Incorporating Domain Knowledge into Topic Modeling via Dirichlet Forest Priors

David Andrzejewski[#*] andrzeje@cs.wisc.edu
Xiaojin Zhu[#] jerryzhu@cs.wisc.edu
Mark Craven[*#] craven@biostat.wisc.edu

Department of Computer Sciences[#]
Department of Biostatistics and Medical Informatics[*]
University of Wisconsin-Madison
Madison, WI 53706 USA

Abstract

Users of topic modeling methods often have knowledge about the composition of words that should have high or low probability in various topics. We incorporate such domain knowledge using a novel Dirichlet forest prior in a Latent Dirichlet Allocation framework. The prior is a mixture of Dirichlet tree distributions with special structures. We present its construction, and inference via collapsed Gibbs sampling. Experiments on synthetic and real datasets demonstrate our model's ability to follow and generalize beyond user-specified domain knowledge.

Problem

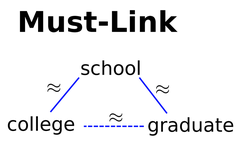
Unsupervised topic models learn a decomposition of text documents into mixtures of multinomials over words ("topics"). However, these topics may not align well with user modeling goals (Topic 13 below contains 2 distinct concepts). The goal of this work is to provide the user with a mechanism for expressing preferences about the topics.

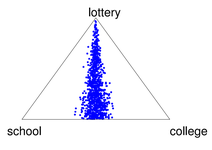
Topic 13	go school cancer into well free cure college . . . graduate . . . law . . . surgery recovery . . .
Topic 14	job go school great into good college . . . business graduate finish grades away law accepted . . .
Topic 15	mom husband cancer hope free son well . . . full recovery surgery pray heaven pain aids . . .

"Primitive" preferences

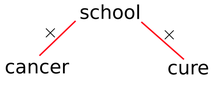
Operation	Meaning
Must-Link (school,college)	$\forall$ topics t, $P(school\|t) \approx P(college\|t)$
Cannot-Link (school,cure)	no topic t has $P(school\|t)$ and $P(cure\|t)$ both high

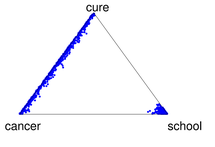
Must-Link





Cannot-Link





Cannot achieve with single Dirichlet !!!

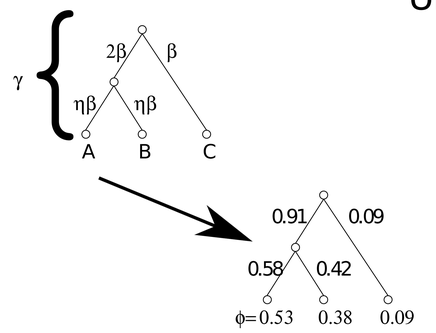
Composite Operations

split	[*go school into college*] vs [*cancer free cure well*] → Must-Link among words for each concept → Cannot-Link between words from different concepts
merge	[*love marry together boyfriend*] in one topic [*married boyfriend engaged wedding*] in another → Must-Link among concept words
isolate	[*the year in 2008*] in many wish topics → Must-Link among words to be isolated → Cannot-Link vs other Top N words for each topic

Must-Link via Dirichlet Tree

Each internal node in a Dirichlet Tree distributes probability mass to its children according to a Dirichlet distribution parameterized by the edge weights, allowing different subsets of variables to have different variances, which cannot be accomplished with a standard Dirichlet.

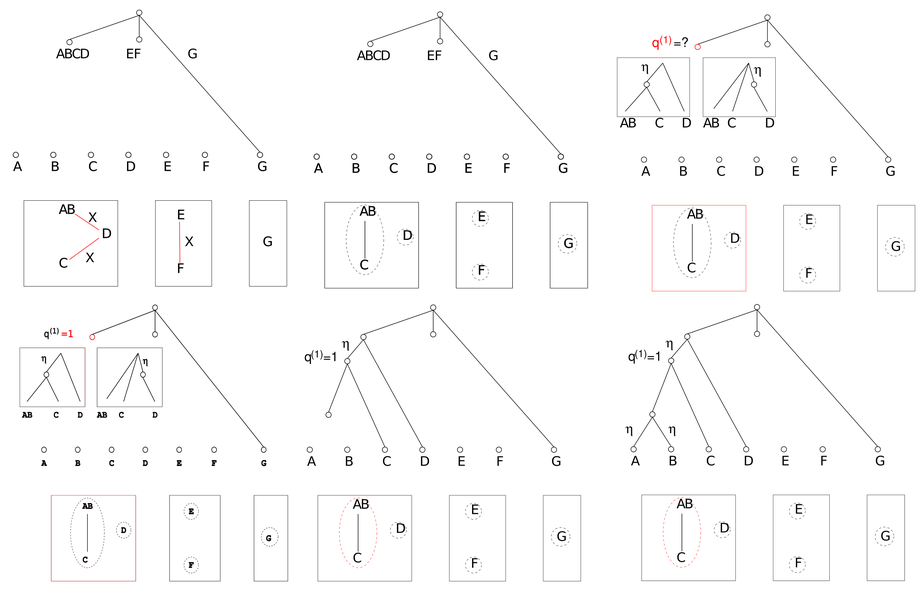
IDEA: Place Must-Linked words together under an internal node with large (eta) edge weights.



Cannot-Link via mixture of Dirichlet Trees (Dirichlet Forest)

Vocabulary	$[A, B, C, D, E, F, G]$
Must-Links	(A, B)
Cannot-Links	(A, D), (C, D), (E, F)

IDEA: Create "Cannot-Link" graph over words. For each connected component, maximal cliques of the subgraph complement correspond to maximal "compatible" word sets. For a single topic, randomly select one maximal clique for each connected component and construct a Dirichlet Tree with special structure that "favors" the chosen cliques. This procedure implicitly defines our Dirichlet Forest mixture.



...then repeat for other connected components, then repeat entire procedure for each topic.

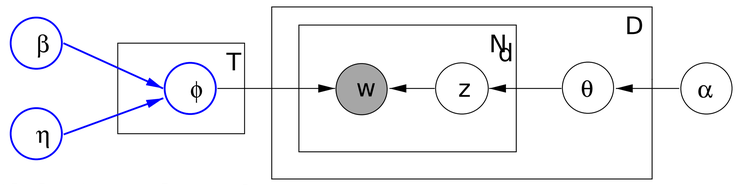
Collapsed Gibbs Sampling

The conjugacy of the Dirichlet Tree allows us to efficiently sample from the posterior of LDA with a Dirichlet Forest prior on topic-word multinomials..

$$p(z_i = v|\mathbf{z}_{-i}, \mathbf{q}_{1:T}, \mathbf{w}) \propto (n^{(d)}_{-i,v} + \alpha) \prod_s^{I_v(\uparrow i)} \frac{\gamma_v^{(C_v(s\downarrow i))} + n^{(C_v(s\downarrow i))}_{-i,v}}{\sum_k^{C_v(s)} \left(\gamma_v^{(k)} + n^{(k)}_{-i,v}\right)}$$

$$p(q_j^{(r)} = q'|\mathbf{z}, \mathbf{q}_{-j}, \mathbf{q}_j^{(-r)}, \mathbf{w}) \propto \left(\sum_k^{M_{rq'}} \beta_k\right) \times \prod_s^{I_{j,r=q'}} \left(\frac{\Gamma\left(\sum_k^{C_j(s)} \gamma_j^{(k)}\right)}{\Gamma\left(\sum_k^{C_j(s)} (\gamma_j^{(k)} + n_j^{(k)})\right)} \prod_k^{C_j(s)} \frac{\Gamma(\gamma_j^{(k)} + n_j^{(k)})}{\Gamma(\gamma_j^{(k)})} \right)$$

Modified Graphical Model



Wish Experiments

We use standard LDA to learn topics from a corpus of New Year's wishes submitted to the website of Times Square in New York City, treating each wish as an individual document. We then interactively encode preferences using the Dirichlet Forest Prior, then re-run inference in order to refine these topics.

1) Original topics

Topic	Top words sorted by $\phi = p(\text{word}\|\text{topic})$
0	love i you me and will forever that with hope
1	and health for happiness family good my friends
2	year new happy a this have and everyone years
3	that is it you we be t are as not s will can
4	my to get job a for school husband s that into
5	to more of be and no money stop live people
6	to our the home for of from end safe all come
7	to my be i find want with love life meet man
8	a and healthy my for happy to be have baby
9	a 2008 in for better be to great job president
10	i wish that would for could will my lose can
11	peace and for love all on world earth happiness
12	may god in all your the you s of bless 2008
13	the in to of world best win 2008 go lottery
14	me a com this please at you call 4 if 2 www

2) isolate([to,and,a...])

Topic	Top words sorted by $\phi = p(\text{word}\|\text{topic})$
0	love forever marry happy together mom back
1	health happiness good family friends prosperity
2	life best live happy long great time ever wonderful
3	out not up do as so what work don was like
MIXED	go school cancer into well free cure college
5	no people stop less day every each take children
6	home safe end troops iraq bring war husband house
7	love peace true happiness hope joy everyone dreams
8	happy healthy family baby safe prosperous everyone
9	better job hope president paul great ron than person
10	make money lose weight meet finally by lots hope
Isolate	and to for a the year in new all my 2008
12	god bless jesus loved know everyone love who loves
13	peace world earth win lottery around save
14	com call if 4 2 www u visit 1 3 email yahoo
Isolate	i to wish my for and a be that the in

3) split([college,...], [cure,...])

LOVE	love forever happy together marry fall
1	health happiness family good friends
2	life happy best live love long time
3	as not do so what like much don was
4	out make money house up work grow able
5	people no stop less day every each take
6	home safe end troops iraq bring war husband
7	love peace happiness true everyone joy
8	happy healthy family baby safe prosperous
9	better president hope paul ron than person
LOVE	lose meet man hope boyfriend weight finally
Isolate	and to for a the year in new all my 2008
12	god bless jesus loved everyone know loves
13	peace world earth win lottery around save
14	com call if 4 www 2 u visit 1 email yahoo 3
Isolate	i to wish my for and a be that the in me get
Split	job go school great into good college
Split	mom husband cancer hope free son well

4) merge([love,marry,...], [meet,married,...])

Topic	Top words sorted by $\phi = p(\text{word}\|\text{topic})$
Merge	love lose weight together forever marry meet
success	health happiness family good friends prosperity
life	life happy best live time long wishes ever years
-	as do not what someone so like don much he
money	out make money up house work able pay own lots
people	no people stop less day every each other another
iraq	home safe end troops iraq bring war return
joy	love true peace happiness dreams joy everyone
family	happy healthy family baby safe prosperous
vote	better hope president paul ron than person bush
Isolate	and to for a the year in new all my
god	god bless jesus everyone loved know heart christ
peace	peace world earth win lottery around save
spam	com call if u 4 www 2 u visit 1
Isolate	i to wish my for and a be that the
Split	job go great school into good college hope move
Split	mom hope cancer free husband son well dad cure