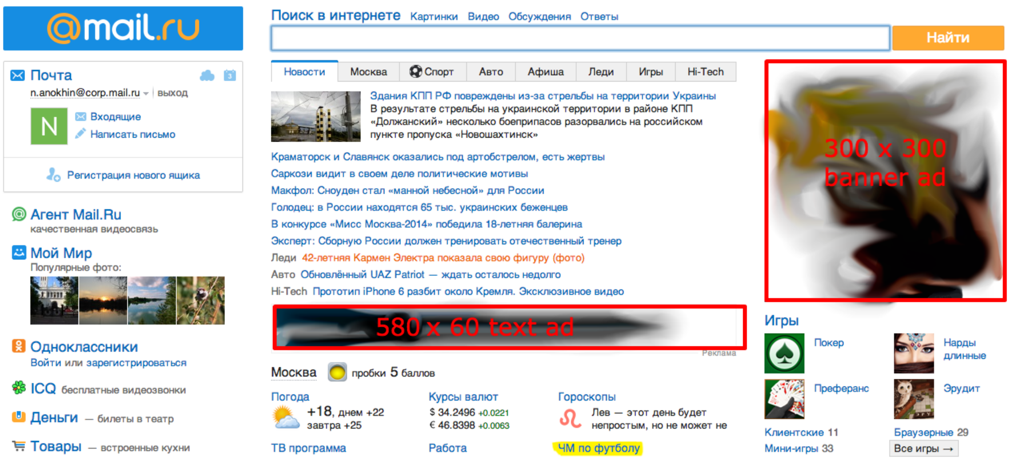
MapReduce programming model for Big Data analysis

Nikolay Anokhin

Advertisement on the Web

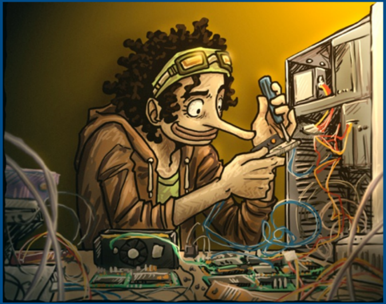


It's all about users (and money)[1]

cars

computers

tourism



online games

shopping

cameras

restaurants

bicycles

[1] Image source: Deviantart

The Data: user access logs

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.ry/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://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/...

$$\downarrow$$

A1B2C3D4: auto, toyota, football, somesite

Multinomial distribution

Let $\theta = (\theta_1, \dots, \theta_k)$ be the probability mass function (PMF) for a set of k events, i.e.

$$\forall i = 1, \dots, k : \ \theta_i \geqslant 0 \quad \text{and} \quad \sum_{i=1}^{k} \theta_i = 1$$

Binomial distribution $(k = 2,\ \theta_1 = q,\ \theta_2 = 1 - q)$

$$p(x|n, q) = \frac{n!}{x!(n-x)!} q^x (1-q)^{n-x}$$

Multinomial distribution

$$p(x_1, \dots, x_k | n, \theta_1, \dots, \theta_k) = \frac{n!}{x_1! \dots x_k!} \prod_{i=1}^{k} \theta_i^{x_i}$$

Dirichlet distribution

Let

1. $\Theta = (\Theta_1, \dots, \Theta_k)$ be a random PMF, i.e. $\forall i: \ \Theta_i \geqslant 0$ and $\sum_{i=1}^{k} \Theta_i = 1$
2. $\alpha = (\alpha_1, \dots, \alpha_k)$ be a vector, s.t. $\forall i: \ \alpha_i > 0$ and $\alpha_0 = \sum_{i=1}^{k} \alpha_i$

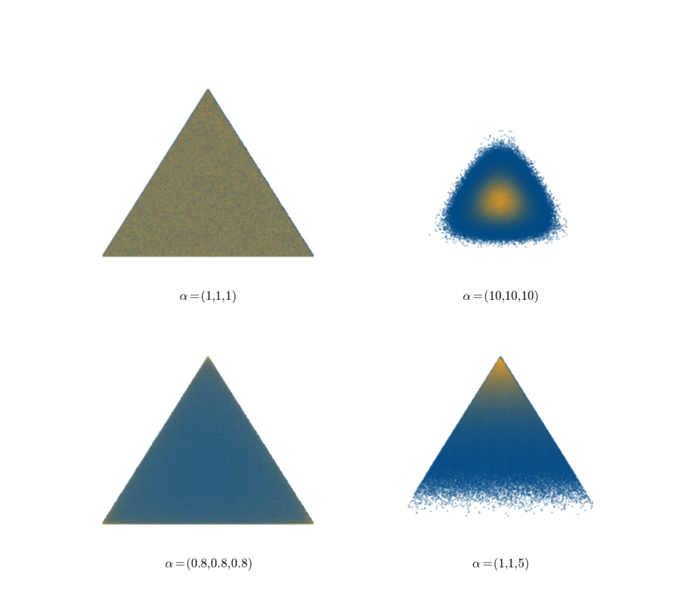
Then Θ is said to have *Dirichlet distribution* with parameter α, iff

$$p(\theta_1, \dots, \theta_k | \alpha_1, \dots, \alpha_k) = \begin{cases} \frac{\Gamma(\alpha_0)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1} & \text{if } \theta - \text{PMF} \\ 0 & \text{otherwise} \end{cases}$$

where

$$\forall s > 0 : \Gamma(s+1) = s\Gamma(s)$$

Dirichlet distribution



Latent Dirichlet Allocation[2]

- 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 $\beta_k \sim \mathrm{Dir}(\eta_k),\ k \in 1, \dots K$

II For each user $u \in 1, \dots, M$:

1. Draw the user's topic distribution $\theta_u \sim \mathrm{Dir}(\alpha)$
2. For each potential token $t \in 1, \dots, N_u$:

   2.1 Choose the token's topic assignment $z_{u,t} \sim \mathrm{Multl}(\theta_u)$

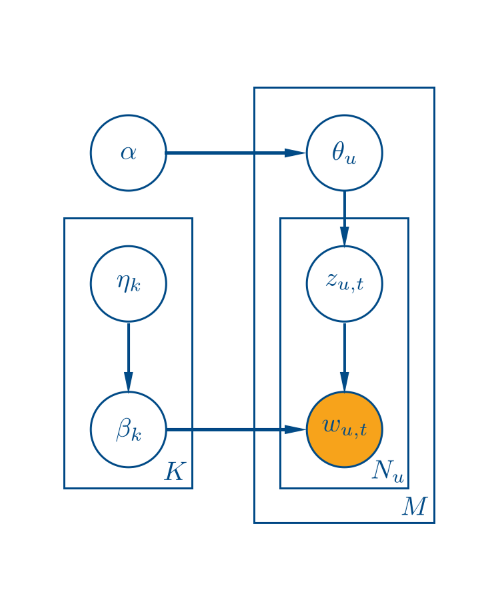
   2.2 Choose the token $w_{u,t} \sim \mathrm{Mult}(\beta_{z_{u,t}})$

[2] Latent Dirichlet Allocation // Blei et. al.

Generative model

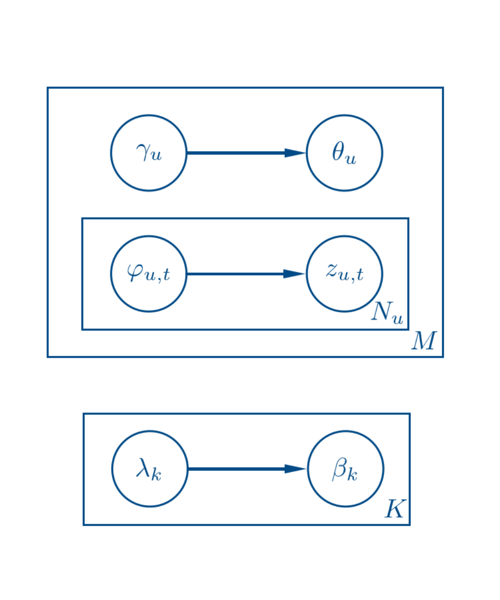
$$p(\mathbf{w}, \theta, \beta, \mathbf{z} | \alpha, \eta) =$$
$$= p(\theta|\alpha) \prod_{t=1}^{N} p(z_t|\theta) p(\beta|\eta) p(w_t|z_t, \beta)$$

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



Variational inference

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



Maximizing the ELBO...

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

...is the same as minimising KL-divergence

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

Variational EM

E1 For each user, given α and λ, update φ and γ

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

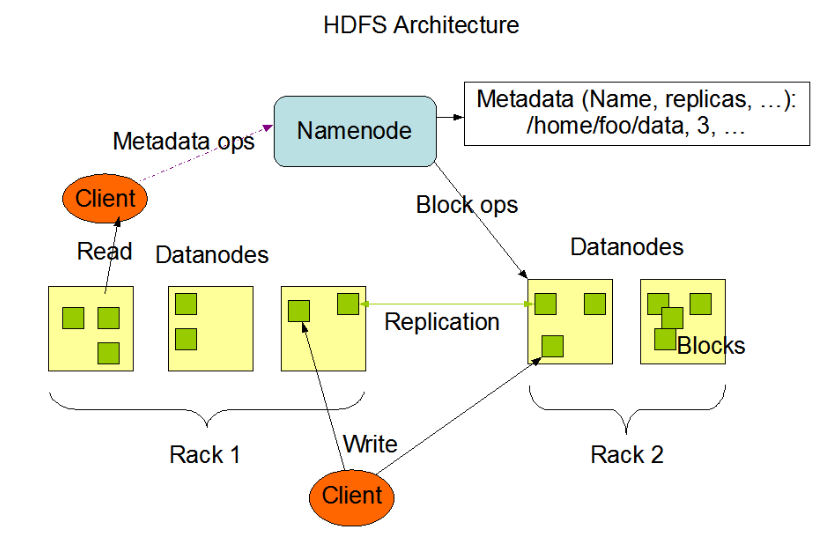
$$\gamma_k = \alpha_k + \sum_{w=1}^{N} \varphi_{t,k}$$

E2 Update λ for each topic, using the obtained φ

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

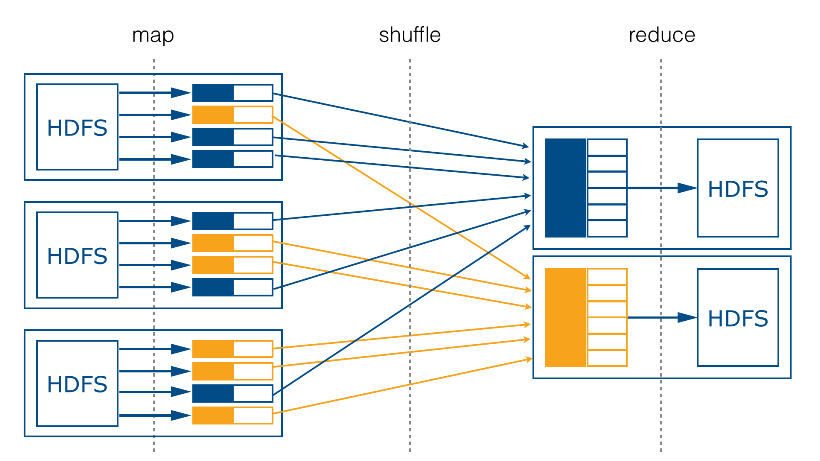
M Maximise lower bound of the data log likelihood w.r.t. to α using Newton-Raphson method

Storing the data — HDFS[3]



[3] Image source: HDFS architecture guide

Processing the data — Hadoop MapReduce[4]




```
map( Key1, Value1 ):  List[( Key2, Value2 )]
reduce( Key2, List[Value2] ):  List[( Key3, Value3 )]
```

[4] MapReduce: Simplified Data Processing on Large Clusters // Jeffrey Dean, Sanjay Ghemawat

LDA – map[5]

Input:

KEY – user ID $u \in [1, M]$

VALUE – user tokens

Configure

Load in α, λ and γ from distributed cache

Normalize λ for every topic

Map

Initialize a zero $V \times K$-dimensional matrix φ

Initialize a zero K-dimensional row vector σ

Read in user logs $\| t_1, t_2, \ldots, w_N \|$

repeat

  for all $t \in [1, N]$ do

    for all $k \in [1, K]$ do

      Update $\varphi_{t,k} = \frac{\lambda_{t,k}}{\sum_t \lambda_{t,k}} \exp\left(\Psi(\gamma_k)\right)$

    end for

    Normalize φ_t, set $\sigma = \sigma + w_t \varphi_{t,*}$

  end for

  Update row vector $\gamma_{u,*} = \alpha + \sigma$

until convergence

for all $k \in [1, K]$ do

  for all $t \in [1, N]$ do

    Emit $< k, t > : w_t \varphi_{t,k}$

  end for

Emit $< k, u > : \gamma_{u,k}$ to file

end for

[5] Mr. LDA: A Flexible Large Scale Topic Modeling Package using Variational Inference in MapReduce // Zhai et. al.

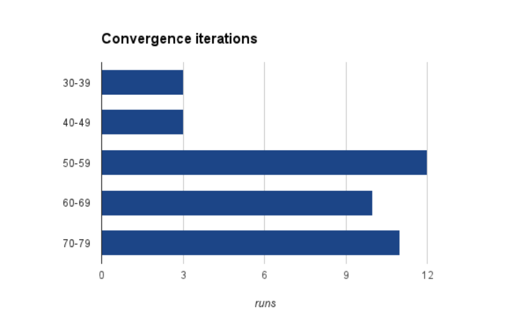
LDA – reduce

Input:
KEY - key pair $< p_{left}, p_{right} >$
VALUE - an iterator $\mathcal{I}$ over sequence of values
Reduce
1: Compute the sum σ over all values in the sequence $\mathcal{I}$ (σ is unnormalized λ)
2: Emit $< p_{left}, p_{right} > : \sigma$

Running LDA

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	

Typical data: 10-days user logs
Typical run time: 6 hours



Modelling results

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	reshebniki	catalog	oblast	novokuznetsk	poster
tululu	otbet	allnokia	komnaty	kurgan	readme

Conclusions and Future Work

- 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