Parallel Algorithms for Mining Large-Scale Data

Edward Chang

Director, Google Research, Beijing

http://infolab.stanford.edu/~echang/

Story of the Photos

- The photos on the first pages are statues of two "bookkeepers" displayed at the British Museum. One bookkeeper keeps a list of good people, and the other a list of bad. (Who is who, can you tell ? ☺)
- When I first visited the museum in 1998, I did not take a photo of them to conserve films. During this trip (June 2009), capacity is no longer a concern or constraint. In fact, one can see kids, grandmas all taking photos, a lot of photos. Ten years apart, data volume explodes.
- Data complexity also grows.
- So, can these ancient "bookkeepers" still classify good from bad? Is their capacity scalable to the data dimension and data quantity ?

Outline

- Motivating Applications
  - Q&A System
  - Social Ads
- Key Subroutines
  - Frequent Itemset Mining [ACM RS 08]
  - Latent Dirichlet Allocation [WWW 09, AAIM 09]
  - Clustering [ECML 08]
  - UserRank [Google TR 09]
  - Support Vector Machines [NIPS 07]
- Distributed Computing Perspectives

Query: *What are must-see attractions at Yellowstone*

What are must-see attractions at Yellowstone - Google Search - Mozilla Firefox

Who is the First Emperor... | Google.com Mail - Inbox ... | Program (RUGS'08) | What are must-see at... | Seeing the opportunity f...

Google

What are must-see attractions at Yellowstone [Search] Advanced Search Preferences

Web Results **1** - **10** of about **12,000** for **What are must-see attractions at Yellowstone**. (**0.18** seconds)

[Three **Must See Attractions at Yellowstone** National Park « The View ...]
Jan 15, 2008 ... Smith presents Three **Must See Attractions at Yellowstone** National Park posted at The View West. Interested in **Yellowstone** National Park? ...
theviewwest.com/2008/01/15/three-**must-see-attractions-at-yellowstone**-national-park/ - 26k - Cached - Similar pages

[Three **Must See Attractions At Yellowstone** National Park]
Jan 15, 2008 ... Three **Must See Attractions At Yellowstone** National Park.
ezinearticles.com/?Three-**Must-See-Attractions-At-Yellowstone**-National-Park&id=929265 - 47k - Cached - Similar pages

[**Yellowstone** National Park: Top Ten **Attractions**]
YELLOWSTONE NATIONAL PARK by **Yellowstone** Net. Top 10 Things to See in YNP What are the "**Must See**" **attractions** to view in **Yellowstone**? Start here! ...
www.**yellowstone**.net/topten.htm - 16k - Cached - Similar pages

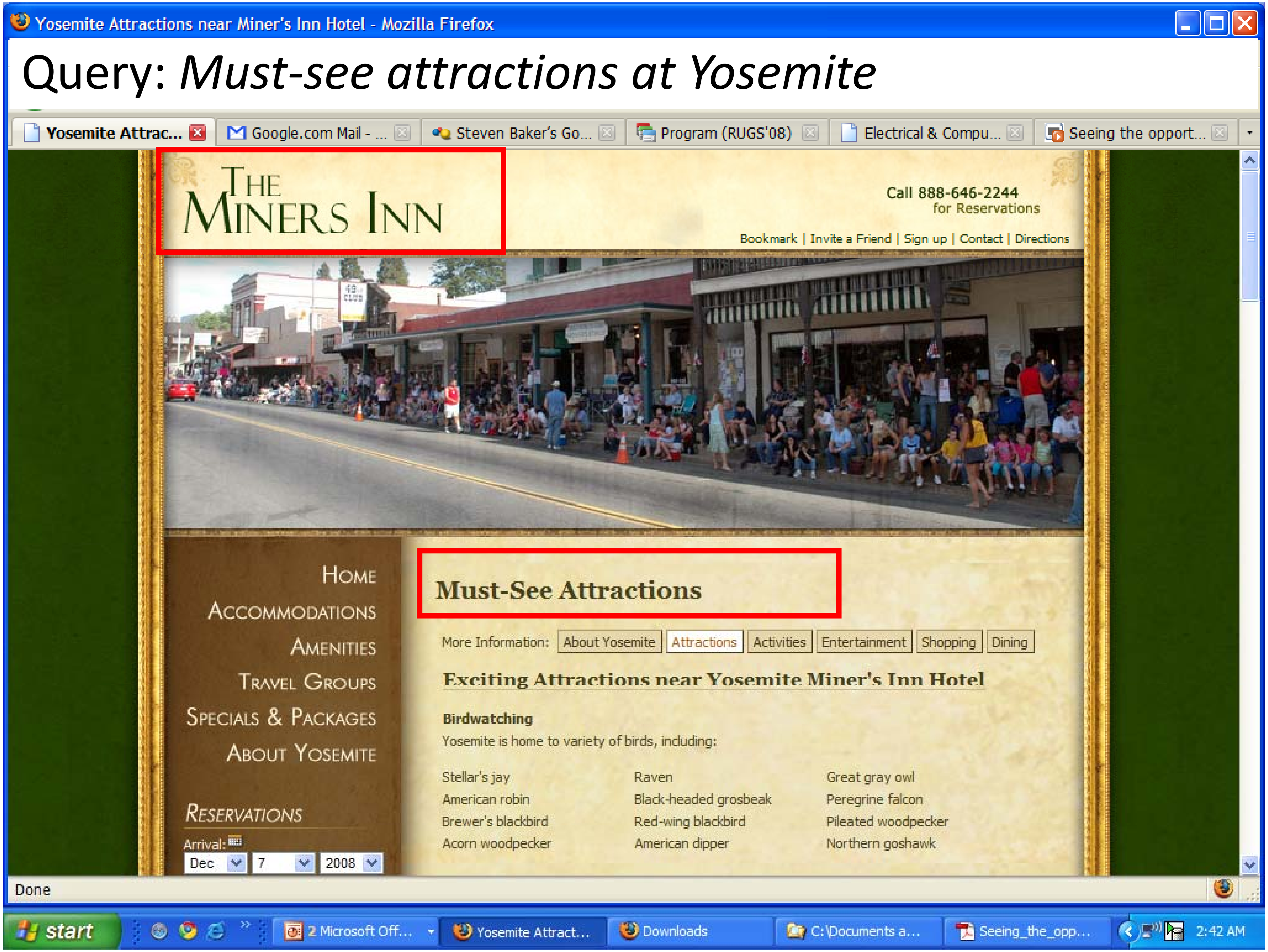
[**Yellowstone Must-see Attractions**]
Yellowstone's Must-See Attractions. The locations of all sites listed below are shown on the map that you receive as you enter the park. ...
www.geocities.com/dmonteit/**must_see**.html - 8k - Cached - Similar pages

> [What to See in **Yellowstone**]
> **Must-See Attractions** -- Text Only Version · Upper Geyser Basin and Old Faithful · Grand Canyon of the **Yellowstone** · Fountain Paint Pots Trail · Wildlife ...
> www.geocities.com/dmonteit/whattosee.html - 10k - Cached - Similar pages
> More results from www.geocities.com »

[**Must See** in **Yellowstone** National Park]

Query: *What are must-see attractions at Yellowstone*

Query: *Must-see attractions at Yosemite*



Copenhagen Things To Do - Copenhagen Attractions - VirtualTourist.com - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Query: *Must-see attractions at Copenhagen*

Copenhagen Things To Do - Cope... | Corp Gmail - MOMA Single Sign On | Gmail: Email from Google

Become a Virtual Tourist Member Today! Sign Up for Free | Sign In

virtual Tourist
Real Travelers - Real Info

Copenhagen Things To Do

Search: Destinations GO!

email to friend | help

Home | Travel Guides | Book Travel | Meet Members | Travel Deals | Trip Planner | Forums

Home » Travel Guides » Europe » Denmark » Københavns Kommune » Copenhagen » Things To Do

Copenhagen Things To Do

Copenhagen Hotels
Real reviews from real travelers.

by Mariajoy

Reviews and photos of Copenhagen attractions posted by real travelers and locals. The best tips for Copenhagen sightseeing.

Copenhagen Map

- 1,115 Members Living in Copenhagen
- 7,311 Copenhagen Photos
- 14 Copenhagen Videos
- 4,114 Copenhagen Tips

Search Things To Do

Search

Copenhagen | Københavns Kommune | Denmark

Copenhagen Things To Do 1 - 25 of 64

Copenhagen Tips:

Copenhagen Overview
Copenhagen Hotels
Copenhagen Flights
Copenhagen Things To Do
Copenhagen Nightlife
Copenhagen Transportation
Copenhagen Restaurants

Done

start | 2 Windows E... | M Hotel Singap... | 2 Microsoft O... | 2 Firefox | 86% | 10:23 PM

Query: *Must-see attractions at Beijing*


谁是姚明 - Google 搜索 - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.google.cn/search?hl=zh-CN&q=谁是姚明&btnG=Google+搜索&meta=&aq=f&oq
Most Visited Helpdesk Presentations (Fensi ...
谁是姚明 - Google 搜索
GOOG: 390.43 +1.69 (0.43%) - ...
Google.com - Calendar
谁是姚明
Who is Yao Ming
所有网页
高级搜索 | 使用偏好
网页
约有 7,730,000 项符合 谁是姚明 的查询结果，以下是第1-10 项 （搜索用时 0.27 秒）
谁是姚明的相关焦点
谁把开拓者推给火箭队？ 姚明：要解决绕前防守 - 9小时前
比赛结束之后，姚明在更衣室接受了媒体采访，他表示，火箭队要想在季后赛中走得更远，就必须解决一个问题，“破绕前”。“我想这个问题，一直是处理不好，打破对方的绕前 ...
搜狐 - 1913 篇相关文章 »
对话魔兽：姚明最难防守我要成第二位黑人总统 - 搜狐 - 22 篇相关文章 »
郭晶晶凭何压张怡宁 谁是下一任广告天后(图) - 央视国际 - 25 篇相关文章 »
姚明官方Flash 谁是姚明？ -姚明官方网站
“我对姚明最欣赏的地方就是他对待比赛的那种谦虚和热情共存的态度，在如今的联盟中，具有这些优点的球员已经看不到了。” ——杰夫-范甘迪，火箭队主帅 ...
yaoming.sports.sohu.com/20071208/n253876711.shtml - 16k - 网页快照 - 类似网页
谁是火箭最该走的人呢？ 姚明之家 篮坛风云 新浪论坛 新浪网
2009年4月5日 ... 谁是火箭最该走的人呢？ 姚明之家 篮坛风云 新浪论坛 新浪网.
sports.sina.com.cn/bbs/2009/0405/114144169.html - 140k - 网页快照 - 类似网页
赛季评分：姚明钻石引领高分猜猜谁是唯一10分？ 姚明-火箭 体坛周报-体坛网
在常规赛结束的时候，还是让我们看看现在和这个赛季曾经的火箭球员的表现，给他们的赛季表现打一个分吧。
rockets.basketball.titan24.com/09-04-16/210058.html - 30k - 网页快照 - 类似网页
火箭球迷热论大前锋人选到底谁是姚明左膀右臂
MrButtocks最后认为诺瓦克应该充当第一替补大前锋，因为从04－05赛季就可以看出三分球对
Done
start
4 Microsoft Office ...
谁是姚明 - Google ...
Downloads
Google - Mozilla Fire...
9:50 AM


谁是姚明 - Google 搜索 - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.google.cn/search?hl=zh-CN&q=谁是姚明&btnG=Google+搜索&meta=&aq=f&oq

Most Visited Helpdesk Presentations (Fensi ...

谁是姚明 - Google 搜索 | GOOG: 390.43 +1.69 (0.43%) - ... | Google.com - Calendar

zhidao.baidu.com/question/2033849 - 30k - 网页快照 - 类似网页

郅联谁是姚明最佳搭档史上最强内线竟成头号难题-经典体育-清谈茶馆 ...
4 个帖子 - 3 个作者 - 新贴子： 2007年9月7日
春秋中文社区# K# C#]# K0 h- b# f1 }5 Q! m8]姚明和王治郅还能打出6年前那样的数据吗?
(X% Z9 G3 i2 P7 U9 `6 P8 f, e7 N+ x6 K. Z2 P5 a2 b1 F1 ...
www.cqzgbbs.net/thread-564511-1-1.html - 类似网页

网友调查：您认为火箭队中谁是姚明的最佳替补-搜狐体育
网友调查：您认为火箭队中谁是姚明的最佳替补. 2009年04月11日13:02 [我来说两句] [字号：
大 中 小]. 来源：搜狐体育. 搜狐体育讯 ...
sports.sohu.com/20090411/n263328639.shtml - 128k - 网页快照 - 类似网页

谁是火炬手？姚明和刘翔的得票数一直领先_YNET.com北青网
谁是火炬手？姚明和刘翔的得票数一直领先. 来源：体育新报(2007/09/07 13:56)◇字号： [大
中 小] 发表评论 · 姚明亲自点燃雅典奥运会北京站的圣火 · 手持祥云火炬 ...
www.ynet.com/zyz/view.jsp?oid=23585326 - 21k - 网页快照 - 类似网页

谁是NBA第四中国秀？姚明不忍朱芳雨来火箭受罪
谁是NBA第四中国秀？姚明不忍朱芳雨来火箭受罪. 2005年04月26日16:31:59 来源：篮球先
锋报. 【字号大 中 小】 【我要打印】 【我要纠错】 ...
news.xi... ...tent_2880317.htm - 61k - 网页快照 - 类似网页

Q&A Yao Ming

相关搜索： ... 出国家队 姚明 姚明资料

相关服务： 到天涯问答提问谁是姚明 到天涯来吧讨论谁是姚明

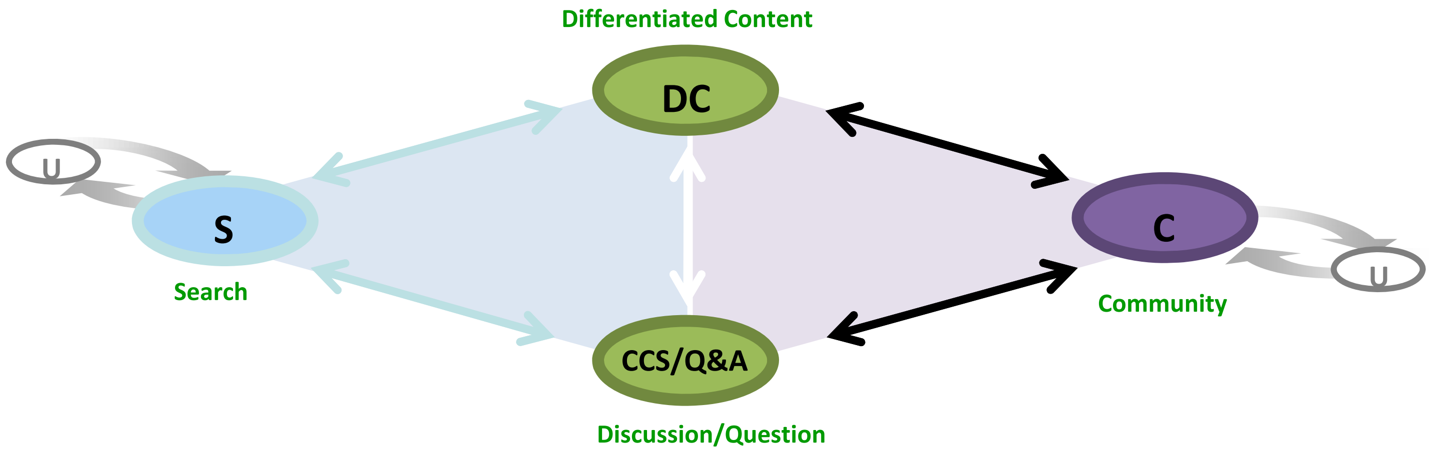
Gooooooooooogle ▶
1 2 3 4 5 6 7 8 9 10 下一页

Done

start | 4 Microsoft Office ... | 谁是姚明 - Google ... | Downloads | Google - Mozilla Fire... | 9:52 AM

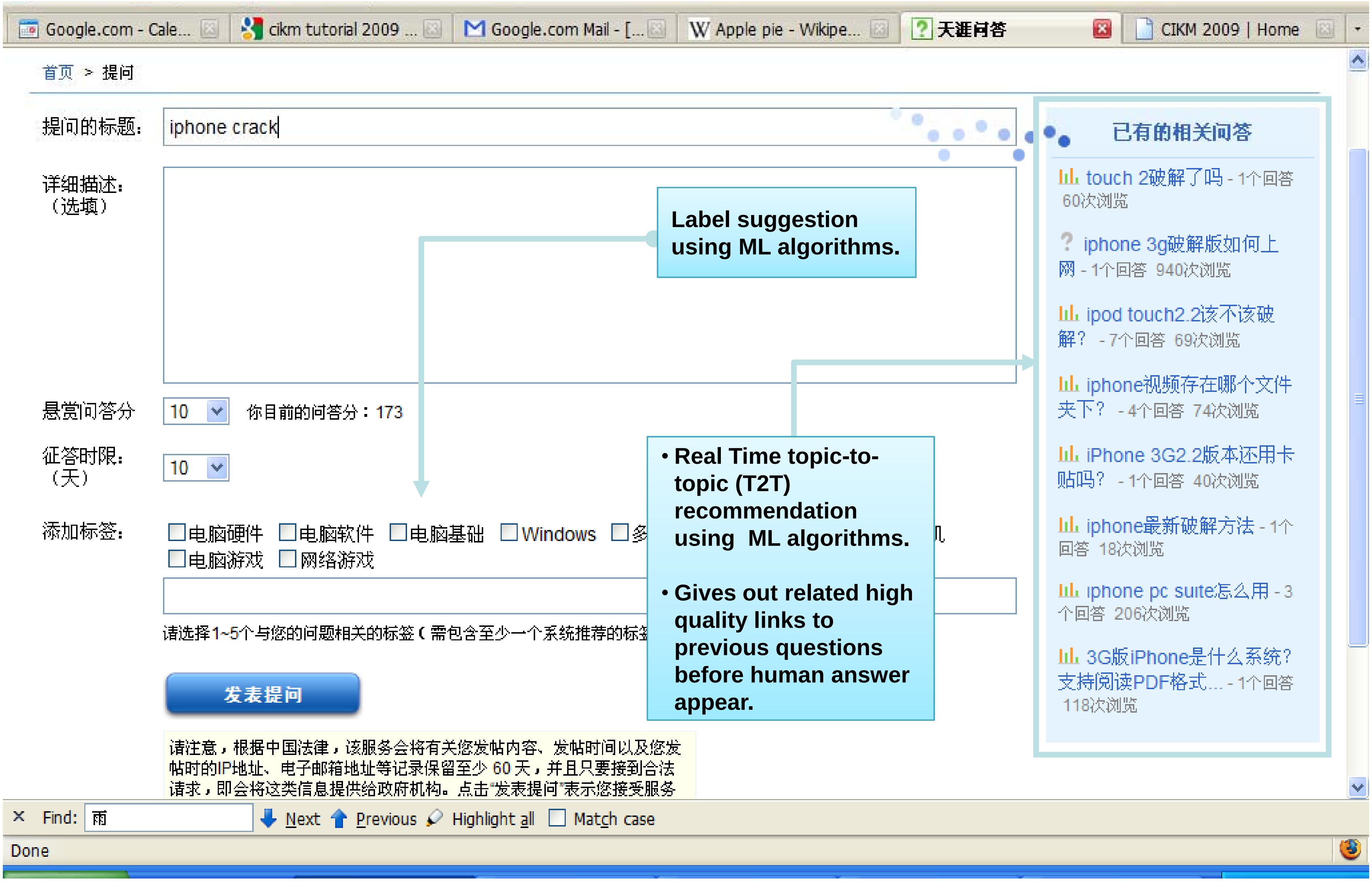

天涯问答 - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://wenda.tianya.cn/wenda/aask?subject=谁是姚明&utm_source=google&utm_medium=
Google
Most Visited Helpdesk Presentations (Fensi ...
谁是姚明 - Google 搜索
GOOG: 390.43 +1.69 (0.43%) - ...
Google.com - Calendar
天涯问答
天涯问答
Google 谷歌 技术提供
搜索引擎里找不到想要的答案？让天涯社区成千上万的专家高手来帮助你！
已有3,634,299个匿名用户已经在此提问，平均每个问题收到第一个答案时间不超过3分钟
让天涯问答专家帮助你答疑解惑
您可以匿名提问。要获得更多精彩知识，请 登录 或者 注册
提问的标题：
谁是姚明
Who is Yao Ming
详细描述：（选填）
提问形式：
还不是天涯用户，使用匿名提问
已经是天涯注册用户，使用天涯ID登录后提问
发表提问
请注意，根据中国法律，该服务会将有关您发帖内容、发帖时间以及您发帖时的IP地址、电子邮箱地址等记录保留至少 60 天，并且只要接到合法请求，即会将这类信息提供给政府机构。点击"发表提问"表示您接受服务条款。服务条款全文»
Yao Ming Related Q&As
2008年中国劳伦斯奖你会选谁呢？ - 112个回答 858次浏览
请问姚明的jj有多大？ - 4个回答 2890次浏览
谁能是最后一个点燃火炬的人? - 6个回答 1344次浏览
姚明一共参加了几届奥运会？分别是哪几... - 3个回答 658次浏览
姚明和刘翔哪个收入高？ - 26个回答 4092次浏览
在哪里能参与北京奥运会奥运物品拍卖？ - 5个回答 4797次浏览
谁是第一个成为NBA状元秀的中国球员？ - 10个回答 1470次浏览
Done

Application: Google Q&A (Confucius)
launched in China, Thailand, and Russia



- ❑ Trigger a discussion/question session during search
- ❑ Provide labels to a Q (semi-automatically)
- ❑ Given a Q, find similar Qs and their As (automatically)
- ❑ Evaluate quality of an answer, relevance and originality
- ❑ Evaluate user credentials in a topic sensitive way
- ❑ Route questions to experts
- ❑ Provide most relevant, high-quality content for Search to index

Q&A Uses Machine Learning



Collaborative Filtering

Based on *membership* so far,
and *memberships* of others

↓

Predict further *membership*

Books/Photos

Users

		1	1	1						
	1		1	1		1		1		1
					1		1			1
	1		1		1	1				
		1								
						1	1			
			1					1		
1	1									
	1								1	
1										1
	1	1	1	1	1					

Collaborative Filtering

Based on *partially* observed matrix

↓

Predict *unobserved* entries

↓

I. Will user **i** enjoy photo **j**?

II. Will user **i** be interesting to user **j**?

III. Will photo **i** be related to photo **j**?

Books/Photos

Users

?		1	1	1		?				
	1	?	1	1		1		1		1
					1	?	1			1
	1		1	?	1	1				
		1					?			
	?					1	1			
			1					1	?	
1	1			?						
	1				?				1	
1								?		1
	1	1	1	1	1	?				

FIM-based Recommendation




To grow the base, we need association rules

- An association rule: $a, b, c \longrightarrow d$
- A Bayesian interpretation: $P(d \mid a, b, c) = \frac{N(a,b,c,d)}{N(a,b,c)}$
- The key is to count the occurrences (*support*) of itemsets $N(\ldots)$

FIM Preliminaries

- Observation 1: If an item A is not frequent, any pattern contains *A* won't be frequent [R. Agrawal]

  → use a threshold to eliminate infrequent items

  ~~{*A*}~~ → ~~{*A, B*}~~

- Observation 2: Patterns containing *A* are subsets of (or found from) transactions containing *A* [J. Han]

  → divide-and-conquer: select transactions containing A to form a conditional database (CDB), and find patterns containing A from that conditional database

  {*A, B*}, {*A, C*}, {*A*} → CDB *A*

  {*A, B*}, {*B, C*} → CDB *B*

- Observation 3: To prevent the same pattern from being found in multiple CDBs, all itemsets are sorted by the same manner (e.g., by descending support)

Preprocessing

f a c d g i m p	f c a m p
a b c f l m o	f c a b m
b f h j o	f b
b c k s p	c b p
a f c e l p m n	f c a m p

f: 4
c: 4
a: 3
b: 3
m: 3
p: 3

---

o: 2
d: 1
e: 1
g: 1
h: 1
i: 1
k: 1
l : 1
n: 1

- According to Observation 1, we count the support of each item by scanning the database, and eliminate those infrequent items from the transactions.
- According to Observation 3, we sort items in each transaction by the order of descending support value.

Parallel Projection

- According to Observation 2, we construct CDB of item A; then from this CDB, we find those patterns containing *A*
- How to construct the CDB of *A*?
  - If a transaction contains *A*, this transaction should appear in the CDB of *A*
  - Given a transaction {*B, A, C*}, it should appear in the CDB of *A*, the CDB of *B*, and the CDB of *C*
- Dedup solution: using the order of items:
  - sort {*B,A,C*} by the order of items → <*A,B,C*>
  - Put <> into the CDB of *A*
  - Put <*A*> into the CDB of *B*
  - Put <*A,B*> into the CDB of *C*

Example of Projection

f c a m p	p: { f c a m / f c a m / c b }
f c a b m	m: { f c a / f c a / f c a b }
f b	b: { f c a / f / c }
c b p	a: { f c / f c / f c }
f c a m p	c: { f / f / f }

Example of Projection of a database into CDBs.
Left: sorted transactions in order of *f*, *c*, *a*, *b*, *m*, *p*
Right: conditional databases of frequent items

Example of Projection

f c a m p	p: { f c a m / f c a m / c b }
f c a b m	m: { f c a / f c a / f c a b }
f b	b: { f c a / f / c }
c b p	a: { f c / f c / f c }
f c a m p	c: { f / f / f }

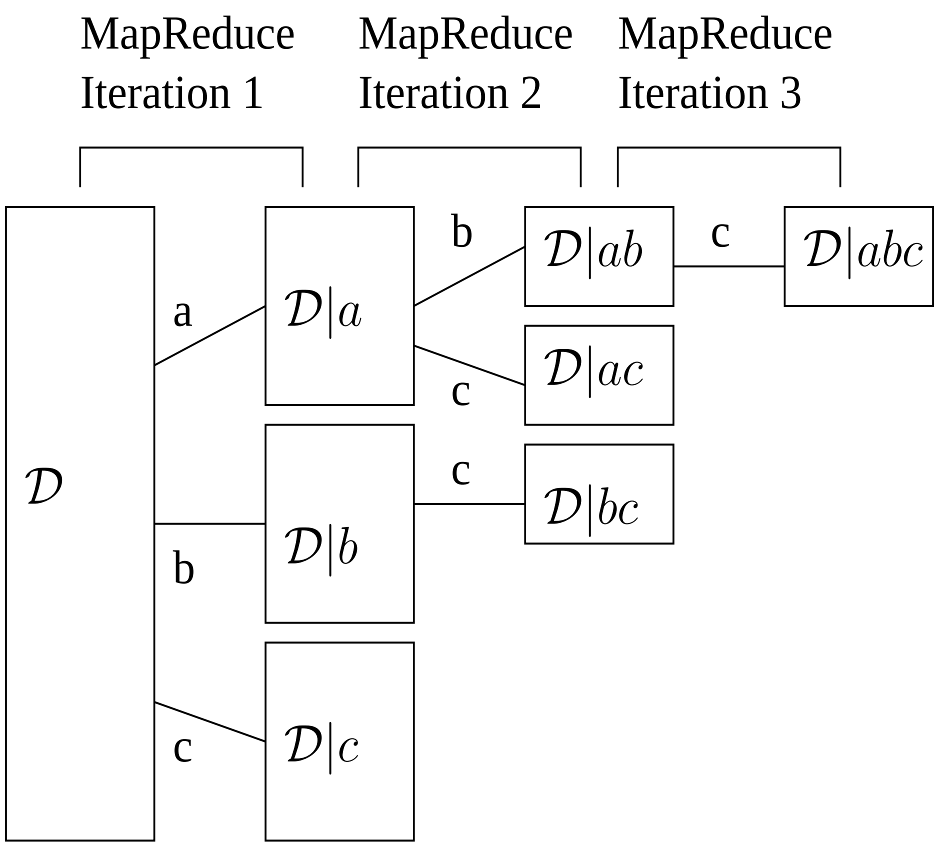
Example of Projection of a database into CDBs.
Left: sorted transactions;
Right: conditional databases of frequent items

Example of Projection

f c a m p	p: { f c a m / f c a m / c b }
f c a b m	m: { f c a / f c a / f c a b }
f b	b: { f c a / f / c }
c b p	a: { f c / f c / f c }
f c a m p	c: { f / f / f }

Example of Projection of a database into CDBs.
Left: sorted transactions;
Right: conditional databases of frequent items

Recursive Projections [H. Li, et al. ACM RS 08]



- Recursive projection form a search tree
- Each node is a CDB
- Using the order of items to prevent duplicated CDBs.
- Each level of breath-first search of the tree can be done by a MapReduce iteration.
- Once a CDB is small enough to fit in memory, we can invoke FP-growth to mine this CDB, and no more growth of the sub-tree.

Projection using MapReduce

Map inputs (transactions) key="": value	Sorted transactions (with infrequent items eliminated)	Map outputs (conditional transactions) key: value
f a c d g i m p	f c a m p	p: f c a m m: f c a a: fc c: f
a b c f l m o	f c a b m	m: f c a b b: f c a a: f c c: f
b f h j o	f b	b: f
b c k s p	c b p	p: c b
a f c e l p m n	f c a m p	b: c p: f c a m m: f c a a: f c c: f

Reduce inputs (conditional databases) key: value		Reduce outputs (patterns and supports) key: value
p:	{fcam/fcam/cb}	p:3, pc:3
m:	{ f c a / f c a / f c a b }	m f : 3 m c : 3 m a : 3 m f c : 3 m f a : 3 m c a : 3 m f c a : 3
b:	{ f c a / f / c }	b : 3
a:	{ f c / f c / f c }	a : 3 a f : 3 a c : 3 a f c : 3
c:	{ f / f / f }	c : 3 c f : 3

Collaborative Filtering

Based on *membership* so far,
and *memberships* of others

↓

Predict further *membership*

Indicators/Diseases

Individuals

		1	1	1						
	1		1	1		1		1		1
					1		1			1
	1		1		1	1				
		1								
						1	1			
			1					1		
1	1									
	1								1	
1										1
	1	1	1	1	1					

Latent Semantic Analysis

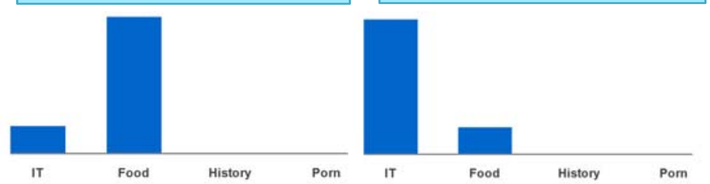
- Search
  - Construct a latent layer for better for semantic matching
- Example:
  - iPhone crack
  - Apple pie

Documents

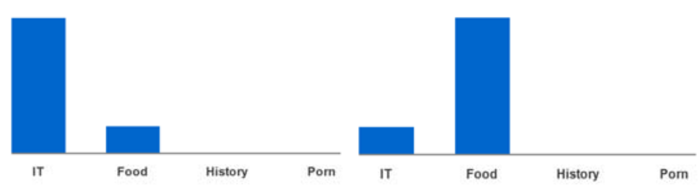
1 recipe pastry for a 9 inch double crust 9 apples, 2/1 cup, brown sugar

How to install apps on Apple mobile phones?

Topic Distribution



Topic Distribution



User quries

iPhone crack

Apple pie

Users/Music/Ads/Question

Users/Music/Ads/Answers

?	?	1	3	1	?	?	?	?	?	?
?	2	?	1	2	?	1	?	3	?	1
?	?	?	?	?	1		5			1
	5		3		1	1				
		1								
						1	4			
			2					1		
1	2									
	1								5	
1										1
	1	4	1	3	6					

- Other Collaborative Filtering Apps
  - Recommend Users → Users
  - Recommend Music → Users
  - Recommend Ads → Users
  - Recommend Answers → Q
- Predict the ? In the light-blue cells

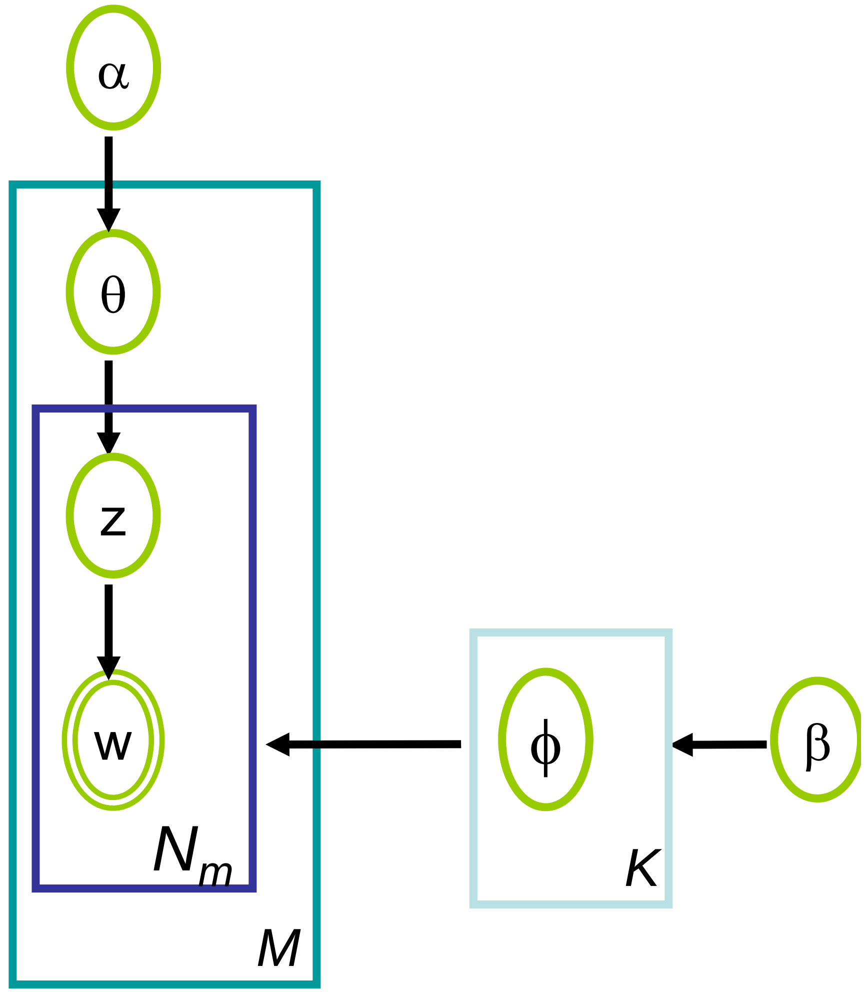
Documents, Topics, Words

- A document consists of a number of topics
  - A document is a probabilistic mixture of topics
- Each topic generates a number of words
  - A topic is a distribution over words
  - The probability of the i^{th} word in a document

$$P(w_i) = \sum_{j=1}^{T} P(w_i | z_i = j) P(z_i = j)$$

Latent Dirichlet Allocation [M. Jordan 04]

- α: uniform Dirichlet ϕ prior for per document d topic distribution (corpus level parameter)
- β: uniform Dirichlet ϕ prior for per topic z word distribution (corpus level parameter)
- θ_d is the topic distribution of doc d (document level)
- z_{dj} the topic if the j^{th} word in d, w_{dj} the specific word (word level)



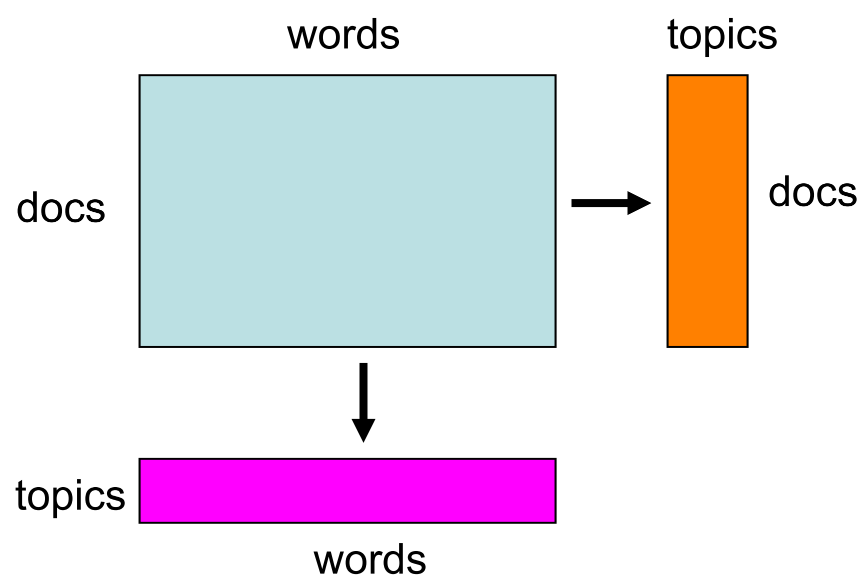
LDA Gibbs Sampling: Inputs And Outputs

Inputs:

1. training data: documents as bags of words
2. parameter: the number of topics

Outputs:

1. by-product: a co-occurrence matrix of topics and documents.
2. model parameters: a co-occurrence matrix of topics and words.



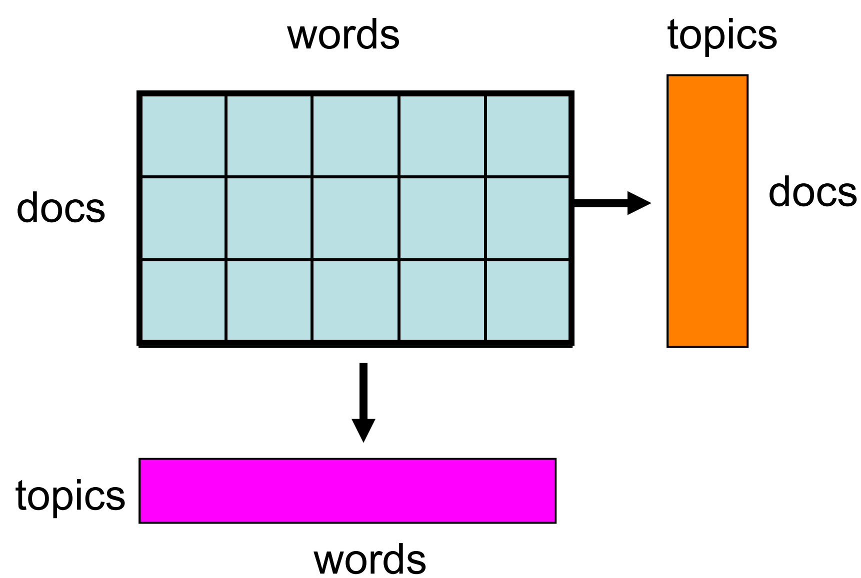
Parallel Gibbs Sampling [aaim 09]

Inputs:

1. training data: documents as bags of words
2. parameter: the number of topics

Outputs:

1. **by-product**: a co-occurrence matrix of topics and documents.
2. model parameters: a co-occurrence matrix of topics and words.



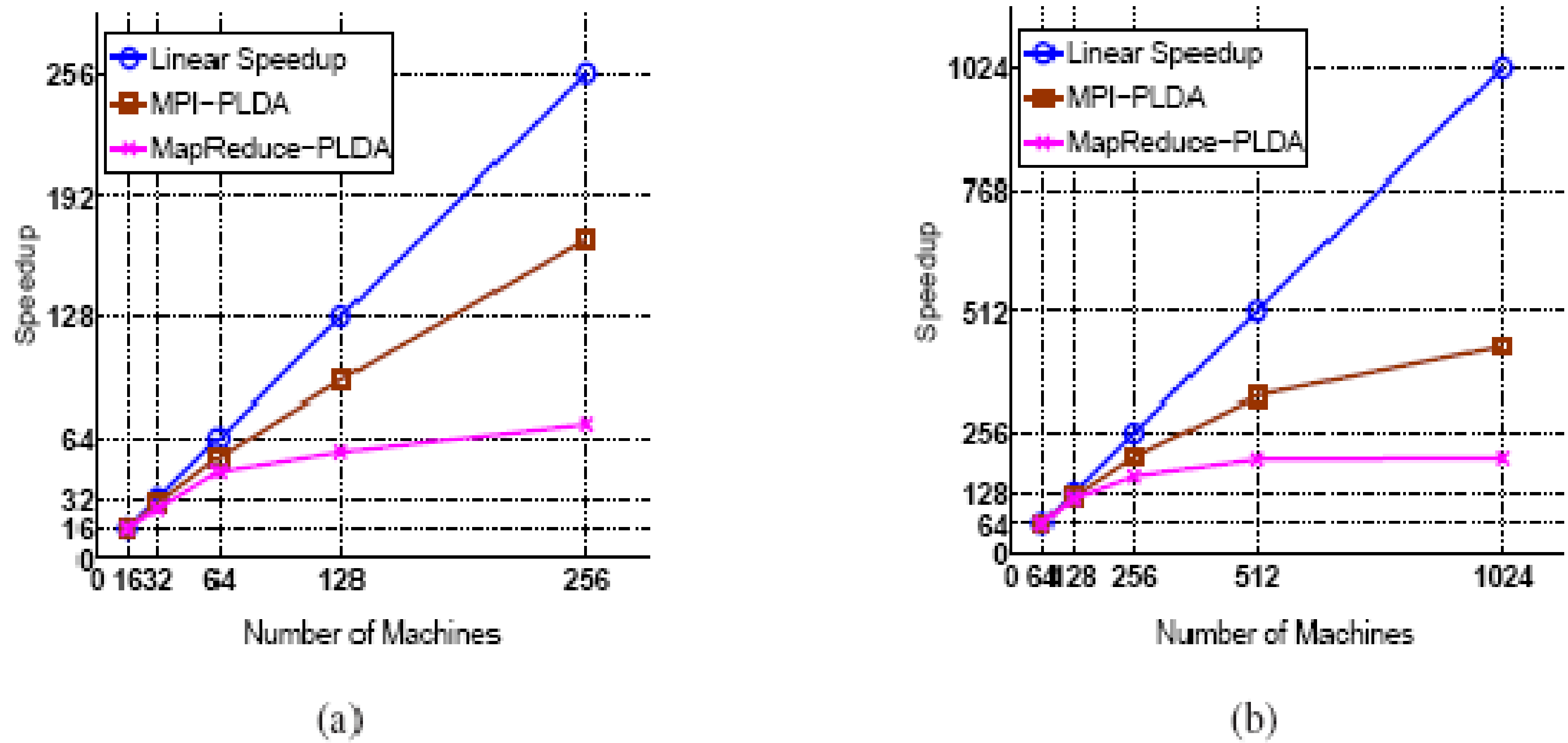


Fig. 4: The speedup of (a) Wikipedia: $K = 500$, $V = 20000$, $D = 2,122,618$, $TotalWordOccurrences = 447,004,756$, $iterations = 20$, $\alpha = 0.1$, $\beta = 0.1$. (b) Forum Dataset: $K = 500$, $V = 50000$, $D = 2,450,379$, $TotalWordOccurrences = 3,223,704,976$, $iterations = 10$, $\alpha = 0.1$, $\beta = 0.1$.

Table 6: Speedup Performance of MPI-PLDA and MapReduce-PLDA

# Machines	MPI-PLDA		MapReduce-PLDA	
	Running Time	*Speedup*	*Running Time*	*Speedup*
16	11940s	16	12022s	16
32	6468s	30	7288s	26
64	3546s	54	4165s	46
128	2030s	94	3395s	57
256	1130s	169	2680s	72

(a) Widipedia dataset (Runtime of 20 iterations)

# Machines	MPI-PLDA		MapReduce-PLDA	
	Running Time	*Speedup*	*Running Time*	*Speedup*
64	9012s	64	10612s	64
128	4792s	120	5817s	117
256	2811s	205	4132s	164
512	1735s	332	3390s	200
1024	1323s	436	3349s	203

(b) Forum dataset (Runtime of 10 iterations)

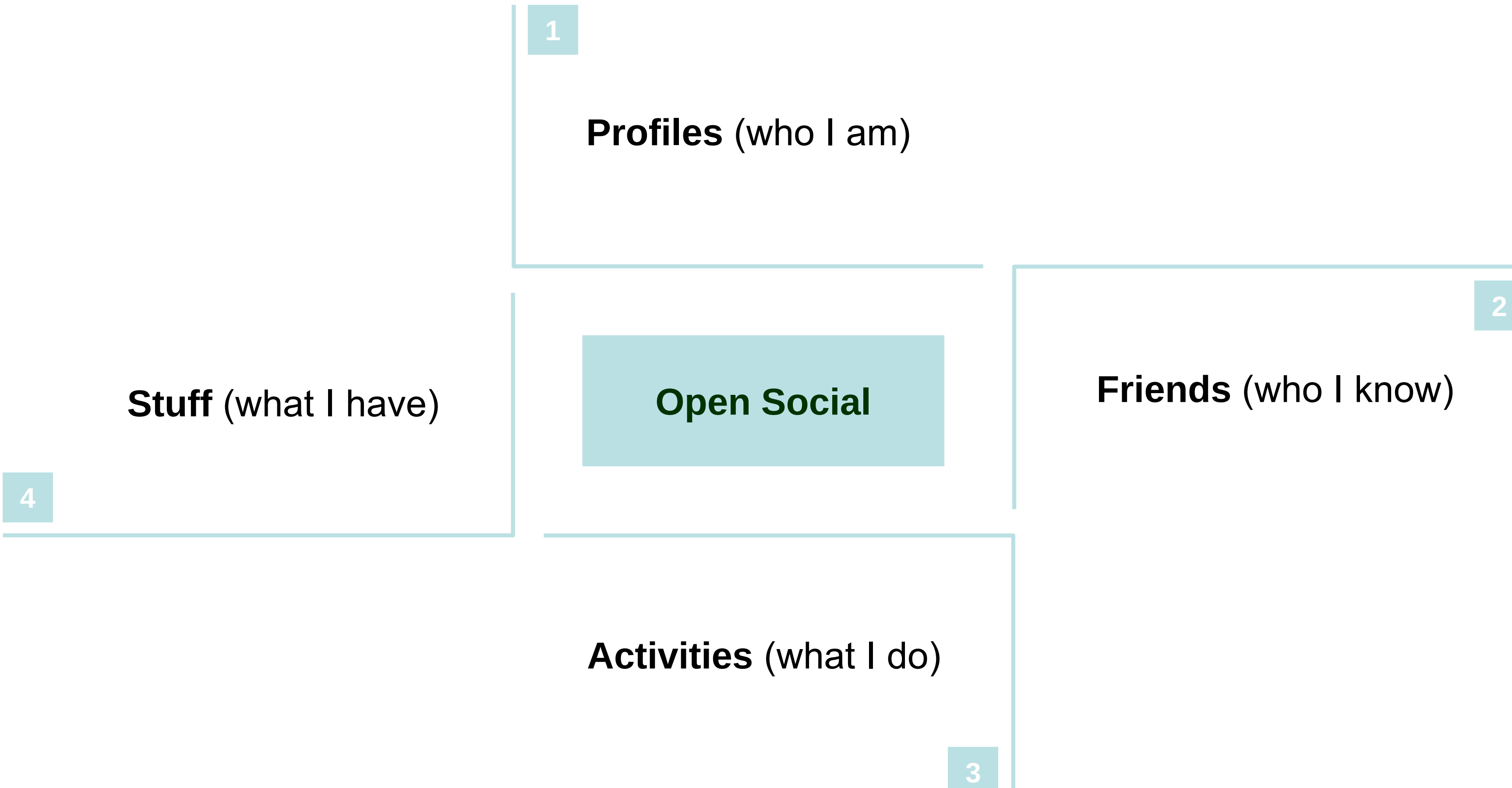
Outline

- Motivating Applications
  - Q&A System
  - Social Ads
- Key Subroutines
  - Frequent Itemset Mining [ACM RS 08]
  - Latent Dirichlet Allocation [WWW 09, AAIM 09]
  - Clustering [ECML 08]
  - UserRank [Google TR 09]
  - Support Vector Machines [NIPS 07]
- Distributed Computing Perspectives


天涯来吧 - 我的朋友圈 - Mozilla Firefox
File
Edit
View
History
Bookmarks
Tools
Help
http://laiba.tianya.cn/laiba/FlashFriends
Google
Google.com - In...
google gadget -...
OpenSocial Plat...
Fensi Dashboard
天涯来吧 - 我...
Darren Huang - ...
Google.com - C...
我的主页
资料
朋友
来吧
帖子
影集
日记
礼物
New!
评价
留言簿
我的朋友圈 | 我的朋友
上一步
回到自己
永远来吧
易岱
机械舞
汤汤汤圆
暗香浮月
开复
沉网
武汉牛牛
水星cici
梦侠人
无聊最寂寞
邀请朋友加入来吧
邮件
发送邀请
我要群发邀请»
看看我的朋友在哪
Transferring data from laiba.tianya.cn...
start
3 Firefox
C:\Docum...
Document...
2 Micros...
Google R...
EN
10:04 PM


天涯来吧 - 我的朋友圈 - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://laiba.tianya.cn/laiba/FlashFriends
Google.com - In...
google gadget -...
Google.com
Fensi Dashboard
天涯来吧 - 我...
Darren Huang - ...
Google.com - C...
我的主页 资料 朋友 来吧 帖子 影集 日记 礼物 New! 评价 留言簿
我的朋友圈 | 我的朋友
上一步
回到自己
永远来吧 (离线)
我的好友
批量上传照片吧!
男 37岁 北京
夜幕诱惑 诱的就是你!... [08-1-3]
这有没GOOGLE公司的伙... [07-8-20]
白领的家庭聚会 [07-12-30]
白领的家庭聚会 [07-12-30]
白领的家庭聚会 [07-12-30]
移除好友
永远来吧
机械舞
暗香浮月
水星cici
梦侠人
无聊最寂寞
邀请朋友加入来吧
邮件
发送邀请
我要群发邀请»
看看我的朋友在哪
Transferring data from laiba.tianya.cn...
start
3 Firefox
C:\Docum...
Document...
2 Micros...
Google R...
EN
10:12 PM

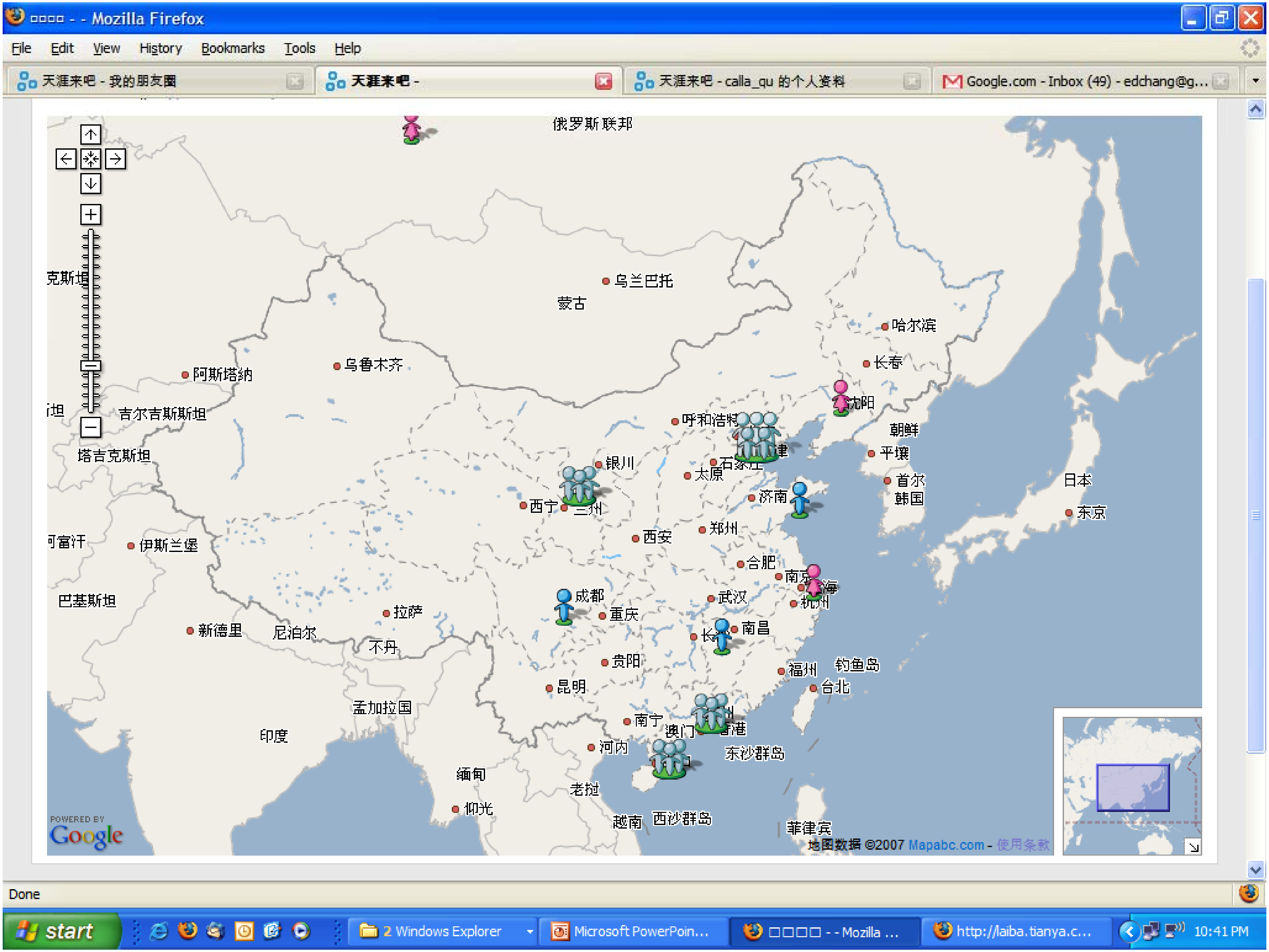
Open Social APIs

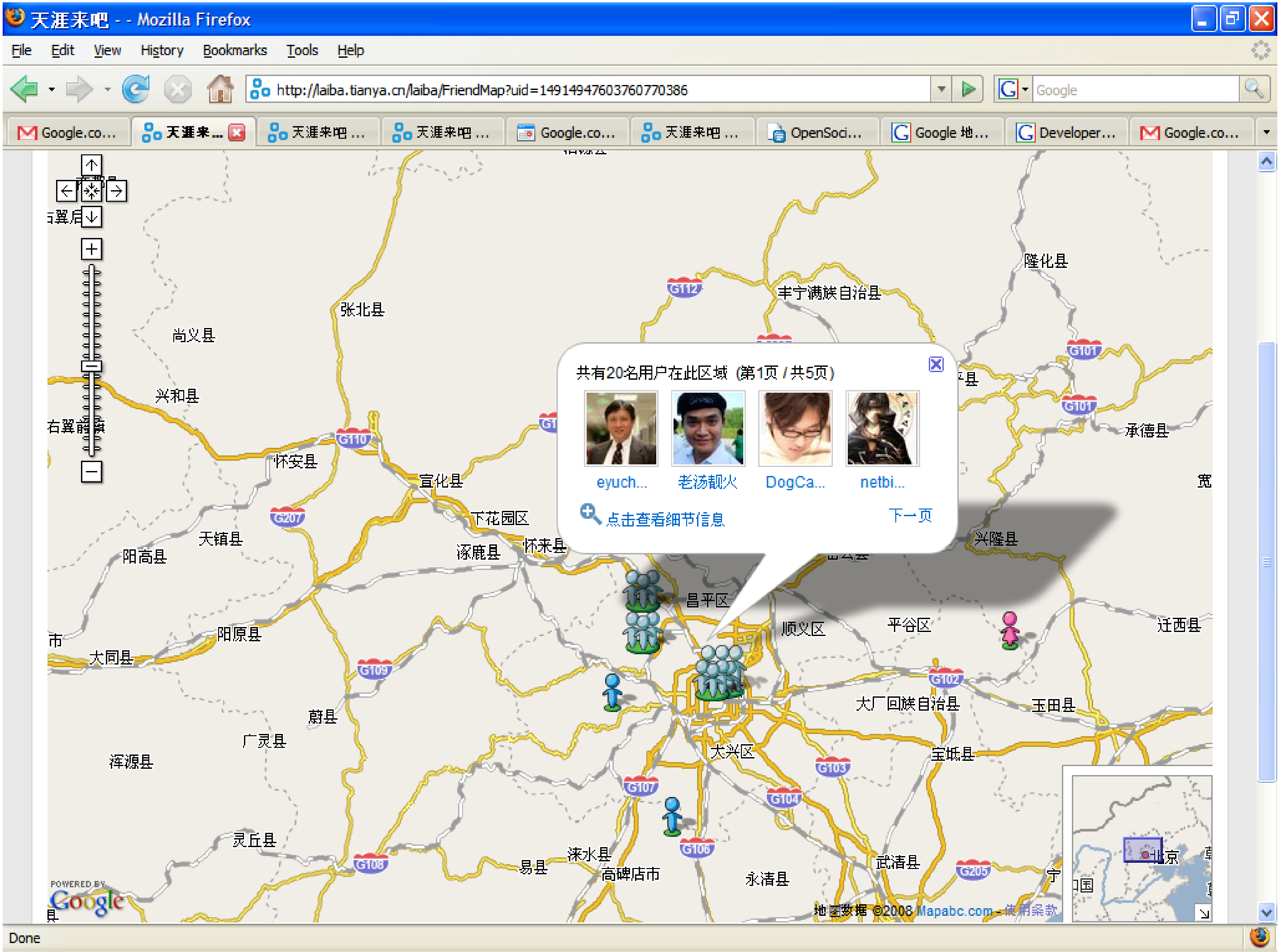



Facebook | Home - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.facebook.com/home.php?ref=logo
Google
Google.com Mail - I...
Google.com - Cale...
Flare | Demos
Google
Facebook | Home
Dart Search Notific...
facebook
Home
Profile
Friends
Inbox
Edward Chang
Settings
Logout
Search
News Feed
UCSB
Status Updates
Photos
Links
More
Create
What's on your mind?
Share
Activities
Bo Zhang http://www.optrip.com
BETA
OpTrip - Plan a Trip
Source: www.optrip.com
Yesterday at 5:02pm · Comment · Like · Share
Tom Wang at 9:28pm June 1
are u working there? btw, I can't connect to it.
Write a comment...
Ling Ling Wed - dinner at BJ's to meet up some old friends from Beijing; funny thing was everyone thought BJ's was a Peking duck restaurant!
Thur - salsa @ Alberto's in MV. Thanks Wing for coming and Stephen was visiting from NYC.
Sat - BBQ @ Lynn and Jeff's new crib in SOMA and we also watched UP in 3D IMAX :)
Sun - reunion brunch with International staffing at Stanford mall and also did shopping there with Dandan.
Requests
See All
6 friend requests
1 how will you die? request
1 best friend request
41 other requests
Suggestions
See All
Chunlei Niu
You and Chunlei both worked at Google.
Add as Friend
Sponsored
Recommendations
Ads
Learn how to connect your business to real customers through Facebook Ads.
Highlights
Last week of May
by Ling Ling
Random...
by Rosalind Chang
TouchGraph Photos
Applications
Applications
Chat (1)
Done
start
4 Windows Explorer
2 Firefox
2 Microsoft Office ...
4 Microsoft Office ...
9:02 AM


Facebook | Kaihua Zhu's Photos - Mobile Uploads - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.facebook.com/photo.php?pid=2038339&id=553855783&ref=nf
Google
Google.com Mail - I...
Google.com - Cale...
Flare | Demos
Google
Facebook | Kaih...
Dart Search Notific...
facebook
Home
Profile
Friends
Inbox
Edward Chang
Settings
Logout
Search
Kaihua Zhu's Photos - Mobile Uploads
Photo 3 of 5 | Back to Album | Kaihua's Photos | Kaihua's Profile
Previous Next
Advertise
Find Your Target Audience
Facebook has over 200 million active users. Quickly find out how many of them match your target audience for free!
verizon
The NEW Verizon Small Business Center
Applications
Chat (0)
Done
start
4 Windows Explorer
2 Firefox
2 Microsoft Office ...
4 Microsoft Office ...
8:59 AM


□□□□ - □□□□□ - Mozilla Firefox
File Edit View History Bookmarks Tools Help
我的主页 资料 朋友 来吧 帖子 影集 日记 礼物 New! 评价 留言簿
我的朋友圈 | 我的朋友
上一步
回到自己
cxwangyi
DogCaptain
NelsonYao
bornde
eyuchang
cyr_215
calla_qu
2dowhat
Mindy_xiaomi...
80x86
YitengZhang
Google
邀请朋友加入来吧
邮件
姓名
发送邀请
我要群发邀请»
看看我的朋友在哪
Transferring data from laiba.tianya.cn...
start
2 Windows Explorer
Microsoft PowerPoi...
□□□□ - □□□...
http://laiba.tianya....
9:30 PM


□□□□ - - Mozilla Firefox
File Edit View History Bookmarks Tools Help
天涯来吧 - 我的朋友圈
天涯来吧 -
天涯来吧 - calla_qu 的个人资料
Google.com - Inbox (49) - edchang@g...
俄罗斯联邦
乌兰巴托
蒙古
哈尔滨
长春
沈阳
乌鲁木齐
阿斯塔纳
吉尔吉斯斯坦
塔吉克斯坦
呼和浩特
朝鲜
平壤
首尔
韩国
日本
东京
银川
太原
石家庄
济南
西宁
兰州
郑州
西安
合肥
南京
上海
杭州
武汉
成都
重庆
南昌
伊斯兰堡
巴基斯坦
新德里
尼泊尔
拉萨
不丹
贵阳
昆明
福州
钓鱼岛
台北
孟加拉国
印度
南宁
澳门
香港
东沙群岛
河内
缅甸
老挝
仰光
越南
西沙群岛
菲律宾
POWERED BY Google
地图数据 ©2007 Mapabc.com - 使用条款
Done
start
2 Windows Explorer
Microsoft PowerPoin...
□□□□ - - Mozilla ...
http://laiba.tianya.c...
10:41 PM



天涯来吧 - - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://laiba.tianya.cn/laiba/FriendMap?uid=1491494760376O770386
共有20名用户在此区域（第1页 / 共5页）
eyuch...
老汤靓火
DogCa...
netbi...
点击查看细节信息
下一页
张北县
尚义县
兴和县
怀安县
宣化县
下花园区
涿鹿县
怀来县
天镇县
阳高县
阳原县
大同县
蔚县
广灵县
浑源县
灵丘县
易县
涞水县
高碑店市
永清县
大兴区
昌平区
顺义区
平谷区
兴隆县
迁西县
承德县
隆化县
丰宁满族自治县
大厂回族自治县
玉田县
宝坻县
武清县
地图数据 ©2008 Mapabc.com - 使用条款
Done

Task: Targeting Ads at SNS Users

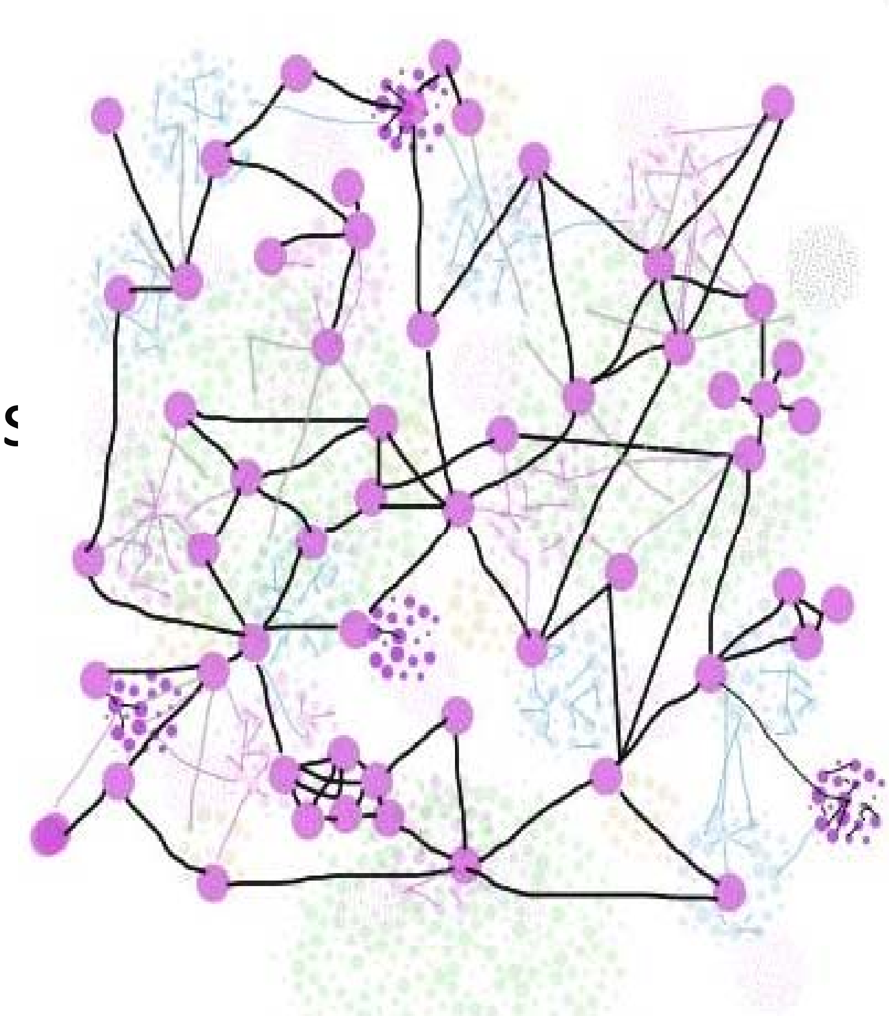
Users

Ads

Mining Profiles, Friends & Activities for Relevance

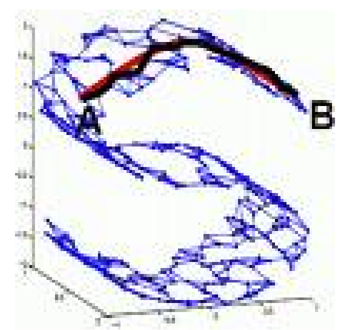
Consider also User *Influence*

- Advertisers consider users who are
  - Relevant
  - *Influential*
- SNS Influence Analysis
  - Centrality
  - Credential
  - Activeness
  - etc.



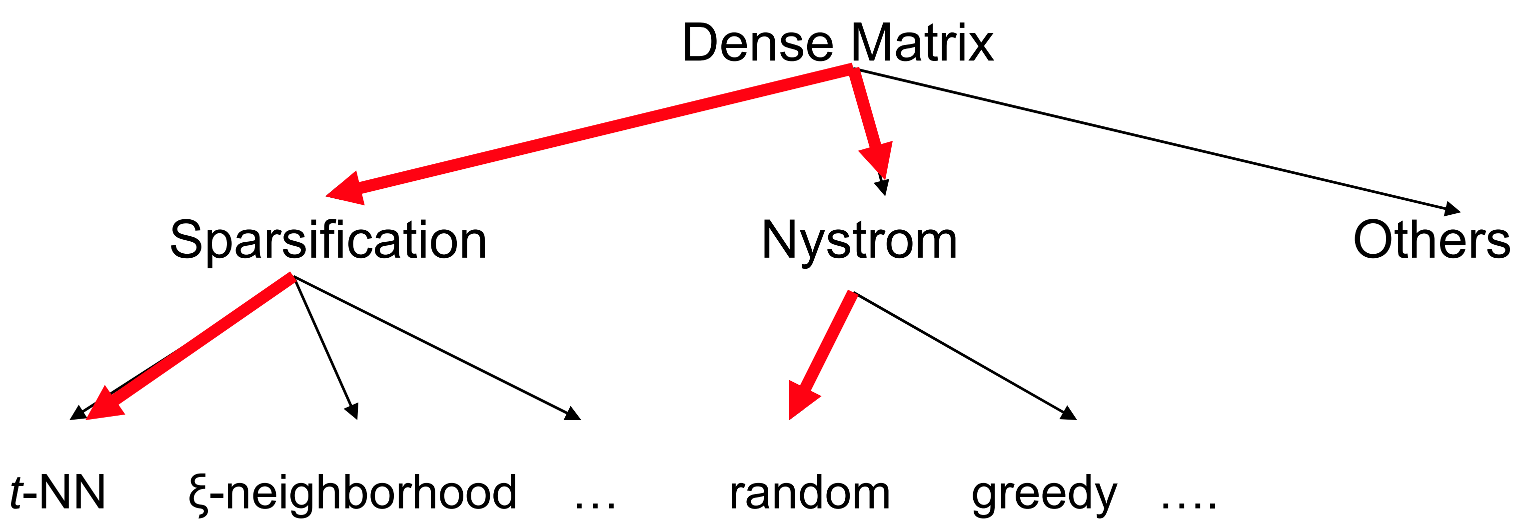
Spectral Clustering [A. Ng, M. Jordan]

- Important subroutine in tasks of machine learning and data mining
  - Exploit *pairwise similarity* of data instances
  - More effective than traditional methods e.g., k-means
- Key steps
  - Construct pairwise similarity matrix
    - e.g., using Geodisc distance
  - Compute the Laplacian matrix
  - Apply eigendecomposition
  - Perform *k*-means



Scalability Problem

- Quadratic computation of nxn matrix
- Approximation methods



Sparsification vs. Sampling

- Construct the dense similarity matrix S
- Sparsify S
- Compute Laplacian matrix L

$$L = I - D^{-1/2} S D^{-1/2}, \quad D_{ii} = \sum_{j=1}^{n} S_{ij}$$

- Apply *ARPACLK* on L
- Use *k*-means to cluster rows of V into *k* groups

- Randomly sample *l* points, where *l* << *n*
- Construct dense similarity matrix [A B] between *l* and *n* points
- Normalize A and B to be in Laplacian form

$R = A + A^{-1/2} B B^T A^{-1/2}$;

$R = U \Sigma U^T$

- *k*-means

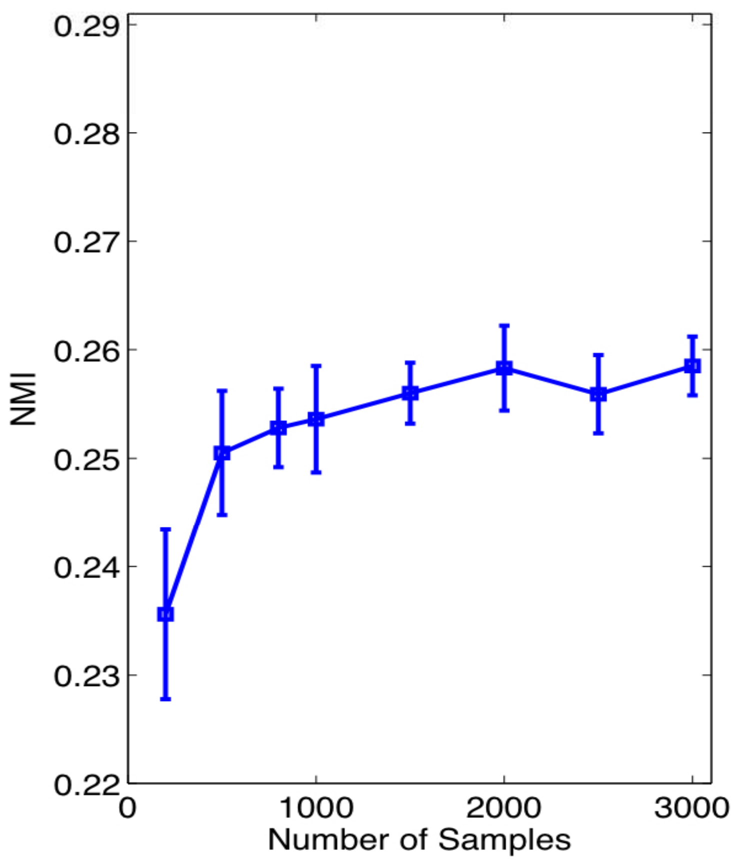
Empirical Study [song, et al., ecml 08]

- Dataset: RCV1 (Reuters Corpus Volume I)
  - A filtered collection of *193,944* documents in *103* categories
- Photo set: PicasaWeb
  - *637,137* photos
- Experiments
  - Clustering quality vs. computational time
    - Measure the similarity between CAT and CLS
    - Normalized Mutual Information (NMI)

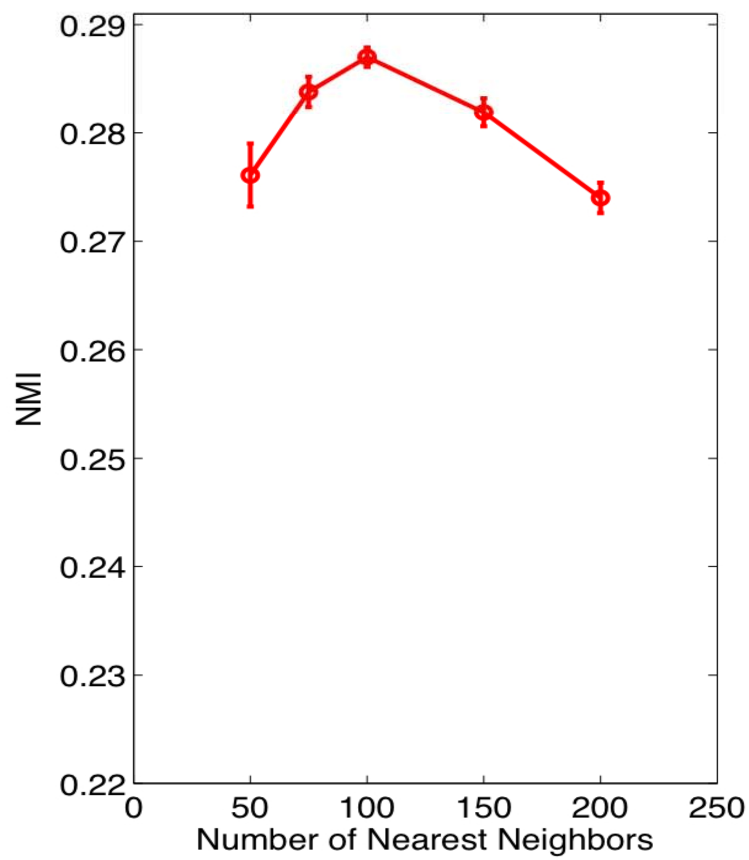
$$NMI(CAT;CLS) = \frac{I(CAT;CLS)}{\sqrt{H(CAT)H(CLS)}}$$

  - Scalability

NMI Comparison (on RCV1)



Nystrom method



Sparse matrix approximation

Speedup Test on *637,137* Photos

- K = 1000 clusters

	Eigensolver		k-means	
Machines	Time (sec.)	Speedup	Time (sec.)	Speedup
1	—	—	—	—
2	8.074×10^4	2.00	3.609×10^4	2.00
4	4.427×10^4	3.65	1.806×10^4	4.00
8	2.184×10^4	7.39	8.469×10^3	8.52
16	9.867×10^3	16.37	4.620×10^3	15.62
32	4.886×10^3	33.05	2.021×10^3	35.72
64	4.067×10^3	39.71	1.433×10^3	50.37
128	3.471×10^3	46.52	1.090×10^3	66.22
256	4.021×10^3	40.16	1.077×10^3	67.02

- Achiever linear speedup when using 32 machines, after that, sub-linear speedup because of increasing communication and sync time

Sparsification vs. Sampling

	Sparsification	Nystrom, random sampling
Information	Full n x n similarity scores	None
Pre-processing Complexity (bottleneck)	$O(n^2)$ worst case; easily parallizable	O(nl), l << n
Effectiveness	Good	Not bad (Jitendra M., PAMI)

Outline

- Motivating Applications
  - Q&A System
  - Social Ads
- Key Subroutines
  - Frequent Itemset Mining [ACM RS 08]
  - Latent Dirichlet Allocation [WWW 09, AAIM 09]
  - Clustering [ECML 08]
  - UserRank [Google TR 09]
  - Support Vector Machines [NIPS 07]
- Distributed Computing Perspectives

Matrix Factorization Alternatives

Factorization	Cost
QR	$O(\frac{4}{3}n^3)$
LU	$O(\frac{2}{3}n^3)$
Cholesky	$O(\frac{1}{3}n^3 + 2n^2)$
LDL^T	$O(\frac{1}{3}n^3)$
Incomplete Cholesky	$O(p^2 n)$
Kronecker	$O(2n^2)$

exact

approximate

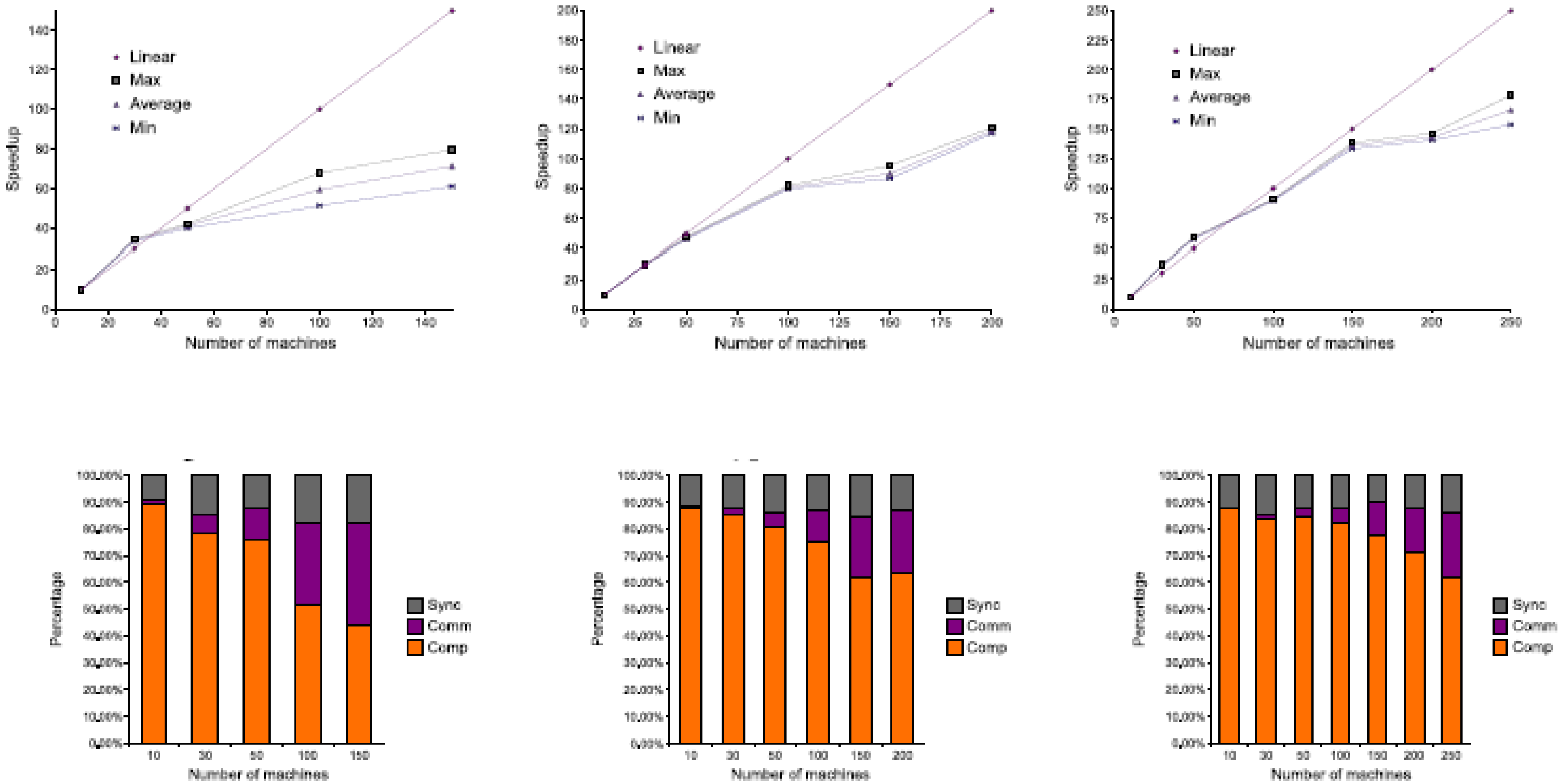
PSVM [E. Chang, et al, NIPS 07]

- Column-based ICF
  - Slower than row-based on single machine
  - Parallelizable on multiple machines
- Changing IPM computation order to achieve parallelization

Speedup

	Image (200k)			CoverType (500k)			RCV (800k)		
Machines	Time (s)		Speedup	Time (s)		Speedup	Time (s)		Speedup
10	1,958	(9)	10*	16,818	(442)	10*	45,135	(1373)	10*
30	572	(8)	34.2	5,591	(10)	30.1	12,289	(98)	36.7
50	473	(14)	41.4	3,598	(60)	46.8	7,695	(92)	58.7
100	330	(47)	59.4	2,082	(29)	80.8	4,992	(34)	90.4
150	274	(40)	71.4	1,865	(93)	90.2	3,313	(59)	136.3
200	294	(41)	66.7	1,416	(24)	118.7	3,163	(69)	142.7
250	397	(78)	49.4	1,405	(115)	119.7	2,719	(203)	166.0
500	814	(123)	24.1	1,655	(34)	101.6	2,671	(193)	169.0
LIBSVM	4,334	NA	NA	28,149	NA	NA	184,199	NA	NA

Overheads



Linear
Max
Average
Min
Speedup
Number of machines
Sync
Comm
Comp
Percentage
Number of machines

Comparison between Parallel Computing Frameworks

	MapReduce	Project B	MPI
GFS/IO and task rescheduling overhead between iterations	Yes	No +1	No +1
Flexibility of computation model	AllReduce only +0.5	AllReduce only +0.5	Flexible +1
Efficient AllReduce	Yes +1	Yes +1	Yes +1
Recover from faults between iterations	Yes +1	Yes +1	Apps
Recover from faults within each iteration	Yes +1	Yes +1	Apps
Final Score for scalable machine learning	3.5	4.5	5

Concluding Remarks

- Applications demand scalable solutions
- Have parallelized key subroutines for mining massive data sets
  - Spectral Clustering [ECML 08]
  - Frequent Itemset Mining [ACM RS 08]
  - PLSA [KDD 08]
  - LDA [WWW 09]
  - UserRank
  - Support Vector Machines [NIPS 07]
- Relevant papers
  - http://infolab.stanford.edu/~echang/
- Open Source PSVM, PLDA
  - http://code.google.com/p/psvm/
  - http://code.google.com/p/plda/

Collaborators

- Prof. Chih-Jen Lin (NTU)
- Hongjie Bai (Google)
- Wen-Yen Chen (UCSB)
- Jon Chu (MIT)
- Haoyuan Li (PKU)
- Yangqiu Song (Tsinghua)
- Matt Stanton (CMU)
- Yi Wang (Google)
- Dong Zhang (Google)
- Kaihua Zhu (Google)

References

[1] Alexa internet. http://www.alexa.com/.
[2] D. M. Blei and M. I. Jordan. Variational methods for the dirichlet process. In Proc. of the 21st international conference on Machine learning, pages 373-380, 2004.
[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. Journal of Machine Learning Research, 3:993-1022, 2003.
[4] D. Cohn and H. Chang. Learning to probabilistically identify authoritative documents. In Proc. of the Seventeenth International Conference on Machine Learning, pages 167-174, 2000.
[5] D. Cohn and T. Hofmann. The missing link - a probabilistic model of document content and hypertext connectivity. In Advances in Neural Information Processing Systems 13, pages 430-436, 2001.
[6] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. Journal of the American Society of Information Science, 41(6):391-407, 1990.
[7] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 39(1):1-38, 1977.
[8] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. IEEE Transactions on Pattern recognition and Machine Intelligence, 6:721-741, 1984.
[9] T. Hofmann. Probabilistic latent semantic indexing. In Proc. of Uncertainty in Arti cial Intelligence, pages 289-296, 1999.
[10] T. Hofmann. Latent semantic models for collaborative filtering. ACM Transactions on Information System, 22(1):89-115, 2004.
[11] A. McCallum, A. Corrada-Emmanuel, and X. Wang. The author-recipient-topic model for topic and role discovery in social networks: Experiments with enron and academic email. Technical report, Computer Science, University of Massachusetts Amherst, 2004.
[12] D. Newman, A. Asuncion, P. Smyth, and M. Welling. Distributed inference for latent dirichlet allocation. In Advances in Neural Information Processing Systems 20, 2007.
[13] M. Ramoni, P. Sebastiani, and P. Cohen. Bayesian clustering by dynamics. Machine Learning, 47(1):91-121, 2002.

References (cont.)

[14] R. Salakhutdinov, A. Mnih, and G. Hinton. Restricted boltzmann machines for collaborative
ltering. In Proc. Of the 24th international conference on Machine learning, pages 791-798, 2007.

[15] E. Spertus, M. Sahami, and O. Buyukkokten. Evaluating similarity measures: a large-scale study in the orkut social network. In Proc. of the 11th ACM SIGKDD international conference on Knowledge discovery in data mining, pages 678-684, 2005.

[16] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Gri♬ ths. Probabilistic author-topic models for information discovery. In Proc. of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 306-315, 2004.

[17] A. Strehl and J. Ghosh. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. Journal on Machine Learning Research (JMLR), 3:583-617, 2002.

[18] T. Zhang and V. S. Iyengar. Recommender systems using linear classi
ers. Journal of Machine Learning Research, 2:313-334, 2002.

[19] S. Zhong and J. Ghosh. Generative model-based clustering of documents: a comparative study. Knowledge and Information Systems (KAIS), 8:374-384, 2005.

[20] L. Admic and E. Adar. How to search a social network. 2004

[21] T.L. Griffiths and M. Steyvers. Finding scientific topics. Proceedings of the National Academy of Sciences, pages 5228-5235, 2004.

[22] H. Kautz, B. Selman, and M. Shah. Referral Web: Combining social networks and collaborative filtering. Communitcations of the ACM, 3:63-65, 1997.

[23] R. Agrawal, T. Imielnski, A. Swami. Mining association rules between sets of items in large databses. SIGMOD Rec., 22:207-116, 1993.

[24] J. S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In Proceedings of the Fourteenth Conference on Uncertainty in Artifical Intelligence, 1998.

[25] M.Deshpande and G. Karypis. Item-based top-n recommendation algorithms. ACM Trans. Inf. Syst., 22(1):143-177, 2004.

References (cont.)

[26] B.M. Sarwar, G. Karypis, J.A. Konstan, and J. Reidl. Item-based collaborative filtering recommendation algorithms. In Proceedings of the 10th International World Wide Web Conference, pages 285-295, 2001.

[27] M.Deshpande and G. Karypis. Item-based top-n recommendation algorithms. ACM Trans. Inf. Syst., 22(1):143-177, 2004.

[28] B.M. Sarwar, G. Karypis, J.A. Konstan, and J. Reidl. Item-based collaborative filtering recommendation algorithms. In Proceedings of the 10th International World Wide Web Conference, pages 285-295, 2001.

[29] M. Brand. Fast online svd revisions for lightweight recommender systems. In Proceedings of the 3rd SIAM International Conference on Data Mining, 2003.

[30] D. Goldbberg, D. Nichols, B. Oki and D. Terry. Using collaborative filtering to weave an information tapestry. Communication of ACM 35, 12:61-70, 1992.

[31] P. Resnik, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. Grouplens: An open architecture for aollaborative filtering of netnews. In Proceedings of the ACM, Conference on Computer Supported Cooperative Work. Pages 175-186, 1994.

[32] J. Konstan, et al. Grouplens: Applying collaborative filtering to usenet news. Communication ofACM 40, 3:77-87, 1997.

[33] U. Shardanand and P. Maes. Social information filtering: Algorithms for automating “word of mouth”. In Proceedings of ACM CHI, 1:210-217, 1995.

[34] G. Kinden, B. Smith and J. York. Amazon.com recommendations: item-to-item collaborative filtering. IEEE Internet Computing, 7:76-80, 2003.

[35] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. Machine Learning Journal 42, 1:177-196, 2001.

[36] T. Hofmann and J. Puzicha. Latent class models for collaborative filtering. In Proceedings of International Joint Conference in Artificial Intelligence, 1999.

[37] http://www.cs.carleton.edu/cs_comps/0607/recommend/recommender/collaborativefiltering.html

[38] E. Y. Chang, et. al., Parallelizing Support Vector Machines on Distributed Machines, NIPS, 2007.

[39] Wen-Yen Chen, Dong Zhang, and E. Y. Chang, Combinational Collaborative Filtering for personalized community recommendation, ACM KDD 2008.

[40] Y. Sun, W.-Y. Chen, H. Bai, C.-j. Lin, and E. Y. Chang, Parallel Spectral Clustering, ECML 2008.