



# Using browsing history to identify Internet users' interests

Nikolay Anokhin  
data scientist



# Advertisement on the Web

The screenshot shows the Mail.Ru website interface. At the top, there is a search bar with the text "Поиск в интернете" and a "Найти" button. Below the search bar, there are navigation tabs for "Новости", "Москва", "Спорт", "Авто", and "Афиша". The "Спорт" tab is currently selected. The main content area displays a news article about a former Ukrainian governor, Pavlo Gubarenko, who has been hospitalized in a serious condition. The article text includes: "Губарев после покушения госпитализирован в тяжелом состоянии", "Бывший «народный губернатор» Донецкой области и один из лидеров ДНР и Новороссии Павел Губарев находится без", "МВФ прогнозирует отмену антироссийских санкций", "Гелетей отчитался о выполнении задач Порошенко", "Россия ответит на санкции ЕС созданием новой госкорпорации", "Премьер Австралии хочет «жестко поговорить» с Путиным", "Инфляция: причины роста и прогнозы", "Нобелевская премия по экономике присуждена Жану Тиролю", "Спорт. Российский боксер Поветкин исключен из рейтинга WBA", "Авто. Представлен российский квадроцикл «Гепард»", "Работа. Карьера или семья?". To the right of the article is a large yellow banner advertisement for "Cuddles" featuring a cartoon rabbit character and the text "300x300 banner ad". Below the article, there are sections for "Москва" (Moscow) with a weather forecast, "Курсы валют" (Exchange rates), "Гороскопы" (Horoscopes), "ТВ программа" (TV program), "Работа" (Jobs), and "Почта для бизнеса" (Business mail). On the left side of the page, there is a sidebar with various services: "Почта" (Mail), "Агент Mail.Ru" (Agent), "Мой Мир" (My World), "Одноклассники" (Odnoklassniki), "ICQ" (Instant messaging), "Деньги" (Money), and "Товары" (Goods).

# It's all about users (and money)

- ✗ clothing
- ✗ travelling
- ✗ cars
- ✗ dating



- ✓ computers
- ✓ gadgets
- ✓ photography
- ✓ data mining

User ID	Timestamp	URL	Etc.
A1B2C3D4	2014-07-01 13:11:37	http://auto.mail.ru/toyota	M/27/...
A1B2C3D4	2014-07-01 13:20:45	http://example.com?id=football	M/27/...
A1B2C3D4	2014-07-02 00:25:10	http://somesite.com/index.php	M/27/...
...			
F9E8D7C6	2014-06-30 18:01:12	http://my-little-pony.com/	F/19/...
F9E8D7C6	2014-06-30 18:10:51	http://afisha.mail.ru/twilight	F/19/...

**Text log files – about 300 G/day (and growing)**

## Some immediate conclusions

User ID	Timestamp	URL	Etc.
A1B2C3D4	2014-07-01 13:11:37	http:// <a href="#">auto.mail.ru</a> / <a href="#">toyota</a>	M/27/...
A1B2C3D4	2014-07-01 13:20:45	http://example.com?id= <a href="#">football</a>	M/27/...
A1B2C3D4	2014-07-02 00:25:10	http:// <a href="#">somesite.com</a> /index.php	M/27/...



A1B2C3D4: auto, toyota, football, somesite

- ▶ Let there be  $M$  users, each user  $u$  is represented by a bag of  $N_u$  tokens
- ▶ Let the number of *topics* (user interests) be given and equal to  $K$

## Generative model

- I For each topic draw a topic distribution  $\beta_k \sim \text{Dir}(\eta_k)$ ,  $k \in 1, \dots, K$
- II For each user  $u \in 1, \dots, M$ :
  - 1 Draw the user's topic distribution  $\theta_u \sim \text{Dir}(\alpha)$
  - 2 For each potential token  $t \in 1, \dots, N_u$ :
    - 2.1 Choose the token's topic assignment  $z_{u,t} \sim \text{Multl}(\theta_u)$
    - 2.2 Choose the token  $w_{u,t} \sim \text{Mult}(\beta_{z_{u,t}})$

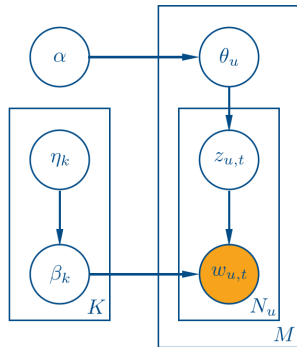
---

<sup>1</sup>Latent Dirichlet Allocation // Blei et. al.

$$\begin{aligned} p(\mathbf{w}, \theta, \beta, \mathbf{z} | \alpha, \eta) &= \\ &= p(\theta | \alpha) \prod_{t=1}^N p(z_t | \theta) p(w_t | z_t, \beta) p(\beta | \eta) \end{aligned}$$

Posterior of hidden variables

$$p(\theta, \beta, \mathbf{z} | \mathbf{w}, \alpha, \eta) = \frac{p(\theta, \beta, \mathbf{z}, \mathbf{w} | \alpha, \eta)}{p(\mathbf{w} | \alpha, \eta)}$$



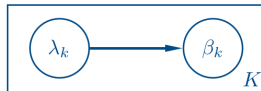
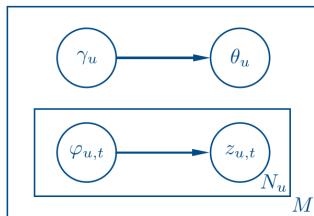
$$q(\theta, \beta, \mathbf{z}) = \prod_{k=1}^K \text{Dir}(\beta_k | \lambda_k) \times \prod_{u=1}^M \text{Dir}(\theta_u | \gamma_u) \prod_{t=1}^N \text{Mult}(z_{u,t} | \varphi_{u,t})$$

Maximizing the ELBO...

$$\mathcal{L} = E_q [\log(p(\mathbf{w}, \theta, \beta, \mathbf{z}))] - E_q [\log q(\theta, \beta, \mathbf{z})]$$

...is the same as minimising KL-divergence

$$KL(q||p) = E_q \left[ \log \frac{q(\theta, \beta, \mathbf{z})}{p(\theta, \beta, \mathbf{z} | \mathbf{w})} \right]$$





**E1** For each user, given  $\alpha$  and  $\lambda$ , update  $\varphi$  and  $\gamma$

$$\varphi_{t,k} \propto E_q[\beta_{t,k}] \exp(\Psi(\gamma_k))$$

$$\gamma_k = \alpha_k + \sum_{w=1}^N \varphi_{t,k}$$

**E2** Update  $\lambda$  for each topic, using the obtained  $\varphi$

$$\lambda_{t,k} = \eta_{t,k} + \sum_{u=1}^M w_t^{(u)} \varphi_{t,k}^{(u)}$$

**M** Maximise lower bound of the data log likelihood w.r.t. to  $\alpha$  using Newton-Raphson method

---

<sup>2</sup>Mr. LDA: A Flexible Large Scale Topic Modeling Package using Variational Inference in MapReduce // Zhai et. al.

```
1 function configure()
2     load alpha, lambda and gamma from distributed cache
3     normalize lambda for every topic
4
5 function map(u, tokens)
6     initialize a zero V x K-dimensional matrix Phi
7     initialize a zero K-dimensional row vector sigma
8     read user logs as tokens w[1], w[2], . . . , w[N]
9     repeat
10         for all t in 1..V do
11             for all k 1..K do
12                 Update  $\text{Phi}[t,k] = \lambda[t,k] / (\sum_t \lambda[t,k]) * \exp(\text{Psi}(\text{gamma}[u,k]))$ 
13                 normalize phi[t,*]
14                 sigma = sigma + w[t] * phi[t,*]
15             Update row vector gamma[u,*] = alpha + sigma
16         until convergence
17     for all k in 1..K do
18         for all t in 1..V do
19             emit <k, t> : w[t] * phi[t,k]
20     emit <k, u> : gamma[u,k]
```

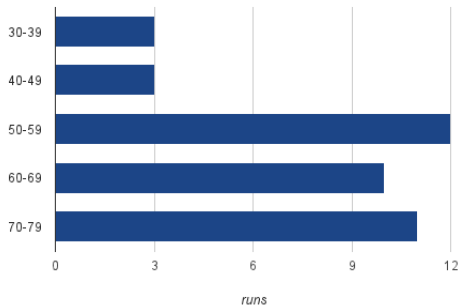
```
1 function map(<p_left, p_right>, Sigmas)
2     # sigma is unnormalized lambda
3     S = sum(Sigmas)
4     emit <p_left, p_right>: S
```

## Typical machine config

processors	2 x Intel(R) Xeon(R) 2.00GHz
cores	12
threads	24
RAM	32 GB
HDD	4-8 TB

**30 machines in cluster**

## Convergence iterations



Typical data: 10-days user logs  
Typical run time: 20 hours

topic1	topic2	topic3	topic4	topic5	topic6
book	klass	mobile	avito	krasnoyarsk	china
books	reshebnik	svyaznoy	kvartiry	tyumen	meta
loveread	class	phone	doma	tomsk	shared
knigi	megabotan	telefony	prodam	kemerovo	links
read	resh	nokia	dachi	surgut	maincat
author	slovo	phones	kottedzhi	barnaul	linkwall
litmir	algebra	iphone	nedvizhimost	nizhnevartovsk	nakanune
labirint	yazyk	samsung	sdam	krsk	razvezlo
authors	reshebniiki	catalog	oblast	novokuznetsk	poster
tululu	otbet	allnokia	komnaty	kurgan	readme

- ▶ LDA is an appropriate model for Internet user's interests
  - ▶ Variational EM is an efficient algorithm for LDA parameter estimation
  - ▶ Variational EM is easy to parallelise using MapReduce paradigm
- 
- ▶ Profile prediction for a new user
  - ▶ Topics as features in data mining tasks

# Q&A

Nikolay Anokhin

n.anokhin@corp.mail.ru