Search disambiguation techniques in multimedia collections

Wan, Kong Wah

2012

Wan, K. W. (2012). Search disambiguation techniques in multimedia collections. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/50768>

<https://doi.org/10.32657/10356/50768>

SEARCH DISAMBIGUATION TECHNIQUES IN MULTIMEDIA COLLECTIONS

WAN KONG WAH

SCHOOL OF COMPUTER ENGINEERING

A THESIS SUBMITTED TO THE NANYANG TECHNOLOGICAL UNIVERSITY
IN PARTIAL FULFILLMENT FOR THE REQUIREMENT FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

2012

Acknowledgements

Writing the acknowledgement page is the easiest part of this report! My first thanks must go to my employer, the Institute for Infocomm Research (I²R), for graciously sponsoring me to pursue this PhD degree on a part-time basis. Over the course of this work, I had been fortunate to have managers who are not only supportive of my study, but who have also kindly adjusted their overall project and resource schedule in order to factor in my inevitable absences of leave and time off. I want to make special mention of two of these managers: Dr YAU Wei Yun (IPTV program manager) and Dr LIM Joo Hwee (CVIU Department Head). Thank you!

Next, I want to acknowledge my appreciation for the support and technical guidance from my three supervisors. Even though I had been in the R&D field (with I²R) for over ten years, most of my previous work had been on technology application and transfer to industry. Enrolling in this PhD program is truly my first in engaging in cutting edge research, a baptism of fire of sort. I want to thank Assoc Prof TAN for his kind patience and gentle prodding along the way, Assoc Prof LIM for his numerous tips on what angle I can take to extend the state-of-art, and Assoc Prof CHIA for guiding me to think critically on various research issues.

No cutting edge research work can be accomplished by solo individuals. I had the fortune of having numerous collaborators, who not only provide additional insights and critical comments on various research issues, but also provide the data-set upon which ideas can be validated, refined and thoroughly tested. I remember Dr XU Changsheng, Dr LI Hai Zhou, Prof TIAN Qi, ZHENG Yantao, Dr DUAN Lingyu, Dr NGO Chong Wah, Dr XIAO Wu, Dr Sujoy ROY and Dr YAP Ghim Eng.

Last but not least, I want to thank my wonderful wife, Esther, and my two lovely boys, Zheng Yu and Zheng Yang, for their quiet support during these five years of study. This is not a trivial accolade, because in coping with both my work and study, the burden of helping the two boys falls upon none other but my dear wife. I dedicate this work to them.

Contents

List of Tables	xiii
List of Figures	xiv
List of Abbreviations	xx
1 INTRODUCTION	1
1.1 Background and Motivation	2
1.1.1 Architecture of a keyword-based Multimedia Retrieval System	2
1.1.2 Two Main Types of Dissatisfactory Retrieval	3
1.1.3 Contemporary Approach to Ameliorate the Two Dissatisfactory Retrieval	4
1.2 Research Issues	6
1.2.1 The Limitations of Current Query Refinement Methods	7
1.2.2 Issues and Challenges	8
1.2.3 Thesis Objectives	9
1.3 Approach and Methodology	12
1.3.1 A Faceted Topic Retrieval Framework for Multimedia Content	13
1.3.2 Non-Parametric Modeling in the Faceted Document Model	15
1.3.3 Multimodal Topic Analysis in the Faceted Document Model	17
1.3.4 Topic-based Query Refinement	19
1.4 Thesis Contributions	21
1.5 Outline of Thesis	22
2 LITERATURE REVIEW	25
2.1 Faceted Retrieval	26

2.2	Faceted Modeling	27
2.2.1	Conventional Modeling	28
2.2.2	Semantic Models	29
2.2.3	Modeling Facets	31
2.3	Probabilistic Topic Models	34
2.3.1	Generative Model	34
2.3.2	Topic Modeling of Text	36
2.3.3	Topic Modeling of Image/Video	36
2.3.4	Other Applications of Topic Models	38
2.3.5	Learning a Topic Model	39
2.3.6	Constraining Topic Formation	41
2.4	Bayesian Non-Parametrics	43
2.4.1	Finite Mixture Modeling	44
2.4.2	Dirichlet Process Mixture Models	45
2.4.3	Dirichlet Process	46
2.4.4	Hierarchical Dirichlet Process	49
2.5	Search Disambiguation	50
2.5.1	Mining Query Intent	50
2.5.2	Search Diversification	51
2.5.3	Sample Diversification in Active Learning	52
2.5.4	Sense Disambiguation	53
2.5.5	Document and Query Expansion	54
3	A FACETED TOPIC RETRIEVAL MODEL FOR MULTIMEDIA CONTENT	55
3.1	Introduction	55
3.2	Preliminaries and Definitions	56
3.3	Problem Definition	57
3.4	Evaluating Faceted Tropic Retrieval	57
3.4.1	S-recall	58
3.4.2	S-precision	58

3.5	A Probabilistic Framework for Faceted Topic Retrieval	59
3.5.1	Faceted Retrieval as an Optimization Problem	59
3.6	Faceted Document Model	61
3.6.1	A Probabilistic Topic Modeling Approach	61
3.6.2	Multi-modal Modeling of Complementary Data Streams	62
3.6.3	Modeling Data Complexity from Query Ambiguity	67
3.7	Faceted Retrieval Model	75
3.7.1	A Conventional Greedy Approach	75
3.7.2	Our Set-based Probabilistic Approach	76
3.8	Summary	77
4	FACETED CLUSTERING OF POLYSEMOUS IMAGES	79
4.1	Motivation	80
4.2	A Bayesian Approach	81
4.2.1	HDP-based Faceted Clustering of Polysemous Images	82
4.2.2	Augmenting with Wiki-Sense-Disambiguated Images	84
4.2.3	Presenting Image Clusters in their Wiki-Primary Senses	86
4.2.4	Algorithm	88
4.3	Experimental Results	88
4.3.1	Methodology and Dataset	88
4.3.2	Visual Features	91
4.3.3	Baselines	92
4.3.4	Evaluation	93
4.3.5	Analysis and Discussion	94
5	FACETED RETRIEVAL OF NEWS VIDEO	101
5.1	Motivation	102
5.2	A Faceted Document Model of News Video	103
5.2.1	Text features	103
5.2.2	Visual features	104
5.2.3	Joint Modeling of Visual and Text	105

5.2.4	Inference and Estimation	107
5.2.5	Re-ranking	109
5.3	Experimental Evaluation	110
5.3.1	Dataset	110
5.3.2	Comparing Retrieval Engines	110
5.3.3	Results and Analysis	112
5.4	Discussions	114
6	TOPIC BASED QUERY SUGGESTIONS	117
6.1	Motivation	117
6.2	Modeling Challenges in the Video Domain	119
6.2.1	Noisy Textual Content	119
6.2.2	Joint Modeling of Multimodal Content Cues	119
6.2.3	Labeling Query Suggestions	121
6.3	Existing Works	121
6.4	Topic Based Query Suggestions	122
6.4.1	Joint Modeling	123
6.4.2	Selecting Topics and Terms for Query Suggestions	130
6.5	Experiments	131
6.5.1	Perplexity Comparison	131
6.5.2	Qualitative Inspection	133
6.5.3	An Objective Measure of Topic Coherence	133
6.5.4	User Study	135
6.6	Discussion	136
7	CONCLUSIONS	137
7.1	Summary of Thesis	137
7.2	Contributions	139
7.3	Future Works	140

A List of Publications **143**

A.1 Journal Papers 143

A.2 Conference Papers 143

B VARIATIONAL INFERENCE OF JOINT MULTINOMIAL CORRESPONDENCE

MODEL **145**

B.1 The Joint Multinomial Correspondence Model 145

B.2 Principles of Variational Methods 145

B.3 Variational Inference 149

 B.3.1 Variational Simplification 149

 B.3.2 Variational Lower Bound 150

 B.3.3 Factor Expansion 151

 B.3.4 Parameter Estimation 154

 B.3.5 Variational Expectation Maximization 157

Abstract

In the last few years, we have witnessed an explosive growth in multimedia content. Online repositories such as Flickr (image) and YouTube (video) contain hundreds of millions of images and videos and are still growing by the day. However, the utility of these data sources is only as good as their accessibility. A primary motivation of this thesis is towards better utilization of this overwhelming wealth of information. Specifically, we look at ways to disambiguate search in a multimedia retrieval system. The need for search disambiguation often arises in practice because the typical issued query is short, imprecise and ambiguous. and it is difficult to infer the exact query intent. Search disambiguation is the general class of techniques to elucidate the multi-faceted information needs arising from an ambiguous query. The objective is to provide an overview of search results grouped by multiple facets, which can better clarify a user’s search intention by enabling him to zoom into any specific facet of the query topic.

There has been a lack of a principled approach to search disambiguation. Many multimedia retrieval systems address the issue of ambiguous queries by results clustering, near-duplicate removal and diversification. The aim is to simply prevent result pages from being cluttered by too many similar Web articles. In this thesis, we take a more principled approach and propose the following two-prong methodology.

First, we propose a novel faceted topic retrieval framework wherein multimedia documents are modeled as comprising of facets or topics. The notion of facets can be interpreted broadly to encompass any binary property of a document that represents a fact or a topic that is contained in the query need. In contrast with traditional term-based retrieval models (e.g. TF-IDF), documents are now ranked by the relevance of their composite facets/topics to the query. The goal is then to return a set of documents that are not only relevant to the query, they also cover the many different facets of the information need. We base our faceted retrieval framework on probabilistic topic models, a class of algorithms designed to discover the latent thematic structures in a document collection.

Second, we augment our faceted retrieval framework with two modeling capabilities to better process ambiguous queries in multimedia collection such as video and images. Firstly, because different queries may have varying level of ambiguity, and hence varying level of polysemy in the return results, we develop a non-parametric Bayesian method to cluster search results. The main

advantage of our clustering method is that the number of mixture components is not fixed a priori, but is determined during the posterior inference process. This allows our model to grow with the level of polysemy (and visual diversity) in the return multimedia results. Secondly, we extend the basic probabilistic topic model (the Latent Dirichlet Allocation, LDA) to jointly model the complementary information in the visual and textual streams. We show how the joint modeling can improve the quality of facet detection in news video, and in turn yield better user satisfaction in faceted topic retrieval.

We demonstrate the utility of our two-prong approach on several large-scale standard multimedia datasets. The empirical results obtained from both subjective and objective evaluation show the viability of our faceted approach in describing content. In subjective testing experiments, we apply our faceted topic approach to compute query suggestions that achieve a higher level of user acceptance than a conventional method based on lexicographical matching. In objective testing, we find that the return documents using our retrieval methods show a better faceted topical recall and precision performance over conventional method such as the Maximal Marginal Relevance. As a clustering engine, our methods have also shown better capability in handling ambiguous queries. Specifically, when presented with queries of different level of ambiguity, our methods are better able to cluster the polysemous images into their visual senses, and achieve better clustering performance metric over existing solutions such as spectral clustering and SVM-based classifiers. The faceted topic retrieval framework and the disambiguation techniques described in this thesis have opened up doors for a new way of searching for multimedia content.

List of Tables

1.1	Top-10 probability words of 2 learned β_k topics, for “War on Fallujah” (top) and “Mideast peace” (bottom). The topics are picked introspectively. Meaningful words highlighted in bold.	19
3.1	Notations used to describe the inference process. Boldfaced font denotes array or matrix.	66
4.1	Keywords and their Wiki-senses used in our experiments	86
4.2	Area Under Curve (AUC) of all senses of each keyword	94
5.1	Sample expansion words given the words on the left.	104
5.2	FTR performance on TRECVID-2005 dataset. The best result in each FDM setting is in bold. An asterisk indicates significant statistical difference over the MMR baseline.	112
5.3	Top-10 probability words of 2 learned FDM β_k topics, for “War on Fallujah” (top) and “Mideast peace” (bottom). The topics are picked introspectively. Meaningful words highlighted in bold.	113
6.1	Sample topic-based query suggestions. The first line is the original query.	133
6.2	Top-10 probability words of two learned cLDA-VT θ_k topics, for “War on Fallujah” (top) and “Mideast peace” (bottom). Meaningful words introspectively highlighted in bold.	133
6.3	A sample of topics with their high topic coherence scores.	134
6.4	Spearman rank correlation values for the learned topics from standard LDA, cLDA-T and cLDA-VT. Note that for the standard LDA model, only the ASR words are used.	135
6.5	User preferences for query suggestions. See text for details.	136

List of Figures

1.1	Basic architecture of an multimedia information retrieval system.	3
1.2	A multimedia information retrieval system augmented with query refinement strategies such as query expansion, query suggestion and results clustering.	5
1.3	Our overall approach and contributions highlighted in purple.	12
1.4	Using the average image to illustrate the visual polysemy of images. Top: for images retrieved using the query “mouse computing”, they are very diverse. Bottom: average images of three HDP generated clusters on the “mouse computing” images. Note the fine-grained elucidation of “notebook”, “pointer” and “desktop” images.	16
1.5	Example clusters on two polysemous keywords: “Tiger” and “Watch”. For each pol- ysemous keyword, we show enclosed in a green box two rows of sample polysemous images returned by Google Image Search. Below the green box, we show the auto- matic clusters of images enclosed in red boxes. For “tiger”, the clusters correspond to “bengal tiger”, “tigger pooh”, “tiger tanks”, “tiger sharks”, “white tiger” and “tiger woods”. For “watch”, the clusters correspond to “clock”, “watch tower” and “wrist watch”.	17
1.6	Our proposed faceted document model for hypothesizing facets in news video. The visual and textual features in news video are jointly modeled to yield facets that better correlate with the actual news events.	18

1.7	An example of our Topic-based query disambiguating search [167]. Left: the normal simple word prefix based query suggestions as the user type into the query box. Right: as the user selects a suggested name, the system further computes subtopics that are semantically related with the name.	20
2.1	Graphical representation of LDA. Shaded node w denotes the observed random variable for the occurrence of a word (visual word). z is the topic index for all words in a document: z_d denotes the topic assignment variable for document d , where $z_{d,n}$ is the topic assignment for the n th word in document d . ϕ_d is the topic mixture variable for document d , $\phi_{d,k}$ is the topic proportion for the topic k in document d . D denotes the number of document (images) in the collection, N_d the number of (visual) words in the document (image), and K the number of topics. α and β are the Dirichlet hyper-parameters controlling the document-topic distribution and the topic-word distribution respectively.	35
2.2	Results of latent modeling of two text documents into a mixture of 100 topics. We show only the three relevant ones (topic number 28, 47, 16, each colored differently). Selected top probability words from each topic are shown. The words in the text document are color-coded to indicate their topic affinity. For the word “court”, its meaning is disambiguated correctly in each text. Best viewed in color.	37
2.3	Results of latent modeling of images using a mixture of 50 topics. Images are described by visual words, densely sampled from overlapping grids over the images. Topics are multinomial distribution of these visual words. We show two images, on which visual words from two topics (colored blue and green) are superimposed. We can see that the green topics describe people while blue topics describe the sky regions. Best viewed in color.	38
2.4	Graphical representation of the Link-PLSA-LDA model that jointly models documents and citations [115].	41
2.5	Graphical representation of the Corr-LDA [25].	42

2.6	Left: Graphical model of the finite mixture model (G is a Gaussian); Right: Graphical model of the Dirichlet process mixture model. Note the addition of a nonparametric Dirichlet process prior on G , with parameters G_0 and concentration parameter α	44
2.7	Figure taken from [63]: Density plots (blue = low, red = high) for the Dirichlet distribution over the probability simplex in $\mathbb{R}^3$ for various values of the parameter α . When $\alpha = [c, c, c]$ for some $c > 0$, the density is symmetric about the uniform pmf (top-left). When $0 < c < 1$, there are sharp peaks of density almost at the vertices of the simplex and the density is very small away from the vertices. The top-right plot shows an example of this case for $\alpha = [0.1, 0.1, 0.1]$, one sees only blue (low density) because all of the density is converged near the edges of the simplex. When $c > 1$, the density becomes concentrated in the center of the simplex (bottom-left). Finally, if α is not a constant vector, the density is not symmetric (bottom-right).	47
3.1	Left: Graphical representation of our joint model. Note that our joint model is different from the Corr-LDA of Blei <i>et al</i> [25] shown in Figure 2.5: we model the generation of <i>both</i> text-words w and visual-words v by the multinomial $P(w z, \theta)$ and $P(v y, \gamma)$ respectively; while in the Corr-LDA, the visual-words are generated by a Gaussian $P(v z, \mu, \delta)$. See text for explanation. Right: graphical representation of the variational distribution used to simplify the graphical model on the left. The factorized distribution assumes the independence amongst ϕ, z and y	65
3.2	Top: Top-ranked “apple” images from Google Image; Bottom: six possible visual senses (clusters): <i>logo, fruit, iphone, iMac, drawings, people</i>	68
3.3	The framework of our approach. Given the ambiguous query “mouse”, a set of polysemous images is retrieved. Their latent visual senses are elucidated by HDP, helped by a parallel injection of images of suggested semantic senses by Wikipedia.	69

3.4	Chinese Restaurant Franchise (CRF) for three images with 8 image feature points. Top shows the global restaurant serving dishes shared by all tables in the three local restaurants shown below. The generative process in each restaurant is a Chinese Restaurant Process (CRP). For example, in the 1st local restaurant, the 1st, 3rd, 6th and 8th customer were the first to sit at an empty table, while the others sat at tables already occupied. For this restaurant, the 9th customer will sit at table 1, 2, 3, 4 with probabilities $\frac{3}{8+\alpha_0}, \frac{2}{8+\alpha_0}, \frac{2}{8+\alpha_0}, \frac{1}{8+\alpha_0}$ respectively, or will sit at a new (5th) table with probability $\frac{\alpha_0}{8+\alpha_0}$. Similarly, in the global restaurant, each customer (j, i) sitting at a table corresponds to table i in the local restaurant j . The table in the global restaurant that (j, i) chooses to sit at, determines the dish that is served at table i in the local restaurant j . In this example, for any new customer entering into a local restaurant j , if he sits down at a new table, the dish for that new table will be ψ_1, ψ_2, ψ_3 or ψ_4 with probability $\frac{4}{11+\gamma}, \frac{4}{11+\gamma}, \frac{2}{11+\gamma}$, or $\frac{1}{11+\gamma}$ respectively, or a new dish ψ_5 with probability $\frac{\gamma}{11+\gamma}$	72
4.1	The framework of our approach. Given the ambiguous query “mouse”, a set of polysemous images is retrieved. Their latent visual senses are elucidated by HDP, helped by a parallel injection of images of suggested semantic senses by Wikipedia.	82
4.2	ROC plots of the first primary sense of the ten polysemous keywords	95
4.3	Example visual sense clusters on four polysemous keywords: “Tiger”, “Watch”, “Bass” and “Mouse”. For each polysemous keyword, we show enclosed in a green box two rows of sample polysemous images returned by Google Image Search. Note that because the first Google page of top rank images usually depict the dominant sense of the query keyword, they are fairly homogeneous. Hence, for our illustrative purpose to show sample polysemous images, we randomly take images that Google ranks from 50 onwards. Below the green box, we show the automatic clusters of images enclosed in red boxes.	96

4.4	Visual senses of the ten polysemous keywords. For each keyword, five visual senses are shown. Each visual sense represents a cluster of the images retrieved from the polysemous keyword. From each cluster, we show two images which are visually informative. As can be seen, the primary semantic senses of the polysemous word are captured by the cluster images.	97
4.5	Google search results on a Wiki-disambiguated query “Mouse computing”. While results are more homogeneous than that of “Mouse”, polysemy clearly remains an issue.	97
4.6	Using the average image to illustrate the visual polysemy of images. Top: even for images retrieved using the Wiki-disambiguated query “Mouse computing”, they are very diverse. This will be a challenge to a supervised learning method. Bottom: average images of three HDP generated clusters on “Mouse computing” images. Note the fine-grained elucidation of “notebook”, “pointer” and “desktop” images.	98
4.7	Comparative AUC performance across various values of the number of mixture components.	98
4.8	Including more sense-specific images improve AUC performance of VS-HDP.	100
5.1	FDM for hypothesizing facets in news video.	103
5.2	TFIDF weighting of keyframe-based visual words. Within the 4 videos, four NDI clusters are shown and colored differently. For the red NDI cluster, it appears in V_2, V_3, V_4 , hence its $df=3$, $tf=2$ for V_3 , $tf=1$ for V_2, V_4 . All non-NDI in a video have $df=1$ and $tf=1$. Best viewed in color.	104
5.3	Top shows the spatial division of the image at each level of the RGB pyramid. At each level, the RGB suffix refers to the number of bits in the color channel. E.g. at level-1, we have 4 divisions each with 3 bits red, 4 bits green, 3 bits blue, totaling 128 bins. Each bin count uses 2 bytes. Therefore, each image is represented by 768 bytes. Bottom shows sample NDI given the leftmost image as query.	105
5.4	Graphical representation of FDM-VT. The red box encloses the additional observations from the visual modality, which constrain the topic formation in the text modality.	106
5.5	Variational Simplification for learning of FDM-VT.	108

5.6	S-recall versus Precision plot for three FDM models and MMR baseline. Best viewed in color.	113
6.1	Left: Existing query suggestions; Right: Presenting suggestions by thematic topics: three topics relevant to the query “fallujah” are shown, each titled and underscored by the highest probability word in the discovered topic.	118
6.2	Joint topic modeling framework for hypothesizing topics in news video.	123
6.3	Top: standard LDA model. Middle: Joint Text model (cLDA-T) Right: Joint Visual+Text model (cLDA-VT)	124
6.4	Variational Simplification for learning cLDA-VT.	128
6.5	Predictive Perplexity plot of LDA, cLDA-T and cLDA-VT on a held-out set from the TRECVID-2005 video dataset. Lower is better.	132
B.1	A concave $f(x)$ and the linear bound $\lambda^T x - f^*(\lambda)$	146
B.2	Left: Graphical model of the LDA; Right: Graphical model of the variational distribution used to approximate the posterior in LDA.	149

List of Abbreviations

ASR	Automatic Speech Recognition
AUC	Area Under Curve
BNP	Bayesian Non-Parametric
BOVW	Bag Of Visual Words
BOW	Bag Of Words
cLDA-T	combined Latent Dirichlet Allocation - Text only model
cLDA-VT	combined Latent Dirichlet Allocation - Visual and Text model
Corr-LDA	Correspondence Latent Dirichlet Allocation
CRF	Chinese Restaurant Franchise
CRP	Chinese Restaurant Process
DoG	Difference of Gaussian
DP	Dirichlet process
DPMM	Dirichlet Process Mixture Model
EM	Expectation Maximization
FDM	Faceted Document Model
FDM	Faceted Document Model
FDM-T	Faceted Document Model – Text only
FDM-VT	Faceted Document Model – Visual and Text
FRM	Faceted Retrieval Model
FTR	Faceted Topic Retrieval
FTRM	Faceted Topic Retrieval Model
HDP	Hierarchical Dirichlet Process
IR	Information Retrieval
KL Divergence	Kullback-Leibler Divergence
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
MCMC	Markov Chain Monte Carlo
MI	Mutual Information

MMR	Maximal Marginal Relevance
MSER	Maximally Stable Extremal Regions
NDI	Near Duplicate Image
OCR	Optical Character Recognition
ODP	The Open Directory Project
PLSA	Probabilistic Latent Semantic Analysis
PMI	Pointwise Mutual Information
PRF	Pseudo Relevance Feedback
ROC	Receiver Operating Characteristic
SIFT	Scale-Invariant Feature Transform
SN	Semantic Network
SPEC-CLUST	Spectral Clustering
SS	Semantic Space
SS-SVM	Sense-Specific Support Vector Machine
SVD	Singular Value Decomposition
SVM	Support Vector Machine
TCS	Topic Coherence Score
TF-IDF	Term Frequency - Inverse Document Frequency
TM	Topic Models
TREC	Text REtrieval Conference
TRECVID	TREC Video Retrieval Evaluation
VS-HDP	Visual Sense-Hierarchical Dirichlet Process
VS-LDA	Visual Sense-Latent Dirichlet Allocation
VSM	Vector Space Model
WSD	Word Sense Disambiguation

Chapter 1

INTRODUCTION

Search disambiguation refers to the general suite of techniques for handling ambiguous queries. The need for search disambiguation is often attributed to the lack of clarity of the query intent, as a direct consequence to the low number of query words typed on average by Web users. Many search engines such as Google and Bing have already started to tackle the problem of search disambiguation. The typical strategy to handle ambiguous queries is results clustering and diversification, so as to prevent result pages from being cluttered by too many similar Web articles.

In this thesis, we focus on the problem of search disambiguation in multimedia collections. While the same idea of result clustering is applicable to the multimedia domain, we show that some fundamental techniques need to be introduced to produce good results. Specifically, result clustering must be adaptive to the varying level of inherent ambiguity in the query, and can grow with the complexity of the data.

Existing multimedia search systems mostly adopt the term-based indexing paradigm that is conveniently “inherited” from their counterpart in textual search [164, 96, 66, 61, 141, 135, 168, 169]. We further argue that such a term-based indexing paradigm is not suitable for search disambiguation. Instead, we propose a novel *faceted topic retrieval* framework for multimedia content. Under this framework, we go beyond term-based indexing to one that is based on *latent topics* describing the content. We also develop joint topic modeling to combine the textual stream in a multimedia content with the complementary information in the visual stream. This multimodal combination can yield more robust topics for faceted retrieval, leading a search engine that can better meet the information needs of the users, and ultimately a better search experience.

1.1 Background and Motivation

Following the explosive growth of the World Wide Web, finding useful information from such a large data repository becomes a critical necessity. In the past three decades, the focus of the development of information retrieval has been on relevance ranking and term weighting [142, 132, 138], where various techniques based on some form of *TF-IDF weighting*¹ have shown compelling ranked retrieval results on standard TREC benchmarks. These success stories have naturally inspired the development of search methodology on multimedia collections. The conventional multimedia search algorithms are predominantly a port of their relatively successful counterpart in the textual Web [96, 66, 61, 141, 135, 168, 169]. These methods assume that the salient audio/visual content can first be adequately expressed in a textual form, (e.g. automatic speech transcript, or image-OCR), upon which efficient term-based indexing techniques can then be applied.

While this approach has seen moderate success in retrieving multimedia content [144, 102, 96, 61, 141, 135], the proliferation of multimedia content on the Internet has exacerbated the various problems with this simple approach. One such problem relates to the issue of search ambiguity.

1.1.1 Architecture of a keyword-based Multimedia Retrieval System

To better understand the issues relating to search ambiguity, and our contributions in resolving them, we first review the basic processes of a keyword-based multimedia retrieval system in Figure 1.1.

The user first attempts to articulate his information need by typing in some keywords. However, because of the limited expressiveness of natural language, these few keywords may not adequately express his original query intent. Coupled with the fact that the conventional types of low level features are ill-suited to represent multimedia content at the same semantic level as the user query [144], the return results from the initial typed query are often not satisfactory. We discuss the two main types of dissatisfactory retrieval in the next section.

¹a term weighting scheme factored on the frequency of occurrence of the term in the corpus (TF: term frequency) and its novelty (IDF: inverse document frequency)

Figure 1.1: Basic architecture of an multimedia information retrieval system.

1.1.2 Two Main Types of Dissatisfactory Retrieval

1.1.2.1 Too Few Relevant Results

The first type of dissatisfactory retrieval relates to the situation when there are too few relevant results. The main reason can be two-fold: (a) low recall, meaning that the desired results are not ranked and presented at the first few pages; (b) low precision, meaning that those top-ranked results are not relevant. In both cases, the main issue is the *term mismatch* problem, wherein a relevant document will fail to be retrieved if it does not share at least one common term with the query [52, 164].

1.1.2.2 Too Many Similar Results

The second type of dissatisfactory retrieval relates to the situation when the top retrieved results are too similar to one another. For example, a recent search on YouTube using the query “larry page” returns mostly videos of his seminar speeches. Interesting snippets of his recent activities, e.g. video

commentary about his CEO role in Google, rivalry between him and Facebook, etc, are glaringly absent from the first pageful of hits ².

The lack of diversity in the return results is also evident in the image search domain. For example, the majority of the returned images from issuing the (ambiguous) query “apple” to Google Image feature the Apple Inc company logo in various shapes and colors, and none of any Apple products such as Mac or iPod ³.

The presentation of redundant results is not only a waste of the limited screen space. More importantly, it causes other diverse but potentially relevant results to be relegated to the next result page. This is clearly a bad practice, considering that most users have limited attention span and rarely go beyond the first two result pages [68].

1.1.3 Contemporary Approach to Ameliorate the Two Dissatisfactory Retrieval

To resolve these problems of dissatisfactory retrieval described in the previous subsection, the contemporary approach is to use *query refinement* strategies such as *query expansion*, *query suggestion* and *Result Clustering* [12, 143, 151, 54, 53]. For ease of comparison with the basic retrieval architecture in Figure 1.1, Figure 1.2 shows the new refinement modules highlighted in red.

1.1.3.1 Query Expansion

For query expansion, to ameliorate the term mismatch problem, the original query is augmented with new terms that are obtained from analyzing either the query log, or the retrieved documents using the initial query, e.g. finding the most common terms from the query log or from the return results of the initial query. There are a few problems with this expansion approach. First, the expansion strategy is heavily dependent on the *availability* and *quality* of the data source (query log or the initial retrieved documents) upon which the expansion is based. Second, this expansion heuristic only works well if the initial query is a good expression of the user’s information need, resulting in the presence of many other relevant words in the initial retrieved documents. If the initial query is

²The lack of diversity in the outputs is a consequence of ranking the videos by meta-text similarity. As a comparison, the same query of “larry page” when issued to Google Web Search also exhibits high content similarity in the first pageful of hits: seven of the ten hits are of a biographical nature, e.g. a profile writeup of him on Forbes, Stanford, New York Times, etc. Remarkably, the writeups in many of these pages are very similar.

³All three YouTube (“larry page”), web (“larry page”) and image (“apple”) queries were issued on 20 May 2011

Figure 1.2: A multimedia information retrieval system augmented with query refinement strategies such as query expansion, query suggestion and results clustering.

badly articulated, irrelevant documents may be retrieved and immaterial expansion terms introduced to the original query, resulting in a query drift [183].

1.1.3.2 Query Suggestion

In contrast to the query expansion technique which is often done transparent to the user, in query suggestion, the term mismatch problem is ameliorated using an *assistive* approach. This is typically implemented as an *online* mechanism to provide suggestions, correct misspellings [53], and auto-complete a user's query [47] as he types into the search bar. However, because of the constrain for real-time response, the new terms in the query suggestions are mostly mined only from query log, by finding other queries that co-occur in sessions with the typed query, or queries that are lexicographically similar. Given a large query log, there may be hundreds of such similar queries. However, given the limited available space, only a small number of related queries can be shown.

1.1.3.3 Result Clustering

To ameliorate the problem with too many similar results, a clustering approach can be used to group results. This becomes important in situations when users are unsure of what they want. So when they issue imprecise and/or ambiguous queries, a clustering approach can present the results in groups that are likely to represent distinct information needs. For example, for the query “larry page”, the similar results describing his profile can be grouped to cater for users looking for his biographical information. Other results describing about his recent speeches and movement, etc, can be similarly clustered. Furthermore, with only limited available space in the results page, and in the absence of disambiguating information about the specific user search intent, it is better to present a diverse set of results that cater to different information needs.

While the clustering problem has a long history and there exist a large number of clustering techniques in pattern recognition, data mining, etc, there remains two significant challenges [83, 19]: (a) clustering high dimensional data, and (b) the need for a predefined number of clusters. In the former, as the dimensionality increases, the data points become increasingly sparse and it becomes harder to define a meaningful measure of similarity between points. In the latter, the number of cluster is an important parameter in most clustering algorithms. While the utility of this number serves a pragmatic objective to specify a stopping criteria to the clustering process, more importantly, it represents a general need to have some “insight” into the data, e.g. how a “good” clustering looks like, the expected number of clusters in the data, etc. This presents a problem to clustering search results from an ambiguous query. Because different queries have varying level of ambiguity, there may be a varying number of clusters in the return results. As a result, the number of clusters must be automatically computed, a non-trivial task.

1.2 Research Issues

In this section, we detail the research issues that this thesis sets out to explore and solve. First, in subsection 1.2.1, we summarize the limitations of the query refinement methods described in the previous section. Then, in subsection 1.2.2, we examine the root causes of the problems, before describing the main research issues in subsection 1.2.3.

1.2.1 The Limitations of Current Query Refinement Methods

1.2.1.1 *The Query-Centric Expansion is too Overly Reliant on Query Log*

The typical query expansion approach is top-down and query-centric, and assumes the availability and quality of secondary data sources, such as a query log. This is a very narrow assumption and is unduely restrictive. In many application scenarios, the *only* data source at our disposal is the document corpus. Further, query logs cannot be transferred across document corpus. Hence, a “bottom-up” and document-centric approach to query refinement, *without* the need for a query log, is more appropriate.

1.2.1.2 *Term-based Expansion is too Simplistic*

Most expansion techniques are based on simplistic term-based co-occurrence to locate words that are semantically related to the original query. There is a need to go beyond surface co-occurrences to deeper semantic-bearing methods that can capture the broader semantic relationships amongst words, e.g. the *topics* or *senses* of related words. For example, given the query word “court”, words such as “judge”, “jury”, “tennis” and “hard” can all be easily associated using simple surface co-occurrence. However, while all documents containing these co-occurring words are related to the query, it is not wise to present them all as equally relevant documents. This is because these documents are *topically* distinct. For example, documents containing the words “judge” and “jury” are of the *legal* sense, while documents containing the words “tennis” and “hard” are of the *sports* sense. If we were to rank and present document results without considering their topical affiliation, we may have a situation wherein the top ranked documents are dominated by a particular sense, resulting in a perception of mis-retrieval and duplicates for a person who is actually looking for documents of other senses. Hence, documents need to be judiciously categorized according to their topical content. By providing query refinement at the topic/sense level, we can improve retrieval by reducing duplicates and diversifying results according to different information needs.

1.2.1.3 *Limitations of Conventional Clustering Approach*

Applying standard clustering to the search results is unlikely to yield satisfactory retrieval. This is because the dataset is high dimensional and we do not know in advance the number of clusters. What

is needed is a non-parametric clustering method that can better deal with high dimensional data, and does not require a-priori the number of clusters. This becomes critical when we want to cluster search results from an ambiguous query, in order to identify clusters describing distinct information needs. The varying degree of ambiguity in the query means that for each ambiguous query, we need to apply a different number of clusters to group the return results.

1.2.2 Issues and Challenges

We view the aforementioned problems of dissatisfactory retrieval as symptomatic of the problem of Search Ambiguity. In this subsection, we examine the three root causes of Search Ambiguity.

1.2.2.1 *Lexical ambiguity in short queries expressed using natural language*

It has been observed that the typical length of a user query is only 2-5 words [171]. Because words are ambiguous, there can be many possible meanings behind the query intent [93]. For example, given the query “mouse”, the user may be referring to the rodent mouse, or the computer mouse. This ambiguity can cause documents to be retrieved that are not relevant. The search disambiguation problem here is concerned with finding a set of documents that represent the common or probable *senses* of the ambiguous query.

1.2.2.2 *Concept Synonymy and Topic Structures*

Words and natural languages are expressions of the concepts in our mental thought process. The hierarchical and inter-linked structures governing these concepts mean that certain concepts are synonymous, or are *subtopics* of one another. For example, “machine learning” can have various subtopics such as “machine vision”, “supervised learning”, “generative model”, etc. The direct implication for retrieval is that a document can still be relevant even though it does not contain the same words as those in the query. Hence, a relevance measure that is based on simple keyword matches will not be ideal, as there may be many relevant documents to “machine learning” without them sharing the common words “machine” and “learning”. The search disambiguation problem here is concerned with finding documents that cover as many different subtopics of a general topic as possible.

1.2.2.3 Noise and Errors during Feature Extraction as Sources of Ambiguity

When documents in a search collection are short or noisy, there may not be sufficient information to find a confident match to the query [18, 163]. This is especially true for the multimedia search domain, wherein the content is typically represented by textual features from meta-data description, automatic speech recognition and/or video OCR. The resulting textual descriptions are noisy because the feature extraction step is generally inaccurate. For example, state-of-the-art speech recognition accuracy is about 70% on clean broadcast news video. Hence, we can expect that some keyword matches are correct, while others are wrong. Another example of noisy features is the set of low level features typically adopted by the vision community to represent images. Because of the semantic gap [144], these features are non-semantic and can be seen as forms of “ambiguity” that further compound the problem of search ambiguity. Any proposed solution to the search disambiguation problem here must be robust enough to cope with noisy features.

1.2.3 Thesis Objectives

Motivated by the observations described in the previous subsections, the primary objective of this thesis is the development of computational methods for handling ambiguous queries. This is addressed from the perspective of (a) a new document and retrieval model to elucidate the wide range of topics or facets that are relevant to the user queries, and (b) a non-parametric clustering model to adaptively handle the varying levels of ambiguity inherent in the user queries. In the following, we describe each of these perspectives.

1.2.3.1 A New Document and Retrieval Model for Multimedia Content

Given an ambiguous query, the critical task for a retrieval system is to return diverse and novel information elucidating the wide range of topics or facets of the query. As discussed in the previous subsections, the conventional term-based methods are ill-designed for this task. The issue is that these algorithms have too narrowly defined the relevance of a document to be based on the exact matching of the terms. In doing so, they have inadvertently transferred the burden of the search efficacy to the ability of the search user to input the appropriate query term. However, this is a very weak assumption. A person who types in the query “Arafat” ought to be also presented with documents

describing “Palestinian”, “Israel”, etc, even if these documents do not have the word “Arafat”. Note that this is an important point that differentiates the approach proposed in this thesis from the approach adopted in conventional text-based search systems such as <http://search.yippy.com>. The latter is based on search result clustering to elucidate the topics related to the query. However, to retrieve a list of top-ranked search results for clustering, it still depends on term-based matching of the query term.

The shortcoming of the term-based retrieval framework is especially evident for multimedia content, wherein the feature extraction process is itself fraught with noise. For example, the word “Arafat” was spoken in the multimedia content, but it was not recognized correctly by the speech recognition engine ⁴ ; or the word “Arafat” appear as text in the multimedia content, but it was not recognized correctly by the video OCR engine. The implication of this point is that the multimedia content may not be adequately represented by the salient keywords that one would expect. As a result, the direct matching of the query terms to the indexed speech content will not receive any hit.

In summary, the traditional term-based indexing paradigm is not effective to achieve satisfactory retrieval in a practical multimedia retrieval system. What is needed is a new document and retrieval model of multimedia content, that goes beyond simple term-based similarity to the matching of higher-level semantic topics. Ideally, a multimedia document is represented as a mixture of semantic topics, wherein each topic is a probability distribution of words. In the multimedia domain, the topical analysis of a document can capitalize on the multiple complementary data streams in the multimedia content, such as visual features and text features. Hence, the latent topic mining of multimedia content must also be amenable to multimodal analysis.

Under this new representation, retrieval of multimedia documents can now be based on topical similarity with the topical content of the query. To better meet the distinct information needs of users, the new topic-based retrieval model will also facilitate the diversification of search results by augmenting a measure of *topical diversity* in the initial pages of results, to the traditional measure of relevance. This is a new measure of relevance that considers both document relevance to the query and inter-document novelty and diversity.

⁴a typical instance of the Out-Of-Vocabulary (OOV) problem in current generation speech recognition technology in recognizing personal names [15]

1.2.3.2 *Non-Parameteric Clustering that Grows With the Data*

Different ambiguous queries exhibit varying degrees of ambiguity [156]. The number of distinct interpretations of an ambiguous query depends on its level of ambiguity, and can vary by a wide margin. A case in point: using the Wikipedia Disambiguation page, the keyword “bond” has 167 possible meanings ⁵, while the keyword “bank” has 23 meanings ⁶. Hence, given an ambiguous query, the level of *polysemy* in the return results will also differ ⁷. A standard approach to elucidate the thematic groups in return results is via clustering. However, if we were to adopt this standard clustering approach, we are immediately faced with a non-trivial issue. A standard clustering algorithm usually requires the a-prior specification of the number of clusters. How do we specify the number of clusters for a given ambiguous query?

More critically, the direct application of the clustering approach to the multimedia domain brings to surface several new significant challenges. These challenges arise as a result of the *open-ended* nature of the visual appearances of objects in images. For example, even for images of the same topical sense, they can exhibit different visual appearances. This results in images of the same topical category having distinct visual features and distributed far apart in the visual feature space. Conversely, for images of different topical senses, they may exhibit the same visual appearances, resulting in different topical category sharing the same visual features and in near proximity in the visual feature space. The wide visual diversity of objects in images has direct implication to learning and clustering algorithms: most images of a particular topical category generally do not have a consistent and distinct visual attributes. How then can we learn a clustering of images according to their topical senses?

What is needed is an adaptive clustering solution that *grows* with the data complexity. Rather than directly modeling object visual content, we need a learning scheme that goes beyond mere visual appearances, to develop a *generative model* of the image appearance diversity of a topical category. Under this generative model, the diverse visual appearances of a topical category can come from a *countably infinite* pool of shared visual appearance patterns. Different topical categories can still be visually similar and share similar visual appearance patterns. However, what differentiates the

⁵obtained by [http://en.wikipedia.org/wiki/Bond_\(disambiguation\)](http://en.wikipedia.org/wiki/Bond_(disambiguation))

⁶obtained by [http://en.wikipedia.org/wiki/Bank_\(disambiguation\)](http://en.wikipedia.org/wiki/Bank_(disambiguation))

⁷This refers to the co-existence of various topical senses in the return documents

Figure 1.3: Our overall approach and contributions highlighted in purple.

topical categories is their different *distribution* and *combination* of appearance patterns. The topical categorization problem now falls to the analysis of the distribution and combination of the visual thematic structures of images within a topical category. Hence, images of the same topical category that have vastly different visual appearances can still be in near proximity in the distribution and combination space.

1.3 Approach and Methodology

This section describes our approach to tackle the various issues described in the previous section. Our overall innovations are illustrated in Figure 1.3, and best appreciated by comparing against Figure 1.1 and Figure 1.2.

1.3.1 A Faceted Topic Retrieval Framework for Multimedia Content

As shown by the purple region in Figure 1.3, we propose a novel *faceted topic retrieval framework* for multimedia content. Note that compared with Figure 1.2, it has completely replaced the term-based TF-IDF retrieval module. Under this framework, multimedia documents are represented in terms of *topics*, which are in turn probability distribution of *words*. Note that we have intentionally italicized the *words* to refer to their general meaning of encapsulating some semantic units of meaning. In this regard, the *words* can be viewed as a collective sum of the textual words, audio words and visual words that are typically adopted as a basic multimodal representation by the multimedia community.

There are two components in the faceted topic retrieval framework: the *Faceted Document Model* and the *Faceted Retrieval Model*. Under the Faceted Document Model, we again have two sub-components: *Multimodal Joint Modeling*, and *Hierarchical Dirichlet Processing modeling*. Each component represents a contribution in this thesis. In the following, we describe these components.

1.3.1.1 Faceted Document Model

The Faceted Document Model (FDM) is the heart of our proposed Faceted Topic Retrieval Framework. Its importance can be appreciated by the fact that its counterpart in the traditional term-based indexing paradigm is the Vector Space Model. The FDM models each multimedia content as a probability distribution over topics, where each topic is in turn a probability distribution of words. The words are the basic units of feature representations of the multimedia content. We have mostly adopted conventional features such as RGB, patch-based SIFT descriptors, etc. On the other hand, the topics are multinomial distributions over the words. Hence, topics are able to encode a higher level of semantic representation of the multimedia content.

It is also important to note that just as the *words* can refer to the visual words, or the textual words of the multimedia content, the *topics* can also be viewed in a similarly generalized form. A topic can refer to the textual topic or the visual topic. In our probabilistic formulation, to be discussed in the next few paragraph, the topic representation from the multiple modalities are concatenated during the learning/modeling phase to form a *multimodal* topic. For example, the multimodal topic of an image can be formed by combining its visual features, with its text features (from its accompanying close caption/ASR or its surround web text). After the learning/modeling

phase, we can separate the multimodal topic into its respective modalities and inspect the results. In the later experimental sections, we show that after separation, the nature of the respective modalities are nicely retained. That is, when we inspect the textual part of the learnt multimodal topic, we can discern that meaningful textual words are learnt. Similarly, when we inspect the visual part of the learnt multimodal topic, we can also discern that meaningful visual features are learnt.

Each latent topic can be perceived as a facet that underlies the document corpus⁸. Collectively, FDM supports the analysis and retrieval of multimedia content at the topics/facets level. This is the reason why we coin the term “faceted topic retrieval”.

We adopt the Bayesian topic modeling paradigm to probabilistically mine topics in a multimedia collection. The advantage of the Bayesian approach is two-fold: (a) it is grounded on statistical foundations, where efficient learning algorithms have been well studied and readily available. (b) the analytic framework is amenable for *multimodal joint modeling* of multimedia content, where an additional information stream can be easily incorporated in the learning algorithm. We reserve further details of this module to subsection 1.3.3.

Further, we augment the traditional topic modeling framework to allow for an *unbounded* number of latent topics to be learnt from the multimedia documents. The motivation is two-fold: (a) to cater to the varying degree of ambiguity in an ambiguous query; (b) to cater to the open-ended visual appearance diversity of images. We reserve further details of this module to subsection 1.3.2.

1.3.1.2 *Faceted Retrieval Model*

Because multimedia documents are now represented as probability distributions of topics, retrieval can be processed at the topical level, instead of the word level. The topical similarity between documents can be compared using distributional comparison methods such as the Kullback Leibler divergence measure. In this way, topical diversity can be facilitated in the return results by ensuring that the top ranked results are from different topics. The rationale is that by promoting topical diversity amongst the top-retrieved documents, the retrieval system is in a better position to serve the varied information needs of many users. The inclusion of diversity as a retrieval objective requires the formulaton of a new metric to measure retrieval. In this new metric, our ranking formula combines

⁸In this thesis, we use the term “facets” and “topic” interchangeably. Where the context can better be understood by viewing topics in hierarchical form, we will also use the term “subtopic”.

both the traditional notion of document relevance and the topical affiliation of a document.

1.3.2 Non-Parametric Modeling in the Faceted Document Model

In Section 1.2.3, we have motivated the need for a clustering approach that caters to the varying level of ambiguity in a query. We have argued for the need for a *non-parametric* method that can grow with the complexity of the multimedia data. The heart of this approach is this: for each topical sense category of the multimedia data, formulate a generative mixture model of its (visual) appearance. The mixture components of each topical category can come from a pool of *countably infinite* size. By tapping on this infinite pool of shared mixture components, as more data are seen, the model can adaptively grow to better represent the new data. In practice, over the course of data presentation, only a finite number of these mixture components from the infinite pool is used: if the new data is similar with some older data, then the same mixture component is likely to be re-used to represent this new data; if the new data is different in the feature space with all existing data, then a new mixture component will likely be initialized to represent this new data.

To realize this generative model, we need to address two main issues. First, there should exist *countably infinitely* many mixture components of appearance patterns, as the nature of the visual appearance diversity of images is open-ended. Second, all topical categories should *share* a common pool of visual appearance patterns, as images from different topical categories can be visually similar. We resolve the two aforementioned issues by developing a hierarchical generative probabilistic model, based on the Hierarchical Dirichlet Process (HDP) mixture model. The first issue is handled by the *stick breaking construction* process and *chinese restaurant process* [59, 152, 24], that allow the probabilistic prior of a distribution over a countably infinite number of mixture components encoding the various appearance patterns. The second issue is handled by the *Chinese Restaurant Franchise* model [152, 77, 166], which allows the mixture components to be shared within and across different topical categories.

Figure 1.4 shows an example of the HDP clustering result on a set of polysemous images returned by issuing the query “mouse computing”. As mentioned earlier in subsection 1.3.1, images can be modeled as comprising of a mixture of *visual* topics. These usually correspond to the common visual appearance patterns that are found across all images. By introducing a non-parametric prior into our clustering, we enable an unlimited number of such visual appearance patterns to be defined and

Figure 1.4: Using the average image to illustrate the visual polysemy of images. Top: for images retrieved using the query “mouse computing”, they are very diverse. Bottom: average images of three HDP generated clusters on the “mouse computing” images. Note the fine-grained elucidation of “notebook”, “pointer” and “desktop” images.

to be shared across all images from all topical categories. As Figure 1.4 shows, this non-parametric approach can handle different ambiguous queries with varying level of ambiguity. For example, when the query “mouse computing” is issued, some return images are probably related to the query term “mouse”, while others are related to the query term “computing”. Collectively, there is a high level of polysemy in the return images.

However, by using the non-parametric clustering approach, coupled with the use of a dictionary resource such as Wikipedia, we can handle the varying level of polysemy well. As a result, the polysemous “mouse computing” images have been automatically clustered into the relatively meaningful clusters, such as “notebook”, “pointer” and “desktop”⁹. Figure 1.5 also shows the image clusters for the two queries with different level of ambiguity: “tiger” and “watch”. For “tiger”, the six clusters can be interpreted as “bengal tiger”, “tiger pooh’s friend”, “tiger tanks”, “tiger sharks”, “white tiger” and “tiger woods”. For “watch”, the three clusters are “clock”, “watch tower” and “wrist

⁹the titles are our own interpretation after inspecting the images in the cluster

(a). clusters of polysemous images from the query “tiger”.

(b). clusters of polysemous images from the query “watch”.

Figure 1.5: Example clusters on two polysemous keywords: “Tiger” and “Watch”. For each polysemous keyword, we show enclosed in a green box two rows of sample polysemous images returned by Google Image Search. Below the green box, we show the automatic clusters of images enclosed in red boxes. For “tiger”, the clusters correspond to “bengal tiger”, “tigger pooh”, “tiger tanks”, “tiger sharks”, “white tiger” and “tiger woods”. For “watch”, the clusters correspond to “clock”, “watch tower” and “wrist watch”.

watch”. Note that the clustered groups correspond well to the common dictionary (visual) senses of the original query. This has the appealing effect of automatically elucidating the ambiguous queries into their respective common interpretations.

1.3.3 Multimodal Topic Analysis in the Faceted Document Model

Our proposed facted document model of multimedia content can support the analysis and retrieval of multimedia documents at the topic or facet level. Here, we further enhance our model with a modeling capability to hypothesize topical facets in news video by combining the complementary information in the visual keyframes and the speech transcripts. Figure 1.6 shows our multimodal

Figure 1.6: Our proposed faceted document model for hypothesizing facets in news video. The visual and textual features in news video are jointly modeled to yield facets that better correlate with the actual news events.

modeling approach.

In contrast to the previous subsection where images are modeled as a mixture of *visual* topics, here the notion of facets is on the *textual* topics, where a facet is interpreted broadly to encompass any binary property of a news video that presents a fact or a topic that is contained in the information need. For example, given the query “War on Terror in Iraq”, the facets may include “War in Fallujah”, “Military Strategy in Pentagon”, “Bin Laden Tape”, etc. The facets of a query can be contained in one or more video, and a video can contain one or more facets. A video is deemed relevant to the query if it contains any of the facets.

Our key motivation in applying a joint modeling approach to mining facets in video is to derive more robust facets that better correlate to actual news events. While we reserve more details of our results in a later chapter, we briefly show in Table 1.1 the qualitative improvement in some detected facets as a result of including visual features during topic learning. On each of two queries “War on Fallujah” and “Mideast peace”, we introspectively pick a learned faceted topic that contains high probability words that are meaningful to the issued queries. Observe that words from our joint model are more intuitive and correlate better to the query topic.

Table 1.1: Top-10 probability words of 2 learned β_k topics, for “War on Fallujah” (top) and “Mideast peace” (bottom). The topics are picked introspectively. Meaningful words highlighted in bold.

Text-only modeling	Visual+text joint modeling
iraq peopl time good meet unit state baghdad chines dai	iraq iraqi peopl militari govern arm kill forc attack baghdad
arafat know yasser leader thing peac minist peopl just israel	palestinian arafat peac israel presid leader elect yasser east middl

In our joint modeling of visual and text features, there are several key deviations from existing modeling approaches. First, our modeling granularity is coarser: our visual features are not at the local patch-level but rather at the keyframe-shot-level. This most closely resembles Wu *et.al*’s video representation with visual shot duplicates [172]. Second, because topic models are unsupervised methods, often best results are obtained by incorporating a priori knowledge about the desired output (e.g. must-link constrains in clustering). Adding observations from cross-media types as a way to constrain topic modeling is proposed by several authors, e.g. the image annotation model of Blei and Jordan [25] and correlating name captions with face recognition [85]. Specifically in our case, we utilize recurring shot features as constrains to the probabilistic latent topic modeling process to learn facets that better correlate with actual news event. To compute recurring visual shots, we use the Near Duplicate Image (NDI) detection method of Chum *et.al* [45]. Third, our modeling objective is faceted topic retrieval of news stories.

1.3.4 Topic-based Query Refinement

In Section 1.2.1, we have highlighted several limitations of the current query refinement methods. The major drawbacks includes the over-reliance of a query log, and the simplistic term-based methods to mine query expansion terms. In this subsection, we propose methods that compute a more general semantic analysis of a given query. The goal of these methods is to associate semantically related words that are affiliated to the various *senses* or *thematic interpretations* of the original query. We empirically show that our approach is better able to address ambiguous queries and help users to find their information.

Figure 1.7 illustrates how our method provides query suggestions that go beyond simple word prefix matching [168]. The video database is a collection of 150 hours of video reporting on the

Figure 1.7: An example of our Topic-based query disambiguating search [167]. Left: the normal simple word prefix based query suggestions as the user type into the query box. Right: as the user selects a suggested name, the system further computes subtopics that are semantically related with the name.

recent Singapore General Election 2011. Each video comprises of a short text description, and is further augmented with automatic speech transcript and video OCR. As the user starts to type into the query box, our system first attempts to suggest the top few named entities using simple prefix matching. This is as per what most existing query suggestion systems do. However, as the user selects a particular name, our system now performs something different. It lists several *subtopics* related to the selected name (“Tin Pei Ling”), such as “Goh Chok Tong”¹⁰, “Marine Parade”¹¹, and “NSP Seah”¹². There are two advantages of this approach to query suggestions. Firstly, the offered suggestions represent some thematic topics that the system thinks that the user may be interested. Compared to a simple lexicographical matching approach to suggesting queries, they are intuitively better able to help a user clarify and locate his information need. Second, in contrast to a query log-based approach, these suggestions are directly mined from the document corpus. Hence the user can be assured of the availability of the information represented by the suggestions [167].

¹⁰the name of the leader of the PAP team in which Tin Pei Ling was a member of

¹¹the constituency that Tin Pei Ling was contesting in

¹²another election candidate by the name of Seah, belonging to another party called NSP, that Tin Pei Ling was often compared with

1.4 Thesis Contributions

In the following, we summarize our key contributions (also shown in Figure 1.3). The related publications are also enclosed in square brackets. Each publication is uniquely numbered, and prefixed by the letter J if it is a journal publication, or the letter C if it is a conference publication.

1. We propose a Faceted Topic Retrieval Model (FTRM) for representing multimedia content that supports retrieval of information that are novel and yet diverse [J1, J2, C1, C2, C3]. There are two distinct components in FTRM:
 - (a) a Faceted Document Model (FDM) to hypothesize facets in multimedia content.
 - (b) a Faceted Retrieval Model (FRM) to rank faceted documents with respect to their relevance to the original query and diversity in the results.

The former is concerned about how to hypothesize and represent a multimedia document as a vector of facets or subtopics. The latter is concerned about how to rank faceted documents to incorporate both novelty and diversity in the results.

2. We further enhance the Faceted Document Model (FDM) with two novel modeling capabilities:
 - (a) Multimodal modeling capability that jointly model the visual and textual features of news video for improve subtopic analysis and retrieval [J1, J2, C1, C2, C3, C4]
 - (b) A Non-Parametric modeling capability, based on Hierarchical Dirichlet Process, that grows with the complexity of the data [J1, C3]. The complexity of data can vary because, for example, given different level of ambiguity in a query, the retrieved images from issuing the query can exhibit varying level of polysemy.
3. We demonstrate the practical utility of FTRM framework in the following search ambiguity problems:
 - (a) Result Clustering of polysemous images retrieved from an ambiguous query, that does not depend on the prior knowledge of the number of clusters [J1]. Our clustering method can handle queries of varying level of ambiguity, and automatically elucidates the polysemous images into its various dictionary senses [J1, C4, C6, C7].

- (b) We combine the visual and textual features as a unified generative model of topical facets in news video [C1, C3, C5, C7]. The joint modeling enables the formation of topics that better correlate with the actual news event, leading to improve retrieval of novel but yet diverse news video.
- (c) a novel online Query Suggestion mechanism that provides faceted topical suggestions related to the query, rather than simple lexicographical query matches [J2, C1]. Based on our FTRM of elucidating subtopics in a document collection, it automatically selects keywords that are semantically affiliated to the original query. These keywords are then augmented to the original query as an expanded query to search into the video database. Our method helps the users to better refine their query and to identify more specific topics of interest.
- (d) a topic-based Query Expansion scheme that selects candidate expansion keywords based on their affiliations to the same thematic topic as the query, as opposed to conventional approaches that are based on simple frequency statistics [J2, C1, C5, C6, C8].

1.5 Outline of Thesis

Chapter 2 introduces the necessary background knowledge that is foundational to the thesis. It also reviews the literature on search disambiguation that are relevant and share similar ideas with the thesis.

Chapter 3 is the main meat of the thesis. It provides the full disclosure of the proposed faceted topic retrieval model for multimedia content. It first defines the problem, before presenting several new ways of evaluating the performance of a retrieval system that incorporates both document relevance and topical diversity in the search results. The set-based probabilistic retrieval framework is next described, with details on the Bayesian topic modeling methodology used to mine topics in documents. Particular emphasis is placed on three critical aspects: (a) how we augment the standard topic modeling framework to achieve multimodal modeling of the complementary information stream in multimedia content? (b) how we can model the open-ended complexity of the return results from ambiguous queries using the Hierarchical Dirichlet Process (HDP)? (c) how can we develop an efficient algorithm to incorporate topical diversity in the top-retrieved results?

Chapter 4 turns the attention to applying the faceted retrieval framework to the image domain. In particular, we consider the problem of clustering a set of polysemous images returned by an ambiguous query. The focus is on the use of the Bayesian nonparameteric method to facilitate clustering without the need to know in advance the number of clusters. This allows our faceted image clustering algorithm to handle a varying level of image polysemy arising from different levels of ambiguity in the query. We show how to construct a visual sense classifier that is able to cluster and re-rank the polysemous images into its sense categories. For increased robustness, we need some help from external resources such as Wikipedia Disambiguation and some sense-specific images as seed for the clustering step. We demonstrate that using our proposed faceted model based approach, we can allow sense specific diversity in the image search results.

Chapter 5 discusses the faceted retrieval framework for the video domain. In particular, we focus on the problem of faceted news video search. Because the utility of our topical retrieval system depends on the quality of the mined topics, it is critical that the learnt topics must capture the salient words that have a meaningful correlation with the actual news events. Hence, the objective is the development of more robust topics by capitalizing on the multiple complementary information that is embedded in the multimedia content. We describe the joint modeling of the visual and text information in a video content to constrain the formation of topics. We show that this results in better retrieval experience.

Chapter 6 describes a faceted topic based query suggestion method to allow a user to interactively select and refine his information need. While most existing query suggestion methods are mined from query log and use some simple measure of query frequency or lexicographical matching, we propose an alternative method of presenting query suggestions by their thematic topics. Our method adopts a document-centric approach to mine topics in the corpus, and does not require the availability of a query log. Empirical results show that this alternate way of organizing query suggestions can better elucidate the high level query intent, and more effectively help a user meet his information need.

Chapter 7 concludes the thesis, highlighting its particular areas of novelty and contribution.

Chapter 2

LITERATURE REVIEW

This thesis is relevant to a range of research topics. These include faceted retrieval, faceted modeling of multimedia content, probabilistic topic models, non-parametric Bayesian methods and search disambiguation. This chapter presents an overview of the necessary background knowledge and foundational concepts that underly our proposed methods. The chapter is outlined as follows:

- Because a major contribution in this thesis is a faceted topic retrieval model for multimedia content, in Section 2.1, we discuss some related works in faceted retrieval.
- A key task in faceted retrieval is the detection of meaningful facets in content. Hence, we survey related works in facet modeling and detection in Section 2.2.
- We have based our facet model on the probabilistic topic model. The issue of joint modeling and inference of the complementary information in a multimedia content is important to our work. Section 2.3 examines work on joint topic modeling and constraining a topic’s development.
- Another important aspect in our facet model is its ability to grow with the ambiguity in the query topic. In Section 2.4, we look at how Bayesian non-parametric methods can help us to achieve this goal.
- Many ideas related to Search Disambiguation are also explored in other research fields, e.g. sense disambiguation, diversification. In Section 2.5, we briefly point out some of these works.

2.1 Faceted Retrieval

Since the early days of information retrieval, researchers have debated on the concept of relevance. The general consensus was that while the relevance of a document may be measured on a binary or continuous scale, it is judged independently from other documents. In other words, two documents are deemed to be relevant as long as they contain information that the user needs, *even if* the two documents are identical. This independence assumption has by and large served the community well and has spurred the development of advanced retrieval models over the past decades. However, in recent years, the retrieval community is increasingly aware of the need to not only retrieve relevant documents, but also *diverse* or *novel* information about a topic. With such a goal in mind, these retrieval models are found to fall short.

The problem is that the relevance dependencies between documents are not captured. In such a setting, the retrieval system disregards the fact that a user may not find it useful to read a second document that contains the same (or nearly the same) piece of information as an earlier retrieved document. To the user, the second document is only useful if it contains new and yet relevant information. The key factor to the utility of a document is not *only* whether it contains any relevant information, but that whether it also contains *new* information. In fact, we can conjecture that if a third document contains both the two pieces of information found in the first and second document, its usefulness is *higher* than that of the first two documents. Using the conventional evaluation metrics such as Precision/Recall, Discounted Cumulative Gain, etc, this notion of retrieval utility is not adequately captured. As a result, the task to retrieve diverse or novel information would require a new formulation of evaluation and operation methods on different types of relevance metrics [49, 176, 40].

The need for diversity in the result sets is well resonated in the literature. Goffman [67] was amongst the first to suggest that the relevance of a document should be dependent on the relevance of previously retrieved documents. Subsequently, several researchers have focused on post-processing search results to reduce redundancy, and also formulated new diverse ranking methods. In result re-ranking, Clarke *et al* argue the interest of the overall user population must first be taken into account [49]. Carbonell and Goldstein [39] developed a diversity ranking algorithm to diversify search results as a way to address ambiguous queries. They proposed a Maximal Marginal Relevance (MMR)

ranking function to tradeoff between maximizing relevance while minimizing similarity amongst the retrieved documents. Zhai *et al* [176] further extend MMR to a general framework to score documents with probability of relevance and novelty. They termed their method as *subtopic* retrieval, wherein a subtopic can be viewed as a meaningful interpretation of the query topic. Because there are usually many meaningful interpretations, or subtopics, for a given query, they develop a method to re-order a set of initially retrieved documents such that many of these subtopics are probabilistically contained in the documents in the top ranks. They have proposed a new evaluation metric for subtopic retrieval that we have also adopted in this thesis. Similar probabilistic formulation of faceted retrieval have also been proposed by Chen and Karger [43] and Carterette *et al* [40]. The former applied a greedy algorithm to penalize redundancy in the result set. The latter used a set-based probabilistic method to control facet duplicates. They have also coined the new diversity/novelty-driven retrieval task as *faceted topic retrieval*, to emphasize the focus to find documents that cover the different *facets* of an information need [40].

A natural evaluation of such a faceted topic retrieval system is to assess its ability to cover as many facets as possible with the smallest number of documents. In this thesis, we focus on developing such a model for faceted topic retrieval for multimedia content. It is important to note that facets are highly correlated, and often appear together in the same documents. The utility of faceted retrieval lies in the *coverage* of the information need present in the retrieved documents. The challenge of a faceted topic retrieval system is to maximize this coverage using the least number of retrieved documents: having one document that cover two facets is better than having two documents that cover one different facet each.

2.2 Faceted Modeling

With the proliferation of digital content, there is great interest in developing computation tools to help organize, understand and navigate through these vast amount of data. Currently, this need is only partially met by the conventional term-based indexing and retrieval techniques. However, often times, in-depth knowledge discovery can only happen when we examine the evolutionary and correlation patterns among multiple facets across a text corpus. In this section, we look at how facets can be modeled in multimedia content. We first start in Section 2.2.1 by examining the basic methods

to model content using low level features (without semantics). Then we look at ways in Section 2.2.2 to incorporate (some) semantic information. Finally, we discuss several ways of modeling facets in Section 2.2.3.

2.2.1 Conventional Modeling

2.2.1.1 *Vector Space Model for Text*

We first look at how text is conventionally modeled. Since the early days, discrete data such as text collection has been widely represented by the vector space model [139](VSM). VSM models a document using a N -dimensional vector of the count of each unique word w (after stop-word removal and stemming) appearing in the document. This is commonly called a bag-of-words (BOW) model where N is the size of the vocabulary. By ignoring the sequence of word appearance in a document, the BOW model computes the similarity between texts simply as a cosine of the angle between their BOW-count vectors. This formulates the intuition that text describing similar topics would likely re-use the set of words often used to describe those topics. Hence, the similarity between text, a foundational basis for any information retrieval task, is implicitly defined at the word level. However, this can result in ambiguities. For example, the word “jaguar” can occur a few times on both a document about the animal, and a document about the car. Hence, documents should ideally be compared at the topic level. To resolve word ambiguity, each word should be interpreted in its (topical) context, wherein a topic is a distribution of frequent co-occurring words. Models that attempt to automatically capture this co-occurrence have therefore attracted interest in the information retrieval community.

2.2.1.2 *Modeling in the Multimedia Domain*

The early visual retrieval systems have mostly focused on an attribute-based representation with a high level of image abstraction and modeling of visual content at the conceptual level [149]. However, the main drawback is that most of these visual descriptions need to be manually entered, This has prompted the development of perceptual features like color, texture, shape, etc, and a visual retrieval model that is defined by a similarity measure on these features [144]. Despite these efforts, progress is hampered because of the significant computation load in calculating these visual descriptors, and

by their poor descriptive power to represent the semantic concepts that are used by humans¹. At the same time, it is often possible, and easier, to obtain some form of textual data from the multimedia content. These can be in the form of textual (named) entities that is embedded in the video frames or the automatic transcripts from the speech and sound. When such text-based descriptions are used during indexing, they provide a higher level of semantics and retrieval robustness. Hence, most existing multimedia search engines are mainly text-based.

2.2.2 Semantic Models

To address the semantic gap problem, various studies have considered semantic models of multimedia that go beyond the basic constituent elements of video information. For instance, many systems seek to represent a video not only by merely concatenating its constituent text vector and visual descriptors, but also attempt to exploit semantic context to define higher-level concepts and mine their relationships. By considering these concepts and their relationships, retrieval of relevant information can be improved [92]. For example, if the word “bear” appears in a document², it might become more likely that words like “polar” and “feed” will also appear in that document. The context of this document is the animal and the retrieval of relevant information concerning the animal “bear” can benefit from retrieval of information on “polar” and “feed”. However, this association task is often complicated by the fact that words have multiple senses or meanings. If words like “stock” or “market” appear in the document, then it is likely that “bear” refers to the *economics* context, and words like “finance” and “downturn” should also correspondingly increase in probability and relevance. Traditionally, the study of modeling these kind of semantic relations can be categorized into the following two approaches [71].

2.2.3.1 Semantic Networks

The first approach is exemplified by the associative semantic networks [50], abbreviated as SN. This approach emphasizes deep, abstract conceptual structure and the relations among concepts, typically expressed in abstract propositions, such as *is-a* and *has*, etc. The focus of SN is developing conceptual hierarchies, rather than the associative relations between words and concepts. The se-

¹this representation problem is also known by the name “semantic gap”

²for a video content, this can be a detection result of a “bear” visual concept classifier applied to an appropriate keyframe.

mantic knowledge in the concept relations is typically hand-crafted by modelers, and the learning of large-scale conceptual relations from data corpus remains an unsolved problem.

2.2.3.2 *Semantic Space*

A second approach of studying semantic representation attempts to model the structure of associative relations between words in natural language use (i.e. from large text corpus) and their contextual dependence of these relations. Hence, by analyzing many documents in the *animal* context, words like “jaguar” can become more likely to be associated with “hunt” and “food”. At the same time, if these documents are mainly from the *car* context, then “jaguar” can also be likely associated with “price” and “torque” in the *car* context. These expectations reflect the fact that “jaguar” has multiple senses, or multiple concepts it can refer to. Methods that exemplify this approach include the recent Semantic Space [106] (abbreviated as SS) and Latent Semantic Analysis (LSA) [94].

The modeling goal of SS is to formalize a semantic space that encodes concept proximity. For example, given the *animal* context, “jaguar”, “hunt” and “food” are close together, while given the *car* context, “jaguar”, “price” and “torque” are closer together. Compared to the first approach (i.e., SN), these methods of semantic association are based on comparatively simple models (e.g. each word is a feature), but provide a direct means of investigating the influence of the statistics of language use (directly from real linguistic corpus) on semantic representation.

The Latent Semantic Analysis (LSA) [57, 94] approach is also a corpus-driven method that attempts to address word ambiguity problems by dimensionality reduction. More specifically, LSA maps documents and their terms to a representation in the so-called *latent semantic space*. Using the vectorial representation of the documents in a collection as starting inputs, LSA applies a Singular Value Decomposition (SVD) on the corresponding term-document matrix. The dimension reduction results from the operation of SVD to project the term-document matrix onto a linear subspace that captures most of the variance in the original dataset. More specifically, terms that frequently co-occur in the original dataset are likely to be mapped onto the same “terms” in the new reduced space. The main claim that LSA makes is that the similarities between documents can be better estimated in the reduced latent space representation than in the original VSM representation. Documents which originally share frequently co-occurring terms will now have a similar representation in the latent space, even if they have no terms in common. Hence, the issue of word polysemy and synonymy is

alleviated in the following sense: for two documents to have high matching score, they no longer need to have many common terms; they only need to share many terms that also frequently co-occur in other documents in the collection.

Continuing on the line of deriving semantic information from word-document co-occurrence and dimensionality reduction, the probabilistic Topic Models (abbreviated as TM) have recently emerged as powerful modeling tools for discrete data collection [27, 25, 70]. Similar in many ways to LSA, the TM approach also emphasizes on a light-weight representation (each word is a feature) and statistical learning from large text corpora. However, TM differs from LSA in that while LSA expresses the semantic properties of words and documents in Euclidean space, TM expresses them in terms of a set of probabilistic topics. Topics are distributions over words that encode the semantic context of documents. Continuing the earlier example, “jaguar”, “price” and “torque” are likely to be high probability words in a “jaguar-car” topic, while “jaguar”, “hunt” and “food” are likely to be high probability words in a “jaguar-animal” topic. Topic models are formulated as generative models, and semantic modeling is cast as the problem of statistical inference: given some observed discrete data (e.g. the set of words in a document collection), infer the latent semantic topics from which it was generated. In recent years, generative models have become popular in both computational linguistics and machine learning [42, 108] as they are able to harness the strength of both structured representations (encoding of knowledge and meaning) and statistical methods (availability of efficient learning algorithms). By representing semantic structures as topics, TM is a novel semantic representation that goes beyond networks and spaces [71]. It is also the foundational framework in this thesis to model facets. We reserve further discussion of TM to Section 2.3.

2.2.3 Modeling Facets

The construction and modeling of facets for diversity retrieval is a relatively new area. However, facets in a document can be modeled in three ways: a clustering model, a classification model and a generative model. In the following, we discuss each of them.

2.2.3.1 Clustering Model

Clustering of search results has long been proposed to improve retrieval effectiveness [175, 4, 170, 55, 75]. The key intuition is that relevant documents tend to be clustered together. Hence, the different

clusters of documents would likely represent distinct topics relevant to the original query. Each document can have a probabilistic membership to every topic clusters, according to the distance of the document from the topic cluster centroid.

Most clustering methods can be divided into two categories: discriminative approaches [82, 160] and generative approaches [23]. In the former, an objective function is defined with the pairwise document similarity measure as a parameter, and the objective is to maximize the average intra-cluster similarity while minimizing the average inter-cluster similarity. In the latter, a generative model of document is defined and learnt from the documents in an unsupervised manner. Probabilistic topic models, which is the modeling framework of this thesis, falls under this category. We reserve further discussion of this category to subsection 2.2.3.3.

There are two main disadvantages of clustering [74]: (a) incoherent groups due to conflation of many dimensions, and (b) the difficulty of labeling the groups. To resolve the latter, candidate labels can first be generated, and then choosing a subset of them as cluster labels. There are a variety of ways to generate candidate labels, such as document titles, document phrases or N-grams, most frequent and/or salient words in a cluster, and named entities, etc. It remains an issue that selected final labels may still convey little semantic value.

2.2.3.2 *Classification Model*

The facets that are present in a document can be modeled from a classification approach. For example, Dakka *et al* [56] treat facet classification as a text classification problem, and introduced a supervised approach for extracting useful facets from a collection of text data. In text classification, each document is represented using a set of words. Using a training document set with words across multiple categories, a Support Vector Machine (SVM) [51] classifier can be trained to assign new documents to the appropriate categories. For better facet generalization, an external knowledge resource such as WordNet Hypernyms ³ is used to assign new hyponyms ⁴ words to each word in a facet. For example, the words “cat” and “dog” are hyponyms of the “Animals” facet.

In the visual domain, such a classification framework is still applicable. The general methodology would in principle still be valid by exploiting a similar external visual knowledge resource such as the

³A hypernym is a word that is more generic and superordinate than a given word, whose meaning includes the meaning of the given word. e.g. “Musical-instrument” is a hypernym of “guitar”, “piano”, etc

⁴A term that denotes a subcategory of a more general class; the opposite of hypernym

ImageNet [81]. Another related visual classification approach has become popular in recent years – the visual concept detection and annotation task [3, 8]. This is a multi-label classification task that is aimed at the automatic annotation of an image with multiple annotations. The task can be solved using a supervised training approach, by either using the visual information alone, or with the combined use of meta-data such as Flickr user tags, meta-file-description, and/or speech transcripts.

The main problems with a classification approach are two-fold: First, the facets that could be identified are, by definition, limited to the facets that appear in the training set. Second, the selection of the facet classes and the initial set of keywords associated with each of these facet classes is manually done.

2.2.3.3 *Generative Model*

Generative models [88, 98, 78, 27, 25] aim to probabilistically describe how a dataset may be formed, often by modeling the dataset as a joint probability of the observed data and some unobserved data. Modeling the dataset in this way corresponds to making an assumption or a hypothesis of how the data is *generated*, hence the reason for the name *generative model*. Assuming that the modeling hypothesis is true, then general inference algorithms such as Expectation Maximization, simulated annealing or Gibbs Sampling can be used to converge to a clustering solution that best agrees with the observed data.

In recent years, this class of modeling technique has found widespread acceptance mainly because it offers solutions in an unsupervised manner. In particular, the probabilistic topic models have been used to automatically discover hidden themes in a document collection [28, 69, 71]. The main attraction of topic models to cluster documents are as follows. First, it assumes that words in a document are chosen from a set of unknown latent topics, which correspond to the clusters of the documents. Note that in the context of our discussion on facets in this subsection, these unknown topics can also be seen as the potential facets of the dataset. Second, the modeling of the relationships amongst the observed words, documents and latent topics is based on a theoretically robust probabilistic framework. Third, upon learning convergence and obtaining the topics, the same probabilistic framework also enables the inference of the cluster probability for an unseen text.

Because we have also grounded our work on probabilistic topic models, we shall reserve further extensive discussion to the next section.

2.3 Probabilistic Topic Models

We focus on a specific instance of a latent topic model: the Latent Dirichlet Allocation (LDA) [27]. In recent years, LDA has seen much success in many interesting applications in both the text and visual domains. One of the key advantages of LDA is its fully generative probabilistic model, allowing principled extensions and variations capable of describing and modeling rich problem domain structure [7, 134, 31, 70, 150]. LDA and many other topic models share the same modeling premise that a document is a mixture of topics and each topic in turn is a mixture of words.

For example, consider the sentences in Figure 2.2. In the sentences, different words are highlighted. Words such as “dollars” and “financing” are highlighted in purple; they are words about *economy*. Words such as “winners” and “champion” are highlighted in pink; they are words about *sports*. Hence, we can see that the document that contains these sentences blends a mixture of the topics *economy* and *sports*. Each word has a probabilistic membership to both topics. The collective topical weightage of all the words in the document would then provide an idea of the overall significance of every topic in the document. Applying the weighted computation to every document in the collection, we would also obtain the main thematic topics underlying the collection. This is the main intuition behind LDA. It assumes that topics are distributions over words, and that there exist some number of topics for the whole collection. It is best described by a generative process, the imaginary random process by which the model assumes all the documents come about. We discuss this in the next subsection.

2.3.1 Generative Model

Suppose we represent each document d as a sequence of N_d words: $w_1, w_2, \dots, w_{N_d}$. Note that when we generalize to the image domain, an image is the equivalent of the document d , and the quantized visual words from its local image feature descriptors are the equivalent of words in text. The process of generating d is as follows:

1. For each topic $t = 1..K$, sample a V -dimension multinomial θ_t from a Dirichlet distribution with parameter β
2. For each document $d = 1..N$, draw a K -dimension multinomial ϕ_d from a Dirichlet distribution with parameter α

Figure 2.1: Graphical representation of LDA. Shaded node w denotes the observed random variable for the occurrence of a word (visual word). z is the topic index for all words in a document: z_d denotes the topic assignment variable for document d , where $z_{d,n}$ is the topic assignment for the n th word in document d . ϕ_d is the topic mixture variable for document d , $\phi_{d,k}$ is the topic proportion for the topic k in document d . D denotes the number of document (images) in the collection, N_d the number of (visual) words in the document (image), and K the number of topics. α and β are the Dirichlet hyper-parameters controlling the document-topic distribution and the topic-word distribution respectively.

3. For each document d , sample visual word w_i repeatedly for N_d times as follows:

- Draw a topic $t \in 1..K$ from ϕ_d
- Draw w_i from θ_t

α and β are the two corpus-level LDA parameters. V is the number of unique visual words in the visual vocabulary. The generative process does not make any assumptions about the order of (visual) words as they appear in documents. This is known as the bag-of-words assumption that is commonly adopted in many Natural Language Processing (NLP) and vision tasks. The generative process corresponds to the following joint distribution of the hidden and observed variables:

$$P(Z, W | \alpha, \beta) = \prod_{i=1}^K p(\theta_i | \beta) \prod_{d=1}^D p(\phi_d | \alpha) \left(\prod_{n=1}^{N_d} p(z_{d,n} | \phi_d) p(w_{d,n} | \theta_{z_{d,n}}) \right) \quad (2.1)$$

The graphical representation of the LDA model is shown in Figure 2.1. The repeated sampling of topics and words is conveniently illustrated using plate notations with the number in the lower right corner referring to the number of replication. Shaded (un-shaded) variables indicate observed (hidden, latent, or unobserved) variables. Arrows indicate conditional dependencies between variables.

2.3.2 Topic Modeling of Text

A key principle in the probabilistic topic modeling approach is to move beyond the standard bag-of-words representation for discrete data collection. For text, the main idea is to capture patterns of word co-occurrences into latent topic variables represented by multinomial distributions. A text document is therefore not only represented as histogram of word occurrences, but also modeled as a mixture of latent topics encoding the patterns of word co-occurrences in the collection. These topics succinctly elucidate the structures in the text collection. Modeling in this way offers the advantage of being able to handle word synonymy and polysemy. Figure 2.2 shows an example of the decomposition of two text documents into their respective mixtures of 100 latent topics. These are news article taken from our News crawler in June 2009. The first text article is on General Motors filing for bankruptcy. The second text article is on a result of a tennis game. The figure also shows three relevant topics represented by the words in the two texts. Selected high probability words are shown to best indicate the meaning of the topics. For example, Topic-28 is likely to be the world economic and financial situation, Topic-47 on legal and judiciary court matters, Topic-16 on the tennis sports. The word court is used in each of the two articles. In the first article, court is assigned to Topic-47. In the second article, court is assigned to Topic-16. Its meaning is correctly disambiguated in both articles.

2.3.3 Topic Modeling of Image/Video

Compared to text, an image contains a much richer array of visual information from its pixel representation. Borrowing the bag-of-word representation from the text domain, a basic image representation used to describe images is the histogram, encoding counts of visual features (e.g. color, texture, shape). Because a color histogram only has limited discriminative power between images, the visual community has in recent years proposed various scale-invariant and location-invariant features [20, 127]. To represent an image, salient image regions are first identified. These can be sampled uniformly given a fixed grid, or detected at various locations and scales with an interest point detector. These regions are then described using a robust, scale-invariant, location-invariant descriptor. Following, these descriptors are then quantized into a fixed set of possible visual words. Visual words can be seen as the approximation of the words in a text document. Figure 2.3 shows a

Figure 2.2: Results of latent modeling of two text documents into a mixture of 100 topics. We show only the three relevant ones (topic number 28, 47, 16, each colored differently). Selected top probability words from each topic are shown. The words in the text document are color-coded to indicate their topic affinity. For the word “court”, its meaning is disambiguated correctly in each text. Best viewed in color.

subset of visual words in two images. Each image can be described as a bag of visual words (BOVW).

Because of possible noise in the quantization process, different image descriptors from different image regions may be quantized into the same visual word. For example, the local image region near to a person with a sky-blue clothing may be visually similar to a patch in the sky region of an image. As a result, it is very likely that these two region descriptors will be quantized to the same visual word. Hence these two image regions will be described by the same visual word. This is analogous to the issue of polysemy in text: the same visual word may refer to two different image contexts [159]. In the same light, we can also see the visual analogy of synonymy, where because of differences in imaging conditions such as illumination and camera viewing angle, the image regions corresponding to the same part of a physical object can be assigned to different visual words.

A possible approach to resolve these visual ambiguities is to decompose the BOVW representation into a mixture of visual-word distributions, similar to what is described in the text domain in the last subsection. Hence, in an image collection, of say, buildings, sky, sea, etc, interesting patterns of visual word co-occurrence underlying these images can be elucidated by multinomial distributions

Figure 2.3: Results of latent modeling of images using a mixture of 50 topics. Images are described by visual words, densely sampled from overlapping grids over the images. Topics are multinomial distribution of these visual words. We show two images, on which visual words from two topics (colored blue and green) are superimposed. We can see that the green topics describe people while blue topics describe the sky regions. Best viewed in color.

over visual words [44, 180, 174]. These are the latent visual topics underlying the image collection. From Figure 2.3, we can immediately see an interesting application of this latent approach. The visual words shown in these images are from two different latent topics (colored blue and green). From the geometric layout of the visual words, these two topics are likely to model the sky region and people region. Both images can now be represented as a vector of the count of the topics assigned to the visual words. In an image retrieval application, given an image with sky and human, we can compute their quantized visual words and infer the latent topic mixtures. Then similar sky-with-human images can be retrieved based on the similarity of the topic mixture.

2.3.4 Other Applications of Topic Models

Traditionally topic models have been mostly applied on well-formed text such as publications and newswire, where competitive document retrieval results have been reported in [170, 37], though there needs to be some modification to the LDA retrieval model. Of late, they are increasingly found applicable on non-conventional document sources such as extracting topic changes and topic identity in multi-party spoken discourse [124] and to identify roles in email and social networks [110].

The application of topic modeling to the visual domain has also received much attention. LDA

was used in [64] to discover visual categories in images like cars, people, cows, etc. An LDA variant that combines spatial constraints into topic modeling is reported in [123] to better retrieve particular objects. The paper in [136] examines simultaneous classification and segmentation of visual objects. LDA is applied on multiple image over-segmentations to retrieve the most probable regions containing salient visual objects. The paper in [58] applies topic modeling to cluster images retrieved from Google Image Search. The main contribution is on automatic learning of object classifiers by harnessing the vast amount of images from Google Image Search. The idea was further developed in [140] for improved classifier precision. The paper in [76] uses LDA to locate recurrent key-places in a movie (e.g. cigarette shelves, room tapestry). Key-places are assumed to be perceptually consistent and recurring in the video. Modeling key-frame as a mixture of LDA topics, groups of recurrent visual words are extracted that correspond to characteristic visual elements of key-places.

2.3.5 Learning a Topic Model

The goal of learning a LDA model is to find the set of latent variables that can best explain the observed words in documents, assuming that the model is capable of generating the text collection. This is a standard Bayesian posterior inference process that involves inferring the probability distribution over words θ associated with each topic, the distribution over topics ϕ for each document, and the topic responsible for generating each word. The hyper-parameters α and β are used as a prior to smooth the distribution over topics ϕ and the distribution over words θ respectively. These hyper-parameters can also be inferred from the observed data. There are two major posterior inference methods: (a) Gibbs sampling [69], and (b) variational methods [89]. We briefly discuss each of these in the following subsections.

2.3.5.1 Bayesian Inference and Gibbs sampling

Given the set of observed words $\mathbf{W} = w_{ij}$, the task of Bayesian inference is to compute the following: (a) the posterior distribution of the latent topic indices $\mathbf{Z} = z_{ij}$, (b) the document mixing proportions ϕ_d for each document $d = 1..N$, and (c) the K topics θ_t for $t = 1..K$. The collapsed Gibbs Sampling is an efficient inference procedure that uses the principle of conjugacy to marginalize and integrate out θ and ϕ . This leaves only the latent variables $\mathbf{Z}$ to be sampled. After the sampler burn-in, we can estimate θ and ϕ given samples of $\mathbf{Z}$.

The Gibbs Sampling inference procedure is computed using pseudo counts. Denote $N_{vdt} = \#\{i : w_{di} = v, z_{di} = t\}$ to be the number of times the word v in document d is assigned to the topic t . Hence, $N_{dt} = \sum_v N_{vdt}$ is the number of times a word in document d has been assigned to topic t , and $N_{vt} = \sum_d N_{vdt}$ is the number of times the word v is assigned to the topic t . Given the current set of all variables minus the variable z_{di} , the conditional probability of z_{di} is then given by:

$$p(z_{di} = t | \mathbf{Z}^{-di}, \mathbf{W}, \alpha, \beta) = \frac{1}{C} a_{dt} b_{vt} \quad (2.2)$$

$$\text{where } a_{dt} = N_{dt}^{-di} + \alpha, \quad (2.3)$$

$$b_{vt} = \frac{N_{vt}^{-di} + \beta}{N_t^{-di} + |W|\beta}, \quad (2.4)$$

$$\text{and } C = \sum_t a_{dt} b_{vt} \text{ is a normalization constant.} \quad (2.5)$$

Note that we use the superscript $-di$ to denote that the corresponding d ith variable is excluded from the pseudo count summation.

The Gibbs Sampling inference procedure proceeds by iteratively drawing a sample for z_{di} according to equation 2.2 for the i th word in document d . Given the sampled value of z_{di} , the pseudo counts N_{dt} , N_t and N_{vt} are updated. Then we estimate θ and ϕ as follows:

$$\theta_{vt} = \frac{N_{vt} + \beta}{N_t + |W|\beta} \quad (2.6)$$

$$\phi_{dt} = \frac{N_{dt} + \alpha}{N_d + K\alpha} \quad (2.7)$$

2.3.5.2 Variational Methods

Variational methods are a class of deterministic algorithms that offer an alternative to performing Bayesian inference in graphical models [89]. They differ from the sampling-based methods in the previous subsection by simplifying the inference process using approximation. Rather than approximating the posterior with samples, variational methods assume a parameterized family of distributions over the same latent structures, and then find the member of that family that is closest to the posterior. Thus, the inference problem is transformed to an optimization problem. This allows variational methods to tap on the advances of optimization techniques, enabling increased practical applications to be developed in probabilistic modeling.

Figure 2.4: Graphical representation of the Link-PLSA-LDA model that jointly models documents and citations [115].

In this thesis, we have based the learning of our joint modeling method on variational principles. More details can be found in Appendix B.

2.3.6 Constraining Topic Formation

The basic LDA topic model can be extended to exploit additional information or structure so as to learn richer and more informative models. In particular, these additional information can act as a constrain or side information to guide the topic formation during learning. These constrains can be generally categorized into: (a) joint modeling of complementary data, (b) modifying the ϕ -side topic formation, and (c) modifying the θ -side word formation. In the following, we discuss each of them.

2.3.6.1 Joint Modeling of Complementary Data

These models extend the LDA by modeling additional observed data beyond the words and documents, such as citation links, document labels, etc. During the learning, the latent topics are forced to simultaneously account for these complementary data and the document text, using the inherent correlations amongst these data. The presence of these additional information guides the formation of better topics, wherein the high probability words are more coherent, and that the topics can better infer new document annotations or predicting links between documents. We look at two representative models of this type.

Figure 2.5: Graphical representation of the Corr-LDA [25].

Documents+Links : There are inter-document links arising from the author information and citations. A topic model that jointly models the generation of document, authors and citations can capture capture the intrinsic relationships amongst these entities, and produce topics that are more coherent, robust and have better predictive power. The Link-PLSA-LDA model [115] is such a model that assumes the generation of each hyperlink in a document is a multinomial sampling of the target document from the topic-specific distribution of the documents.

Image+Text : Correspondence LDA [25] is a joint model of images and their associated caption text. In addition to the standard LDA model wherein each latent topic is associated with a multinomial distribution for generating caption words, each latent topic is now also coupled with a multi-variate Gaussian for generating image patches. The new graphical connections, shown in Figure 2.5, make explicit the correlation between the caption content of an image patch and its visual content. These constrain the topic learning process to produce topics that better capture the image-caption coupling, which yield better performance in annotating new images.

2.3.6.2 Modifying ϕ -side Topic Formation in Documents

In the standard LDA, the document-topic mixture for each document is independently sampled from a Dirichlet prior. This assumes that the topics in each document occur independently. However, our intuition tells us that there exists correlations amongst the topics, e.g. there is a higher likelihood for a document with a topic on “machine learning” to be also associated non-nominally to a topic on “posterior sampling”. ϕ -side-based LDA variants are concerned with capturing a richer representation of the document-topic associations.

In the Correlated Topic Models [26], the Dirichlet prior is replaced by a logistic normal dis-

tribution, allowing a covariance structure that encodes the pairwise correlations between the topic components. This provides a more realistic model of the latent topic structure where the presence of one latent topic may be correlated with the presence of another. In the Labeled LDA [129], each document is associated with a binary vector of labels. Each label is associated with its own special topic which can only be used in documents bearing that label. This constrain allows the model to learn special topics that are strongly associated with their corresponding labels.

2.3.6.3 *Modifying θ -side Word Formation in Topics*

In the basic LDA, words are generated by topics, which are simple multinomials drawn from a Dirichlet prior with hyperparameter β . Models that relax this multinomial word generation assumption may be able to capture the richer linguistic structures of the actual word observations.

An example of this approach is the Hidden Topic Markov model of Gruber *et al* [72]. This models the topics of document words as a Markov chain. It assumes that all words in the same sentence have the same topic, and successive sentences are more likely to have the same topics. Since the topics are hidden, the well-known Hidden Markov Models is used for learning and inference. The authors show that incorporating this dependency enables the formation of better topics that can disambiguate words belong to different topics, and result in better topic boundary detection in sentences.

Another example is the Syntactic Topic Models of Boyd-Graber and Blei [30]. This explicitly models each word in a sentence as generated by a distribution that combines document-specific topic weights and parse-tree-specific syntactic transitions. Specifically, each sentence is represented by a tree of latent topics which generate the words. To encode the topic of a word, a joint distribution is defined over the topic of the word’s parent (as per the syntax parse tree), the usual topic weights for a document, and the nodes’ parents successor weights. This allows the model to learn syntactically relevant topics capturing grammatical regularities in language.

2.4 Bayesian Non-Parametrics

To properly address the clustering of multimedia return results from ambiguous queries, we need to formulate methods with *unlimited* modeling capability. In this section, we review some related development in Bayesian Non-Parametrics that allows for infinite mixture modeling. In subsection 2.4.1,

Figure 2.6: Left: Graphical model of the finite mixture model (G is a Gaussian); Right: Graphical model of the Dirichlet process mixture model. Note the addition of a nonparametric Dirichlet process prior on G , with parameters G_0 and concentration parameter α .

we start off by looking at the traditional problem of modeling with a finite number of mixture components. Then in subsection 2.4.2, we see how this can be extended to an *countably infinite* number of mixture components. We examine Dirichlet Process, the theoretical foundations of infinite modeling, in subsection 2.4.3, before looking in subsection 2.4.4 at the related issue of sharing these infinite pool of mixture components, by using the Hierarchical Dirichlet Process.

2.4.1 Finite Mixture Modeling

Mixture modeling is an effective density estimation method to represent data. In *finite* mixture modeling, we assume that there is a *pre-specified* number K of clusters, each with a parameter $\theta_p, p = 1..K$, from which a set of N observed data $\mathbf{Y} = \{y_1, \dots, y_n\}$ is generated in the following way. For each $n = 1..N$, first choose a cluster c_n according to a cluster probability $P(c_n)$. Then generate y_n according to the corresponding cluster distribution $F(y_n|\theta_{c_n})$.

Casting this finite mixture modeling in a Bayesian framework, we define a distribution $G_0(\theta)$, say Gaussian, over the cluster parameters, so that all θ_p are drawn from G_0 . This is in effect a Gaussian mixture model (see left of Figure 2.6). Let the cluster assignment be $\mathbf{C} = \{c_1, \dots, c_N\}$. Assume each observation is conditionally independent given its cluster assignment, the data likelihood can be obtained by marginalizing over θ : $P(\mathbf{Y}|\mathbf{C}) = \int \left(\prod_{k=1}^K G_0(\theta_k) \prod_{n=1}^N F(y_n|\theta_{c_n}) \right) d\theta$, and the cluster assignments can be calculated from the posterior: $P(\mathbf{C}|\mathbf{Y}) = \frac{P(\mathbf{Y}|\mathbf{C})P(\mathbf{C})}{\sum_{\mathbf{C}} P(\mathbf{Y}|\mathbf{C})P(\mathbf{C})}$. As discussed in an earlier paragraph, the calculation of the posterior is intractable, and approximate methods such as Markov Chain Monte Carlo (MCMC) or variational inference discussed in subsection 2.3.5 can be used.

In the above finite mixture modeling, an important question is: “what is the value of K ?” This

question is mostly addressed by fitting several models to the data, each with a different number K of mixtures, and then selecting one using some model comparison metrics, such as cross-validation. However, such an approach can be limiting, especially if we wish to model new observations from as yet unseen components. Bayesian Non-Parametric (BNP) models provide a different approach to this problem: fit a *single* model that allows its complexity to grow according to the complexity of the data. To do this, K is also treated as a random variable and is inferred during posterior inference.

2.4.2 Dirichlet Process Mixture Models

The Dirichlet Process Mixture Model (DPMM) is a BNP clustering process that allows K to grow with data. K can grow up to *infinity* (that is, an infinite number of clusters), but in practice, only a small K is needed to account for the observed data. It does this by specifying a prior over the infinite clusters $c_{1..∞}$ (see right of Figure 2.6) that favors assigning data to a small number of clusters:

$$P(c_n = k | c_{1..n-1}) \sim \begin{cases} m_k & \text{if } k \leq K \text{ (k is an existing assigned cluster)} \\ \alpha & \text{otherwise (k is a new cluster)} \end{cases} \quad (2.8)$$

where m_k is the number of data assigned to cluster k , K is the number of assigned clusters where $m_K \geq 0$. The parameter α is called the *concentration parameter*. A larger α will produce more clusters (with fewer data assigned per cluster). Our aim is the posterior on cluster parameter θ_n , given data y_n . Using Bayes rules and the prior $P(\theta_n | \theta^{(-n)})$ from equation 2.8, we have:

$$P(\theta_n | \theta^{(-n)}, y_n) \sim P(y_n | \theta_n) P(\theta_n | \theta^{(-n)}) \quad (2.9)$$

$$\text{where the prior is: } P(\theta_n = \theta | \theta^{(-n)}) \sim \alpha G_0(\theta) + \sum_{j \neq n} \delta(\theta - \theta_j) \quad (2.10)$$

Multiplying with the data likelihood, which is simply $F(y_n | \theta_n)$, the posterior is:

$$\begin{aligned} P(\theta_n | \theta^{(-n)}, y_n) &\sim \alpha G_0(\theta_n) F(y_n | \theta_n) + \sum_{j \neq n} F(y_n | \theta_n) \delta(\theta - \theta_j) \\ &= \begin{cases} \theta_j & \text{with probability } F(y_n | \theta_j) \\ \frac{G_0(\theta) F(y_n | \theta)}{\int G_0(\theta) F(y_n | \theta)} & \text{with probability } \alpha \int G_0(\theta) F(y_n | \theta) \end{cases} \end{aligned} \quad (2.11)$$

Again, the above posterior is intractable, and MCMC-sampling-based methods such as Gibbs sampling need to be applied to yield an approximate solution [116].

2.4.3 Dirichlet Process

The foundational for the DPMM lies in the ability to model a discrete random variable of unknown cardinality. In the previous subsection, we have seen an example of a clustering solution whose number of clusters can grow with the data. In the case of LDA, an equivalent problem is how many topics K should a given set of texts has? The heart of the DPMM is the stochastic Dirichlet process (DP). In this section, we discuss the DP based on the Dirichlet Distribution.

2.4.3.1 Dirichlet Distribution

Let $\mathbf{X} = \{X_1, X_2, \dots, X_N\}$ be a random probability mass function, $X_i \geq 0, \forall i = 1, \dots, N$, and $\sum_{i=1}^N X_i = 1$. Let $\alpha = \{\alpha_1, \alpha_2, \dots, \alpha_N\}$ with $\alpha_i > 0, \forall i$, and let $\alpha_0 = \sum_{i=1}^N \alpha_i$. Then $\mathbf{X}$ is said to be Dirichlet-distributed with parameter α . This is denoted by $\mathbf{X} \sim \text{Dir}(\alpha)$. The density of $\mathbf{X}$ is:

$$f(\mathbf{X}; \alpha) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^N \Gamma(\alpha_i)} \prod_{i=1}^N X_i^{\alpha_i - 1} \quad (2.12)$$

Figure 2.7 shows plots of the Dirichlet distribution density over the two-dimensional simplex in $\mathbb{R}^3$ for some values of the parameter value α .

The Dirichlet distribution is the conjugate prior for the Multinomial distribution. To understand the multinomial distribution, suppose we have n independent events, and for each event, the probability of outcome i is y_i . Then the multinomial distribution specifies the probability that the outcome i occurs y_i times. A multinomial distribution is parameterized by n and $\mathbf{Y} = [y_1, y_2, \dots, y_N]$ ⁵:

$$f(x_1, x_2, \dots, x_N | n, y_1, y_2, \dots, y_N) = \frac{n!}{x_1! x_2! \dots x_N!} \prod_{i=1}^N y_i^{x_i}. \quad (2.13)$$

In light of the conjugate relationship with the multinomial events $\mathbf{Y} = [y_1, y_2, \dots, y_N]$, we can now view the parameter vector α of the Dirichlet distribution as a discrete distribution G_0 over a set of

⁵Note that when $N = 2$, the multinomial distribution reduces to the binomial distribution.

Figure 2.7: Figure taken from [63]: Density plots (blue = low, red = high) for the Dirichlet distribution over the probability simplex in $\mathbb{R}^3$ for various values of the parameter α . When $\alpha = [c, c, c]$ for some $c > 0$, the density is symmetric about the uniform pmf (top-left). When $0 < c < 1$, there are sharp peaks of density almost at the vertices of the simplex and the density is very small away from the vertices. The top-right plot shows an example of this case for $\alpha = [0.1, 0.1, 0.1]$, one sees only blue (low density) because all of the density is converged near the edges of the simplex. When $c > 1$, the density becomes concentrated in the center of the simplex (bottom-left). Finally, if α is not a constant vector, the density is not symmetric (bottom-right).

categories $k \in [1, N]$ scaled by a scalar value α_s :

$$\text{Dir}(\alpha) \equiv \text{Dir}(\alpha_s G_0), \text{ where } G_0 = \sum_{n=1}^N \pi_{0n} \delta(x - n), \text{ and } \sum_{n=1}^N \pi_{0n} = 1 \quad (2.14)$$

$\delta(x)$ is discrete Dirac function that has a non-zero unit integral only at $x = 0$, and G_0 is called a base distribution of the Dirichlet distribution. Hence, all samples drawn from this base distribution will have the same form: $G \sim \text{Dir}(\alpha_s G_0) \Rightarrow G = \sum_{n=1}^N \pi_i \delta(x - n)$, and $\sum_{n=1}^N \pi_n = 1$.

2.4.3.2 Dirichlet Process as an extention of the Dirichlet Distribution

The DP extends the Dirichlet distribution by generalizing the base distribution from the discrete G_0 to *any* (including non-discrete) distribution H_0 :

$$H \sim \text{DP}(\alpha_s H_0) \Rightarrow H = \sum_{n=1}^{\infty} \pi_n \delta(x - \theta_n), \text{ where } \sum_{n=1}^{\infty} \pi_n = 1. \quad (2.15)$$

In other words, there can be an infinite number of points θ_n within the support set of H_0 , over which H is a discrete distribution. Suppose we partition the (infinite) θ_n points into finitely many disjoint subsets $S_1, S_2, \dots, S_J, J < \infty$, and we sum up the weights π_n for all points within each subset. Denote the weight sum of subset j by $H(S_j) = \sum_{\theta_n \in S_j} \pi_n$, and the subset integral over the base distribution by $H_0(S_j)$, then the distribution of the subset weight sum has a Dirichlet distribution as follows:

$$H(S_1), H(S_2), \dots, H(S_J) \sim \text{Dir}(\alpha_s H_0(S_1), \alpha_s H_0(S_2), \dots, \alpha_s H_0(S_J)). \quad (2.16)$$

To calculate the posterior of a DP, we first recall that because the Dirichlet is conjugate to Multinomial, given a set of multinomial samples $\mathbf{X}$, the posterior of a dirichlet is simply updated using pseudo-counting: that is, if $\mathbf{X} \sim \text{Multi}(n, \mathbf{x})$ and $Q \sim \text{Dir}(\alpha)$, then $P(Q|X = \mathbf{x}) \propto \text{Dir}(\alpha + \mathbf{x})$. In the DP, this generalizes to pseudo-counting the n new observation θ_{new} to the scaled base distribution $\alpha_s H_0$, where the i th data $\theta_{\text{new},i}$ is accordingly added into the respective subset weight sum:

$$\begin{aligned} H(S_1), \dots, H(S_J) | \theta_{\text{new},i} &\sim \text{Dir}(\alpha_s H_0(S_1) + n_1, \dots, \alpha_s H_0(S_J) + n_J), \text{ where } n_j = \sum_{i=1}^n \mathbb{1}_{S_j}(\theta_{\text{new},i}) \\ \Rightarrow H | \theta_{\text{new}} &\sim \text{DP}(\alpha_s H_0 + \sum_{i=1}^n \delta(x - \theta_{\text{new},i})). \end{aligned} \quad (2.17)$$

To calculate the predictive distribution of a DP, we similarly recall the predictive distribution of a Dirichlet distribution: given n previous data points $x_{i=1}^n$ from a distribution $G \sim \text{Dir}(\alpha_s G_0)$, the probability of a new value x is: $p(x = k | x_i, \alpha_s G_0) = \frac{n_k + \alpha_s G_0(k)}{n + \alpha_s}$. To generalize this to the DP, consider a sample $H \sim \text{DP}(\alpha_s H_0)$, and n points $\{\theta_i\}_{i=1}^n \sim H$. The probability of choosing a subset S_j from $H | \{\theta_i\}$ follows a Dirichlet distribution. Using the posterior of a DP in equation 2.17, the subset probability mass is obtained using expectation over the DP posterior:

$$\begin{aligned} p(\theta \in S_j | \{\theta_i\}, \alpha_s H_0) &= \frac{\alpha_s H_0(S_j)}{n + \alpha_s} + \frac{\sum_i \mathbb{1}_{S_j}(\theta_i)}{n + \alpha_s} \\ \Rightarrow \theta | \{\theta_i\}, \alpha_s H_0 &\sim \frac{\alpha_s H_0 + \sum_{i=1}^{K^*} n_k \delta(\theta - \theta_i)}{n + \alpha_s} \end{aligned} \quad (2.18)$$

where K^* may increase as new data are observed: either an existing sample has already been seen (second term in the numerator), or a new θ_{K^*+1} is drawn from $\alpha_s H_0$ (the first term in the numerator), followed by $K^* = K^* + 1$.

2.4.3.3 Review: Dirichlet Process as a prior over the infinite sample space

Recall our objective of modeling a component of unknown cardinality. This means that we need to deal with a probability distribution over an infinite sample space, which obviously is unmanageable. To address this, the DP tries to restrict the class of distributions under consideration to a more manageable set: discrete probability distributions of finite subsets of the infinite sample space. Note that in so doing, the DP is *still* dealing with a probability distribution over the infinite sample space, where each discretized subset is written as an infinite sum of weighted indicator functions.

The appeal of this approach can be seen with the following example. Suppose we try to sample a distribution of favourite color by surveying a group of people. There is possibly an infinite number of color choices and it is not good to limit the available color choices to some finite common colors, e.g. “red”, “green”. Here, being able to model the infinite color space is desirable. DP achieves this by assigning a finite “atom” probability to the observed data in the sample space. For example, as an favourite color is mentioned, it assigns a non-zero probability to it. Over time, common favourite color names such “red”, “green”, etc, will receive increased probability, while other color names that are difficult to name would be assigned zero probability. There is a “clustering” or “rich-gets-richer” effect in the DP that it enforces the behavior of samples that tend to repeat existing sample observations: a new sample will have a higher chance of being the same as those are already observed. This corresponds well to our intuition that the higher probability samples are the popular color names. It is in this sense that the DP is a prior over the infinite sample space.

2.4.4 Hierarchical Dirichlet Process

In many real-world applications, there are natural structures in the data, wherein the data are subdivided into groups, and within each group, there are an unknown number of clusters that embody the latent structures in the group. Often times, these clusters are shared among the groups. Examples include the modeling of shared latent topics that underly sets of documents, and modeling of shared salient visual features for object and scene recognition in computer vision. In these problem settings, the goal is beyond discovering the clusters in each group which, as seen from the previous subsections, can be modeled using the nonparametric Bayesian approach based on DPMM. Here, the problem is that how to allow these clusters to be *shared* among the groups.

A naive approach is to apply the DP approach to sample the topic parameter for every observed word. Hence, as more words are seen, new topics may be created as needed. However, the question is whether these new topics can be shared: if we sample a topic parameter $\theta \sim G_j$, where G_j is a group-specific DP sampled from a common distribution G_0 : $G_j \sim \text{DP}(\alpha_s G_0)$, then it will not work if G_0 is a continuous distribution, since each drawing of G_j from G_0 will be unique, and hence it is not possible to share θ among different G_j [152].

The solution is therefore to force G_0 to be a discrete function. Then a draw G_j from G_0 will comprise of repeated atoms. At the same time, we want G_0 to have as broad a support as possible. This requirement is met if G_0 itself is a draw from a $\text{DP}(\gamma, H)$. Hence, atoms θ are shared among the multiple DPs, resulting in the desired sharing of atoms among groups.

$$G_0|H \sim \text{DP}(\gamma, H) \quad (2.19)$$

$$G_j|G_0 \sim \text{DP}(\alpha_0, G_0) \quad (2.20)$$

A hierarchical Dirichlet process can be used as the prior distribution over the latent factors for grouped data. For each group j , let $\{\theta_{ji}\}_{i=1}^{n_j}$ be random variables distributed as G_j . Each θ_{ji} is a latent factor corresponding to the observed data point x_{ji} . The likelihood is given by:

$$\theta_{ji}|G_j \sim G_j \quad (2.21)$$

$$x_{ji}|\theta_{ji} \sim F(\theta_{ji}) \quad (2.22)$$

2.5 Search Disambiguation

There has been no concerted effort in the literature to address search disambiguation. In the following, we highlight the sporadic ad-hoc related work in mining query intent, search diversification, sample diversification in active learning, sense disambiguation, and document/query expansion.

2.5.1 Mining Query Intent

It is a well-known fact that web queries are short and imprecise [17]. Therefore, to accurately identify the user's intent behind an ambiguous web query becomes one of the most challenging tasks for the

modern information retrieval systems. The key strategy to mine user intent is to utilize a large query log. Broder [34] first classified web queries into three categories (informational, transactional, and navigational) that can be refined and used to characterize user intent [133, 86, 101]. Interestingly, two recent papers on query intent analysis have focused on refining Broder’s seminal query taxonomy to include the following new additions: (a). a *polysemy facet* to indicate whether the query is ambiguous or not [120], and (b). a *faceted topic* to indicate the ODP ⁶ topic class to which the query would best belong [10, 36]. In the former, the authors wrote of the class of polysemous queries in section 2.1 on page 3: “*General and polysemous queries as defined in our Ambiguity facet are too general and underspecified. An appropriate disambiguation or query expansion should be used, either in a relevance feedback framework (in which the user chooses between alternatives) is beyond the scope of the query intent mining.*” For the latter, the authors saw the need for a faceted topic assignment to queries, so as to achieve a better user intent match.

2.5.2 Search Diversification

In recent years, the bulk of research focus in image and video search has been on improving search precision, while the issue of search disambiguation and diversification has been relatively neglected. Many techniques such as [79, 41] aim to improve precision by reordering search results according to their visual and contextual similarity to results with high estimated relevance. This results in pushing clusters of visually and contextually similar shots up to higher ranks. However, the net effect is that most search results at the top tend to be similar. For a user who is looking for a wider range of background and story setting of the queried person or object, the result list appears poor owing to its lack of diversity. In many situations, users may be actually looking for a particular context in which his queried person or object appears. Hence, while the system would have technically achieved a high search precision, its returned list of highly similar results will be of little use to the user. This issue has also been perceived by other researchers in movie recommendation, leading it to be coined as the portfolio effect [6]. The negative impact of this effect to users is related to the law of diminishing marginal returns commonly known in economics. It explains the decreasing level of enjoyment over a product when it is repeatedly consumed over and over again.

In web search, presenting search results in topical clusters can greatly alleviate the problem of

⁶The Open Directory Project at www.dmoz.org/

information overload and unwieldy search results. This has led to both commercial venture [4] and open source effort [1] that consider mainly server-side content clustering. In [182], Zeigler introduce an intra-list similarity metric to assess topical diversity of recommendation and a client-side result diversification algorithm to balance between accuracy and diversity of recommendation lists. They found that while accuracy has been compromised, overall user satisfaction is improved. The paper in [128] capitalizes on a large user query log to diversify results by mining information from query reformulation. Zhang *et al* [178] propose a ranking scheme that optimizes two metrics: diversity and information richness. A similar re-ranking approach is applied in [147] to achieve diversity in image retrieval. The notion of topic richness is defined to enrich topic coverage in retrieval results, while maintaining acceptable retrieval performance.

Graph-based diversification approaches provide sound theoretical basis for result diversification. In [181], the authors defined a random walk with absorbing states over sentences for text summarization. Sentences which have been ranked so far become absorbing states. These absorbing states drag down the importance of similar un-ranked states, thus encouraging diversity in sentence selection. The paper in [87] cast the image-ranking problem into the task of identifying authority nodes on an inferred visual similarity graph. By analyzing the visual link structure using an iterative PageRank-based procedure, a numerical weight is assigned to each image measuring its relative importance to the other images being considered. The authority nodes form the cluster-centers of similar images, and facilitate diverse image presentations to search.

The TREC Novelty Track [73, 146] and ImageCLEF Photographic Retrieval Task [8] are international benchmarking forums to encourage research in finding novel sentences (images) in a set of relevant sentences (images). The task resembles the current work in that the results must be relevant as well as novel.

2.5.3 Sample Diversification in Active Learning

In recent years, related notions of disambiguation and diversification has also been pursued in the Active Learning research community. In active learning, the key task is to select the most useful training samples from an unlabeled data pool, in order to minimize the number of training instances needed to learn the separation decision boundary. During a particular iteration, based on the current pool of labeled training samples L , the unlabeled data samples that are nearest to the current sepa-

rating hyperplane are the most ambiguous with respect to the classifier that is trained on L ⁷. Hence, a simple approach, due to Tong and Chang [154], is to attempt to achieve maximum disambiguation by selecting these most ambiguous samples.

However, a better approach, due to Brinker [33], is to further introduce a diversity metric into the sample selection process. Called the *angle diversity* algorithm, the main idea is to not only select samples that are close to the separating hyperplane, these samples should also be diversely different from one another. This latter aspect mirrors the selection of diversely different samples in the MMR ranking algorithm, introduced in Section 2.1. An angular distance metric is used to measure the diversity of the data samples. Specifically, the angle between the unlabeled instance x_i and the current labeled pool L is defined as the maximal angle a_i from x_i to any instance $x_j \in L$. The distance d_i from x_i to the separating hyperplane is also measured. Similar to the MMR algorithm, a parameter α is then introduced to weight the contribution of d_i and a_i . The angle diversity algorithm then chooses as the next training sample the sample that has the smallest weighted sum. Brinker [33] has shown good performance using this diversification approach.

2.5.4 Sense Disambiguation

The need to disambiguate information exists across many domains. In the text community, word sense disambiguation (WSD) is the task of determining the meaning of an ambiguous word in its context. Effective WSD can improve systems for tasks such as information retrieval, machine translation, and summarization.

In [148], person name appearances in web pages and scientific documents are disambiguated by a probabilistic topic modeling approach. An additional observed variable is augmented into LDA to learn the topic distribution with regard to person-names and words. By treating these distributions as feature vectors and clustering them, the method is able to discern whether two different names correspond to the same person, based on the similarity of the topic word distributions. In a similar fashion, the method in [31] embeds LDA with a hidden variable denoting the sense of a word to simultaneously disambiguating a corpus and learning the domains in which to consider each word.

The paper in [91] introduces the interesting concept of visual senses of images. Given a pol-

⁷Seen in another perspective, these most ambiguous samples are also the most informative samples to achieve a better separation of the training samples, and so ultimately improve the classification performance

polysemous keyword, images returned from a search engine will be polysemous. The idea is then to learn the primary visual sense model of these polysemous images by utilizing words from a dictionary (WORDNET [5]) and probabilistically comparing them with the latent topics mined from the surrounding text of all images. We further extend this idea in Chapter 4.

2.5.5 Document and Query Expansion

A query refinement approach can also be used to help resolve search ambiguity. The main idea behind these methods is to analyze the original queries and attempt to generate alternative queries that better express the underlying information need. The set of alternative queries is usually called the *expanded* queries and this class of query refinement methods is called *expansion* methods. The expansion approach can be applied at the query and document level.

Current automatic expansion techniques of query and document can be generally categorized into global analysis and local analysis. An expansion method based on global analysis attempts to build a knowledge resource (e.g. thesaurus) to assist users to reformulate their queries [125]. Such a knowledge resource can be computed by analyzing the statistics of term co-occurrences in the documents. From this, a user will be able to obtain synonyms or related terms given a user query. Another group of techniques for query expansion is based on local analysis. The idea is to extract expansion terms from a subset of the initial retrieval results [173]. This subset can be determined by the user (i.e. relevance judgments), or automatically by the system (i.e., just select the top-ranked documents). The latter approach is also known as pseudo relevance feedback (PRF), and assumes that the top-ranked documents are relevant.

Compared to global thesaurus-based expansion, local analysis is more query-oriented. However, retrieval performance of local analysis-based expansion can suffer if the precision of top-ranked documents is low. Furthermore, using unrelated expansion terms from irrelevant results may result in topic drift. To remedy these problems, several authors have deployed user query logs. Huang *et al* [80] focus on user reformulations of queries differing only in one segment. To use these segments as candidate expansion terms, they need also to be repeated in similar sessions. Cut *et al* [54] present a probabilistic approach to extract expansion terms from click-through data. The main idea is that if a term in a query often leads to the selection of documents containing another term, then there is a strong relationship between the two terms.

Chapter 3

A FACETED TOPIC RETRIEVAL MODEL FOR MULTIMEDIA CONTENT

3.1 Introduction

A key challenge for modern Information Retrieval (IR) systems is dealing with ambiguous queries. Studies have shown that most users pose short and broad queries, expecting the IR system to elucidate their possible interpretations [171]. The query “apple” is a classic example of an ambiguous query. In the absence of any disambiguating information about the user’s query intention, an IR system may best return a mix of documents discussing the fruit, the company and the product.

To better address the aforementioned problem, a retrieval system should be designed to account for the breadth of available information and to resolve any ambiguity inherent in the query [49, 176, 40]. In this chapter, we define a Faceted Topic Retrieval (FTR) model for multimedia content, and augment conventional retrieval with both *novelty* and *diversity* considerations.

In the remaining of this chapter, we first present some key concepts in Section 3.2, before defining our problem formally in Section 3.3. We discuss the evaluation framework for faceted topic retrieval in Section 3.4. Then, in Section 3.5, we begin with motivating our approach from a general probabilistic framework of faceted topic retrieval. This is then followed by the details on our proposed Faceted

Document Model (FDM) in Section 3.6, where we augment the FDM with two novel modeling abilities: (a) joint modeling of the complementary information in a multimedia content; and (b) modeling the varying levels of ambiguity in a query topic. Lastly, in Section 3.7, we discuss the Faceted Retrieval Model (FRM).

3.2 Preliminaries and Definitions

In this section, we first define the key concepts in the problem of faceted topic retrieval as follows:

Definition 1 (Word): We define a *word* as an uni-modal feature of a multimedia content. The set of unique words representing this particular uni-modal stream is called the *vocabulary* of this uni-modal stream. As a multimedia content typically comprises of the text, visual and audio stream, to represent each of these domains, we use a separate set of uni-modal words, each of which is also called a vocabulary in that domain. We use w^t, w^v and w^a to denote a word in the text, visual and audio domain, and V^t, V^v and V^a to denote the corresponding vocabulary respectively. The multi-modal vocabulary of the document is $V^d = \{V^t, V^v, V^a\}$. We also generalize this notation to any *auxiliary* information related to the multimedia content. For example, a web captioned image has visual w^v and text w^t as its intrinsic representation. But another source of auxiliary information of the image is its surround web text, or the anchor text of its incoming links. This can be represented as $w^{t'}$ with vocabulary $V^{t'}$.

Definition 2 (Document): We define a *document* d in a multimedia collection as a sequence of words. To make explicit the nature of its multi-modal representation, we denote our multi-stream document as $d = (\mathbf{w}^t, \mathbf{w}^v, \mathbf{w}^a)$, where $\mathbf{w}^t = \{w_1^t, w_2^t, \dots, w_{N_t}^t\}$ is the text-word vector of dimension $N_t = |V^t|$, $\mathbf{w}^v = \{w_1^v, w_2^v, \dots, w_{N_v}^v\}$ is the visual-word vector of dimension $N_v = |V^v|$, $\mathbf{w}^a = \{w_1^a, w_2^a, \dots, w_{N_a}^a\}$ is the audio-word vector of dimension $N_a = |V^a|$, and the combined feature vector dimension is $N_d = |V^d|$.

Definition 3 (Corpus or Collection): A *corpus* or *collection* D of J multimedia documents is denoted as $D = \{\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_J\}$. Following most existing work on topic modeling, we do not model the order of words. Thus each document in D is simply treated as a bag of words.

Definition 4 (Facet): A *facet* θ in the multimedia corpus D is a multinomial distribution of words. A facet is synonymous to the topic in the probabilistic topic model framework (see also

Figure 2.1), which can also be viewed as a subtopic or theme in the document. We assume that D covers a number of such facets. If we assume that there are K facets in D , the set of facets is denoted by: $\Theta = \{\theta_1, \theta_2, \dots, \theta_K\}$, where each θ is a probabilistic distribution of words $\{p(w|\theta)\}_{w \in V^d}$ that represents a semantically coherent topic. By coherence, we mean that by looking at the high probability words in a facet, one can often understand the theme the topic has captured. For example, a facet about “Finance” would assign a high probability to words like “money”, “bank” and “economy”.

3.3 Problem Definition

We now define the specific problem we want to address. The problem of Faceted Topic Retrieval (FTR) is to return a set of documents that cover the different *facets* of an information need. Informally, the notion of facets can be interpreted to be any binary property of a document that represents a fact or a topic that is contained in the information need. For example, given the query “War on Terror in Iraq”, the facets may include “War in Fallujah”, “Military Strategy in Pentagon”, “Bin Laden Tape”, etc. The facets of a query can be contained in one or more documents, and a document can contain one or more facets. A document is deemed relevant to the query if it contains any of the facets. There are two sub-problems in FTR:

1. to develop a *Faceted Document Model* (FDM) that given a set of documents D , hypothesizes the set of facets in D ;
2. to develop a *Faceted Retrieval Model* (FRM) that finds the smallest set of documents that cover all of the facets.

Before we dive into the details of our methods, in the next Section, we shall first discuss how our faceted retrieval model can be evaluated.

3.4 Evaluating Faceted Tropic Retrieval

The main performance criteria for a faceted topic retrieval engine is how well it is able to retrieve all the facets that are present in the corpus. Zhai *et al* [176] first proposed an evaluation metric for the subtopic retrieval problem that has been adopted by the diversity retrieval community [121, 40].

Hence, we follow the same measures for our faceted topic retrieval task. This evaluation measure is called the S-recall and S-precision.

3.4.1 S-recall

Suppose we have a set of M facets $\mathbf{F} = \{F_1, F_2, \dots, F_M\}$, and a collection of J documents $D = \{d_1, d_2, \dots, d_J\}$. Each document $d_j, j = 1..J$ is judged as to whether they contain each of the facets $F_m, m = 1..M$. The S-recall at rank k is defined as:

$$\text{S-rec}@k = \frac{1}{M} \sum_{m=1}^M \mathbb{1}_{\text{top}_k(D)}(F_m) \quad (3.1)$$

where top_k denotes the Union set of all documents up to the rank k :

$$\text{top}_k(D) = \bigcup_{\{j=1..k\}} d_j \quad (3.2)$$

and the indicator function $\mathbb{1}_{\text{top}_k(D)}(F_m)$ returns 1 if F_m is present in any of the top- k -ranked documents. At any fixed k , S-rec@ k rewards returning a list of k documents that contain as many of the M facets as possible. The maximum value of S-recall at a particular rank k is the number of facets that can be found in the top- k documents. At the minimum rank of k_{min} , at which all M facets are contained in the top- k_{min} documents, we have the perfect recall of 1. At this rank k_{min} , the value of the S-recall is denoted simply by $S-rec$. Note that a faceted topic retrieval is arguably best assessed by the number k_{min} , the smaller the better, since this means that the user only needs to browse over a smaller number of documents to retrieve all the facets related to the information need. Hence, $S-rec$ is the most natural measure to evaluate a faceted topic retrieval system.

3.4.2 S-precision

While S-recall measures the ability of the system to find facets, Zhai et al define S-precision to assess whether the system is also ranking them well or not. The S-precision at rank k , denoted by $S-prec@k$, is analogous to the traditional precision metric at rank k , which measures the number of relevant documents r found in the top- k retrieved documents, computed simply by r/k . In the case of $S-prec@k$, instead of r , the numerator is now the minimum rank k' at which the current

value of $S-rec@k$ could *possibly* be achieved. That is, suppose the current retrieval system is called *curr*. Given the current rank k , it has a value of $S-rec_{curr}@k$. However, there are two ways that this result may not be the best compared to that of a perfect faceted topic retrieval system. First, at the current rank k , the $S-rec_{perfect}@k$ of the perfect system may be higher. Or at the given $S-rec_{curr}@k$ of the current retrieval system, the perfect system can achieve it at a lower k' . Hence, the retrieval precision of the current system can be assessed against a perfect system by any of the above two ways. We select the latter measure to define S-precision, i.e. S-precision is taken to be k'/k .

Note that this variant of precision has the same properties as traditional precision: it ranges from 0 to 1; it is greater when more unique facets have been retrieved; it approaches zero as $k \rightarrow \infty$.

3.5 A Probabilistic Framework for Faceted Topic Retrieval

Our discussion of the evaluation framework for faceted topic retrieval in Section 3.4 suggests a probabilistic set-based ranking approach for faceted topic retrieval. The goal is to retrieve the smallest *set* of documents that contains all the facets with the highest probability. Given a set of attested M facets $F = \{F_1, F_2, \dots, F_M\}$ in a collection of J documents $D = \{d_1, d_2, \dots, d_J\}$, every document $d_j, \forall j = 1..J$ has some conditional probability of containing any facet $F_m, \forall m = 1..M$, denoted by $P(F_m|d_j)$.

Suppose we have a subset $D' \subset D$. Denote the probability that D' contains F as $P(F \in D')$. Clearly, $P(F \in D) = 1$, since all facets $F_m, m = 1..M$ are contained in the document collection D . In a retrieval setting, it is of course not practical to return the entire document collection D . It is more desirable to find the *smallest* subset D' that contains all $F_m \in F$. On the other hand, the problem can also be framed as: maximize $P(F \in D')$ given that D' must be of a certain size.

3.5.1 Faceted Retrieval as an Optimization Problem

The proper formulation of $P(F \in D')$ is the core of our *set*-based probabilistic framework. We make use of the following two key observations in our formulation: (a) for any $F_m \in F$, F_m can be in any $d_j \in D'$; (b) it is not required that F_m be in *all* the d_j , only in at least one. With these observations,

and applying the sum-rule of probability, we have:

$$\begin{aligned}
P(F_m \in \{d_1, d_2\}) &= P(F_m \in d_1 \cup F_m \in d_2) \\
&= P(F_m \in d_1) + P(F_m \in d_2) - P(F_m \in d_1 \cap F_m \in d_2) \\
&= P(F_m \in d_1) + P(F_m \in d_2) - P(F_m \in d_1)P(F_m \in d_2) \\
&= -(-P(F_m \in d_1) - P(F_m \in d_2) + P(F_m \in d_1)P(F_m \in d_2)) \\
&= 1 - (1 - P(F_m \in d_1))(1 - P(F_m \in d_2))
\end{aligned} \tag{3.3}$$

The third line in the above results from assuming that facets are present in documents independently¹. Hence, the probability that a facet F_m is present in *at least* one document in D' is:

$$P(F_m \in D') = 1 - \prod_{d_j \in D'} (1 - P(F_m \in d_j)) \tag{3.4}$$

Finally, the probability that all facets $F_m \in F$ are contained in D' is:

$$\begin{aligned}
P(F \in D') &= \prod_{m=1}^M P(F_m \in D') \\
&= \prod_{m=1}^M \left\{ 1 - \prod_{d_j \in D'} (1 - P(F_m \in d_j)) \right\}
\end{aligned} \tag{3.5}$$

Note that the problem of maximizing $S\text{-}rec@k$ is equivalent to maximizing $P(F \in D)$. The latter can be resolved by finding a subset $D' \subset D, |D'| = k$ that maximizes $P(F \in D)$. This is related to the issue of faceted *retrieval*, and forms the second sub-problem of our Faceted Topic Retrieval Model – the problem of Faceted Retrieval Model. We shall reserve the discussion of this problem until a later section in Section 3.7. Instead, we first turn our attention to the first sub-problem of our Faceted Topic Retrieval Model – the problem of hypothesizing the set of facets F in the document collection D . To achieve this, we need documents to be represented by latent facets. In the next section, we propose such a Faceted Document Model.

¹In the spirit of the Naive Bayes, this independence assumption is admittedly made mainly for computational convenience. However, note the modular nature of our formulation in that faceted retrieval is separate from facet hypothesizing. This allows faceted retrieval to capitalize on technological advances in the hypothesis of (independent) facets in documents, such as the development of the Correlated Topic Model in [26]. We revisit this issue again in Section 5.4.

3.6 Faceted Document Model

In Faceted Topic Retrieval, the Faceted Document Model (FDM) is a model for multimedia document so as to facilitate the hypothesizing of facets in a multimedia document, and for estimating document-facet probabilities. The goal is that given a document collection and a short query, the FDM is to produce some hypothetical set of facets existing in the document collection so that documents can be scored with a relevance to the query. There are clearly many ways that this can be achieved, e.g. soft clustering [83], relevance modeling [97], key-phrase mining [157], topic modeling [111], etc. Some of these approaches may involve supervision, while others are unsupervised. Given that multimedia document collections are typically huge in size, an unsupervised approach is generally preferred due to the difficulty in collecting adequate training examples.

3.6.1 A Probabilistic Topic Modeling Approach

In this section, we propose a faceted document model based on the Latent Dirichlet Allocation (LDA) of Blei and Jordan [25], an instance of the general class of probabilistic topic models. Compared to the conventional approach that extracts key-words or key-phrases as document facet, (see the discussion in Section 2.2.3 of Chapter 2 on modeling facets), our LDA-based approach builds a facet model $\Theta = \{\theta_1, \theta_2, \dots, \theta_M\}$, where each facet θ_m is a probabilistic distribution of words $P(w|\theta)$.

3.6.1.1 Learning Facets

Different facets θ_m will encode the probability of words differently according to $P(w|\theta_m)$, hence capturing different vocabularies amongst the high probability words, and ultimately developing a particular facet-word coverage that is semantically meaningful. For example, a facet $P(w|\theta_i)$ on “sports” may assign high probability to frequently co-occurring words like “tennis”, “football”, while a facet $P(w|\theta_j)$ on “economy” may assign high probability to other related words. The basis for comparing facets is hence built upon comparing between probability distributions of words, where $P(d|\theta) = \prod_{w \in d} P(w|\theta)$, for which many robust distance measures are available, such as the KL-divergence. Further, the intuition is then that given a document d , if $P(d|\theta_i) > P(d|\theta_j)$, it means that d is more likely to contain words related to “sports” than to “economy”.

3.6.1.2 Estimating Document-Facet Probabilities

In the probabilistic topic modeling approach, the facets can be naturally associated with the θ -side topic-word distributions $P(w|z_m, \theta)$ (see Figure 2.1). As discussed in Section 2.3.5, these topic-word distributions can be learned using Gibbs Sampling [69] or Variational Methods [25]. Upon learning convergence, we can associate each facet $F_m, m = 1..K$ to one of the K topic-word distributions: $P(w|F_m) \equiv P(w|z_m, \theta), m = 1..K$. Then given a document d , we can find $P(d|F_m) = \prod_{w \in d} P(w|F_m) = P(w|z_m, \theta)$ and $P(F_m|d) \propto P(d|F_m)P(F_m)$.

3.6.2 Multi-modal Modeling of Complementary Data Streams

We introduce the first of our two novel modeling capabilities that we incorporate into the Faceted Document Model. This relates to the multi-modal modeling of the accompanying data streams in a multimedia content. The motivation for doing this is to better ground the facet (topic) formation during the probabilistic learning process. This is achieved by exploiting the fact that these data streams, while separately encoded, are information that is complementary to one another. Hence, by jointly modeling these information, the hope is to derive facets (topics) that are more coherent.

Another motivation is to produce facets that are more robust to noise in the multimedia streams. As discussed in Chapter 1, the feature extraction process in a multimedia content is less than perfect, and the resulting noisy features actually contribute to part of the Search Ambiguity problem. However, from the perennial wisdom that a picture is worth a thousand words, information from one feature type (e.g. visual) can often serve as a useful description for the other types (e.g. automatic speech transcripts) [25, 13]

3.6.2.1 The Challenges

Despite proven success of mining semantic topics (facets) in text collections, the use of probabilistic topic models, such as LDA, for faceted topic retrieval of multimedia content poses some new challenges. For example, in the case of news video, naively feeding the noisy speech transcripts (usually obtained from an Automatic Speech Recognition, or ASR engine, where typical word error rates are above 20%) as textual input to LDA will likely not yield good results. Note that the majority of the successful application of LDA are reported on clean text (e.g. newswire and publications). Apart

from the few sporadic work in [37, 124], the utility of LDA to noisy text source remains unclear.

Another challenge relates to the quality of LDA output: LDA often produces word distributions that are coarse, with no apparent meaning amongst high probability words. This can degrade the quality of detected facets and undermine faceted retrieval. The common approach to deal with this problem is to introduce side-information into the modeling [25, 85].

3.6.2.2 Our Joint Modeling Approach

In this thesis, we adopt a similar approach to constrain facet development. For example, in the case of analyzing facets in news video, we use the visual information in the shot keyframes as an auxiliary data source to jointly model alongside the noisy text transcript.

Without loss of generality for any multimedia content, we consider the joint modeling of two modalities, namely speech and shot keyframes in news video. There are two motivating intuitions: First, video footages are often repeated for similar or related news stories, and hence are highly correlated with the spoken (ASR) words. Second, different facets of a query may use different sets of words, and the same set of recurring visual shots becomes a “bridge” between these words, allowing them to be learned as distinct facets. To compute recurring visual shots, we use the Near Duplicate Image (NDI) detection method of Chum *et al* [45]. Our multimodal learning procedure is then run to jointly account for the NDI shots and the ASR words as distinct but correlated sets of observations.

Figure 3.1 illustrates the graphical representation of our joint model. Note that while our joint model shares many similarities with the Corr-LDA of Blei *et al* [25] (see Figure 2.5), there are two important differences:

1. We model the generation of *both* text-words w and visual-words v by the multinomial $P(w|z, \theta)$ and $P(v|y, \gamma)$ respectively; while in the Corr-LDA, the visual-words are generated by a Gaussian $P(v|z, \mu, \delta)$. Our use of a multinomial to represent the visual words distribution is also consistent with the recent trend in the vision community in using a bag of quantized visual features to represent visual content.
2. We anchor the latent visual words variable y on the variable z that generates the text-words. Whereas in the Corr-LDA, it is the other way round: the caption words are anchored on the visual word generation. In other words, our model captures the notion that the visual features

are conditional on text words that are present in the video. Hence the text words are generated first, and the visual words are then generated to provide visual reinforcement. This can become important if the noise in the visual features are greater than that in the text features (speech recognition errors).

The complete generative process is as follows:

1. draw topic proportions $\phi \sim \text{Dir}(\alpha)$
2. For each topic $k = 1 \dots K$,
 - draw multinomial $\theta_k \sim \text{Dir}(\beta)$ for text words
 - draw multinomial $\gamma_k \sim \text{Dir}(\beta)$ for visual words
3. There are N_d text words. For each text word $w_n, n = 1..N_d$:
 - draw topic assignment $z_n | \phi \sim \text{Multinomial}(\phi)$
 - draw text word $w_n | z_n \sim \text{Multinomial}(\theta_{z_n})$
4. There are M_d visual words. For each visual word $v_m, m = 1..M_d$:
 - draw discrete index variable $y_m \sim \text{Uniform}(1..N_d)$
 - draw visual word $v_m \sim \text{Multinomial}(\gamma_{z_{y_m}})$

From Figure 3.1, we can write the joint distribution for a multimedia document in terms of the visual words v , textual (speech transcript) words w and latent variable z as:

$$p(\mathbf{z}, \mathbf{y}, \phi, \mathbf{v}, \mathbf{w} | \alpha, \theta, \gamma) = p(\phi | \alpha) \left(\prod_{n=1}^{N_d} p(z_n | \phi) p(w_n | z_n, \theta) \right) \left(\prod_{m=1}^M p(y_m | N_d) p(v_m | y_m, \mathbf{z}, \gamma) \right) \quad (3.6)$$

3.6.2.3 Inference and Estimation

The parameters we want to estimate are: α, θ and γ . Marginalizing over the latent variables ϕ, z and y , we re-write equation 3.6 as:

$$p(\mathbf{v}, \mathbf{w} | \alpha, \theta, \gamma) = \int_{\phi} \sum_z \sum_y p(\mathbf{v}, \mathbf{w}, \phi, \mathbf{z}, \mathbf{y} | \alpha, \theta, \gamma) d\phi \quad (3.7)$$

Figure 3.1: Left: Graphical representation of our joint model. Note that our joint model is different from the Corr-LDA of Blei *et al* [25] shown in Figure 2.5: we model the generation of *both* text-words w and visual-words v by the multinomial $P(w|z, \theta)$ and $P(v|y, \gamma)$ respectively; while in the Corr-LDA, the visual-words are generated by a Gaussian $P(v|z, \mu, \delta)$. See text for explanation. Right: graphical representation of the variational distribution used to simplify the graphical model on the left. The factorized distribution assumes the independence amongst ϕ, z and y .

Unfortunately, equation 3.7 is intractable due to the coupling between ϕ and γ in the summation over the latent variables z . We turn to the use of variational inference to approximate the posterior distribution over latent variables given the observed data (visual words v and text w). We follow the general recipe of variational inference and derive the inference algorithm. While the full details of the derivation are given in Appendix B, we highlight the key inferencing steps.

Variational Simplification: To simplify the original graphical representation in equation 3.7, we introduce free variational parameters χ, δ and λ and write the factorized variational distribution on the latent variables as:

$$q(\phi, \mathbf{z}, \mathbf{y} | \chi, \delta, \lambda) = q(\phi | \chi) \prod_{n=1}^{N_d} q(z_n | \delta_n) \prod_{m=1}^M q(y_m | \lambda_m) \quad (3.8)$$

where χ is the variational Dirichlet parameter, δ_n is the variational multinomial parameter over K topics, and λ_m is the variational multinomial parameter over M visual words.

We minimize the KL-divergence between this factorized distribution and the true posterior. Taking derivatives with respect to the variational parameters, we obtain the following coordinate ascent algorithm (for the full derivation details, see Appendix B):

Symbol	Description
v	visual word
v_n^j	visual word indicator variable: $v_n^j = 1$ if v_n is the same as the j th token in the visual vocabulary, else $v_n^j = 0$.
w	text word
w_m^j	text word indicator variable: $w_m^j = 1$ if w_m is the same as the j th token in the text vocabulary, else $w_m^j = 0$
y	Discrete index variable $y \in \{1, \dots, M\}$
z	Latent topic
z_n^i	Topic indicator variable: $z_n^i = 1$ if z_n is the same as the i th latent topic, else $z_n^i = 0$
ϕ	Multinomial of topic distribution in document. $\phi_i, i = 1..K$
α	Dirichet prior for ϕ . $\alpha_i, i = 1..K$
γ	Multinomial of a topic's visual word distribution. $\gamma_{ij}, i = 1..K, j = 1..V_v$
θ	Multinomial of a topic's text word distribution. $\theta_{ij}, i = 1..K, j = 1..V_t$
β	Dirichet prior to θ
χ	Variational Dirichlet. $\chi_i, i = 1..K$
δ	Variational Multinomial. $\delta_{ni}, n = 1..N_d, i = 1..K$
λ	Variational Multinomial. $\lambda_{mn}, m = 1..M_d, n = 1..N_d$
M_d	Number of text words in document d
V_t	Size of text vocabulary
N_d	Number of visual words in document d
V_v	Size of visual vocabulary
D	Number of documents in the training set
K	Number of topics

Table 3.1: Notations used to describe the inference process. Boldfaced font denotes array or matrix.

1. Update the posterior Dirichlet parameters:

$$\chi_i = \alpha_i + \sum_{n=1}^{N_d} \delta_{ni} \quad (3.9)$$

2. For each text word, update the posterior distribution over the latent factors:

$$\delta_{ni} \propto \theta_{i,w_n} \exp \left(\Psi(\chi_i) - \Psi\left(\sum_{j=1}^K \chi_j\right) + \sum_{m=1}^M \lambda_{mn} \log \gamma_{iv_m} \right) \quad (3.10)$$

3. For each visual word, update the posterior distribution over the anchoring text words:

$$\lambda_{mn} \propto \exp \left(\sum_{i=1}^K \delta_{ni} \log \gamma_{iv_m} \right) \quad (3.11)$$

These update equations are invoked repeatedly until the change in KL divergence is small.

Parameter Estimation: So far, we have maximized the lower bound with respect to the variational parameters χ, δ, λ . Next, we maximize the bound with respect to the model parameters α, θ, γ . Given a training set $\mathbb{T} = \{(\mathbf{v}_d, \mathbf{w}_d)\}_{d=1}^D$, we find the maximum likelihood estimation for visual word topics γ , text topics θ and the prior α . The overall log likelihood of $\mathbb{T}$ is bounded by:

$$\mathcal{L}(\mathbb{T}) = \sum_{d=1}^D \log p(\mathbf{v}_d, \mathbf{w}_d | \alpha, \theta, \gamma) \geq \sum_{d=1}^D \mathcal{L}(\chi_d, \delta_d, \lambda_d; \alpha, \theta, \gamma) \quad (3.12)$$

In Appendix B, we show that the maximization-step for the conditional multinomial θ is:

$$\theta_{ij} \propto \sum_{d=1}^D \sum_{n=1}^{N_d} \delta_{dni} w_{dn}^j \quad (3.13)$$

$w_{dn}^j = 1$ if the n th text word in document d is the j th word in the text vocabulary, else $w_{dn}^j = 0$.

The maximization-step for the conditional multinomial γ is:

$$\gamma_{ij} \propto \sum_{d=1}^D \sum_{m=1}^{M_d} v_{dm}^j \sum_{n=1}^{N_d} \delta_{dni} \lambda_{dmn}, \quad (3.14)$$

$v_{dm}^j = 1$ if the m th visual word in document d is the j th word in the visual word vocabulary, else $v_{dm}^j = 0$. Finally, for the Dirichlet prior α , we follow the approach in [27] and use a Newton-Raphson algorithm to maximize α .

3.6.3 Modeling Data Complexity from Query Ambiguity

A key challenge for modern information retrieval systems is to deal gracefully with ambiguous queries. If the return documents from an ambiguous query are to be clustered according to the possible meanings of the query, we need to know how many clusters are there in the return documents. That is, we need to know the number of possible interpretations to the query. This falls to the problem of assessing query ambiguity, a long standing topic of interest to the information retrieval community.

One way to assess query ambiguity is to use a dictionary or thesaurus to resolve the lexical ambiguity of the query words, and elucidate plausible interpretations of the query. Hence, given the ambiguous query “apple”, one may conjecture three groups corresponding to “apple company”,

Figure 3.2: Top: Top-ranked "apple" images from Google Image; Bottom: six possible visual senses (clusters): *logo, fruit, iphone, iMac, drawings, people*

"apple product", "apple farm". However, the main problem with this approach is that the resulting groupings are often too coarse-grained. For example, under the "apple product" interpretation of the query, there may be "apple iMac", "apple iPOD", etc (see Figure 3.2).

A better way is to view query ambiguity as dependent on the documents in the collection. Following the above query example, if the collection is a corporate corpus of documents from the Apple company, we would least expect the cluster corresponding to "apple farm" to form. Hence, disambiguating a query and elucidating clusters of possible interpretation in the return documents should be a data-driven approach. Because different ambiguous queries may trigger varying degrees of polysemy in the return documents, we need a way to handle clustering without knowing the number of clusters in advance.

To capture the varying levels of ambiguity in a query topic, we need to augment our facet

Figure 3.3: The framework of our approach. Given the ambiguous query “mouse”, a set of polysemous images is retrieved. Their latent visual senses are elucidated by HDP, helped by a parallel injection of images of suggested semantic senses by Wikipedia.

document model with the ability to grow with the polysemy (hence, complexity) of the data. To achieve this, we base our faceted document modeling approach on the Hierarchical Dirichlet Process (HDP), a Bayesian nonparametric method.

In the following subsections, without loss of generality, we describe our HDP-based approach on modeling facets in image collections. Figure 3.3 illustrates our image facet clustering framework.

3.6.3.1 The Hierarchical Dirichlet Process (HDP)

The HDP is an extension of the Dirichlet Process (DP), a type of stochastic process first introduced in the 1960’s [59]. But it has recently become an important tool as a prior for infinite mixture models. HDP extends DP in such a way that the dependencies amongst a set of DPs can be specified in a tree structure [152]. HDPs are useful priors for hierarchical mixture models, in which data are organized in groups that share the mixture components. We review the DP and then describe how we model images and clusters with an HDP.

Dirichlet Process: We begin by considering a Bayesian finite mixture model with J mixture components and each component j has a mixture weight π_j and a parameter vector θ_j . Assume we

have N data points denoted x_i , where $1 \leq i \leq N$. Each point is assigned to a mixture component indexed by the indicator z_i . Hence, $p(z_i = j|\pi) = \pi_j$, or $z_i|\pi \sim \text{Multi}(\cdot|\pi)$. The data likelihood is $p(x_i|z_i = j, \theta) = F(x_i|\theta_j)$, where $F(\cdot|\cdot)$ is a pdf parameterized by θ , and $\theta_j \sim H(\cdot)$ for some base distribution H . The mixture weight π has a symmetric Dirichlet prior: $\pi|J, \alpha \sim \text{Dir}(\cdot|\alpha/J)$. From [130], the probability of assigning data point i to component j given all other assignments is:

$$p(z_i = j|z_{-i}, \alpha, J) \propto \frac{n_j^{-i} + \alpha/J}{N - 1 + \alpha} \quad (3.15)$$

where z_{-i} denotes the assignment of all data points excluding the point i , and n_j^{-i} denotes the number of points assigned to component j excluding point i .

Now consider an infinite mixture model and how we can define a prior for the infinite dimensional parameters. The key is to understand that the infinite dimension weights must *sum to one*. We can construct such a weight distribution by a *stick-breaking* process $\text{Stick}(\alpha)$, where α is a concentration parameter. We imagine starting with a stick with unit length and breaking it at a random point. We take the right piece and break it again at a random point. The process is repeated infinitely, producing a set of random weights π that has a *countably* infinite number of dimensions $j = 1..\infty$, and whose components all sum to one. H would need to be sampled a countably infinite number of times to generate the component parameter values θ_j .

Now consider the distribution over all possible component parameter values θ . This distribution is non-zero at a countably infinite number of values. We denote this distribution by $G(\psi) = \sum_{j=1}^{\infty} \pi_j \delta(\psi - \theta_j)$. Each such G can be seen as a sample from a stochastic process that can be proven to be the DP. In general, a DP is characterized by a scalar parameter α and a base distribution H . A sample from a DP, denoted as $G|\alpha, H \sim \text{DP}(\alpha, H)$, is a distribution that is non-zero over a countably infinite number of values. As we have seen, this is exactly what is required to parameterize an infinite mixture model.

Hierarchical Dirichlet Process: As we have seen how DP acts as a prior for infinite mixture models, HDP can be similarly seen as a prior for the *hierarchical* infinite mixture models. Here, we consider a two-level hierarchy model but it can be easily generalized to more levels. In HDP, we assume that we have T groups of data, each consisting of N_t data points x_{ti} , $1 \leq t \leq T$ and

$1 \leq i \leq N_t$. Each data group is modeled by an infinite mixture model. These models are not independent: the mixtures share component parameters θ and a common DP prior.

The dependencies amongst the infinite mixture models can be again understood using the stick-breaking distribution. Starting at the top level, imagine drawing a sample $G|\alpha_0, H \sim \text{DP}(\alpha_0, H)$. Then we can write as: $G(\psi) = \sum_{j=1}^{\infty} \beta_j^0 \delta(\psi - \theta_j)$ where $\beta^0|\alpha_0 \sim \text{Stick}(\alpha_0)$ are the infinite dimensional mixing weights.

We next form a second DP using G as a base distribution, with scalar parameter α_1 , and generate samples from this DP for each of the T mixture models: $G_t|\alpha_1, G \sim \text{DP}(\alpha_1, G)$. Each sample can be written as: $G_t(\psi) = \sum_{j=1}^{\infty} \pi_{tj} \delta(\psi - \theta_j)$. As each G_t inherits the same non-zero points θ_j as G , we have constructed T dependent infinite mixture models. Each model has a separate weight π_t but shared θ .

3.6.3.2 *An Image Facet or Cluster as a Visual Sense of the Query*

Following recent trends, we adopt a multinomial bag of words (textual words and visual words) for images. Another motivation for using multinomial distributions is that they are computationally amenable for the latent topic modeling framework [28]. This framework was used in our earlier work in [165] to model the visual senses of images. In that work, images are first represented as multinomial data, and clustered according to their latent topics. We now further extend the modeling framework to include HDP as a non-parametric prior. HDP offers the advantage of sample an infinite number of latent topics for each visual sense cluster. This is desirable because different queries inherently have varying levels of ambiguity, leading to different levels of image polysemy. As the complexity of the image data grows, such as when there is a greater visual diversity, our HDP model allows us to grow the visual sense model.

Given a collection of polysemous images retrieved from an ambiguous image search query, these images induce a variable number of latent topics proportionate to the number of underlying visual senses (or clusters). Images belonging to the same visual sense (or cluster) are expected to have similar words (textual words and visual words). We now consider each image j as consisting of n_j image feature points $(x_{j1}, \dots, x_{jn_j})$. We assume that these features are exchangeable and to be modeled with a mixture model. While each mixture model has mixing proportions specific to the image, we require that all images share the same set of mixture components. The main idea behind

Figure 3.4: Chinese Restaurant Franchise (CRF) for three images with 8 image feature points. Top shows the global restaurant serving dishes shared by all tables in the three local restaurants shown below. The generative process in each restaurant is a Chinese Restaurant Process (CRP). For example, in the 1st local restaurant, the 1st, 3rd, 6th and 8th customer were the first to sit at an empty table, while the others sat at tables already occupied. For this restaurant, the 9th customer will sit at table 1, 2, 3, 4 with probabilities $\frac{3}{8+\alpha_0}, \frac{2}{8+\alpha_0}, \frac{2}{8+\alpha_0}, \frac{1}{8+\alpha_0}$ respectively, or will sit at a new (5th) table with probability $\frac{\alpha_0}{8+\alpha_0}$. Similarly, in the global restaurant, each customer (j, i) sitting at a table corresponds to table i in the local restaurant j . The table in the global restaurant that (j, i) chooses to sit at, determines the dish that is served at table i in the local restaurant j . In this example, for any new customer entering into a local restaurant j , if he sits down at a new table, the dish for that new table will be ψ_1, ψ_2, ψ_3 or ψ_4 with probability $\frac{4}{11+\gamma}, \frac{4}{11+\gamma}, \frac{2}{11+\gamma}$, or $\frac{1}{11+\gamma}$ respectively, or a new dish ψ_5 with probability $\frac{\gamma}{11+\gamma}$.

the constraints is to allow statistical strength to be shared amongst images, that also facilitate generalization to new images [152].

We use the HDP as a non-parametric prior to allow component sharing amongst the mixture models. The HDP is a distribution over a set of random probability measures over the parameter space of infinite mixture models. There is one measure G_j for each image j , and a global measure G_0 . G_0 is distributed as $\text{DP}(\gamma, H)$, with H the base measure and γ the concentration parameter. Each G_j is conditionally dependent given G_0 , with distribution $G_j \sim \text{DP}(\alpha_0, G_0)$.

The goal of learning is to update the parameters in HDP. We adopt the Gibbs' sampling [152] as the learning algorithm. We choose the simpler Chinese Restaurant Franchise (CRF) metaphor to describe the learning process.

Imagine there are an infinite number of Chinese restaurants each with an infinite number of

tables (see Figure 3.4). All customers sitting on the same table in any restaurant will share the same dish of food. Let x_{ji} be the i^{th} customer in the j^{th} restaurant. In this metaphor, the j^{th} restaurant represents the j^{th} image, and x_{ji} is the i^{th} observed image features. There are n_j such observed features. Recall then that $x_{ji}|\theta_{ji} \sim F(\theta_{ji})$. Hence, for the j^{th} image, there are $\theta_{j1}, \dots, \theta_{jn_j}$ such parameters. To generate $\theta_{j1}, \dots, \theta_{jn_j}$, imagine n_j customers (each corresponds to a θ_{ji}) in the Chinese restaurant. The first customer sits at the first table. A subsequent customer sits at an occupied table with probability proportional to the number of customers already seated there, or at the next unoccupied table with probability proportional to α_0 . Suppose customer i sat at table t_{ji} . The above conditional distribution can be written as:

$$t_{ji}|t_{j1}, \dots, t_{ji-1}, \alpha_0 \sim \sum_{t=1}^{T_j} \frac{n_{jt}}{\sum_k n_{jk} + \alpha_0} \delta_t + \frac{\alpha_0}{\sum_k n_{jk} + \alpha_0} \delta_{t^{new}} \quad (3.16)$$

where T_j is the current number of tables in restaurant j , and n_{jt} is the number of customers currently sitting at table t . Note that because all customers share the same dish of food, all x_{ji} share the same mixture component indexed by the table t_{ji} . Once all customers have sat down, the seating plan corresponds to a partition of $\theta_{j1}, \dots, \theta_{jn_j}$. Because this is an exchangeable process [152], the probability of a partition does not depend on the order in which the customers sit down. Now we associate each table t a draw ϕ_{jt} from G_0 , and assign $\theta_{ji} = \phi_{jt_{ji}}$.

Performing this process independently for each image j , we have an assignment of each θ_{ji} to a sample $\phi_{jt_{ji}}$ from G_0 . Because all ϕ_{jt} are independent draws from G_0 , which is distributed according to $DP(\gamma, H)$, we may apply the same Chinese restaurant partitioning process to all ϕ_{jt} . This corresponds to partitioning at the next-level hierarchy. Continuing the metaphor, we associate ϕ_{jt} with a customer seated at table k_{jt} at the global level restaurant. However, each k_{jt} is actually a (common) dish that is served to a table at the local restaurant. We can then write the conditional probability as:

$$k_{jt}|k_{11}, \dots, k_{1n_1}, k_{21}, \dots, k_{jt-1}, \gamma \sim \sum_{k=1}^K \frac{m_k}{\sum_{k'} m_{k'} + \gamma} \delta_k + \frac{\gamma}{\sum_{k'} m_{k'} + \gamma} \delta_{k^{new}} \quad (3.17)$$

where K is the current number of dishes, and m_k represents the number of local restaurant tables that has ordered dish k . Just as a new table at a local restaurant can be generated from G_0 , a

new dish at the global level “restaurant” can be generated from H : we draw ψ_k from H and assign $\phi_{jt} = \psi_{k_{jt}}$.

The state space of HDP consists of values of $\mathbf{t}, \mathbf{k}$ and $\boldsymbol{\psi}$. In the global “restaurant”, the number of k_{jt} and ψ_k variables is not fixed. We can think of the actual state space as consisting of countably infinite number of k_{jt} and ψ_k . Only finitely many are actually assigned to image feature data and represented explicitly.

Sampling the feature clusters (tables in the local restaurants): Let $f(\cdot|\psi)$ and h be the density functions for $F(\psi)$ and H respectively, n_{jt}^{-i} be the number of tables t_{ji} ’s equal to t except t_{ji} , and m_k^{-jt} be the number of dishes k_{jt} ’s equal to k except k_{jt} . From equation 3.16, we can compute the conditional prior distribution of t_{ji} . Combined with the likelihood of generating image feature x_{ji} given $t_{ji} = t$ (which is simply $f(x_{ji}|\psi_{k_{jt}})$), we obtain the conditional posterior for t_{ji} . Hence, the probability of a new customer x_{ji} sitting at table t is:

$$p(t_{ji} = t | \mathbf{t}_{-ji}, \mathbf{k}, \mathbf{x}, \boldsymbol{\psi}) \propto \begin{cases} \alpha_0 f(x_{ji}|\psi_{k_{jt}}) & \text{if } t = t^{new} \\ n_{jt}^{-i} f(x_{ji}|\psi_{k_{jt}}) & \text{if } t \text{ is used} \end{cases} \quad (3.18)$$

Sampling the global latent topics (dishes or tables in the global restaurant): Sampling the k_{jt} variables is similar to sampling the t_{ji} variables described above. We generate a new mixture parameter $\psi_{k^{new}} \sim H$. Changing k_{jt} changes the component membership of all feature data in table t . The conditional distribution for k_{jt} is:

$$p(k_{jt} = k | \mathbf{t}, \mathbf{k}_{-jt}, \boldsymbol{\psi}, \mathbf{x}) \propto \begin{cases} \gamma \prod_{i:t_{ji}=t} f(x_{ji}|\psi_k) & \text{if } k = k^{new} \\ m_k^{-t} \prod_{i:t_{ji}=t} f(x_{ji}|\psi_k) & \text{if } k \text{ is used} \end{cases} \quad (3.19)$$

Sampling $\boldsymbol{\psi}$: Conditioned on the indicator variables $\mathbf{k}$ and $\mathbf{t}$, ψ_k for each mixture component are mutually independent. The posterior distribution is dependent only on the image feature assigned to component k and is given by:

$$p(\boldsymbol{\psi}_k | \mathbf{t}, \mathbf{k}, \boldsymbol{\psi}_{-\mathbf{k}}, \mathbf{x}) \propto h(\boldsymbol{\psi}_k) \prod_{ji:k_{jt_{ji}}=k} f(x_{ji}|\boldsymbol{\psi}_k) \quad (3.20)$$

3.7 Faceted Retrieval Model

At the start of this chapter, we introduce our faceted topic retrieval modeling framework. We have mentioned that there are two sub-problems in faceted topic retrieval. We have discussed in Section 3.6 the first problem of a faceted document model, where documents are represented in terms of the latent facets or topics. In this section, we discuss the second sub-problem: a Faceted Retrieval Model (FRM), whose goal is to find the smallest set of documents that cover all of the facets.

3.7.1 A Conventional Greedy Approach

We first take a look at a conventional greedy approach to finding a set of documents that cover the most number of facets: *Maximal Marginal Relevance* by Goldstein and Carbonell *et al* [39]. In choosing a document d to be ranked at the i th position, MMR uses a greedy approach that combines the similarity of d to the query and its similarity to all documents ranked at positions 1 to $i-1$.

More formally, let Sim_1 be a scoring function between document and query, and Sim_2 be a similarity function between documents. In MMR, the top rank-1 document d_1 is first chosen solely based on its relevance to the query, as per other conventional relevance-based ranking methods.

$$d_1 = \operatorname{argmax}_{d \in D} \{\text{Sim}_1(d, Q)\}. \quad (3.21)$$

For positions from rank 2 onwards, MMR now selects documents that are not only based on its relevance to the query, but also whether these documents are also similar to those that are already retrieved. In other words, in considering documents to be placed at the rank 2 position and onwards, MMR will favour these documents that are *not* too similar to those at the higher ranks. This is where MMR differs from the conventional methods of ranking documents by relevance. More formally, for $i = 2, \dots$, we have:

$$\text{MMR}(d_i, Q) = \alpha \text{Sim}_1(d_i, Q) - (1 - \alpha) \max_{j=1..i} \text{Sim}_2(d_i, d_j) \quad (3.22)$$

α is a parameter that specifies a linear weighting on Sim_1 and Sim_2 . The minus term in the above equation introduces a penalty term to the score of d_i , in proportion to how similar is d_i to the documents that are already ranked above. For a retrieval system, α is a parameter that controls the balance between how much importance it should place on document being relevant to

the query (Sim_1), and how much importance it should place on documents being dis-similar to all other documents that have been retrieved so far (Sim_2). When $\alpha=1$, MMR is equivalent to the conventional document-query ranking approach.

3.7.2 Our Set-based Probabilistic Approach

Recall our earlier discussion in Section 3.5 that given a set of facets F , and a set of documents $D = d_1, d_2, \dots, d_J$, the probability that all facets $F_m \in F, m = 1..M$ are contained in D is:

$$P(F \in D) = \prod_{m=1}^M P(F_m \in D) = \prod_{m=1}^M \left\{ 1 - \prod_{d_j \in D} (1 - P(F_m \in d_j)) \right\} \quad (3.23)$$

In a practical retrieval setting, users are expected to focus only on the first few result pages. Hence, the goal of faceted retrieval is to find a smaller subset $D' \subset D$ without compromising too much the probability of facet presence $P(F \in D')$. In particular, if we set $|D'| = k$, then maximizing the faceted retrieval metric $S\text{-rec}@k$ is equivalent to the following constrained optimization problem.

$$\begin{aligned} \max \quad & P(F \in D') = \prod_{m=1}^M \left\{ 1 - \prod_{d_j \in D'} (1 - P(F_m \in d_j)) \right\} \\ \text{subject to:} \quad & |D'| = k \end{aligned} \quad (3.24)$$

Note that in our set-based probabilistic approach to faceted topic retrieval, it is NP-hard to choose a subset of documents D' to maximize $P(F \in D')$. To see this, define a document selection indicator function $\delta(d_j) = 1$ if document d_j is selected, and 0 otherwise. Then define a new optimization function as:

$$\begin{aligned} \max \quad & \prod_{m=1}^M \left\{ 1 - \prod_{d_j \in D'} (1 - P(F_m \in d_j))^{\delta(d_j)} \right\} \\ \text{subject to:} \quad & \sum_{d_j \in D'} \delta(d_j) \leq k \end{aligned} \quad (3.25)$$

Note that the constraint in the above optimization function limits the number of the selected documents to be equal or less than k , making the optimization function identical to the original optimization function in equation 3.24. However, it is clear that this new optimization function is a

typical 0-1 integer programming problem and is therefore NP-hard.

Hence, we need an approximation to solve the faceted retrieval optimization problem. One approach is to follow the MMR greedy methodology as follows. Given the current set of retrieved documents D' , choose from amongst the documents that have not yet been retrieved, the single document d' that maximizes $P(F \in d' \cup D')$. That is, choose the first document that has the greatest expected number of facets, then choose the second document that has the greatest expected number of facets that are different from those already found in the first, and so on. It also provides a natural ranking of documents in order of their selection.

3.8 Summary

In this chapter, we propose a faceted topic retrieval model for multimedia content. Under this framework, multimedia documents are represented in terms of topics, which are probability distributions of words and other low level semantic units of representation or meaning. We define the problem in Section 3.3, and motivate an set-based probabilistic evaluation framework in Section 3.4. There are two components in our proposed faceted topic retrieval framework: the Faceted Document Model and the Faceted Retrieval Model.

We describe the Faceted Document Model in Section 3.6. Our goal is to hypothesize facets of documents in a multimedia collection. To achieve this, we develop a Bayesian topic model of a multimedia document that mines the co-occurrence of observed features in the textual and visual streams. This is a key contribution in this thesis, and we shall devote more details in Chapter 5 on how our joint topic model can mine topical facets in news video retrieval. To cater to the varying complexity of the multimedia collection, we also augment our Bayesian topic model with the ability to grow with the data complexity. Specifically, we base our faceted document modeling approach on the Hierarchical Dirichlet Process, a Bayesian nonparametric method. This is another key contribution in this thesis, and we shall devote more details in Chapter 4 on how our nonparametric method can effectively cope with ambiguous queries and to cluster polysemous images.

We describe the Faceted Retrieval Model in Section 3.7. Our goal is to find the smallest set of documents that cover all of the facets. To achieve this, we develop a greedy approximation method that follows after the conventional MMR approach. In Chapter 5, we also report on the application

of this greedy approximation on news video retrieval.

Finally, in Chapter 6, we will also describe the application of our faceted topic retrieval model on deriving query suggestions for video search.

Chapter 4

FACETED CLUSTERING OF POLYSEMOUS IMAGES

In the last chapter, we have proposed a faceted topic retrieval model for multimedia content analysis. Over the next three chapters, we turn our focus to examining how our proposed model is applied to three novel classes of faceted topic retrieval problems:

1. In this chapter, we look at faceted clustering of polysemous images arising from an ambiguous query. The focus is on the use of the Bayesian nonparameteric method to facilitate clustering without the need to know in advance the number of clusters. Specifically, we use the Hierarchical Dirichlet Process to enable our clustering model to adaptively grow with the polysemy in the image dataset. This allows our image faceted clustering algorithm to handle varying level of image polysemy arising from different level of ambiguity in the query.
2. In chapter 5, we examine faceted topic of news video. The focus is on the joint modeling of the visual and text information in a video content to constrain the formation of topics that better correlate to actual news events.
3. In chapter 6, we propose a faceted topic based query suggestion method to allow a user to interactively select and refine his information need. We show that our method improves over existing query suggestion methods that are based on simple lexicographical matching of query words.

4.1 Motivation

A common approach to handle query ambiguity is through automatic clustering of web search results. The theoretical underpinning for this approach is based on Rijsbergen’s cluster hypothesis [131] that states that the associations between documents convey information about the relevance of documents to requests. A key application of the hypothesis is that the extent to which documents are relevant to a query is related to the extent to which they belong to the same cluster. In particular, a search engine may assume that the different information needs of an ambiguous query can be captured by the topical clusters of the search results.

The validity of Rijsbergen’s cluster hypothesis has been borne out by its many successful applications in web search result clustering [175, 4], cluster-topic-based document retrieval models [170], and exploratory browsing interfaces [55]. However, the above successes are mainly limited to web *text* document search, and have not been replicated to image and video search. For example, on the ambiguous query “apple” which can refer to the fruit, the company or the product, most top image search engines¹ return results that haphazardly alternate between the first two meanings. Hence, a user looking for images of iMac would likely be disappointed. In the absence of any disambiguating information on the search query, a better approach, exemplifying the cluster hypothesis, is to present users with clusters of images embodying the multiple dominant interpretations of the original query. The resulting clusters can then be seen as the *visual* senses of the query.

Clustering images into their visual senses is akin to *semantic* clustering of the images, which remains a challenging problem. A common approach is to assume that visually similar images are close to each other in feature space, and construct models in which features are generated from a mixture of probability distributions. However, because feature vectors are usually very high dimensional, a parametric characterization of their distribution is difficult. In particular, the appropriate number of mixture components (K) is often not known *a priori*, hence necessitating some form of model selection.

¹We tried Google, Yahoo and Microsoft Bing Image Search

4.2 A Bayesian Approach

In this chapter, following the foundational work that we have discussed in Section 3.6.3.1 and Section 3.6.3.2, we adopt an alternate, fully Bayesian approach to cluster polysemous images. Specifically, we attempt to build an *infinite* mixture model [130], where the number of mixture component K is itself a random variable and can potentially be unlimited. Of course, at any point in time, only a finite number of mixture components have data assigned to them. But as more data is seen, more components may be used. Hence, this is a non-parametric model in that the model parameters (e.g. K) grow with the amount (and complexity) of data. This is desirable in our case for visual sense clustering of polysemous images, because different queries inherently have varying levels of ambiguity, leading to different levels of image polysemy. As the complexity (polysemy) of the image data grows, our non-parametric model will also grow to accommodate the greater visual diversity.

We follow recent trends in the vision community to represent images as bags of words. Here, the words refer to both the *textual* words associated with the images, and the *visual* words encoding the visual elements in the images. These two modalities are related since they collectively describe the image content. By incorporating the two modalities in our framework, we exploit their co-occurrence patterns. We adopt the latent topic model [28, 126] to encode these co-occurrence patterns. We define latent topics as multinomial distributions over the words, and make use of the Hierarchical Dirichlet Process (HDP) [152] to learn these topics and derive compact representations of images and clusters.

Our clustering method is completely unsupervised. To constrain the clustering of the images towards their *semantic* senses, we introduce into the clustering process additional images that represent these semantic senses. These *sense-specific* images are obtained by first consulting an online resource (Wikipedia) to solicit possible ways to *expand* the ambiguous query, and then issuing the expanded query to an image search engine. This process of query expansion can be seen as a way to disambiguate query. Together, the polysemous images from the ambiguous query and the sense-specific images from the disambiguated queries are input to the HDP-based clustering step. Collectively, HDP computes latent topics that are shared by all images. Figure 4.1 is a pictorial illustration of our approach.

For the sake of continuity, we first review the key insights that we have derived in Section 3.6.3.1 and Section 3.6.3.2.

Figure 4.1: The framework of our approach. Given the ambiguous query “mouse”, a set of polysemous images is retrieved. Their latent visual senses are elucidated by HDP, helped by a parallel injection of images of suggested semantic senses by Wikipedia.

4.2.1 HDP-based Faceted Clustering of Polysemous Images

Consider a collection of polysemous images returned from an ambiguous image search query. Within these images, there is a variable number of latent topics that is proportionate to the number of underlying clusters. We would expect that images belonging to the same cluster have similar words (both textual words and visual words). Denote each image j as consisting of n_j image feature points $(x_{j1}, \dots, x_{jn_j})$. We assume that these features are exchangeable and to be modeled with a mixture model. While each image would have different image-specific mixing proportions, we require that all images share the *same* set of mixture components. The motivation for imposing this constraint is to allow statistical strength to be shared amongst images, that also facilitate generalization to new images [152].

To allow component sharing amongst the mixture models, we use the HDP as a non-parametric prior. The HDP is a distribution over a set of random probability measures over the parameter space of infinite mixture models. There is one measure G_j for each image j , and a global measure G_0 . G_0 is distributed as $\text{DP}(\gamma, H)$, with H the base measure and γ the concentration parameter. Each G_j is conditionally dependent given G_0 , with distribution $G_j \sim \text{DP}(\alpha_0, G_0)$. The goal of learning is to update the parameters in the HDP.

Let x_{ji} be the i^{th} observed image feature vector in the j^{th} image. There are n_j such observed feature vectors. Supposed that $x_{ji}|\theta_{ji} \sim F(\theta_{ji})$, for some parameterized pdf $F(.|.)$. Hence, for the j^{th} image, there are $\theta_{j1}, \dots, \theta_{jn_j}$ such parameters. To generate $\theta_{j1}, \dots, \theta_{jn_j}$, consider the n_j feature vectors, each of which corresponding to a θ_{ji} in the j^{th} image. The conditional distribution of an image feature vector x_{ji} being assigned to the t_{ji} cluster can be written as:

$$t_{ji}|t_{j1}, \dots, t_{ji-1}, \alpha_0 \sim \sum_{t=1}^{T_j} \frac{n_{jt}}{\sum_k n_{jk} + \alpha_0} \delta_t + \frac{\alpha_0}{\sum_k n_{jk} + \alpha_0} \delta_{t^{new}} \quad (4.1)$$

where T_j is the current number of clusters in image j , and n_{jt} is the number of feature vectors currently assigned to cluster t . Note that all image feature vectors x_{ji} share the same mixture component indexed by the cluster t_{ji} . Once all image feature vectors have been assigned to clusters, the new clustering corresponds to a partition of $\theta_{j1}, \dots, \theta_{jn_j}$. Because this is an exchangeable process [152], the probability of a partition does not depend on the order in which the feature vectors were observed and assigned to a cluster. Now we associate each cluster t a draw ϕ_{jt} from G_0 , and assign $\theta_{ji} = \phi_{jt_{ji}}$.

Performing this process independently for each image j , we have an assignment of each θ_{ji} to a sample $\phi_{jt_{ji}}$ from G_0 . Because all ϕ_{jt} are independent draws from G_0 , which is distributed according to $DP(\gamma, H)$, we may apply the same partitioning process to all ϕ_{jt} . This corresponds to partitioning at the next-level hierarchy. Continuing the metaphor, we associate ϕ_{jt} with an image feature vector clustered to k_{jt} at the global level cluster-set. However, each k_{jt} is actually a (common) cluster-set at the local level image cluster-set. We can then write the conditional probability as:

$$k_{jt}|k_{11}, \dots, k_{1n_1}, k_{21}, \dots, k_{jt-1}, \gamma \sim \sum_{k=1}^K \frac{m_k}{\sum_{k'} m_{k'} + \gamma} \delta_k + \frac{\gamma}{\sum_{k'} m_{k'} + \gamma} \delta_{k^{new}} \quad (4.2)$$

where K is the current number of global level cluster and m_k represents the number of image clusters that has adopted the same cluster parameter-set k . Just as a new image feature vector at a local image can be generated from G_0 , a new cluster parameter-set at the global level image-pool can be generated from H : we draw ψ_k from H and assign $\phi_{jt} = \psi_{k_{jt}}$.

The state space of HDP consists of values of $\mathbf{t}, \mathbf{k}$ and $\boldsymbol{\psi}$. At the global image-pool level, the number of k_{jt} and ψ_k variables is not fixed. We can think of the actual state space as consisting of

countably infinite number of k_{jt} and ψ_k . Only finitely many are actually assigned to image feature data and represented explicitly.

Sampling the feature clusters at the local image level: Let $f(\cdot|\psi)$ and h be the density functions for $F(\psi)$ and H respectively, n_{jt}^{-i} be the number of clusters t_{ji} 's equal to t except t_{ji} , and m_k^{-jt} be the number of image clusters k_{jt} 's equal to k except k_{jt} . From equation 4.1, we can compute the conditional prior distribution of t_{ji} . Combined with the likelihood of generating image feature x_{ji} given $t_{ji} = t$ (which is simply $f(x_{ji}|\psi_{k_{jt}})$), we obtain the conditional posterior for t_{ji} . Hence, the probability of a new image feature vector x_{ji} being assigned to cluster t is:

$$p(t_{ji} = t | \mathbf{t}_{-ji}, \mathbf{k}, \mathbf{x}, \psi) \propto \begin{cases} \alpha_0 f(x_{ji}|\psi_{k_{jt}}) & \text{if } t = t^{new} \\ n_{jt}^{-i} f(x_{ji}|\psi_{k_{jt}}) & \text{if } t \text{ is used} \end{cases} \quad (4.3)$$

Sampling the global latent topics (at the global image-pool level): Sampling the k_{jt} variables is similar to sampling the t_{ji} variables described above. We generate a new mixture parameter $\psi_{k^{new}} \sim H$. Changing k_{jt} changes the component membership of all feature data in cluster t . The conditional distribution for k_{jt} is:

$$p(k_{jt} = k | \mathbf{t}, \mathbf{k}_{-jt}, \psi, \mathbf{x}) \propto \begin{cases} \gamma \prod_{i:t_{ji}=t} f(x_{ji}|\psi_k) & \text{if } k = k^{new} \\ m_k^{-t} \prod_{i:t_{ji}=t} f(x_{ji}|\psi_k) & \text{if } k \text{ is used} \end{cases} \quad (4.4)$$

Sampling ψ : Conditioned on the indicator variables $\mathbf{k}$ and $\mathbf{t}$, ψ_k for each mixture component are mutually independent. The posterior distribution is dependent only on the image feature assigned to component k and is given by:

$$p(\psi_k | \mathbf{t}, \mathbf{k}, \psi_{-k}, \mathbf{x}) \propto h(\psi_k) \prod_{ji:k_{jt_{ji}}=k} f(x_{ji}|\psi_k) \quad (4.5)$$

4.2.2 Augmenting with Wiki-Sense-Disambiguated Images

Because of its unsupervised nature, the resulting HDP-based image clusters may not coincide with the semantic meaning of the original query. To constrain the clustering of the images towards their semantic senses, we introduce into the clustering step, images that are exemplars of the various senses of the query. These sense-specific images can also be seen as a kind of “seed” to facilitate

clustering [32]. To achieve this, we use Wikipedia to suggest possible ways to expand the ambiguous query. These disambiguated queries are then used to retrieve sense-specific images. Together, the polysemous images from the ambiguous query and the sense-specific images from the disambiguated queries are input to the HDP-based clustering step. Collectively, HDP computes latent topics that are shared by all images.

Wikipedia is a free online encyclopedia, representing the outcome of a continuous collaborative effort of a large number of volunteers. Because of the open and collaborative environment the quality and quantity is well trusted. As a large-scale repository of structured knowledge, Wikipedia is a valuable resource for a diverse array of research activities [112]. One structure of particular interest to this chapter is the *disambiguation* page. It gives a detailed list of possible senses (meanings) of ambiguous words by attaching the expression (*disambiguation*) to the name of the ambiguous entity, e.g., *bar_(disambiguation)*, which identifies the disambiguation page of the entity “bar”. The advantage of this disambiguation page is that it not only gives the word senses in a structured categorized way, but also links up pages that have further details. All these advantages motivate us to use it for disambiguating keyword based image search. Given an ambiguous query keyword we issue the query to Wikipedia to extract different (senses) meanings of the word automatically.

Because Wikipedia may suggest many superfluous disambiguations, we use a simple variant of the method in [158] to evaluate the mutual information content of each suggested disambiguation. Each such disambiguation S comprises of one or more terms $t \in S$ that augment the original query to give it a more specific meaning. We score each candidate disambiguation S by the average *pointwise mutual information* (PMI) of the disambiguated terms $t \in S$ with the original query keyword q , weighted by the relative importance of each term t over the Web corpus [48]. The Mutual Information (MI) of S reflects its “semantic distance” from the ambiguous query q and is computed as follows:

$$\text{MI}(S, q) = \sum_{t \in S} \text{PMI}(t, q | \text{corpus}) \times w(t) \quad (4.6)$$

where $w(t)$ denotes the relative importance² of the term t and normalized to sum to 1: $\sum_{t \in S} w(t) = 1$.

²Many standard statistical measures of term weights can be used. The only requirement is that the normalized weights be used in the equation. In our case, we use a variant of TF-IDF by the CLucene open source indexing engine.

Keyword	Wikipedia word senses
bank	Bank finance, Bank building, River Bank, Bank sea floor, Blood bank, Gene bank, Piggy bank
bar	Bar rod, Bar pole, Dessert Bar, Bar Law, Candy Bar, Barbell
bass	Bass Drum, Bass guitar, Bass Flute, Bass Fish, Bass Rock, Bass Strait, Bass Instrument, Acoustic Bass Guitar
mouse	Mouse computing, Mickey Mouse, Mouse Rodent, Stanley Mouse, Mouse anime
plant	Tree, chemical plant, implant, herb, bush, grass, vines, ferns, mosses, forest
speaker	Speaker government, loudspeaker, Orator, computer speaker, BBC speaker
temple	temple anatomy, hindu temple, mount temple, temple mount, temple Jerusalem
tiger	Bengal Tiger, Tiger Woods, Tiger Shark, Tiger Snake, Tiger Beer, Tiger Mac OS, Tiger Tank, Tony the Tiger, White Tiger, Detroit Tiger
watch	wrist watch, guard, watch tower, wall clock, pocket watch
window	window house, computer window, windows operating system, window snyder, window blind

Table 4.1: Keywords and their Wiki-senses used in our experiments

The PMI between two terms is computed as follow:

$$\text{PMI}(t, q | \text{corpus}) = \log\left(\frac{P(t, q | \text{corpus})}{P(t | \text{corpus}) \times P(q | \text{corpus})}\right) \quad (4.7)$$

where the probability of term(s) is approximated by maximum likelihood: $P(t, q | \text{corpus}) = \frac{\#(t, q)}{\#(\text{corpus})}$ is the fraction of documents in the corpus where both t and q are found, and $P(t | \text{corpus}) = \frac{\#(t)}{\#(\text{corpus})}$ is the fraction of documents where the term t is found.

Treating MI in equation 4.6 as a form of saliency, we then rank and retrieve the top most salient senses that exceed a certain threshold. We manually set this threshold such that we were able to set a limit to the number of salient senses for each keyword (to at most ten), and yet at the same time, able to obtain fairly meaningful disambiguated senses of the keyword. Each of the senses is then used to retrieve images from the web using Google Image Search (See table 4.1 for examples of automatic disambiguated senses of keywords).

4.2.3 Presenting Image Clusters in their Wiki-Primary Senses

Given an ambiguous query, the return images will comprise of images from various senses. After the HDP parameters are learnt, we have an approximation of the posterior distribution of the latent variables conditioned on the observed image features. This distribution is over the cluster partition

assigning the image feature vectors to clusters and a truncated vector π_j defining the mixture proportions of each image j over the finite subset of K mixture components that are actually assigned to image feature points. These mixture components $\psi_{1..K}$ parameterize the latent structures in the image features. Because the HDP machinery enforces statistical sharing of strength, the mixture proportion vectors $\pi_{1..J}$ can compactly express the features for each image $1..J$ in terms of the underlying polysemy in the image collection. That is, we can expect that images corresponding to the same visual sense would have a greater similarity in the π distribution. This is the basis for music similarity described in [77], where the KL divergence of π_i and π_j is used to compute the distance between song i and j .

Given a query keyword P , we treat the disambiguated senses suggested by Wikipedia as the *primary* senses S_i of P , $i = 1, 2, \dots, N_P$, where N_P is the number of Wiki-senses of P . For example, in Figure 4.1, the “mouse” keyword has three primary senses, corresponding to the three wiki-disambiguated-senses: “computing”, “Mickey” and “rodent”. Because these are usually the *semantic* senses of the original query, we propose classifying the polysemous images from the original ambiguous query to one of the N_p primary senses. We define the likelihood of the i^{th} sense S_i given the global latent topic $z = z_j$ as:

$$P(S_i|z = z_j) = \frac{1}{|S_i|} \sum_{a \in S_i} P(a|z = z_j) \quad (4.8)$$

$$\approx \frac{1}{|S_i|} \sum_{a \in S_i} \mathbb{KL}(W_a, Z_j) \quad (4.9)$$

where W_a is the word (concatenated textual and visual) distribution of image a , Z_j is the word distribution of topic z_j , and $\mathbb{KL}(\cdot)$ is the Kullback Leibler divergence between the two. For an image d , the model computes the probability of d belonging to the i^{th} sense S_i as:

$$P(S_i|d) = \sum_{j=1}^K P(S_i|z = z_j)P(z = z_j|d) \quad (4.10)$$

Equation 4.10 assigns visual sense probabilities to an image according to how similar it is to the sense-specific images. $P(S_i|d)$ can be used to re-rank the images so as to preserve their original polysemous order: images belonging to some sibling senses are given lower probabilities and pushed

to the back of the rank list.

4.2.4 Algorithm

The pseudo-code of our algorithm is given in Algorithm 1 and Algorithm 2. The main routine is *faceted_cluster*(q), which takes as input the original ambiguous query keyword. It then uses Wikipedia Disambiguation to suggest related senses A to q , each of which is then issued to Google Image Search Engine to retrieve sense-specific images A_{img} . These images are combined with the original polysemous image collection I , and input to the HDP cluster routine *hdp_cluster*($\cdot$). The inference algorithm in *hdp_cluster*($\cdot$) is based on Gibbs sampling. This requires a certain burn-in lag time, before the posterior samples can be reliably sampled. In our implementation, the burn-in lag time is determined in an ad-hoc trial and error fashion. It remains a difficult issue to determine this burn-in time automatically. After HDP inference, the routine *faceted_cluster*(q) receives a set of return variables: the mixture components Ψ , the number of mixture component K and the inferred image mixture proportions π . It then uses these variables to re-rank the polysemous images in I into the Wiki-primary senses.

4.3 Experimental Results

4.3.1 Methodology and Dataset

In this section, we evaluate how well our visual sense models can distinguish between images depicting the various senses of a given polysemous query keyword. We focus on objects, and define a set of ten polysemous keywords. Two factors motivate our choice of object-based keywords. Firstly, following [104], we focus on the core senses of images, which are typically exemplified as objects displayed predominantly in the images. Secondly, the three most relevant previous works on sense discrimination [104] and sense model [137, 165] have also focused on object-based keywords. In comparison, our experiments on ten polysemous object keywords provide a more comprehensive evaluation in terms of dataset size and extent of query class. ([104] uses three object keywords, while [137] reports on five). For each keyword, we automatically mine the dominant senses from Wikipedia (refer to Section 4.2.2). Table 4.1 shows the ten keywords and their respective senses (64 in total).

For each of the ten keywords, we create an image dataset by issuing the keyword as a search

Algorithm 1 : HDP based Clustering Algorithm $hdp_cluster(D)$

Input: An image collection D of images indexed by j , the i^{th} image feature vector x_{ji} ;
Data: t_{ji} is the local cluster to which x_{ji} is assigned to, t_{ji} is unbounded (metaphor: x_{ji} is the i^{th} customer in the j^{th} local restaurant and he chooses to sit on table t_{ji});
 n_{jt} is the size of cluster t in image j ;
 T_j is the number of cluster currently allocated to image j ;
 k_{jt} is the global cluster-parameter-set adopted by the t^{th} local cluster in image j (metaphor: a dish from a global menu that is shared by all local franchise restaurants);
 m_{jk} is the number of global cluster-parameter-set k that are assigned to image j (metaphor: number of tables in restaurant j that order dish k);
 $m_k = \sum_j m_{jk}$ is the total number of times that the global cluster-parameter-set k has been assigned (metaphor: the total number of times that dish k has been ordered);
 h is the density function of H , the base measure;
Output: $\Psi = \{\psi_k\}$ is the global cluster-parameter-sets that are assigned, k is variable, unbounded ;
 $\pi_{j1,j2..jk}$ is the mixture proportions of each image j ;
Initialize link-list L to hold Ψ with pointers to k_{jt}, n_{jt} ; set n_{jt}, m_{jk} to zero ;
// Gibbs sampling over burn-in period and sampling cycle
while *not finished* **do**
 foreach *images* $j \in D$ **do**
 foreach *image feature vectors* x_{ji} **do**
 // calculate image feature likelihood to each existing assigned cluster
 foreach $t = 1$ *to* T_j **do**
 | Calculate the likelihood of x_{ji} given cluster t : $f_{k_{jt}}(x_{ji})$; Multiple by $n_{jt}^{-j_i}$;
 end
 // calculate image feature likelihood to a new cluster
 Calculate $f_{k^{new}}(x_{ji})$ which is simply the prior density of x_{ji} ; Multiple by α_0 ;
 // sample t by mapping the above $T_j + 1$ likelihoods into $[0..1]$ and sampling from $[0..1]$
 $p(t_{ji} = t | \mathbf{t}_{-ji}, \mathbf{k}, \mathbf{x}, \boldsymbol{\psi}) \propto \alpha_0 f_{k^{new}}(x_{ji} | \psi_{k_{jt}})$ if $t = t^{new}$;
 $p(t_{ji} = t | \mathbf{t}_{-ji}, \mathbf{k}, \mathbf{x}, \boldsymbol{\psi}) \propto n_{jt}^{-j_i} f_{k_{jt}}(x_{ji} | \psi_{k_{jt}})$ if t is used ;
 // sample k similarly
 $p(k_{jt} = k | \mathbf{t}, \mathbf{k}_{-jt}, \boldsymbol{\psi}, \mathbf{x}) \propto \gamma \prod_{i:t_{ji}=t} f_{k^{new}}(x_{ji} | \psi_k)$ if $k = k^{new}$;
 $p(k_{jt} = k | \mathbf{t}, \mathbf{k}_{-jt}, \boldsymbol{\psi}, \mathbf{x}) \propto m_k^{-j_t} \prod_{i:t_{ji}=t} f_{k_{jt}}(x_{ji} | \psi_k)$ if k is used ;
 If some $n_{jt} = 0$, i.e., cluster t becomes empty, delete k_{jt} from the link list L ;
 Similarly, if some $m_{jk} = 0$, delete the global cluster-parameter-set k altogether ;
 end
 end
 // Check Convergence and sample Ψ and output mixture proportion for each image
 if *converged and after the sampling cycle* **then**
 foreach $t = 1$ *to* T_j **do**
 | $p(\psi_k | \mathbf{t}, \mathbf{k}, \boldsymbol{\psi}_{-k}, \mathbf{x}) \propto h(\psi_k) \prod_{j:i:k_{jt_{ji}}=k} f_{k_{jt}}(x_{ji} | \psi_k)$;
 end
 foreach *image* j **do**
 | Set π_j to the normalized histogram of the cluster ψ_k assignment of all x_{ji} ;
 end
 end
end

Algorithm 2 : Faceted Image Clustering Algorithm *faceted_cluster(q)*

Input: An (ambiguous) query keyword q

Data: Wikipedia corpus C

Output: K image clusters

Set I to be the set of return (polysemous) images by issuing q to Google Image Search Engine ;

Set D_q = Set of Candidate Disambiguated Queries by issuing q to Wikipedia-Disambiguation-Page ;

// assess the semantic relatedness of each candidate disambiguated query A with q

foreach $A \in D_q$ **do**

 // calculate normalized term weight $w(t)$ of every term t in A

 Set R = random subset of C ;

foreach $t \in A$ **do**

 | $w(t)$ = term frequency of t in R ;

end

 Normalize $w(t)$ to $[0..1]$;

 // calculate Mutual Information (MI) between A and q

$MI(A, q) = \sum_{t \in A} PMI(t, q|C) \times w(t)$;

 where $PMI(t, q|corpus) = \log\left(\frac{P(t, q|corpus)}{P(t|corpus) \times P(q|corpus)}\right)$;

end

Filter all A from D_q with MI scores below a threshold; ;

// D_q is now deemed to be embedding the Wiki-primary senses of the original query q

// Using D_q , incorporate Wikipedia sense-specific images into I

Set $I' = I$;

foreach $A \in \text{the filtered } D_q$ **do**

 Set A_{img} = Retrieve images by issuing A to Google Image Search Engine ;

 Add to I' ;

end

// Invoke the HDP based clustering algorithm (Algorithm 1);

// Get the mixture components Ψ , image mixture proportions π , and the number of components K

$\{\Psi, \pi, K\} = hdp_cluster(I')$;

// Presenting the original polysemous images I according to the Wiki-Primary sense

foreach image $d \in I$ **do**

 // calculate the probability of d belonging to each Wiki-Primary sense in D_q

foreach $A \in D_q$ **do**

 | $P(A|d) = \sum_{k=1}^K P(A|\psi_k)P(\psi_k|d)$;

 | where $P(A|\psi_k) = \frac{1}{|\{a \in A_{img}\}|} \sum_{a \in A_{img}} \mathbb{KL}(\pi_a, \psi_k)$;

end

 Assign d to the Wiki-primary sense with the highest $P(A|d)$;

end

query to Google Image Search. We retrieve about 500 images for each keyword. In total, there are 5013 keyword images. We do the same for the 64 keyword senses, each time issuing to the image search engine with the expanded keywords as search query. We retrieve about 200 images for each keyword sense, totaling 12,336 sense images.

All images were automatically downloaded by following the image URLs on the Google image result index page. For each image, we also retrieve their surrounding text. To do this, we first remove all HTML and meta-tags, and retain only the non-markup words. Then, centered on the HTML location of the image, we collate all the adjacent 100 words and take them to be the textual content of the image.

Our image labeling procedure is briefly described as follows. For each image of a particular keyword, three human labellers were given the list of the word senses, and they were asked to choose only one (dominant) sense. An extra “None” label is defined on images where the object was too small or occluded.

The keyword images and the sense-specific images serve different purposes. For the keyword images, our intent is as follow. Because each keyword is polysemous and can have different meanings, the return images will contain a mix of images representing the various meanings of the keyword. The efficacy of our visual sense models can be evaluated on how well they can classify each of these images into their respective meaning.

For the sense-specific images, they are used for the following. Firstly, as mentioned in Section 4.2.2, we constrain our unsupervised clustering of the polysemous images with exemplar images depicting the various senses of the polysemous keyword. Secondly, these sense-specific images are used to create a sense-specific image classifier that becomes a baseline comparison for our visual sense model. For example, by issuing the expanded query “Mouse Computing” to Google Image Search, we can retrieve images of computer-mouse and learn a computer-mouse classifier. Clearly we can compare this classifier to our visual sense model by their classification results on the polysemous “Mouse” images.

4.3.2 Visual Features

We use a dense over-sampling approach to represent images. Local image regions are extracted from three sources: Difference of Gaussian (DoG) interest points [105], Maximally Stable Extremal

Regions (MSER) [109], and overlapping rectangular grid. DoG and MSER regions can be viewed as complementary to each other, sampling blob-like regions and high contrast image structures. Similar dense representation has been successfully used in the context of supervised object and scene category recognition [90, 99]. For each region, the SIFT [105] descriptors are then computed using the VLFeat [161] toolkit, and assigned to the nearest visual word from a visual codebook learned on a separate dataset using K-means clustering.

4.3.3 Baselines

Our goal is to re-rank the polysemous images from the ambiguous keyword query into its various visual senses. Naturally, our first baseline is the existing Google rank of the polysemous images. The choice of this as a baseline seems contrived because by design, Google never intends to rank images into their visual senses. Nonetheless, we retain this baseline for the sake of comparison.

As our second baseline, we use an unsupervised method based on spectral clustering to group images that are iconographically coherent. Spectral methods use eigen-decomposition to compute non-linear clustering of high-dimensional manifold data [119]. In the image domain, they have found successful applications in clustering web images [104, 35]. Spectral clustering works by first constructing a graph Laplacian $L = D^{-1/2}WD^{-1/2}$ of the image data, where W is the pairwise image-similarity matrix, and D is the diagonal matrix such that $D_{ii} = \sum_j W_{ij}$. The first p eigenvectors of L are then taken and arranged as columns in a new matrix Y . The rows of Y are then normalized, and the final cluster assignment is obtained by using the traditional k-means to cluster the rows of Y . The number of k in the final k-means clustering is set to be the same as the number of Wiki-primary senses. Following [119], we encode the affinity matrix W_{ij} between image i and image j as a combination of cosine similarity between text feature vectors and χ^2 distance between visual words histograms in images: $W_{ij} = \exp(-(1 - \cos_{ij}^t) - (1 - \chi_{ij}^2))$. For brevity, we shall call this baseline method SPEC-CLUST in short.

Because we have sense-specific images by issuing Wiki-disambiguated queries, our third baseline is to bootstrap sense-specific classifiers from these images. We call this method Sense-Specific SVM (SS-SVM in short). While we expect that these images can be more homogeneous as a result of increase query specification, polysemy will nonetheless be a problem in learning the sense-specific SVM (see Figure 4.5). In contrast, our approach in this chapter resolves these issues by incorporating

a latent model of the visual senses of the original polysemous keyword. The key idea is that in these images, there is a rich source of information about the various senses (visual or textual content) of the word. These visual senses capture the salient visual (and textual) characteristics of images associated with the keyword, and offer a more robust model than learning on just the Wiki-sense-specific images.

Our fourth baseline is the visual sense model described in [165]. In that work, a latent image model based on LDA (we shall call it VS-LDA, in short for Visual Sense-LDA) is similarly trained on polysemous images. VS-LDA differs from VS-HDP in that for each polysemous keyword, the number of global latent topics (K) is *hard-coded* to be twice the number of Wiki-senses of that keyword. This number is based on the intuition that there are *more* visual topics spanning the polysemous image data-sets than that specified by Wikipedia. In contrast, this number is now completely automated by the non-parametric approach described in this chapter. We shall call our new model VS-HDP in short.

4.3.4 Evaluation

We now evaluate how well the four algorithms (SPEC-CLUST, SS-SVM, VS-LDA and VS-HDP) can re-rank the polysemous keyword image dataset using a classification approach. For each of the Wiki-primary senses of every query keyword, we train sense-specific image classifiers and apply them to the polysemous images³. Images are re-ranked by their classification confidence, where lowly-scored images are moved down to the rank. For example, for each query keyword, a multi-class SS-SVM is trained on the sense-specific images for each sense class of the query keyword. The trained SS-SVM is then used to classify polysemous images into the various senses. For VS-LDA and VS-HDP, we similarly train the two models on the polysemous images. compute $P(S_i|d)$ for each image d using Equation 4.10, and rank the corresponding images according to the probability of each sense S .

We evaluate the retrieval performance using receiver operating characteristic (ROC) by thresholding $P(S|d)$ for every sense S of a keyword. Due to space constraint, Figure 4.2 shows the ROCs for the first Wiki-primary sense of each keyword. The dark-blue lines are the ROCs for the original Google search ranks. The cyan lines are the ROCs using the sense-specific SVMs to re-rank the

³For SPEC-CLUST, our classification-based evaluation framework poses some problem. Because SPEC-CLUST is an unsupervised clustering method, it is unknown which output cluster correspond to the query sense currently being evaluated. We resolve this by manually inspecting the clusters and introspectively labeling the one that best group all images with the query sense. We treat the centroid image of the labeled cluster as the sense prototype, and use the W_{ij} similarity values as classification scores to rank all other images

Google search image order. The orange lines show the ROCs of the SPEC-CLUST baseline, while the red and green lines are the ROCs obtained by VS-LDA and VS-HDP respectively. Table 4.2 shows the total Area Under Curve (AUC) for all senses of each keyword.

Keyword	Google rank	SPEC-CLUST	SS-SVM	VS-LDA	VS-HDP
bank	3.18402	4.8934	5.1769	5.4252	5.50356
bar	2.48154	3.3478	3.94088	4.2184	4.6842
bass	3.71719	5.5345	5.47068	5.9037	6.21209
mouse	1.73422	2.2279	2.41485	2.5194	2.4917
plant	5.03565	6.8972	7.45235	8.0146	8.06207
speaker	2.07675	2.4125	2.82301	3.1466	3.04853
temple	2.1008	2.9134	2.90866	3.1084	2.88802
tiger	3.87138	6.8976	7.72587	8.1875	8.04058
watch	2.06972	3.3231	3.7881	4.0434	3.98498
window	1.93045	3.4665	4.05622	4.2138	4.24679
Total-AUC	28.2017	41.9139	45.7575	48.781	49.1625

Table 4.2: Area Under Curve (AUC) of all senses of each keyword

From the results, there are a few notable observations. Firstly, all baseline models, including both VS-LDA and VS-HDP visual sense models, are able to retrieve far more positive class images than the original Google ordering. While this should not come too much as a surprise, since many authors have also found similar deficiency in Google rank [140], our results are achieved without the need for any training data. Perhaps more surprising is that the re-ranking results of our visual sense model, a generative model by design, outperform those of SS-SVM, a discriminative model. We postpone the discussion on this to the next section. Finally, between the two visual sense models, VS-HDP model has produced the better overall performance. This points to the advantage in adapting the sense modeling to the level of polysemy in the images. We show the clustering results on some keywords in Figure 4.3. Figure 4.4 summarizes the first five visual senses for all ten keywords.

4.3.5 Analysis and Discussion

Difficulty for SS-SVM to learn on polysemous images. A likely reason for the shortcoming of SS-SVM is that the sense-specific images on which it is trained are also fraught with polysemy (see Figure 4.5), making SS-SVM learning difficult. To further illustrate this problem, we show an example of the Wiki-disambiguated query “Mouse-computing” in Figure 4.6. In the figure, we compute the average

Figure 4.2: ROC plots of the first primary sense of the ten polysemous keywords

Figure 4.3: Example visual sense clusters on four polysemous keywords: “Tiger”, “Watch”, “Bass” and “Mouse”. For each polysemous keyword, we show enclosed in a green box two rows of sample polysemous images returned by Google Image Search. Note that because the first Google page of top rank images usually depict the dominant sense of the query keyword, they are fairly homogeneous. Hence, for our illustrative purpose to show sample polysemous images, we randomly take images that Google ranks from 50 onwards. Below the green box, we show the automatic clusters of images enclosed in red boxes.

bank					
bar					
bass					
mouse					
plant					
speaker					
temple					
tiger					
watch					
window					

Figure 4.4: Visual senses of the ten polysemous keywords. For each keyword, five visual senses are shown. Each visual sense represents a cluster of the images retrieved from the polysemous keyword. From each cluster, we show two images which are visually informative. As can be seen, the primary semantic senses of the polysemous word are captured by the cluster images.

Figure 4.5: Google search results on a Wiki-disambiguated query “Mouse computing”. While results are more homogeneous than that of “Mouse”, polysemy clearly remains an issue.

Figure 4.6: Using the average image to illustrate the visual polysemy of images. Top: even for images retrieved using the Wiki-disambiguated query “Mouse computing”, they are very diverse. This will be a challenge to a supervised learning method. Bottom: average images of three HDP generated clusters on “Mouse computing” images. Note the fine-grained elucidation of “notebook”, “pointer” and “desktop” images.

Figure 4.7: Comparative AUC performance across various values of the number of mixture components.

image to visualize the visual polysemy (and diversity) of images. The top figure shows the average image of images retrieved using the expanded query “Mouse-computing”. Note that even with the help of this disambiguation so as to retrieve mouse images in the “computing” sense, there is little visual structure in these sense-specific “Mouse-computing” images. In contrast, the images in the three HDP clusters shown in the bottom figure can be seen to exhibit more visual structures. Using HDP, further fine-grained image senses are elucidated, wherein “Mouse-computing” images are now clustered into the “notebook”, “pointer-device” and “desktop” senses.

Effect of K on AUC results. We take a closer look at the two most competitive models, the VS-LDA and VS-HDP. While the overall performance difference of the two models over the ten keywords are statistically insignificant, the main advantage of the VS-HDP model is that it circumvents the need for a prior choice of the number of mixture components K . In contrast, the VS-LDA model needs to define K a-priori, and the optimal value of K is necessarily determined via repeated trials and cross-validation. We show this in Figure 4.7 on two query words (“Bass” and “Speaker”), but results are similar for the other query words. We plot the comparative performance of the parametric models against the non-parametric VS-HDP across varying K . Apart from VS-LDA, we also use a truncated version of the VS-HDP model ⁴ as another parametric model. Note the different values of K for which their normalized AUC is nearest to the best obtained from VS-HDP: 35 for “Bass”, and 25 for “Speaker”. The higher K for “Bass” reflects its greater ambiguity than “Speaker” (“Bass” has eight Wiki-senses, while “Speaker” has five). For VS-LDA, its AUC results are most sensitive to the value of K . It peaks around the empirical optimal point before deteriorating.

Effect of injecting sense-disambiguated images. The appeal of unsupervised clustering methods is that there is no need for manual labels. However, their resulting clusters are generally of lower quality than those generated with supervision. On the other hand, it is often possible to improve clustering results by including some form of limited supervision. For example, pairwise constraints are often

⁴Note that in HDP, even though the CRP prior allows an infinite multinomial distribution over the mixture components, each image nevertheless learns a posterior distribution $\hat{\pi}$ over only a *finite* subset of cluster partition that are actually assigned to image feature vectors. Following [77], we use a truncated DP approach, and truncate β at $K + 1$, so that $\beta_k = 0$ for all $k > K + 1$. Hence, $\hat{\pi}_j \sim \text{DP}(\alpha^\pi, \beta_{1,\dots,\infty})$ becomes $\hat{\pi}_j \sim \text{Dirichlet}(\alpha^\pi, \beta_{1,\dots,K+1})$. Specifically, for each image j , we let $\pi_{j,1..K} = \hat{\pi}_{j,1..K}$, and $\pi_{j,K+1} = 1 - \sum_{k=1}^K \hat{\pi}_{j,k}$, where $\pi_{j,K+1}$ denotes the probability of sampling from a mixture component that has not been used to explain any image feature vector.

Figure 4.8: Including more sense-specific images improve AUC performance of VS-HDP.

used to indicate that two particular data points should belong to same (or different) clusters. In this chapter, the use of sense-disambiguated images (suggested by Wikipedia disambiguation links) can be seen as an injection of supervised labels onto the unsupervised HDP clustering framework. The inclusion of these sense-specific images constrains the clustering so that fine-grained sense clusters can be obtained (see Figure 4.6). On first thought, it would seem that these sense-specific images already encode the semantic senses that we want. However, this is not the case, since even with these expanded queries, return images are not homogeneous. The important thing to note is that it is not just the inclusion of *more* data that matters (because much of these data are also noisy), it is the judicious combination of the adaptive strength-sharing machinery of HDP, that results in the fine-grained sense clustering. Figure 4.8 charts the performance improvement as more sense-specific images are used to constrain VS-HDP clustering. AUC results rapidly improve early on but plateau off when hundreds of images are used. We note that this is the range where the return Google images start to be of lower precision, i.e. noisier. In our dataset, we crawled 200 sense-specific images for each keyword sense (see Section 4.3.1)

Chapter 5

FACETED RETRIEVAL OF NEWS VIDEO

In this chapter, we present a faceted topic retrieval system for news *video*. Faceted topics in news video are mined using probabilistic topic modeling, a suite of statistical methods designed to automatically analyze a large corpus of documents to discover its thematic information. The main advantage of the probabilistic topic modeling approach is its unsupervised methodology: there is no need for prior document labels. This allows for the automatic organization and summarization of huge amount of content.

A probabilistic topic model associates with each document a probability distribution over topics, which are in turn distributions over words. The semantic content in a topic can be discerned from the high probability words that it contains. For instance, if the high probability words in a topic are words like “football”, “tennis” and “score”, then we may postulate that this is a “sports” topic. In the same way, in a “finance” topic, examples of high probability words would be words like “bank”, “economy” and “Dow Jones”¹. There are two main learning goal in topic models: (a) learn a number of (distinct) topics by assigning (different) probabilities to words, and (b) learn the most probable topic proportions of documents.

This modeling paradigm is particularly appealing to our objective of modeling topics and facets in news content. The Latent Dirichlet Allocation (LDA) is the simplest probabilistic topic model,

¹Note that the topic labels such as “sports” and “finance” are post-hoc labels that are given by a human assessor, by inspecting the high probability words.

and following the seminal work of Blei and Jordan [25], it has been widely studied and is one of the most popular topic model. LDA and many of its extensions have been widely studied and deployed in many applications and domains [85, 37, 124]. In this thesis, we have also grounded our faceted topic retrieval framework on the LDA.

5.1 Motivation

Despite the general success in applying LDA to mine semantic topics in many application domains, its use for the faceted topic retrieval (FTR) in the video domain poses the following new challenges. First, the naive approach of directly applying LDA to process the textual speech transcripts from Automatic Speech Recognition (ASR) will likely not produce good results. The reason is that with ASR word error rates generally above 20%, these ASR transcripts will be noisy. We note that with the exception of a limited number of reported works in some non-traditional text domains such as meeting dialogue [124], the successful application domains of LDA are mostly limited to those with clean text, such as newswire and research publications. In a work that is more closely similar to ours, Cao *et al* examined LDA modeling and retrieval of ASR-based spoken news transcripts [37], and reported mixed results.

Another issue in using LDA for faceted retrieval of video concerns the quality of LDA output. Specifically, the utility of LDA to extract coherent topics remains an open issue [113]. Often times, LDA produces word distributions that are coarse, with no apparent meaning amongst high probability words. The lack of topic coherence results in facets that are only weakly related to documents, and this will have a direct impact on the quality of faceted topic retrieval. The common approach to deal with this problem is to introduce side-information into the modeling [25, 85]. In this chapter, this complementary side-information comes in the form of visual information in the shot keyframes, and we use it to constrain facet development. Our approach is motivated by two intuitions: First, video footages are often repeated for similar or related news stories, and hence are highly correlated with the spoken (ASR) words. Second, different facets of a query may use different sets of words, and the same set of recurring visual shots become a “bridge” between these words, allowing them to be learned as distinct facets. In this chapter, the recurring visual shots are computed using Near Duplicate Image (NDI) detection [45]. Combining with the ASR textual words with the NDI visual

Figure 5.1: FDM for hypothesizing facets in news video.

shots, we define a multimodal Faceted Document Model for news video to jointly account for both information streams as distinct but correlated sets of observations.

5.2 A Faceted Document Model of News Video

In this section, we present a Facet Document Model (FDM) to hypothesize a set of k facets $\{f_1, f_2 \dots f_k\}$ in a news video collection D . Our goal is to compute $P(f_j \in D)$. Our overall approach is shown in Figure 5.1. Two feature extraction tracks act on an input news story video simultaneously to compute text and visual features respectively. Their outcomes are then combined using a generative model to compute facets. We discuss each of these key modules below.

5.2.1 Text features

While important keywords are generally detected by ASR, the presence of the many misrecognized words can result in erroneous topic formation. Apart from stemming and discarding rare words, we use the document expansion approach of Wan [164] to alleviate the problem. The main idea is to introduce additional words to form an expanded text document vector. These words are selected based on their high mutual information in a parallel news corpus. Such a corpus is readily obtainable since news content is widely available over the internet. For the TRECVID-2005 dataset used in

Table 5.1: Sample expansion words given the words on the left.

Condoleezza	state, bush, secretary, security, stanford
Jintao	china, brazil, sino, economic, aids
Basketball	points, conference, group, match, nba

Figure 5.2: TFIDF weighting of keyframe-based visual words. Within the 4 videos, four NDI clusters are shown and colored differently. For the red NDI cluster, it appears in V_2, V_3, V_4 , hence its $df=3$, $tf=2$ for V_3 , $tf=1$ for V_2, V_4 . All non-NDI in a video have $df=1$ and $tf=1$. Best viewed in color.

this paper, we build the parallel corpus by issuing to Google Archives News our query description and restricting retrieval time-range to the period when the dataset was collected (Nov-Dec 2004).

Table 5.1 shows some examples of expansion words.

5.2.2 Visual features

To capture a higher level of visual information, we generalize the common practice of modeling images as bag-of-features to represent video as *bag-of-keyframes*. Following the approach of Wu *et al* in [172], a keyframe is classified as whether it is a Near Duplicate Image (NDI) to other keyframe(s) or not. By assigning unique IDs to keyframes, they can be treated as visual words. All keyframes in a NDI-cluster are visually similar and are given the same ID. We can now generalize the TD-IDF weighting in text domain to these visual words. Figure 5.2 shows some examples of how visual similarity in videos is succinctly represented by the term frequencies (tf) and document frequencies (df) of visual words.

To compute near duplicates images, we use a color histogram combined with a spatial pyramid

Figure 5.3: Top shows the spatial division of the image at each level of the RGB pyramid. At each level, the RGB suffix refers to the number of bits in the color channel. E.g. at level-1, we have 4 divisions each with 3 bits red, 4 bits green, 3 bits blue, totaling 128 bins. Each bin count uses 2 bytes. Therefore, each image is represented by 768 bytes. Bottom shows sample NDI given the leftmost image as query.

over the image to jointly encode global and local information. This approach is inspired by Chum *et al* [45], who applied the method to efficiently handle NDI detection amidst jitter and noise. Figure 5.3 shows the spatial pyramid configuration which is arranged so that an increasingly granular grid (i.e. from global to more localized) of color information is stored as we move up the level. The setup provides a highly compressed representation for each image that makes histogram comparison efficient. Given a query image, the NDIs are defined to be those within a specified Euclidean distance from the query. To compute a NDI cluster, we maintain a NDI list that initially only contains the query image. This list is then repeatedly populated with NDI of the new members in the list. The final NDI cluster is then given by the transitive closure of the NDI list.

5.2.3 Joint Modeling of Visual and Text

We now consider a corpus $\mathcal{D}$ of news video, each comprising of text $\mathbf{W}$ and visual words $\mathbf{V}$. Each video d is modeled as a mixture of latent topics, to *simultaneously* account for $\mathbf{W}$ and $\mathbf{V}$ as distinct set of observations. Our model is motivated by Blei and Jordan’s Corr-LDA model for text and images [25], and also Jain *et al*’s People-LDA model in [85]. We call our model FDM-VT, denoting

Figure 5.4: Graphical representation of FDM-VT. The red box encloses the additional observations from the visual modality, which constrain the topic formation in the text modality.

the use of both visual and text modality. Figure 5.4 shows the graphical representation of FDM-VT ².

The generative process of FDM-VT is as follows:

- Draw a multinomial ϕ over K topics: $\phi \sim \text{Dir}(\alpha)$
- For each topic $k = 1 \dots K$,
 - draw multinomial $\beta_k \sim \text{Dir}(\eta_w)$ for text words
 - draw multinomial $\gamma_k \sim \text{Dir}(\eta_v)$ for visual words
- For each text word index n in d , $n = 1$ to N_d ,
 - sample a topic z_n from ϕ : $z_n \sim \text{Multinomial}(\phi)$
 - sample a text word w_n from β_{z_n}
- For each visual word index m in d , $m = 1$ to M_d ,
 - draw an indexing variable $y_m \sim \text{Uniform}(1, \dots, N_d)$
 - sample a visual word $v_m \sim \text{Multinomial}(\gamma_{z_{y_m}})$

where N_d and M_d are respectively the numbers of text words and visual words in video d , and η_w and η_v are Dirichlet priors for the text and visual words distribution respectively. The above FDM-VT

²Note that this model differs slightly from the one shown in Figure 3.1. In the latter, there is a directed link from z to v , which is absent here. The source of the link to y is from N_d in Figure 3.1, while in this model, it is from ϕ . While there are structural changes resulting from these topological deviations, these differences are eventually subsumed by the variational distribution used in the approximate inference. Both have identical variational simplification process in Equation 3.8 and Equation 5.3. The learning algorithm is exactly the same for both graphical models.

model results in the following joint distribution on text $\mathbf{W}$, visual $\mathbf{V}$ and the latent topics:

$$P(\mathbf{W}, \mathbf{V}, \phi, \mathbf{z}, \mathbf{y}) = P(\phi|\alpha) \left(\prod_{n=1}^{N_d} P(z_n|\phi) P(w_n|z_n, \beta) \right) \left(\prod_{m=1}^{M_d} P(y_m|\phi) P(v_m|y_m, \gamma) \right) \quad (5.1)$$

The main difference of our model from the Corr-LDA in Blei and Jordan [25] is that we use two multinomial distributions for the text and visual words. The sampling of visual words is essentially the same as the sampling text words. However, within a video, ϕ is a higher-level factor that is held fixed and it governs the ensemble of all text word and visual word observations. The topic-word multinomial β and γ now learn the combined co-occurrence of important text words across video documents and also the complementary visual words.

5.2.4 Inference and Estimation

The parameters we want to estimate are: α, θ and γ . Marginalizing over the latent variables ϕ, z and y , we re-write equation 3.6 as:

$$p(\mathbf{v}, \mathbf{w}|\alpha, \theta, \gamma) = \int_{\phi} \sum_z \sum_y p(\mathbf{v}, \mathbf{w}, \phi, \mathbf{z}, \mathbf{y}|\alpha, \theta, \gamma) d\phi \quad (5.2)$$

The above is intractable due to the coupling between ϕ and γ in the summation over the latent variables z . We follow the general recipe of variational inference and derive the inference algorithm. First, we approximate the posterior distribution over latent variables given the observed data (visual words v and text w). In the following, we highlight the key inferencing steps, while reserving the full details of the derivation in Appendix B.

Variational Simplification: We first define a simplification to the original graphical representation by introducing three variational parameters χ, δ and λ (see Figure 5.5) We then write the factorized variational distribution on the latent variables as:

$$q(\phi, \mathbf{z}, \mathbf{y}|\chi, \delta, \lambda) = q(\phi|\chi) \prod_{n=1}^{N_d} q(z_n|\delta_n) \prod_{m=1}^M q(y_m|\lambda_m) \quad (5.3)$$

Figure 5.5: Variational Simplification for learning of FDM-VT.

where χ is the variational Dirichlet parameter, δ_n is the variational multinomial parameter over K topics, and λ_m is the variational multinomial parameter over M visual words.

Next, we minimize the KL-divergence between this factorized distribution and the true posterior. Taking derivatives with respect to the variational parameters, we obtain the following coordinate ascent algorithm:

1. Update the posterior Dirichlet parameters:

$$\chi_i = \alpha_i + \sum_{n=1}^{N_d} \delta_{ni} \quad (5.4)$$

2. For each text word, update the posterior distribution over the latent factors:

$$\delta_{ni} \propto \theta_{i,w_n} \exp \left(\Psi(\chi_i) - \Psi\left(\sum_{j=1}^K \chi_j\right) + \sum_{m=1}^M \lambda_{mn} \log \gamma_{iv_m} \right) \quad (5.5)$$

3. For each visual word, update the posterior distribution over the anchoring text words:

$$\lambda_{mn} \propto \exp \left(\sum_{i=1}^K \delta_{ni} \log \gamma_{iv_m} \right) \quad (5.6)$$

These update equations are invoked repeatedly until the change in KL divergence is small.

Parameter Estimation: After we have maximized the lower bound with respect to the variational parameters χ, δ, λ , we next maximize the bound with respect to the model parameters α, θ, γ . Given a training set $\mathbb{T} = \{(\mathbf{v}_d, \mathbf{w}_d)\}_{d=1}^D$, we find the maximum likelihood estimation for visual word

topics γ , text topics θ and the prior α . The overall log likelihood of $\mathbb{T}$ is bounded by:

$$\mathcal{L}(\mathbb{T}) = \sum_{d=1}^D \log p(\mathbf{v}_d, \mathbf{w}_d | \alpha, \theta, \gamma) \geq \sum_{d=1}^D \mathcal{L}(\chi_d, \delta_d, \lambda_d; \alpha, \theta, \gamma) \quad (5.7)$$

The final maximization-step for the conditional multinomial θ is:

$$\theta_{ij} \propto \sum_{d=1}^D \sum_{n=1}^{N_d} \delta_{dni} w_{dn}^j \quad (5.8)$$

$w_{dn}^j = 1$ if the n th text word in document d is the j th word in the text vocabulary, else $w_{dn}^j = 0$.

The maximization-step for the conditional multinomial γ is:

$$\gamma_{ij} \propto \sum_{d=1}^D \sum_{m=1}^{M_d} v_{dm}^j \sum_{n=1}^{N_d} \delta_{dni} \lambda_{dmn}, \quad (5.9)$$

$v_{dm}^j = 1$ if the m th visual word in document d is the j th word in the visual word vocabulary, else $v_{dm}^j = 0$. Finally, for the Dirichlet prior α , we follow the approach in [27] and use a Newton-Raphson algorithm to maximize α .

5.2.5 Re-ranking

After the FDM-VT learning has converged, the K latent topic distributions of both the text words and visual words are given by β_k and γ_k , $k=1..K$. In particular, the probability of a text-word w in the k^{th} latent topic is given by $P(w|\beta_k)$. Given $d_i \in \mathbf{D}$, we have:

$$P(d_i|\beta_k) = \prod_{w \in d_i} P(w|\beta_k). \quad (5.10)$$

We now assume that the multinomial β_k represents the k^{th} actual facet f_k in $\mathbf{D}$. The probability of d_i containing the k^{th} facet is now given by:

$$P(f_k \in d_i) = P(d_i|\beta_k) = \prod_{w \in d_i} P(w|\beta_k). \quad (5.11)$$

We use the greedy heuristic to maximize facet retrieval: for each topic $k = 1..K$, we find the document d_i that maximizes $P(d_i|\beta_k)$ and *re-rank* it by its relative order in the original ad-hoc retrieval rank.

5.3 Experimental Evaluation

5.3.1 Dataset

A few corpora have been created to facilitate research in novelty and diversity tasks in text (TDT5 [40], TREC-2006 Q&A [49]) and images (ImageCLEFPhoto [121]). However, no corresponding dataset for video is available. To evaluate our methods, we turn to a subset of the TRECVID-2005 video dataset [3]. It comprises of about 127 hours of Chinese and English news broadcast from five different sources (e.g. CCTV4, CNN). The dataset also includes computed story boundaries from the CMU Informedia Lab.

For queries, we used the topic labels provided by Wu *et al*, who also used the same TRECVID dataset [172]. In their work, 33 story topics are defined on which each news story is manually labeled with one of the topic labels. Sample topic labels include “Bush visits Canada”, “Mideast-Peace”, “Arafat-health”. For the full list of topics, please refer to their paper.

For our FTR evaluation, three human assessors further annotated facets on these video topics. For each of the 33 topics, the assessors were given the videos labeled as relevant and asked to name the facets and provide some keywords. Assessors were given free reign to name and describe facets, but they were advised to consult online resources such as the Wikipedia and news commentary sites. On average, there are 43.5 relevant videos per query topic, 5.7 facets per query topic, and 3.1 facets per video. There are a total of 70 unique facets. While the assessors generally agreed on the number of facets per video, they disagreed more on the facets in a query topic. We mostly resolve this by taking a liberal and all-inclusive approach to accept all facet suggestions.

5.3.2 Comparing Retrieval Engines

We compare FTR performance with a well-known diversity retrieval model of Carbonell and Goldstein [39], called Maximal Marginal Relevance (MMR). This is a greedy method that selects the i^{th} document d_i according to a combination of its similarity to the query Q and its similarity to all higher-ranked documents at position 1 to $k-1$:

$$\mathbf{MMR}(d_i, Q) = \alpha \text{Sim}_1(d_i, Q) - (1 - \alpha) \max_{1 \leq j < i} \text{Sim}_2(d_i, d_j)$$

where $\text{Sim}_1()$ is a standard language model-based document-query similarity function with Dirichlet smoothing [2], $\text{Sim}_2()$ is a Cosine similarity function between documents, and α is a trade-off parameter between 0 and 1. When $\alpha = 1$, MMR ranking falls back onto the standard relevance ranking based on query-document similarity. We set $\alpha = 0.5$.

For our FDM modeling of video, we have the following ways to hypothesize facets:

1. Unimodal FDM modeling only on text:

- (a) **FDM-T-ASR:** on only the ASR speech transcripts.
- (b) **FDM-T-ASR+DE:** on the resulting text vector after document expansion on ASR transcripts.

2. Joint FDM modeling on text and visual features:

- (a) **FDM-VT-ASR:** visual features and ASR
- (b) **FDM-VT-ASR+DE:** visual features and text from document expansion of ASR transcripts.

For each of the above, we also have the following options:

- 1. **FDM on entire corpus:** hypothesize facets in the entire corpus (1028 videos). In our dataset, we have a total of 70 unique facets. So we set the number of latent topic in FDM modeling K to be 70.
- 2. **FDM on top- D documents:** For each query, hypothesize facets in the top retrieved videos. In our dataset, we have on average 43.5 relevant videos per query. So we round off $D = 50$. We also have an average of 5.7 facets per query, and so we set K to be 6.

While the latter setup simulates the results obtainable in interactive retrieval, real-time response is severely limited by the huge CPU demand on visual computing. Hence, the first setup is generally the only permissible option for real-life deployment. In all five retrieval methods, re-ranking (Section 5.2.5) is performed on the top-100 videos. For fair comparison, the same $\text{Sim}_1()$ document scoring function is used.

Table 5.2: FTR performance on TRECVID-2005 dataset. The best result in each FDM setting is in bold. An asterisk indicates significant statistical difference over the MMR baseline.

	$Prec@20$	$S-rec@20$	F1
MMR	0.3276	0.4704	0.3862
modeling on entire corpus ($K = 70$)			
FDM-T-ASR	0.2410	0.4472	0.3132
FDM-T-ASR+DE	0.3187	0.4591	0.3762
FDM-VT-ASR	0.2772	0.4521	0.3437
FDM-VT-ASR+DE	0.3306	0.4612	0.3851
modeling on top- D documents ($D = 50, K = 6$)			
FDM-T-ASR	0.2810	0.4618	0.3494
FDM-T-ASR+DE	0.3487	0.4707	0.4006
FDM-VT-ASR	0.3272	0.4834	0.3903
FDM-VT-ASR+DE	0.3706	0.4992	0.4254*

5.3.3 Results and Analysis

For each comparative FTR method, we report S-recall at the 20th rank. This assumes that users are interested in the first two pages of search results. While S-recall measures the retrieving of facets, it is nonetheless desirable to have a ranked list also populated by *relevant* documents. To capture both intent, we introduce an extra F1-measure that combines a standard precision metric with S-recall:

$$F1 = \frac{2 \cdot (Prec@20 \cdot S-rec@k)}{(Prec@20 + S-rec@k)}. \quad (5.12)$$

Table 5.2 reports both the S-recall and precision at rank 20.

All methods are able to retrieve nearly 50% of the facets, with FDM-VT-ASR+DE achieving the best overall F-score. Notably, this is obtained by modeling on the top- D documents, and the results from modeling on the entire corpus are marginally *worse* than the MMR baseline. This points to the general difficulty in topic modeling of noisy documents.

Consistent with results obtained elsewhere [164], modeling benefits from augmenting text by document expansion. The query with the lowest F-score is “Bush second term plan” (0.0478). While the S-recall is decent, this query suffers from very low precision score as it has a lot of overlap with queries such as “War-on-Terror”. The query with the highest F-score is “HIV aids” (0.7545). Again high precision plays a major part, since the topic is unique in the corpus.

Figure 5.6 compares the S-recall-vs-precision plots of three FDM models with the MMR baseline. While all four coincide closely, the text-only FDM-T-ASR model significantly underperforms

Figure 5.6: S-recall versus Precision plot for three FDM models and MMR baseline. Best viewed in color.

Table 5.3: Top-10 probability words of 2 learned FDM β_k topics, for “War on Fallujah” (top) and “Mideast peace” (bottom). The topics are picked introspectively. Meaningful words highlighted in bold.

FDM-T (text-only)	FDM-VT (Visual+text)
iraq peopl time good meet unit state baghdad chines dai	iraq iraqi peopl militari govern arm kill forc attack baghdad
arafat know yasser leader thing peac minist peopl just israel	palestinian arafat peac israel presid leader elect yasser east middl

compared to the rest, and the visual+text FDM-VT-ASR+DE model improves over the MMR baseline. In Table 5.3, we qualitatively show how the text-words multinomial β_k has benefited from the inclusion of visual features during topic learning. On each of two queries “War on Fallujah” and “Mideast peace”, we introspectively pick a learned FDM topic that contains high probability words that are meaningful to the queries. Observe that words from multimodal model are more intuitive and correlate better to the query topic.

5.4 Discussions

While the past decade has seen tremendous progress in robust document ranking, a major challenge remains when users underspecify their true information needs. As a step towards addressing this problem, we propose a faceted topic retrieval (FTR) model to mine the wide range of facets in documents. Implicit in our methodology to resolve query ambiguity is that by returning a set of diverse or novel results that are pertinent to the query, the average user will find something useful.

Accurate elucidation of facets depends on learned topics that are highly correlated with actual queried topics or events. In the case of media-rich content such as video, we argue that this can best be achieved by joint modeling in the complementary visual and text modalities. We applied and tested this approach to hypothesize facets in a news video search application. Using a variant of the LDA topic model to combine the visual keyframes and the speech transcripts, we show the efficacy of our method on the standard TRECVID-2005 news corpus annotated with facets. We find that the joint model can achieve significant F-score improvement over a text-alone system and that it compares favorably with standard diverse ranking algorithms such as the MMR. Motivated by our findings, we have implemented our FTR model on a live news search prototype that is undergoing commercial trial. As a further work, we will write a follow-up paper to provide more details and empirical findings in our real-world deployment.

A few serious criticisms can be levied against some of our assumptions. The first is the assumption that facets occur in documents independently. While this is admittedly a very weak assumption in practice, the performance impact is expected to be minimal, and ameliorated by the fact that the retrieval function is a relative ranking function, and hence the estimation of the facet hypothesis can afford to be poor while retrieval ranking remains high. This line of thought is inspired by a similar argument for the surprisingly superior performance of the Naive Bayes classifier despite its seemingly debilitating feature independence assumption [62, 122].

In our retrieval model, we have also conveniently made use of privileged information about our dataset such as the average number of relevant documents and number of facets per query topic, to provide a number for D and K in Section 5.3.2. While these parameters are not immediately available in a real-world setting, initial values for these parameters can still be determined using empirical methods such as interactive feedback and query log analysis. We have also seen in the previous

chapter how Bayesian non-parametric techniques such as the Hierarchical Dirichlet Process can be used to automatically determine K . The downside of these methods, however, is their computation time. But with recent significant advances in the development of parallel topic modeling [145, 177], one can remain optimistic of the applicability of these techniques for real-world deployment.

Last but not least, it appears that further work can be explored along the direction of how best to exploit the use of other visual features for better joint modeling.

Chapter 6

TOPIC BASED QUERY SUGGESTIONS

Query suggestion is an assistive technology mechanism commonly used in search engines to enable a user to formulate their search queries by predicting or completing the next few query words that the user is likely to type. In most implementations, the suggestions are mined from query log and use some simple measure of query similarity such as query frequency or lexicographical matching. In this chapter, we propose an alternative method of presenting query suggestions by their thematic topics. Our method adopts a document-centric approach to mine topics in the corpus, and does not require the availability of a query log. The heart of our algorithm is a probabilistic topic model that assumes that topics are multinomial distributions of words, and jointly learns the co-occurrence of textual words and the visual information in the video stream. Empirical results show that this alternate way of organizing query suggestions can better elucidate the high level query intent, and more effectively help a user meet his information need.

6.1 Motivation

Query formulation has long been a critical component in information retrieval, and most modern search engines now have mechanisms to help users refine their queries. One such mechanism is the auto-query completion that automatically suggests possible complete queries as a user type into the query box [60]. However, most existing query suggestion methods suffer from the following two

Figure 6.1: Left: Existing query suggestions; Right: Presenting suggestions by thematic topics: three topics relevant to the query “fallujah” are shown, each titled and underscored by the highest probability word in the discovered topic.

problems:

First, the query suggestions in most systems are obtained by utilizing a large-scale query log [11, 29]. These methods typically work by looking at the clickthrough patterns of similar queries. However, the applicability of such methods for customized search domains such as personal desktop search or intranet search is severely limited. This is because often times there is either no available query log, or the size of the user base and the number of past queries is too small for any meaningful query mining.

Second, query suggestions are mostly obtained using simple lexicographical matches and then ranked by their frequencies. The resulting suggestions often appear ad-hoc and disorganized. An example is shown on the left in Figure 6.1. Our approach is to identify the *topics* among the documents that are relevant to the original query, and to group the query suggestions according to these topics. This new layout is shown on the right in Figure 6.1. Note that query suggestions are now clustered into topics that are labeled by a keyword for easy navigation. The query suggestion “fallujah birth-defects” is also a new suggestion that gets presented as being part of the “fallujah uranium” topic. This allows a user to uncover possible new query interests that were hidden previously. The new presentation is clearly an improvement over the original.

Motivated by the above observations, in this chapter, we propose a document-centric topic-based query suggestion method for the video search domain. We base our topic-elucidation methodology on the latent dirichlet allocation (LDA), which aims to uncover the hidden thematic structures in documents [25]. Using LDA to model query suggestions offers the following two advantages: (a). the

relationships among the observed words, documents and latent topics are based on a theoretically sound probabilistic framework. (b). the learnt topics are multinomial distributions of words, from which the high probability words are natural candidates for query word suggestion.

6.2 Modeling Challenges in the Video Domain

While latent topic modeling has recently become a popular tool for unsupervised topic discovery in large document collections, there are a few challenges to its application to model topical structures in multimedia collections.

6.2.1 Noisy Textual Content

The most straightforward way of applying LDA to the multimedia domain is to naively feed the speech transcripts (usually from Automatic Speech Recognition, or ASR) of the video content as textual input to the topic modeling algorithm. However, this will likely not yield good results. This is because the ASR transcripts are noisy (ASR word error rates are generally above 20%), and may impact (reduce) the co-occurrence of semantically related keywords in the ASR content. As a result, the set of words that are learnt by a topic may not convey much meaning to a human reader. If such phenomena of semantic malaise occurs across multiple topics, it will lead to a general failure of the topic modeling framework to capture the “true” topical structures in the multimedia content [113, 118]. Note that most successful applications of LDA are reported on clean text such as newswire and scientific publications. While there has been a few sporadic work on noisy ASR content [124, 38], the utility of LDA to noisy text source remains unclear.

6.2.2 Joint Modeling of Multimodal Content Cues

One way to deal with the modeling of the noisy textual content in a video is to exploit the complementary information available in the other textual/auditory/visual content cues. A multimodal joint modeling approach can be taken by treating these additional information cues as side-information into the joint modeling [85, 25]. In this chapter, we follow these trends and explore the use of two additional content cues to constrain topic development: (1) textual information in a parallel news corpus, and (2) visual information in the shot keyframes. We show that while the inclusion of com-

plementary textual information will help produce topics that better reflect the topical structures of the video content, combining *both* textual and visual information will yield the greatest gain in topic coherence. We shall define how we measure topic coherence in the experimental section.

6.2.2.1 *Joint Modeling of Textual Information*

We exploit the fact that apart from the traditional news bulletin video broadcast, a news story is also published on online web portals. Hence, we attempt to *expand* the noisy ASR content of a news video content by associating it with the top few closest matched news articles from a parallel news corpus. Our motivating principle is as follow. While the ASR content is noisy and generally unreliable, there may still be some content keywords that are correctly recognized. Using these content keywords, there is a good likelihood of matching the ASR text to the correct news articles describing the same news event. By augmenting the words from these matched news articles to the original noisy ASR text, we have a more accurate textual description of the video content, upon which we can better obtain topics of a greater semantic coherence.

6.2.2.2 *Joint Modeling of Visual Information*

We also jointly model the visual information in the shot keyframes to obtain better topics. There are two motivating intuitions: First, video footages are often repeated for similar or related news stories, and hence are highly correlated with the spoken (ASR) words. Second, different topics of a query may use different sets of words, and the same set of recurring visual shots become a bridge between these words, allowing them to be learned as distinct topics. To compute recurring visual shots, we use the Near Duplicate Image (NDI) detection method of Chum *et al* [45].

6.2.2.3 *Variational Inference and Parameter Estimation*

We derive a variational inference and parameter estimation algorithm for the combined modeling of the multimodal content cues that jointly account for the ASR words, the parallel news corpus, and the visual NDI shots as distinct but correlated sets of observations. As baseline, we use the model that is learnt by applying standard LDA solely on the ASR text transcripts. We then learned two different joint models by incrementally adding more content cues on top of the ASR text transcript baseline: (1) combining the ASR text transcripts with matched text articles from a parallel news

corpus, and (2) combining (1) with visual information from the shot keyframes. We compare the results from these two models using an objective topic coherence measure and subjective user tests.

6.2.3 Labeling Query Suggestions

The third challenge relates to how to choose words from a learnt topic to label suggestions. The difficulty arises because the semantic *theme* of a learnt topic is only *collectively* conveyed by the high probability words, without any preferential order. Hence, a judicious choice of words from these high probability words is still needed to yield meaningful suggestions.

6.3 Existing Works

Query Completion: Machine-predicted text was first used mainly as an assistive mechanism for the physically handicap, but in recent years, has also found widespread benefit for mobile web search [47]. Today, query completion is featured in all major search engines. However, the focus of most work is on deriving a set of semantically related queries [179], rather than clustering them by topics. Jain and Mishne [84] proposed clustering query suggestions by simple phrasal keywords. There are only a few works on query suggestions without query log. Bhatia *et al* [22] proposed a probabilistic model for generating query suggestions from the corpus, again using phrasal keywords. However, several limitations of phrasal keywords as a text summary have been highlighted in [107], such as low recall and ignoring the order of the query terms.

Topic Coherence: Because the quality of a query suggestion is dependent on the corresponding quality of a topic, the notion of semantic coherence in a topic is relevant to this chapter. While it is well established that the task of topic modeling is to learn sets of words that capture the latent semantics of a document or document collection, the task of evaluating the quality of a learnt topic is not as well developed. In many cases, an *quantitative* evaluation methodology were used to demonstrate the effectiveness of the learned topics, e.g. model perplexity [27, 25], but less attention is given to the *qualitative* understanding of the semantic nature of the learnt topics. In this chapter, we propose an computational measure of the semantic coherence of a learnt topic, and use it to compare the ability to the multimodal joint modeling methods to obtain topics of better quality.

Joint Topic Modeling: Topic models are first derived as multinomial distributions of unimodal text data, and the joint modeling of multiple data types such as visual and text is not straightforward. The authors in [114, 103] model annotated images using the visual features and text annotations, for automatic annotation and retrieval respectively. Two ways of combining the two modalities are explored: feature concatenation and hierarchical modeling. The former treats the two modalities equally, while the latter first models each individually and then fuses them at a later stage. Our work in this chapter differs from the above in that we perform joint modeling of visual features and text in the video domain. Our modeling granularity is coarser: our visual features are not at the local patch-level but rather at the keyframe-shot-level. This most closely resembles Wu *et al*'s video representation with visual shot duplicates [172].

Because topic models are unsupervised methods, often best results are obtained by incorporating a priori knowledge about the desired output (e.g. must-link constraints in clustering). Adding observations from cross-media types as a way to constrain topic modeling has been proposed by several authors. Jain *et al* guide topic formation of news photo caption by correlating the names with a face recognizer [85]. Blei and Jordan describe an image annotation model to learn the correspondence between an image region and a word in the caption [25].

6.4 Topic Based Query Suggestions

In this section, we describe our topic based query suggestion method. Figure 6.2 shows the overall processing flow. Given an input news video story, we apply feature extraction on the video simultaneously to compute textual (ASR words from speech recognition) and visual features (visual words). Additional textual features (text words) are also processed from a parallel news corpus. All these are then jointly combined using a generative model to compute topics. We focus on two particular joint models. The first is a text based joint models, where the ASR words are jointly modeled with the text words from the parallel news corpus. We name this the cLDA-T model, where the prefix *c* denotes we are defining a *combined* LDA model, and the hyphenated suffix *T* denotes the *text* model. The second is a combined visual+text model, where the ASR words, text words and visual words are jointly modeled. We name this the cLDA-VT model, where the hyphenated suffix *VT* denotes

Figure 6.2: Joint topic modeling framework for hypothesizing topics in news video.

the *visual+text* model. After learning, the output of each of joint model is a sets of topics computed from the collection of news video stories. Given a query, we formulate a way to rank and select the topics according to their relevance to the query. For each of the selected topic, we then select the high probability terms to label the topic. In the following, we describe the two joint models in subsection 6.4.1 and the topic and term selection in subsection 6.4.2

6.4.1 Joint Modeling

Our overall joint modeling approach is shown in Figure 6.3. The goal of joint modeling is to obtain better topics in terms of topic coherence. To achieve this, we define our joint model as a variant of the standard Latent Dirichlet Allocation (LDA) model (see Top of Figure 6.3). Compared to the standard LDA model which only deals with a single stream of observed data, we now present to our

Figure 6.3: Top: standard LDA model. Middle: Joint Text model (cLDA-T) Right: Joint Visual+Text model (cLDA-VT)

joint model an additional observed stream of complementary data, on which we formulate a new generative model that accounts for *both* streams of data.

6.4.1.1 Joint Text Modeling (cLDA-T)

We describe first the joint text model cLDA-T, which is shown in the middle of Figure 6.3. The cLDA-T model combines the ASR words and the text words from a parallel news corpus. As shown in the figure, we introduce a dependency relation between the N_d ASR words of a document d and the P text words in a news article from the parallel news corpus. We assume that this news article has been separately identified by a matching process ¹. This dependency is formed by making the latent topics of both these text sources depend on the document topic proportion ϕ . This means that we are imposing the constraint that the latent topic Z_{ASR} of the ASR words and the latent topic Z_{text} of the matched news article must come from the same topic proportion ϕ . By sharing the same ϕ , the joint text model is able to simultaneously account for the co-occurrence of both text words arising from the video content. This results in the formation of topics that are more coherent. The generative process of cLDA-T is as follows:

- Draw a multinomial ϕ over K topics: $\phi \sim \text{Dir}(\alpha)$
- For each topic $k = 1 \dots K$,
 - draw multinomial $\theta_{ASR,k} \sim \text{Dir}(\beta)$ for ASR words
 - draw multinomial $\theta_{text,k} \sim \text{Dir}(\beta)$ for text words (from matched news article)
- For each ASR word index n in d , $n = 1$ to N_d
 - sample a topic z_n from ϕ : $z_n \sim \text{Multinomial}(\phi)$
 - sample a ASR word $w_{ASR,n}$ from θ_{ASR,z_n}
- For each text word index $p = 1$ to P
 - sample a topic z_p from ϕ : $z_p \sim \text{Multinomial}(\phi)$
 - sample a text word $w_{text,p}$ from θ_{text,z_p}

¹this can be done by a simple text match using a standard cosine similarity score on the word features.

where N_d and P are respectively the number of ASR words and text words in the video document d , and β is a Dirichlet prior for both the ASR word and text word distribution. The above cLDA-T model results in the following joint distribution on the ASR words $\mathbf{W}_{\text{ASR}}$, the text words $\mathbf{W}_{\text{text}}$ and the latent topics:

$$P(\mathbf{W}_{\text{ASR}}, \mathbf{W}_{\text{text}}, \phi, \mathbf{Z}_{\text{ASR}}, \mathbf{Z}_{\text{text}}) = P(\phi|\alpha) \left(\prod_{n=1}^{N_d} P(z_n|\phi) P(w_{\text{ASR},n}|z_n, \theta_{\text{ASR}}) \right) \left(\prod_{p=1}^P P(z_p|\phi) P(w_{\text{text},p}|z_p, \theta_{\text{text}}) \right) \quad (6.1)$$

The inference and parameter estimation step is similar to Section 5.2.4 and we will not repeat them here.

6.4.1.2 Joint Modeling of Visual and Text (cLDA-VT)

For the joint modeling of visual and text features, we have already described in details in Section 5.2 of Chapter 5. For continuity, we provide a brief summary here.

As our visual features, we represent video as a *bag of keyframes*. In this approach, we treat the shot keyframe as the basic unit of visual information. After the keyframe extraction step, we take the entire collection of keyframes and perform clustering. Each keyframe is then assigned with unique IDs of its nearest cluster. These unique IDs are treated as the visual words. Our overall motivation is to capture a higher level of visual information. We can now generalize the TD-IDF weighting in text domain to these visual words. The reader is urged to refer back to Figure 5.2 of Chapter 5 to see some examples of videos represented by the term frequencies (tf) and document frequencies (df) of visual words.

As our clustering step, we adopt the Near Duplicate Image (NDI) approach of Chum *et al* [46]. In this approach, near duplicates images are computed based on a spatial pyramid of color histogram that jointly encodes the global and local information of the keyframe content. Given a query image, the NDIs are defined to be those within a specified Euclidean distance from the query. To compute a NDI cluster, we maintain a NDI list that initially only contains the query image. This list is then repeatedly populated with NDI of the new members in the list. The final NDI cluster is then given by the transitive closure of the NDI list. Again, the reader is urged to refer back to Figure 5.3 for some examples.

Given a corpus $\mathbf{D}$ of news video, each comprising of ASR words $\mathbf{W}_{\text{ASR}}$, text words $\mathbf{W}_{\text{text}}$ and visual words $\mathbf{V}$, we can write the generative process of cLDA-VT as follows:

- Draw a multinomial ϕ over K topics: $\phi \sim \text{Dir}(\alpha)$
- For each topic $k = 1 \dots K$,
 - draw multinomial $\theta_{ASR,k} \sim \text{Dir}(\beta)$ for text words
 - draw multinomial $\theta_{text,k} \sim \text{Dir}(\beta)$ for text words
 - draw multinomial $\gamma_k \sim \text{Dir}(\eta)$ for visual words
- For each ASR word index n in d , $n = 1$ to N_d
 - sample a topic z_n from ϕ : $z_n \sim \text{Multinomial}(\phi)$
 - sample a ASR word $w_{ASR,n}$ from θ_{ASR,z_n}
- For each text word index $p = 1$ to P
 - sample a topic z_p from ϕ : $z_p \sim \text{Multinomial}(\phi)$
 - sample a text word $w_{text,p}$ from θ_{text,z_p}
- For each visual word index m in d , $m = 1$ to M_d
 - draw discrete index variable $y_m \sim \text{Uniform}(1..N_d)$
 - sample a visual word v_m from $\text{Multinomial}(\gamma_{z_{y_m}})$

The above cLDA-VT model results in the following joint distribution on text $\mathbf{W}_{\text{ASR}}$, $\mathbf{W}_{\text{text}}$ and visual $\mathbf{V}$ and the latent topics:

$$\begin{aligned}
 P(\mathbf{W}_{\text{ASR}}, \mathbf{W}_{\text{text}}, \mathbf{V}, \phi, \mathbf{Z}_{\text{ASR}}, \mathbf{Z}_{\text{text}}, \mathbf{y}) &= P(\phi|\alpha) \left(\prod_{n=1}^{N_d} P(z_n|\phi) P(w_{ASR,n}|z_n, \theta_{ASR}) \right) \\
 &\quad \left(\prod_{p=1}^P P(z_p|\phi) P(w_{text,p}|z_p, \theta_{text}) \right) \\
 &\quad \left(\prod_{m=1}^{M_d} P(y_m|\phi) P(v_m|y_m, \gamma) \right) \tag{6.2}
 \end{aligned}$$

The main difference of our model from the Corr-LDA in Blei and Jordan [25] is that we use two multinomial distributions for the text and visual words. The sampling of visual words is essentially

Figure 6.4: Variational Simplification for learning cLDA-VT.

the same as the sampling text words. However, within a video, ϕ is a higher-level factor that is held fixed and it governs the ensemble of all text word and visual word observations. The topic-word multinomial θ and γ now learn the combined co-occurrence of important text words across video documents and also the complementary visual words.

Inference and Parameter Estimation

The cLDA-VT incorporates information from three modalities. At first glance, the learning of this joint model is more complicated than cLDA-T. However, the inference and parameter estimation can be generalized from that of cLDA-T. We present the key steps here, and leave the details to Appendix B.

We first define a simplification to the graphical representation of cLDA-VT by introducing four variational parameters χ, δ, Ω and λ (see Figure 6.4). We then write the factorized variational distribution on the latent variables as:

$$q(\phi, \mathbf{z}, \mathbf{y} | \chi, \delta, \lambda) = q(\phi | \chi) \prod_{n=1}^{N_d} q(z_n | \delta_n) \prod_{p=1}^P q(z_p | \Omega_p) \prod_{m=1}^M q(y_m | \lambda_m) \quad (6.3)$$

where χ is the variational Dirichlet parameter, δ_n and Ω_p are the variational multinomial parameters over K topics, and λ_m is the variational multinomial parameter over M visual words.

Next, we minimize the KL-divergence between this factorized distribution and the true posterior. Taking derivatives with respect to the variational parameters, we obtain the following coordinate ascent algorithm:

1. Update the posterior Dirichlet parameters:

$$\chi_i = \alpha_i + \sum_{n=1}^{N_d} \delta_{ni} + \sum_{p=1}^P \Omega_{pi} \quad (6.4)$$

2. For each ASR word and text word, update the posterior distribution over the latent factors:

$$\delta_{ni} \propto \theta_{i,w_n} \exp \left(\Psi(\chi_i) - \Psi\left(\sum_{j=1}^K \chi_j\right) + \sum_{m=1}^M \lambda_{mn} \log \gamma_{iv_m} + \sum_{p=1}^P \Omega_{pn} \log \theta_{iw_p} \right) \quad (6.5)$$

$$\Omega_{pi} \propto \theta_{i,w_p} \exp \left(\Psi(\chi_i) - \Psi\left(\sum_{j=1}^K \chi_j\right) + \sum_{n=1}^{N_d} \delta_{np} \log \theta_{iw_n} \right) \quad (6.6)$$

3. For each visual word, update the posterior distribution over the anchoring text words:

$$\lambda_{mn} \propto \exp \left(\sum_{i=1}^K (\delta_{ni}) \log \gamma_{iv_m} \right) \quad (6.7)$$

These update equations are invoked repeatedly until the change in KL divergence is small.

Parameter Estimation: After we have maximized the lower bound with respect to the variational parameters $\chi, \delta, \lambda, \Omega$, we next maximize the bound with respect to the model parameters α, θ, γ . The final maximization-step for the conditional multinomial θ is:

$$\theta_{ASR,ij} \propto \sum_{d=1}^D \sum_{n=1}^{N_d} \delta_{dni} w_{dn}^j \quad (6.8)$$

$$\theta_{text,ij} \propto \sum_{d=1}^D \sum_{p=1}^P \Omega_{dpi} w_{dp}^j \quad (6.9)$$

$w_{dn}^j = 1$ if the n th ASR word in document d is the j th word in the text vocabulary, else $w_{dn}^j = 0$.

$w_{dp}^j = 1$ if the p th text word in document d is the j th word in the text vocabulary, else $w_{dp}^j = 0$.

The maximization-step for the conditional multinomial γ is:

$$\gamma_{ij} \propto \sum_{d=1}^D \sum_{m=1}^{M_d} v_{dm}^j \sum_{n=1}^{N_d} \delta_{dni} \lambda_{dmn}, \quad (6.10)$$

$v_{dm}^j = 1$ if the m th visual word in document d is the j th word in the visual word vocabulary, else $v_{dm}^j = 0$. Finally, for the Dirichlet prior α , we follow the approach in [27] and use a Newton-Raphson algorithm to maximize α .

6.4.2 Selecting Topics and Terms for Query Suggestions

In this section, we describe the topic and term selection step. Even though we have described two joint models in the previous sections, we confine our discussion here on the cLDA-VT joint model without loss of generality. After the cLDA-VT learning has converged, the K latent topic distributions of both the text words and visual words are given by θ_k and γ_k , $k=1..K$. In particular, the probability of a text-word w in the k th latent topic is given by $P(w|\theta_k)$. We now make the assumption that the text multinomial θ_k represents the k th topic f_k in $\mathbf{D}$.

Topic Selection: From the K topics $\{f_1, f_2, \dots, f_K\}$, we select a few as suggestion clusters. We can do this by a ranking approach, where the k th topic f_k is ranked by its relevance to the query as follows. Given a query Q , note the following:

$$p(f_k|Q) \propto p(Q|f_k)p(f_k) \quad (6.11)$$

Hence, by assuming $p(f_k)$ are uniformly distributed and query word independence, we can write:

$$p(f_k|Q) \propto p(Q|f_k) = \prod_{q \in Q} p(q|f_k). \quad (6.12)$$

Based on the value of $p(f_k|Q)$, we can select the top S topics as suggestion clusters.

Terms Selection: Next, from each of the top S topics, we wish to select a few query suggestions as representative queries of the topic. For example, for the “fallujah battle” topic in Figure 6.1, we have selected three suggestions: “battle”, “terror” and “al-qaeda”. While each of these suggestions is anchored on the “battle” topic of the query “fallujah”, they also bear distinctive aspects within the topic. Our proposed way to achieve this is to select from amongst the high probability words candidate terms that are “compatible” with the current topic. By compatibility, we mean the

following: suppose a candidate term t is also associated with a multinomial distribution of words $p(t|C)$, where C is a parallel context corpus. Then we can compare this multinomial distribution with the topic multinomial word distribution f_k using a suitable distribution measure such as the KL divergence:

$$\begin{aligned}
\mathbb{KL}(t, f_k) &= - \sum_w p(w|f_k) \log \frac{p(w|f_k)}{p(w|t, C)} \\
&= \sum_w p(w|f_k) \log \frac{p(w, t|C)}{p(t|C)p(w|f_k)} \\
&\propto \sum_w p(w|f_k) \text{PMI}(w, t|C)
\end{aligned} \tag{6.13}$$

where PMI is the *pointwise mutual information* between the candidate term t and the terms in the topic model over the context corpus C . The PMI of two words is usually used as a measure of the semantic relationship between them. Intuitively, equation 6.13 assigns greater weights to a candidate term if it has a stronger semantic relationship to the important topic words. Hence, selected candidate terms are better representative of the entire topic f_k .

6.5 Experiments

To evaluate our methods, we used a subset of the TRECVID-2005 dataset [3]. It comprises of about 127 hours of Chinese and English news broadcast from five different sources (e.g. CCTV4, CNN). The dataset includes computed story boundaries from CMU Infromedia. For queries, we used the annotated set of 33 queries from Wu *et al* [172]. Sample queries include “Bush visits Canada”, “Mideast Peace”, “Arafat health”. For the full list of queries, refer to [172]. For the subsequent experiments described in this section, we had set the number of topics K to be 50.

6.5.1 Perplexity Comparison

A commonly used quantitative evaluation of a topic model is how well it can predict the words in a test-set W_{test} after learning from a train-set W_{train} : that is, we are interested in which model provide a better predictive distribution $p(w \in W_{test} | w \in W_{train})$. We adopt the popular *perplexity* measure in [25]. Denoting Ψ to be the model parameters of a learnt LDA or cLDA-VT model, the perplexity

Figure 6.5: Predictive Perplexity plot of LDA, cLDA-T and cLDA-VT on a held-out set from the TRECVID-2005 video dataset. Lower is better.

measure can be written as:

$$\text{Perplexity}(\Psi) = \left(\prod_{d=1}^D \prod_{w \in W_{test}} p(w|\Psi, w \in W_{train}) \right)^{\frac{-1}{\sum_{d=1}^D (|doc_d| - |W_{train}|)}}, \quad (6.14)$$

Mathematically, the perplexity of a word distribution is defined as the inverse of the per-word geometric average of the probability of the observations. Informally, it can be thought of as the effective number of equally likely words according to the model. Note that a lower number denotes more predictive power. Figure 6.5 compares the predictive perplexity obtained by the regular LDA and cLDA-VT. In the former, only the ASR words are fed to the inference process. As can be seen from the graph, there is less uncertainty in the joint models cLDA-T and cLDA-VT than in the standard LDA. This indicates that the topics learnt as a result of joint modeling are more robust and descriptive. Comparing between the two joint models, we observe that cLDA-T is slightly inferior to cLDA-VT. Though this difference is not significant, it does demonstrate the utility of visual+text joint modeling.

Table 6.1: Sample topic-based query suggestions. The first line is the original query.

Fallujah	Middle-east	George Bush
fallujah battle	Middle-east Yasser Arafat	George Bush visits
fallujah terror	Middle-east Palestinian	George Bush canada
fallujah al-qaeda	Middle-east Israel	George Bush falungong
fallujah uranium	Middle-east peace	George Bush protests
fallujah birth-defects	Middle-east minister	George Bush pentagon
fallujah massacre	Middle-east election	George Bush rice
fallujah blackwater	Middle-east leader	George Bush cheney

Table 6.2: Top-10 probability words of two learned cLDA-VT θ_k topics, for “War on Fallujah” (top) and “Mideast peace” (bottom). Meaningful words introspectively highlighted in bold.

LDA (ASR-text-only)	cLDA-VT (Visual+text)
iraq people time good meet united states bagh- dad chinese	iraq iraqi people military govern arm kill force attack baghdad
arafat know yasser leader thing peace minister people just israel	palestinian arafat peace israel president leader election yasser east middle

6.5.2 Qualitative Inspection

Given a query topic, sample results of our topic-based query suggestions are shown in Table 6.1. In Table 6.2, we qualitatively show the learned topics have benefited by the inclusion of visual features during topic learning. On the two queries “War on Fallujah” and “Mideast peace”, we introspectively pick a learned cLDA-VT topic that contains high probability words that are meaningful to the queries. Words from the joint cLDA-VT model are more intuitive and correlate better to the query topic.

6.5.3 An Objective Measure of Topic Coherence

While the previous subsection has *qualitatively* examined the output topics from an *intrinsic* point of view, in this subsection, we explore a more extrinsic and *quantitative* measure of the quality of a topic. First, we make the assumption that good topics are those that are semantically coherent to a human reader. Then, given a set of learned topics, we formulate an objective measure to predict the *coherence* of each topic. Next, we ask the human readers to decide if the learned topics are coherent. This is done by assessing each topic for its interpretability and association with a dominant semantic concept, and giving it a coherence score. Given a good topic coherence measuring function, we

Table 6.3: A sample of topics with their high topic coherence scores.

Topics	TCS value
iran nuclear energi korea agreement north atom council secur uranium	0.765
aid patient china treatment peopl prevent infect hiv hospit control	0.713
elect vote ukraine presidenti state result peopl opposit govern parliament	0.634
palestinian arafat yasser leader minist cairo ramallah funer hospit death	0.577
presid dai nbc state elect morn campaign right vice american	0.304
herman appropri girl wish husband diagnosi milk elizabeth conced edward	0.264

would expect that there is a high rank correlation between the predicted topic coherence score and the manually assigned coherence score.

6.5.3.1 Topic Coherence Score

Following the success of our formulation of the PMI-based semantic distance between queries in Equation 4.7 and Equation 4.6, we adopt a similar approach to base our prediction function of topic coherence on the Wikipedia external resource. Recall that a topic is a probability distribution of words. Given a learned topic, we take the top 15 highest probability words $w_1, w_2, \dots, w_{15}$ as the representative keywords of the topic. From these 15 words, we take all possible combinations $\binom{15}{2} = 105$ of word-pairs and assign each word-pair (w_i, w_j) a score based on PMI as follows:

$$\text{PMI}(w_i, w_j | \text{Wiki}) = \log\left(\frac{P(w_i, w_j | \text{Wiki})}{P(w_i | \text{Wiki}) \times P(w_j | \text{Wiki})}\right) \quad (6.15)$$

where the probability of term(s) is approximated by maximum likelihood: $P(w_i, w_j | \text{Wiki}) = \frac{\#(w_i, w_j)}{\#(\text{Wiki})}$ is the fraction of documents in the corpus where both w_i and w_j are found, and $P(w_i | \text{Wiki}) = \frac{\#(w_i)}{\#(\text{Wiki})}$ is the fraction of documents where the term w_i is found, and Wiki is the Wikipedia corpus ². Given a topic k , we can now define the Topic Coherence Score (TCS) of k as follows:

$$\text{TCS}(k) = \frac{1}{105} \sum_{i, j \in 1..15, i < j} \text{PMI}(w_i, w_j | \text{Wiki}) \quad (6.16)$$

Table 6.3 shows some example topics with their TCS scores.

²we use the English Wikipedia 2010-10-11 dump

Table 6.4: Spearman rank correlation values for the learned topics from standard LDA, cLDA-T and cLDA-VT. Note that for the standard LDA model, only the ASR words are used.

LDA	cLDA-T	cLDA-VT
0.21	0.47	0.53

6.5.3.2 Evaluation Results of Topic Coherence

For manual assessment of topic coherence, three human assessors were asked to score each of the $K = 50$ produced topics. Scoring was done on a 5-point scale where 1 denotes least coherent and 5 denotes very coherent. The assessors were provided with guidelines and examples on how to judge whether a topic was coherent or not. These include factors such as the ease for which the assessors can discern the dominant theme underlying the words in the topic, and whether the words in the topics are useful search keywords for a particular subject.

At the end of the topic coherence assessment, we have the manually assigned topic coherence scores (from 1 to 5) for all 50 topics. At the same time, we also have the automatically predicted topic coherence scores (TCS) of these 50 topics. We can now evaluate whether our TCS prediction can faithfully reproduce the human assessment score. This is done by using the Spearman Rank Correlation ³. The results are shown in Table 6.4. As expected, the TCS values from standard LDA are the worst. The joint text modeling cLDA-T improves on the LDA results significantly. By using visual features, the joint visual+text cLDA-VT model further achieves incremental improvement.

6.5.4 User Study

We are also interested to see if users would find it beneficial to be provided with our query suggestions. To do this, we conducted a small-scale user study. We asked three subjects to look at two versions of suggestions displayed for all 33 queries. The first version uses a simple lexicographical match and spelling corrector to find the top ten similar queries as query suggestions. They form the non-clustered suggestions as are currently obtained on most search engines. The second version uses the results from our automatic topic-based query suggestions. The subjects were then asked to write on a scale of 1 to 5 their preference liking to both methods. We consider a method to be significantly

³via the `spear(.)` function in Numerical Recipes in C

Table 6.5: User preferences for query suggestions. See text for details.

Category	Percentage of time
Topic-based clustered suggestion significantly better	30
Topic-based clustered suggestion marginally better	22
No preference	28
Unclustered suggestions marginally better	18
Unclustered suggestions significantly better	2

better or worse liked for a user if the difference in his rating on the two methods is equal or more than two, and marginally better or worse liked if the rating difference is one. The results are shown in Table 6.5. They indicate that our topic-based query suggestion method has potential value.

6.6 Discussion

Because the average length of a typical query is only two or three words long, most search engines are faced with a difficult task of discerning a user’s search intent. Query auto-completion is an important mechanism to ameliorate the problem, and to facilitate a user in finding his information need. However, most existing query suggestion methods are based on simple lexicographical matching, and suggestions are placed in an ad-hoc manner. In this chapter, we propose an alternative form of query suggestions that presents to users a list of possible complete queries grouped by their thematic topics. We argue for the benefits of such an arrangement, and demonstrate empirical results that show a high level of user acceptance and satisfaction with the idea.

As opposed to current query suggestion techniques, our method is document-centric and do not require a query log. Such a platform is particularly helpful when the user is not aware of how to phrase his query, or when the query words he chooses are not found in the documents. By mining topics on the documents directly, the resulting query suggestions are guaranteed to be found in the documents, making the suggestive framework useful not only for predicting what the user is likely to type, but also for uncovering new queries that may be of interest to the user.

Several aspects of the current implementation will be considered for future works. One relates to the run-time efficiency. Another pertains to the use of more sophisticated methods for query suggestion term selection.

Chapter 7

CONCLUSIONS

7.1 Summary of Thesis

This thesis has focused on the issue of search disambiguation in multimedia collections. A key motivation of this work is based on observation that while there is an exponentially growing amount of multimedia content available on the internet, this pace of growth is not commensurate with the level of satisfaction enjoyed by the typical search user. The main reason for this discrepancy is related to the nature of the average search users.

We conjecture that there has been a significant change in the average search behaviour which necessitates the development of a new model of information retrieval and its evaluation. In the early days of information retrieval, most document collections were well-authored and well-structured that were generally well understood by the users. These users in turn were typically computer professionals that are knowledgeable in phrasing queries to maximize the retrieval of relevant documents. In contrast, most document collections today are heterogeneous in nature and a typical search user is oblivious of the need to properly formulate a query intent. As a result, the average number of keywords in a search query is short and ambiguous, wherein there is a general expectation that it is the job of the search engine to disambiguate the vague query.

In this thesis, we define search disambiguation as the general class of techniques to elucidate the multi-faceted information needs arising from an ambiguous query. A key tenet of this work is the faceted view of information. When applied to the ambiguous query, this view purports the notion that there are various facets to the query, corresponding to the different possible interpretation of

true information need. When applied to the data corpus, the faceted view looks at documents in a collection as being composed of various thematic structures. We develop a computational model of the multi-faceted view of content based on the probabilistic topic model. To support our multi-faceted view of information, we also develop a faceted topic retrieval model that goes beyond the traditional term-based matching, to facet-based matching of documents to query. By describing content in terms of facets, our retrieval framework makes it easy for a search engine to group the search results by multiple facets. Because facets represent a higher level of information than the mere words and terms, our faceted model naturally alleviates the problem of result duplicates, and provides a way for a user to zoom into any specific facet group and to progressively clarify and narrow his search intent.

Given an ambiguous query, our faceted retrieval model aims to present the users with clusters of documents embodying the multiple dominant interpretations of the original query. This approach exemplifies the famous cluster hypothesis, first popularized by Rijsbergen in the text domain. In this thesis, we have extended the cluster hypothesis in two following ways.

First, we extend the cluster hypothesis to the multimedia domain. For example, in Chapter 4, we cluster a collection of polysemous images into groups that can be seen as depicting the various visual senses of the query. In Chapter 5 and Chapter 6, we look at how we can capitalize on the availability of multimodal content cues in news video, to learn facets that are semantically more coherent, and better correlate with actual news events. We extend the basic probabilistic topic model to jointly model the co-occurrence of the complementary data streams as additional observations.

Second, we extend the cluster hypothesis to handle the unbounded nature of query ambiguity. To cluster a collection of documents, most clustering methods would require the a priori specification of the number of clusters. This presents an important problem in the traditional application of the cluster hypothesis to clustering results from an ambiguous query. If the goal of the cluster hypothesis is to produce groups that can meaningfully represent the various topics in the return results, then the clustering algorithm must know the number of these topics in the first place. This becomes tricky when the query is ambiguous. This is because different queries have varying degrees of ambiguity, and this will result in varying levels of heterogeneity or polysemy in the return results. To address this issue, we develop a non-parametric Bayesian method to cluster search results. The main advantage of our clustering method is that the number of mixture components is not fixed a priori, but is

determined during the posterior inference process. This allows our model to grow with the level of polysemy (and visual diversity) in the return multimedia results.

A major consideration in the design of the search disambiguation methods in this thesis is that they must not assume the availability of any contextual information that can potentially shed light on the user search intent. We work with the two most basic sources of information available: the query statement, and the data collection. This is a key differentiation of our work from other disambiguation effort that requires the availability of special corpus such as, for example, a query log, from which the search intent can be better inferred. In this aspect, an intrinsic appeal of our methods is that it has a low working requirement. Of course, the presence of external knowledge resources will help improve results in our methods. But we do not exert any other infrastructural demand on them, other than their general availability over the internet.

7.2 Contributions

The work in this thesis consists of three major parts: proposing a faceted topic retrieval framework for multimedia collection, building a joint model of the complementary information available in the multimedia content, and developing a non-parametric Bayesian clustering method to address the unbounded nature of query ambiguity. The contributions are as follows:

First, we propose a novel faceted topic retrieval framework wherein multimedia documents are modeled as comprising of facets or topics. We define a facet as a new information unit of content that represents a fact or a topic that is contained in the query need. With our faceted topic retrieval framework, documents are now ranked by the relevance of their composite facets/topics to the query. This differs from the traditional term-based retrieval models (e.g. TF-IDF). With documents now described in terms of facets, we facilitate the retrieval of documents that are not only relevant to the query, but are also representative of the many different facets of the information need. We base our faceted retrieval framework on probabilistic topic models, a class of algorithms designed to discover the latent thematic structures in a document collection.

Second, we augment our faceted retrieval framework with an ability to process ambiguous queries in multimedia collection. The need for this arises from the fact that different queries have varying level of ambiguity. Hence, there is varying level of heterogeneity and polysemy in the return results.

This presents a problem to a traditional clustering algorithm which requires the specification of the number of clusters. To address this issue, we develop a non-parametric Bayesian method to cluster search results. The main advantage of our clustering method is that the number of mixture components is not fixed a priori, but is determined during the posterior inference process. This allows our model to grow with the level of polysemy (and visual diversity) in the return multimedia results.

Thirdly, we extend the basic probabilistic topic model (the Latent Dirichlet Allocation, LDA) to jointly model the complementary information in the visual and textual streams. We show that our joint modeling approach can produce topics that are more semantically coherent and better correlate with actual reality. This in turn improves the quality of facet detection in multimedia content, and ultimately yields better user satisfaction in faceted topic retrieval.

7.3 Future Works

While the works done in this thesis has made contributions to the field of multimedia modeling and retrieval, there are several limitations. First, we make the implicit assumption that facets occur in documents independently. In Chapter 3, this assumption eases our formulation of the faceted retrieval model, where we can adopt a greedy approach to simultaneously maximize the novelty and diversity of facets present in the top ranked documents. It is clear that facets may be correlated in real documents. For example, the facet “Death of Arafat” is expected to correlate highly with the facet “rockets fired at Israel”. There are two main directions we can take to relax this assumption:

1. The first is to replace the Dirichlet with a logistic normal, as per what has been done in the Correlated Topic Model by Blei and Lafferty *et al* [26]. When a document is modeled as a draw from a Dirichlet distribution, the topic proportions are modeled as nearly independent. Hence, this hampers the ability to model the *topic-level* correlations in real-world corpus. By using a logistic normal, the co-occurrence between topics can be captured in the full covariance structure of the normal distribution. However, the choice of prior becomes a significant challenge because the inference in logistic normal based topic models are now based on conjugate gradient update, wherein parameter estimation for a full-rank covariance matrix can be very inaccurate in high dimensional space.

2. The second direction is to consider the use of linear latent factor model inspired by work in the field of Factor Analysis [14] and Probabilistic Principal Component Analysis [153]. These methods aim to uncover hidden factors that explain correlations in the observed data. Adopting this approach, we can assume the presence of a two-tier hierarchical facet structure in the corpus, and that the leaf-node facets co-occur because of some independent parent node sources. The correlation structure between the leaf-node facets can then be modeled by linearly mixing these parent sources to form correlated topic vectors.

Second, the inference and learning of our probabilistic faceted models is a time consuming process. It is clear that for practical deployment, we need to improve the run-time efficiency of the inference process. In this regard, we can look to several recent advances in distributed topic modeling such as [9, 117]. The main idea in these distributed algorithms is to partition the data onto separate CPUs and to perform the inference in parallel. The challenge lies in how to map the data onto the individual processors while preserving sampling integrity, and how to merge newly-created topics from the processors. Another related issue is the handling of new documents in a practical multimedia document retrieval system. For example, in the retrieval experiments done in Chapter 5, we have conveniently made use of privileged information about our TRECVID-2005 dataset such as the average number of relevant documents and number of facets per query topic. In a real-world setting, we have daily ingest of new documents and discarding of unwanted documents. Hence, the document collection is much more dynamic. It is not immediately clear how these parameters can be determined in an efficient way.

Third, in the joint modeling of visual and text information in Chapter 6, we have mostly made use of conventional features (visual and textual). Further work can be explored along the direction of how best to exploit the use of other visual features for better joint modeling. For example, there have been advances in the development of new visual features such as the Space-Time Interest Points [95] and global geometric correspondence [100], and learning methods (e.g. classesemes [155], group sparse coding [21, 65]) that can yield better results in image categorization. These methods are able to handle the problem areas wherein images of the same category exhibit wide diversity of intra-class visual features, and images of different categories have similar inter-class visual features.

Appendix A

List of Publications

A.1 Journal Papers

- (J1) Kong-Wah WAN, Ah-Hwee TAN, Joo-Hwee LIM and Liang-Tien Clement CHIA, “A Non-parametric Visual-Sense Model of Images - Extending the Cluster Hypothesis Beyond Text”, In Multimedia Tools and Applications, accepted for publication
- (J2) Kong-Wah WAN, Ah-Hwee TAN, Joo-Hwee LIM and Liang-Tien CHIA, “Topic Based Query Suggestions for Video Search”, In K. Schoeffmann, B. Merialdo, A. Hauptmann, C. Ngo, Y. Andreopoulos, and C. Breiteneder, editors, Advances in Multimedia Modeling, volume 7131 of Lecture Notes in Computer Science, pp 288-299, Springer Berlin / Heidelberg, 2012.

A.2 Conference Papers

- (C1) Kong-Wah WAN, Ah-Hwee TAN, Joo-Hwee LIM and Liang-Tien CHIA, “Topic-based Query Suggestions for Video Search”, In Proc International Multimedia Modeling Conference, 2011 (Best Paper Award)
- (C2) Kong-Wah WAN, Ah-Hwee TAN, Joo-Hwee LIM and Liang-Tien CHIA, “Faceted topic retrieval of news video using joint topic modeling of visual features and speech transcripts”, In Proc of International Conference on Multimedia & Expo, pp 843-848, 2010
- (C3) Kong-Wah WAN, Ah-Hwee TAN, Joo-Hwee LIM, Liang-Tien CHIA and Sujoy ROY, “A Latent

Model for Visual Disambiguation of Keyword-based Image Search”, In Proc British Machine Vision Conference, 2009

- (C4) Sujoy ROY, Mun-Thye MAK and Kong-Wah WAN, “Wikipedia Based News Video Topic Modeling for Information Extraction”, In Proc of Multimedia Modeling, pp 411-420, 2011
- (C5) Kong-Wah WAN and Sujoy ROY, “Identifying and learning visual attributes for object recognition”, In Proc International Conference on Image Processing, pp 3893-3896, 2010
- (C6) Kong-Wah WAN, “Exploiting Story-Level Context To Improve Video Search”, In Proc International Conference on Multimedia and Expo, 2008

Appendix B

VARIATIONAL INFERENCE OF JOINT MULTINOMIAL CORRESPONDENCE MODEL

B.1 The Joint Multinomial Correspondence Model

From figure 3.1, we can write the joint distribution for a multimedia document in terms of the visual words v , textual (speech transcript) words w and latent variable z as:

$$p(\mathbf{z}, \mathbf{y}, \phi, \mathbf{v}, \mathbf{w} | \alpha, \theta, \gamma) = p(\phi | \alpha) \left(\prod_{n=1}^{N_d} p(z_n | \phi) p(w_n | z_n, \theta) \right) \left(\prod_{m=1}^M p(y_m | N_d) p(v_m | y_m, \mathbf{z}, \gamma) \right) \quad (\text{B.1})$$

B.2 Principles of Variational Methods

Variational methods are a class of deterministic algorithms that offer an alternative to performing Bayesian inference in graphical models [89]. They differ from the sampling-based methods by simplifying the inference process using approximation. Rather than approximating the posterior with samples, variational methods assume a parameterized family of distributions over the same latent structures, and then find the member of that family that is closest to the posterior. Thus, the inference problem is transformed to an optimization problem. This allows variational methods to tap on the advances of optimization techniques, enabling increased practical applications to be developed

Figure B.1: A concave $f(x)$ and the linear bound $\lambda^T x - f^*(\lambda)$.

in probabilistic modeling.

To transform the graphical model into a simplified one, we turn to the general principle of convex duality, which translates a concave function $f(x)$ into an optimization problem. Using Taylor's linearization, we have: $f(x) \leq f(\xi) + f'(\xi)(x - \xi) = f'(\xi)x - (f'(\xi)\xi - f(\xi))$. If we define $\lambda = f'(\xi)$, we have: $f(x) \leq \lambda x - (\lambda(f')^{-1}(\lambda) - f \circ (f')^{-1}(\lambda))$. If we define $f^*(\lambda)$ to be the second term on the right hand side, then we have:

$$f(x) \leq \lambda x - f^*(\lambda). \quad (\text{B.2})$$

Since we know that $f(x)$ is concave and bounded above, the bound is *exact* somewhere:

$$f(x) = \min_{\lambda} \{ \lambda^T x - f^*(\lambda) \} \quad (\text{B.3})$$

Further, we know by construction that if we shift λx down by $f^*(\lambda)$, we will obtain that bound. So, considering λ fixed, we can write:

$$f^*(\lambda) = \min_x \{ \lambda^T x - f(x) \} \quad (\text{B.4})$$

The idea behind equation B.3 and equation B.4 is that given x , we can find a linear function that has exactly the value $f(x)$ at x_0 , and greater than or equal to $f(x)$ everywhere else. This works as long as $f(x)$ is concave, and $\lambda^T x - f^*(\lambda)$ is an variational upper bound of $f(x)$. Figure B.1 illustrates this variational framework.

To see how the variational framework can be applied to the learning of a graphical model, consider

the set of N observed data: $\mathbf{d} = \{x^{(1)}, \dots, x^{(N)}\}$, where $x^{(n)} = (x_1^{(n)}, \dots, x_d^{(n)})$. The data maximum likelihood (ML) is given by: $p(\mathbf{d}|\theta) = \prod_{n=1}^N p(x^{(n)}|\theta)$. For computational convenience, we take the log likelihood:

$$\mathcal{L}(\theta) = \log p(\mathbf{d}|\theta) = \sum_{n=1}^N \log p(x^{(n)}|\theta) = \sum_{n=1}^N \sum_{i=1}^d \log p(x_i^{(n)}|\theta_i) \quad (\text{B.5})$$

Assuming the prior $p(\theta)$ can be factorized as: $p(\theta) = \prod_i p(\theta_i)$, we can write the MAP posterior estimate as: $p(\theta|\mathbf{d}) \approx p(\mathbf{d}|\theta)p(\theta)$. Again, by taking log, we have:

$$\mathcal{L}'(\theta) = \sum_{i=1}^d \mathcal{L}_i(\theta_i) + \log p(\theta_i) \quad (\text{B.6})$$

However, some of the variables are hidden. Assume the data $\{\mathbf{x}, \mathbf{y}\}$ is now composed of observed $\mathbf{x}$ and hidden $\mathbf{y}$. We cannot decompose the log likelihood as before. Rather, we now marginalize over the hidden variables:

$$\mathcal{L}(\theta) = \log p(\mathbf{x}|\theta) = \log \sum_{\mathbf{y}} p(\mathbf{x}, \mathbf{y}|\theta) \quad (\text{B.7})$$

Maximizing the above is intractable. There is where we apply the variational principles, and define a variational lower bound $q(\mathbf{y})$, and maximize it instead:

$$\begin{aligned} \mathcal{L}(\theta) &= \log \sum_{\mathbf{y}} p(\mathbf{x}, \mathbf{y}|\theta) = \log \sum_{\mathbf{y}} q(\mathbf{y}) \frac{p(\mathbf{x}, \mathbf{y}|\theta)}{q(\mathbf{y})} \\ &\geq \sum_{\mathbf{y}} q(\mathbf{y}) \log \frac{p(\mathbf{x}, \mathbf{y}|\theta)}{q(\mathbf{y})} = \sum_{\mathbf{y}} q(\mathbf{y}) \log p(\mathbf{x}, \mathbf{y}|\theta) - \sum_{\mathbf{y}} q(\mathbf{y}) \log q(\mathbf{y}) \end{aligned} \quad (\text{B.8})$$

The inequality in equation B.8 is due to $\log(\frac{\sum_i x_i}{N}) \geq \sum_i \frac{\log x_i}{N}$, a result from Jensen's inequality applied to a concave function. To maximize $\mathcal{L}(\theta)$, we use expectation maximization:

$$\text{E step: } q_{k+1} \leftarrow \operatorname{argmax}_q \sum_{\mathbf{y}} q(\mathbf{y}) \log p(\mathbf{x}, \mathbf{y}|\theta_k) - \sum_{\mathbf{y}} q(\mathbf{y}) \log q(\mathbf{y}) \quad (\text{B.9})$$

$$\text{M step: } \theta_{k+1} \leftarrow \operatorname{argmax}_{\theta} \sum_{\mathbf{y}} q_{k+1}(\mathbf{y}) \log p(\mathbf{x}, \mathbf{y}|\theta) \quad (\text{B.10})$$

The maximization of the above E-step is when the bound in equation B.8 becomes exact. That is,

q_{k+1} is chosen such that $\log p(\mathbf{x}|\theta) - \sum_{\mathbf{y}} q(\mathbf{y}) \log \frac{p(\mathbf{x}, \mathbf{y}|\theta)}{q(\mathbf{y})}$. By simple arithmetic re-arrangement,

$$\begin{aligned}
\log p(\mathbf{x}|\theta) - \sum_{\mathbf{y}} q(\mathbf{y}) \log \frac{p(\mathbf{x}, \mathbf{y}|\theta)}{q(\mathbf{y})} &= \log p(\mathbf{x}|\theta) - \sum_{\mathbf{y}} q(\mathbf{y}) \log \frac{p(\mathbf{y}|\mathbf{x}, \theta)p(\mathbf{x}|\theta)}{q(\mathbf{y})} \\
&= \log p(\mathbf{x}|\theta) - \sum_{\mathbf{y}} q(\mathbf{y}) \log \frac{p(\mathbf{y}|\mathbf{x}, \theta)}{q(\mathbf{y})} - \sum_{\mathbf{y}} q(\mathbf{y}) \log p(\mathbf{x}|\theta) \\
&= - \sum_{\mathbf{y}} q(\mathbf{y}) \log \frac{p(\mathbf{y}|\mathbf{x}, \theta)}{q(\mathbf{y})} \\
&= \mathbb{KL}(q(\mathbf{y}), p(\mathbf{y}|\mathbf{x}, \theta))
\end{aligned} \tag{B.11}$$

Hence, the maximum of the E-step in equation B.9 happens by setting $q_{k+1} = p(\mathbf{y}|\mathbf{x}, \theta)$. Putting this into equation B.10, the M-step becomes:

$$\text{M step: } \theta_{k+1} \leftarrow \underset{\theta}{\operatorname{argmax}} \sum_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}, \theta_k) \log p(\mathbf{x}, \mathbf{y}|\theta) \tag{B.12}$$

The form of $p(\mathbf{y}|\mathbf{x}, \theta_k)$ can be complicated in general, and hence intractable. The heart of applying variational method in graphical models lies in defining the variational factorization distribution as the approximation distribution $q(x)$. Interestingly, Neal commented in [16] on page 58-60 that “*there is something of a dark art in discovering a factorization amongst the hidden variables, such that the approximation remains faithful at an “acceptable” level*”. In the case of LDA, Blei *et al* [27] defined this by first removing all connecting edges between the interior nodes: $\phi \rightarrow \mathbf{Z}$, $\theta \rightarrow \mathbf{W}$ and $\mathbf{Z} \rightarrow \mathbf{W}$, and defining the variational distribution (Figure B.2):

$$q(\phi, \mathbf{Z}|\gamma, \eta) = q(\phi|\gamma) \prod_{n=1}^N q(z_n|\eta_n) \tag{B.13}$$

The rationale for this seemingly drastic step is based on naive mean field approximation [162]. This approximation method makes use of a subgraph that is completely disconnected. Thus, the approximating distribution is fully factorized. In LDA, the difference between the variational approximation $q(\phi, \mathbf{Z}|\gamma, \eta)$ and the true posterior $P(\phi, \mathbf{Z}|\mathbf{W}, \alpha, \beta)$ is then minimized using the KL divergence between the two:

$$(\gamma^*, \eta^*) = \underset{(\gamma, \eta)}{\operatorname{argmax}} \mathbb{KL}(q(\phi, \mathbf{Z}|\gamma, \eta) || P(\phi, \mathbf{Z}|\mathbf{W}, \alpha, \beta)) \tag{B.14}$$

To minimize the KL divergence, several numerical methods can be used, such as gradient ascent, or

Figure B.2: Left: Graphical model of the LDA; Right: Graphical model of the variational distribution used to approximate the posterior in LDA.

iterative fixed-point methods.

B.3 Variational Inference

The parameters we want to estimate are: α, θ and γ . Marginalizing over the latent variables ϕ, z and y , we re-write equation B.1 as:

$$p(\mathbf{v}, \mathbf{w} | \alpha, \theta, \gamma) = \int_{\phi} \sum_z \sum_y p(\mathbf{v}, \mathbf{w}, \phi, \mathbf{z}, \mathbf{y} | \alpha, \theta, \gamma) d\phi \quad (\text{B.15})$$

Unfortunately, equation B.15 is intractable due to the coupling between ϕ and γ in the summation over the latent variables z . We turn to the use of variational inference to approximate the posterior distribution over latent variables given the observed data (visual words v and text w). We follow the general recipe of variational inference and derive the inference algorithm.

B.3.1 Variational Simplification

To simplify the original graphical representation in equation B.15, we introduce free variational parameters χ, δ and λ and write the factorized variational distribution on the latent variables as:

$$q(\phi, \mathbf{z}, \mathbf{y} | \chi, \delta, \lambda) = q(\phi | \chi) \prod_{n=1}^{N_d} q(z_n | \delta_n) \prod_{m=1}^M q(y_m | \lambda_m) \quad (\text{B.16})$$

where χ is the variational Dirichlet parameter, δ_n is the variational multinomial parameter over K topics, and λ_m is the variational multinomial parameter over M visual words.

B.3.2 Variational Lower Bound

We now attempt to lower bound equation B.15 by the variational distribution in equation B.16. Plugging equation B.16 into equation B.15, and applying Jensen's inequality, which says that if X is a random variable and f is a concave function, then $f(\mathbb{E}[X]) \geq \mathbb{E}[f(X)]$, we have:

$$\begin{aligned}
\log p(\mathbf{v}, \mathbf{w} | \alpha, \gamma, \theta) &= \log \int_{\phi} \sum_{\mathbf{z}, \mathbf{y}} p(\mathbf{v}, \mathbf{w}, \phi, \mathbf{z}, \mathbf{y} | \alpha, \gamma, \theta) d\phi \\
&= \log \int_{\phi} \sum_{\mathbf{z}, \mathbf{y}} \frac{p(\mathbf{v}, \mathbf{w}, \phi, \mathbf{z}, \mathbf{y} | \alpha, \gamma, \theta) q(\phi, \mathbf{z}, \mathbf{y} | \chi, \delta, \lambda)}{q(\phi, \mathbf{z}, \mathbf{y} | \chi, \delta, \lambda)} d\phi \\
&= \log \mathbb{E}_q \left[\frac{p(\mathbf{v}, \mathbf{w}, \phi, \mathbf{z}, \mathbf{y} | \alpha, \gamma, \theta)}{q(\phi, \mathbf{z}, \mathbf{y} | \chi, \delta, \lambda)} \right] \\
&\geq \mathbb{E}_q [\log p(\mathbf{v}, \mathbf{w}, \phi, \mathbf{z}, \mathbf{y} | \alpha, \gamma, \theta)] - \mathbb{E}_q [\log q(\phi, \mathbf{z}, \mathbf{y} | \chi, \delta, \lambda)] \\
&= \mathcal{L}(\chi, \delta, \lambda; \alpha, \gamma, \theta)
\end{aligned} \tag{B.17}$$

where $\mathbb{E}_q$ is the expectation with respect to the variational distribution q : $\mathbb{E}_q[f(X)] = \int f(x)q(x)dx$, for any arbitrary function $f(X)$ of X . Equation B.17 provides a lower bound on the log likelihood of the joint probability of a document.

Next, we maximize this with respect to all six parameters given by the three variational parameters χ, δ, λ , and three model parameters α, γ, θ . To do this, we apply Naive mean field approximation principle, and factorize p and q as follow:

$$\begin{aligned}
p(\mathbf{v}, \mathbf{w}, \phi, \mathbf{z}, \mathbf{y} | \alpha, \gamma, \theta) &= p(\phi | \alpha) p(\mathbf{z} | \phi) p(\mathbf{w} | \mathbf{z}, \theta) p(\mathbf{y} | N_d) p(\mathbf{v} | \mathbf{y}, \mathbf{z}, \gamma) \\
q(\phi, \mathbf{z}, \mathbf{y} | \chi, \delta, \lambda) &= q(\phi | \chi) q(\mathbf{z} | \delta) q(\mathbf{y} | \lambda)
\end{aligned} \tag{B.18}$$

Using the fact that $\mathbb{E}[\log(XY)] = \mathbb{E}[\log X] + \mathbb{E}[\log Y]$, and we accordingly rewrite each term in equation B.17 to get:

$$\begin{aligned}
\mathcal{L}(\chi, \delta, \lambda; \alpha, \gamma, \theta) &= \mathbb{E}_q [\log p(\phi | \alpha)] + \mathbb{E}_q [\log p(\mathbf{z} | \phi)] + \mathbb{E}_q [\log p(\mathbf{w} | \mathbf{z}, \theta)] + \\
&\quad \mathbb{E}_q [\log p(\mathbf{y} | N_d)] + \mathbb{E}_q [\log p(\mathbf{v} | \mathbf{y}, \mathbf{z}, \gamma)] \\
&\quad - \mathbb{E}_q [\log q(\phi | \chi)] - \mathbb{E}_q [\log q(\mathbf{z} | \delta)] - \mathbb{E}_q [\log q(\mathbf{y} | \lambda)].
\end{aligned} \tag{B.19}$$

B.3.3 Factor Expansion

We now expand every term in equation B.19 in terms of the model parameters α, ϕ, θ or variational parameters χ, δ, λ . We factorize the first term $\mathbb{E}_q[\log p(\phi|\alpha)]$ as:

$$\begin{aligned}
\mathbb{E}_q[\log p(\phi|\alpha)] &= \mathbb{E}_q\left[\log \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \left(\prod_{i=1}^K \phi^{\alpha_i-1}\right)\right] \\
&= \mathbb{E}_q\left[\log \Gamma\left(\sum_{i=1}^K \alpha_i\right) - \sum_{i=1}^K \log \Gamma(\alpha_i) + \sum_{i=1}^K (\alpha_i - 1) \log \phi_i\right] \\
&= \mathbb{E}_q[\log \Gamma(\sum_{i=1}^K \alpha_i)] - \mathbb{E}_q[\sum_{i=1}^K \log \Gamma(\alpha_i)] + \mathbb{E}_q[\sum_{i=1}^K (\alpha_i - 1) \log \phi_i] \\
&= \log \Gamma(\sum_{i=1}^K \alpha_i) - \sum_{i=1}^K \log \Gamma(\alpha_i) + \sum_{i=1}^K (\alpha_i - 1) \mathbb{E}_q[\log \phi_i] \\
&= \log \Gamma(\sum_{i=1}^K \alpha_i) - \sum_{i=1}^K \log \Gamma(\alpha_i) + \sum_{i=1}^K (\alpha_i - 1) \left(\Psi(\chi_i) - \Psi\left(\sum_{j=1}^K \chi_j\right) \right) \quad (\text{B.20})
\end{aligned}$$

Note that in line three in the above, we have $\mathbb{E}_q[\log \Gamma(\sum_{i=1}^K \alpha_i)] = \log \Gamma(\sum_{i=1}^K \alpha_i)$ because α_i is constant with respect to $q(\cdot)$, and hence $\mathbb{E}_q[\text{constant}] = \text{constant}$. Further, the last line in the above results from the following important fact:

$$\mathbb{E}_q[\log \phi_i | \chi] = \Psi(\chi_i) - \Psi\left(\sum_{j=1}^K \chi_j\right). \quad (\text{B.21})$$

We factorize the second term $\mathbb{E}_q[\log p(\mathbf{z}|\phi)]$ in equation B.19 as:

$$\begin{aligned}
\mathbb{E}_q[\log p(\mathbf{z}|\phi)] &= \mathbb{E}_q\left[\log \prod_{n=1}^{N_d} \prod_{i=1}^K \phi_i^{z_n^i}\right] \\
&= \sum_{n=1}^{N_d} \sum_{i=1}^K \mathbb{E}_q[z_n^i \log \phi_i] \\
&= \sum_{n=1}^{N_d} \sum_{i=1}^K \mathbb{E}_q[\log \phi_i] \mathbb{E}_q[z_n^i] \\
&= \sum_{n=1}^{N_d} \sum_{i=1}^K \left(\Psi(\chi_i) - \Psi\left(\sum_{j=1}^K \chi_j\right) \right) \delta_{ni} \quad (\text{B.22})
\end{aligned}$$

The third line in the above uses the property of the mean of a Multinomial distribution: $\mathbb{E}_q[z_n^i] = \delta_{ni}$, since in the variational distribution $q(\cdot)$, z_n is a sample of the variational Multinomial δ .

We now factorize the third term $\mathbb{E}_q[\log p(\mathbf{w}|\mathbf{z}, \theta)]$ in equation B.19 as:

$$\begin{aligned}
\mathbb{E}_q[\log p(\mathbf{w}|\mathbf{z}, \theta)] &= \mathbb{E}_q\left[\log \prod_{n=1}^{N_d} \prod_{i=1}^K \prod_{j=1}^{V_t} \theta_{ij}^{(z_n^i \cdot w_n^j)}\right] \\
&= \sum_{n=1}^{N_d} \sum_{i=1}^K \sum_{j=1}^{V_t} \mathbb{E}_q[z_n^i] \cdot w_n^j \log \theta_{ij} \\
&= \sum_{n=1}^{N_d} \sum_{i=1}^K \sum_{j=1}^{V_t} \delta_{ni} \cdot w_n^j \log \theta_{ij}
\end{aligned} \tag{B.23}$$

We now factorize the fourth term $\mathbb{E}_q[\log p(\mathbf{y}|N_d)]$ in equation B.19 as:

$$\mathbb{E}_q[\log p(\mathbf{y}|N_d)] = \text{constant} \tag{B.24}$$

We now factorize the fifth term $\mathbb{E}_q[\log p(\mathbf{v}|\mathbf{y}, \mathbf{z}, \gamma)]$ in equation B.19 as:

$$\begin{aligned}
\mathbb{E}_q[\log p(\mathbf{v}|\mathbf{y}, \mathbf{z}, \gamma)] &= \mathbb{E}_q\left[\log \prod_{m=1}^M \prod_{i=1}^K \prod_{j=1}^{V_v} \prod_{n=1}^{N_d} \gamma_{ij}^{z_n^i \cdot y_m^n \cdot v_m^j}\right] \\
&= \sum_{m=1}^M \sum_{i=1}^K \sum_{j=1}^{V_v} \sum_{n=1}^{N_d} \mathbb{E}_q[z_n^i \cdot y_m^n \cdot v_m^j \log \gamma_{ij}] \\
&= \sum_{m=1}^M \sum_{i=1}^K \sum_{j=1}^{V_v} \sum_{n=1}^{N_d} \mathbb{E}_q[z_n^i] \cdot \mathbb{E}_q[y_m^n] \cdot \mathbb{E}_q[v_m^j \log \gamma_{ij}] \\
&= \sum_{m=1}^M \sum_{i=1}^K \sum_{j=1}^{V_v} \sum_{n=1}^{N_d} \delta_{ni} \cdot \lambda_{mn} \cdot v_m^j \log \gamma_{ij}
\end{aligned} \tag{B.25}$$

The last line in the above again uses the property of the mean of a Multinomial distribution: $\mathbb{E}_q[y_m^n] = \lambda_{mn}$, since in the variational distribution $q(\cdot)$, y_m is a sample of the variational Multinomial λ .

We now factorize the sixth term $\mathbb{E}_q[\log q(\phi|\chi)]$ in equation B.19 as:

$$\begin{aligned}
\mathbb{E}_q[\log q(\phi|\chi)] &= \mathbb{E}_q \left[\log \frac{\Gamma(\sum_{i=1}^K \chi_i)}{\prod_{i=1}^K \Gamma(\chi_i)} \prod_{i=1}^K \phi_i^{\chi_i-1} \right] \\
&= \mathbb{E}_q \left[\log \Gamma(\sum_{i=1}^K \chi_i) - \sum_{i=1}^K \log \Gamma(\chi_i) + \sum_{i=1}^K (\chi_i - 1) \log \phi_i \right] \\
&= \log \Gamma(\sum_{i=1}^K \chi_i) - \sum_{i=1}^K \log \Gamma(\chi_i) + \sum_{i=1}^K (\chi_i - 1) \left(\Psi(\chi_i) - \Psi(\sum_{j=1}^K \chi_j) \right) \quad (\text{B.26})
\end{aligned}$$

We now factorize the seventh term $\mathbb{E}_q[\log q(\mathbf{z}|\delta)]$ in equation B.19 as:

$$\begin{aligned}
\mathbb{E}_q[\log q(\mathbf{z}|\delta)] &= \mathbb{E}_q \left[\log \prod_{n=1}^{N_d} \prod_{i=1}^K \delta_{ni}^{z_{ni}^i} \right] \\
&= \sum_{n=1}^{N_d} \sum_{i=1}^K \mathbb{E}_q[z_n^i] \log \delta_{ni} \\
&= \sum_{n=1}^{N_d} \sum_{i=1}^K \delta_{ni} \log \delta_{ni} \quad (\text{B.27})
\end{aligned}$$

We now factorize the eighth term $\mathbb{E}_q[\log q(\mathbf{y}|\lambda)]$ in equation B.19 as:

$$\begin{aligned}
\mathbb{E}_q[\log q(\mathbf{y}|\lambda)] &= \mathbb{E}_q \left[\log \left(\prod_{m=1}^M \prod_{n=1}^{N_d} \lambda_{mn}^{y_{mn}^n} \right) \right] \\
&= \sum_{m=1}^M \sum_{n=1}^{N_d} \mathbb{E}_q[y_m^n] \log \lambda_{mn} \\
&= \sum_{m=1}^M \sum_{n=1}^{N_d} \lambda_{mn} \log \lambda_{mn} \quad (\text{B.28})
\end{aligned}$$

Now we plug all eight expanded forms back into equation B.19 to obtain:

$$\begin{aligned}
\mathcal{L}(\chi, \delta, \lambda; \alpha, \gamma, \theta) &= \log \Gamma(\sum_{i=1}^K \alpha_i) - \sum_{i=1}^K \log \Gamma(\alpha_i) + \sum_{i=1}^K (\alpha_i - 1) \left(\Psi(\chi_i) - \Psi(\sum_{j=1}^K \chi_j) \right) + \\
&\quad \sum_{n=1}^{N_d} \sum_{i=1}^K \left(\Psi(\chi_i) - \Psi(\sum_{j=1}^K \chi_j) \right) \delta_{ni} +
\end{aligned}$$

$$\begin{aligned}
& \sum_{n=1}^{N_d} \sum_{i=1}^K \sum_{j=1}^{V_t} \delta_{ni} \cdot w_n^j \log \theta_{ij} + \\
& \text{constant} + \\
& \sum_{m=1}^M \sum_{i=1}^K \sum_{j=1}^{V_v} \sum_{n=1}^{N_d} \delta_{ni} \cdot \lambda_{mn} \cdot v_m^j \log \gamma_{ij} \\
& - \log \Gamma\left(\sum_{i=1}^K \chi_i\right) + \sum_{i=1}^K \log \Gamma(\chi_i) - \sum_{i=1}^K (\chi_i - 1) \left(\Psi(\chi_i) - \Psi\left(\sum_{j=1}^K \chi_j\right) \right) \\
& - \sum_{n=1}^{N_d} \sum_{i=1}^K \delta_{ni} \log \delta_{ni} \\
& - \sum_{m=1}^M \sum_{n=1}^{N_d} \lambda_{mn} \log \lambda_{mn}
\end{aligned} \tag{B.29}$$

B.3.4 Parameter Estimation

In equation B.29, we now have the variational lower bound expressed in terms of the three variational parameters and the three model parameters. We now detail how to maximize the variational lower bound w.r.t. all six parameters. First, we maximize w.r.t variational parameters χ, δ and λ .

We start with the variational Dirichlet χ by isolating the terms in equation B.29 that contains χ :

$$\begin{aligned}
\mathcal{L}_{[\chi_i]} &= \sum_{i=1}^K (\alpha_i - 1) \left(\Psi(\chi_i) - \Psi\left(\sum_{j=1}^K \chi_j\right) \right) \\
&+ \sum_{n=1}^{N_d} \sum_{i=1}^K \left(\Psi(\chi_i) - \Psi\left(\sum_{j=1}^K \chi_j\right) \right) \\
&- \log \Gamma\left(\sum_{i=1}^K \chi_i\right) + \sum_{i=1}^K \log \Gamma(\chi_i) - \sum_{i=1}^K (\chi_i - 1) \left(\Psi(\chi_i) - \Psi\left(\sum_{j=1}^K \chi_j\right) \right) \\
&= \sum_{i=1}^K \left(\Psi(\chi_i) - \Psi\left(\sum_{j=1}^K \chi_j\right) \right) \cdot \\
&(\alpha_i + \sum_{n=1}^{N_d} \delta_{ni} - \chi_i) - \log \Gamma\left(\sum_{j=1}^K \chi_j\right) + \sum_{j=1}^K \chi_j \log \Gamma(\chi_j)
\end{aligned} \tag{B.30}$$

Taking derivative with respect to χ_i , we have:

$$\frac{\partial \mathcal{L}_{[\chi_i]}}{\partial \chi_i} = \Psi'(\chi_i)(\alpha_i + \sum_{n=1}^{N_d} \delta_{ni} - \chi_i) - \Psi'(\sum_{j=1}^K \chi_j) \cdot \sum_{i=1}^K \left(\alpha_j + \sum_{n=1}^{N_d} \delta_{nj} - \chi_j \right) \quad (\text{B.31})$$

Setting this to zero yields a minimum at:

$$\chi_i = \alpha_i + \sum_{n=1}^{N_d} \delta_{ni} \quad (\text{B.32})$$

We turn our attention to the second variational multinomial δ . We follow the derivation steps in [27] and maximize equation B.29 w.r.t. δ by an Lagrangian approach, since this is an optimization problem that is constrained by $\sum_{i=1}^K \delta_{ni} = 1$. We form the Lagrangian by isolating the terms which contain δ_{ni} and adding the Lagrange multiplier $(\sum_{i=1}^K \delta_{ni} - 1)$:

$$\begin{aligned} \mathcal{L}_{[\delta_{ni}]} &= \delta_{ni} \left(\Psi(\chi_i) - \Psi(\sum_{j=1}^K \chi_j) \right) + \delta_{ni} \log \theta_{iw_n} \\ &\quad + \delta_{ni} \left(\sum_{m=1}^M \lambda_{mn} \log \gamma_{iv_m} \right) - \delta_{ni} \log \delta_{ni} + \eta \left(\sum_{i=1}^K \delta_{ni} - 1 \right) \end{aligned} \quad (\text{B.33})$$

Taking derivative w.r.t δ_{ni} , we have:

$$\frac{\partial \mathcal{L}_{[\delta_{ni}]}}{\partial \delta_{ni}} = \Psi(\chi_i) - \Psi(\sum_{j=1}^K \chi_j) + \log \theta_{iw_n} + \sum_{m=1}^M \lambda_{mn} \log \gamma_{iv_m} - \log \delta_{ni} - 1 + \eta \quad (\text{B.34})$$

Setting this to zero we have:

$$\delta_{ni} \propto \theta_{iw_n} \exp \left(\Psi(\chi_i) - \Psi(\sum_{j=1}^K \chi_j) + \sum_{m=1}^M \lambda_{mn} \log \gamma_{iv_m} \right) \quad (\text{B.35})$$

Finally, we maximize the third variational multinomial λ , by a similar Lagrangian approach, now with Lagrange multiplier $\sum_{n=1}^{N_d} \lambda_{mn} - 1$:

$$\mathcal{L}_{[\lambda_{mn}]} = \sum_{i=1}^K \delta_{ni} \cdot \lambda_{mn} \log \gamma_{iv_m} - \lambda_{mn} \log \lambda_{mn} + \eta \left(\sum_{n=1}^{N_d} \lambda_{mn} - 1 \right) \quad (\text{B.36})$$

Taking derivative w.r.t. λ_{mn} , we have:

$$\frac{\partial \mathcal{L}_{[\lambda_{mn}]}}{\partial \lambda_{mn}} = \sum_i 1^K \delta_{ni} \log \gamma_{iv_m} - \log \lambda_{mn} - 1 + \eta \quad (\text{B.37})$$

Setting this to zero, we have:

$$\lambda_{mn} \propto \exp \left(\sum_{i=1}^K \delta_{ni} \log \gamma_{iv_m} \right) \quad (\text{B.38})$$

We now turn to maximizing the three model parameters α, γ, θ . Given a training set $\mathbb{T} = \{(\mathbf{v}_d, \mathbf{w}_d)\}_{d=1}^D$, we find the maximum likelihood estimation for visual word topics γ , text topics θ and the prior α . The overall log likelihood of $\mathbb{T}$ is bounded by:

$$\mathcal{L}(\mathbb{T}) = \sum_{d=1}^D \log p(\mathbf{v}_d, \mathbf{w}_d | \alpha, \theta, \gamma) \geq \sum_{d=1}^D \mathcal{L}(\chi_d, \delta_d, \lambda_d; \alpha, \theta, \gamma) \quad (\text{B.39})$$

We want to find α, θ, γ that maximize this lower bound. We start with the conditional multinomial θ . We again isolate terms in equation B.29 that contains θ . Then we add Lagrange multipliers $\sum_{j=1}^{V_t} \theta_{ij} - 1$:

$$\mathcal{L}_{[\theta_{ij}]} = \sum_{d=1}^D \sum_{n=1}^{N_d} \sum_{i=1}^K \delta_{dni} \log \theta_{iwn} - \eta \left(\sum_{j=1}^{V_t} \theta_{ij} - 1 \right) \quad (\text{B.40})$$

Setting to zero the above derivative w.r.t. θ_{ij} , we have:

$$\theta_{ij} \propto \sum_{d=1}^D \sum_{n=1}^{N_d} \delta_{dni} w_{dn}^j \quad (\text{B.41})$$

We proceed onto the second model parameter, which is the conditional multinomial γ . Again, we isolate terms in equation B.29 containing γ . Then we add Lagrange multipliers $\sum_{j=1}^{V_v} \gamma_{ij} - 1$:

$$\mathcal{L}_{[\gamma_{ij}]} = \sum_{d=1}^D \sum_{m=1}^{M_d} \sum_{i=1}^K \sum_{n=1}^{N_d} \delta_{dni} \cdot \lambda_{mn} \log \gamma_{iv_m} - \eta \left(\sum_{j=1}^{V_v} \gamma_{ij} - 1 \right) \quad (\text{B.42})$$

Setting to zero the above derivative w.r.t. γ_{ij} , we have:

$$\gamma_{ij} \propto \sum_{d=1}^D \sum_{m=1}^{M_d} v_{dm}^j \sum_{n=1}^{N_d} \delta_{dni} \lambda_{dmn} \quad (\text{B.43})$$

Finally, for the Dirichlet prior α , we follow the approach in [27] and use a Newton-Raphson algorithm to maximize α .

B.3.5 Variational Expectation Maximization

We summarize the derivation steps described in the previous subsection. The entire algorithm constitutes a variational expectation maximization (EM) approach to compute the maximum likelihood estimation of the model parameters. The E-step corresponds to computing the posterior distribution of the variational parameters $\{\chi_d, \delta_d, \lambda_d\}_{d=1}^D$ for each multimedia document d , given the current settings of the model parameters α, γ, θ . The M-step then computes the maximum likelihood estimates of α, γ, θ using the expected sufficient statistics taken from the variational distribution. The variational EM alternates between these two steps until the bound on the expected log likelihood converges.

Bibliography

- [1] Carrot2, <http://project.carrot2.org/index.html>.
- [2] *Lemur Toolkit*. <http://www.lemurproject.org>.
- [3] *TRECVID 2005*. <http://www-nlpir.nist.gov/projects/trecvid>.
- [4] Vivisimo web clustering. In <http://vivisimo.com/>.
- [5] Wordnet, a lexical database for the english language. In <http://wordnet.princeton.edu/>.
- [6] K. Ali and V. Stam. TiVo: Making show recommendations using a distributed collaborative filtering architecture. In *Proc. ACM International Conference on Knowledge Discovery and Data Mining*, pages 394–401, 2004.
- [7] D. Andrzejewski and X. Zhu. Latent Dirichlet Allocation with topic-in-set knowledge. In *Proc. The Annual Conference of the North American Chapter of the Association for Computational Linguistics Workshop on Semi-supervised Learning for Natural Language Processing*, 2009.
- [8] T. Arni, P. Clough, M. Sanderson, and M. Grubinger. Overview of the ImageCLEFphoto 2008 photographic retrieval task. In *Working Notes of the 2008 CLEF Workshop*, 2008.
- [9] A. Asuncion, P. Smyth, and M. Welling. Asynchronous distributed learning of topic models. In *Advances in Neural Information Systems*, pages 81–88, 2008.
- [10] R. Baeza-Yates, L. Calderon-Benavides, and C. Gonzalez-Caro. The intention behind web queries. In *Proc. String Processing and Information Retrieval*, volume 4209, pages 98–109. Springer, 2006.

- [11] R. Baeza-Yates, C. Hurtado, and M. Mendoza. Query recommendation using query logs in search engines. *Lecture Notes in Computer Science*, 3268:588–596, 2004.
- [12] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.
- [13] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M. Jordan. Matching words and pictures. *Journal Machine Learning Research*, 3:1107–1135, 2003.
- [14] D. Bartholomew, M. Knott, and I. Moustaki. *Latent variable models and factor analysis: a unified approach*, volume 43. John Wiley & Sons, 2011.
- [15] I. Bazzi and J. Glass. Modeling out-of-vocabulary words for robust speech recognition. *Proc International Conference on Spoken Language Processing*, 2000.
- [16] M. Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.
- [17] S. Beitzel, E. Jensen, O. Frieder, D. Grossman, D. Lewis, A. Chowdhury, and A. Kolcz. Automatic web query classification using labeled and unlabeled training data. In *Proc. ACM SIGIR conference on Research and development in information retrieval*, pages 581–582. ACM, 2005.
- [18] S. Beitzel, E. Jensen, and D. Grossman. A survey of retrieval strategies for ocr text collections. In *Proc. of the Symposium on Document Image Understanding Technologies*, 2003.
- [19] R. Bellman. Adaptive control processes: a guided tour. *Princeton University Press, Princeton, New Jersey, USA*, 19:94, 1961.
- [20] J. Belongie, J. Shi, and T. Leung. Textons, contours and regions: Cue integration in image segmentation. In *Proc. International Conference on Computer Vision*, 1999.
- [21] S. Bengio, F. Pereira, Y. Singer, and D. Strelow. Group sparse coding. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, pages 82–89. 2009.
- [22] S. Bhatia, D. Majumdar, and P. Mitra. Query suggestions in the absence of query logs. In *Proc ACM conference on research and development in information retrieval*, 2011.

- [23] J. Bilmes. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Technical report, University of California at Berkeley, 1998.
- [24] D. Blei, T. Griffiths, and M. Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57(2):1–30, 2010.
- [25] D. Blei and M. Jordan. Modeling annotated data. In *Proc. ACM SIGIR conference on research and development in information retrieval*, pages 127–134, 2003.
- [26] D. Blei and J. Lafferty. A correlated topic model of science. *Annals of Applied Statistics*, 1(1):17–35, 2007.
- [27] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. In *Journal of Machine Learning Research, Vol. 3*, pages 993–1022, 2003.
- [28] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
- [29] P. Boldi, F. Bonchi, C. Castillo, D. Donato, and S. Vigna. Query suggestions using query-flow graphs. In *Proc workshop on Web Search Click Data*, pages 56–63, 2009.
- [30] J. Boyd-Graber and D. Blei. Syntactic topic models. In *Proc of Neural Information Processing Systems*, 2008.
- [31] J. Boyd-Graber, D. Blei, and X. Zhu. A topic model for word sense disambiguation. In *Proc. Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1024–1033, 2007.
- [32] P. Bradley and U. Fayyad. Refining initial points for k-means clustering. In *Proc. International Conference on Machine Learning*, pages 91–99, 1998.
- [33] K. Brinker. Incorporating diversity in active learning with support vector machines. In *Proc. International Conference on Machine Learning*, pages 59–66, 2003.
- [34] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36:3–10, 2002.

- [35] D. Cai, X. He, Z. Li, W. Ma, and J. Wen. Hierarchical clustering of www image search results using visual, textual and link information. In *Proc. Multimedia*, pages 952–959, 2004.
- [36] L. Caldern-Benavides, C. Gonzalez-Caro, and R. Baeza-Yates. Towards a deeper understanding of the users query intent. In *Proc. In Query Representation and Understanding, workshop at SIGIR 2010*, 2010.
- [37] J. Cao, J. Li, Y. Zhang, and S. Tang. LDA-based retrieval framework for semantic news video retrieval. In *Proc. International Conference on Semantic Computing*, pages 155–160, 2007.
- [38] J. Cao, J. Li, Y. Zhang, and S. Tang. LDA-based retrieval framework for semantic news video retrieval. In *Proc International Conference on Semantic Computing*, pages 155–160, 2007.
- [39] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proc. ACM SIGIR conference on research and development in information retrieval*, pages 335–336, 1998.
- [40] B. Carterette and P. Chandar. Probabilistic models of ranking novel documents for faceted topic retrieval. *Proc. CIKM*, pages 1287–1296, 2009.
- [41] S. Chang, W. Hsu, W. Jiang, L. Kennedy, D. Xu, A. Yanagawa, and E. Zavesky. Evaluating the impact of 374 visual based LSCOM concept detectors on automatic search. In *TREC Video Retrieval Evaluation Online Proceedings*, 2006.
- [42] E. Charniak. *Statistical language learning*. Cambridge, MA: MIT Press, 1993.
- [43] H. Chen and D. Karger. Less is more: Probabilistic models for retrieving fewer relevant documents. *Proc. SIGIR*, pages 429–436, 2006.
- [44] O. Chum and J. Matas. Unsupervised discovery of co-occurrence in sparse high dimensional data. In *Computer Vision and Pattern Recognition*, pages 3416–3423, 2010.
- [45] O. Chum, J. Philbin, M. Isard, and A. Zisserman. Scalable near identical image and shot detection. *Proc. ACM International Conference on Image and Video Retrieval*.

- [46] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *Proc. International Conference on Computer Vision*, 2007.
- [47] K. Church and B. Thiesson. The wild thing! In *Proc Association for Computational Linguistics*, pages 93–96, 2005.
- [48] R. Cilibrasi and P.M.B. Vitanyi. The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370–383, 2007.
- [49] C. Clarke, M. Kolla, G. Cormack, O. Vechtomova, A. Ashkan, S. Butcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proc. ACM SIGIR conference on research and development in information retrieval*, pages 659–666, 2008.
- [50] A. Collins and E. Loftus. A spreading activation theory of semantic processing. In *Psychological Review*, 82, pages 407–428, 1975.
- [51] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.
- [52] F. Crestani. Exploiting the similarity of non-matching terms at retrieval time. *Journal of Information Retrieval*, 2:25–45, 2000.
- [53] S. Cucerzan and E. Brill. Spelling correction as an iterative process that exploits the collective knowledge of web users. In *Proc. Conference on Empirical Methods in Natural Language Processing*, 2004.
- [54] H. Cui, J. Wen, J. Nie, and W. Ma. Probabilistic query expansion using query logs. In *Proc. World Wide Web*, 2002.
- [55] D. Cutting, D. Karger, J. Pedersen, and J. Tukey. Scatter/gather: a cluster-based approach to browsing large document collections. In *Proc. ACM SIGIR conference on research and development in information retrieval*, 1992.
- [56] W. Dakka. Automatic construction of multifaceted browsing interfaces. In *Proc. ACM Conference on Information and Knowledge Management*, pages 768–775, 2005.

- [57] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman. Indexing by latent semantic analysis. In *Journal of the American Society of Information Science*, 41(6), pages 391–407, 1990.
- [58] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from Google image search. In *Proc. International Conference on Computer Vision*, 2005.
- [59] T. Ferguson. A bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1, 1973.
- [60] A. Feuer, S. Savev, Aslam, and A. Javed. Evaluation of phrasal query suggestions. In *Proc ACM Conference on information and knowledge management*, pages 841–848, 2007.
- [61] G. Friedland, L. Gottlieb, and A. Janin. Joke-o-mat: browsing sitcoms punchline by punchline. In *Proc. ACM Multimedia*, 2009.
- [62] J. Friedman. On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1(1):55–77, 1997.
- [63] B. Frigiyik, A. Kapila, and M. Gupta. Introduction to the dirichlet distribution and related processes. Technical report, Department of Electrical Engineering, University of Washington, 2010.
- [64] M. Fritz and B. Schiele. Decomposition, discovery and detection of visual categories using topic models. In *Proc. Computer Vision and Pattern Recognition*, 2008.
- [65] S. Gao, L. Chia, and I. Tsang. Multi-layer group sparse coding - for concurrent image classification and annotation. In *Proc. Computer Vision and Pattern Recognition*, pages 2809–2816, 2011.
- [66] J. Garofolo, C. Auzanne, and E. Voorhees. The trec spoken document retrieval track: A success story. In *in Text Retrieval Conference (TREC) 8*, pages 16–19, 2000.
- [67] W. Goffman. On relevance as a measure. *Information Storage and Retrieval*, 2:201–203, 1964.

- [68] L. Granka, T. Joachims, and G. Gay. Eye-tracking analysis of user behavior in www search. In *Proc. ACM SIGIR conference on Research and development in information retrieval*, pages 478–479, 2004.
- [69] T. Griffiths and M. Steyvers. Finding scientific topics. In *Proc. of the National Academy of Sciences 101 (suppl. 1)*, pages 5228–5235, 2004.
- [70] T. Griffiths, M. Steyvers, D. Blei, and J. Tenenbaum. Integrating topics and syntax. In *Proc. Neural Information Processing Systems*, 2005.
- [71] T. Griffiths, M. Steyvers, and J. Tenenbaum. Topics in semantic representation. In *Psychological Review, Vol. 114, no. 2*, pages 211–244, 2007.
- [72] A. Gruber, M. Rosen-zvi, and Y. Weiss. Hidden topic markov models. In *Proc. of Artificial Intelligence and Statistics*, 2007.
- [73] D. Harman. Overview of the TREC 2002 novelty track. In *Proc. of the Eleventh Text Retrieval Conference (TREC)*, 2002.
- [74] M. Hearst. Clustering versus faceted categories for information exploration. 49, 2006, pages=59-61.
- [75] M. Hearst and J. Pedersen. Reexamining the cluster hypothesis: scatter/gather on retrieval results. In *Proc. ACM SIGIR conference on Research and development in information retrieval*, pages 76–84, 1996.
- [76] M. Heritier, S. Foucher, and L. Gagnon. Key-places detection and clustering in movies using latent aspects. In *Proc. International Conference Image Processing, Vol. 2*, pages 225–228, 2007.
- [77] M. Hoffman, D. Blei, and P. Cook. Content-based musical similarity computation using the hierarchical dirichlet process. In *Proc. International Conference on Music Information Retrieval*, 2008.
- [78] T. Hofmann. Probabilistic latent semantic analysis. In *Proc. Conference on Uncertainty in Artificial Intelligence*, 1999.

- [79] W. Hsu, L. Kennedy, and S. Chang. Video search reranking through random walk over document-level context graph. In *Proc. ACM Multimedia*, pages 971–980, 2007.
- [80] C. Huang, L. Chien, and Y. Oyang. Relevant term suggestion in interactive web search based on contextual information in query session logs. In *Journal of the American Society for Information Science and Technology*, 54(7), pages 638–649, 2003.
- [81] ImageNet. <http://www.image-net.org/>.
- [82] P. Indyk. A sublinear-time approximation scheme for clustering in metric spaces. In *Proc. Symposium on Foundations of Computer Science*, pages 154–159, 1999.
- [83] A. Jain and R. Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.
- [84] A. Jain and G. Mishne. Organizing query completions for web search. In *Proc ACM international conference on Information and knowledge management*, pages 1169–1178, 2010.
- [85] V. Jain, E. Learned-Miller, and A. McCallum. People-LDA: Anchoring topics to people using face recognition. In *Proc. International Conference on Computer Vision*, 2007.
- [86] B. Jansen, D. Booth, and A. Spink. Determining the informational, navigational, and transactional intent of web queries. *Information Processing and Management*, 44:1251–1266, 2008.
- [87] Y. Jing and S. Baluja. PageRank for product image search. In *Proc. World Wide Web*, 2008.
- [88] M. Jordan. Why the logistic function? a tutorial discussion on probabilities and neural networks. Technical report, Massachusetts Institute of Technology, 1995.
- [89] M. Jordan, G. Zoubin, J. Tommi, and S. Lawrence. An introduction to variational methods for graphical models. In *Proc. of the NATO Advanced Study Institute on Learning in graphical models*, pages 105–161, 1998.
- [90] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *Proc. International Conference on Computer Vision*, volume 1, pages 604–610, 2005.
- [91] Saenko K and T. Darrell. Unsupervised learning of visual sense models for polysemous words. In *Proc. Neural Information Processing Systems*, 2008.

- [92] W. Kintsch. The role of knowledge in discourse comprehension: A construction-integration model. In *Psychological Review*, 95, pages 163–182, 1988.
- [93] R. Krovetz and B. Croft. Lexical ambiguity and information retrieval. *ACM Transactions on Information Systems*, 10:115–141, April 1992.
- [94] T. Landauer and S. Dumais. A solution to platos problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. In *Psychological Review*, 104, pages 211–240, 1997.
- [95] I. Laptev and T. Lindeberg. Space-time interest points. In *Proc. International Conference on Computer Vision*, pages 432–439, 2003.
- [96] M. Larson, R. Ordelman, F. de Jong, J. Kohler, and W. Kraaij. Multimedia with a speech track: searching spontaneous conversational speech. *SIGIR Forum*, 44:76–81, August 2010.
- [97] V. Lavrenko and W. Croft. Relevance-based language models. In *Proc. ACM SIGIR conference on research and development in information retrieval*, pages 120–127, 2001.
- [98] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *Proc. Neural Information Processing Systems*. MIT Press, 2003.
- [99] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 2169–2178, 2006.
- [100] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. Computer Vision and Pattern Recognition*, pages 2169–2178, 2006.
- [101] U. Lee, Z. Liu, and J. Cho. Automatic identification of user goals in web search. In *Proc. World Wide Web*, pages 391–400, 2005.
- [102] M. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions Multimedia Computing, Communications, Applications*, 2(1):1–19, 2006.

- [103] R. Lienhart, S. Romberg, and E. Horster. Multilayer plsa for multimodal image retrieval. *Proc. ACM International Conference on Image and Video Retrieval*, pages 1–8, 2009.
- [104] N. Loeff, C. Alm, and D. Forsyth. Discriminating image senses by clustering with multimodal features. In *Proc. Association of Computational Linguistics*, pages 547–554, 2006.
- [105] D. Lowe. Distinctive image features from scale-invariant keypoints. In *Journal of Computer Vision*, 60(2), pages 91–110, 2004.
- [106] K. Lund and C. Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. In *Behavior Research Methods, Instruments and Computers*, 28, pages 203–208, 1996.
- [107] N. Madnani and B. Dorr. Generating phrasal and sentential paraphrases: A survey of data-driven methods. In *Journal of Computational Linguistics*, 36(3):341–387, 2010.
- [108] C. Manning and H. Schutze. *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press, 1999.
- [109] J. Matas, O. Chum, M. Urba, and T. Pajdla. Robust wide baseline extremal regions. In *Proc. British Machine Vision Conference*, pages 384–396, 2002.
- [110] A. McCallum, X. Wang, and A. Corrada-Emmanuel. Topic and role discovery in social networks with experiments on enron and academic email. In *Journal of Artificial Intelligence Research*, 2007.
- [111] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai. Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs. In *Proc. international conference on World Wide Web*, pages 171–180, 2007.
- [112] R. Mihalcea. Using Wikipedia for automatic word sense disambiguation. In *Proc. The Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2007.

- [113] D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum. Optimizing semantic coherence in topic models. In *Proc. Conference on Empirical Methods in Natural Language Processing*, 2011.
- [114] F. Monay and D. Gatica-perez. Modeling semantic aspects for cross-media image retrieval. *IEEE PAMI*, 29:1802–1817, 2007.
- [115] R. Nallapati and W. Cohen. Link-plsa-lda: A new unsupervised model for topics and influence of blogs. In *International Conference on Weblogs and Social Media*, 2008.
- [116] R. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, pages 249–265, 2000.
- [117] D. Newman, A. Asuncion, P. Smyth, and M. Welling. Distributed algorithms for topic models. *Journal of Machine Learning Research*, 10:1801–1828, December 2009.
- [118] D. Newman, J. Lau, K. Grieser, and T. Baldwin. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108, 2010.
- [119] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14*, pages 849–856, 2001.
- [120] V. Nguyen and M. Kan. Functional faceted web query analysis. In *Query Log Analysis: Social And Technological Challenges.*, 2007.
- [121] M. Paramita, M. Sanderson, and P. Clough. Diversity in photo retrieval: overview of the ImageCLEFphoto task 2009. *CLEF working notes*, 2009.
- [122] D. Pedro and P. Michael. On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29(2-3):103–130, 1997.
- [123] J. Philbin, J. Sivic, and A. Zisserman. Geometric LDA: A generative model for particular object discovery. In *Proc. British Machine Vision Conference*, 2008.

- [124] M. Purver, K. Krding, T. Griffiths, and J. Tenenbaum. Unsupervised topic modelling for multi-party spoken discourse. In *Proc. International Conference on Computational Linguistics and the annual meeting of the Association for Computational Linguistics*, pages 17–24, 2006.
- [125] Y. Qiu and H. Frei. Concept based query expansion. In *Proc. ACM SIGIR conference on research and development in information retrieval*, pages 160–170, 1993.
- [126] P. Quelhas, F. Monay, J. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. Van Gool. Modeling scenes with local descriptors and latent aspects. In *Proc. International Conference on Computer Vision*, 2005.
- [127] P. Quelhas, F. Monay, J. Odobez, D. Gatica-Perez, T. Tuytelaars, L. V., and V. Gool. Modeling scenes with local descriptors and latent aspects. In *Proc. International Conference on Computer Vision*, 2005.
- [128] F. Radlinski and S. Dumais. Improving personalized web search using result diversification. In *Proc. ACM SIGIR conference on research and development in information retrieval*, pages 691 – 692, 2006.
- [129] D. Ramage, D. Hall, R. Nallapati, and C. Manning. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pages 248–256. Association for Computational Linguistics, 2009.
- [130] C. Rasmussen. The infinite gaussian mixture model. In *Neural Information Processing Systems*, 2000.
- [131] C. Rijsbergen. *Information Retrieval*. University of Glasgow, 1979.
- [132] S. Robertson. The probability ranking principle in ir. *Journal of Documentation*, 33(4):294–304, 1977.
- [133] D. Rose and D. Levinson. Understanding user goals in web search. In *Proc. World Wide Web*, pages 13–19. ACM, 2004.

- [134] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *Proc. Conference on Uncertainty in Artificial Intelligence*, pages 487–494, 2004.
- [135] S. Rudinac, M. Larson, and A. Hanjalic. Exploiting visual reranking to improve pseudo-relevance feedback for spoken-content-based video retrieval. In *Workshop on Image Analysis for Multimedia Interactive Services*, pages 17–20, 2009.
- [136] B. Russell, A. Efros, J. Sivic, W. Freeman, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *Proc. Computer Vision and Pattern Recognition*, 2006.
- [137] K. Saenko and T. Darrell. Unsupervised learning of visual sense models for polysemous words. In *Proc. Neural Information Processing Systems*, 2008.
- [138] G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986.
- [139] G. Salton, A. Wong, C. S., and S. Yang. A vector space model for automatic indexing. In *Communications of the ACM*, 18, pages 613–620, 1975.
- [140] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting image databases from the web. In *Proc. International Conference on Computer Vision*, 2007.
- [141] F. Seide, K. Thambiratnam, L. Lu, and R. Yu. Multimedia retrieval through indexing speech: An enterprise perspective. In *Proc of the third workshop on Searching spontaneous conversational speech*, pages 1–2, 2009.
- [142] A. Singhal. Modern information retrieval: A brief overview. In *IEEE Data Engineering Bulletin*, 24(4):35–43, 2001.
- [143] A. Singhal and F. Pereira. Document expansion for speech retrieval. In *Proc. ACM SIGIR conference on Research and development in information retrieval*, pages 34–41, 1999.

- [144] A. Smeulders, M. Worring, S. Santini, and A. Gupta. Content-based image retrieval at the end of the early years. In *IEEE Transactions in Pattern Analysis Machine Intelligence*, 22(12), pages 1349–1380, 2000.
- [145] A. Smola and S. Narayanamurthy. An architecture for parallel topic models. *Proc. of the Very Large Data Bases Endowment*, 3(1):703–710, 2010.
- [146] I. Soboroff. Overview of the TREC 2004 novelty track. In *Proc. of the Thirteenth Text Retrieval Conference (TREC)*, 2004.
- [147] K. Song, Y. Tian, W. Gao, and T. Huang. Diversifying the image retrieval results. In *Proc. ACM Multimedia*, pages 707–710, 2006.
- [148] Y. Song, J. Huang, I. Councill, J. Li, and C. Giles. Efficient topic-based unsupervised name disambiguation. In *Proc. ACM/IEEE Joint Conference on Digital libraries*, pages 342–351, 2007.
- [149] S. Spaccapietra and R. Jain, editors. *Visual Information Management*.
- [150] E. Sudderth, A. Torralba, W. Freeman, and A. Willsky. Describing visual scenes using transformed objects and parts. In *Proc. International Journal of Computer Vision*, 77(13), 2008.
- [151] T. Tao, X. Wang, Q. Mei, and C. Zhai. Language model information retrieval with document expansion. In *Proc. of the Annual Conference of the North American Chapter of the Association of Computational Linguistics*, pages 407 – 414, 2006.
- [152] Y. Teh, M. Jordon, M. Beal, and D. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1556–1581, 2007.
- [153] M. Tipping and C. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, 61:611–622, 1999.
- [154] S. Tong and E. Chang. Support vector machine active learning for image retrieval. In *Proc. international conference on Multimedia*, pages 107–118, 2001.

- [155] L. Torresani, M. Szummer, and A. Fitzgibbon. Efficient object category recognition using classemes. In *European Conference on Computer Vision (ECCV)*, pages 776–789, September 2010.
- [156] S. Townsend and B. Croft. Quantifying query ambiguity. In *Proc. International Conference on Human Language Technology Research*, pages 104–109, 2002.
- [157] P. Turney. Learning algorithms for keyphrase extraction. *Information Retrieval*, 2:303–336, 2000.
- [158] P. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proc. Association of Computational Linguistics*, pages 417–424, 2002.
- [159] J. van Gemert, C. Veenman, A. Smeulders, and J. Geusebroek. Visual word ambiguity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(7):1271–1283, 2010.
- [160] V. Vapnik. *Statistical learning theory*. John Wiley, 1988.
- [161] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms, 2008. <http://www.vlfeat.org/>.
- [162] M. Wainwright and M. Jordan. *Graphical Models, Exponential Families, and Variational Inference*. 2008.
- [163] D. Walker, W. Lund, and E. Ringger. Evaluating models of latent document semantics in the presence of ocr errors. In *Proc. Conference on Empirical Methods in Natural Language Processing*, pages 240–250. Association for Computational Linguistics, 2010.
- [164] K. Wan. Exploiting story-level context to improve video search. In *Proc. International Conference on Multimedia and Expo*, 2008.
- [165] K. Wan, A. Tan, J. Lim, and L. Chia. A latent model for visual disambiguation of keyword-based image search. In *Proc. British Machine Vision Conference*, 2009.
- [166] K. Wan, A. Tan, J. Lim, and L. Chia. A non-parametric visual-sense model of images – extending the cluster hypothesis beyond text. *Multimedia Tools and Applications*, pages 1–26, 2010. 10.1007/s11042-010-0615-y.

- [167] K. Wan, A. Tan, J. Lim, and L. Chia. Topic based query suggestions for video search. In K. Schoeffmann, B. Merialdo, A. Hauptmann, C. Ngo, Y. Andreopoulos, and C. Breiteneder, editors, *Advances in Multimedia Modeling*, volume 7131 of *Lecture Notes in Computer Science*, pages 288–299. Springer Berlin / Heidelberg, 2012.
- [168] Kong-Wah WAN, Ah-Hwee TAN, Joo-Hwee LIM, and Liang-Tien CHIA. Topic-based query suggestions for video search. In *Proc International Multimedia Modeling Conference*, 2011.
- [169] Kong-Wah WAN, Yan-Tao ZHENG, and Lekha CHAISORN. Known-item video search via query-to-modality mapping. In *Proc ACM Multimedia*, 2011.
- [170] X. Wei and W. Croft. LDA-based document models for ad-hoc retrieval. In *Proc. ACM SIGIR conference on research and development in information retrieval*, pages 178–185, 2006.
- [171] J. Wen, J. Nie, and H. Zhang. Query clustering using user logs. In *ACM Transactions Information Systems, Vol. 20, no. 1*, pages 59–81, 2002.
- [172] X. Wu, A. Hauptmann, and C. Ngo. Novelty detection for cross-lingual news stories with visual duplicates and speech transcripts. In *Proc. ACM Multimedia*, pages 168–177, 2007.
- [173] J. Xu and W. Croft. Query expansion using local and global document analysis. In *Proc. ACM SIGIR conference on research and development in information retrieval*, pages 4–11, 1996.
- [174] J. Yuan, Y. Wu, and M. Yang. Discovery of collocation patterns: from visual words to visual phrases. pages 1–8, 2007.
- [175] H. Zeng, Q. He, Z. Chen, W. Ma, and J. Ma. Learning to cluster web search results. In *Proc. ACM SIGIR conference on research and development in information retrieval*, 2004.
- [176] C. Zhai, W. Cohen, and J. Lafferty. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *Proc. ACM SIGIR conference on research and development in information retrieval*, pages 10–17, 2003.
- [177] K. Zhai, J. Boyd-Graber, N. Asadi, and M. Alkhouja. Mr. LDA: A flexible large scale topic modeling package using variational inference in MapReduce. In *Proc. on World Wide Web*, 2012.

- [178] B. Zhang, H. Li, Y. Liu, L. Ji, W. Xi, W. Fan, Z. Chen, and W. Ma. Improving web search results using an affinity graph. In *Proc. ACM SIGIR conference on research and development in information retrieval*, pages 504–511, 2005.
- [179] Z. Zhang and O. Nasraoui. Mining search engine query logs for query recommendation. In *Proc International World Wide Web Conference*, 2006.
- [180] Y. Zheng, S. Neo, T. Chua, and Q. Tian. Visual synset: a higher-level visual representation for object-based image retrieval. *The Visual Computer*, 25:13–20, 2009.
- [181] J. Zhu, A. Goldberg, J. Van, and D. Andrzejewski. Improving diversity in ranking using absorbing random walks. In *Proc. The Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2007.
- [182] C. Ziegler, S. Mcnee, J. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In *Proc. International Conference on World Wide Web*, pages 22–32, 2005.
- [183] L. Zighelnic and O. Kurland. Query-drift prevention for robust query expansion. In *Proc. ACM SIGIR conference on Research and development in information retrieval*, pages 825–826. ACM, 2008.