28a), let $n_j^{(d)}$ be the number of word tokens in document d assigned to topic j, integrating out θ,z can be generated as:

$p(\mathbf{z}|\alpha) = (\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T})^D \cdot \prod_{d=1}^{D} \frac{\prod_{j=1}^{T} \Gamma(n_j^{(d)}+\alpha)}{\Gamma(n^{(d)}+T\alpha)}$

For each topic j=1,,T, we sampled a Dirichlet tree $q_j = (q_j^{(1)}, \cdots, q_j^{(R)})$ from the Dirichlet Forest prior (β, η)

$p(q_j) = \prod_{r=1}^{R} p(q_j^{(r)})$

In which each $q_j^{(r)}, r = 1, \cdots, R$ is sampled by:$q_j^{(r)} \propto |clique_{q_j^{(r)}}| (r = 1, \cdots, R)$

Finally, the model can be gented by:

$p(w, z, q_{1:T} | \alpha, \beta, \eta) = p(w | q_{1:T}, z, \beta, \eta) \cdot p(z|\alpha) \cdot \prod_{j=1}^{T} p(q_j)$

Gibbs Sampling For Model Estimation

In this section, we introduce the Markov Chain Monte Carlo and Gibbs sampling process of the proposed topic model. Let $n_{-i,j}^{(d)}$ be the number of word tokens in document assigned to topic j, excluding the word w_i . Let $n_{-i,j}^{(k)}$ denotes the number of word tokens in

the corpus that are under node k in topic j's Dirichlet tree q_j, excluding the word at position i. For candidate topic labels $t = 1, \cdots, T$, we have:
$p(z_i = t | z_{-i}, q_{1:T}, w) \propto (n_{-i,t}^{(d)} + \alpha) \prod_s^{Parent_i} \frac{r_t^{(c_t(s))} + n_{-i,t}^{(c_t(s))}}{\sum_k^{c_t(s)} (r_t^{(k)} + n_{-i,t}^{(k)})}$
In which $Parent_i$ denotes the subset of internal nodes in topic v Dirichlet tree that are ancestors of leaf w_i and $c_t^{(s)}$ is the unique node that is s's immediate child and is also an ancestor of w_i (including w_i itself). Since the R branches (each corresponding to an object category, featured by both Must-Links and Cannot-Links) are independent, sampling the Dirichlet tree q_j is factorized to sampling the cliques for each $q_j^{(r)}$. For candidate cliques of connected component $r : q' = 1, \cdots, Q(r)$ we have:
$p(q_j^{(r)} = q^{'} | z, q_{-j}, q_j^{(-r)}, w) \propto (\sum_k^{|clique_{q_j^{'(r)}}|} \beta_k) \prod_s^{I_{j,r=q'}} (\frac{\Gamma(\sum_k^{c_j(s)} (r_j^{(k)})}{\Gamma(\sum_k^{c_j(s)} (r_j^{(k)} + n_j^{(k)})} \cdot \prod_k^{c_j^{(s)}} \frac{\Gamma(r_j^{(k)} + n_j^{(k)})}{\Gamma(r_j^{(k)})})$
In which $I_{j,r=q}$ denotes the internal nodes below the r-th branch of tree q_j when clique $q_j^{(r)}$ is selected.

3.6 EXPERIMENT RESULTS

3.6.1 Spatial Weighting for the 'Bag-of-Visual-Words'

In order to evaluate the effectiveness of the proposed image representation method, we carry outcontent-based image retrieval (CBIR) experiment over different image representation methods. In this experiment, we use the 15-scene Benchmark database [35], from which we select 8 outdoor categories, consist of a total of 2,689 outdoor images from the LabelMe dataset [87].In our experiment, we randomly select 1/6 images from each image category to build the lexicon of visual words. In total, we extract 175,535 visual tokens. At the retrieval stage, we use the selected 1/6 images as query images to retrieve images from the remaining 5/6 images in the data set. The retrieval results are ranked according to the similarity between the query image and all the images in the retrieval set following the similarity measurement in Section 3.2.3.

In the experiment, we compare the performance of our approach (i.e. spatial weighting for the bag-of-visual-words) with the performance of SIFT features and the 'blobworld'.

For SIFT features, we will adopt its latest variation, i.e. the 'bag-of-visual words' by tf-idf term weighting [51]. In recent study, there has been an intense focus on applying term weighting schemes (like tf, idf) to the bag-of-visual-words feature vectors [91], [3]. Extensive study in [51] suggests that when the vocabulary size of visual words is around 1000, the tf-idf weighting performs best. Recall that in our approach, we also use SIFT features to make the experimental results comparable, we choose to compare our image representation approach with the tf-idf weighted 'bag-of-visual-words' approach [3] under the same visual words vocabulary size. As a region-based image representation method that encodes color, shape and texture information by multivariate Gaussians, the 'blobworld' approach is essentially the same as the Gaussian Mixture Model (GMM). It represents image content by the parameter sets of Gaussian mixture components. Following the method in [12], each coherent region is modeled as a multivariate Gaussian. After learning the parameters sets (that is, the mean vector μ_i and the covariance matrix $\sum_i$.) of Gaussians, the KL-divergence for multivariate Normal densities is used as the similarity measurement. The performance will be evaluated by the averaged precision-recall of the content-based image retrieval across all the 8 outdoor categories.

Fig. 3.29 represents the over-all precision-recall of our approach, the 'bag-of-visual-words' approach and the 'blobworld' approach. As expected, our approach achieves high

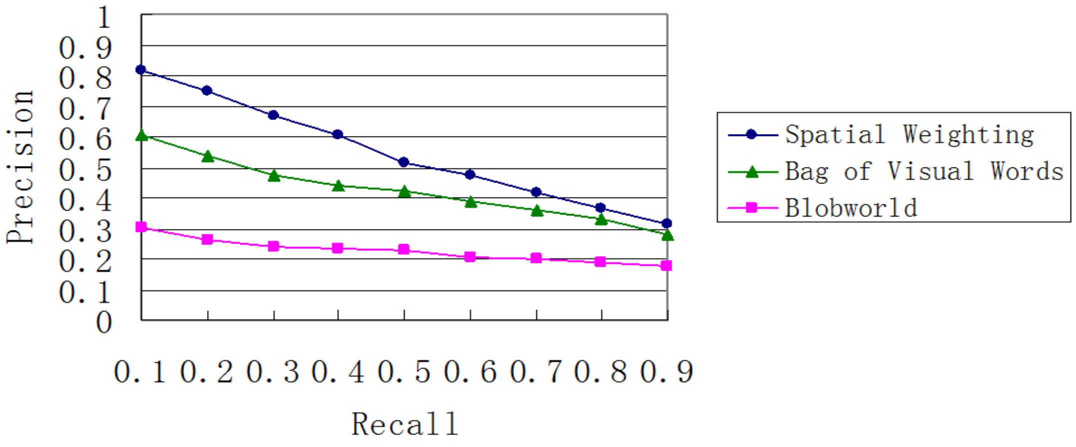


Figure 3.29: Comparison of over-all precision-recall of our approach and comparative approaches.

semantic consistency in CBIR and outperforms both two comparative approaches. Since the contents of different image categories are widely different, it is helpful to compare the performances in individual categories (a briefly description of the three selected categories is represented in Table 3.5.)

The precision-recalls of our approach (spatial weighting, short for S), the 'bag-of-visual-words' approach (short for V) and the 'blobworld' approach (short for B) in selected categories are shown in Fig. 3.30. In categories whose image compositions are highly varied and thus more complicated (such as 'coast' and 'tall-building'), our approach is about 10-20 percentage points better than the 'bag-of-visual-word' approach, while in the category whose image compositions are relatively uniform (like 'forest'), the 'bag-of-visual-word' approach performs as well as our approach. Compared to the other two approaches, the 'blobworld' approach works well only when the colors and image compositions are uniform.

Table 3.5: Description of Three Selected Categories

Image Category	Coast	Forest	Tall-building
Number of Images	360	328	356
Examples			

Conclusions

The experiment results suggest that, visual words from different kinds of regions may make the ‘bag of visual words’ noisy and thus less differentiable. Take the ‘coast’ images for example, the ‘primary’ information about sea and sand beach may be ‘contaminated’ by visual words from other ‘inessential’ parts like boats, buildings and coconut trees. Therefore, the significant improvement in our approach can be explained by the introducing of spatial weighting, which weights visual words according to actual spatial constitution of regions in images. Moreover, the experiment results also suggest that the Gaussian mixture model alone is insufficient to distinguish images which are highly varied in colors and compositions. However, the Gaussian mixture model is still able to provide enough information about the spatial constitutions of images.

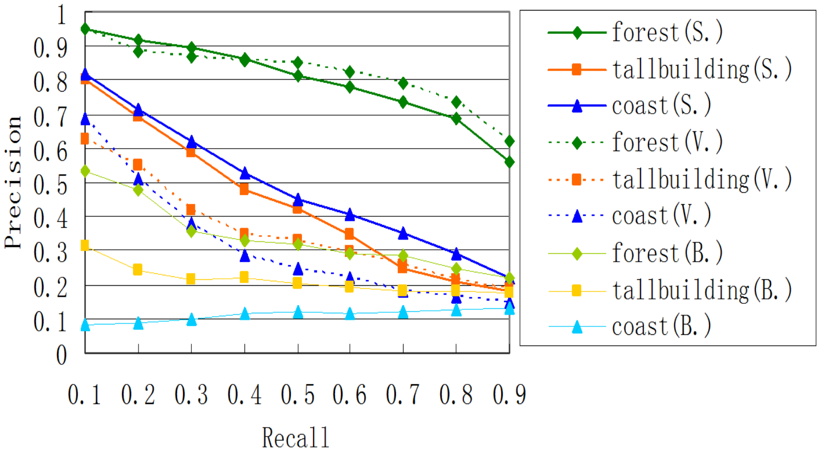


Figure 3.30: Precision-recall in different image categories

3.6.2 Probabilistic Models for Topic Learning from Images and Captions in Online Biomedical Literatures

In this section, we apply the proposed HPB model to topic learning and compare the performance of HPB model with that of the extended Correspondence LDA (Corr-LDA) model under the same biomedical image annotation scenario using cross-validation. For topic learning, we look into the average log-likelihood of two models and visualize the discovered latent themes. The performance of automatic image annotation is evaluated by perplexity and annotation accuracy.

Data Collection and Settings

The data used in our experiment is from the online journal ‘Breast Cancer Research’ in the publicly available PubMed Central database (http://www.pubmedcentral.nih.gov/). In this journal, all the research articles (in digital version) between year 2002 and 2008 are downloaded and parsed. After that, a total of 2320 image-caption pairs are extracted from the original biomedical literatures and make up the dataset for experiment. As introduced in Section 3.3.1, words, visual words and ontology-based biomedical concepts are indexed from image-caption pairs. In total, we indexed 132,978 text tokens which belong to 4113 unique words, 379,526 visual words from a vocabulary size of 1000, and 53,825 concepts, with 1938 unique concepts appear. The original dataset is divided into 5 subsets with equal size. Of the 5 subsets, one subset (20%) is retained as the validation data for testing the model, and the remaining 4 subsets (80%) are used as training data. For image annotation evaluation, the cross-validation process repeats 5 times, with each of the 5 subsets used once as the validation data. After that, we take the average results for evaluation.

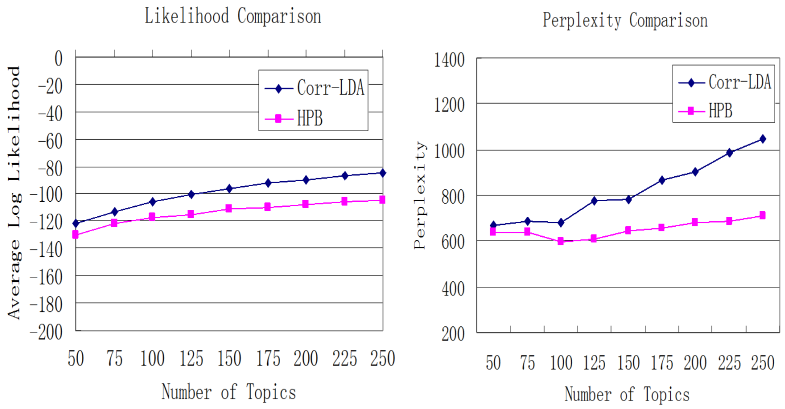
Topic Learning and Representation

The topic learning process of the proposed probabilistic model is achieved by running the collapse Gibbs sampling process over training dataset until converge (basically, it takes less than 100 iterations to converge in model estimation). When the topic model is estimated from the training dataset, we will be able to visualize the uncovered latent themes and tell the correlation among words, visual words and biomedical concepts.

Likelihood Comparison

Log-likelihood is a standard criterion for generative models. It can be calculated by integrating out the topic variables after the convergence of Gibbs sampling. Generally, the higher log-likelihood the model assigned to the data, the better predictive power and generalization ability the model has.

The average word likelihood of the extended Corr-LDA model and the HPB model is compared. The marginal likelihood $p(w|z)$ of the extended Corr-LDA model can be calculated by integrating out latent variables φ:



(a) Likelihood comparison (after convergence) **(b) Perplexity comparison**

Figure 3.31: The comparison of the extended Corr-LDA model and the HPB model

$$p(\mathbf{w}|\mathbf{z}) = \prod_{t=1}^{T}[\int_{\varphi_{z_t}} p(\mathbf{w}|z_t,\varphi_{z_t})p(\varphi_{z_t}|z_t)]\mathrm{d}\varphi_{z_t}$$

$$= \prod_{t=1}^{T}[\frac{\Gamma(W\beta)}{\Gamma(\beta)^w}\int_{\varphi_{z_t}}\prod_{i=1}^{W} p_{w_i}^{n_t^{(w_i)}+\beta-1}]\mathrm{d}\varphi_{z_t}$$

$$= [\frac{\Gamma(W\beta)}{\Gamma(\beta)^w}]^T \cdot \prod_{t=1}^{T}\frac{\prod_{w_i}(n_t^{(w_i)}+\beta)}{\Gamma(n_t^{(w_i)}+W\beta)}$$

The average word likelihood can be obtained by taking the logarithm of $p(w|z)$ and averaging the resulting summation by W.

For the HPB model, the marginal likelihood $p(w|z)$ is as follows:

$$p(\mathbf{w}|\mathbf{z}) = [\frac{\Gamma(W\beta)}{\Gamma(\beta)^w}]^T \cdot \prod_{t=1}^{T}\frac{\prod_{w_i}(n_t^{(w_i)}+\beta)}{\Gamma(n_t^{(w_i)}+W\beta)} \cdot \frac{\Gamma(W\beta_2)}{\Gamma(\beta_2)^w} \cdot \frac{\prod_{w_i}(n_0^{(w_i)}+\beta_2)}{\Gamma(n_0^{(w_i)}+W\beta_2)}$$

The average word likelihood of the HPB2 model is the same as the HPB model.

As illustrated in Fig. 3.31a, for both models, the likelihood increase as the number of topic increase, which means that a relatively larger topic numbers may potentially result in better modeling of testing data. However, it should be noted that there is a trade-off between topic numbers and convergence time of models. And, as we would see, the increase of

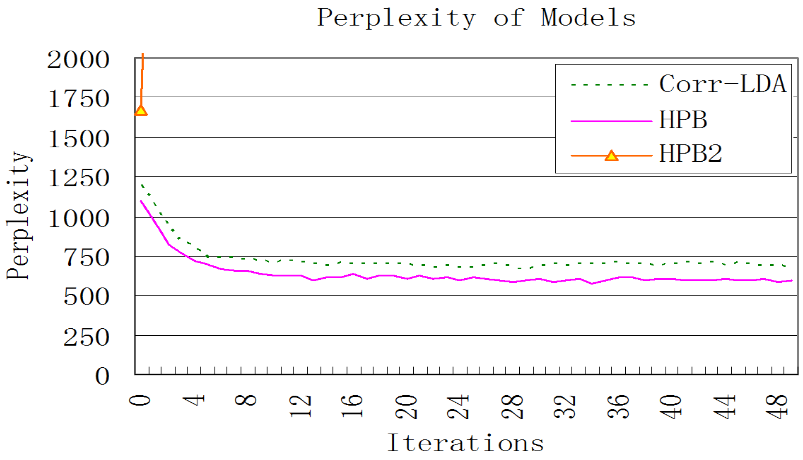


Figure 3.32: Perplexity over the iterations (number of topics equals 100)

topic number does not always lead to the improvement of predictive results.

In general, the log-likelihood of the extended Corr-LDA model and the HPB model are close, the difference between two models can be explained by the introduction of background topic in the HPB model.

Illustration of Discovered Latent Themes

One major objective of the proposed models is to uncover the latent topics from image-caption pairs and facilitate knowledge organization and understanding in online biomedical literatures. With this consideration, we visualize the discovered latent topics by providing the top-ranked words, top-ranked concepts (Fig. 3.20 and 3.33) and most related images (Fig. 3.33, with probability under each image). For this example, the latent topics are learnt by the HPB model, in which the topic number is 125. As illustrated in Fig. 3.20, the background topic depicts the contextual information of the biomedical journal, such as breast cancer, human body and tumor. The regular latent topics, on the other hand, reveal some

Topic28

Top words	Probability
P53	0.184625
mammary	0.096425
heterozygous	0.06976
adenocarcinoma	0.051299
carcinomas	0.038992
panels	0.036941
enhances	0.036941
deficiency	0.032839
knockout	0.032839
irradiated	0.032839
early	0.018481
developing	0.018481
tumours	0.01643
Balb	0.01643
genotypes	0.01643
Spontaneously	0.01643
MSM	0.01643
cHeA	0.014379
spares	0.012327
adjuvant	0.012327

Top Concepts	Concept Names
C0001418	Adenocarcinoma
C0858252	Breast Adenocarcinoma
C0007097	Carcinomas
C0206745	Knockout Mice
C0599772	knockout gene
C0025919	BALB C Mice
C1446974	cheA protein, E coli
C1307090	F1-20 protein, mouse
C0598034	BRCA2 Gene
C0677850	adjuvant therapy
C0001551	Immunoadjuvants
C0029463	Osteosarcoma
C1336745	Thymic Lymphoma
C0009085	Clustering
C0349966	Figs
C0315310	Salmonella uppsala
C0439828	Variable
C0557775	Van
C0439234	yr
C0906368	MPR1 transport protein

Topic70

Top words	Probability
immunostaining	0.057799
Receptor	0.040886
Intense	0.031021
Chains	0.029611
Infiltrative	0.028202
Complex	0.026792
Immunohistochemical	0.025383
Lobular	0.023974
comparative	0.022564
Interleukin	0.022564
Infiltrating	0.019745
corresponding	0.018336
Androgen	0.018336
Thymus	0.016927
Labeled	0.015517
IDC	0.014108
magnification	0.012699
Observed	0.012699
comparison	0.011289
Distinct	0.011289

Top Concepts	Concept Names
C0597357	Receptor
C0444498	In situ
C0439855	Complex
C0021764	Interleukin
C0009491	Comparative Study
C0291890	NR0B1
C0154073	Stage 0 Skin Cancer
C0002844	Androgenic Agents
C0205417	Lobular
C0034804	Receptors, Estrogen
C0264793	DCM
C0206692	Carcinoma, Lobular
C0796396	I-125
C0022262	Isotopes
C0123356	ILS
C0021010	Gm Allotype
C1101536	IL-2Ralpha
C0079004	B-Cell Subset
C0010834	Cytoplasm
C0456981	Specific antigen

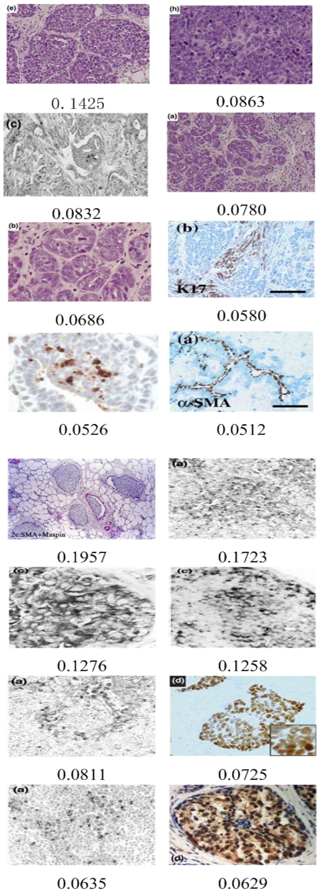


Figure 3.33: Illustration of discovered latent themes by HPB model

domain specific knowledge. As illustrated in Fig. 3.33, the top-ranked words, concepts and images of the uncovered latent topics have high semantic consistency. The top ranked words and concepts not only contain domain specific terms such as receptor, carcinomas, breast adenocarcinoma and Immunohistochemical, which help users to interpret the topics, but also provide many protein names and gene names that are related to the uncovered latent topic.

Image Annotation and Evaluation

The proposed probabilistic models are able to establish direct correlation among caption words, visual words and biomedical concept in biomedical image-caption pairs. Therefore, given the image content, a good model should be able to predict the missing captions. Next we automatically annotate caption words and concepts for images in the testing dataset based on the uncovered latent topics from training dataset, with both captions and concepts in testing dataset regarded as unknown (missing). The performance of automatic annotation is evaluated by perplexity and annotation accuracy using cross-validation.

In our experiment, we resort to the word caption perplexity as standard criteria of the annotation performance.

The perplexity of a set of testing image-caption pairs (for all $d \in D_{test}$) is defined as the exponential of the negative normalized predictive log-likelihood using the training model, in which the topic-word conditional probability: $p(w_i|z_{wi} = t)$ is obtained from the Gibbs sampling process of training dataset.

$ppx = exp\{-\frac{1}{N_W}\sum_{j=1}^{D}\sum_{i=1}^{W} log[\sum_{t=1}^{T} E(p(w = w_{j,i}|z = t))E(p(z = t|d = j))]\}$

With uncovered latent topics from training image-caption pairs, the estimation of prior probability of topic in a testing image can be approximated by running collapse Gibbs sampling over all the extracted visual words (no words or concepts used) in testing dataset (eq. (3.35)) using fixed visual word-topic conditional probability obtained from the Gibbs sampling process of training dataset.

$$p(y''_i = t|v_i, \mathbf{v}_{-i}, \mathbf{y}''_{-i}) \propto p(v_i|y''_i = t) \cdot \frac{\alpha + n^d_{-i,t}}{T\alpha + n^d_{-i,t}} \tag{3.35}$$

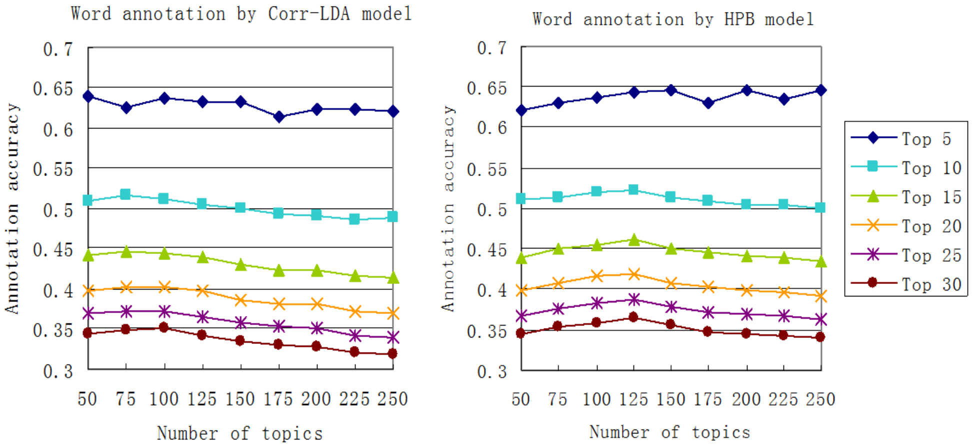


Figure 3.34: Comparison of word annotation accuracy over different topic numbers

After the convergence of the Gibbs sampling process, the probability for the ’missing’ caption words and concepts of an image can be calculated via the production of topic-word/concept conditional probability and the prior probability for each topic.

Recall that for HPB model, we assume no background topic for visual words, the prior for background topic in a document is approximated by averaging probability over the training dataset. Fig. 3.31b represents the perplexity of CorrLDA and HPB model over different topic numbers. The perplexity of HPB model is lower than that of the CorrLDA model, which indicates that HPB model generated from training data set is ‘less surprised’ by the testing data, thus, it demonstrates better ability in annotation. What is more, as the topic number increases, the perplexities of both models decrease first, and then increase, with 100 topics having the lowest perplexity. It appears that the increase of topic number does not always lead to persistent improvement of predictive ability.

Fig. 3.32 illustrates the perplexities over the iterations when the topic number is 100.

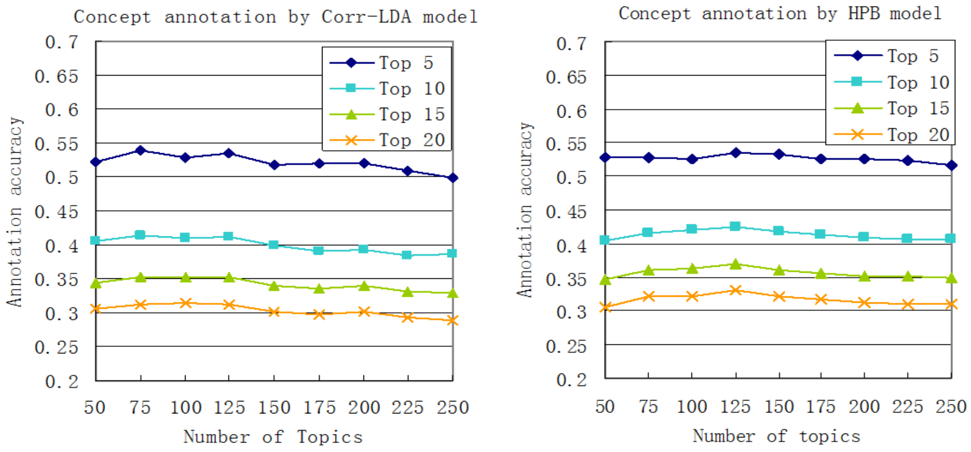


Figure 3.35: Comparison of concept annotation accuracy over different topic numbers

Although the HPB model appears to be more sophisticated than the Corr-LDA model, they converged in similar number of iterations. Recall that we have a variation of HPB model (named as the HPB2 model), which assumes that background words and concepts are related to certain image content (visual words). As shown in Fig. 3.32, the perplexity of HPB2 increases sharply and quickly exceeds 10000, which indicates that the Gibbs sampling process for this model fails to converge. Finally, over 90% of the entities in documents are assigned to the background topic (as a comparison, only about 1/10 of the words will be assigned to background topic when the Gibbs sampling process of HPB model converges). According to the perplexity results, there is no evidence that there exist a direct correlation between image content and background information in the caption.

When the prior probability of topics in a testing image is estimated (eq. (3.35)), the word and caption annotation for each document can be achieved by ranking words and concepts

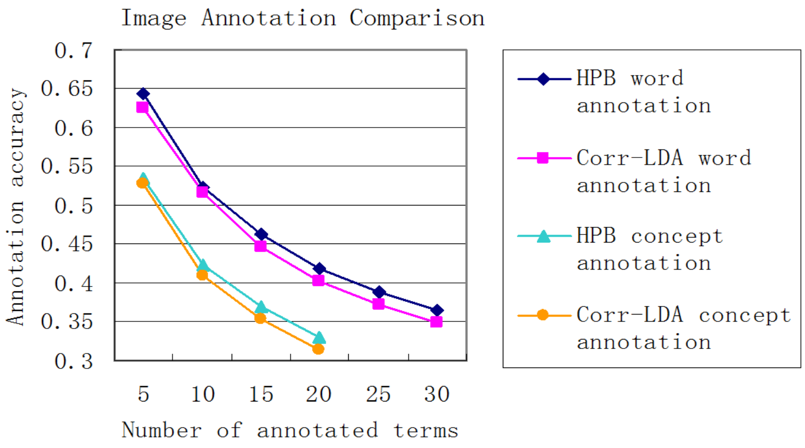


Figure 3.36: Image annotation comparison

with regard to the following probability.

$$\begin{cases} p(w_i|d_j) = \sum_{t=1}^{T} p(w = w_i|z = t)p(z = t|d = j) \\ p(c_i|d_j) = \sum_{t=1}^{T} p(c = c_i|z = t)p(z = t|d = j) \end{cases} \tag{3.36}$$

The words and concepts that achieve highest probability value in eq. (3.36) are used as the annotation of images. After that, the image annotations are compared to the original words and concepts in testing image-caption pairs for validation. During annotation evaluation, the cross-validation process repeats 5 times, and the results are averaged to produce the final annotation accuracy.

The accuracy of word and concepts annotation over different topic numbers is illustrated in Fig. 3.34 and Fig 3.35. Specifically, Fig. 3.34 represents the annotation accuracy from top 5 annotation words to top 30, while Fig. 3.35 provides the annotation accuracy from top 5 concepts to top 20. According to the experiment results, the HPB achieves best annotation performance when topic number is 150, while the Corr-LDA model achieves best performance with 100 topics. As the topic number increases, the annotation accuracy of both models increase first, and then decrease, which is consistent with the results in perplexity comparison.

The annotation accuracy of extended Corr-LDA model and the proposed HPB model is compared using their best annotation performance (i.e. 100 topics for Corr-LDA model, and 125 topics for HPB model). As illustrated in Fig. 3.36, the HPB model is consistently better than the extended Corr-LDA model in both word annotation and concept annotation tasks, which is consistent with the perplexity comparison results. Furthermore, according to Fig. 3.34-3.36, the performance of HPB model drop slower than the Corr-LDA model when considering the annotation accuracy of large number of annotation terms. The result indicates that HPB model is more robust and is able to achieve better performance in annotating less frequent terms.

Conclusions

The contribution of this part of thesis is twofold. First, we proposed a novel HPB model to integrate background information in topic learning, incorporating contextual information to interpret the uncovered latent topic and improve the image annotation accuracy. Second, in our experiments, we discovered that there is no direct correlation between image content and the background information in the captions. In other word, the extracted visual words from images have nothing to do with the background topic. It is unnecessary to incorporate contextual information when modeling the image contents.

3.6.3 Probabilistic Topic-Connection Model for Co-existing Image Features and Annotations

The image dataset used in our study is downloaded from the ImageNet database (http://www.image-net.org/) under the granted access permission, following the term of access. The ImageNet is built on the hierarchical ontology structure provided by WordNet, in which each node involves a group of images that depict a particular concept named as a synonym set, or 'synset'. Specifically, we download a total of 508 synsets under the 'flower' sub-tree, 1473 synsets under the 'mammal' sub-tree and 1118 synsets under the 'tree' sub-tree. Following the term mapping schema in Section 3.3.2, we map each synset to a Wikipedia article that describes the same concept. Then, we parse the structured content of Wikipedia articles and apply a rule-based method to identify the explanative sections. Unrelated sections such as 'External-links' and 'References' are removed from the articles. After that, to ensure the quality of text description, we filter out articles with insufficient words ($<$ 200 words). The qualified articles then serve as text description for corresponding ImageNet synsets. In total, we obtain comprehensive text descriptions for 1452 synsets (330, 562 and 560 synsets for sub-trees 'flower', 'mammal' and 'tree', respectively).

For each of the 1452 synsets, we randomly select 5 images from the corresponding image group and adjust them to normalized size (640×480 pixels). After that, we replicate the text descriptions to each of the 5 images, resulting in 5 image-text pairs. As introduced in Section 3.3.2, we make index for single-words and multiple word phrases in the text descriptions, and extract visual-word features as well as MESR region features from images. In total, we indexed 5,699,505 word tokens which belong to 35,744 different words, 624,205 multiple word phrases from a total number of 13078 unique phrases, 7,945,075 visual words (an average of 1095 visual words per image)from a vocabulary size of 2000, and a total of 924,924 MSER region features(an average of 127 MSER regions per image). The original dataset is divided into 5 subsets with equal size. Of the 5 subsets, one subset (20%) is retained as the validation data for testing the model, and the remaining 4 subsets (80%) are used as training data. For image annotation evaluation, the cross-validation process repeats 5 times, with each of the 5 subsets used once as the validation data. After that, we take the average results for evaluation.

The estimation of the proposed probabilistic topic model is achieved by performing Gibbs Sampling over training dataset until convergence (generally, the model takes less than 100 iterations to converge). Once the topic model is estimated from the training dataset, we will be able to evaluate it by log-likelihood and perplexity. The inferred latent topics are also visualized.

Log-Likelihood Comparison

Log-likelihood is one of the standard criteria for generative model evaluation. It provides a quantitative measurement of how well a topic model fits the training data. The score of log-likelihood (which is a negative number) is the higher the better. In practice, the log-likelihood of elements given latent topics can be calculated by integrating out all the latent variables. In our study, we are interested in which topic model is more suitable to study the latent patterns of image features. Thus, instead of calculating word-likelihood, we choose to evaluate the log-likelihood of visual words for both models. In the proposed probabilistic topic-connection (PTC)model, The marginal likelihood of visual words v given all the visualtopics y is $p(v|y)$, which can be calculated by integrating out latent variables ψ:

$$p(\mathbf{v}|\mathbf{y}) = \prod_{j=1}^{T_2}[\int_{\psi_j} p(\mathbf{v}|y_j,\psi_j)p(\psi_j|y_j)\mathrm{d}\psi_j] = [\frac{\Gamma(V\beta_v)}{\Gamma(\beta_v)^V}]^{T_2} \cdot \prod_{j=1}^{T_2} \frac{\prod_v (C_{vj}^{VT_2}+\beta_v)}{\Gamma(\sum_{v'} C_{v'j}^{VT_2}+V\beta_v)} \quad (3.37)$$

The final log-likelihood of visual words is obtained by taking the logarithm of eq. (3.37) and averaging the resulting summation by V.

For the extended Corr-LDA model, the log-likelihood can be calculated in a similar way, the only difference is, instead of using their own latent topics, the visual words in Corr-LDA model directly use latent topics generated from text words.

In Fig. 3.37a, we plot the log-likelihood for both models under different topic number (to make this comparison fair, the number of word topic and visual topics are made equal). It shows that our model has higher log-likelihood than Corr-LDA model, which means that our model fits training data better. It also shows that the log-likelihood of both models increase as the number of topic increase, which suggests that a relatively greater topic number may potentially fit the training data better. However, it should be noted that there is a trade-off between topic numbers and convergence time of model estimation, and the unbounded increase of topic number may results in an over-fitting problem.

Perplexity Comparison

The perplexity is a standard criterion for topic models that evaluates how well the model predicts the new data. Specifically, the perplexity of a set of testing documents is defined as the exponential of the negative normalized per-word predictive log-likelihood using parameters from the trained topic model. The score of perplexity is the lower the better.

With uncovered latent topics from training image-text pairs, the problem of estimating topic priors in testing images can be approximated by performing Gibbs sampling over observations of image features, while keeping all the topic-entity conditional probability fixed. It should be noted that we need to know the posterior probability of word topic indicators given visual topics: $p(s|y)$ when estimating the new document-level word topic prior. In our study, this probability is approximated by counting the number of evidences across the training dataset.

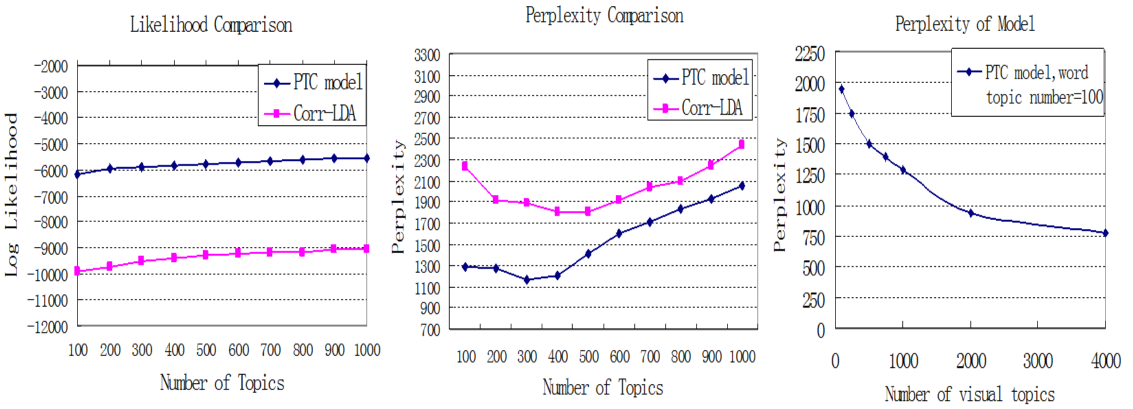
Upon the convergence of the Gibbs sampling process over testing data, the word perplexity of testing image-text pairs is:

$$Perplexity = exp[\frac{-\sum_{d_{test}} logp(\mathbf{w}^d, \mathbf{p}^d | \mathbf{v}^d, \mathbf{r}^d)}{\sum_{d_{test}} (N_w^d + N_p^d)}] \tag{3.38}$$

One advantage of our model is that it assigns different topic numbers to different types of data, which makes this model more suitable to deal with image and associated text. Fig.3.37 b represents the perplexity comparison between our model and the Corr-LDA model as the increase of word topic number, in which the number of visual topics in our model is fixed to 1000. It shows that the perplexity of our model is consistently lower than Corr-LDA model, which suggests that our model is ‘less surprised’ by the testing data, thus demonstrates better performance. Also, it shows that the predictive ability of our model may benefit from greater visual topic number, as it tend to have lower perplexity as the visual topic number increases (Fig. 3.37c).

Illustration of Inferred Latent Topics

In order to better interpret the uncovered latent topics, we visualize the word topics by providing the top-ranked words, top-ranked phrases and most related images. As represented



(a) Likelihood comparison **b) Perplexity comparison (# of visual topics=1000)** **(c) Perplexity vs. visual topic numbers**

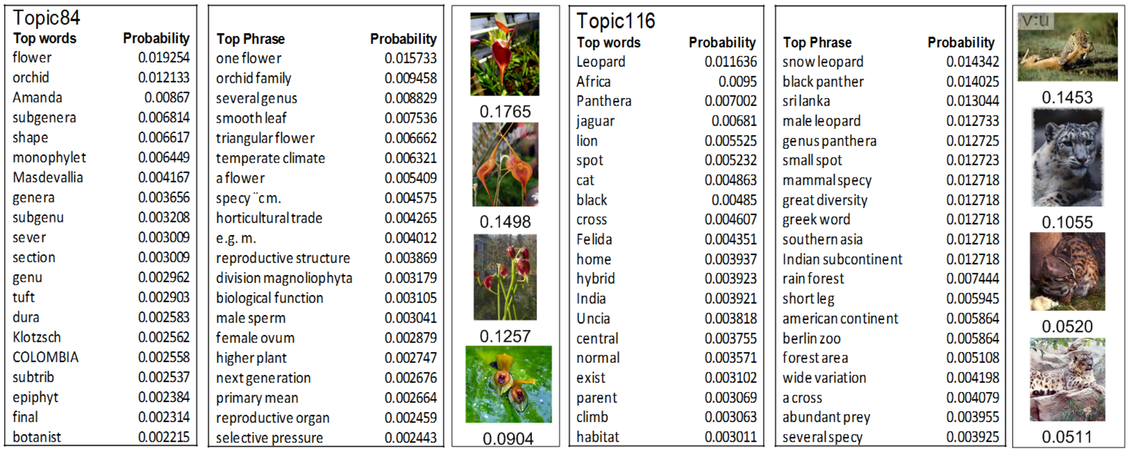
Figure 3.37: The likelihood and perplexity comparison of the proposed PTC model and the extended Corr-LDA model

in Fig. 3.38, the words and phrases are sorted by their probability of being generated from a word topic, while images are sorted by the probability of containing that word topic (by counting the topic indicator variables of image features).

In Fig.3.38, we present two examples of uncovered latent word topics. The former one is a topic related to the concept of ‘orchid’, while the later one is a topic related to the concept of ‘leopard’. By providing a combination of words, multiple word phrases and images, it becomes much easier to interpret the domain knowledge captured by each topic. As we can see, the uncovered latent topics show high consistency to semantic concepts.

Conclusions

In this section, a probabilistic topic-connection model is proposed to deal with the problem of modeling images and associated text description. Specifically, new latent variables have been introduced to allow for more flexible sampling of word topics and visual topics, in which one word topic may connect to multiple visual topics.The proposed model provides



Topic84

Top words	Probability
flower	0.019254
orchid	0.012133
Amanda	0.00867
subgenera	0.006814
shape	0.006617
monophylet	0.006449
Masdevallia	0.004167
genera	0.003656
subgenu	0.003208
sever	0.003009
section	0.003009
genu	0.002962
tuft	0.002903
dura	0.002583
Klotzsch	0.002562
COLOMBIA	0.002558
subtrib	0.002537
epiphyt	0.002384
final	0.002314
botanist	0.002215

Top Phrase	Probability
one flower	0.015733
orchid family	0.009458
several genus	0.008829
smooth leaf	0.007536
triangular flower	0.006662
temperate climate	0.006321
a flower	0.005409
specy ¨cm.	0.004575
horticultural trade	0.004265
e.g. m.	0.004012
reproductive structure	0.003869
division magnoliophyta	0.003179
biological function	0.003105
male sperm	0.003041
female ovum	0.002879
higher plant	0.002747
next generation	0.002676
primary mean	0.002664
reproductive organ	0.002459
selective pressure	0.002443

Topic116

Top words	Probability
Leopard	0.011636
Africa	0.0095
Panthera	0.007002
jaguar	0.00681
lion	0.005525
spot	0.005232
cat	0.004863
black	0.00485
cross	0.004607
Felida	0.004351
home	0.003937
hybrid	0.003923
India	0.003921
Uncia	0.003818
central	0.003755
normal	0.003571
exist	0.003102
parent	0.003069
climb	0.003063
habitat	0.003011

Top Phrase	Probability
snow leopard	0.014342
black panther	0.014025
sri lanka	0.013044
male leopard	0.012733
genus panthera	0.012725
small spot	0.012723
mammal specy	0.012718
great diversity	0.012718
greek word	0.012718
southern asia	0.012718
Indian subcontinent	0.012718
rain forest	0.007444
short leg	0.005945
american continent	0.005864
berlin zoo	0.005864
forest area	0.005108
wide variation	0.004198
a cross	0.004079
abundant prey	0.003955
several specy	0.003925

Figure 3.38: Illustration of uncovered latent topics by PTC model

better representation of the connection between latent semantic topics and latent image patterns, thus achieves better performance compared to the traditional Corr-LDA model.

3.6.4 Perspective Hierarchical Dirichlet Process for User-Tagged Image Modeling

In this section, we investigate the performance of the proposed pHDP model in automatic image tagging experiments using the MIR-Flickr dataset, which is composed of 25000 images covering a wide spectrum of image categories(contributed by a total of 9862 Flickr users). In total,there are 302447 tags, with a vocabulary size (number of unique tags) of 64037; thus the average number of tags per image is 8.94. In the image tagging experiment, we use a 50% subset of the MIR-Flickr collection as training data and the other 50% as testing data (with tags removed). On constructing the two subsets, we ensure that tagged images from the same user are equally split to both subsets. The values of global concentration parameter γ and the user perspective number L are determined by perplexity

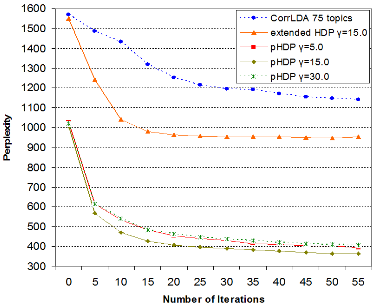


Figure 3.39: The comparison of Perplexity changing over iterations

comparison on a serial of values. Other hyper-parameters (such as Dirichlet distribution priors: $\alpha_u, \xi_y, \xi_t, \eta, \zeta$) are set in prior and fixed during the experiments. The prediction of image tags for the testing images is achieved by performing another Gibbs sampling on testing images to estimate the document-level distribution of switch variable and semantic components, with a fixed set of semantic components and user perspectives estimated from the training dataset. On the convergence of Gibbs sampling, the probability of tagging an image j from user u with tag t_j is:

$$p(t_j) = p(x_{jt} = 0,1) \sum_{k=1}^{K} p(t_j|z_k) p_{test}(z_k|j) + p(x_{jt} = 2) \sum_{l=1}^{L} p(t_j|p_l) p_{test}(p_l|u) \tag{3.39}$$

The performance is evaluated by perplexity and tagging accuracy.

Perplexity Comparison

The perplexity is a standard criterion for generative probabilistic models that evaluates how well the model predicts the testing data. The perplexity of a testing image dataset D_{test} is:

$$perplexity(D_{test}) = exp[\frac{-\sum_{j=1}^{D_{test}} log(p(\mathbf{t}_j))}{\sum_{j=1}^{D_{test}} N_j^t}] \tag{3.40}$$

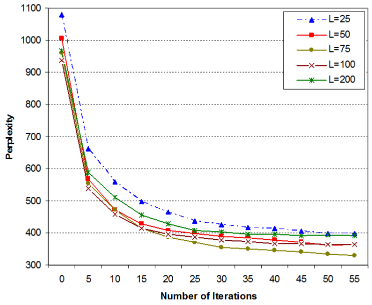


Figure 3.40: Perplexityas perspective number changing(γ=15.0)

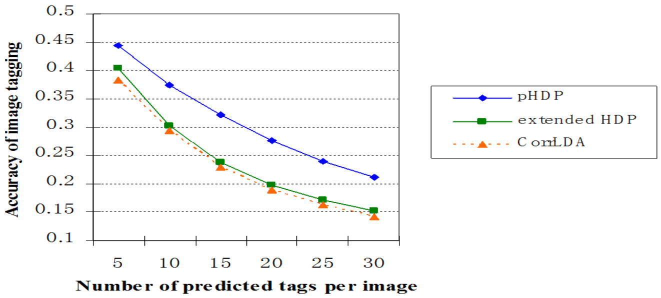


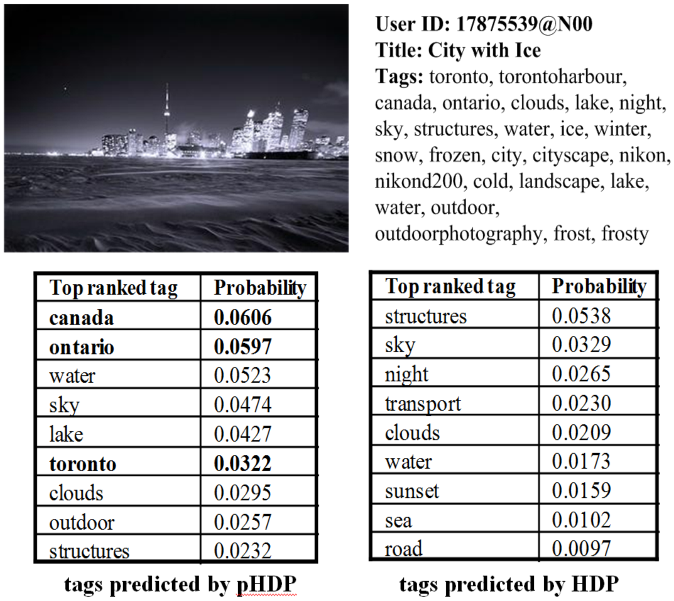
Figure 3.41: Average image tagging accuracy comparison of the proposed pHDP models and baseline models

The perplexity score for a model is the lower the better. Fig. 3.39 shows the perplexity changing of the proposed pHDP model and baseline modes (CorrLDA model and HDP model) over the iterations during the Gibbs sampling process. We test pHDP model on a serial of γ values. For CorrLDA model and HDP model, we only show their perplexity scores under the optimal parameter settings (i.e. CorrLDA model with 75 topics and HDP model with $\gamma = 15.0$). The results show that pHDP model achieve best performance with $\gamma = 15.0$ and it outperforms both HDP model and CorrLDA model. Fig 3.40 represents the perplexity of pHDP model under different perspective numbers. The optimal perspective number is L=75.

Image Tagging Accuracy Evaluation

Using Eq.(3.39), we calculate the probability of tagging an image j from user u with different tags. Tags with highest probability are used for tagging. After that, the predicted top-ranked image tags are compared with the ground truth for validation. If a predicted tag finds exact match in the ground truth tags, it will be considered as one hit. The ratios of hit numbers over the predicted tag numbers are averaged to produce the final annotation accuracy.

Fig.3.42, Fig.3.43, Fig.3.44 illustrates examples of image tagging results. Fig.3.42 is an image shows a winter night in Toronto,Ontario, Canada.The ground truth image tags involve both location tags ('ontario', 'canada') and topic tags (such as 'clouds', 'lake', 'night', 'sky' and 'water'). However, the image content alone provides little clues about the location. Further studies indicate that other images contributed by the same user are also tagged with 'ontario' and 'canada'. This may suggest that the user lives in Ontario, Canada and contribute pictures taken from the same location. During the pHDP modeling, this user contextual information is captured as a part of the user's perspectives. When tagging a new image from the same user, the pHDP model will smooth the document-level predictive tag distribution with users perspective and allow for tagging with location tags (Fig. 3.42 and 3.43, highlighted in bold). Fig 3.44 is contributed a user from Malaysia. Similarly, the user contextual information is captured in users perspectives. Thus the pHDP model succeeds in tagging image with both location tags (such as 'malaysia') and type tags (camera settings, like 'nikon'). As shown in Fig. 3.44, tags predicted by the pHDP model also involve subjective tag, like 'interestingness', which demonstrates that the pHDP model is also capable



User ID: 17875539@N00
Title: City with Ice
Tags: toronto, torontoharbour, canada, ontario, clouds, lake, night, sky, structures, water, ice, winter, snow, frozen, city, cityscape, nikon, nikond200, cold, landscape, lake, water, outdoor, outdoorphotography, frost, frosty

Top ranked tag	Probability
canada	**0.0606**
ontario	**0.0597**
water	0.0523
sky	0.0474
lake	0.0427
toronto	**0.0322**
clouds	0.0295
outdoor	0.0257
structures	0.0232

tags predicted by pHDP

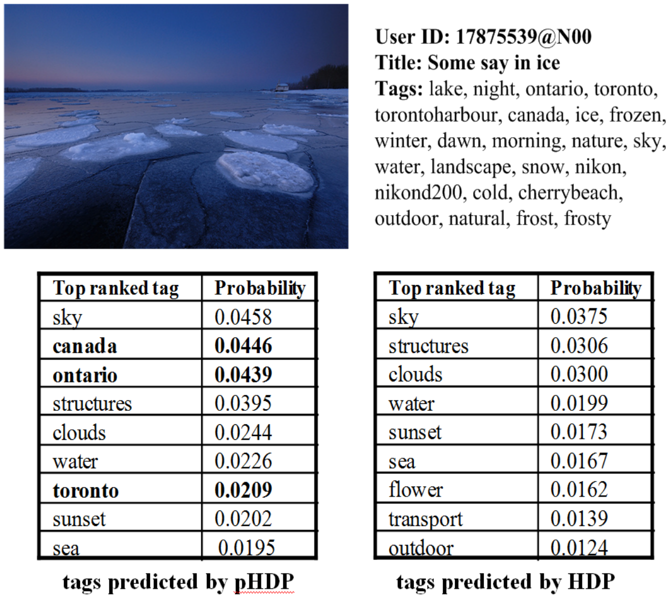
Top ranked tag	Probability
structures	0.0538
sky	0.0329
night	0.0265
transport	0.0230
clouds	0.0209
water	0.0173
sunset	0.0159
sea	0.0102
road	0.0097

tags predicted by HDP

Figure 3.42: Tagging results of image entitled ‘City with ice’

of modeling user’s subjective feelings. As a comparison, the HDP model fails to predict either location tags or subjective tags since it only relies on image content to make tag predictions.

Fig 3.44 shows the overall image tagging accuracy (averaged over the testing dataset) of different models under their optimal parameter settings. It should be noted that the accuracy is calculated based on exact match, so it won’t take into account the synonym tags. In other words, predicted tags like ‘human’ and ‘female’ will be considered as unmatched with respect to the ground truth tag ‘girl’. According to the result, the HDP model doesn’t show much improvement in the tagging accuracy compared to the CorrLDA model. It’s reasonable because the CorrLDA model is in essence a finite case of HDP model. Under optimal parameter settings, their performance should be similar. The pHDP model, as it



User ID: 17875539@N00
Title: Some say in ice
Tags: lake, night, ontario, toronto, torontoharbour, canada, ice, frozen, winter, dawn, morning, nature, sky, water, landscape, snow, nikon, nikond200, cold, cherrybeach, outdoor, natural, frost, frosty

Top ranked tag	Probability
sky	0.0458
canada	**0.0446**
ontario	**0.0439**
structures	0.0395
clouds	0.0244
water	0.0226
toronto	**0.0209**
sunset	0.0202
sea	0.0195

tags predicted by pHDP

Top ranked tag	Probability
sky	0.0375
structures	0.0306
clouds	0.0300
water	0.0199
sunset	0.0173
sea	0.0167
flower	0.0162
transport	0.0139
outdoor	0.0124

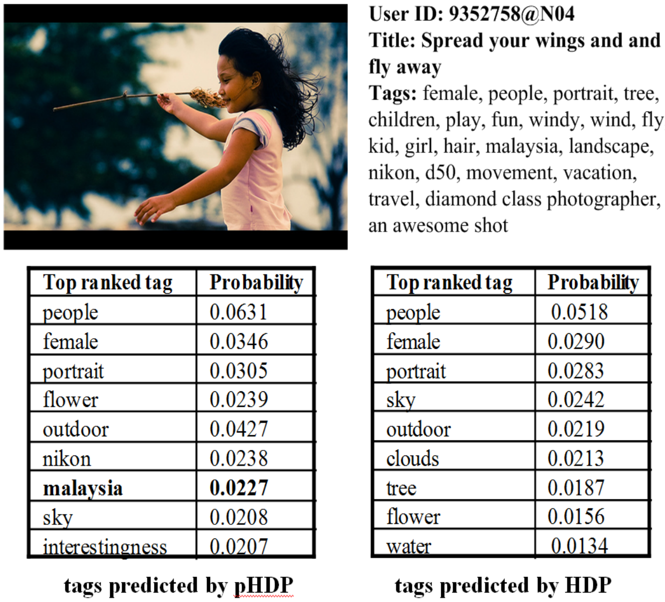
tags predicted by HDP

Figure 3.43: Tagging results of image entitled ‘Some say in ice’

integrates the user perspective information, significantly outperforms both CorrLDA model and HDP model in predicting image tags for different users.

Conclusions

In this section, we proposed a perspective Hierarchical Dirichlet Process (pHDP) model to deal with user-tagged image modeling. The contribution is two fold. Firstly, we associate image features with image tags. Secondly, we incorporate the user’s perspectives into the image tag generation process and introduce new latent variables to determine if an image tag is generated from user’s perspectives or from the image content. Therefore, the model is capable of extracting both embedded semantic components and user’s perspectives from user-tagged images. Based on the proposed pHDP model, we achieve automatic image tagging with user’s perspective. Experimental results show that the pHDP model achieves



User ID: 9352758@N04
Title: Spread your wings and and fly away
Tags: female, people, portrait, tree, children, play, fun, windy, wind, fly, kid, girl, hair, malaysia, landscape, nikon, d50, movement, vacation, travel, diamond class photographer, an awesome shot

Top ranked tag	Probability
people	0.0631
female	0.0346
portrait	0.0305
flower	0.0239
outdoor	0.0427
nikon	0.0238
malaysia	**0.0227**
sky	0.0208
interestingness	0.0207

tags predicted by pHDP

Top ranked tag	Probability
people	0.0518
female	0.0290
portrait	0.0283
sky	0.0242
outdoor	0.0219
clouds	0.0213
tree	0.0187
flower	0.0156
water	0.0134

tags predicted by HDP

Figure 3.44: Tagging results of image entitled ‘Spread your wings and fly away’

better image tagging performance compared to state-of-the-art topic models.

3.6.5 Modeling Semantic Relations between Visual Attributes and Object Categories via Dirichlet Forest Prior

In this section, we evaluate the performance of the proposed methods, including automatic attribute identification, object categorization, and modeling the semantic relations between visual attributes and object categories.

Datasets and Experimental Setup

In learning the visual attributes of object categories, we use the Animals with Attributes (AwA) dataset introduced by [57] (which is a fraction of ImageNet database). The AwA dataset consists of 50 mammal object categories with a human provided attribute inventory and corresponding object class-attribute associations. In our experiment, we split the dataset into 80% training and 20% testing (i.e. 24,295 trainingimages and 6,180 test images) for learning the attribute classifiers.We map all the 50 AwA categories to the corresponding synsets (identified withWordNet ID) under the ImageNet hierarchical taxonomy, from which we are able to calculate the semantic metric among different categories. Also, with the help of word entities from the corresponding WordNet ID (wnid), we download 75 Wikipedia articles that share the same lexicalized entry as WordNet entities (considering synonyms). The Wikipedia articles are then considered as the knowledge base of natural language description for corresponding ImageNet synsets or AwA categories.

Object Categorization and Attribute Identification

The first part of our experiments is semi-supervised learning for object categorization. As mentioned in Section 3.5.1, bounding box detection is performed to ensure that we identify clean attributes from each object category. We are the learning process from 50 labeled images bounding box per class (i.e. 2500 image bounding boxes in total) for training, The semi-supervised SVM use 50 labeled bounding boxes and an addition of 50 unlabeled bounding box samples (from preliminary detection in Section 3.5.1). We use another 100 images from each class (other than the 5000 training image) for testing. For each test image i, if it is correctly classified then the flat error $e_i = 0$, if it is not correctly classified then$e_i = 1$. Fig.3.45 shows the Receiver Operating Characteristic (ROC) curves of the object categorization results. Each curve is the result of the one-vs-all semi-supervised SVM categorical classifier on HOG-LBP feature. The Area Under Curves (AUC) are also provided. Higher AUC indicate better classify performance.For example, the AUC scores of giant panda (AUC=0.912) and zebra (AUC=0.975) are among the highest, indicating that

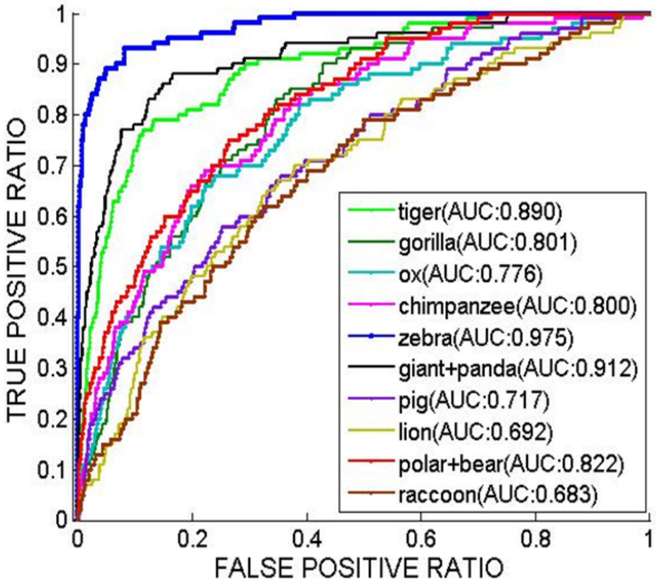


Figure 3.45: Part of the object classification results plotted in receiver operating characteristic (ROC) curves. Each curve is the result of the one-vs-all semi-supervised SVM categorical classifier on HOG-LBP feature.

these object categories are well represented by HOG-LBP feature and are well separated in the feature space. The categorization results of raccoon and lion is fair, which are possibly caused by the high diversity of object appearance among training samples.

In Fig.3.46, we compare our categorization method (semi-supervised SVM with HOG-LBP features) to two state-of-the-art approaches, i.e. SVM classifier using spatial bag of word (sBoW) features and SVM using HOG-LBP features. Specifically, Fig. 3.46a shows the Average Flat Error (AFE) with respect test images with top predictive scores. The AFE score is defined as: $e = \frac{1}{N}\sum_{i=1}^{N} e_i, e \in [0,1], N = 5000$, the lower AFE score indicates better classification performance. Fig. 3.46 b represents the Average Hierarchical Error (AHE) of different approaches. Supposing that the image from class i is mis-classified as class j,

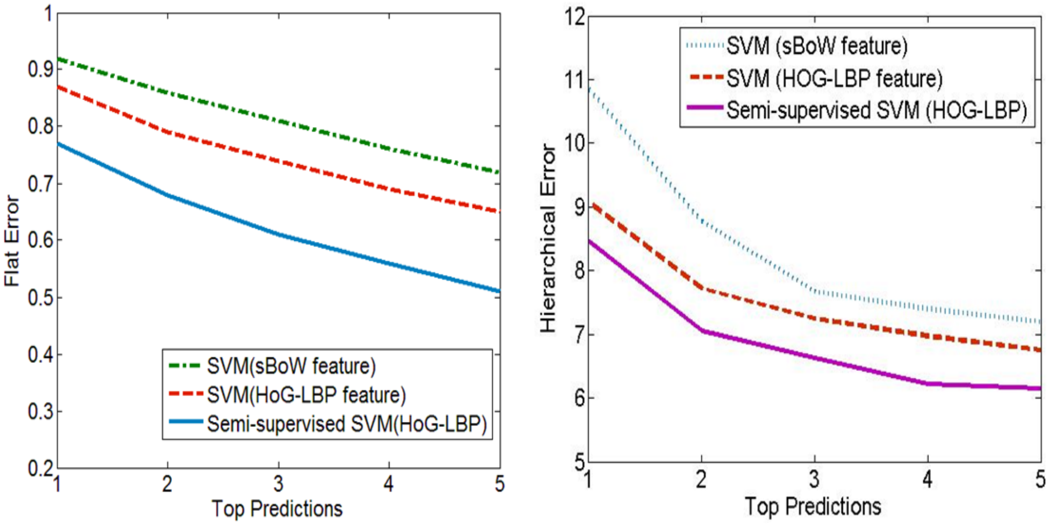


Figure 3.46: Performance comparison of proposed method (semi-supervised SVM + HOG-LBP features) with state-of-the-art approaches (a) Average Flat Error (AFE)(b)Average Hierarchical Error (AHE)

and $\pi(i,j)$ is the lowest common ancestor between class i and j in the hierarchy of ImageNet taxonomy. The height $h(\pi(i,j))$ of node $\pi(i,j)$ on the hierarchy is then defined as the length of the longest path to one of its leaf node. Leaf nodes have height 0. The AHE is the average of $h(\pi(i,j))$ for all the testing images, , the lower AHE score indicate higher semantic accuracy in object categorization. As shown in Fig. 3.46, the proposed object categorization method consistently outperforms the state-of-the-art approaches under both AFE and AHE comparisons.

Fig. 3.47 shows the confusion matrix of object categorization in all the 50 AWA categories. By comparing the confusion matrix of semi-supervised SVM classification and the confusion matrix categorization results are smoothed by the signature of the most relevant attributes, we can see that, the performance of object categorization can be signif-

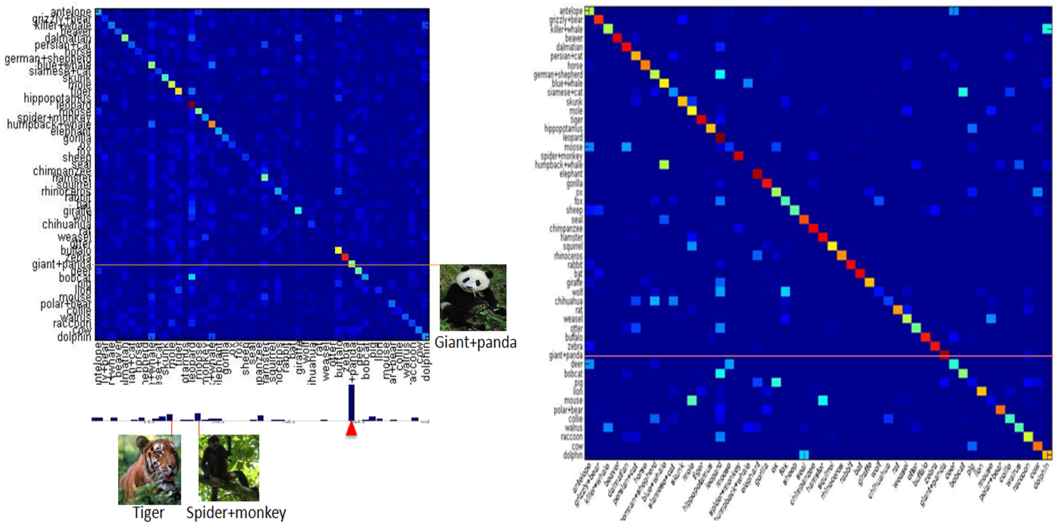


Figure 3.47: Confusion matrix of classifying 50 AWA animal classes.Upper: confusion matrix by semi-supervised SVM classifier. Lower: confusion matrix after object-attribute signatures being introduced in, categorization results are smoothed by (eq. 9), with K=10 most relevant visual attributes used.

icantly improved when attribute signatures are introduced. For example, in the original semi-supervised SVM classification results (upper part of Fig.3.47), the giant panda category is to some extends confused with the tiger category and the spider monkey category. However, after introducing in the object-attribute signatures and smoothing the categorization results with posterior object-attribute prediction model, the categorization ambiguity is mostly eliminated (lower part of Fig.3.47). The significantly reduced categorization ambiguity across the 50 AWA animal classes (Fig. 3.47) evidences the effectiveness of identified attribute-object relations.

Model Estimation and Illustration

In this section, we perform both quantitative and qualitative evaluation on the performance of proposed topic model. The quantitative evaluation includes comparing both log-likelihood and perplexity, while qualitative evaluation is achieved by visualizing the inferred latent topics and evaluate its relevance to the object-attribute relations.

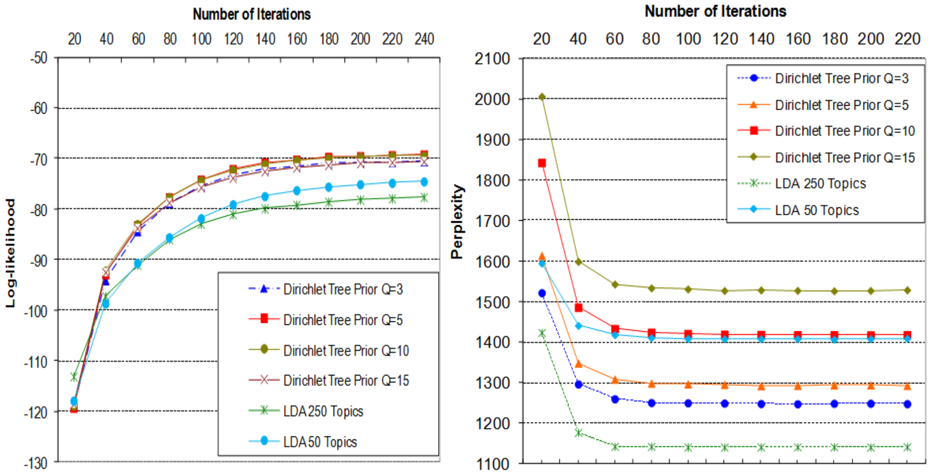
Log-likelihood is one of the standard criteria in generative model evaluation. It provides a quantitative measurement of how well a topic model fits the training data. The score of log-likelihood (which is a negative real number) is the higher the better. In practice, the log-likelihood of words given latent topics can be calculated by integrating out all the latent variables:

$$\begin{aligned} p(\mathbf{w}|\mathbf{z}) &= \prod_{t=1}^{T} \Big(\int_{\varphi_{z_t}} p(\mathbf{w}|z_t, \varphi_{z_t}) \, p(\varphi_{z_t}|z_t) \, \mathrm{d}\varphi_{z_t} \Big) \\ &= \Big(\frac{\Gamma(W\beta)}{\Gamma(\beta)^{W}}\Big)^{T} \prod_{t=1}^{T} \frac{\prod_{w_i} \Gamma\left(n_t^{(w_i)} + \beta\right)}{\Gamma\left(n_t^{(\cdot)} + W\beta\right)} \cdot \frac{\Gamma(W\eta)}{(\eta)^{W}} \cdot \frac{\prod_{w_i} \Gamma\left(n_0^{(w_i)} + \eta\right)}{\Gamma\left(n_t^{(\cdot)} + W\eta\right)} \end{aligned} \tag{3.41}$$

The perplexity is another standard criterion for generative probabilistic models that evaluates how well the model predicts the testing data. The perplexity of a testing dataset D_{test} is:

$$perplexity\left(D_{test}\right) = exp\left(\frac{-\sum_{j=1}^{D_{test}} log\left(p\left(t_j\right)\right)}{\sum_{j=1}^{D_{test}} N_j^t}\right) \tag{3.42}$$

The perplexity score for a model is the lower the better. Fig. 3.48a represents the log-likelihood comparison results between our proposed model and the LDA model over the iterations. As we can see from Fig. 3.48a, our proposed topic model has consistently higher log-likelihood than standard LDA model, which can be explained by the introduced Dirichlet Forest priors, which make our model fit better to training data than the LDA model. Fig. 3.48b shows the comparison of perplexity between our model and the LDA

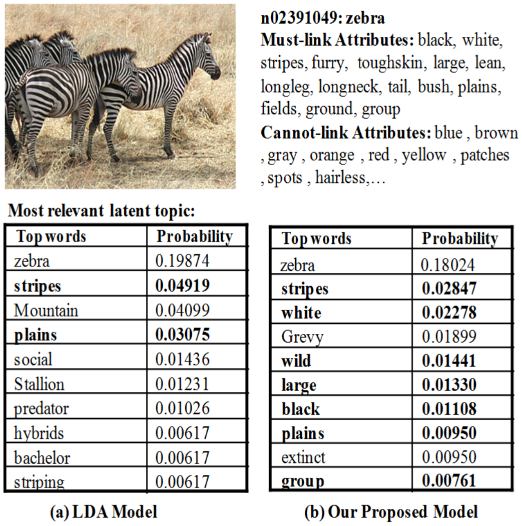


a) Log-likelihood comparison (using T=50)

b)Perplexity comparison (using T=50)

Figure 3.48: Log-likelihood and perplexity comparison between proposed model and LDA model over the iterations

model over the iterations. Our model achieves best perplexity scores when Q=3, while the LDA model achieves best perplexity scores when topic number is 250. Although LDA model has relative lower perplexity score compared to our model, however, as we can see in the next section, the LDA model may not be able to accurately link object category to its attributes. On the convergence of the Markov Chain Monte Carlo and Gibbs sampling process, the conditional probability of each word/entity given each inferred latent topic can be obtained. In Fig. 3.49 - Fig.3.51, we illustrate the qualitative evaluations of 3 ImageNet object categories (i.e. n02391049:zebra, n02129165:lion and n02581957:dolphin), including the category names, the identified visual attributes, the Must-Links and Cannot-Link from aggregated attribute-object concurrence map. We also visualize the most relevant inferred latent topics with respect to each object category name entity. The relevance between object category name entities and the inferred latent topics can be obtained by calculating the



n02391049: zebra
Must-link Attributes: black, white, stripes, furry, toughskin, large, lean, longleg, longneck, tail, bush, plains, fields, ground, group
Cannot-link Attributes: blue , brown , gray , orange , red , yellow , patches , spots , hairless,…

Most relevant latent topic:

Top words	Probability
zebra	0.19874
stripes	**0.04919**
Mountain	0.04099
plains	**0.03075**
social	0.01436
Stallion	0.01231
predator	0.01026
hybrids	0.00617
bachelor	0.00617
striping	0.00617

(a) LDA Model

Top words	Probability
zebra	0.18024
stripes	**0.02847**
white	**0.02278**
Grevy	0.01899
wild	**0.01441**
large	**0.01330**
black	**0.01108**
plains	**0.00950**
extinct	0.00950
group	**0.00761**

(b) Our Proposed Model

Figure 3.49: The most relevant latent topic of object category ‘zebra’

Mutual Information(MI) score.

The calculation of MI between a specific word entity and a latent topic is shown as eq. (3.43), in which R_g and Z_t are binary indicator variables corresponding to the word and the latent topic, respectively. The variable pair(R_g, Z_t) indicates the case that latent topic Z_t being assigned to word entity R_g.

$$MI\left(R_g, Z_t\right) = p\left(R_g, Z_t\right) log \frac{p\left(R_g, Z_t\right)}{p\left(R_g\right) p\left(Z_t\right)} \tag{3.43}$$

Given the training data, both the joint probability $p(R_g, Z_t)$ and the marginal probabilities $p(R_g)$ and $p(Z_t)$ can be empirically estimated by counting the number of evidences over the training dataset.

As we can see from Fig. 3.49 - Fig. 3.51, for the LDA model, the inferred latent topic that is most relevant to the object category doesn’t contain much visual attribute names associated

N02129165 : lion
Must-link Attributes: brown, yellow, furry, big, lean, pads, paws, tail, claws, walks, muscular, quadrapedal, desert, bush, ground
Cannot-link Attributes : black , white , blue , gray , orange , red , patches , spots , stripes , hairless , toughskin, fly, swim,…

Most relevant latent topic:

Top words	Probability
lion	0.18134
cubs	0.03486
Asiatic	0.01932
subspecies	0.01649
hunting	0.01272
spotted	0.01272
incidents	0.00942
American	0.00926
Barbary	0.00848
ligers	0.00613

(a) LDA model

Top words	Probability
lion	0.17802
wild	**0.01949**
kills	0.01860
cats	0.01861
desert	**0.01684**
Asiatic	0.00975
forest	**0.00798**
food	**0.00798**
big	**0.00709**
away	0.00709

(b) Our Proposed Model

Figure 3.50: The most relevant latent topics of object category ‘lion’

with that object category. On the contrary, the latent topics inferred from our model have a lot of important visual attributes among the top-ranked words of the most relevant latent topic. Specifically, for the object category ‘dolphin’ in Fig. 3.51, the most relevant latent topic inferred from our model involve visual attributes that are highly consistent with the identified Must-Link relations associated with the dolphin category such as ‘blue’, ‘water’, ‘white’, ‘fish’, etc. while the most relevant latent topic inferred by LDA model doesn‘t involve any visual attributes associated with dolphin in its top-ranked words. Similarly, for the object category ‘zebra’ in Fig. 3.49, the most relevant latent topic inferred from our model involve most visual attributes associated with the dolphin category such as ‘stripes’, ‘black’, ‘white’, ‘large’, ‘group’, etc. , suggesting that the Must-Link relationships between object categories and visual attributes are well preserved by the Dirichlet-tree distribution in our proposed model.

N02581957: dolphin
Must-link Attributes: white, blue, gray, hairless, toughskin, big, lean, flippers, tail, swims, fish, coastal, ocean, water,
Cannot-link Attributes: black , brown , orange , red , yellow , patches , spots , stripes , furry, desert , bush , plains , forest, mountain,…

Most relevant latent topic:

Top words	Probability
dolphins	0.14474
River	0.03321
whale	0.01860
species	0.01329
attacks	0.01296
meat	0.01111
tuna	0.01062
sounds	0.00931
mammals	0.00798
Delphinidae	0.00798

(a) LDA Model

Top words	Probability
blue	**0.15112**
dolphin	0.10289
tons	0.04181
water	**0.03645**
white	**0.02144**
large	**0.01930**
metric	0.01930
brevicauda	0.01501
fish	**0.01287**
harpoon	0.00858

(b) Our Proposed Model

Figure 3.51: Comparison the most relevant latent topics of object category ‘dolphin’

It’s also worth mentioning that, in Fig. 3.50, one of the attributes (i.e. ‘spotted’) that cannot be linked to lion category is among the top-ranked words of the most relevant latent topic inferred by the conventional LDA model. As a comparison, none of the latent topics inferred by our proposed model violate Cannot-Link relations. The experiment results indicate thatthe Cannot-Link relations between object categories and visual attributes can be effectively encoded by the Dirhchlet Forest prior introduced in Section 3.5.2, which enables the topic model to purify the inferred latent topics, filter out the ‘noisy’ and ‘self-contradictory’ information from the textual descriptions and produce consistently well topical abstraction of object categories and associated visual attributes.

Conclusions

In this section, we deal with two research issues, i.e.the automation of visual attribute identification and semantic relation learning between visual attributes and object categories.

The contribution is two-fold, firstly, we provide uniform framework to reliably extract both categorical attributes and depictive attributes. Secondly, we incorporate the obtained semantic associations between visual attributes and object categories into a text-based topic model and extract descriptive latent topics from natural language knowledge base. Specifically, we showthat in mining large scale text corpora of natural language descriptions, the relation between semantic visual attributes and object categories can be encoded as Must-Links and Cannot-Links, which can be represented by Dirichlet-Forest prior. To reduce the amount of manual supervision and labeling in large-scale image categorization, a semi-supervised training framework using soft-margin semi-supervised SVM classifier is introduced. Last but not least, automatically extracted visual attributes are used in a posterior object-attribute prediction model to further improve the performance of object categorization. Experimental results show that the proposed model achieves better ability in describing object-related attributes and makes the inferred latent topics more descriptive.

4. PROBABILISTIC TOPIC MODEL FOR BIOINFORMATICS STUDIES

In this chapter, I would like to introduce the background of Generative Latent Space Models and review the related works on topic modeling.

4.1 OVERVIEW AND OBJECTIVE

In the system biology community, there has been a long time focus on studying gene-expression data in isolated organisms and cultures. However, relatively less effort has been made to study the genome-wide gene-expression data from uncultured environment samples (like the ocean, soil and human body) and understand the underlying biological processes. Recently, the development of new sequencing techniques and meta-genomics has dramatically changed the way of genomics data acquiring and analyzing. Next generation sequencing methods (such as Roche/454 Sequencing and Illumina Sequencing) are able to extract very large amount (100 $\sim$ 1000 MB) of DNA fragment sequences from an environmental sample (like the ocean, soil and human body) in only a single run (the acquired data is also known a 'meta-genomic data'). With the fast advancing sequencing technology, large amount of sequenced genomes and meta-genomes from uncultured microbial samples becomes available. Based on the meta-genome sequences, bioinformatics researchers have done a lot of work to study the underlying biology process such as signal transduction, translation, and molecular functions like the biochemical activity of gene product. However, our knowledge about the biological functions encoded in the meta-genome sequence is still limited. Current functional annotation (genome-level annotation of biological functions) is still far from satisfied. The lack of high quality functional annotation of the major functionality encoded in the gene-expression data of given genome/meta-genome posed a great challenge in the task of interpreting the biological process of meta-genome.

The major objectives of analyzing and interpreting the large amount of meta-genomic data involve answering two questions. The first question is, ‘Given a large number of genome fragments from an environmental sample, what genomes are there?’ Answering this question requires mapping the meta-genomic reads to taxonomic units (usually a homology-based sequence alignment, this task is also known as taxonomic classification or taxonomic analysis). The second question is, ‘What are the major functions of these genomes?’ The answers to this question involve annotating the major functional units (such as signal transduction, metabolic capacity and gene regulatory) on the genome-level (a.k.a. functional analysis).

Toward these two questions, we present a set probabilistic topic models to identify functional groups from microbial samples. Probabilistic topic models have been developed for applications in various domains such as text mining [95], information retrieval [15] and computer vision [2], [93]. In bioinformatics domain, generative topic model has been previously used to learn protein-protein relations from MEDLINE abstracts of biomedical literatures [9], [110]; it has also been applied to identify gene relations from microarray profiles [36]; the generative topic model is also used to describe the process of constructing mRNA module collections [40]. In [40], the author uses a topic-model-based Gene Program algorithm to allocate mRNA from each tissue to different gene expression programs, in which each tissue is considered as a sample from a population of related tissues. In the model, gene sets have different chances of being co-expressed in different subset of samples, which also encodes the assumption that similar sample groups are more likely to share similar gene sets. The model provides the flexibility in allocating the expression data and discovering co-expressed gene sets. In our approach, the probabilistic topic models are derived from either taxonomic or functional-element abundance data(such as high abundance of specific functional group, high expression level of specific taxon, gene cluster, or specific metabolic pathway) acquired from either composition-based genome classification or homology-based alignment.

4.2 A BRIEF REVIEW OF COMPOSITION-BASED AND HOMOLOGY-BASED FUNCTIONAL ANALYSIS

4.2.1 Composition-based Approaches

Recently, the composition-based approaches [84], [85], [4], which break down the DNA fragments into N-mer sub-reads, have achieved good performance in the task of genome classification. After extracting N-mers from DNA fragments, the composition-based taxonomic analysis is achieved by picking up N-mer features for each genome fragment and scoring against each taxon (or calculating the probability that a DNA fragment comes from a particular taxon [84]).

The taxonomy analysis of ‘N-mer’ sub-reads is achieved by picking up N-mer features for each genome fragment and scoring against each taxon (or calculating the probability that a DNA fragment comes from a particular taxon). Recent composition-based approaches for taxonomic classification usually rely on supervised learning algorithms such as the Naive Bayesian classifier (NBC) [84], and Phymm [4] to classify short fragments. State-of-the-art supervised learning methods of taxonomic classification are reviewed in [55].The Phymm algorithm uses interpolated Markov model (IMM) to solve the phylogenetic classification problem [4]. The interpolated Markov model is an extension of traditional fixed-order Markov models. In this model, the estimation of the probability for the next state depends on probabilities of all the different orders (that is, probabilities from 0, 1, 2, ..., n previous nucleotides, with different weights); as a comparison, a regular n-th order Markov chain only depends on the n previous nucleotides.During the training stage, the IMM model compute probability of different nucleotide patterns from each species When the model is learnt from the training dataset, it is able to compute the probability of a given nucleotide sequence generated by a specific IMM model distribution of corresponding taxon.The Naive Bayesian Classifier (NBC) algorithm [84], on the other hand, calculates the N-mer frequencies for each taxon and uses these frequencies as features to train the Naive Bayesian Classifier (NBC). Based on the Bayesian's Theorem, the estimation of this classifier can be achieved by calculating and optimizing the scoring function, which is a product of conditional probability of each Nmer given a genomic class. The prior probability of Nmers and prior probability of genome classes are omitted in the final scoring function. The obtained Classifier is then used to determine the assignment of genome classes to DNA fragments using the posterior probability that a DNA fragment comes from a specific class. After assigning to the NCBI taxonomy, the encoded functionality of meta-genomic reads can be readily available by querying the Gene Ontology (GO) database (Fig.4.1).

When considering each genome fragment as a 'document', the 'N-mers' can to some extend be considered as a kind of 'code words' that compose a genome fragment (we

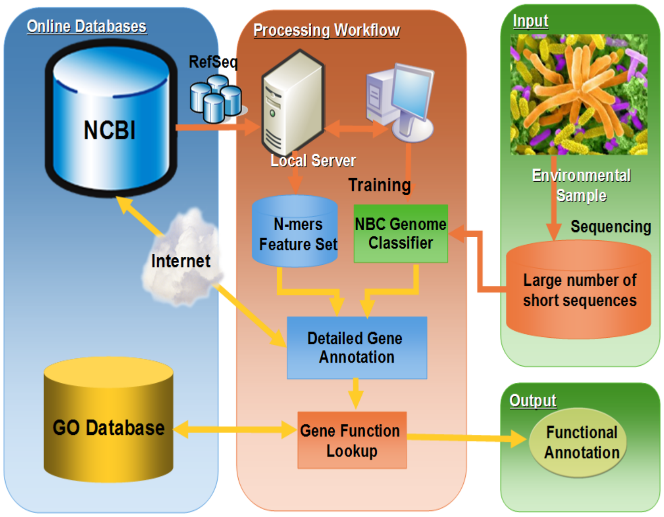


Figure 4.1: A framework of composition-based genomic data analyzing

may consider the A,T,C,G nucleotides as letters, so N-mers bear an analogy with N-letter words). Take one step further; we may assume that a genome fragment bears an analogy with a text document. The analogy between Nmers and text words (both of them can be represented as vector of term frequencies) has inspired researchers to introduce some text mining techniques such as TD-IDF term weighting to Nmers features [97]. The analogy between Nmers and text words also shows a potential of identifying the functional cores (core genes) via a mixture of N-mer sub-read distributions (like we are able to identify the major semantic topics of a text document via a mixture of text word distributions). To the best of our knowledge, although the composition-based approaches have been intensively used in solving taxonomic classification problems, however, little efforts have been made to exploit the latent Nmer patterns that build up the core genes and encode the major metabolic functionalities.

4.2.2 Homology-based Approaches

In analyzing meta-genomics data, the identification of both taxonomic unit and functional unit are difficult because the large amount of meta-genome fragments are usually very short ($<$ 100 bp) and may be from a variety of organisms. The challenge is that, when dealing with large genome-wide gene expression data, the samples may be from different individuals with different genetic and environmental background. What's more, the samples usually represent collections of diverse cell-types mixed together in different proportions. Therefore, in processing the meta-genomic reads, it's required that the raw reads be firstly assembled to longer contigs ($>$ 500 bp). After that, the protein-encoding sequences (CDs) are predicted from assembled contigs, results in a non-redundant catalogue of CDs (gene regions). The construction of non-redundant CDs catalogue gives rise to a 'minimal genome' and provides opportunity to identify bacterial functions that play important roles in microbial samples. Based on the non-redundant CDs catalogue, both taxonomic unit identification (identifying the existence of certain microbial species in the meta-genome) and functional unit identification (identifying the existence of certain gene product) can be readily achieved by matching amino acid sequence in CDs regions to standard reference sequences using homology-based alignments.

The homology-based approaches that align meta-genomics read to standard reference (of known species) in standard databases (such as NCBI NR database) via BLASTP or BLASTX algorithms have been intensively used to deal with both taxonomic analysis and functional analysis problems [21].

One example of homology-based classification approach is the Metagenome Analyzer (a.k.a. MEGAN) [49]. MEGAN is computer software that achieves taxonomical analysis over large databases. It compares DNA fragments against the database of reference sequence, and extracts taxonomical information from the high score BLAST hits. Based on the taxonomical information, the BLAST hits will be matched to different species and strains of the NCBI taxonomy (the algorithm collect all high score BLAST hits and assign taxon ID to each hit based on NCBI taxonomy, the NCBI Taxonomy contains over 460,000 taxa from different taxonomical ranks such as Kingdom, Phylum, Class, Order,...,). After that, the algorithm will look for the lowest common ancestor (LCA) taxon of all those BLAST hits and then assigned the input sequence fragment with that taxon. In practice, BLASTX algorithm will be used to compare all reads against the standard references database (like the NR (non-redundant) protein database from NCBI). Before loading BLASTX hits of a specific meta-genomic fragment to perform a LCA algorithm, the MEGAN algorithm uses a bit-score threshold to limit the number of BLAST hits. It also discards all the 'isolated assignment' by looking into the number of hit with regard to each taxon. Despite these efforts, however, the number of resulting hits may still be up to tens of thousands. The LCA algorithm used in MEGAN is able to visualize the hierarchical structure of the phylogenetic tree. However, it should be pointed out that the resulting taxon assignment may only has a limited resolution, as reads with too much BLAST hits may always be assigned to relative high taxon levels such as genus, family instead of species and strains [48]. Recently, Richter and Huson introduced a new framework to achieve detailed functional annotation for meta-genomic reads [21]. The framework begins with a homology based BLASTX algorithm to match the meta-genomic fragments to the reference sequences in NCBI database. The BLASTX hits will associate fragments with related protein ID and gene names. After that, with the help of the Gene Ontology (GO) database (http://www.geneontology.org), which uses controlled vocabulary to represent biological processes, cellular components and molecular functions of genes and gene products [29], the framework is able to refer associated gene names to corresponding GO terms, thus provides an overview of gene function and products for meta-genomic reads.

By aligning local amino acid sequence to the reference sequences (of known species) in standard databases (such as NCBI NR database, eggNOG database and KEGG database), researchers are able to acquire a lot of useful information with respect to the functionality encoded in predicted CDs regions, including taxonomic levels, indicator of gene orthologous groups (OGs) and KEGG pathway mappings. The alignment of amino acid sequence also provides an insight about the functionality groups existing in the genomes. Although genes vary from strain to strain, similar genes can have similar functions among different species known as clusters of Orthologous (COGs). The relative abundance of certain COG categories in a microbial sample may indicate whether the sample is rich in particular functions. In practice, COGs can be determined based on their sequence similarity and can be classified into different function categories [48] including signal transduction, metabolic pathways and gene regulatory network. With this consideration, the functional units in microbial samples can be identified by the gene clusters such as COGs shared among species/strains. Understanding the functionality of gene clusters is of practical and theoretical importance. For example, the functionality roles of organ/cell specific gene clusters may be different from gene clusters which are active across diverse cell types. Set of genes that are very specific to a particular cell type or organ may be useful as diagnostic bio-markers. In contrast, gene clusters that are active across diverse cell types can give us insight to uncover functional similarities among organs/cells. Since different microbial samples are taken from different micro-environment and expressing different sets of genes, we may assume that each microbial sample (with multiple cell types) has its own configuration of gene clusters, some clusters will be shared among many cell types while others will be more specific. It has been pointed out that the existence of commonly shared gene clusters across samples suggests functional similarity and biological relevance [36], [40]. Therefore, we aim to develop a method that enables analyzing the genome-level configuration of both taxonomic units and functional units derived from the non-redundant CDs catalogue. As we mentioned, by homology-based alignment, each CDs sequence can be represented as a triplet (i.e. taxonomic levels, indicator of gene orthologous groups and KEGG pathway mappings) each unit may be considered as a functional element at different levels (i.e. taxonomic level, gene level and pathway level). As a result, the functional meta-genome annotation can be achieved by first decomposing the meta-genomic samples into a mixture of functional elements (from three different levels); and then analyzing the genome-level configuration of functional elements to learn how those functional elements are grouped and jointly participate in the biological processes.

4.3 FUNCTIONAL AND TAXONOMIC ANALYSIS OF N-MER SUB-READ PROFILES BY PROBABILISTIC TOPIC MODELING

In 2005, Medini et al. challenged the concept of 'the species' and described the concept of the pan-genome, which is the 'entire genome' of an entire species, instead of each thinking about each strains genome individually [68]. Genetic elements are classified as two types in each strain genes shared by all the strains and genes that are 'dispensable' or only contained within a subset of the strains. For example, strains of E. Coli are hypothesized to only share 1,560 core genes (approximately 1/3 of the genes of any given strain) [34], which means many more are 'dispensable' or as Erhlich et al. put it 'distributed' [27]. We will refer to those essential to each strain as core gene, while those vary from strain to strain as distributed gene.

In this section, we aim to analyze the genome-level composition of DNA sequences. In order to characterize a set of common genomic features shared by strains within the same species and tell their functional roles, we firstly apply a composition-based approach to break down DNA sequences into sub-reads called the 'N-mer' and represent the N-mer sub-read sequences as a 'bag-of-words' model. Then, we employ the Latent Dirichlet Allocation (LDA) model to study latent topics. Specifically, the inferred latent topic means genome-level distributions of N-mer sub-read features corresponding to some functional groups. A collection of inferred latent topics thus represent the major functional groups of the whole genome. The identified latent topics are not necessarily specific to any specific microbe; instead, they may potentially indicate the co-occurrence of diverse gene function categories (these gene function categories may either belong to biology process such as signal transduction, translation, etc. or belong to molecular functions like the biochemical activity of gene product) from different organisms and may suggest functional relationships between genes. With the help of the BioJava toolkit [45], we access to the gene region information of reference sequences from the NCBI database. We also use our data mining framework to investigate two areas: 1) do strains within species share similar core and distributed topics? and 2) do genes with similar functional roles contain similar latent topics?

4.3.1 Construction of N-mer sub-read Profile

The dataset we dealt with involves standard reference sequences (FASTA format) of genomes downloaded from the NCBI database. Each genome is corresponding to a specific strain (say, Escherichia_coli_K12) and has multiple chromosomes. The chromosome (reference sequence) has an average length of about one million base pairs and is uniquely tagged by a Gen-Bank accession number (such as NC_009926).

For each chromosome under the same strain, by parsing the corresponding FASTA file, both GenBank accession number and the nucleotide sequence of the chromosome can be obtained. After that, a sequential indexing process is performed on all the chromosome sequences, in which each N-mer (N=9) is consider as a number from a quaternary (base 4) numeral system and assigned a numerical ID. Compared to regular N-mer indexing approach in [84], which only keeps N-mer frequencies while ignore their orders (Fig. 4.2), the sequential indexing approach generates a sequence of N-mer indices, while keeping their original order in each of the chromosomes. Specifically, a JAVA program is performed to automatically extract all the N-mers as well as their location in the chromosomes. The location of each N-mer is then kept in the index file, which is then used to recover the original order of N-mer sequence.

4.3.2 Gene Region Annotation for N-mer Sequence

With the help of the 'BioJava' package [45], detailed gene region information from the NCBI database such as locations of non-protein encoding regions (such as tRNA) and gene regions (CDS) as well as the corresponding gene names are obtained by querying the

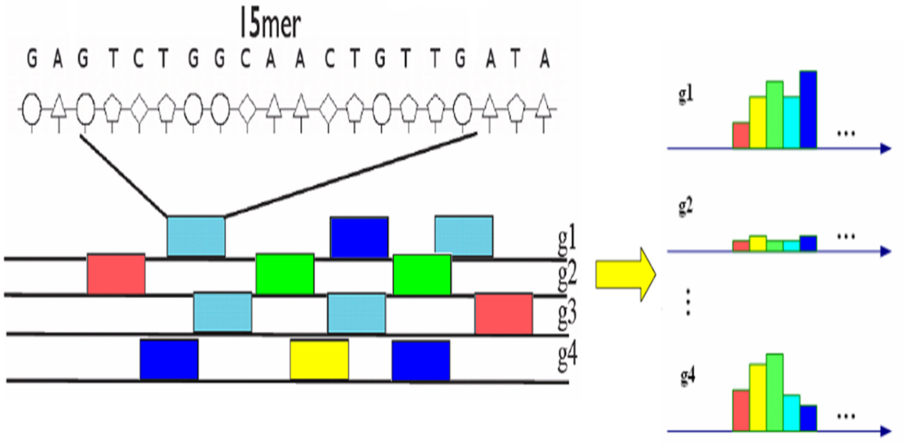
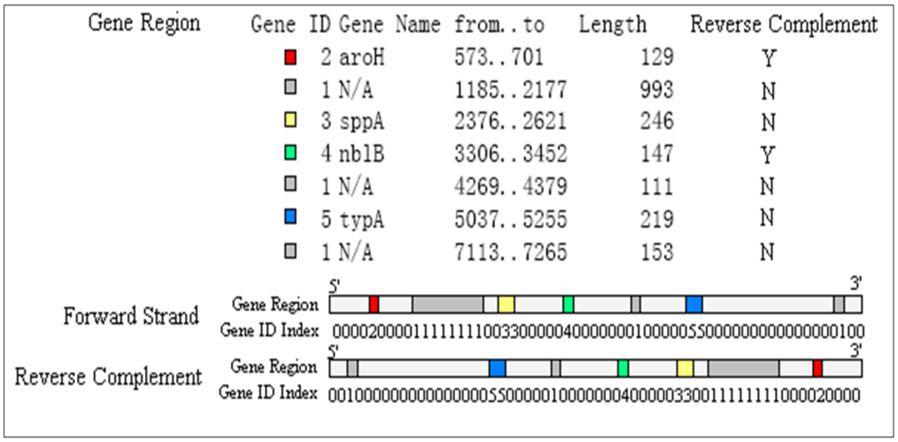


Figure 4.2: Illustration of the N-mer (N=15) presences in genomes, which bear analogy with text documents. The diagram on the left represents different genomes which may share or not share the same N-mers and the diagram on the right is their frequency occurrences in those genomes.

online NCBI database with the GenBank accession number of each reference sequence. Given an N-mer index file (obtained by method introduced in Section 2.1), by matching every N-mer location against the gene regions, it is able to generate a detailed gene region annotation that specify the relation between gene regions and each of the N-mers (Fig. 4.3). More specifically, for each N-mer from a given chromosome, the gene annotation is achieved by comparing the position of current N-mer to all the available gene regions in this chromosome. If an nmer doesn't belong to any protein-encoding region, it will be tagged with '0'. If an nmer is within a gene region with a gene name, it will be tagged with the corresponding gene name ID number; otherwise, it will be tagged with '1' for protein-encoding region that doesnt specify its name. The significance of the sequential indexing is that, by keeping the order of nmers, we will be able to annotate them with gene region information obtained from NCBI database, which may also be used to calculate the



Gene Region	Gene	ID Gene Name	from..to	Length	Reverse Complement
	🟥	2 aroH	573..701	129	Y
	▨	1 N/A	1185..2177	993	N
	🟨	3 sppA	2376..2621	246	N
	🟩	4 nblB	3306..3452	147	Y
	▨	1 N/A	4269..4379	111	N
	🟦	5 typA	5037..5255	219	N
	▨	1 N/A	7113..7265	153	N

Figure 4.3: Illustration of the N-mer (N=15) presences in genomes, which bear analogy with text documents. The diagram on the left represents different genomes which may share or not share the same N-mers and the diagram on the right is their frequency occurrences in those genomes.

relevance between genome-level statistic patterns (a.k.a. latent topics) and gene regions. The relevance between latent topics and region regions will provide us an insight of how N-mer patterns related to their functional roles. It should be noted that the actual protein-encoding region (translation site) may be from the reverse complement instead of from the forward strand. Thus, as illustrated in Fig.4.2, some gene regions may be specified as ‘from reverse complement’. As a result, we need to carry out sequential indexing from both directions, and tag gene region on both forward and reverse complement sequence (Fig. 4.3).

4.3.3 Generative Topic Model for N-mer Sequence

In our study, we use the Latent Dirichlet Allocation (LDA) model [15], an effective generative topic model firstly introduced in text mining domain, to study the functional groups. Latent Dirichlet Allocation (LDA) model [15], is an effective probabilistic topic model firstly introduced in text mining domain to infer latent semantic topics from text documents. The LDA model allows us to study underlying concurrence patterns of the data and extract useful knowledge such as latent semantic topics from the data. Whats more, the learning process of LDA model is totally unsupervised; therefore, it is very suitable for research areas which lack of labeled data. Due to its solid theoretic foundation and promising performance, the LDA model has been popular with the data mining community in recent years. It is widely agreed that the LDA model promises good results across most text data categories including domain specific text data (such as MEDLINE) [110] and general text data (such as the New York Times Dataset) [95] and may also bring good results in other text-like data such as visual code words [2]. Those data categories are also known as ‘bag-of-words’ models since they represent each document by a distribution over fixed vocabulary (in which the order of the vocabulary doesnt matter).

In text mining, the underlying assumption of LDA model is that, a document may deal with multiple topics; and each of these topics can be represented by a unique distribution of words. A latent topic is a high-level concept which explains the co-occurrence patterns of words that appear in one document, it provides an effective way to analyze the composition of documents. Depending on different application context, a latent topic may have different semantic meanings. Based on the definition of latent topics, the objective of LDA model is to assign these latent topics to words in a document (each word wi may only be assigned one topic), so that a document may in turn be represented as a mixture of latent topics. In practice, the latent topic assignment is achieved by manipulating some unseen latent random variables to determine the conditional probability of words given a latent topic $p(word|topic)$ and probability of latent topics given a document $p(topic|document)$. When a generative topic model (such as LDA model) is used to study microbial communities, each sample can be considered as a 'document', which has a mixture of functional groups, while each functional group (also known as a 'latent topic') is a weight mixture of taxa (the taxon label of each genomic read can be considered as a word). Estimating the generative topic model will uncover the distribution over latent functions (latent topic) in each sample.

Due to the analogy between N-mers and text words, the genome sequences are comparable to documents. Therefore, the LDA model for N-mers can be defined as follows. Assuming that there are a total of D genomes (documents) in the data collection, which in total contain N_w N-mers (tokens) from W different numerical N-mer indices (words); and there are a total of T latent topics. For the d-th genome (document), the LDA model samples a latent variable $\theta^d \sim Dir(\alpha)$, in which θ^d is a T-dimensional vector of topic priors in genome d. Then, for the j-th topic, the model samples latent variable$\varphi_j \sim Dir(\beta)$, which serves as prior probabilities for N-mer distributions of different topics. After sampling latent variables θ^d and φ_j , the probability that topic j appears in genome d is defined as $p(z^j|d) \sim Multi(\theta^d)$, in which z^j means topic $z = j$. The probability that a given N-mer w_i is generated by z^j is defined as $p(w_i|z^j) \sim Multi(\varphi^j)$.

The generative process of this model is defined as follows:

1. For the d-th (d=1D) genomes(documents), sample $\theta_d \sim Dir(\alpha)$

2. For the t-th (t=1T) topic, sample $\varphi_t \sim Dir(\beta)$

3. For each of the N_d N-mers (words) w_i in genome d:

a) Sample a topic $z_i \sim Multi(\theta_d)$ and sample $w_i|z_i \sim Multi(\varphi_{z_i})$.

The estimation of LDA model given the observed N-mer sequence data can be estimated via the Gibbs Sampling Monte Carlo processs [95]. The estimation process requires separately sampling topic for each word in each document according to the posterior probability as follows.

$$p(z_{wi} = j|w_i, \mathbf{w}_{-i}, \mathbf{z}_{-w_i}) \propto \frac{\beta + n^{w_i}_{-i,j}}{W\beta + n^{*}_{-i,j}} \cdot \frac{\alpha + n^{d}_{-i,j}}{T\alpha + n^{d}_{-i,*}} \tag{4.1}$$

In which $n^{wi}_{-i,j}$ is the total number of taxon labels assigned to topic j except for w_i , and $n^{d}_{-i,j}$ is the total number of taxon labels in sample d (except for w_i) that have been assigned to topic j. In our model, we assume symmetric priors and set $\alpha = 0.1, \beta = 0.01$. Such a parameter setting is for the consideration of making topic modeling results more diverse. For example, by setting Dirichlet distribution with parameter $\alpha = 0.1$, the topic mixture for each genome will converge on several unique topics instead of having equal probability for every topic. We follow the model selection method in [110] to determine the number of latent topics. In general, a larger topic number may provide higher resolution to the uncovered functional core (either microbial core or gene core) of genome. However, a large topic number may also cause an over-fitting problem to the model. The selection among the models with different topic number is carried out based on the approximated evidence (log likelihood) of samples. After extensive experimental study, the number of latent topics is set to be 50 for taxon-abundance data. Usually, it takes less than 100 iterations for the Gibbs sampling process to converge, thus in the model estimation procedure, we terminate the Gibbs sampling process after 100 iterations.

4.3.4 Relevance between Latent Topics and Gene Regions

The significance of learnt latent topics can be greatly enriched by studying the relevance between latent topics and gene regions. As mentioned in Section 4.2, a sequential indexing process is performed on the N-mer features, which keeps the original order of N-mers in genome sequences, thus enables the relevance study between latent topics and gene regions after the topic model assigning topics labels to N-mers. During the sequential indexing stage, we tagged each N-mers position with an identification number specifying the gene region type. The identification number (which covers all the gene names appear in all the genomes in our study) not only indicates whether the current N-mers is within a gene region or not, but also tell what the gene name is (if gene region has a name).

After estimating the topic model and assigning latent topic to each N-mer, the relevance between latent topics and gene regions can be obtained by calculating the mutual information (MI) between genes and obtained latent topics based on the annotated gene regions and final latent topic assignments. The MI between a specific gene (R_g) and a latent topic (Z_t) is as follows:

$$p(x_{w_{ji}} = 0, z_{w_{ji}} = 0 | w_{ji}, \mathbf{w}_{-ji}, \mathbf{z}_{-ji}, \mathbf{x}_{-ji},) \propto \frac{N^{0}_{d,-i} + \gamma}{N_{d,-i} + 2\gamma} \cdot \frac{\beta_2 + n^{w_i}_{-i,0}}{w\beta_2 + n^{*}_{-i,0}} \tag{4.2}$$

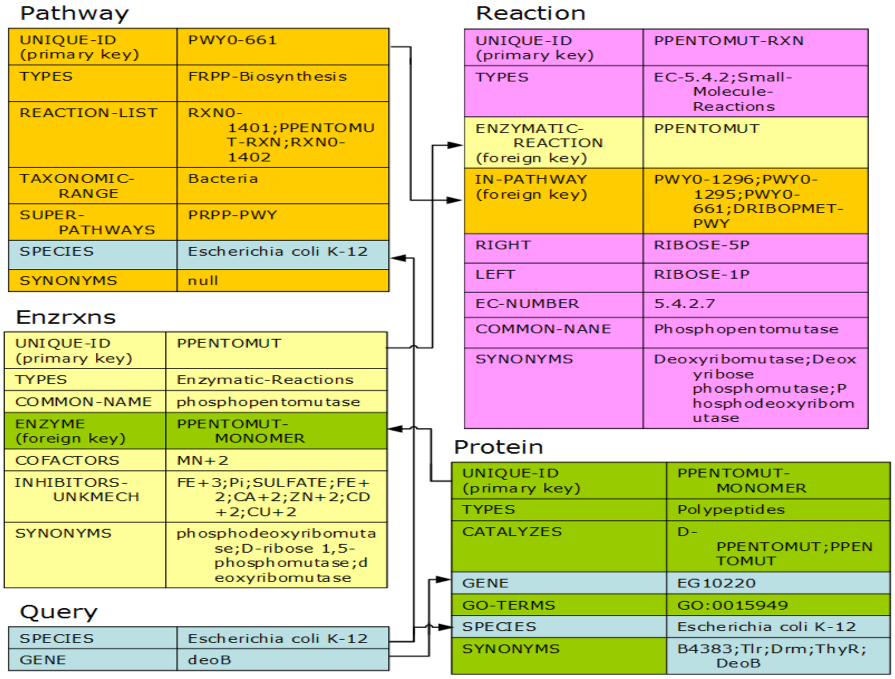
In which the gene region indicator R_g and latent topic indicator Z_t are both nmer-wide binary variables, which indicate whether an nmer is within a gene region and which latent topic has been assigned to this nmer. Given the training data, the joint probability $p(R_g, Z_t)$ and two marginal probabilities $p(R_g)$ and $p(Z_t)$ can be simply estimated by counting the number of evidences over all the training data.

4.3.5 Functional Annotation for Latent Topics

After identifying a set of genome-level latent N-mer patterns (latent topics) that represent different genome sequence components and calculating the relevance between inferred latent topics and gene regions, the functional role of each inferred latent topic can be explain by the function of its most relevant genes.

Traditional functional annotation approaches use the Gene Ontology (GO) terms to explain the functional role of genomic sequences [21]. Typically, the GO terms are obtained by query the GeneGo Database with corresponding gene symbols or protein IDs. One problem with the GO terms is that they are highly incomplete (only covers a limited number of organisms and gene symbols) and usually not very detailed. Therefore, the GO terms are unable to provide us a comprehensive view of gene functions in different species. What's more, in order to fully understand the gene functions, its of great importance to look into the enzyme, pathway information, and metabolic capabilities related to the gene. With this consideration, we utilize the BioCyc database [30], an openly available, highly accurate, valuable database of metabolite pathway and enzyme data that have been experimentally demonstrated in the scientific literatures, to study the functional role of identified genome components (latent topics). Compared to the GO terms, the gene function searched from the BioCyc database is much more comprehensive, which involves information about enzyme, pathway, and metabolic capabilities

Specifically, we have acquired a data accessing license and download flat files of individual databases (covers most of the species we study) from BioCyc DB. After that, we convert the flat data file into data table that enable effective data search (Fig. 4.4). The gene function search begins with querying the BioCyc database with gene symbol and species name,



Pathway

UNIQUE-ID (primary key)	PWY0-661
TYPES	FRPP-Biosynthesis
REACTION-LIST	RXN0-1401;PPENTOMUT-RXN;RXN0-1402
TAXONOMIC-RANGE	Bacteria
SUPER-PATHWAYS	PRPP-PWY
SPECIES	Escherichia coli K-12
SYNONYMS	null

Enzrxns

UNIQUE-ID (primary key)	PPENTOMUT
TYPES	Enzymatic-Reactions
COMMON-NAME	phosphopentomutase
ENZYME (foreign key)	PPENTOMUT-MONOMER
COFACTORS	MN+2
INHIBITORS-UNKMECH	FE+3;Pi;SULFATE;FE+2;CA+2;ZN+2;CD+2;CU+2
SYNONYMS	phosphodeoxyribomutase;D-ribose 1,5-phosphomutase;deoxyribomutase

Query

SPECIES	Escherichia coli K-12
GENE	deoB

Reaction

UNIQUE-ID (primary key)	PPENTOMUT-RXN
TYPES	EC-5.4.2;Small-Molecule-Reactions
ENZYMATIC-REACTION (foreign key)	PPENTOMUT
IN-PATHWAY (foreign key)	PWY0-1296;PWY0-1295;PWY0-661;DRIBOPMET-PWY
RIGHT	RIBOSE-5P
LEFT	RIBOSE-1P
EC-NUMBER	5.4.2.7
COMMON-NANE	Phosphopentomutase
SYNONYMS	Deoxyribomutase;Deoxyribose phosphomutase;Phosphodeoxyribomutase

Protein

UNIQUE-ID (primary key)	PPENTOMUT-MONOMER
TYPES	Polypeptides
CATALYZES	D-PPENTOMUT;PPENTOMUT
GENE	EG10220
GO-TERMS	GO:0015949
SPECIES	Escherichia coli K-12
SYNONYMS	B4383;Tlr;Drm;ThyR;DeoB

Figure 4.4: Gene function search in BioCyc flat tables

after that, based on the searching schema in Fig. 4.4, enzyme and pathway information as well as the metabolic capabilities related to the gene will in turn be retrieved from the database, which explains the function role of genomic patterns that most relevant to the gene.

4.4 ESTIMATING FUNCTIONAL GROUPS IN HUMAN GUT MICROBIOME WITH PROBABILISTIC TOPIC MODELS

In this section, based on the functional elements derived from non-redundant CDs catalogue, we present the generative framework to infer the configuration of functional groups in meta-genome samples and introduce the extended Enterotypr-HDP model to infer functional basis of detected enterotypes.

In using the probabilistic topic model for inferring functional groups of biological process, we define the model as follows. The genome set serves as the document corpus, with individual samples representing the documents. The functional elements (including NCBI taxonomic level indicators, indicator of gene orthologous groups and KEGG pathway indicators) serve as words, which jointly define a fixed words vocabulary of the corpus (take the genes orthologous group (OGs) indicators for example, the COG and NOG terms from the eggNOG database [75]can be used as vocabulary of the model). Consequently, each document can be represented as a bag of words, in which the order of words is not considered. Each inferred latent topic (i.e. functional group such as bacteria groups or group of gene clusters) defines a multinomial distribution over given vocabulary. In other words, each functional group specifies a multinomial distribution over functional elements. The discrete expression levels of functional elements are treated analogously as the word frequency in text documents. The configuration of functional groups in each sample as well as the distribution of functional elements in each functional group are considered generated conditional independently by the topic model. With inferred latent topics, meta-genome samples can be represented as weighted combinations of functional groups. Different functional groups (latent topics) may co-exist in the same sample and may be shared across a set of samples. The samples differ in terms of which functional groups are presenting in and how they weighted.

4.4.1 Modeling Commonly Shared Functional Elementsvia LDA Model with Background Distribution(LDA-B)

Given non-redundant CDs catalog, and derived functional elements, we are interested in identifying the frequent co-occurrence patterns of functional elements. Commonly shared functional elements (such as taxa groups, gene clusters and pathways) across samples may suggest functional similarity and biological relevance among samples. If strong genome-wide co-existing patterns of functional elements do exist, then it may suggest the existence of 'core' genome.

With this consideration, we extend the LDA model by adding a background distribution of commonly shared functional elements. We present graphical representation of the proposed model in Fig. 4.5. Following the convention in depicting graphical representation of topic models, we use round nodes to represent random variables, in which the white nodes stand for latent random variables, while the gray nodes denote observations during the model training. The rounded boxes are used to represent fixed hyper-parameters of the model, while the edges illustrate the conditional dependency underlying the generative process.The generative process of the proposed model is as follows. As shown in Fig. 4.5, a switch variable x is introduced in the model, which fits a Binomial distribution λ (with a Beta prior of γ) and only takes binary values 0, 1. Before sampling the latent topics in sample j, the switch variable x needs to be sampled for each functional element w_{ji}. For a given w_{ji}, if its switch variable x equals to 0, then it should be generated by the background topic z_0, otherwise, if its switch variable x takes the value 1, it should be generated by one of the T regular latent topics. For functional element w_{ji} in sample j, its assigned topic (either background topic z_0 or one of the T regular latent topics)is sampled according to the

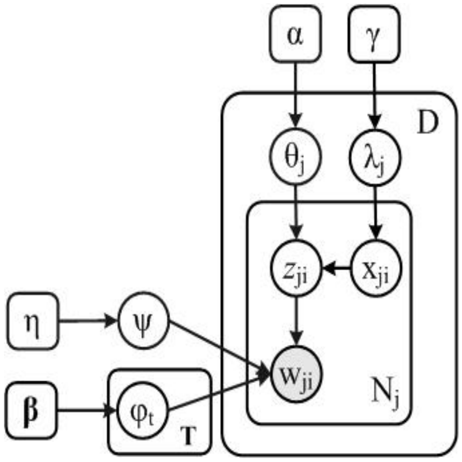


Figure 4.5: Hierarchical structure of proposed LDA-B model

posterior probability:

$$p(x_{w_{ji}}=0,z_{w_{ji}}=0|w_{ji},\mathbf{w}_{-ji},\mathbf{z}_{-ji},\mathbf{x}_{-ji},)\propto\frac{N^0_{d,-i}+\gamma}{N_{d,-i}+2\gamma}\cdot\frac{\beta_2+n^{w_i}_{-i,0}}{w\beta_2+n^*_{-i,0}}\tag{4.3}$$

$$p(x_{w_{ji}}=1,z_{w_{ji}}=k|w_{ji},\mathbf{w}_{-ji},\mathbf{z}_{-ji},\mathbf{x}_{-ji},)\propto\frac{N^1_{d,-i}+\gamma}{N_{d,-i}+2\gamma}\cdot\frac{\beta_2+n^{w_i}_{-i,0}}{w\beta_2+n^*_{-i,0}}\cdot\frac{\alpha+n^d_{-i,k}}{T\alpha+n^*_{-i,*}}\tag{4.4}$$

For formally, the generative process of this model is defined as follows:

1. For the j-th document(meta-genome sample), sample $\theta_j \sim Dir(\alpha)$ and $\lambda_j \sim Beta(\gamma, 1-\gamma)$

2. For the t-th (t=1T) latent topic, sample $\varphi_t \sim Dir(\beta)$, for the background topic, sample $psi \sim Dir(\eta)$

3. For each of the N_j functional element w_{ji} in document(sample) j:

4. For each functional element w_{ji}, sample a switch $x_{ji} \sim Bernoulli(\lambda_j)$

a. If $x_{ij} = 0$, sample $w_i|z_0 \sim Multi(\psi)$

b. If $x_{ij} = 1$, sample a topic $z_{ji} \sim Multi(\theta_j)$ and sample $w_{ji}|z_{ji} \sim Multi(\varphi_{z_i})$

In our model, we assume symmetric priors and set $\alpha = 0.1, \beta = 0.01, \gamma = 0.5$. Such a parameter setting is for the consideration of making topic modeling results more diverse. For example, by setting Dirichlet distribution with parameter$\alpha = 0.1$, the topic mixture for each genome will converge on several unique topics instead of having equal probability for every topic. We follow the model selection method in [110]to determine the optimal latent topic number. In general, a larger topic number may provide higher resolution to the uncovered functional core (either microbial core or gene core) of genome. However, a large topic number may also cause an over-fitting problem to the model. The selection among the models with different topic number is carried out based on the approximated evidence (log likelihood) of samples. Usually, it takes less than 100 iterations for the Gibbs sampling process to converge.

After estimating the topic model and assigning latent topic to each functional element, the relevance between latent topics and functional element indicators (i.e. NCBI taxonomic level indicators, indicator of gene orthologous groups and KEGG pathway indicators) can be obtained by calculating the mutual information (MI) between functional element indicators and obtained latent topics based on the final latent topic assignments to functional elements. The MI between specific functional element indicators and a latent topic is shown in eq. 4, in which Rg and Zt are binary indicator variables corresponding to functional element and latent topic, respectively. The variable pair (Rg, Zt) indicates whether a latent topic has been assigned to a specific functional element.

$$MI(R_g, Z_t) = p(R_g, Z_t) log \frac{p(R_g, Z_t)}{p(R_g)p(Z_t)} \tag{4.5}$$

Given the training data, the joint probability $p(R_g, Z_t)$ and two marginal probabilities $p(R_g)$ and $p(Z_t)$ can be simply estimated by counting the number of evidences over all the training data.

One limitation of LDA-based topic model is that it requires specifying the exact number of mixture components, which remains unchanged during the model estimation. In practice, in order to get an optimal number, the researchers have to try different mixture components numbers and make a choice by comparing the log-likelihood, perplexity and other criteria that indicate how good the model fits the data. The Hieratical Dirichlet Process (HDP) model [108], is a nonparametric extension of the Latent Dirichlet Allocation (LDA)-based topic models, it enables modeling documents with countable infinite mixture components, thus provides the flexibility of modeling data with different semantic component numbers.

4.4.2 Extended Hierarchical Dirichlet Process for Detected Enterotypes

In this section, we introduce the extended background HDP model to infer the functional basis for detected phylogenetic clusters (a.k.a. Enterotypes).

In recent studies [32] [28], there has been some general consensus about the phylogenetic composition in human gut microbiome. However, the composition of gene functions in human gut microbiome and their variations across human population is still not clear. It's unknown whether inter-individual variation may lead to dramatically different gene function composition or whether individual human gut microbiome congregates on several categories with shared functional properties. It has been demonstrated in that researcher may identified distinct clusters in human gut microbiome by analyzing the phylogenetic composition. Specifically, a large fraction of the meta-genome can be matched to the reference genome set on the genus and phylum level. In [28], multidimensional cluster analysis and principle component analysis (PCA) are performed on phylogenetic abundance profiles to further cluster 33 samples into 3 distinct clusters (a.k.a. 'Enterotypes'), which are identified by the levels of one of three genera: Bacteroides, Prevotella and Ruminococcus. It's hoped that the identified Enterotypesmay explain either the host properties (such as IBD) or the complex mixture of functional properties. However, when clustering the samples using purely functional metric (such as the abundance of the orthologous groups derived from predicted genes) the grouping of samples doesn't very much agree with the Enterotypes obtained by phylogenetic clustering, indicating that the abundance of function may not be coinciding with the abundance of genera. Typically, the most abundant molecular functions can be traced back to the most dominant species or genera. However, it should be noted that abundant species or genera cannot reveal the entire functional complexity of the gut microbiota, some identified orthologous groups may also be primary contributed by low-abundance genera. In our study, in order to determine the functional basis of the identified Enterotypes, we introduce the extended background HDP model of inferring sample-level composition of orthologous groups with respect to different Enterotypes.

To indicate the Enterotypes label of each sample, a switch variable x is introduced. The generative process of the Enterotype HDP model is represented in Fig. 4.6. For each orthologous group (OG) indicator w_{ji} in sample j, the value of x_{ji} (which takes values 0-1) is sampled from a binomial distribution λ_j (with a Beta prior ζ). When the value of x_{ji} equals 0, the topical indicator of OG indicator wji is draw uniformly from the functional basis ψ_e learned from the corresponding Enterotypes (the blue arrows in Fig. 4.6 illustrate this procedure). When x_{ji} equals 1, a mixture component of functional properties will be sampled

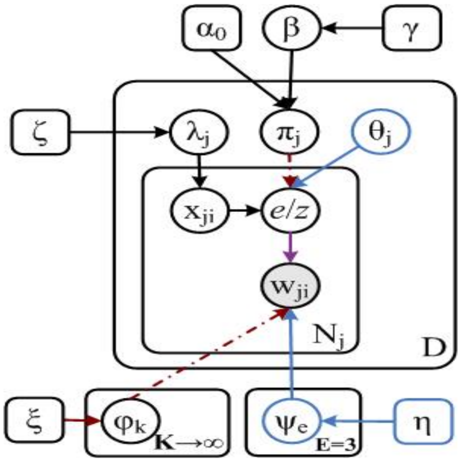


Figure 4.6: Hierarchical structure of proposed Enterotype HDP model

according to the sample-level weights of functional mixture components $\pi_j \sim DP(\alpha_0, \beta)$ for sample j, and OG indicator w_{ji} will be drawn from the distributionφ_k of functional mixture component k(the red dashed arrows in Fig. 4.6 illustrate this procedure). Detailed explanations of notations in the model are summarized in Table 4.1.

The generative process of this model begins with drawing a global probability measure $G_0 \sim DP(\gamma, H)$ and for each sample j, draw a child Dirichlet process $G_j \sim DP(\alpha_0, G_0)$. Following the stick-breaking construction, it is equivalent to firstly drawing a global weight $\beta \sim GEM(\gamma)$ for functional component indicators k, then for each sample j, draw the document-level weights of functional component indicators $\pi_j \sim DP(\alpha_0, \beta)$. The data observations in sample j are generated by repeatedly drawing functional component indicator $z_{ji} = k$ from π_j and then draw each OG indicator w_{ji} from the conditional probability φ_k of the sampled functional component $z_{ji} = k$. For formally, the generative process of this model is defined as follows:

Table 4.1: Notations in Proposed Topic Model

Symbol	Descriptions
D,W	Number of samples;OGs indicators
z	Indicators for functional mixture components
K	The number of functional mixture components at a certain time point
N_j, n_{jk}	Number of OG indicators in document j; number of OG indicators assigned to functional component k in sample j
$\breve{n}_{j,-q}$	The number of OG indicators in sample j generated from Enterotype functional basis(x_{ji}=0), except current instance
$n_{j,-q}$	The number of OG indicators in sample j generated from functional components (x_{ji}=1), except current instance
$C^{WE}_{we,-q}$	The number of times that OG indicator $w = q$ is generated from Enterotype e, except current instance
$C^{WZ}_{wk,-q}$	The number of times that OG indicator $w = q$ is generated from functional component k, except current instance
α_0, γ	Concentration parameters of Dirichlet process
φ_k	The OG indicator distribution of functional component k
π_j	The sample-level weights of functional components for sample j
ξ, η, σ	Hyper-parameters of Dirichlet , Beta distributions
ψ_p	The OG indicator distribution of Enterotype e
x_{ii}, λ_j	Switch variable that decides the source of each OG indicator and the sample-level distribution of different x values
β	The global weight of functional components

1. Draw a global weight $\beta \sim GEM(\gamma)$;

2.For each functional component k, draw conditional OG indicator distribution $\varphi_k \sim Dirichlet(\xi)$;

3.For each Enterotype e, sample conditional OG indicator distribution of its functional basis:$\psi_e \sim Dirichlet(\eta)(e = 1,2,3)$;

4.For the jthsample, draw $\pi_j \sim DP(\alpha_0, \beta)$;

5.For the i-th of the N_j OG indicators in the j-th sample

a. sample functional component indicator $z_{ji} = k \sim Discrete(\pi_j)$

b. sample its OG indicator $w_{ji} \sim Multinomial(\varphi_k)$;

6. For each sample j, sample $\lambda_j \sim Beta(\zeta)$;

7. For each OG indicator w_{ji} in sample j;

a.sample a switch variable $x_{ii} \sim Binomial(\lambda_j)$;

b.if (x = 0) Generate OG indicator w_{ji} from the functional basis of corresponding Enterotypee: $w_{ji} \sim Multinomial(\psi_e)(e = 1,2,3)$

c. if (x = 1) Sample functional indicator $z_{ji} \sim Discrete(\pi_j)$ .Then generate OG indicator wji from functional indicator $z_{ji} = k$ according to the conditional probability $w_{ji} \sim Multinomial(\varphi_k)$.

In the following, we describe the Gibbs sampling scheme for the proposed Enterotype HDP model. The sampling scheme consists of two steps. The first step is sampling for semantic component indicators z as well as the corresponding HDP hyper-parameters β. In order to sample a HDP-like model, one may either follow the Chinese restaurant franchise (CRF) or use direct assignment [108]. In our work, the direct assignment is used(Table 4.2). The second step is sampling for switch variable x, and conditional distribution of OG indicators $varphi_k$ and ψ_e . We derive the sampling equation of switch variable x_{ji} for each OG indicator $w_{ji} = q$ in sample j as follows:

$$p(x_{ji} = 1, z_{ji} = k | w_{ji} = q, \mathbf{z}_{-q}, \mathbf{w}_{-q}, \xi, \zeta) \propto \frac{n_{j,-q} + \zeta}{n_{j,-q} + \breve{n}_j + 2\zeta} \cdot \frac{C^{WZ}_{wk}}{N_W} \cdot \frac{C^{WZ}_{wk,-q} + \xi}{\sum_{w'} C^{WZ}_{w'k,-q} + W\xi} \tag{4.6}$$

$$p(x_{ji} = 0, z_{ji} = e | w_{ji} = q, \mathbf{z}_{-q}, \mathbf{w}_{-q}, \eta, \zeta) \propto \frac{n_{j,-q} + \zeta}{n_j + \breve{n}_{j,-q} + 2\zeta} \cdot \frac{C^{WE}_{we,-q} + \eta}{\sum_{w'} C^{WE}_{w'e,-q} + W\eta} \tag{4.7}$$

Table 4.2: The Posterior Sampling Process

Preliminaries:

Suppose that at current stage of the sampling, only K of functional components have been assigned to the observations, define:

$\beta_u = \sum_{k=K+1}^{\infty} \beta_k, \gamma_v = \gamma/L, \gamma_u = \gamma(L-K)/L$, then we get:

$\beta = \{\beta_1, \cdots, \beta_v,\ \beta_u\} \sim Dirichlet(\gamma_1, \cdots, \gamma_r, \gamma_u)$

Repeat for each data observation until convergence:

Sampling z (may either equals to an existing k or $k_{new} = K+1$):

Firstly, integrate out π_j to get the marginal probability of :

$$p(\mathbf{z}|\beta) = \int_{\pi_j} p(\mathbf{z}|\pi_j) p(\pi_j|\alpha_0, \beta) \mathrm{d}\pi_j$$

$$= \prod_{j=1}^{J} \int \prod_{k=1}^{K} \pi_{jk}^{n_{jk}+\alpha_0\beta_k-1} \cdot \frac{\Gamma(\sum_{k=1}^{K} \alpha_0\beta_k)}{\prod_{k=1}^{K} \Gamma(\alpha_0\beta_k)} \mathrm{d}\pi_j$$

$$= \prod_{j=1}^{J} [\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0+n_j)} \cdot \prod_{k=1}^{K} \frac{\Gamma(\alpha_0\beta_k + n_{jk})}{\Gamma(\alpha_0\beta_k)}]$$

Secondly, get the posterior probability of z_{ii} given the data observations (not counting the current observation v_{ii})

$$p(z_{ji} = k|v_{ji}, \mathbf{z}^{-ji}, \mathbf{v}^{-ji}, \beta) \propto p(z_{ji} = k|\mathbf{z}^{-ji}, \beta) p(v_{ji}|z_{ji}, \mathbf{z}^{-ji}, \mathbf{v}^{-ji}, \beta)$$

$$\propto \left\{ \begin{array}{l} (\alpha_0\beta_k + n_{jk}^{-ji}) f_k^{-v_{ji}}(v_{ji}) \\ (\alpha_0\beta_k) f_{k_{new}}^{-v_{ji}}(v_{ji}) \end{array} \right.$$

For OG indicator w_{ii} , $f_k^{-w_{ji}}(w_{ji}) \propto \frac{C_{wk,-i}^{WZ}+\xi}{\sum_{w'=1}^{W} C_{w'k,-i}^{WZ}+W\xi} f_{k_{new}}^{-w_{ji}}(w_{ji}) \propto \frac{\xi}{W\xi}$

Sampling m: For each j, the auxiliary variable m$m(0 \leq m \leq n_{jk})$ is sampled as:

$p(m_{jk} = m|v_{ji}, \mathbf{m}^{-jk}, \mathbf{z}, \beta) = \frac{\Gamma(\alpha_0\beta_k)}{\Gamma(n_{jk}+\alpha_0\beta_k)} s(n_{jk}, m)(\alpha_0\beta_k)^m$

in which $s(n,m)$ is defined as: $s(0,0) = s(1,1) = 1$, $s(n,0) = 0$ $s(n,m) = 0 for m > n, s(n+1,m) = s(n,m-1) + ns(n,m)$ Sampling β: accumulate m_{ik} for all document j to get $m_1, m_2, \cdots, m_k$, then draw $\beta \sim Dirichlet(m_1, m_2, \cdots, m_k, \gamma)$

After a set of sampling processes based on the posterior distribution calculated above, other parameters can be sampled using the following equations:

$$\psi_e^{(w)} = \frac{C_{we,-q}^{WE}+\eta}{\sum_{w'} C_{we,-q}^{WE}+W\eta} \ \varphi_k^{(w)} = \frac{C_{wk,-q}^{WE}+\xi}{\sum_{w'} C_{wk,-q}^{WZ}+W\xi} \ \lambda_j^{(0)} = \frac{\breve{n}_{j,-q}+\zeta}{n_j+\breve{n}_{j,-q}+2\zeta} \ \lambda_j^{(1)} = \frac{n_{j,-q}+\zeta}{n_{j,-q}+\breve{n}_j+2\zeta}$$

4.5 EXPERIMENT RESULTS

4.5.1 Functional and Taxonomic Analysis of N-Mer Sub-Read Profiles by Probabilistic Topic Modeling

Taxonomic Data Analysis of Human Gut Microbial Samples

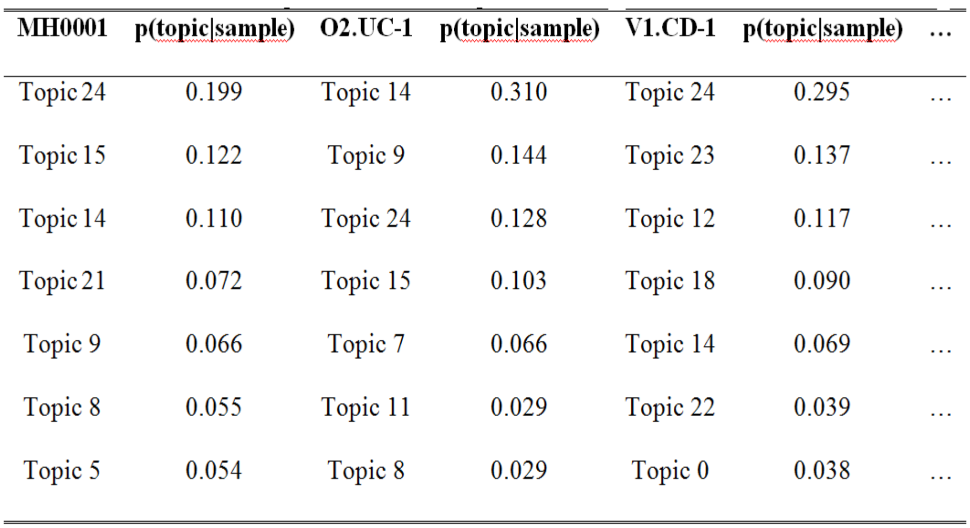
In this section, we conduct a generative topic modeling experiment for taxonomic analysis. Following the methods in Section 3, we apply the LDA topic model to the taxon abundance data of human gut microbial samples. The human gut microbial community taxon abundance data is generated by [32], which is openly accessible via http://gutmeta.genomics .org.cn/. According to [32], the Illumina GA reads from human gut microbial samples are firstly assembled into longer contigs. After that, the MetaGene program was used to predict open reading frames (ORFs) from those contigs. The predicted ORFs were then aligned to each other and grouped to a non-redundant gene set. The gene taxonomic assignment is achieved by carrying out BLASTP alignment against the NR database. The taxonomical level of each gene is determined by the lowest common ancestor (LCA). As a result of gene taxonomic assignment, the taxon abundance data for each sample can be produced.

The human gut microbial samples from [32] belong to both healthy subjects (HS) and patient with inflammatory bowel disease (IBD). Specifically, the IBD patients are from two different groups, one group with Crohns disease (CD), and the other group with ulcerative colitis (UC). In total, there are 85 healthy human gut microbial samples (MH0001 to MH0086), 15 UC samples (O2.UC-1 to O2.UC-24) and 12 CD samples (V1.CD-1 to V1.CD-15).

During topic modeling, we assume symmetric priors and set hyper-parameters following the methods in Section 3. The number of latent topics is set to be 50. In our approach, we apply a Gibbs sampling process to iteratively estimate the model from the genome sequence data. On the convergence of the Gibbs sampling process, we will be able to tell the topic-level distribution of taxa as well as the sample-level distribution of latent topics. Examples of uncovered latent topics are illustrated in Table 4.3-4.5. More specifically, Table 1 illustrates the top-ranked latent topics of three different samples, in which the ID of latent topics are sorted by the probability with respect to different samples. Table 4.4 represents the top-ranked taxa with respect to different latent topics, in which the taxa are sorted by the probability of being generated by latent topics. Table 4.5 shows the commonly shared top-ranked latent topics of different sample categories.

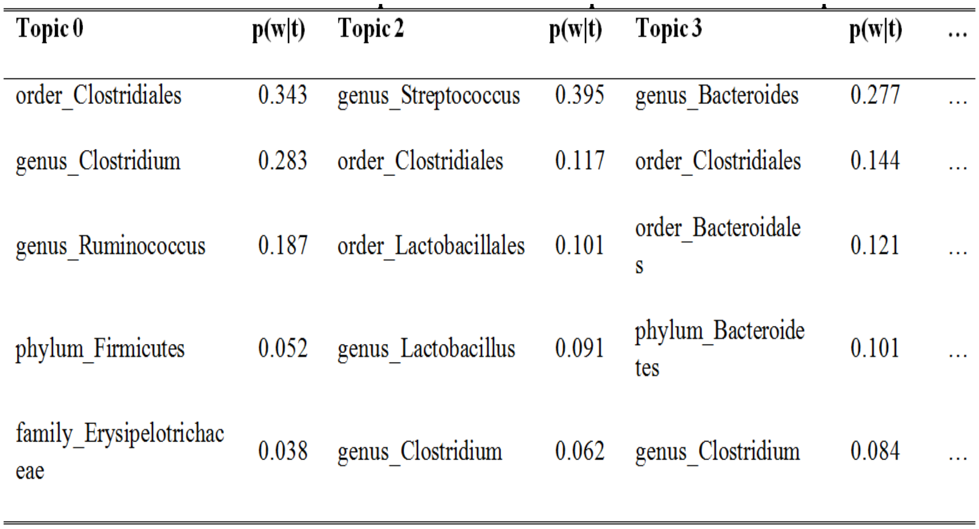
Table 4.6 illustrated the most relevant latent topics of each taxon. For each taxon, latent topics are sorted with respect to the mutual information (MI)score, which severs as a relevance measurement between taxa and latent topics. As shown in Table 4.6, phylum Firmicutes is most relevant to Topic 15. According to Table 4.5 we know that, Topic 15 is a common latent topic in Healthy and UC samples, yet it is not a common latent topic in CD samples. This may suggests that for CD samples, the proportion of bacteria belong to phylum Firmicutes is reduced. Similarly, since genus Clostridium is most relevant to Topic 14 and genus Bacteroides is most relevant to Topic 24, the prevalence of Topic 14 and 24 in samples may indicate the existence and possibly high abundance of genus Clostridium

Table 4.3: Illustration of Top-ranked Latent Topics with Respect to Different Microbial Samples



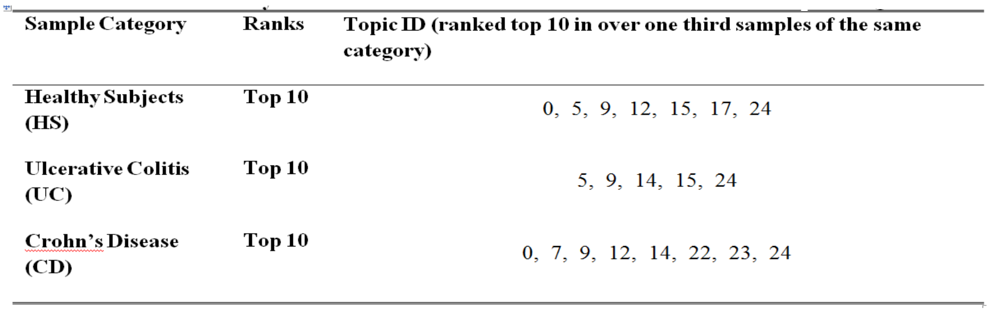
MH0001	p(topic|sample)	O2.UC-1	p(topic|sample)	V1.CD-1	p(topic|sample)	…
Topic 24	0.199	Topic 14	0.310	Topic 24	0.295	…
Topic 15	0.122	Topic 9	0.144	Topic 23	0.137	…
Topic 14	0.110	Topic 24	0.128	Topic 12	0.117	…
Topic 21	0.072	Topic 15	0.103	Topic 18	0.090	…
Topic 9	0.066	Topic 7	0.066	Topic 14	0.069	…
Topic 8	0.055	Topic 11	0.029	Topic 22	0.039	…
Topic 5	0.054	Topic 8	0.029	Topic 0	0.038	…

Table 4.4: Illustration of Top-ranked Taxa with Respect to Different Latent Topics



Topic 0	p(w|t)	Topic 2	p(w|t)	Topic 3	p(w|t)	…
order_Clostridiales	0.343	genus_Streptococcus	0.395	genus_Bacteroides	0.277	…
genus_Clostridium	0.283	order_Clostridiales	0.117	order_Clostridiales	0.144	…
genus_Ruminococcus	0.187	order_Lactobacillales	0.101	order_Bacteroidales	0.121	…
phylum_Firmicutes	0.052	genus_Lactobacillus	0.091	phylum_Bacteroidetes	0.101	…
family_Erysipelotrichaceae	0.038	genus_Clostridium	0.062	genus_Clostridium	0.084	…

Table 4.5: Commonly Shared Top-ranked Latent Topics for Three Different Sample Categories



Sample Category	Ranks	Topic ID (ranked top 10 in over one third samples of the same category)
Healthy Subjects (HS)	Top 10	0, 5, 9, 12, 15, 17, 24
Ulcerative Colitis (UC)	Top 10	5, 9, 14, 15, 24
Crohn's Disease (CD)	Top 10	0, 7, 9, 12, 14, 22, 23, 24

Table 4.6: Illustration of the Most Relevant Latent Topics with Respect to Different Taxa

	Topic ID	MI Score	Topic ID	MI Score	Topic ID	MI Score
family Enterobacteriaceae	Topic 48	0.02476	Topic 121	0.00915	Topic 31	0.00279
genus Clostridium	Topic 50	0.01628	Topic 153	0.01001	Topic 95	0.00765
genus Bacteroides	Topic 156	0.03030	Topic 77	0.02018	Topic 52	0.01661
phylum Bacteroidetes	Topic 132	0.00476	Topic 165	0.00260	Topic 67	0.00257
phylum Firmicutes	Topic 0	0.01256	Topic 99	0.00550	Topic 193	0.00212

and genus Bacteroides, correspondingly.

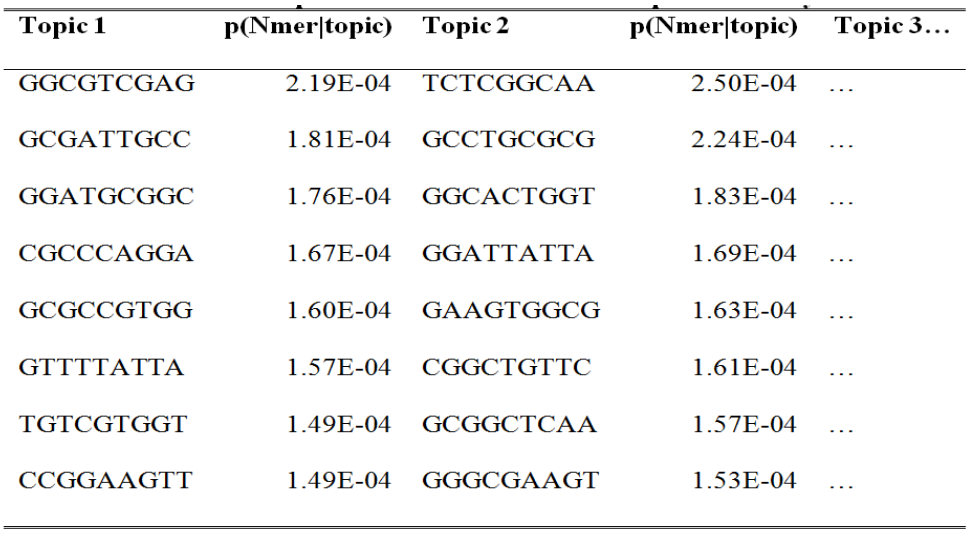
Our conclusion from the results is evidenced by the recent discoveries in fecal microbiota study of inflammatory bowel disease (IBD) patients [42], [31], [96], [33]. It has been reported that there is a significant reduction in the proportion of bacteria belonging to phylum Firmicutes in CD samples, which is consistent with our results. This can be explained by the fact mucosal microbial diversity is reduced in IBDs, particular in CD, which is associated with bacterial invasion of the mucosa. In UC, the inflammation is typically more superficial; therefore, the reduction of phylum Firmicutes in UC is not significant.

Gene Function Analysis of N-mer sequence

In this section, we deal with the problem of uncovering genome-level composition of N-mer latent patterns and explain the functional role of different components. More specifically, in order to deepen our understanding of genome composition and exploit the common function of genome sequences from the same species, we propose to identify the relatively stable part (core genome) and relatively diverse part (distributed genome) from the genome sequences by examining those uncovered latent genomic patterns. We apply the LDA topic model to N-mer sequence data obtained from standard reference sequences (FASTA format) of 635 genomes downloaded from the NCBI database. During topic modeling, we assume symmetric priors and set hyper-parameters. The number of latent topics is set to be 100. In our approach, we apply a Gibbs sampling process to iteratively estimate the model from the genome sequence data. On the convergence of the Gibbs sampling process, each N-mer is assigned a topic label. Also, we are able to tell the topic-level distribution of N-mer as well as the genome-level distribution of topics.

An example of uncovered latent topics is illustrated in Table 4.7 and Table 4.8. More specif-

Table 4.7: Illustration of Top-ranked 9-mers for Latent Topics Learnt by the LDA Model

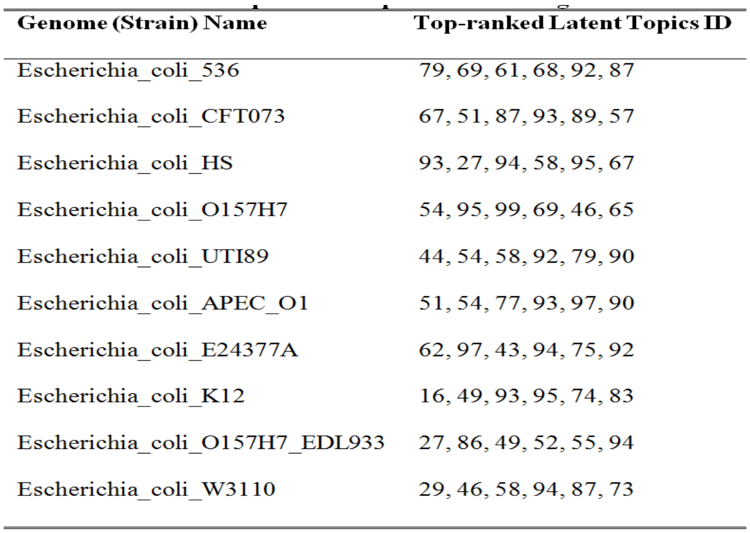


Topic 1	p(Nmer|topic)	Topic 2	p(Nmer|topic)	Topic 3…
GGCGTCGAG	2.19E-04	TCTCGGCAA	2.50E-04	…
GCGATTGCC	1.81E-04	GCCTGCGCG	2.24E-04	…
GGATGCGGC	1.76E-04	GGCACTGGT	1.83E-04	…
CGCCCAGGA	1.67E-04	GGATTATTA	1.69E-04	…
GCGCCGTGG	1.60E-04	GAAGTGGCG	1.63E-04	…
GTTTTATTA	1.57E-04	CGGCTGTTC	1.61E-04	…
TGTCGTGGT	1.49E-04	GCGGCTCAA	1.57E-04	…
CCGGAAGTT	1.49E-04	GGGCGAAGT	1.53E-04	…

ically, Table 4.7 represents the top-ranked N-mer (N=9) for each latent topic, in which the N-mers are sorted by the probability that they are generated by a certain latent topic. As we can see, the topic-level distribution of N-mers demonstrates a unique concurrence pattern, which may consistently present in different genomes. In other words, the presences of a certain latent topic in a genome may indicate the presence of a specific genomic component that be related to specific functional roles. Table 4.8 illustrates the top-ranked latent topic ID for genomes sets E. Coli, in which the latent topics are sorted by the probability that the given genome contains them.

Table 4.8 and Table 4.9 represent the top-ranked latent topic for genomes sets E. Coli and P. Marinus, which each involves at least 10 genomes from the same species. Table 4.10 further shows up the commonly shared top-ranked latent topics for both E. Coli and

Table 4.8: Top-ranked Topics for E. Coli Genomes

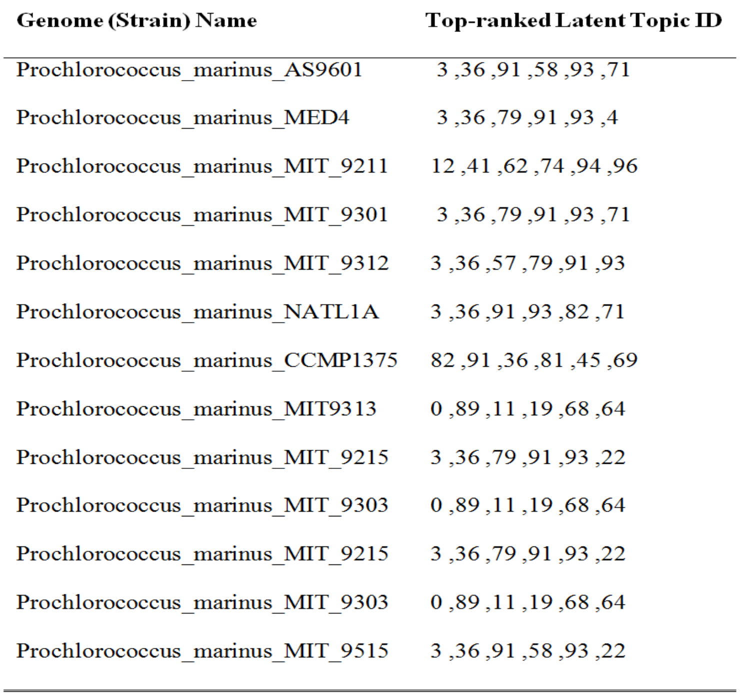


Genome (Strain) Name	Top-ranked Latent Topics ID
Escherichia_coli_536	79, 69, 61, 68, 92, 87
Escherichia_coli_CFT073	67, 51, 87, 93, 89, 57
Escherichia_coli_HS	93, 27, 94, 58, 95, 67
Escherichia_coli_O157H7	54, 95, 99, 69, 46, 65
Escherichia_coli_UTI89	44, 54, 58, 92, 79, 90
Escherichia_coli_APEC_O1	51, 54, 77, 93, 97, 90
Escherichia_coli_E24377A	62, 97, 43, 94, 75, 92
Escherichia_coli_K12	16, 49, 93, 95, 74, 83
Escherichia_coli_O157H7_EDL933	27, 86, 49, 52, 55, 94
Escherichia_coli_W3110	29, 46, 58, 94, 87, 73

P. Marinus genome sets. From the experiment results, we can see that the E. Coli genomes are really diverse, as they rarely share common latent topics among their top-ranked topics. This result suggests that the E. Coli species has more distributed genes than core genes as those strains seldom share common genomic components. One possible reason is that the whole genome sequence in strain of E.Coli has experience massive gene loss and gene gain which cause increasingly large intra-species genomes variation. In the contrary, P. Marinus, another genome set we studied, have some common latent topics shared among its different strains, which may potentially relate to the core genome.

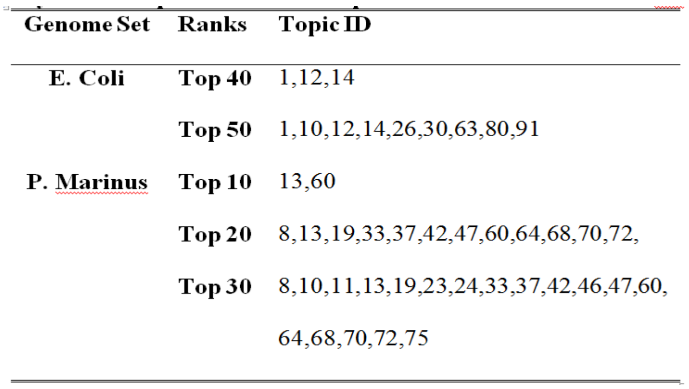
Although the genome-wide top-ranked latent topics provide us some insights about the common genomic patterns (core genome) in a set of genomes, however, in order to fully understand such a common genomic pattern, and identify the functional role of core genome,

Table 4.9: Top-ranked Topics for P. Marinus Genomes



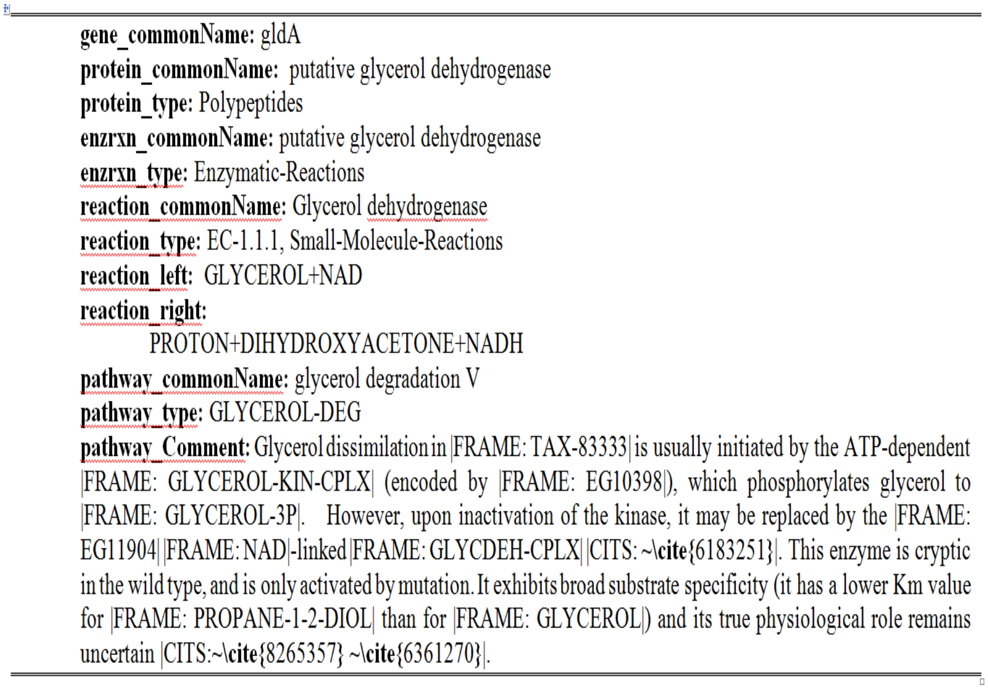
Genome (Strain) Name	Top-ranked Latent Topic ID
Prochlorococcus_marinus_AS9601	3 ,36 ,91 ,58 ,93 ,71
Prochlorococcus_marinus_MED4	3 ,36 ,79 ,91 ,93 ,4
Prochlorococcus_marinus_MIT_9211	12 ,41 ,62 ,74 ,94 ,96
Prochlorococcus_marinus_MIT_9301	3 ,36 ,79 ,91 ,93 ,71
Prochlorococcus_marinus_MIT_9312	3 ,36 ,57 ,79 ,91 ,93
Prochlorococcus_marinus_NATL1A	3 ,36 ,91 ,93 ,82 ,71
Prochlorococcus_marinus_CCMP1375	82 ,91 ,36 ,81 ,45 ,69
Prochlorococcus_marinus_MIT9313	0 ,89 ,11 ,19 ,68 ,64
Prochlorococcus_marinus_MIT_9215	3 ,36 ,79 ,91 ,93 ,22
Prochlorococcus_marinus_MIT_9303	0 ,89 ,11 ,19 ,68 ,64
Prochlorococcus_marinus_MIT_9215	3 ,36 ,79 ,91 ,93 ,22
Prochlorococcus_marinus_MIT_9303	0 ,89 ,11 ,19 ,68 ,64
Prochlorococcus_marinus_MIT_9515	3 ,36 ,91 ,58 ,93 ,22

Table 4.10: Commonly Shared Top-ranked Latent Topics for both E. Coli and P. marinus Genome Sets



Genome Set	Ranks	Topic ID
E. Coli	Top 40	1,12,14
	Top 50	1,10,12,14,26,30,63,80,91
P. Marinus	Top 10	13,60
	Top 20	8,13,19,33,37,42,47,60,64,68,70,72,
	Top 30	8,10,11,13,19,23,24,33,37,42,46,47,60, 64,68,70,72,75

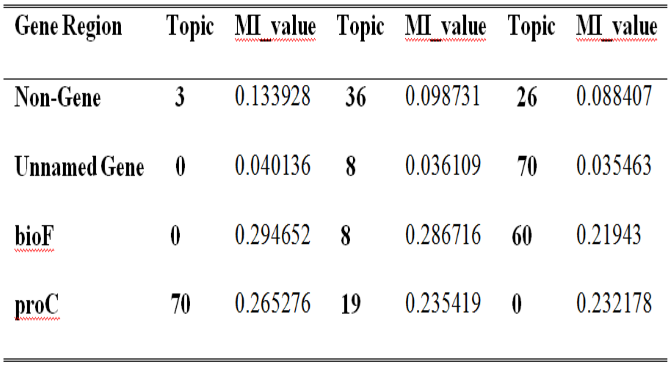
Table 4.11: Annotation (Enzyme and Pathway Information, Metabolic Capabilities) for Gene gldA



gene_commonName: gldA
protein_commonName: putative glycerol dehydrogenase
protein_type: Polypeptides
enzrxn_commonName: putative glycerol dehydrogenase
enzrxn_type: Enzymatic-Reactions
reaction_commonName: Glycerol dehydrogenase
reaction_type: EC-1.1.1, Small-Molecule-Reactions
reaction_left: GLYCEROL+NAD
reaction_right:
PROTON+DIHYDROXYACETONE+NADH
pathway_commonName: glycerol degradation V
pathway_type: GLYCEROL-DEG
pathway_Comment: Glycerol dissimilation in |FRAME: TAX-83333| is usually initiated by the ATP-dependent |FRAME: GLYCEROL-KIN-CPLX| (encoded by |FRAME: EG10398|), which phosphorylates glycerol to |FRAME: GLYCEROL-3P|. However, upon inactivation of the kinase, it may be replaced by the |FRAME: EG11904| |FRAME: NAD|-linked |FRAME: GLYCDEH-CPLX| |CITS: ~**cite**{6183251}|. This enzyme is cryptic in the wild type, and is only activated by mutation. It exhibits broad substrate specificity (it has a lower Km value for |FRAME: PROPANE-1-2-DIOL| than for |FRAME: GLYCEROL|) and its true physiological role remains uncertain |CITS:~**cite**{8265357} ~**cite**{6361270}|.

it is of great importance to study the relationship between latent topics and gene regions and give this relationship a biological explanation to help with the explanation of the functional role of the core genomes and distributed genomes. As mentioned, we exploited the BioCyc database to provide hierarchical functional annotations for gene regions, which involves enzyme and pathway information as well as the metabolic capabilities. An example of gene region functional annotation is illustrated in Table 4.11. Table 4.12 illustrated the most relevant latent topics of each gene region type (with mutual information score specified behind each topic ID). As mentioned in Section 4.4, after assigning topic labels to each

Table 4.12: Illustration of the Most Relevant Latent Topics of Some Gene Region in P. Marinus Genomes



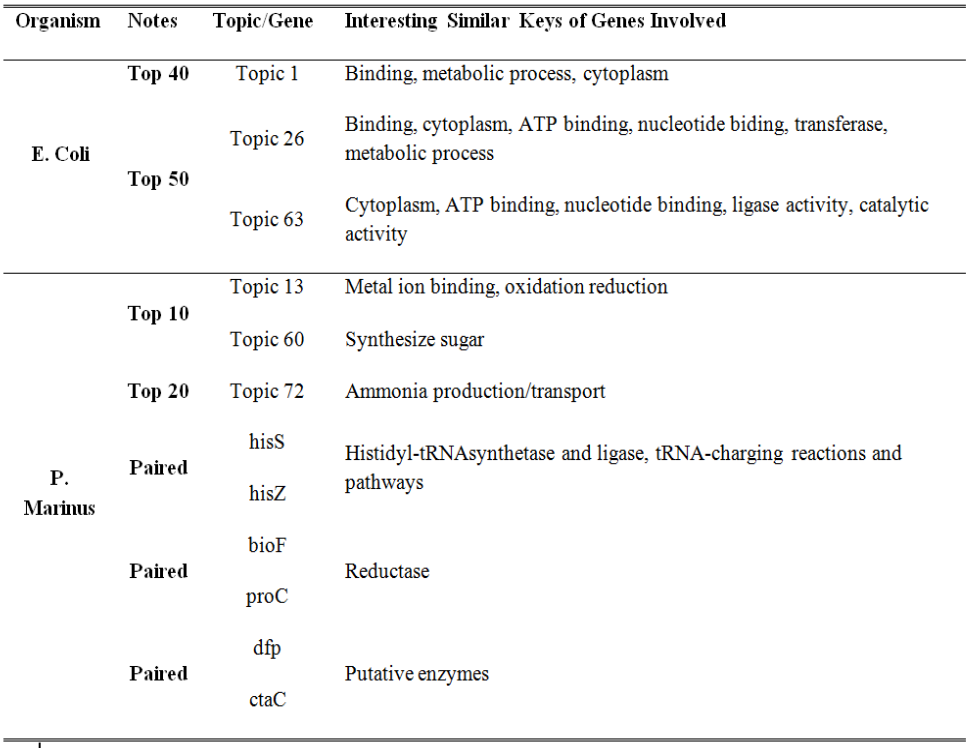
Gene Region	Topic	MI_value	Topic	MI_value	Topic	MI_value
Non-Gene	3	0.133928	36	0.098731	26	0.088407
Unnamed Gene	0	0.040136	8	0.036109	70	0.035463
bioF	0	0.294652	8	0.286716	60	0.21943
proC	70	0.265276	19	0.235419	0	0.232178

of the N-mers, the mutual information between latent topics and gene regions is calculated as the relevance measurement. It should be note that we also calculate mutual information for non-protein-encoding region and unnamed gene regions, under the name 'Non-Gene' and 'Unnamed Gene', respectively. According to Table 4.12, latent topics 3, 36 and 26 are in general more relevant to non-protein-encoding regions in P. Marinus genomes. Latent topics 0, 8, 70, on the other hand, are in general more relevant to protein-encoding regions. In Table 4.13, we highlight the major functions of genes that are most relevant to the commonly shared top-ranked latent topics, which provide us an insight about the functional role of the core genome. We also show that gene pairs relevant to the same latent topics also share some common gene functions, which indicates that the uncovered latent topics are biological informative and useful to the interpretation.

Conclusions

We introduce generative topic model to both homology-based approach and composition-

Table 4.13: Features for Top-ranked Latent Topics and the Paired Genes



Organism	Notes	Topic/Gene	Interesting Similar Keys of Genes Involved
E. Coli	**Top 40**	Topic 1	Binding, metabolic process, cytoplasm
	Top 50	Topic 26	Binding, cytoplasm, ATP binding, nucleotide biding, transferase, metabolic process
		Topic 63	Cytoplasm, ATP binding, nucleotide binding, ligase activity, catalytic activity
P. Marinus	**Top 10**	Topic 13	Metal ion binding, oxidation reduction
		Topic 60	Synthesize sugar
	Top 20	Topic 72	Ammonia production/transport
	Paired	hisS hisZ	Histidyl-tRNAsynthetase and ligase, tRNA-charging reactions and pathways
	Paired	bioF proC	Reductase
	Paired	dfp ctaC	Putative enzymes

based approach to further study the functional core (i.e. microbial core and gene core, correspondingly). We show that generative topic model can be used to model the taxon abundance information obtained by homology-based approach and study the microbial core. Our experimental results show that estimated generative topic model for taxon abundance data is able to uncover the structure of microbial groups in each sample. Secondly, the experimental results demonstrate that the proposed method is capable of characterizing a set of common genomic features (core genomes) shared by the genome sets, thus providing new insights into our understanding of genome composition. The developed framework also utilizes the BioCyc dataset to provide a reliable and comprehensive explanation of the functional roles for genome components, which enable us to acquire the enzyme and pathway information as well as the major metabolic capabilities of genomic components. We also show that latent topic modeling can be used to characterize core and distributed genes within a species and to correlate similarities between genes and their functions.

4.5.2 Estimating Functional Groups in Human Gut Microbiome with Probabilistic Topic Models

In this section, we conduct a probabilistic topic modeling experiment to identify functional groups from microbial samples from two large published gut microbiome datasets: the Illumina-based metagenomics data from 112 Danish individuals [32] and the Sanger-sequenced meta-genome of 39 individuals dataset [28]. Following the methods in Section 4.4,we apply the proposed probabilistic topic models to the functional element abundance data acquired from non-redundant gene catalog of human gut microbial samples.

Experimental Data Collection

The Illumina human gut microbial community taxon abundance data is generated by [32], which is openly accessible via http://gutmeta.genomics.org.cn/. According to [32], the Illumina GA reads from human gut microbial samples are firstly assembled into longer contigs. After that, the Glimmer program was used to predict protein-encoding sequences (CDs) from assembled contigs. The predicted CDs sequences were then aligned to each other and form a non-redundant CDs catalog (a.k.a. minimal gut genome). The non-redundant CDs catalog consists of 3,299,822 non-redundant CDs sequences with an average length of 704 bp.For a given non-redundant CDs sequence, its NCBI taxonomical level is obtained by carrying out BLASTP alignment against the NCBI NR database. The taxonomical level of each non-redundant CDs sequence is determined by the lowest common ancestor (LCA) based algorithm. The taxonomic abundance data for each sample can be computed by counting the indicators of NCBI taxonomical levels. The assignments of gene orthologous indicator and KEGG pathway indicator are achieved by BLASTP alignment of the amino-acid sequence from predicted CDs to the eggNOG database and KEGG database. The human gut microbial samples from [32] belong to both healthy subjects (HS) and patients with inflammatory bowel disease (IBD). Specifically, the IBD patients are from two different groups, one group with Crohn's disease (CD), and the other group with ulcerative colitis (UC). In total, there are 85 healthy samples, 15 UC samples and 12 CD samples.

The Sanger sequenced gut microbime dataset [28] includes 22 European meta-genomes from Danish, French, Italian and Spanish individuals in combine with Sanger gut dataset from 13 Japanese and 4 American individuals. For sequencing processing, the raw Sanger reads are trimmed to remove low-quality reads and possible human DNA contaminations.

The cleaned Sanger reads are then assembled to longer contigs for gene prediction. The phylogentic annotation of samples was performed by aligned Sanger reads against a total of 1,511 reference genomes.The gene functions are annotated via BLASTP against eggNOG and KEGG databases, which yields high through-put gene function profiling, as 63.5% of the predicted genes in the Sanger-sequenced samples can be assigned to the orthologous group. The gene function profile, may then be used to study the composition of eggNOG and KEGG orthologous groups across distinct samples.

Topic Inferring from Predicted Gene Cataloguewith LDA-B Model

As introduced in Section 4.4, the functional elements, which bear an analogy with text words, includes three different types of indicators, i.e. NCBI taxonomic level indicators, indicator of gene orthologous groups and KEGG pathway indicators. The union of unique functional elements jointly defines a fixed word vocabulary. In Illumina dataset, there are 647,136 NCBI taxonomic level indicators, with a vocabulary size of 748; there are a total of 1,293,764 gene orthologous group indicators, with a vocabulary size of 4667; and there are 953,493 KEGG pathway indicators, with a vocabulary size of 237.

It should be pointed out that, in our approach we separately estimated three probabilistic topic models with respect to three different types of functional elements (i.e. NCBI taxonomic level indicators, indicator of gene orthologous groups and KEGG pathway indicators). We apply a Gibbs sampling process to iteratively update the model estimation from the functional element abundance data until converge (basically, it takes less than 100 iterations to converge). During topic modeling, we assume symmetric priors and set hyperparameters.On the convergence of the Gibbs sampling process, we will be able to tell the topic-level distribution of functional elements as well as the sample-level distribution of latent topics. In our experiment, we test different topic numbers on the proposed LDA-B model and compare the log-likelihood. Log-likelihood is one of the standard criteria for generative model evaluation. It provides a quantitative measurement of how well a topic model fits the training data. The score of log-likelihood (which is a negative number) is the higher the better. In practice, the log-likelihood of elements given latent topics can be calculated by integrating out all the latent variables.

$$\begin{aligned} p(\mathbf{w}|\mathbf{z}) &= \prod_{t=1}^{T} \left(\int_{\varphi_{z_t}} p(\mathbf{w}|z_t, \varphi_{z_t})\, p(\varphi_{z_t}|z_t)\, \mathrm{d}\varphi_{z_t} \right) \\ = \left(\frac{\Gamma(W\beta)}{\Gamma(\beta)^W}\right)^T &\prod_{t=1}^{T} \frac{\prod_{w_i} \Gamma\left(n_t^{(w_i)} + \beta\right)}{\Gamma\left(n_t^{(\cdot)} + W\beta\right)} \cdot \frac{\Gamma(W\eta)}{(\eta)^W} \cdot \frac{\prod_{w_i} \Gamma\left(n_0^{(w_i)} + \eta\right)}{\Gamma\left(n_0^{(\cdot)} + W\eta\right)} \end{aligned} \tag{4.8}$$

For the LDA model, the log-likelihood can be calculated in a similar way. In Fig. 4.7(a-c),we show the log-likelihood comparison of the proposed LDA-B models and LDA model on three different types of functional elements under different topic number. It shows that, for both models, the likelihood increases as the number of topic increases, which means that a relatively larger topic numbers may potentially result in better fitting of the data. However, it should be noted that there is a trade-off between topic numbers and convergence time of models. And, as we would see in next section, the increase of topic number does not always lead to the improvement of predictive results. In general, the log-likelihood of LDA model is higher than that of the LDA-B model, which shows LDA model fit the training data better. The difference between two models can be explained by the introducing of background topic in the LDA-B model.

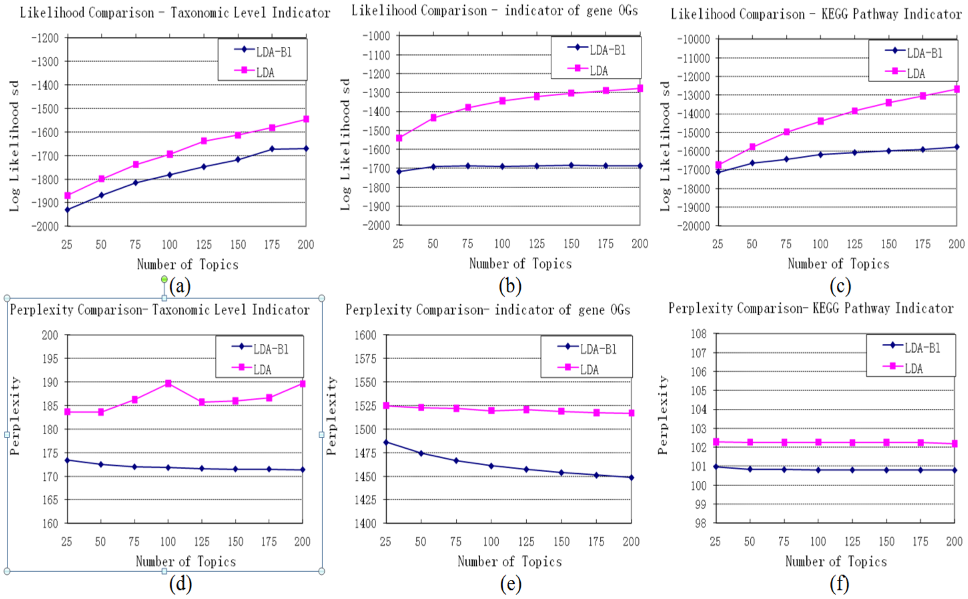
The Perplexity is a widely used criterion for evaluating the predictive ability of probabilistic topic models. The perplexity is calculated for held-out testing data. In our experiment, we use a 50% subset of the functional elements as training data and the other 50% as testing data. On constructing the two subsets, we ensure that functional elements from the same sample are equally split to both subsets. In practice, it is the inverse predicted model likelihood of data in held-out testing data, using parameters inferred from the trained topic model. Thus the smaller perplexity value indicates better model fitting.

$$perplexity\left(D_{test}\right) = exp\left(\frac{-\sum_{j=1}^{D_{test}} log\left(p\left(t_j\right)\right)}{\sum_{j=1}^{D_{test}} N_j^t}\right) \tag{4.9}$$

One advantage of our LDA-B model is that it assigns commonly shared functional elements to the background distribution, which makes the model more suitable to represent genome-wide co-existing patterns of functional elements. Fig. 4.7(d-f)represents the perplexity comparison of the proposed LDA-B models and LDA model on three different types of functional elements as the topic number increasing. It shows that the perplexity of our model is consistently lower than LDA model, which suggests that our model is 'less surprised' by the testing data, thus demonstrates better predictive ability. Also, it shows that the predictive ability of our model may benefit from greater topic number, as it tends to have lower perplexity as the topic number increases. The proposed LDA-B model achieves best log-likelihood and perplexity scores when topic number equals to 200. Therefore, the LDA-B models are inferred with topic number set to 200.

Inferring Functional Basis for Enterotypes with Extended Background HDP Model

In this section, we investigate the performance of the proposed Enterotype-HDP model using the Sanger-sequenced meta-genome samples [28].



(a) (b) (c)

(d) (e) (f)

Figure 4.7: (a-c) The Log-likelihood comparison of the proposed LDA-B models and LDA model on three different types of functional elements (as topic number changing), (e-f) The perplexity comparison on three different types of functional elements (as topic number increasing).

In [28], the predicted gene catalog from Sanger-sequenced meta-genome samples covering a wide spectrum of bacteriaonly 0.14% of the reads could be classified as human contamination. Also, 63.5% of the predicted genes in the Sanger-sequenced samples can be assigned to the orthologous groups. Across the 33 samples in 3 distinct Enterotypes, there are 2,319,439 genes assigned to 13507 eggNOG orthologous groups and 1,543,293 genes mapped to 4,900 KEGG orthologous groups.

The values of global concentration parameter γ are determined by log-likelihood and perplexity comparison on a serial of values. Other hyper-parameters (such as Dirichlet distribution priors:$\xi,\eta$and Beta distribution prior ζ) are set in prior and fixed during the experiments. The prediction of functional basis for each Enerotype and functional mixture components across the samples is achieved by performing Gibbs sampling on sample orthologous-group (OG) profiles (including both eggNOG and KEGG OG indicators) to estimate the sample-level distribution of switch variable and functional components. The output will be a set of functional components and Enterotype functional basis inferred from the training dataset.

Fig. 4.8 shows the log-likelihood comparison of the Enterotype HDP model with different concentration parameterγ.Overall, the log-likelihood of the model increase over the iterations during the Gibbs sampling process, indicating better fitting to the training data. The best (highest) log-likelihood score is achieved with $\gamma = 15.0$. We also compare the perplexity of Enterotype HDP model on a serial of γ values. The results in Fig. 4.9 show that the model achieve best perplexity score when=20.0 .

Illustration of Discovered Latent Themes

One major objective of the proposed models is inferring functional groups from meta-

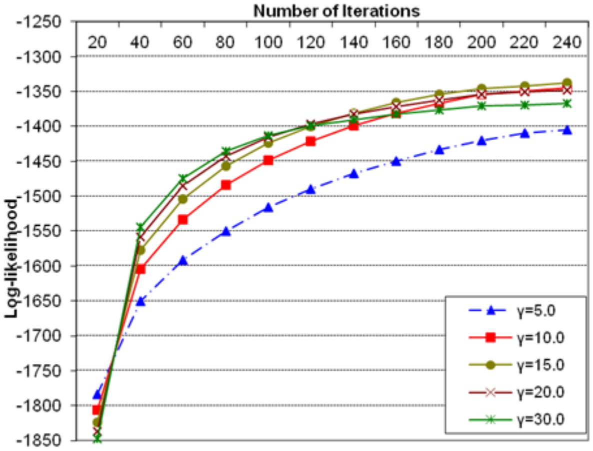


Figure 4.8: Log-likelihood comparison of Enterotype HDP model

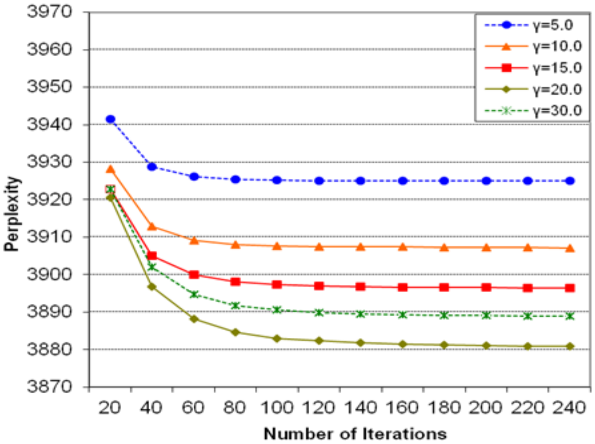


Figure 4.9: Perplexity comparison of Enterotype HDP model

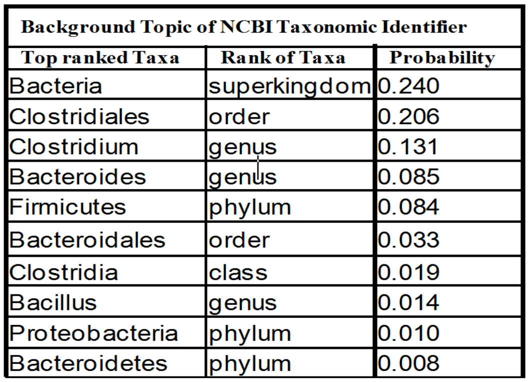
genomes to facilitate knowledge organizing and interpreting the biological processes encoded in meta-genome sequences. Inferred latent topic may provide more details to study both the phylogenetic variation at the genus and phylum level and the functional variations at gene and functional class levels across samples. With this consideration, we visualize the uncovered background topics of NCBI taxonomic level indicators, geneOGs indicators and KEGG pathway indicators from three independent LDA-B models and providing the top-ranked functional elements (Table 4.14-4.16).

Table 4.14 illustrates the background topic of taxonomic level indicators, which provides an insight of the bacteria core of the most common co-existing taxa across meta-genome samples. Table 4.15 represents the background topic of gene OGs indicators. As we can see, the top-ranked functional elements not only involves general biology process and molecular functions such as signal transduction, metabolic capacity, and important protein synthesis (RNA and DNA polymerase, ATP synthase) but also involves gut-specific functions such as adhesion to the host protein or in harvesting sugars of the glycolipids. Table 4.16 shows the background topic of KEGG pathway indicators, which involves the main metabolic pathways such as carbon metabolism and amino acid metabolism.

More examples about uncovered latent topics with respect to NCBI taxonomic indicators are illustrated in Table4.17 -4.19.Specifically, Table 4.17 illustrates the top-ranked latent topics of three different samples, in which the ID of latent topics are sorted by the probability with respect to different samples. Table 4.18 represents the top-ranked taxa with respect to different latent topics, in which the taxa are sorted by the probability of being generated by topics.

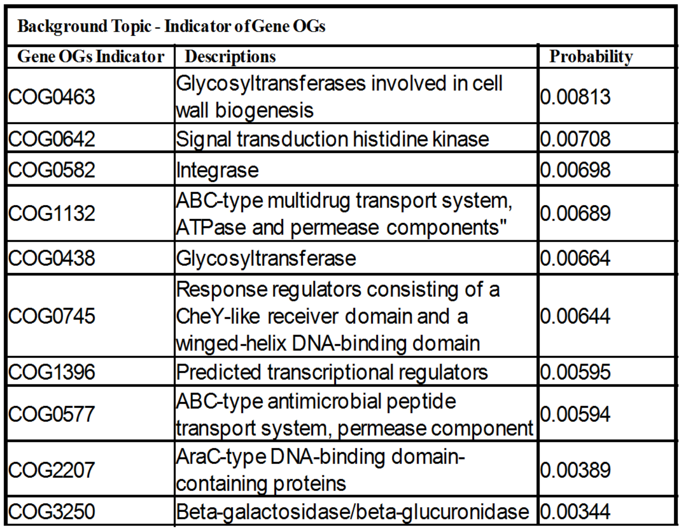
Table 4.19 illustrated the most relevant latent topics of each taxon. For each taxon, latent

Table 4.14: Illustration of the Background Topic of Taxonomic Level Indicators



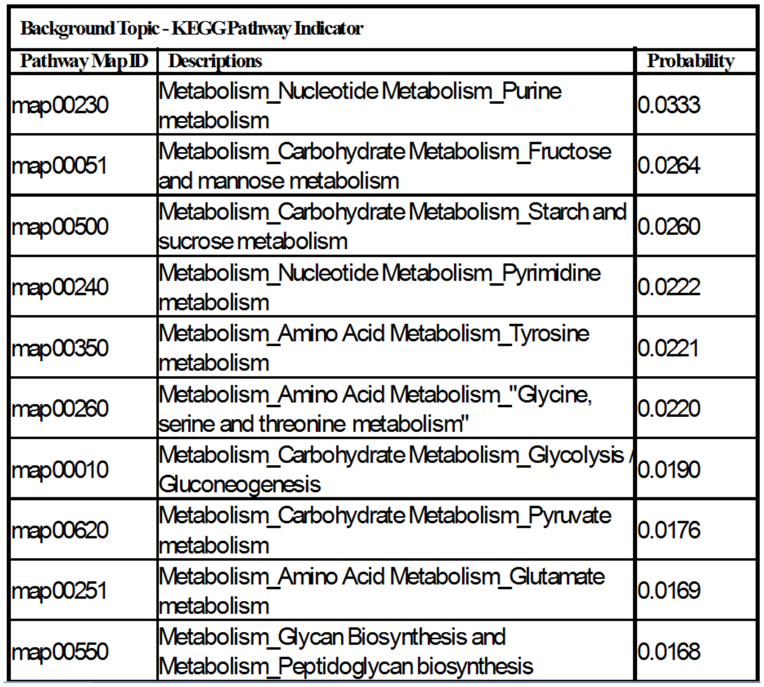
Background Topic of NCBI Taxonomic Identifier		
Top ranked Taxa	**Rank of Taxa**	**Probability**
Bacteria	superkingdom	0.240
Clostridiales	order	0.206
Clostridium	genus	0.131
Bacteroides	genus	0.085
Firmicutes	phylum	0.084
Bacteroidales	order	0.033
Clostridia	class	0.019
Bacillus	genus	0.014
Proteobacteria	phylum	0.010
Bacteroidetes	phylum	0.008

Table 4.15: Illustration of the Background Topic of Gene Ogs Indicators



Background Topic - Indicator of Gene OGs		
Gene OGs Indicator	**Descriptions**	**Probability**
COG0463	Glycosyltransferases involved in cell wall biogenesis	0.00813
COG0642	Signal transduction histidine kinase	0.00708
COG0582	Integrase	0.00698
COG1132	ABC-type multidrug transport system, ATPase and permease components"	0.00689
COG0438	Glycosyltransferase	0.00664
COG0745	Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain	0.00644
COG1396	Predicted transcriptional regulators	0.00595
COG0577	ABC-type antimicrobial peptide transport system, permease component	0.00594
COG2207	AraC-type DNA-binding domain-containing proteins	0.00389
COG3250	Beta-galactosidase/beta-glucuronidase	0.00344

Table 4.16: Illustration of the Background Topic of Kegg Pathway Indicators



Background Topic - KEGG Pathway Indicator		
Pathway Map ID	Descriptions	Probability
map00230	Metabolism_Nucleotide Metabolism_Purine metabolism	0.0333
map00051	Metabolism_Carbohydrate Metabolism_Fructose and mannose metabolism	0.0264
map00500	Metabolism_Carbohydrate Metabolism_Starch and sucrose metabolism	0.0260
map00240	Metabolism_Nucleotide Metabolism_Pyrimidine metabolism	0.0222
map00350	Metabolism_Amino Acid Metabolism_Tyrosine metabolism	0.0221
map00260	Metabolism_Amino Acid Metabolism_"Glycine, serine and threonine metabolism"	0.0220
map00010	Metabolism_Carbohydrate Metabolism_Glycolysis / Gluconeogenesis	0.0190
map00620	Metabolism_Carbohydrate Metabolism_Pyruvate metabolism	0.0176
map00251	Metabolism_Amino Acid Metabolism_Glutamate metabolism	0.0169
map00550	Metabolism_Glycan Biosynthesis and Metabolism_Peptidoglycan biosynthesis	0.0168

topics are sorted with respect to the mutual information score (MI score). As shown in Table 8, phylum Firmicutes is most relevant to the background topic (Topic 0). According to Table4.17, the probability of Topic 0 in Healthy and UC samples (0.475 in MH0001 and 0.363 in O2.UC-1) is much higher than that in CD samples (0.286 in V1.CD-1). This suggests that for CD samples, the proportion of bacteria belong to phylum Firmicutes is significantly reduced. Similarly, since genus Clostridium is most relevant to Topic 50, 153, 95 and genus Bacteroides is most relevant to Topic 156, 77, 52, the prevalence of Topic 95 and 52 in samples O2.UC-1 and sample V1.CD-1 may indicate the existence and possibly high abundance of genus Clostridium and genus Bacteroides, correspondingly. Our conclusion from the results is evidenced by the recent discoveries in fecal microbiota study of inflammatory bowel disease (IBD) patients [96], [42], [31], [33]. It has been reported that there is a significant reduction in the proportion of bacteria belonging to phylum Firmicutes in CD samples, which is consistent with our results. This can be explained by the fact mucosal microbial diversity is reduced in IBDs, particular in CD, which is associated with bacterial invasion of the mucosa. In UC, the inflammation is typically more superficial; therefore, the reduction of phylum Firmicutes in UC is not significant.

In our experiment, the phylogenetic composition inferred from latent topics(Fig. 4.10)agrees with previous observations in [32]and [28]: the Firmicutes and Bacteroidetes phyla constitute the vast majority of the dominant human gut microbiota, and Bacteroidesis among the most abundant yet most variable genus across samples.

In order to facilitate analyzing the composition of microbiome community of human gut across cohorts, and get insights into functional differences between gut microbiomes across different samples, we use the extended HDP model to infer the functional basis of each of

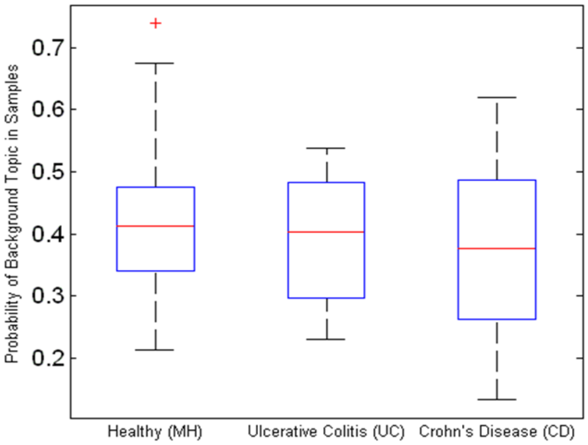


Figure 4.10: Box-plot of background topic probability in samples

the three identified Enterotype in [28]. We illustrate the inferred functional basis ψ_e learned from the corresponding Enterotypes in Table 4.20-4.22.

Conclusions

In this section, based on the functional elements derived from the non-redundant CDs catalogue, we have shown that the configuration of functional groups encoded in the gene-expression data of meta-genome samples can be inferred by applying probabilistic topic modeling to functional elements derived from the non-redundant CDs catalogue (including taxonomic levels, indicators of gene orthologous groups and KEGG pathway mappings). When used to study microbial samples, the proposed model considers each sample as a ‘document’, which has a mixture of ‘latent topic’; while each latent topic is a weighted mixture of functional elements that bear an analogy with ‘words’. We also introduce the extended Enterotypr-HDP model to infer functional basis from detected enterotypes. The latent topics estimated from human gut microbial samples are evidenced by the recent dis-

Table 4.17: Illustration of Top-ranked Latent Topics with Respect to Different Microbial Samples

MH0001	$p(t\|sample)$	O2.UC-1	$p(t\|sample)$	V1.CD-1	$p(t\|sample)$	···
Topic 0	0.475	Topic 0	0.363	Topic 0	0.286	···
Topic 124	0.116	Topic 95	0.101	Topic 61	0.124	···
Topic 181	0.103	Topic 143	0.062	Topic 12	0.116	···
Topic 159	0.040	Topic 83	0.059	Topic 115	0.050	···
Topic 86	0.027	Topic 65	0.056	Topic 52	0.048	···
Topic 72	0.018	Topic 139	0.034	Topic 32	0.037	···
Topic 19	0.017	Topic 59	0.033	Topic 50	0.036	···

Table 4.18: Illustration of Top-ranked Taxa with Respect to Different Latent Topics

Topic 1	$p(w\|t)$	Topic 2	$p(w\|t)$	Topic 3	$p(w\|t)$	···
order Clostridiales	0.343	genus Streptococcus	0.395	genus Bacteroides	0.277	···
genus Clostridium	0.283	order Clostridiales	0.117	order Clostridiales	0.144	···
genus Ruminococcus	0.187	order Lactobacillales	0.101	order Bacteroidales	0.121	···
phylum Firmicutes	0.052	genus Lactobacillus	0.091	phylum Bacteroidetes	0.101	···
family Erysipelotrichaceae	0.038	genus Clostridium	0.062	genus Clostridium	0.084	···

Table 4.19: Illustration of the Most Relevant Latent Topics with Respect to Different Taxa

	Topic ID	MI Score	Topic ID	MI Score	Topic ID	MI Score
family Enterobacteriaceae	Topic 48	0.02476	Topic 121	0.00915	Topic 31	0.00279
genus Clostridium	Topic 50	0.01628	Topic 153	0.01001	Topic 95	0.00765
genus Bacteroides	Topic 156	0.03030	Topic 77	0.02018	Topic 52	0.01661
phylum Bacteroidetes	Topic 132	0.00476	Topic 165	0.00260	Topic 67	0.00257
phylum Firmicutes	Topic 0	0.01256	Topic 99	0.00550	Topic 193	0.00212

coveries in fecal microbiota study, which demonstrate the effectiveness of the proposed method.

Table 4.20: Illustration of the Functional Basis of Gene Ogs in Enterotype 1 of Sanger Sequenced Samples

Orthologous Group	Descriptions	Probability
COG0642	Signal transduction histidine kinase	0.009613
COG1132	ABC-type multidrug transport system, ATPase and permease components	0.006868
COG0745	Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain	0.005773
COG0577	ABC-type antimicrobial peptide transport system, permease component	0.005155
COG3451	Type IV secretory pathway, VirB4 compo-nents	0.004516
COG3250	Beta-galactosidase/beta-glucuronidase	0.004511
COG0550	Topoisomerase IA	0.004370
COG0463	Glycosyltransferases involved in cell wall biogenesis	0.003678
COG1472	Beta-glucosidase-related glycosidases	0.003632
COG1595	DNA-directed RNA polymerase specialized sigma subunit, sigma24 homolog	0.003350
K03169	DNA topoisomerase III [EC:5.99.1.2]	0.003330
K03088	RNA polymerase sigma70 factor, ECF subfamily	0.003151

Table 4.21: Illustration of the Functional Basis of Gene Ogs in Enterotype 2 of Sanger Sequenced Samples

Orthologous Group	Descriptions	Probability
COG1132	ABC-type multidrug transport system, ATPase and permease components	0.006746
COG0642	Signal transduction histidine kinase	0.006288
COG0745	Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain	0.005296
COG3451	Type IV secretory pathway, VirB4 components	0.004350
COG1373	Predicted ATPase (AAA+ superfamily)	0.004011
COG3505	Type IV secretory pathway, VirD4 components	0.003961
COG0577	ABC-type antimicrobial peptide transport system, permease component	0.003841
COG3344	Retron-type reverse transcriptase	0.003704
COG0550	Topoisomerase IA	0.003307
COG1472	Beta-glucosidase-related glycosidases	0.003187
COG0463	Glycosyltransferases involved in cell wall biogenesis	0.003164
COG0178	Excinuclease ATPase subunit	0.003083
K03205	type IV secretion system protein VirD4	0.002699
COG0463	Glycosyltransferases involved in cell wall biogenesis	0.003164

Table 4.22: Illustration of the Functional Basis of Gene Ogs in Enterotype 3 of Sanger Sequenced Samples

Orthologous Group	Descriptions	Probability
COG0745	Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain	0.007661
COG3451	Type IV secretory pathway, VirB4 components	0.007353
COG0577	ABC-type antimicrobial peptide transport system, permease component	0.005765
COG3505	Type IV secretory pathway, VirD4 components	0.004628
COG0550	Topoisomerase IA	0.004314
COG1472	Beta-glucosidase-related glycosidases	0.004213
COG3250	Beta-galactosidase/beta-glucuronidase	0.003617
COG0463	Glycosyltransferases involved in cell wall biogenesis	0.003524
COG1136	ABC-type antimicrobial peptide transport system, ATPase component	0.003474
K03205	type IV secretion system protein VirD4	0.003246
K03169	DNA topoisomerase III [EC:5.99.1.2]	0.002958

5. CONCLUSIONS

In this thesis, a set of novel probabilistic topic models have been proposed to address challenging issues in image mining and bioinformatics studies. The contributions are as follows.

To leverage image, text and user-created tags to enhance the performance of image annotation and retrieval,we have introduced novel image representation, and a wide-range of algorithms and methods including the saliency model (salient regions and key-points) as a complement part of spatial layout model for image representation. Several probabilistic topic models are proposed for effective and robust modeling of the co-existing image features, annotations, user-perspective and the semantic relations between visual attributes and object categories. Specifically,a probabilistic topic-connection (PTC) model is proposed for co-existing image features and annotations, in which new latent variables are introduced to allow for more flexible sampling of word topics and visual topics, allowing one word topic may connect to multiple visual topics. A perspective hierarchical Dirichlet process (pHDP) model is proposed to deal with user-tagged image modeling, associating image features with image tags and incorporating the user's perspectives into the image tag generation process. New latent variables are introduced to determine if an image tag is generated from user's perspectives or from the image content. Moreover, the automatic framework forvisual attributes identification and semantic relation learning between visual attributes and object categories is proposed. The semantic associations between visual attributes and object categories are then incorporated into a text-based topic model to infer descriptive latent topics from natural language knowledge base. It's shown that in mining large scale text corporaof natural language descriptions, the relation between semantic visual attributes and object categories can be encoded as Must-Links and Cannot-Links, which can be represented by Dirichlet-Forest prior.

We also introduce generative topic model to meta-genomics studies. We show that generative topic model can be used to model the taxon abundance information obtained by homology-based approach and study the microbial core. Our experimental results show that estimated generative topic model for taxon abundance data is able to uncover the structure of microbial groups in each sample. Secondly, the experimental results demonstrate that the proposed method is capable of characterizing a set of common genomic features (core genomes) shared by the genome sets, thus providing new insights into our understanding of genome composition. The developed framework also utilizes the BioCyc dataset to provide a reliable and comprehensive explanation of the functional roles for genome components, which enable us to acquire the enzyme and pathway information as well as the major metabolic capabilities of genomic components. We also show that latent topic modeling can be used to characterize core and distributed genes within a species and to correlate similarities between genes and their functions.Based on the functional elements derived from the non-redundant CDs catalogue, our study shows that the configuration of functional groups encoded in the gene-expression data of meta-genome samples can be inferred by applying probabilistic topic modeling to functional elements derived from the non-redundant CDs catalogue (including taxonomic levels, indicators of gene orthologous groups and KEGG pathway mappings). When used to study microbial samples, the proposed model considers each sample as a 'document', which has a mixture of 'latent topic';

while each latent topic is a weighted mixture of functional elements that bear an analogy with ‘words’. The extended Enterotypr-HDP model is introduced to infer functional basis from detected enterotypes. The latent topics estimated from human gut microbial samples are evidenced by the recent discoveries in fecal microbiota study, which demonstrate the effectiveness of the proposed models.

In summary, a broad range of topics in statistical learning, image processing, social network analysis, content-based image retrieval and bioinformatics studies are addressed in this thesis. A set of the robust probabilistic topic models and annotation algorithms are developed. Research outcomes from this thesis will lead to more efficient and effective modeling and simulation mechanism insemantic image annotation, statistical learning, bioinformatics and social network analysis.

Bibliography

[1] Fei-Fei Li 0002, Robert Fergus, Pietro Perona, and Pietro Perona. One-shot learning of object categories. pages 594–611, 2006.

[2] Fei-Fei Li 0002, Pietro Perona, California Institute of Technology, and California Institute of Technology. A bayesian hierarchical model for learning natural scene categories. In *CVPR (2)*, pages 524–531, 2005.

[3] Jun Yang 0003, Yu-Gang Jiang, Alexander G. Hauptmann, Chong-Wah Ngo, and Chong-Wah Ngo. Evaluating bag-of-visual-words representations in scene classification. In *Multimedia Information Retrieval*, pages 197–206, 2007.

[4] Brady A. and Salzberg S. L. Phymm and phymmbl: metagenomic phylogenetic classification with interpolated markov models. *Nature Methods*, 6(9):673–676, 2009.

[5] N. Laird A. Dempster and D. Rubin. Maximum-likelihood from incomplete data via the em algorithm. *J. Royal Statistical Soc.*, B-39, 1977.

[6] S. Santini A. Gupta A. W. M. Smeulders, M. Worring and R. Jain. Content-based image rerieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):13491380, 2000.

[7] Stern H.S. A.Gelman, J.B.Carlin and Rubin D.B. *Bayesian Data Analysis*. Chapman Hall/CRC, 1995.

[8] William W. Cohen Robert F. Murphy Amr Ahmed, Eric P. Xing. Structured correspondence topic models for mining captioned figures in biomedical literature. In *Proceedings of the 15th ACM SIGKDD International conference on Knowledge discovery and data mining*, Paris, France, June 28-July 01 2009.

[9] Eguchi K Aso T. Predicting protein-protein relationships from literature using latent topics. *Genome Inform*, 23:3–12, 2009.

[10] Humphreys B. and Lindberg D. The umls project: making the conceptual connection between users and the information they need. *Bulletin of the Medical Library Association*, 81:170, 1993.

[11] A. Smola B. Schoelkopf. *Learning with Kernels*. MIT Press, 2002.

[12] Carson C Belongie S. Color- and texture-based image segmentation using em and its application to content-based image retrieval. In *ICCV'98*, pages 675–682, 1998.

[13] J.A. Bilmes. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *Technical Report*, 1998.

[14] D.M. Blei and M.I. Jordan. Modeling annotated data. In *The 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 127–134, Toronto, Canada, 2003.

[15] Ng A..and Jordan M. Blei D. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[16] A.Muoz X. Bosch, A.Zisserman. Scene classification using a hybrid generative/discriminative approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(4):712–727, 2008.

[17] Xin Chen, Xiaohua Hu, Zhongna Zhou, Caimei Lu, Gail Rosen, Tingting He, E. K. Park, and E. K. Park. A probabilistic topic-connection model for automatic image annotation. In *CIKM*, pages 899–908, 2010.

[18] Xin Chen, Caimei Lu, Yuan An, Palakorn Achananuparp, and Palakorn Achananuparp. Probabilistic models for topic learning from images and captions in online biomedical literatures. In *CIKM*, pages 495–504, 2009.

[19] Yizong Cheng. Mean shift, mode seeking, and clustering. pages 790–799, 1995.

[20] ed.) Christiane Fellbaum (1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

[21] Richter D. and Huson D. Functional metagenome analysis using gene ontology (megan 4). *Talk at the SIG M3 meeting (ISMB 2009)*, 2009.

[22] Ingrid Daubechies. Ten lectures on wavelets, society for industrial and applied mathematics. 1992.

[23] J. G. Daugman. Two dimensional spectral analysis of cortical receptive field profile. *Vision Research*, 20:847–856, 1980.

[24] Jia Deng, Alexander C. Berg, Kai Li, Fei-Fei Li 0002, and Fei-Fei Li 0002. What does classifying more than 10, 000 image categories tell us? In *ECCV (5)*, pages 71–84, 2010.

[25] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, Fei-Fei Li 0002, and Fei-Fei Li 0002. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[26] Thomas Deselaers, Vittorio Ferrari, and Vittorio Ferrari. Visual and semantic similarity in imagenet. In *CVPR*, pages 1777–1784, 2011.

[27] Hu F. Ehrlich G, Hiller NL. What makes pathogens pathogenic. *Genome Biol.*, page 9:225, 2008.

[28] Arumugam M et al. Enterotypes of the human gut microbiome. *Nature*, 472(7343), 2011.

[29] Ashburner M. et al. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet.*, 25(1):25–29, 2000.

[30] Caspi et al. The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic Acids Research*, 38:473–479, 2010.

[31] Manichanh C et al. Reduced diversity of faecalmicrobiota in crohn's disease revealed by a meta-genomic approach. *Gut.*, 55(2):205211, 2006.

[32] Qin J et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65, 2010.

[33] Walker A. et. al. J high-throughput clone library analysis of the mucosa-associated microbiota reveals dysbiosis and differences between inflamed and non-inflamed regions of the intestine in inflammatory bowel disease. *BMC Microbiology*, 11(7), 2011.

[34] Willenbrock et al. Characterization of probiotic escherichia coli isolates with a novel pan-genome microarray. *Genome Biology*, 8(12), 2007.

[35] Robert Fergus, Fei-Fei Li 0002, Pietro Perona, Andrew Zisserman, and Andrew Zisserman. Learning object categories from google's image search. In *ICCV*, pages 1816–1823, 2005.

[36] Patrick Flaherty, Guri Giaever, Jochen Kumm, Michael I. Jordan, Adam P. Arkin, and Adam P. Arkin. A latent variable model for chemogenomic profiling. pages 3286–3293, 2005.

[37] Per-Erik Forssn, David G. Lowe, and David G. Lowe. Shape descriptors for maximally stable extremal regions. In *ICCV*, pages 1–8, 2007.

[38] P. J. Moreno G. Carneiro, A. B. Chan and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(3):394–410, 2007.

[39] A. Holub G. Griffin and P. Perona. Caltech-256 object category dataset. Technical report, 2007.

[40] Dowell Robin D. Jaakkola Tommi S. Gifford David K. Gerber, Georg K. Hierarchical dirichlet process-based models for discovery of cross-species mammalian gene expression. Technical report, 2007.

[41] Matthieu Guillaumin, Thomas Mensink, Jakob J. Verbeek, Cordelia Schmid, and Cordelia Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *ICCV*, pages 309–316, 2009.

[42] et. al Harry S. Specificities of the fecal microbiota in inflammatory bowel disease. *Inflammatory Bowel Diseases*, 12:106111, 2006.

[43] J.M. Henderson and Hollingworth. High level scene perception. *Annual Review of Psychology*, 50:243271, 1999.

[44] T. Hofmann. Probabilistic latent semantic indexing. *Proceedings of the Twenty-Second Annual International SIGIR Conference*, 1999.

[45] Pocock M. Holland R., Down T. and Prlic A. Biojava: an open-source framework for bioinformatics. *Bioinformatics*, 24:397–2097, 2008.

[46] Jian Hu, Lujun Fang, Yang Cao, Hua-Jun Zeng, Hua Li, Qiang Yang, Zheng Chen, and Zheng Chen. Enhancing text clustering by leveraging wikipedia semantics. In *SIGIR*, pages 179–186, 2008.

[47] Xiaohua Hu, Xiaodan Zhang, Caimei Lu, E. K. Park, Xiaohua Zhou, and Xiaohua Zhou. Exploiting wikipedia as external knowledge for document clustering. In *KDD*, pages 389–396, 2009.

[48] Mitra S. Auch A. Schuster S. Huson D., Richter D. Methods for comparative metagenomics. *BMC Bioinformatics*, 10, 2009.

[49] Qi J. Schuster S. Huson D., Auch A. Megan analysis of metagenomic data. *Genome research 2007*, 2007.

[50] Sethuraman J. A constructive definition of dirichlet priors. *Statistica Sinica*, 4(2):639650, 1994.

[51] Yu-Gang Jiang, Chong-Wah Ngo, Jun Yang 0003, and Jun Yang 0003. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *CIVR*, pages 494–501, 2007.

[52] D. Forsyth N. de Freitas D. Blei and M. Jordan K. Barnard, P. Duygulu. Matching words and pictures. *JMLR*, 3:11071135, 2003.

[53] W. Nejdl K. Bischoff, C.S. Firan and R. Paiu. Can all tags be used for search? In *CIKM08*, pages 203–212, Napa Valley, California, USA, 2008.

[54] Timor Kadir, Michael Brady, and Michael Brady. Saliency, scale and image description. pages 83–105, 2001.

[55] Knight R. Knights D, Costello EK. Supervised classification of human microbiota. *FEMS Microbiol Rev.*, 35(2):343–359, 2011.

[56] Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, Tamara L. Berg, and Tamara L. Berg. Baby talk: Understanding and generating simple image descriptions. In *CVPR*, pages 1601–1608, 2011.

[57] Christoph H. Lampert, Hannes Nickisch, Stefan Harmeling, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, pages 951–958, 2009.

[58] Svetlana Lazebnik, Cordelia Schmid, Jean Ponce, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR (2)*, pages 2169–2178, 2006.

[59] Michael S. Lew, Nicu Sebe, Chabane Djeraba, Ramesh Jain, and Ramesh Jain. Content-based multimedia information retrieval: State of the art and challenges. pages 1–19, 2006.

[60] Jia Li, James Ze Wang, and James Ze Wang. Real-time computerized annotation of pictures. In *ACM Multimedia*, pages 911–920, 2006.

[61] Li-Jia Li, Richard Socher, Fei-Fei Li 0002, and Fei-Fei Li 0002. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *CVPR*, pages 2036–2043, 2009.

[62] Li-Jia Li, Chong Wang, Yongwhan Lim, David M. Blei, Fei-Fei Li 0002, and Fei-Fei Li 0002. Building and using a semantivisual image hierarchy. In *CVPR*, pages 3336–3343, 2010.

[63] David G. Lowe. Distinctive image features from scale-invariant keypoints. pages 91–110, 2004.

[64] Caimei Lu, Xiaohua Hu, Xin Chen, Jung ran Park, Tingting He, Zhoujun Li, and Zhoujun Li. The topic-perspective model for social tagging systems. In *KDD*, pages 683–692, 2010.

[65] V. Sindhwani M. Belkin, P. Niyogi. Manifold regularization: A geometric framework for learning for examples. *Technical Report*, 2004.

[66] C. K. I. Williams J. Winn M. Everingham, L. Van Gool and A. Zisserman. The pascal visual object classes challenge 2008 (voc2008) results. *http://www.pascal-network.org/ challenges/VOC/voc2008/workshop/*, 2008.

[67] Jiri Matas, Ondrej Chum, Martin Urban, Toms Pajdla, and Toms Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *BMVC*, 2002.

[68] D. Medini. The microbial pan-genome. *Current Opinion in Genetics and Development*, 15(6):589–594, 2005.

[69] Krystian Mikolajczyk, Cordelia Schmid, and Cordelia Schmid. A performance evaluation of local descriptors. In *CVPR (2)*, pages 257–263, 2003.

[70] Krystian Mikolajczyk, Cordelia Schmid, and Cordelia Schmid. Scale and affine invariant interest point detectors. pages 63–86, 2004.

[71] Krystian Mikolajczyk, Cordelia Schmid, and Cordelia Schmid. A performance evaluation of local descriptors. pages 1615–1630, 2005.

[72] Krystian Mikolajczyk, Tinne Tuytelaars, Cordelia Schmid, Andrew Zisserman, Jiri Matas, Frederik Schaffalitzky, Timor Kadir, Luc J. Van Gool, and Luc J. Van Gool. A comparison of affine region detectors. pages 43–72, 2005.

[73] T. P. Minka. The dirichlet-tree distribution. *http://research.microsoft.com/ minka/papers/dirichlet/minkadirtree.pdf*, 1999.

[74] T. P. Minka. Estimating a dirichlet distribution. *http://research. Microsoft.com/en-us/um/people/minka/pap- ers/dirichlet*, 2009.

[75] Julien P Letunic I Roth A Kuhn M Powell S von Mering C Doerks T Jensen LJ Bork P Muller J, Szklarczyk D. eggnog v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res 2010*, 38:190–195, 2010.

[76] Ramesh Nallapati, Amr Ahmed, Eric P. Xing, William W. Cohen, and William W. Cohen. Joint latent topic models for text and citations. In *KDD*, pages 542–550, 2008.

[77] David Newman, Chaitanya Chemudugunta, Padhraic Smyth, and Padhraic Smyth. Statistical entity-topic models. In *KDD*, pages 680–686, 2006.

[78] Cam-Tu Nguyen, Natsuda Kaothanthong, Xuan Hieu Phan, Takeshi Tokuyama, and Takeshi Tokuyama. A feature-word-topic model for image annotation. In *CIKM*, pages 1481–1484, 2010.

[79] V. Honavar O. Yakhnenko. Annotating images and image objects using a hierarchical dirichlet process model. *proceedings of the 9th International Workshop on Multimedia Data Mining*, pages 1–7, 2008.

[80] Aude Oliva, Antonio Torralba, and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. pages 145–175, 2001.

[81] J. Platt. Probabilistic outputs for support vector machines and comparison to regularize likelihood methods. *Advances in Large Margin Classifiers*, page 6174, 2000.

[82] Daniel Ramage, Paul Heymann, Christopher D. Manning, Hector Garcia-Molina, and Hector Garcia-Molina. Clustering the tagged web. In *WSDM*, pages 54–63, 2009.

[83] Marcus Rohrbach, Michael Stark, Gyrgy Szarvas, Iryna Gurevych, Bernt Schiele, and Bernt Schiele. What helps where - and why? semantic relatedness for knowledge transfer. In *CVPR*, pages 910–917, 2010.

[84] Gail L. Rosen, Elaine Garbarine, Diamantino Caseiro, Robi Polikar, Bahrad A. Sokhansanj, and Bahrad A. Sokhansanj. Metagenome fragment classification using n-mer frequency profiles. 2008.

[85] Polikar R. Bruns M. A. Russell J. Garbarine E. Essinger S. Rosen G., Sokhansanj B. and Yok N. Signal processing for metagenomics: Extracting information from the soup. *Current Genomics*, 2009.

[86] O. Russakovsky and L. Fei-Fei. Attribute learning in large-scale datasets. *Proceedings of the 12th European Conference of Computer Vision (ECCV)*, 2010.

[87] Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, William T. Freeman, and William T. Freeman. Labelme: A database and web-based tool for image annotation. pages 157–173, 2008.

[88] Shilad Sen, Shyong K. Lam, Al Mamunur Rashid, Dan Cosley, Dan Frankowski, Jeremy Osterhouse, F. Maxwell Harper, John Riedl, and John Riedl. tagging, communities, vocabulary, evolution. In *CSCW*, pages 181–190, 2006.

[89] Heng Tao Shen, Beng Chin Ooi, Kian-Lee Tan, and Kian-Lee Tan. Giving meanings to www images. In *ACM Multimedia*, pages 39–47, 2000.

[90] Christian Siagian, Laurent Itti, and Laurent Itti. Rapid biologically-inspired scene classification using features shared with visual attention. pages 300–312, 2007.

[91] Josef Sivic, Andrew Zisserman, and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, pages 1470–1477, 2003.

[92] F. Smadja. Retrieving collections from text: Xtract. *Computational Linguistics*, 19(1):143–177, 1993.

[93] Erik B. Sudderth, Antonio Torralba, William T. Freeman, Alan S. Willsky, and Alan S. Willsky. Learning hierarchical models of scenes, objects, and parts. In *ICCV*, pages 1331–1338, 2005.

[94] Erik B. Sudderth, Antonio Torralba, William T. Freeman, Alan S. Willsky, and Alan S. Willsky. Describing visual scenes using transformed objects and parts. pages 291–330, 2008.

[95] M. Steyvers T. L. Griffiths. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101:5228–5235, 2004.

[96] Gerald W. Tannock. The bowel microbiota and inflammatory bowel diseases. *International Journal of Inflammation*, 2010, 2010.

[97] Gadia V. and Rosen G. A text-mining approach for classification of genomic fragments. In *IEEE International Workshop on Biomedical and Health Informatics*, Philadelphia, PA, 2008.

[98] Joost van de Weijer, Cordelia Schmid, and Cordelia Schmid. Coloring local feature extraction. In *ECCV (2)*, pages 334–348, 2006.

[99] C.J. Van Rijsbergen. Information retrieval. *Butterworths*, 1975.

[100] R. Wang W. W. Cohen and R. F. Murphy. Understanding captions in biological publications. In *ACM KDD*, 2005.

[101] Changhu Wang, Lei Zhang 0001, Hong-Jiang Zhang, and Hong-Jiang Zhang. Learning to reduce the semantic gap in web image retrieval and annotation. In *SIGIR*, pages 355–362, 2008.

[102] Huan Wang, Xing Jiang, Liang-Tien Chia, Ah-Hwee Tan, and Ah-Hwee Tan. Ontology enhanced web image retrieval: aided by wikipedia and spreading activation theory. In *Multimedia Information Retrieval*, pages 195–201, 2008.

[103] Pu Wang, Carlotta Domeniconi, and Carlotta Domeniconi. Building semantic kernels for text classification using wikipedia. In *KDD*, pages 713–721, 2008.

[104] Xiaogang Wang, Eric Grimson, and Eric Grimson. Spatial latent dirichlet allocation. In *NIPS*, 2007.

[105] Xiaoyu Wang, Tony X. Han, Shuicheng Yan, and Shuicheng Yan. An hog-lbp human detector with partial occlusion handling. In *ICCV*, pages 32–39, 2009.

[106] Zhong Wu, Qifa Ke, Michael Isard, Jian Sun 0001, and Jian Sun 0001. Bundling features for large scale partial-duplicate web image search. In *CVPR*, pages 25–32, 2009.

[107] XiajiongShen Xin Chen, Xiaohua Hu. Spatial weighting for bag-of-visual-words representation and its application in content-based image retrieval. In *the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'09)*, pages 867–874, Bangkok, Thailand, 2009.

[108] M. Beal Y. Teh, M. Jordan and D. Blei. Hierarchical dirichlet process. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[109] Jianguo Zhang, Marcin Marszalek, Svetlana Lazebnik, Cordelia Schmid, and Cordelia Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. pages 213–238, 2007.

[110] Bin Zheng, David C. McLean Jr., Xinghua Lu, and Xinghua Lu. Identifying biological concepts from a protein-related corpus with a probabilistic topic model. page 58, 2006.

[111] Ding Zhou, Jiang Bian, Shuyi Zheng, Hongyuan Zha, C. Lee Giles, and C. Lee Giles. Exploring social annotations for information retrieval. In *WWW*, pages 715–724, 2008.

[112] Xiaohua Zhou, Xiaodan Zhang, Xiaohua Hu, and Xiaohua Hu. Maxmatcher: Biological concept extraction using approximate dictionary lookup. In *PRICAI*, pages 1145–1149, 2006.

Vita

Xin Chen received both B.Eng. and M.Eng. degrees in Electronic Engineering and Information Science from the University of Science and Technology of China, Hefei, China in year 2004 and 2007, respectively. From Sept. 2007, he studied his Ph.D. degree at the College of Information Science and Technology, Drexel University, Philadelphia, PA, under the advice of Prof. Tony Xiaohua Hu.

During his five-year career as a doctoral student and a research assistant in information science, he has accumulated extensive knowledge and hands-on experience in the areas of data mining, machine learning and information retrieval. Particularly, he has been interested in studying user interests and preferences from collaboratively created social annotations (such as social tags from del.icio.us and flickr.com created by online users). As a researcher, besides applying state-of-the-art techniques and methods, he is also dedicated to developing innovative systems, methods and algorithms for data mining.

He also has expertise in relational data modeling and database design. he has completed two summer internship on database and software development in Merck Research Laboratory, in which he sharpened his skill to translate business requirements into effective data models and design relational databases accordingly.

He has published and presented his research work in some top conferences in the field of data mining and knowledge management, such as ACM CIKM 2011, CIKM 2010, CIKM 2009 and ACM SIG KDD 2010. He got Student Travel Award for CIKM 2010 and CIKM 2009.