Nonparametric Bayesian Storyline Detection from Microtexts

Vinodh Krishnan and Jacob Eisenstein

Georgia Institute of Technology

Clustering microtexts into storylines

Strong start for Barcelona

Dog tuxedo bought with county credit card

Messi scores! Barcelona up 1-0

...

Yellow card for Messi

Clustering microtexts into storylines

$z = 1$	Strong start for **Barcelona**
$z = 2$	Dog tuxedo bought with county credit **card**
$z = 1$	**Messi** scores! **Barcelona** up 1-0
	...
$z = 1$	Yellow **card** for **Messi**

Clustering microtexts into storylines

$z = 1$	Strong start for **Barcelona**	Oct 1, 1:15pm
$z = 2$	Dog tuxedo bought with county credit **card**	Oct 1, 1:23pm
$z = 1$	**Messi** scores! **Barcelona** up 1-0	Oct 1, 1:39pm
	...	
$z = 3$	Yellow **card** for **Messi**	**Oct 8, 10:15am**

Clustering microtexts into storylines

$z = 1$	Strong start for **Barcelona**	Oct 1, 1:15pm
$z = 2$	Dog tuxedo bought with county credit **card**	Oct 1, 1:23pm
$z = 1$	**Messi** scores! **Barcelona** up 1-0	Oct 1, 1:39pm
	...	
$z = 3$	Yellow **card** for **Messi**	**Oct 8, 10:15am**

Storyline detection is a multimodal clustering problem, involving **content** and **time**.

About time

Prior approaches to modeling time

- Maximum temporal gap between items on same storyline
- Look for attention peaks (Marcus et al., 2011)
- Model temporal distribution per storyline (Ihler et al., 2006; Wang & McCallum, 2006)

About time

Prior approaches to modeling time

- Maximum temporal gap between items on same storyline
- Look for attention peaks (Marcus et al., 2011)
- Model temporal distribution per storyline (Ihler et al., 2006; Wang & McCallum, 2006)

Problems with these approaches:

- Storylines can have vastly different timescales, might be periodic, etc.
- Methods for determining number of storylines are typically ad hoc.

This work

A non-parametric Bayesian framework for storylines

- The number of storylines is a latent variable.
- No parametric assumptions about the temporal structure of storyline popularity.
- Text is modeled as a bag-of-words, but the modular framework admits arbitrary (centroid-based) models.
- Linear-time inference via streaming sampling

Modeling framework

Prior probability of storyline assignments, conditioned on timestamps

$$P(\boldsymbol{w}, \boldsymbol{z} \mid \boldsymbol{t}) = P(\boldsymbol{z} \mid \boldsymbol{t}) \prod_{k=1}^{K} P(\{\boldsymbol{w}_{i:z_i=k}\})$$

Likelihood of text, computed per storyline

The prior over storyline assignments

We want a prior distribution $P(\boldsymbol{z} \mid \boldsymbol{t})$ that is:

- nonparametric over the number of storylines;
- nonparametric over the storyline temporal distributions.

How to do it?

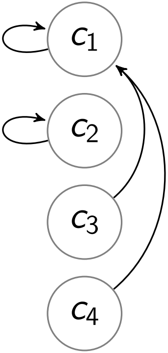
The prior over storyline assignments

We want a prior distribution $P(\boldsymbol{z} \mid \boldsymbol{t})$ that is:

- nonparametric over the number of storylines;
- nonparametric over the storyline temporal distributions.

How to do it? The distance-dependent Chinese restaurant process (Blei & Frazier, 2011)

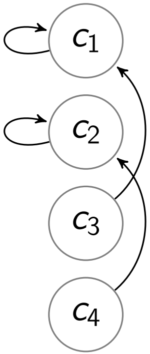
From graphs to clusterings



Key idea of dd-CRP: "follower" graphs define clusterings.

- $\mathcal{Z} = ((1, 3, 4), (2))$

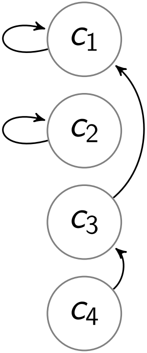
From graphs to clusterings



Key idea of dd-CRP: "follower" graphs define clusterings.

- $\mathcal{Z} = ((1, 3, 4), (2))$
- $\mathcal{Z} = ((1, 3), (2, 4))$

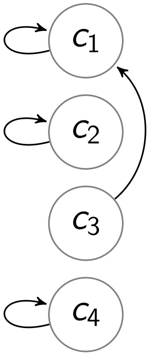
From graphs to clusterings



Key idea of dd-CRP: "follower" graphs define clusterings.

- $\mathcal{Z} = ((1, 3, 4), (2))$
- $\mathcal{Z} = ((1, 3), (2, 4))$
- $\mathcal{Z} = ((1, 3, 4), (2))$

From graphs to clusterings



Key idea of dd-CRP: “follower” graphs define clusterings.

- $\mathcal{Z} = ((1, 3, 4), (2))$
- $\mathcal{Z} = ((1, 3), (2, 4))$
- $\mathcal{Z} = ((1, 3, 4), (2))$
- $\mathcal{Z} = ((1, 3), (2), (4))$

Prior distribution

We reformulate the prior over follower graphs:

$$P(\boldsymbol{z} \mid \boldsymbol{t}) = P(\boldsymbol{c} \mid \boldsymbol{t}) = \prod_{i=1}^{N} P(c_i \mid t_i, t_{c_i})$$

$$P(c_i \mid t_i, t_{c_i}) = \begin{cases} e^{-|t_i - t_{c_i}|/a}, & c_i \neq i \\ \alpha, & c_i = i \end{cases}$$

- Probability of two documents being linked decreases exponentially with time gap $t_i - t_j$.
- The likelihood of a document linking to itself (starting a new cluster) is proportional to α.

Modeling framework

Prior probability of storyline assignments, conditioned on timestamps

$$P(\boldsymbol{w}, \boldsymbol{z} \mid \boldsymbol{t}) = P(\boldsymbol{z} \mid \boldsymbol{t}) \prod_{k=1}^{K} P(\{\boldsymbol{w}_{i:z_i=k}\})$$

Likelihood of text, computed per storyline

Likelihood

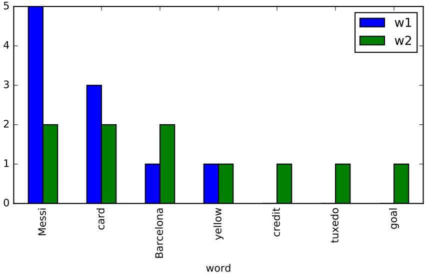
Cluster likelihoods are computed using the Dirichlet Compound Multinomial (Doyle & Elkan, 2009).

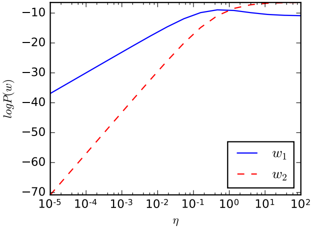
$$\begin{aligned} P(\boldsymbol{w}) &= \prod_{k=1}^{K} P(\{\boldsymbol{w}_i\}_{z_i=k}) \\ &= \prod_{k=1}^{K} \int_{\theta} P_{\text{MN}}(\{\boldsymbol{w}_i\}_{z_i=k} \mid \theta_k) P_{\text{Dir}}(\theta_k; \eta) d\theta_k \\ &= \prod_{k=1}^{K} P_{\text{DCM}}(\{\boldsymbol{w}_i\}_{z_i=k}; \eta), \end{aligned}$$

where η is a concentration hyperparameter.

The Dirichlet Compound Multinomial

The DCM is a distribution over vectors of counts, which rewards compact word distributions.





We set the hyperparameter η using a heuristic from Minka (2012).

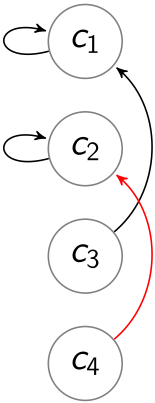
Modeling framework

Prior probability of storyline assignments, conditioned on timestamps

$$P(\boldsymbol{w}, \boldsymbol{z} \mid \boldsymbol{t}) = P(\boldsymbol{z} \mid \boldsymbol{t}) \prod_{k=1}^{K} P(\{\boldsymbol{w}_{i:z_i=k}\})$$

Likelihood of text, computed per storyline

Inference: Gibbs sampling

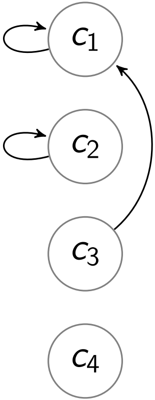


- We iteratively cut and resample each link.
- Each link is sampled from the joint probability,

$$\underset{\text{sample}}{\Pr}(c_i = j \mid \boldsymbol{c}_{-i}, \boldsymbol{w}) \propto \Pr(c_i = j) \times P(\boldsymbol{w} \mid \boldsymbol{c})$$

$$\propto \Pr(c_i = j) \times \frac{P(\{\boldsymbol{w}_k\}_{z_k = z_i \vee z_k = z_j})}{P(\{\boldsymbol{w}_k\}_{z_k = z_i}) \times P(\{\boldsymbol{w}_k\}_{z_k = z_j})}$$

Inference: Gibbs sampling

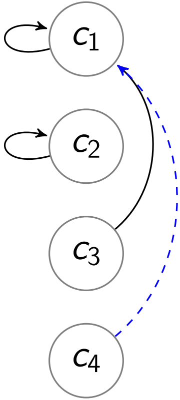


- We iteratively cut and resample each link.
- Each link is sampled from the joint probability,

$$\Pr_{\text{sample}}(c_i = j \mid \boldsymbol{c}_{-i}, \boldsymbol{w}) \propto \Pr(c_i = j) \times P(\boldsymbol{w} \mid \boldsymbol{c})$$

$$\propto \Pr(c_i = j) \times \frac{P(\{\boldsymbol{w}_k\}_{z_k = z_i \vee z_k = z_j})}{P(\{\boldsymbol{w}_k\}_{z_k = z_i}) \times P(\{\boldsymbol{w}_k\}_{z_k = z_j})}$$

Inference: Gibbs sampling

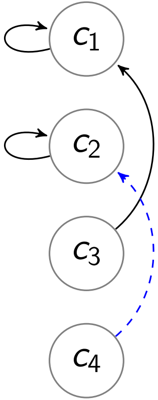


- We iteratively cut and resample each link.
- Each link is sampled from the joint probability,

$$\Pr_{\text{sample}}(c_i = j \mid \boldsymbol{c}_{-i}, \boldsymbol{w}) \propto \Pr(c_i = j) \times P(\boldsymbol{w} \mid \boldsymbol{c})$$

$$\propto e^{-\frac{t_4 - t_1}{a}} \times \frac{P(\{\boldsymbol{w}_1, \boldsymbol{w}_3, \boldsymbol{w}_4)\}}{P(\{\boldsymbol{w}_4\}) \times P(\{\boldsymbol{w}_1, \boldsymbol{w}_3\})}$$

Inference: Gibbs sampling

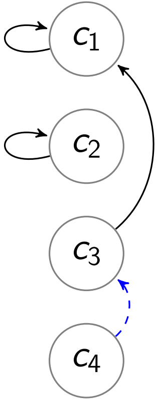


- We iteratively cut and resample each link.
- Each link is sampled from the joint probability,

$$\Pr_{\text{sample}}(c_i = j \mid \boldsymbol{c}_{-i}, \boldsymbol{w}) \propto \Pr(c_i = j) \times P(\boldsymbol{w} \mid \boldsymbol{c})$$

$$\propto e^{-\frac{t_4 - t_2}{a}} \times \frac{P(\{\boldsymbol{w}_2, \boldsymbol{w}_4)\}}{P(\{\boldsymbol{w}_4\}) \times P(\{\boldsymbol{w}_2\})}$$

Inference: Gibbs sampling

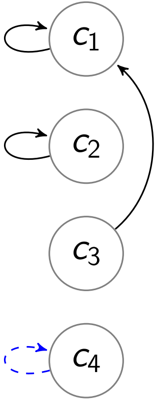


- We iteratively cut and resample each link.
- Each link is sampled from the joint probability,

$$\Pr_{\text{sample}}(c_i = j \mid \boldsymbol{c}_{-i}, \boldsymbol{w}) \propto \Pr(c_i = j) \times P(\boldsymbol{w} \mid \boldsymbol{c})$$

$$\propto e^{-\frac{t_4 - t_3}{a}} \times \frac{P(\{\boldsymbol{w}_1, \boldsymbol{w}_3, \boldsymbol{w}_4)\}}{P(\{\boldsymbol{w}_4\}) \times P(\{\boldsymbol{w}_1, \boldsymbol{w}_3\})}$$

Inference: Gibbs sampling

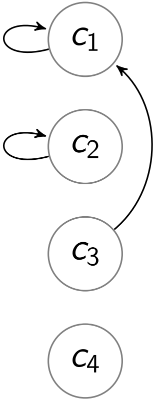


- We iteratively cut and resample each link.
- Each link is sampled from the joint probability,

$$\Pr_{\text{sample}}(c_i = j \mid \boldsymbol{c}_{-i}, \boldsymbol{w}) \propto \Pr(c_i = j) \times P(\boldsymbol{w} \mid \boldsymbol{c})$$

$$\propto \alpha \times \frac{P(\{\boldsymbol{w}_4)\}}{P(\{\boldsymbol{w}_4\})}$$

Inference: Gibbs sampling



- We iteratively cut and resample each link.
- Each link is sampled from the joint probability,

$$\Pr_{\text{sample}}(c_i = j \mid \boldsymbol{c}_{-i}, \boldsymbol{w}) \propto \Pr(c_i = j) \times P(\boldsymbol{w} \mid \boldsymbol{c})$$

$$\propto \Pr(c_i = j) \times \frac{P(\{\boldsymbol{w}_k\}_{z_k = z_i \vee z_k = z_j})}{P(\{\boldsymbol{w}_k\}_{z_k = z_i}) \times P(\{\boldsymbol{w}_k\}_{z_k = z_j})}$$

- Online inference: Gibbs sampling restricted to a moving window (linear-time)

TREC 2014 TTG Results

Model	F_1	F_1^w
dd-CRP clustering models		
1. BASELINE	0.20	0.30
2. OFFLINE	0.29	0.34
3. ONLINE	0.29	0.35

TREC 2014 TTG Results

Model	F_1	F_1^w
dd-CRP clustering models		
1. BASELINE	0.20	0.30
2. OFFLINE	0.29	0.34
3. ONLINE	0.29	0.35
Top systems from Trec-2014 TTG		
4. `TTGPKUICST2` (Lv et al., 2014)	0.35	0.46
5. `EM50` (Magdy et al., 2014)	0.25	0.38
6. `hltcoeTTG1` (Xu et al., 2014)	0.28	0.37

TREC 2014 TTG Results

Model	F_1	F_1^w
dd-CRP clustering models		
1. BASELINE	0.20	0.30
2. OFFLINE	0.29	0.34
3. ONLINE	0.29	0.35
Top systems from Trec-2014 TTG		
4. `TTGPKUICST2` (Lv et al., 2014)	0.35	0.46
5. `EM50` (Magdy et al., 2014)	0.25	0.38
6. `hltcoeTTG1` (Xu et al., 2014)	0.28	0.37

- Online inference as accurate as offline Gibbs
- 2nd of 14 TREC systems on F_1, 4th/14 on F_1^w
- We use the baseline retrieval model, 0.31 MAP vs 0.5-0.6 MAP for best systems.

Summary

- Nonparametric Bayesian storyline detection incorporating content and time.

  Content Centroid-based likelihood (Dirichlet Compound Multinomial)

  Time Distance-based prior (ddCRP)

  Fancier likelihoods and distance functions can be incorporated in future work!

- Our nonparametric model is competitive with TREC TTG systems, despite using a much weaker retrieval model.

Acknowledgments

- National Institutes for Health (R01GM112697-01)
- A Focused Research Award for computational journalism from Google
- CNewsStory 2016 reviewers
- Patrick Violette and Irfan Essa

References I

Blei, D. M. & Frazier, P. I. (2011). Distance dependent chinese restaurant processes. *Journal of Machine Learning Research*, *12*(Aug), 2461–2488.

Doyle, G. & Elkan, C. (2009). Accounting for burstiness in topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, (pp. 281–288). ACM.

Ihler, A., Hutchins, J., & Smyth, P. (2006). Adaptive event detection with time-varying poisson processes. In *KDD*, (pp. 207–216). ACM.

Lv, C., Fan, F., Qiang, R., Fei, Y., & Yang, J. (2014). PKUICST at TREC 2014 Microblog Track: feature extraction for effective microblog search and adaptive clustering algorithms for TTG. Technical report, DTIC Document.

Magdy, W., Gao, W., Elganainy, T., & Wei, Z. (2014). Qcri at trec 2014: applying the kiss principle for the ttg task in the microblog track. Technical report, DTIC Document.

Marcus, A., Bernstein, M. S., Badar, O., Karger, D. R., Madden, S., & Miller, R. C. (2011). Twitinfo: aggregating and visualizing microblogs for event exploration. In *chi*, (pp. 227–236). ACM.

Minka, T. (2012). Estimating a dirichlet distribution. `http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/minka-dirichlet.pdf`.

Wang, X. & McCallum, A. (2006). Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, (pp. 424–433). ACM.

Xu, T., McNamee, P., & Oard, D. W. (2014). Hltcoe at trec 2014: Microblog and clinical decision support.