Modeling Social Annotation Data with Content Relevance using a Topic Model

T. Iwata, T. Yamada and N. Ueda
NIPS 2009

presented by J. Silva, Duke University

Introduction

- Context: social annotations, aka collaborative tagging or folksonomy
- Users freely annotate objects such as webpages, photos, blog posts, videos, music and scientific papers
- Examples: Delicious, Flickr, Technorati, YouTube, Last.fm, CiteULike

Introduction

- Problem: users often write *noisy*, or content-irrelevant annotations (e.g. "great", "to read")
- This paper proposes a generative model for topics and annotations that takes into account relevance/irrelevance of the annotations
- It is an extension of Blei and Jordan's Correspondence Latent Dirichlet Allocation (Corr-LDA), which assumes the annotations are always relevant

Proposed Method

- We have a set of D documents, and each consists of a pair of words and annotations $(\mathbf{w}_d, \mathbf{t}_d)$, where $\mathbf{w}_d = \{w_{dn}\}_{n=1}^{N_d}$ and $\mathbf{t}_d = \{t_{dm}\}_{m=1}^{M_d}$

Table 1: Notation

Symbol	Description
D	number of documents
W	number of unique words
T	number of unique annotations
K	number of topics
N_d	number of words in the dth document
M_d	number of annotations in the dth document
w_{dn}	nth word in the dth document, $w_{dn} \in \{1, \cdots, W\}$
z_{dn}	topic of the nth word in the dth document, $z_{dn} \in \{1, \cdots, K\}$
t_{dm}	mth annotation in the dth document, $t_{dm} \in \{1, \cdots, T\}$
c_{dm}	topic of the mth annotation in the dth document, $c_{dm} \in \{1, \cdots, K\}$
r_{dm}	relevance to the content of the mth annotation of the dth document, $r_{dm} = 1$ if relevant, $r_{dm} = 0$ otherwise

Note: all tables and figures taken from the original paper

Proposed Method

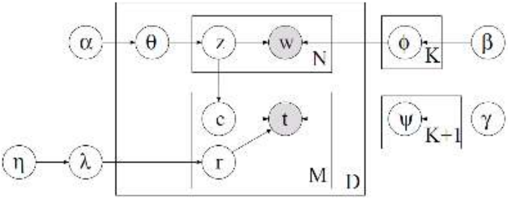


Figure 1: Graphical model representation of the proposed topic model with content relevance

- There are $K+1$ annotation distributions ($\boldsymbol{\Psi}$), since $\boldsymbol{\Psi}_0$ is a topic-unrelated distribution that applies to irrelevant annotations

Proposed Method: Generative Model

1. Draw relevance probability $\lambda \sim \text{Beta}(\eta)$
2. Draw content-unrelated annotation probability $\boldsymbol{\psi}_0 \sim \text{Dirichlet}(\gamma)$
3. For each topic $k = 1, \cdots, K$:
   (a) Draw word probability $\boldsymbol{\phi}_k \sim \text{Dirichlet}(\beta)$
   (b) Draw annotation probability $\boldsymbol{\psi}_k \sim \text{Dirichlet}(\gamma)$
4. For each document $d = 1, \cdots, D$:
   (a) Draw topic proportions $\boldsymbol{\theta}_d \sim \text{Dirichlet}(\alpha)$
   (b) For each word $n = 1, \cdots, N_d$:
      i. Draw topic $z_{dn} \sim \text{Multinomial}(\boldsymbol{\theta}_d)$
      ii. Draw word $w_{dn} \sim \text{Multinomial}(\boldsymbol{\phi}_{z_{dn}})$
   (c) For each annotation $m = 1, \cdots, M_d$:
      i. Draw topic $c_{dm} \sim \text{Multinomial}(\{\frac{N_{kd}}{N_d}\}_{k=1}^{K})$
      ii. Draw relevance $r_{dm} \sim \text{Bernoulli}(\lambda)$
      iii. Draw annotation $t_{dm} \sim \begin{cases} \text{Multinomial}(\boldsymbol{\psi}_0) & \text{if } r_{dm} = 0 \\ \text{Multinomial}(\boldsymbol{\psi}_{c_{dm}}) & \text{otherwise} \end{cases}$

Proposed Method: Inference

- The joint distribution is $P(\mathbf{W}, \mathbf{T}, \mathbf{Z}, \mathbf{C}, \mathbf{R}|\alpha, \beta, \gamma, \eta) = P(\mathbf{Z}|\alpha)P(\mathbf{W}|\mathbf{Z}, \beta)P(\mathbf{T}|\mathbf{C}, \mathbf{R}, \gamma)P(\mathbf{R}|\eta)P(\mathbf{C}|\mathbf{Z})$
- In the expression above, we have: $\mathbf{W} = \{\mathbf{w}_d\}_{d=1}^D$, $\mathbf{T} = \{\mathbf{t}_d\}_{d=1}^D$, $\mathbf{Z} = \{\mathbf{z}_d\}_{d=1}^D$, $\mathbf{C} = \{\mathbf{c}_d\}_{d=1}^D$, $\mathbf{c}_d = \{c_{dm}\}_{m=1}^{M_d}$, $\mathbf{R} = \{\mathbf{r}_d\}_{d=1}^D$, $\mathbf{r}_d = \{r_{dm}\}_{m=1}^{M_d}$
- $\boldsymbol{\theta}$, $\boldsymbol{\Phi}$, $\boldsymbol{\Psi}$ and λ are marginalized out

Proposed Method: Inference

- $P(\mathbf{Z}|\alpha) = \prod_{d=1}^{D} \int P(\mathbf{Z}|\boldsymbol{\theta}_d) P(\boldsymbol{\theta}_d|\alpha) d\boldsymbol{\theta}_d = \left(\frac{\Gamma(\alpha K)}{\Gamma(\alpha)^K}\right)^D \prod_d \frac{\prod_k \Gamma(N_{kd}+\alpha)}{\Gamma(N_{kd}+\alpha K)}$
- $P(\mathbf{W}|\mathbf{Z}, \beta) = \left(\frac{\Gamma(\beta W)}{\Gamma(\beta)^W}\right)^K \prod_k \frac{\prod_w \Gamma(N_{kw}+\beta)}{\Gamma(N_{kw}+\beta W)}$
- $P(\mathbf{T}|\mathbf{C}, \mathbf{R}, \gamma) = \left(\frac{\Gamma(\gamma T)}{\Gamma(\gamma)^T}\right)^{K+1} \prod_{k'} \frac{\prod_t \Gamma(N_{k't}+\gamma)}{\Gamma(N_{k't}+\gamma T)}$, where $k' \in \{0, \ldots, K\}$
- $P(\mathbf{R}|\eta) = \frac{\Gamma(2\eta)}{\Gamma(\eta)^2} \frac{\Gamma(M_0+\eta)\Gamma(M-M_0+\eta)}{\Gamma(M+2\eta)}$
- $P(\mathbf{C}|\mathbf{Z}) = \prod_d \prod_k \left(\frac{N_{kd}}{N_d}\right)^{M'_{kd}}$
- Inference of the latent $\mathbf{Z}|\mathbf{W}, \mathbf{T}$ is done using collapsed Gibbs sampling
- The hyperparameters are estimated by maximizing the joint distribution, using a fixed-point iteration method

Proposed Method: Inference

- We have the following expressions, where $j = (d, n)$, $i = (d, m)$ and $\backslash j$ denotes the count excluding the n-th word in the d-th document

$$P(z_j = k|\boldsymbol{W}, \boldsymbol{T}, \boldsymbol{Z}_{\backslash j}, \boldsymbol{C}, \boldsymbol{R}) \propto \frac{N_{kd\backslash j} + \alpha}{N_{d\backslash j} + \alpha K} \frac{N_{kw_j\backslash j} + \beta}{N_{k\backslash j} + \beta W} \left(\frac{N_{kd\backslash j} + 1}{N_{kd\backslash j}} \frac{N_d - 1}{N_d} \right)^{M'_{kd}}$$

$$P(r_i = 0|\boldsymbol{W}, \boldsymbol{T}, \boldsymbol{Z}, \boldsymbol{C}, \boldsymbol{R}_{\backslash i}) \propto \frac{M_{0\backslash i} + \eta}{M_{\backslash i} + 2\eta} \frac{M_{0t_i\backslash i} + \gamma}{M_{0\backslash i} + \gamma T},$$

$$P(r_i = 1|\boldsymbol{W}, \boldsymbol{T}, \boldsymbol{Z}, \boldsymbol{C}, \boldsymbol{R}_{\backslash i}) \propto \frac{M_{\backslash i} - M_{0\backslash i} + \eta}{M_{\backslash i} + 2\eta} \frac{M_{c_it_i\backslash i} + \gamma}{M_{c_i\backslash i} + \gamma T}.$$

$$P(c_i = k|r_i = 0, \boldsymbol{W}, \boldsymbol{T}, \boldsymbol{Z}, \boldsymbol{C}_{\backslash i}, \boldsymbol{R}_{\backslash i}) \propto \frac{N_{kd}}{N_d},$$

$$P(c_i = k|r_i = 1, \boldsymbol{W}, \boldsymbol{T}, \boldsymbol{Z}, \boldsymbol{C}_{\backslash i}, \boldsymbol{R}_{\backslash i}) \propto \frac{M_{kt_i\backslash i} + \gamma}{M_{k\backslash i} + \gamma T} \frac{N_{kd}}{N_d}.$$

Experiments

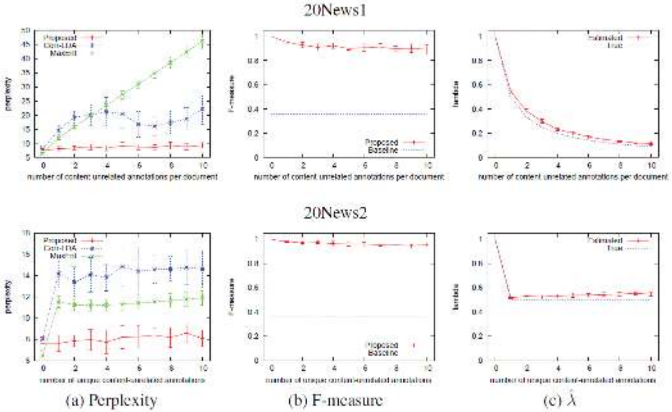
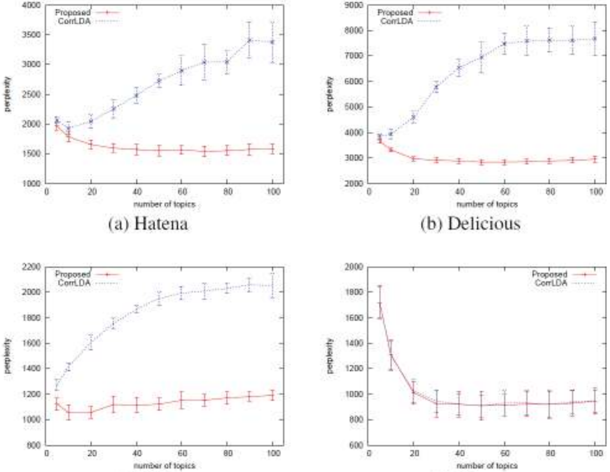


Figure 2: (a) Perplexities of the held-out content-related annotations, (b) F-measures of content relevance, and (c) Estimated content-related annotation ratios in 20News data.

Figure 2 (c) shows the content-related annotation ratios as estimated by the following equation, $\hat{\lambda} = \frac{M - M_0 + \eta}{M + 2\eta}$, with the proposed method. The estimated ratios are about the same as the true ratios.

Experiments



(a) Hatena (b) Delicious (c) Flickr (d) Patent

Figure 3: Perplexities of held-out annotations with different numbers of topics in social annotation data (a)(b)(c), and in data without content unrelated annotations (d).

Experiments

Table 2: The ten most probable content-unrelated annotations (leftmost column), and the ten most probable annotations for some topics (other columns), estimated with the proposed method using 50 topics. Each column represents one topic. The lower half in (a) and (b) shows probable words in the content.

(a) Hatena

unrelated	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9
toread	programming	game	economics	science	food	linux	politics	pc	medical
web	development	animation	finance	research	cooking	tips	international	apple	health
later	dev	movie	society	biology	gourmet	windows	oversea	iphone	lie
great	webdev	Nintendo	business	study	recipe	security	society	hardware	government
document	php	movie	economy	psychology	cook	server	history	gadget	agriculture
troll	java	event	reading	mathematics	life	network	china	mac	food
*	software	xbox360	investment	pseudoscience	fooditem	unix	world	cupidity	mentalhealth
?	ruby	DS	japan	knowledge	foods	mysql	international	technology	mental
summary	opensource	PS3	money	education	alcohol	mail	usa	ipod	environment
memo	softwaredev	animation	company	math	foodie	Apache	news	electronics	science
	development	game	year	science	eat	in	japan	yen	rice
	web	animation	article	researcher	use	setting	country	product	banana
	series	movie	finance	answer	omission	file	usa	digital	medical
	hp	story	economics	spirit	water	server	china	pc	diet
	technology	work	investment	question	decision	case	politics	support	hospital
	management	create	company	human	broil	mail	aso	in	poison
	source	PG	day	ehara	face	address	mr	note	eat
	usage	mr	management	proof	input	connection	korea	price	incident
	project	interesting	information	mind	miss	access	human	equipment	korea
	system	world	nikkei	brain	food	security	people	model	jelly

Experiments

(b) Delicious

reference	money	video	opensource	food	windows	art	shopping	iphone	education
web	finance	music	software	recipes	linux	photo	shop	mobile	learning
imported	economics	videos	programming	recipe	sysadmin	photography	Shopping	hardware	books
design	business	fun	development	cooking	Windows	photos	home	games	book
internet	economy	entertainment	linux	Food	security	Photography	wishlist	iPhone	language
online	Finance	funny	tools	Recipes	computer	Art	buy	apple	library
cool	financial	movies	rails	baking	microsoft	inspiration	store	tech	school
toread	investing	media	ruby	health	network	music	fashion	gaming	teaching
tools	bailout	Video	webdev	vegetarian	Linux	foto	gifts	mac	Education
blog	finances	film	rubyonrails	diy	ubuntu	fotografia	house	game	research
	money	music	project	recipe	windows	art	buy	iphone	book
	financial	video	code	food	system	photography	online	apple	legal
	credit	link	server	recipes	microsoft	photos	price	ipod	theory
	market	tv	ruby	make	linux	camera	cheap	mobile	books
	economic	movie	rails	wine	software	vol	product	game	law
	october	itunes	source	made	file	digital	order	games	university
	economy	film	file	add	server	images	free	pc	students
	banks	amazon	version	love	user	2008	products	phone	learning
	government	play	files	eat	files	photo	rating	mac	education
	bank	interview	development	good	ubuntu	tracks	card	touch	language

(c) Flickr

2008	dance	sea	autumn	rock	beach	family	island
nikon	bar	sunset	trees	house	travel	portrait	asia
canon	dc	sky	tree	party	vacation	cute	landscape
white	digital	clouds	mountain	park	camping	baby	rock
yellow	concert	mountains	fall	inn	landscape	boy	blue
red	bands	ocean	garden	coach	texas	kids	tour
photo	music	panorama	bortescristian	creature	lake	brown	plant
italy	washingtondc	south	geotagged	halloween	cameraphone	closeup	tourguidesoma
california	dancing	ireland	mud	mallory	md	08	koh
color	work	oregon	natura	night	sun	galveston	samui