Scalable Collaborative Filtering for Mining Social Networks

Edward Chang

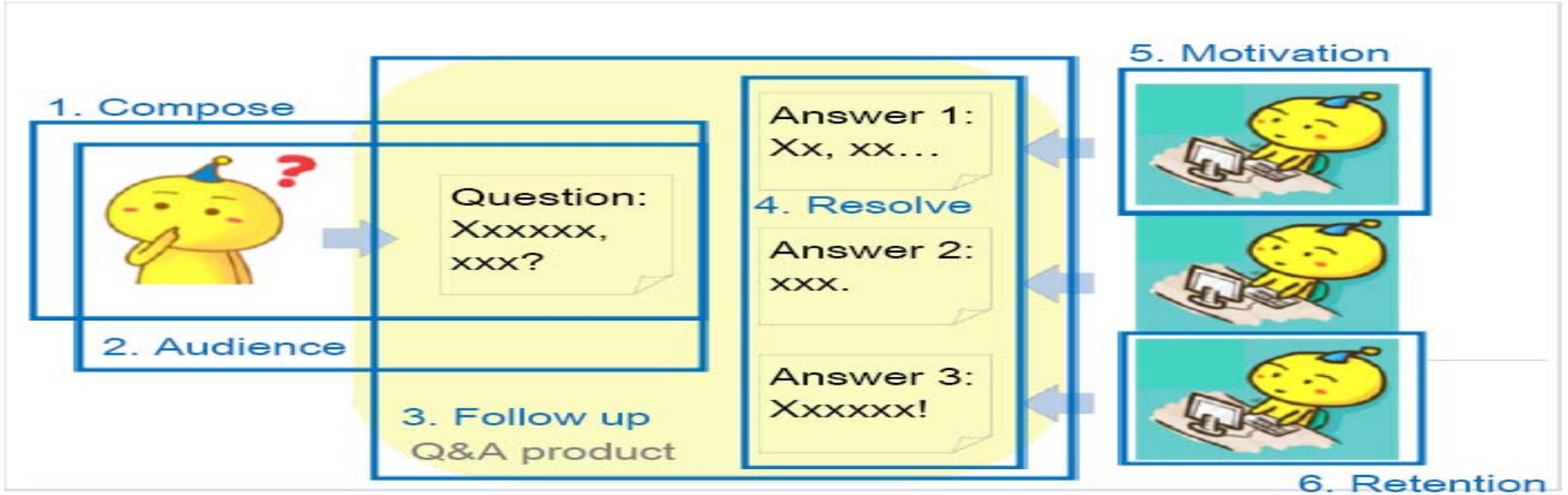
Google Research, Beijing

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

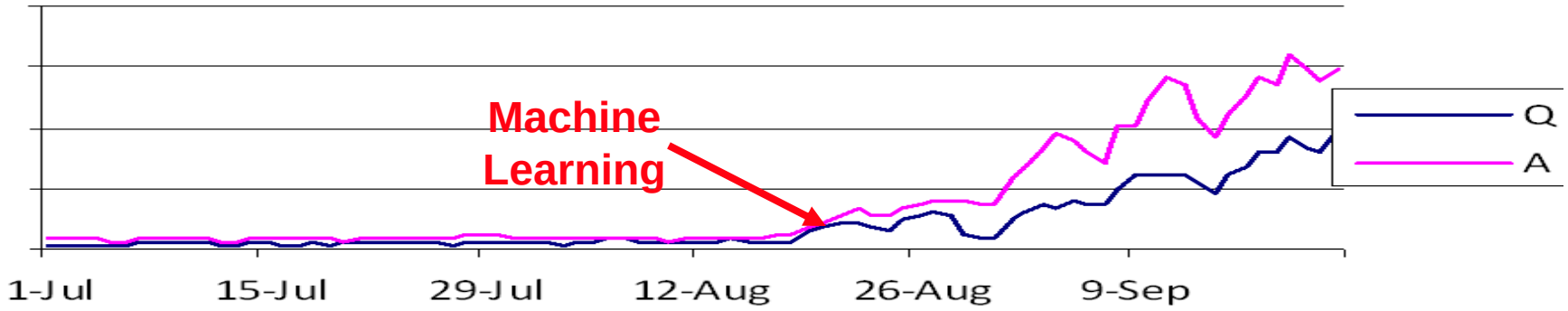
Collaborators

- 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)
- Confucius Team led by Jim Deng (Google Beijing)
- OpenSocial Team led by David Glazer (Google MTV)

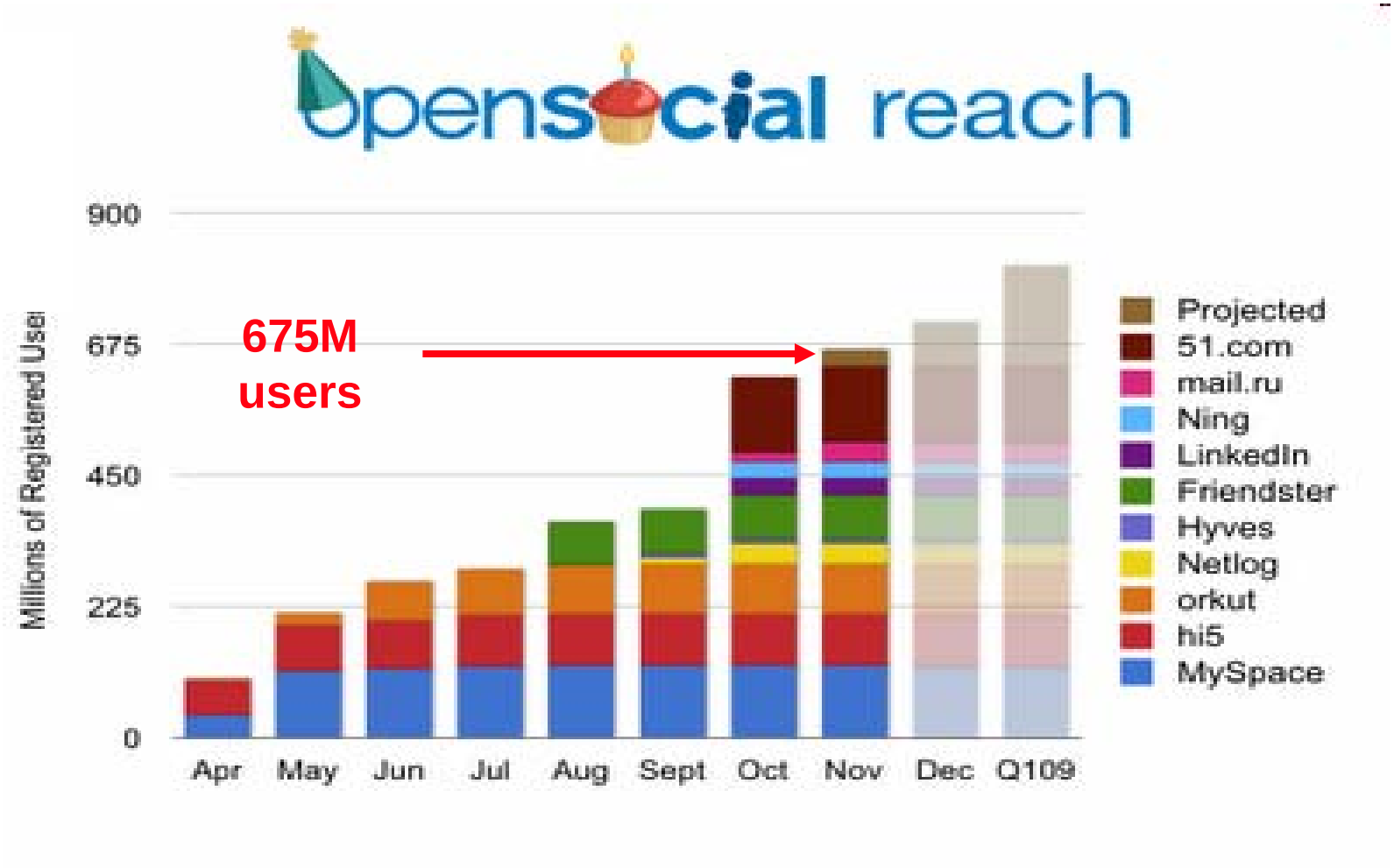
Confucius, a Q&A System



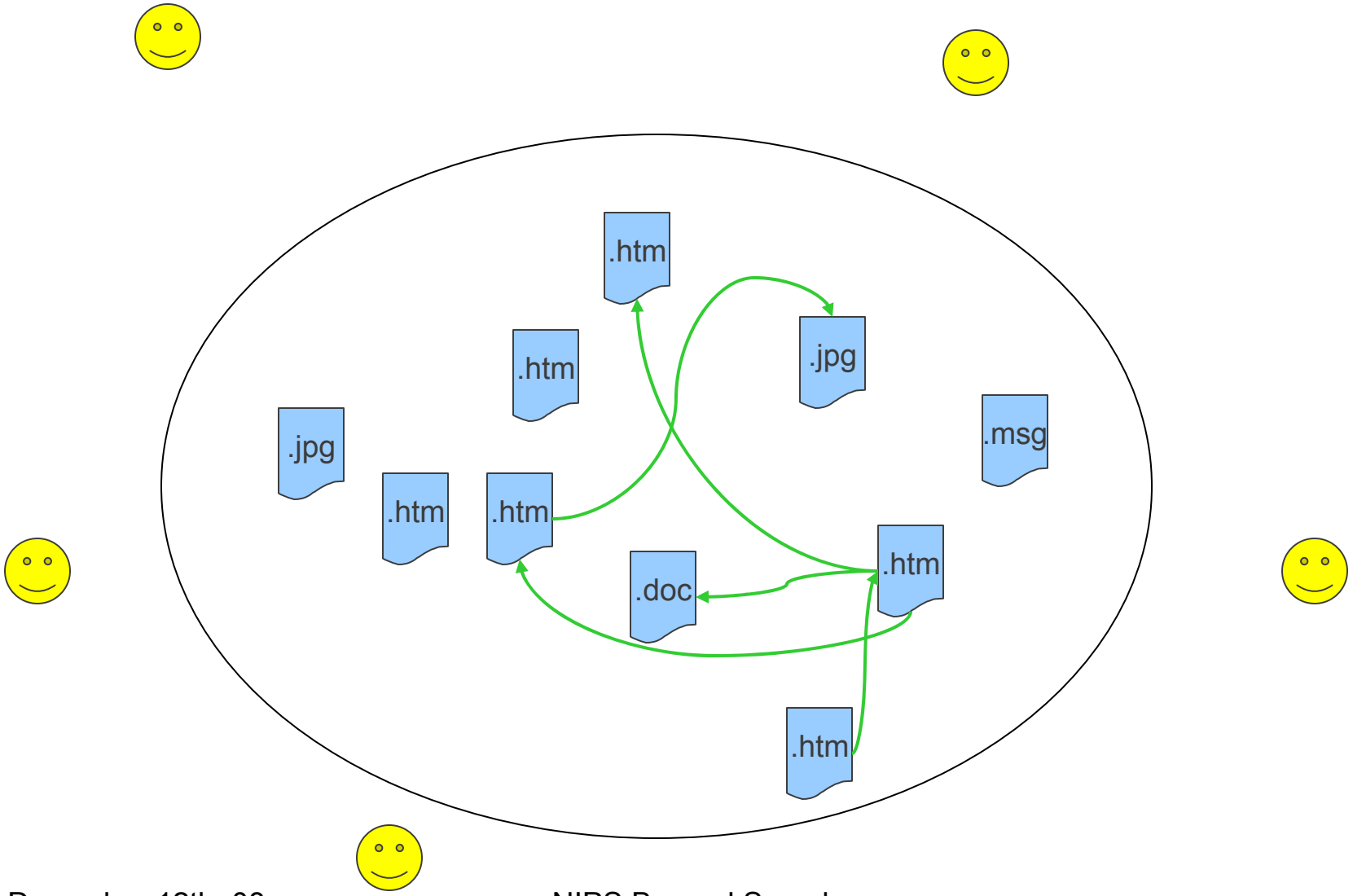
Confucius Growth



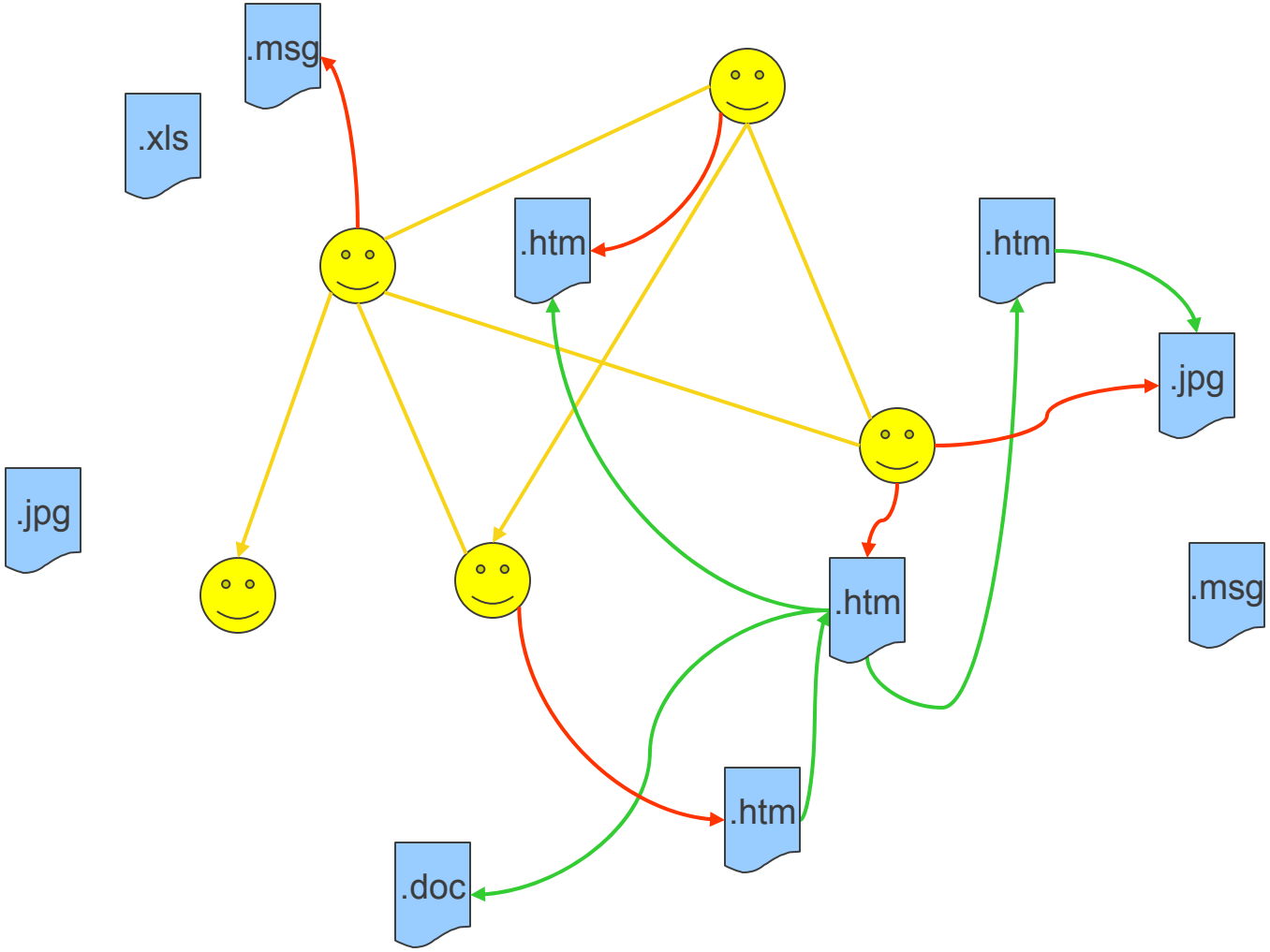
OpenSocial


opensocial reach
900
675
450
225
0
Millions of Registered Users
675M users
Apr
May
Jun
Jul
Aug
Sept
Oct
Nov
Dec
Q109
Projected
51.com
mail.ru
Ning
LinkedIn
Friendster
Hyves
Netlog
orkut
hi5
MySpace

Web 1.0

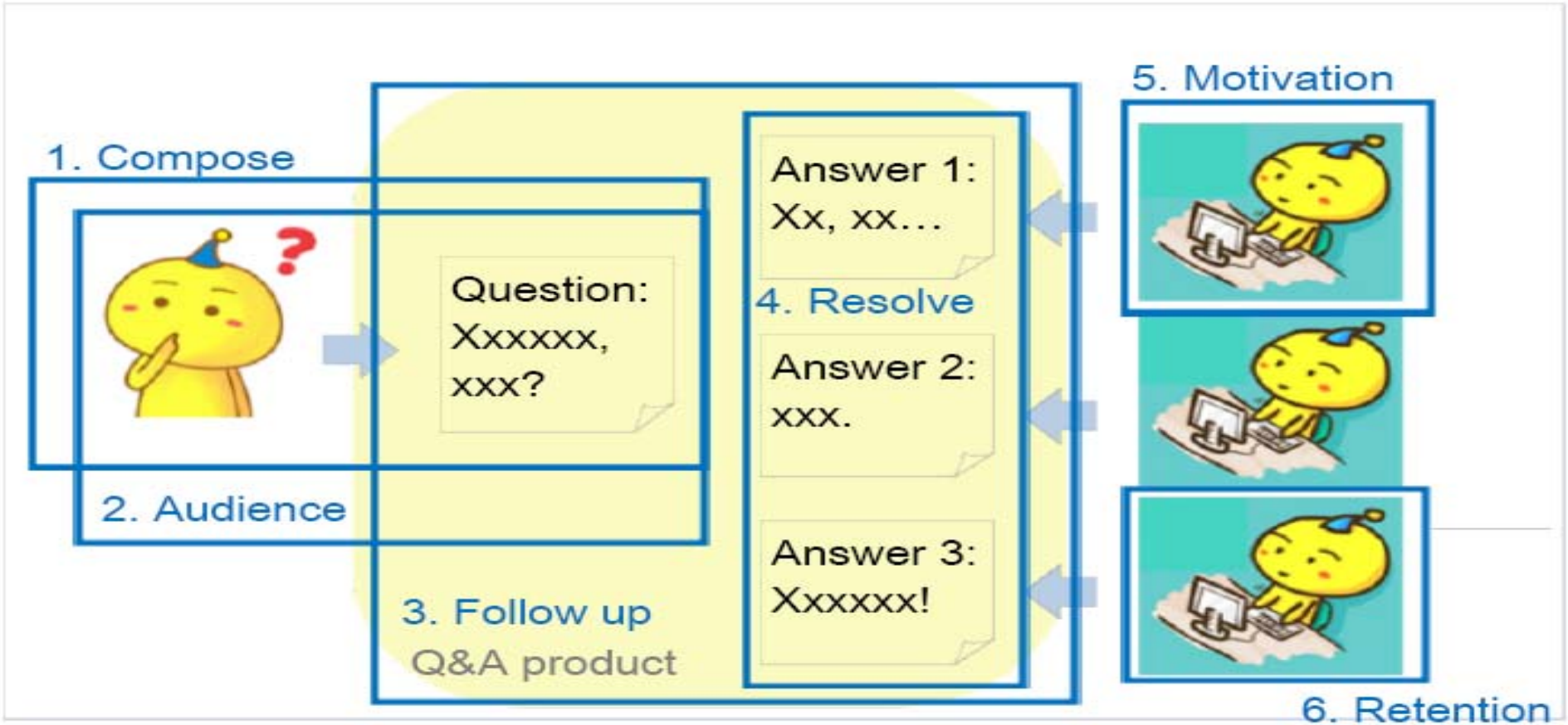


Web 2.0 --- Web with People

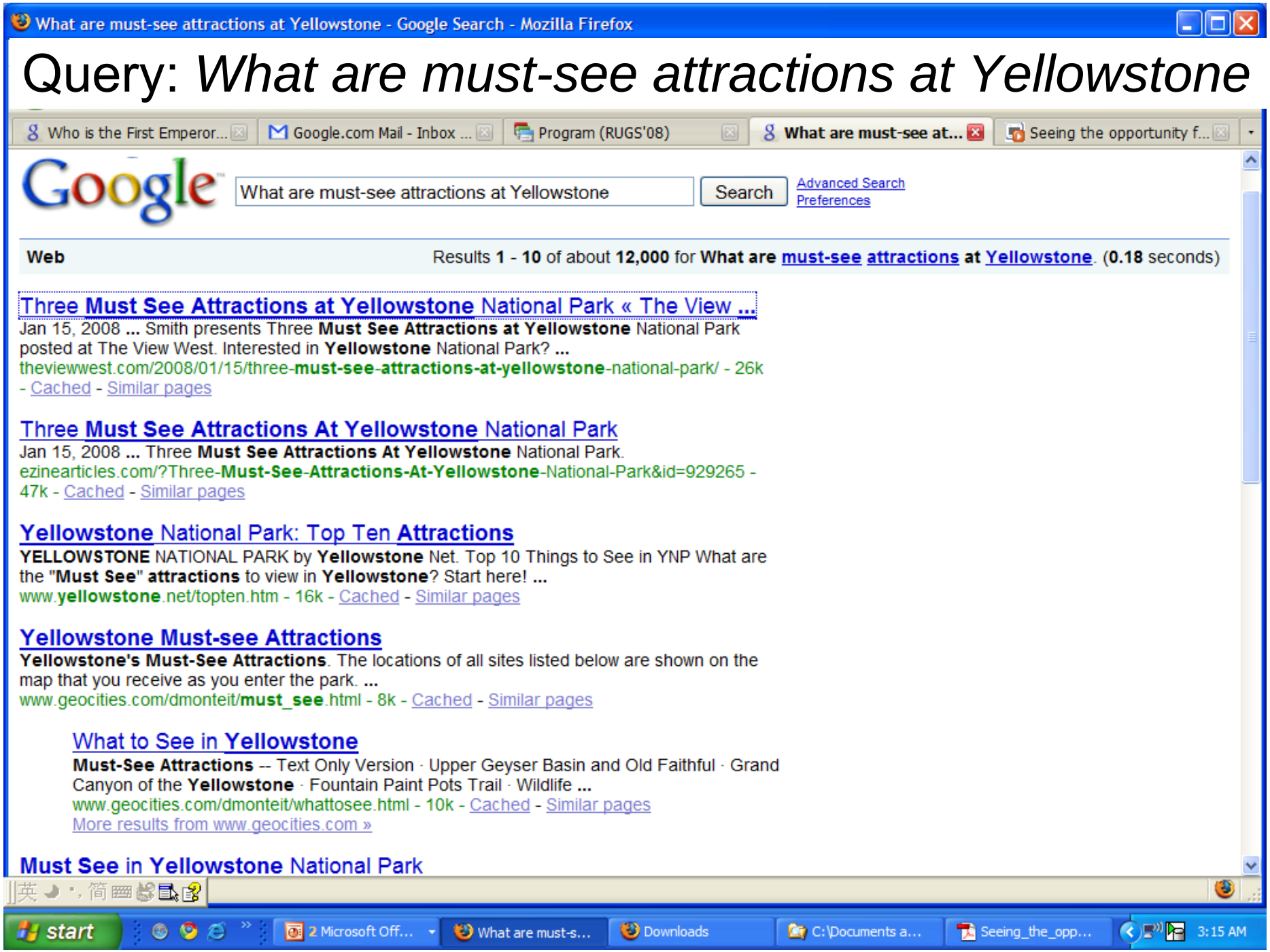


Confucius, a Q&A system

- Allowing people to ask questions for information that cannot be found by Web search



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



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

Three Must See Attractions at Yellowstone National Park « The View West - Mozilla Firefox

Three Must See ... | Google.com Mail - ... | Steven Baker's Go... | Program (RUGS'08) | Electrical & Compu... | Seeing the opport...

At first glance, Mammoth Hot Springs appear as a frozen waterfall. Large terraces abound while being connected by trickling water. The hot acidic water from the thermal aspect below ascends through ancient limestone deposits in the area. As the water dissolves the limestone, it is carried to the surface. When the suspension cools and becomes less acidic at the surface it forms the pools and the cascading features. This area is truly an amazing and dynamic work of art.

Wildlife

- The Church of Jesus Christ of Latter Day Saints
- The View West Bookstore
- WordPress.com
- WordPress.org

ARCHIVES

- May 2008 (1)
- March 2008 (1)
- February 2008 (15)
- January 2008 (19)

BLOG STATS

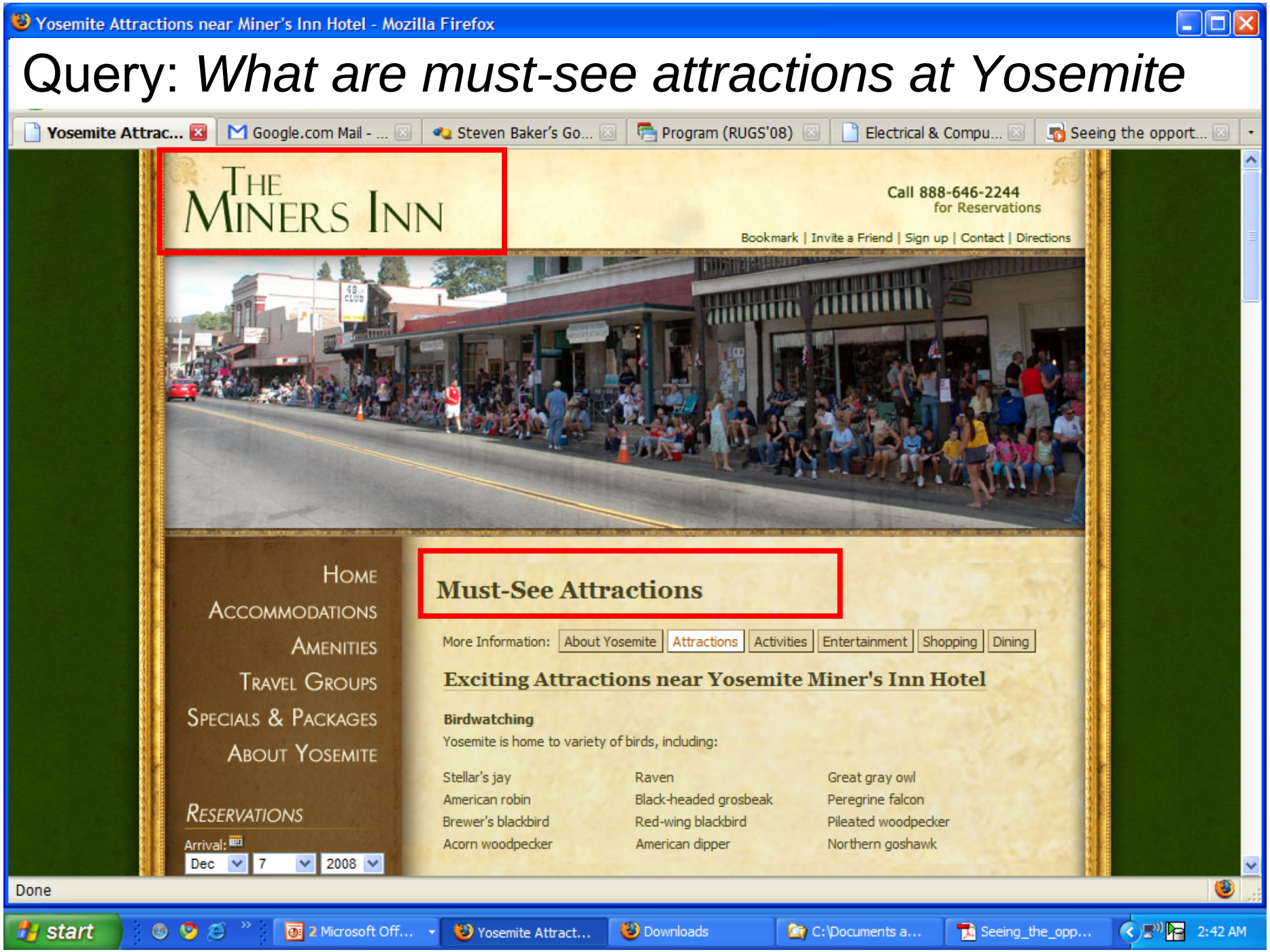
- 4,702 hits

TAGS

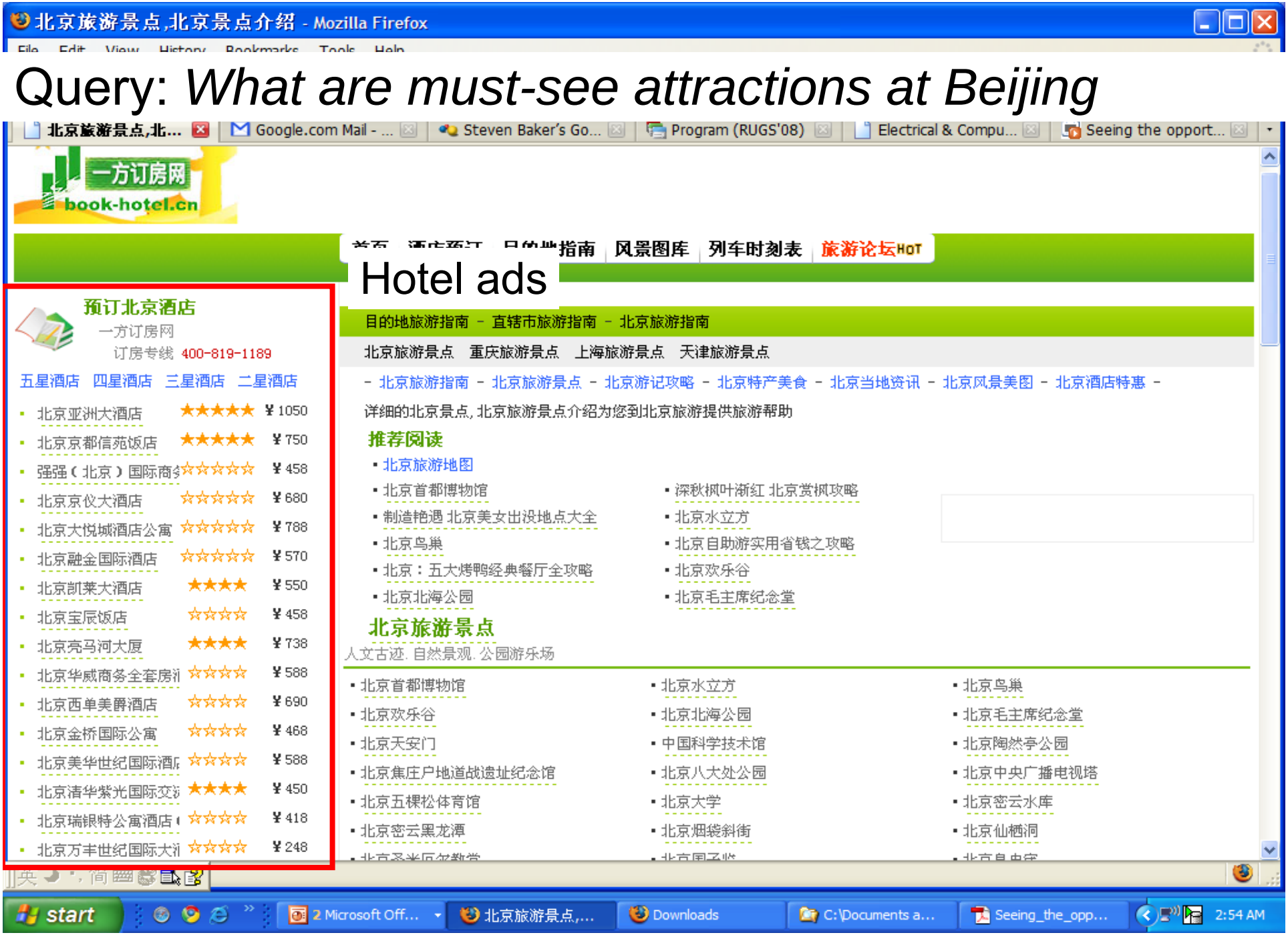
Avalanche avalanche deaths avalanche fatalities baseball Bill Richardson bonneville dam Book Reviews California budget California Deficit education cuts Election 2008 full day kindergarten geysers goose gossage gossage governor Schwarzenegger hall of fame highway 66 idaho snow jaycee carroll kindergarten lava dome LDS church montana avalanche Mount St.

start | 2 Microsoft Off... | Three Must See ... | Downloads | C:\Documents a... | Seeing_the_opp... | 2:40 AM

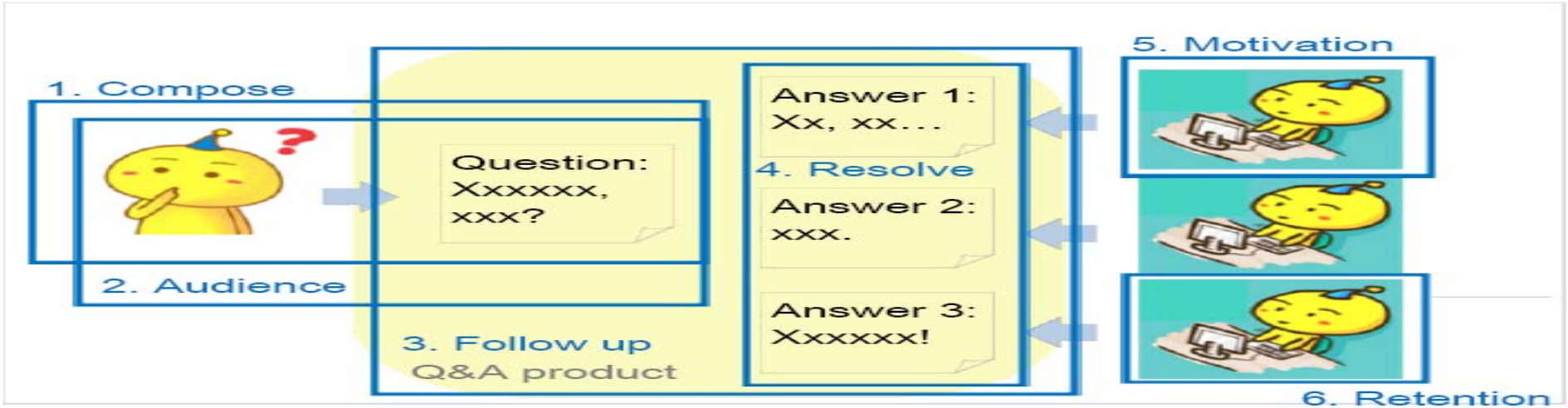
Query: *What are must-see attractions at Yosemite*



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



Key ML Subroutines of Confucius



- ❑ Trigger a question session during search
- ❑ Given a question, provide labels for easy organization
- ❑ Given a question, find similar questions and their answers
- ❑ Evaluate user credentials in a domain sensitive way
- ❑ Given a question, route it to domain experts
- ❑ Evaluate quality of answers to a question
- ❑ Machine-generated answers

Naive User Evaluation

Point system based on *hand-crafted rules*:

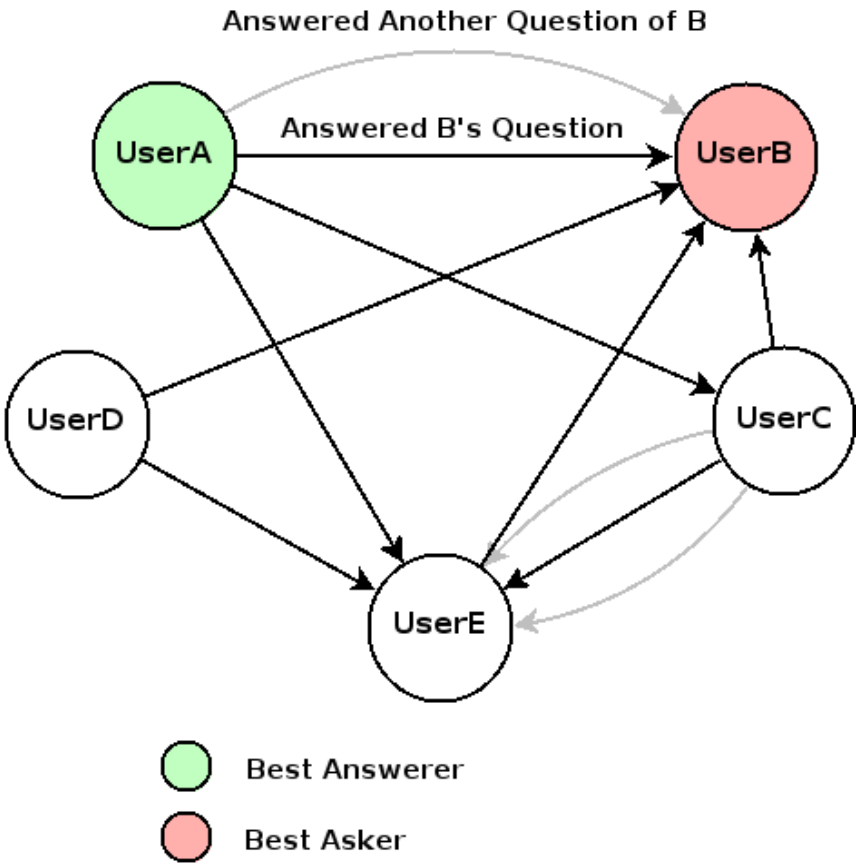
- registration → +100 points
- each time login → +5 points
- ask one question → +bonus points
- ask one question → +2 points
- vote on one answer → +1 points
- best answers → +bonus points
- …

Shortcomings

- Easily Spammed
  - Mutual enforcement, answer “friends” questions

    →1,000 IDs of the same person
  - Copy & paste others’ answers
  - Advertising posts
- Freshness
  - User's recent activities are not emphasized

Link-based User Credential Ranking: HITS



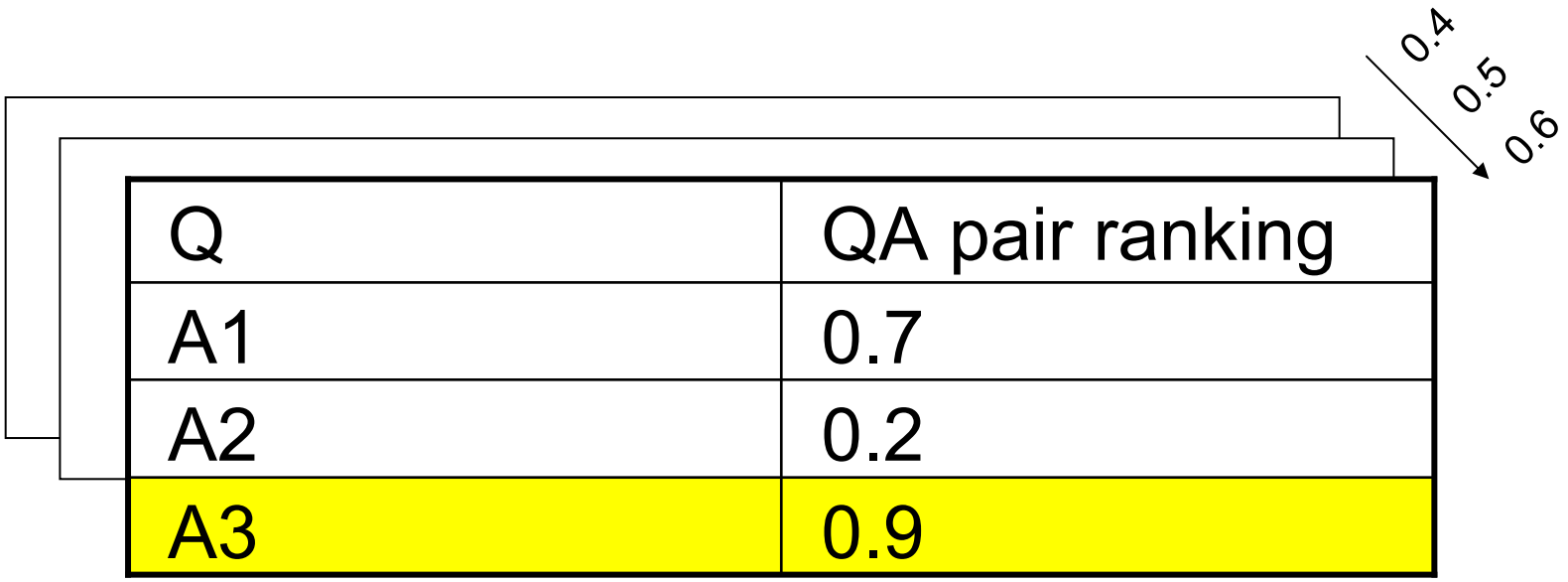
QA pairs → User Relation
Ranking user using HITS*

	In links	Out Links
A	0	4 (3)
B	4	0
C	1	4 (2)
D	0	2
E	5 (3)	1

* HITS is based on Zoltan et al, *Questioning Yahoo Answers*. QAWeb, WWW2008

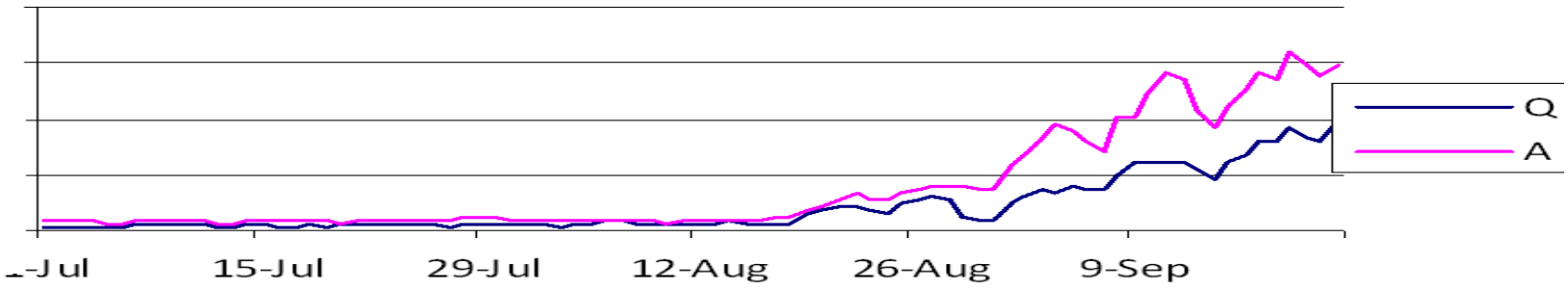
Q&A/Blog/BBS Search

- Lack of links
- Links can be easily spammed
- User credential can help ranking

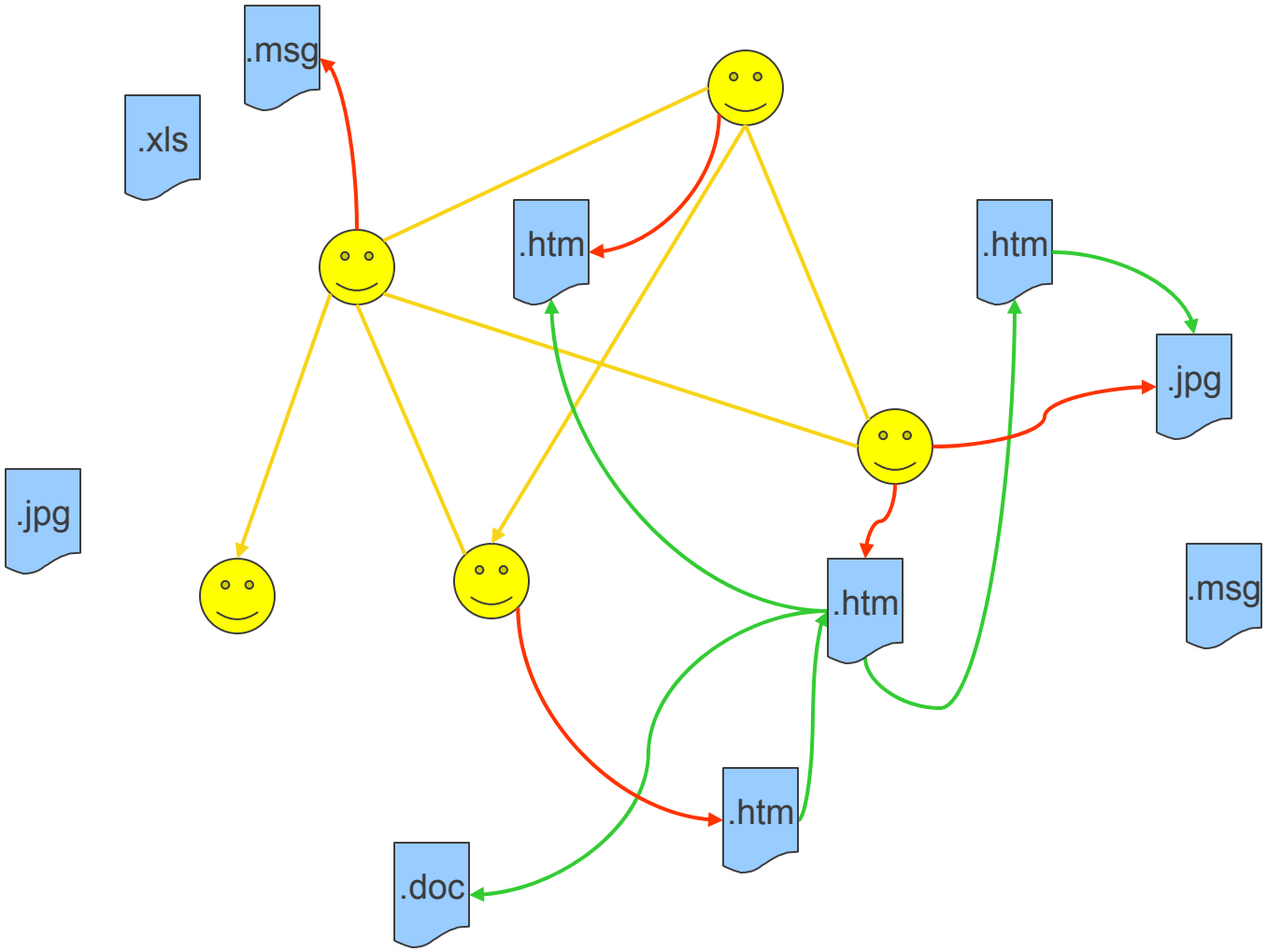


Data Mining
Impact & Opportunities

Confucius Growth

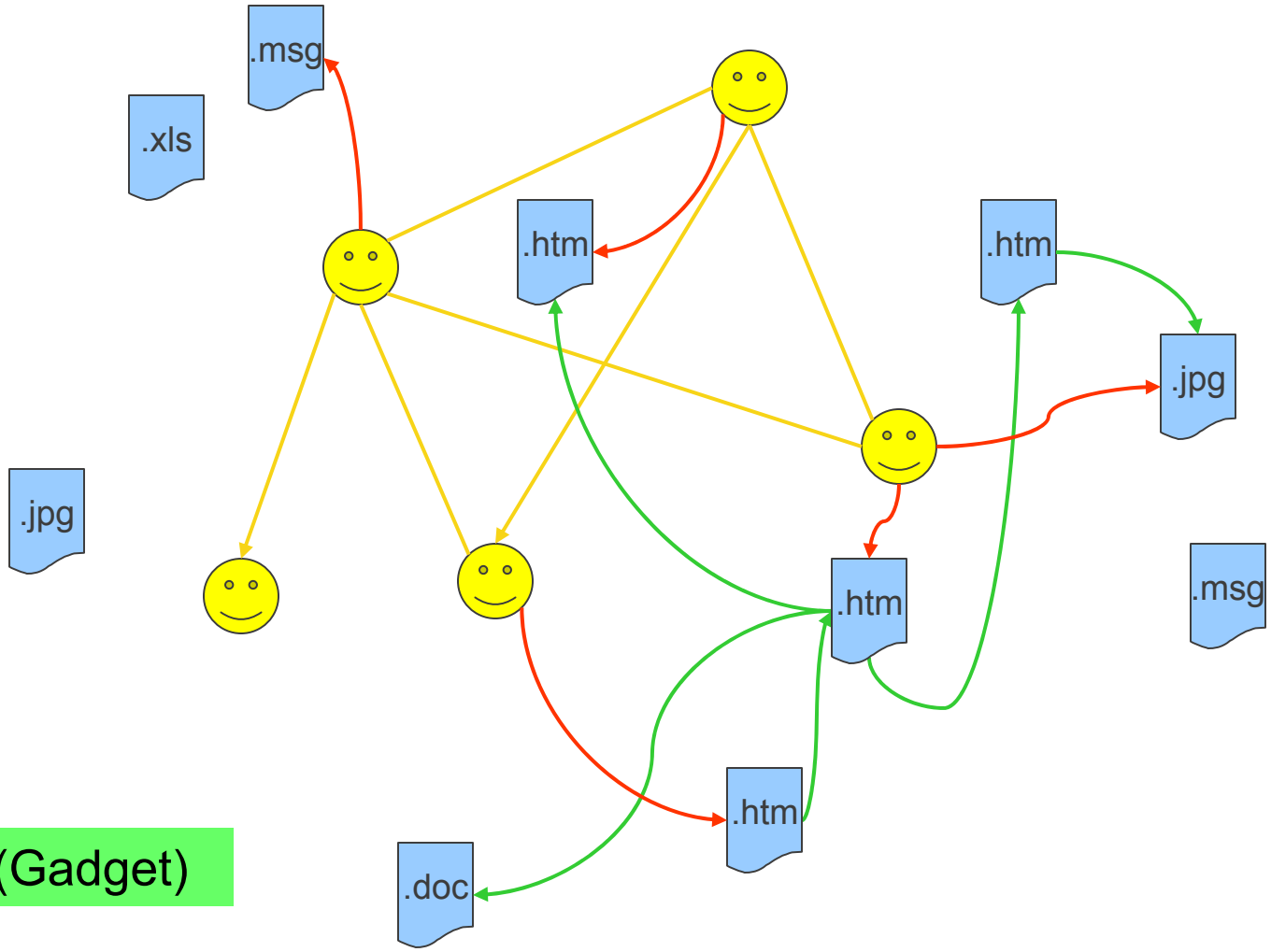


Web 2.0 --- Web with People



+ Social Platforms

App (Gadget)



App (Gadget)

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 -... | 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
易绍
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»

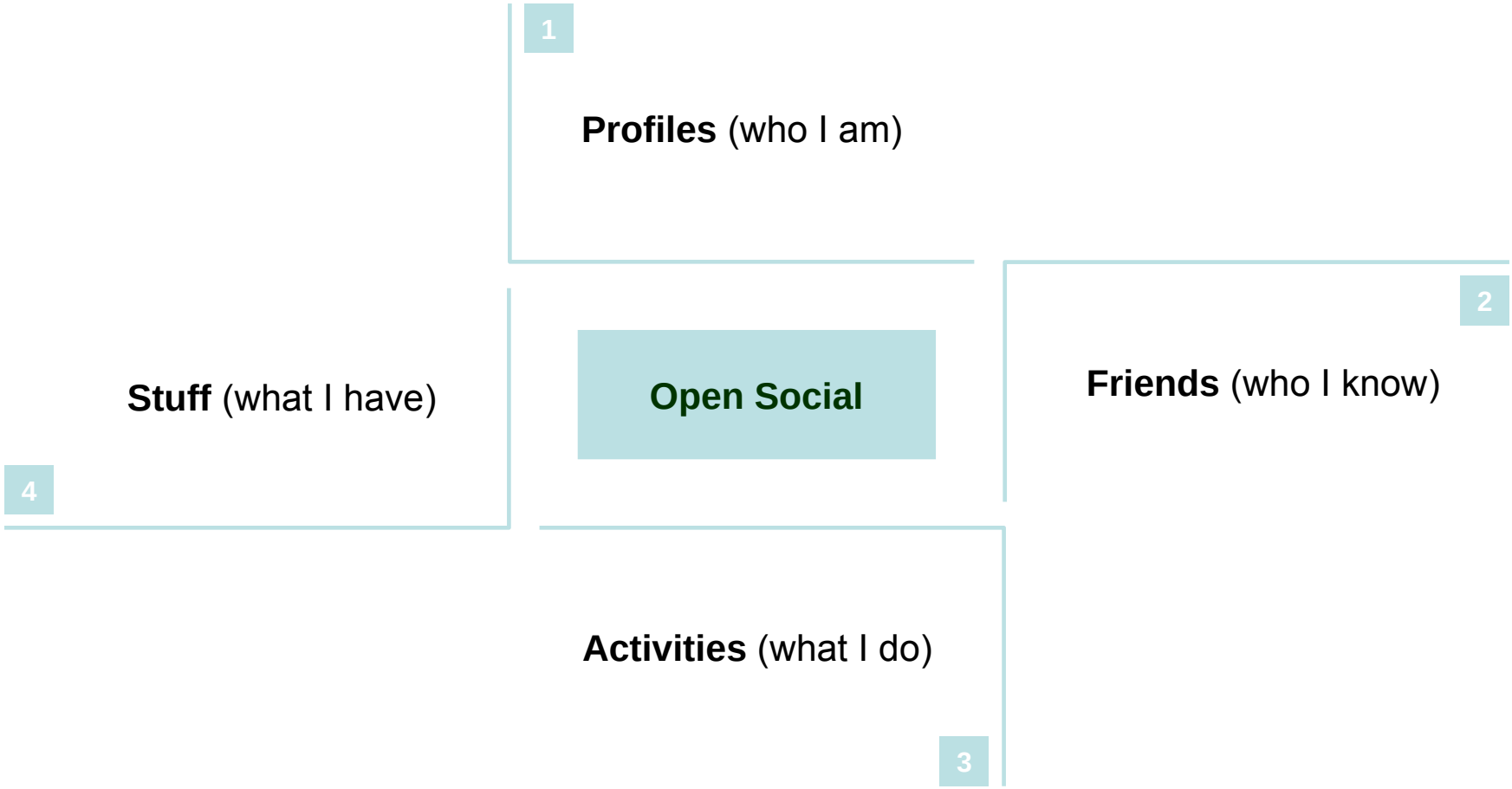
看看我的朋友在哪

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
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
机械舞
暗香浮月
水星cici
梦侠人
无聊最寂寞
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
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

What users are interested in?



between religiousness of a country and its progress.

I love India. It bothers me, therefore, that the society seems to be much more religious now then in 1960s when I was growing up there. Yes, India has made good progress and possibly economically India is in the best situation now than any other time in the last 200 years. But when one thinks at what has happened in many other Asian countries (not only China, but also in Korea, Taiwan, Singapore, Malayasia, and soon in Vietnam) and compares current Indian situation to what could be, it becomes very depressing. And sitting in this wonderful lounge at Beijing Airport, and comparing this to the lounges in the Mumbai or Delhi Airports, this thought is obvious.

TECHNICAL THOUGHTS, GENERAL UPDATES | 1 COMMENT »

Beijing Trip

Posted by Ramesh on December 7th, 2008

The last 3 days I have been in Beijing to attend SKG2008. I was requested to give a keynote talk at this conference — Semantics, Knoledge, and Grid — and I talked about the EventWeb ideas.

Though I came here only after about 7 months, this trip showed me a bit more of how rapidly China is transformed. It does not feel like a developing country — all the facilities and the infrastructure makes it look better than many developed countries. Of course, people tell me that once you go away from a few top places like Beijing and Shanghai, the story is different. Even if that is the case, what China has accomplished seems to be unparalleled in the history. Being Indian, it is natural for me to think about India and I feel very depressed about India. IN fact I feel worried a bit even about USA

Archives
December 2008
November 2008
October 2008
September 2008
August 2008
July 2008
June 2008
May 2008
April 2008
March 2008
February 2008
January 2008
December 2007
November 2007
October 2007
September 2007
August 2007
July 2007
June 2007
May 2007
April 2007
March 2007
February 2007
January 2007
December 2006
November 2006
October 2006
September 2006
August 2006
July 2006

天涯来吧 - eyuchang的相册 - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://laiba.tianya.cn/laiba/AlbumListL?uid=14914947603760770386

Google

来吧主页 > 我的相册

我的主页 资料 朋友 来吧 帖子 相册 日记 礼物New! 评价 留言簿

上传照片

您已经使用了1024MB中的2MB (0%)

我的相册(8) 我的相片收藏

北京研究会 (18)

点击 1014 2008-4-29

北京过年 ... (6)

点击 1395 2008-2-13

天涯谷歌会议 (12)

点击 1405 2008-1-28

三亚 Goo... (12)

点击 1414 2008-1-19

绿色网络生活 (4)

点击 1331 2007-12-29

成都 De... (6)

点击 1361 2007-12-29

Europe ... (4)

点击 2864 2007-10-24

默认相册 (0)

点击 1343

我朋友的相册

panjie...的相册

18照片

calla_qu的相册

1296照片

磨刀美眉的相册

2照片

天涯游的相册

85照片

三皮儿的相册

61照片

钟巍巍的相册

682照片

01liux...的相册

75照片

efatao...的相册

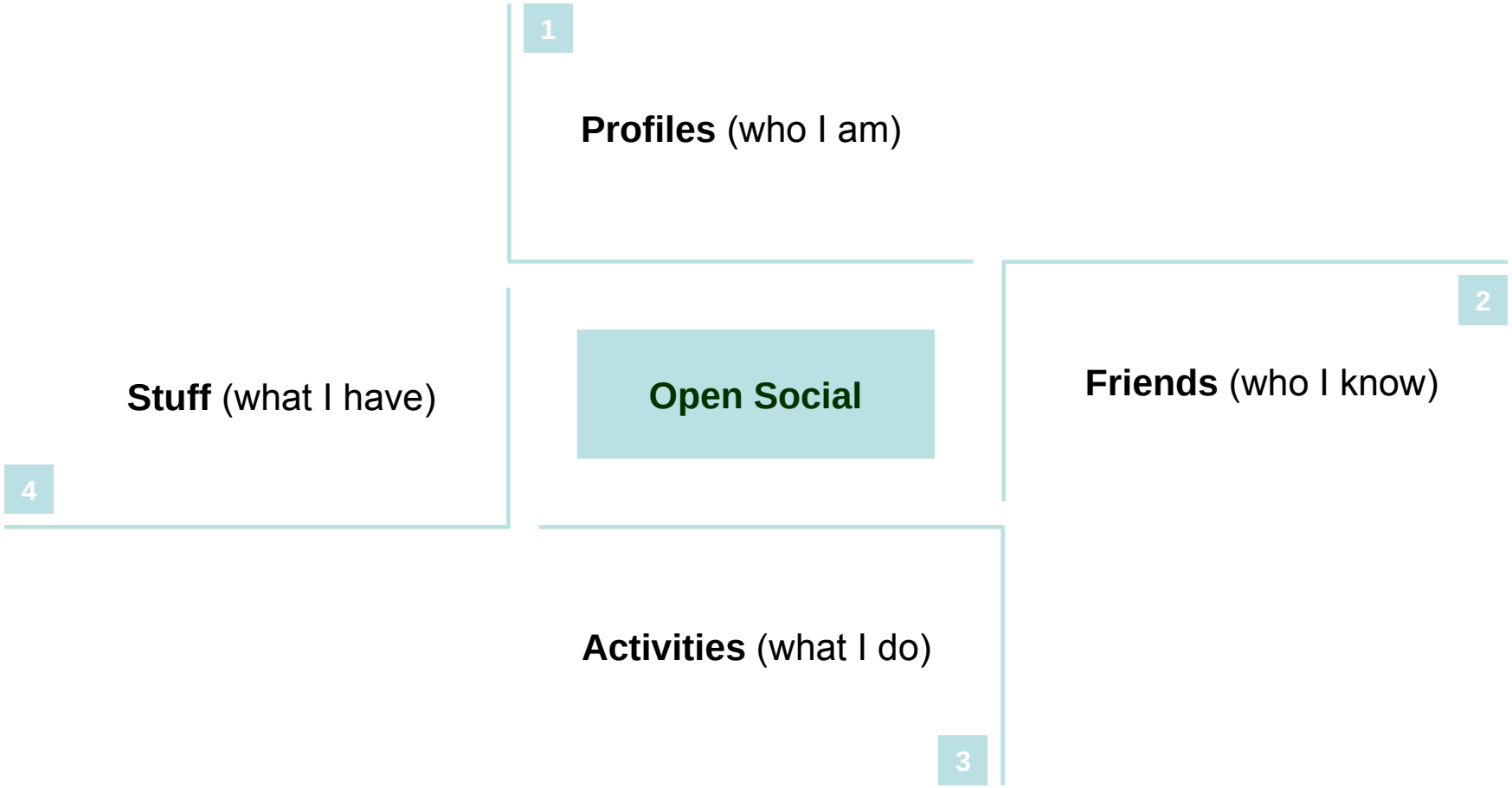
9照片

开复的相册

http://laiba.tianya.cn/laiba/AlbumL?uid=14914947603760770386&aid=UP0000000677/00/0946732164-0000000677&xid=11775521930820846062

start 4 Windows E... Document1 - M... Downloads 天涯来吧 - ey... 2 Microsoft O... EN 3:52 AM

Open Social APIs

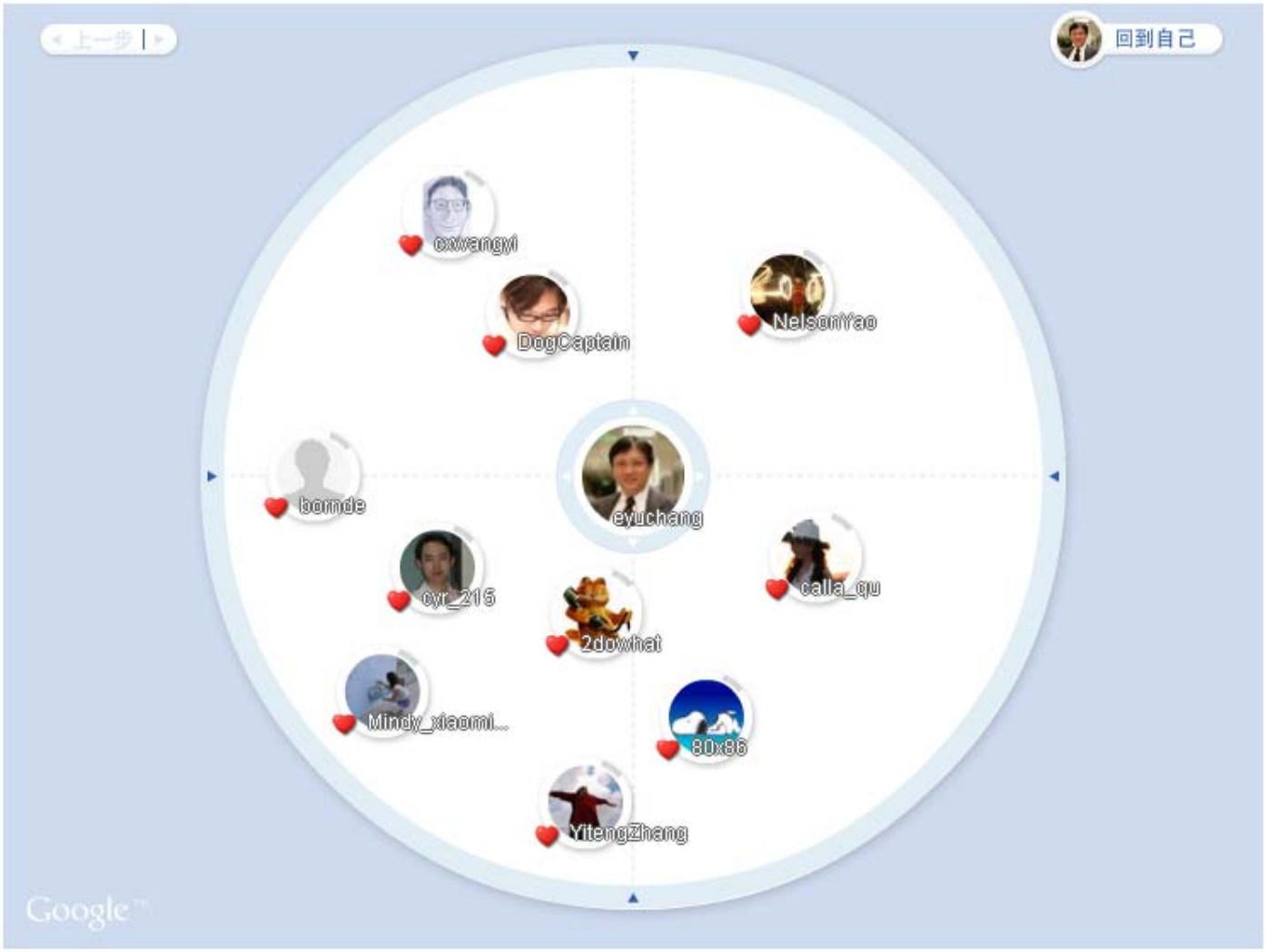


□□□□ - □□□□□ - Mozilla Firefox

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**



邀请朋友加入来吧

邮件 *

姓名

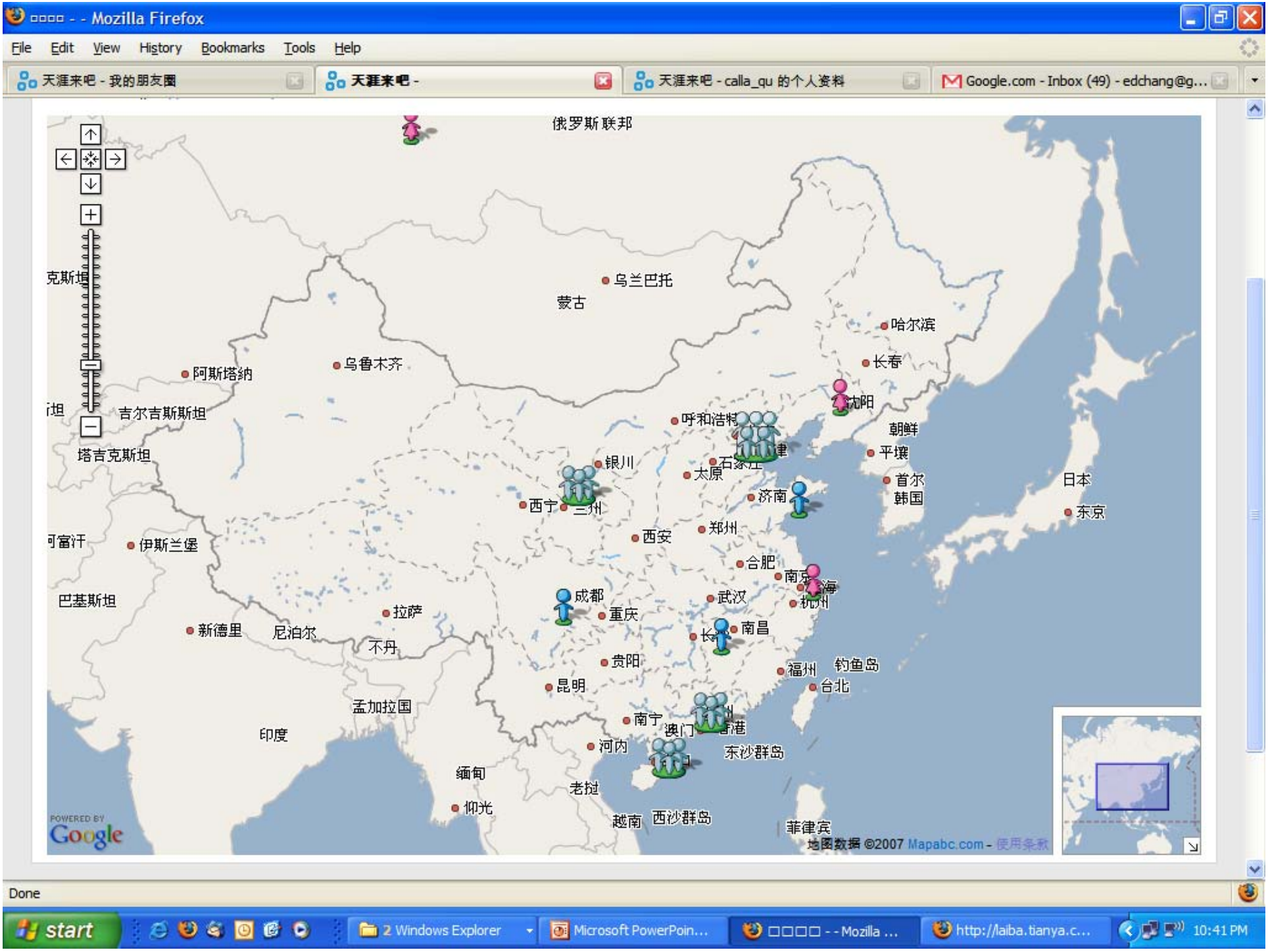
发送邀请

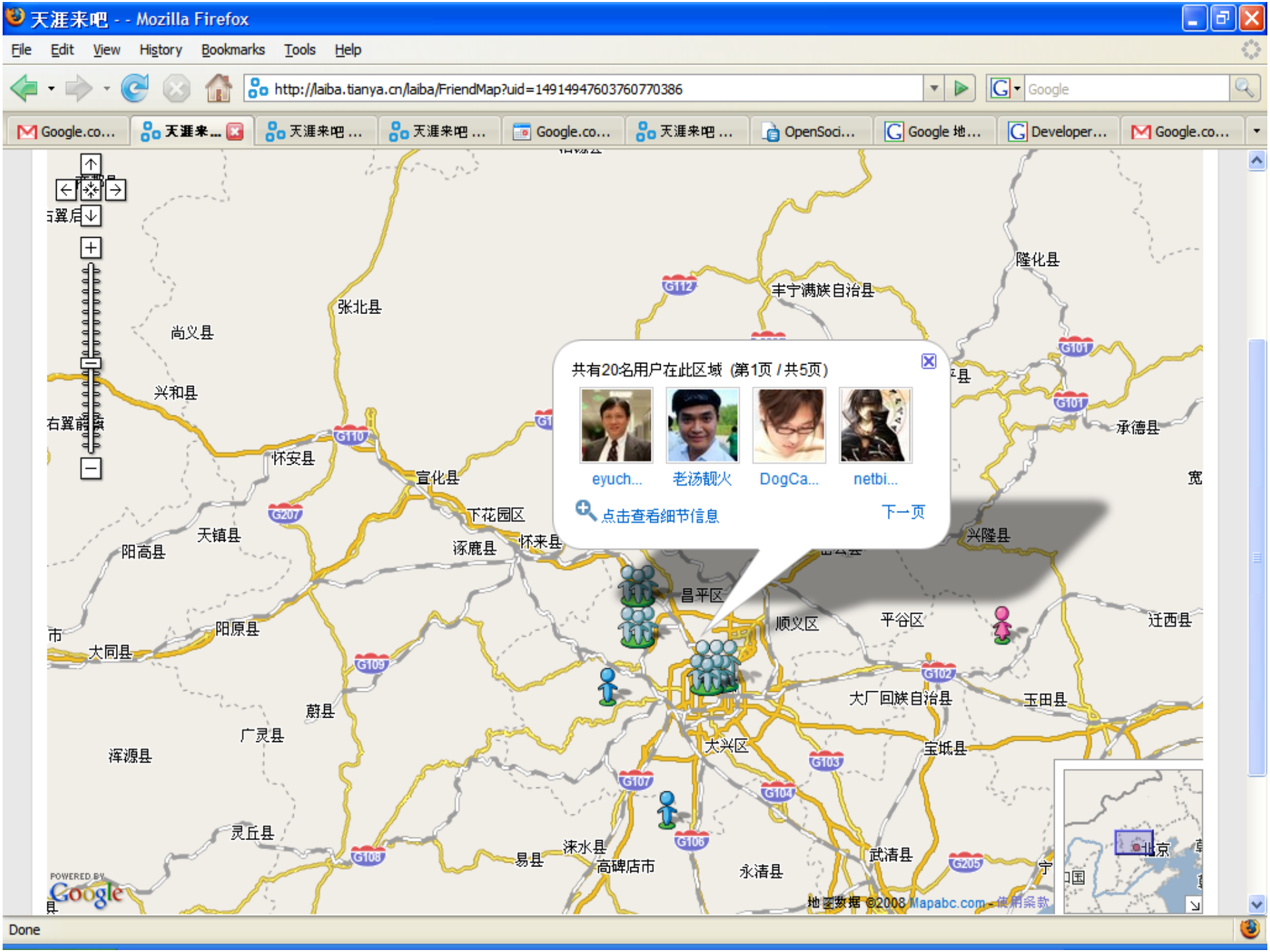
我要群发邀请»

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
朝鲜
平壤
首尔
韩国
日本
东京
乌鲁木齐
阿斯塔纳
吉尔吉斯斯坦
塔吉克斯坦
呼和浩特
银川
太原
西宁
济南
郑州
西安
伊斯兰堡
合肥
南京
杭州
武汉
成都
重庆
南昌
拉萨
新德里
尼泊尔
不丹
贵阳
福州
钓鱼岛
台北
昆明
孟加拉国
印度
南宁
澳门
香港
河内
东沙群岛
缅甸
老挝
仰光
越南
西沙群岛
菲律宾
巴基斯坦
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

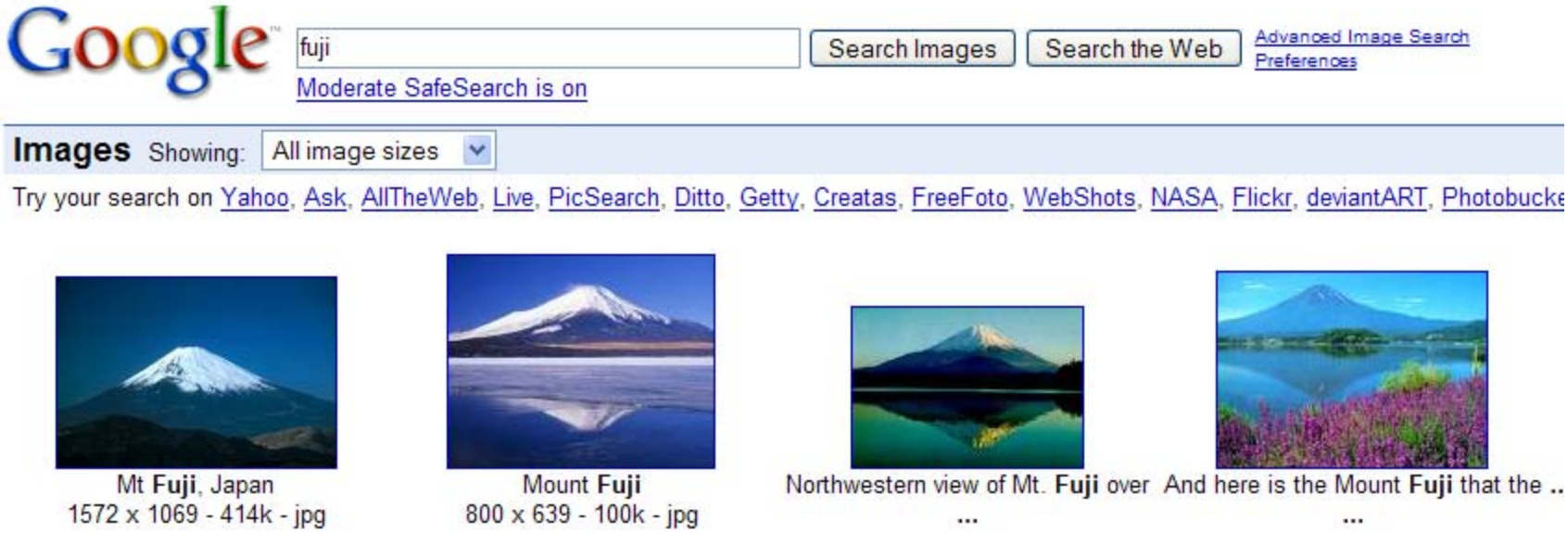

天涯来吧 - - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://laiba.tianya.cn/laiba/FriendMap?uid=14914947603760770386
Google
Google.co...
天涯来...
天涯来吧 ...
天涯来吧 ...
Google.co...
天涯来吧 ...
OpenSoci...
Google 地...
Developer...
Google.co...
共有20名用户在此区域 (第1页 / 共5页)
eyuch...
老汤靓火
DogCa...
netbi...
点击查看细节信息
下一页
张北县
尚义县
兴和县
右翼前旗
丰宁满族自治县
隆化县
承德县
怀安县
宣化县
下花园区
天镇县
阳高县
涿鹿县
怀来县
兴隆县
昌平区
顺义区
平谷区
迁西县
阳原县
大同县
蔚县
广灵县
浑源县
大兴区
大厂回族自治县
玉田县
宝坻县
灵丘县
涞水县
易县
高碑店市
永清县
武清县
北京
POWERED BY Google
地图数据 ©2008 Mapabc.com - 使用条款
Done


orkut
LinkedIn
hi5
salesforce.com
OpenSocial

Personalized Search Example

- Infer relevance through social networks

- Query “fuji” can return
  - Fuji mountain
  - Fuji apples
  - Fuji cameras


Google
fuji
Search Images
Search the Web
Advanced Image Search
Preferences
Moderate SafeSearch is on
Images
Showing:
All image sizes
Try your search on Yahoo, Ask, AllTheWeb, Live, PicSearch, Ditto, Getty, Creatas, FreeFoto, WebShots, NASA, Flickr, deviantART, Photobucke
Mt Fuji, Japan
1572 x 1069 - 414k - jpg
Mount Fuji
800 x 639 - 100k - jpg
Northwestern view of Mt. Fuji over ...
And here is the Mount Fuji that the ..
...


Google
fuji
Search Images
Search the Web
Advanced Image Search
Preferences
Moderate SafeSearch is on
Images
Showing:
All image sizes
Try your search on Yahoo, Ask, AllTheWeb, Live, PicSearch, Ditto, Getty, Creatas, FreeFoto, WebShots, NASA, Flickr, deviantART, Photobuck
(Apples, Fuji) Fuji apples are an ...
765 x 792 - 37k - jpg
www.all-creatures.org
fuji apple
300 x 294 - 17k - jpg
www.wisegeek.com
Organic - Apples, Fuji
375 x 375 - 67k - jpg
www.cleanfoodconnection.com
fuji apple Manufacturer
800 x 600 - 81k - jpg
www.supplierlist.com

Google fuji Search Images Search the Web Advanced Image Search Preferences

Moderate SafeSearch is on

Images Showing: All image sizes

Try your search on Yahoo, Ask, AllTheWeb, Live, PicSearch, Ditto, Getty, Creatas, FreeFoto, WebShots, NASA, Flickr, deviantART, Photobucket

... "as is typical of **Fuji cameras** ...
400 x 400 - 78k - jpg
www.livingroom.org.au

fujifilm digital **camera**, digtal, ...
464 x 254 - 13k - jpg
www.fujifilm-cameras.com

Fuji cameras, one with face ...
425 x 313 - 35k - jpg
www.gadgetell.com

Fuji fujifilm finepix A800
425 x 290 - 34k - jpg
www.gadgetell.com

□□□□ - □□□□□ - 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

jifilm digital **camera**, digtal, ...
464 x 254 - 13k - jpg
www.fujifilm-cameras.com

Fuji cameras, one with face ...
425 x 313 - 35k - jpg
www.gadgetell.com

Fuji Camera
450 x 333 - 440k - bmp
emergencygadget.com

cheap **fuji** digital **camera**
400 x 318 - 14k - jpg
cheap-digital-camera.com.au

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

Personalized Recommendation

Recommendation Systems

- Photo/Video Recommendation
- Friend Recommendation
- Community/Forum Recommendation
- Ads Matching

- Performance Requirements
  - Scalability, scalability, scalability

Outline

- Applications
  - Confucius
  - OpenSocial
- Key Subroutines for Mining Massive SNS
  - Clustering [ECML 08]
  - Frequent Itemset Mining [ACM RS 08]
  - Combinational Collaborative Filtering [KDD 08]
    - with PLSA
    - with LDA
  - Support Vector Machines [NIPS 07]
- Distributed Computing Perspectives

Outline

- Applications
  - Confucius
  - OpenSocial
- Key Subroutines
  - Clustering [ECML 08]
  - Frequent Itemset Mining (FIM)
  - Combinational Collaborative Filtering
    - with PLSA
    - with LDA
  - Support Vector Machines
- Distributed Computing Perspectives

Task: Targeting Ads at SNS Users

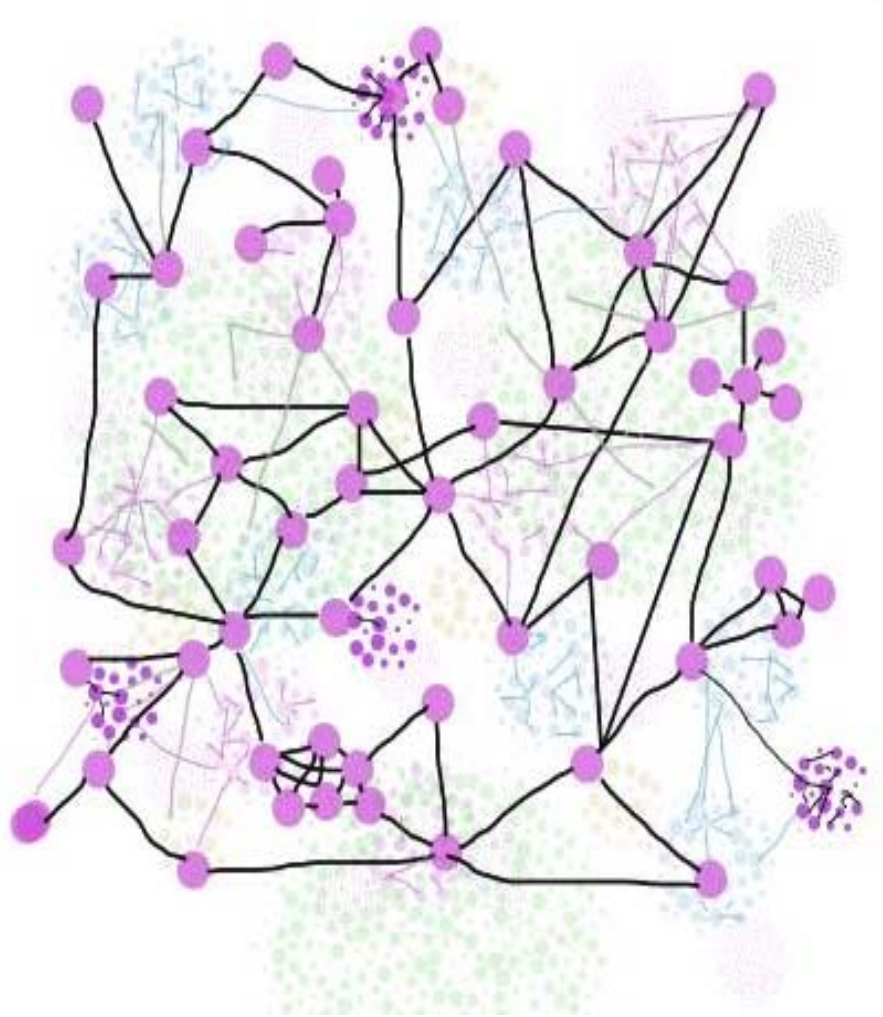
Users

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.



Outline

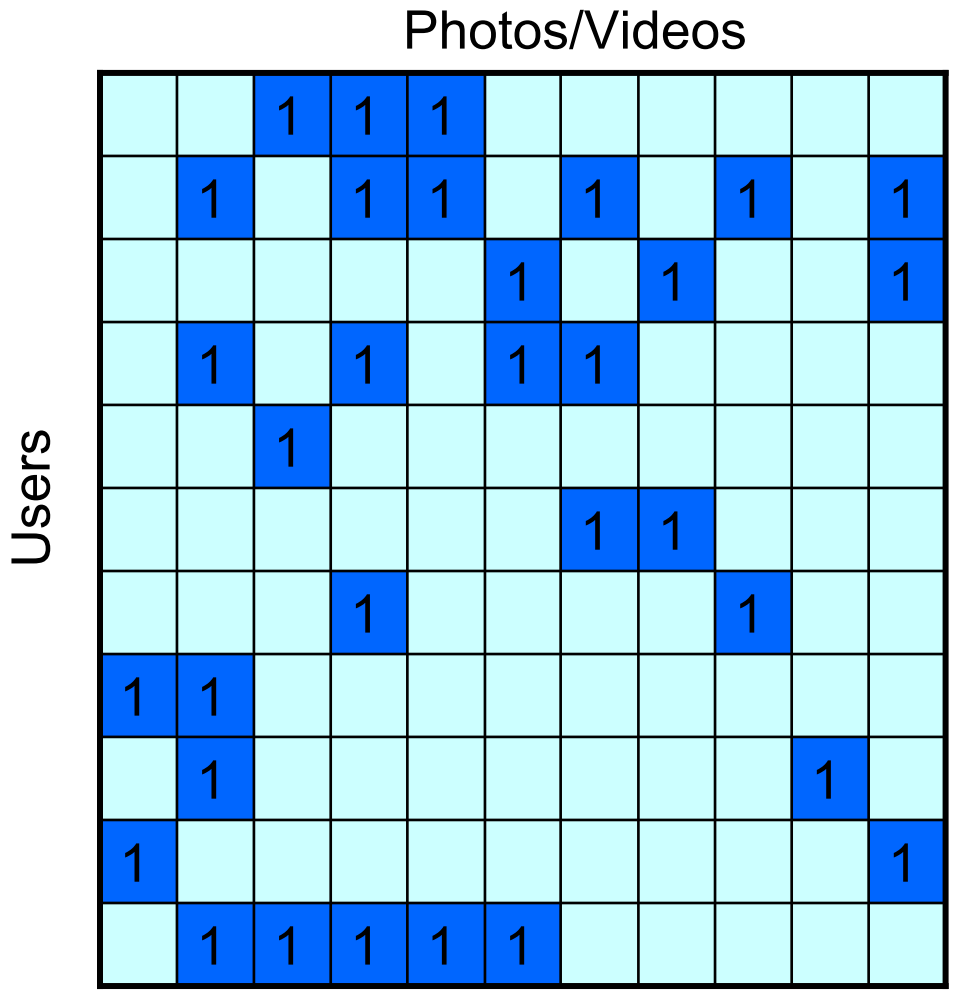
- Emerging Applications
  - Social networks
  - Personalized Information retrieval
- Key Subroutines
  - Clustering [ECML 08]
  - Frequent Itemset Mining (FIM)
  - Combinational Collaborative Filtering
    - with PLSA
    - with LDA
  - Support Vector Machines

Collaborative Filtering

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

↓

Predict further *membership*



Some Queries

Based on *partially* observed matrix

↓

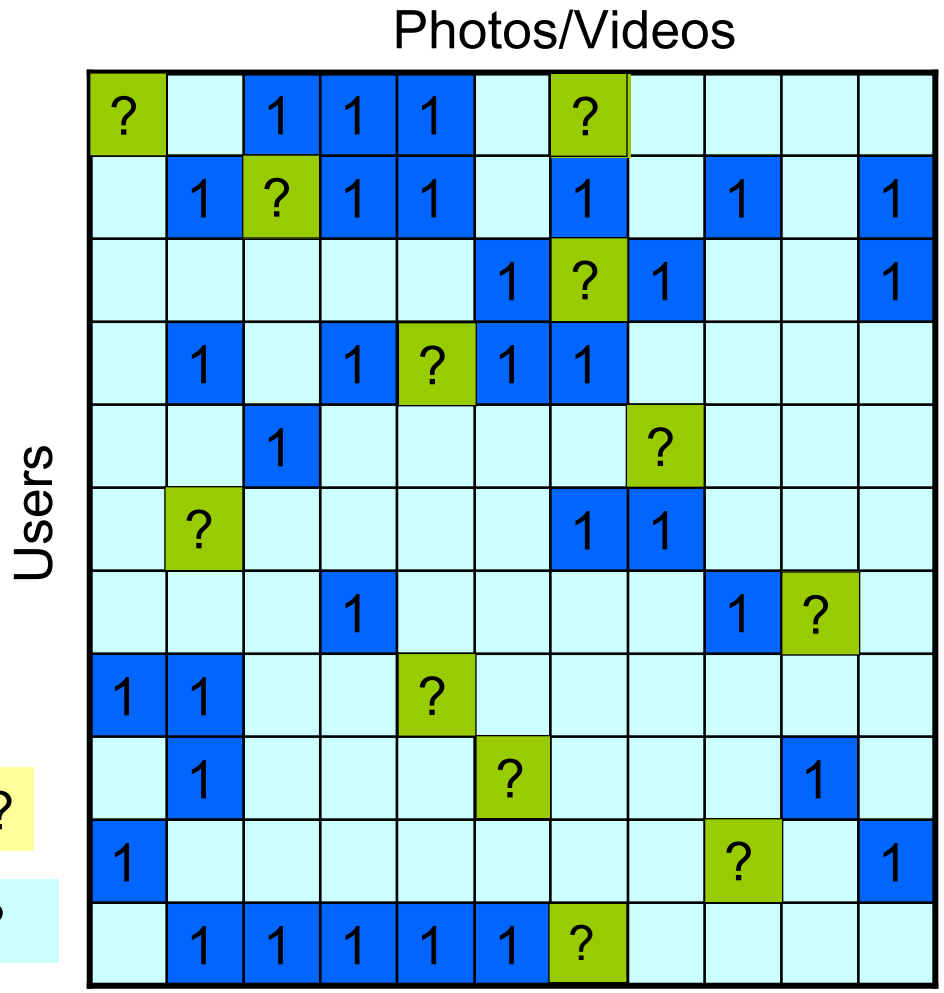
Predict *unobserved* entries

↓

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**?



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: Some patterns may be found in multiple CDBs

Preprocessing

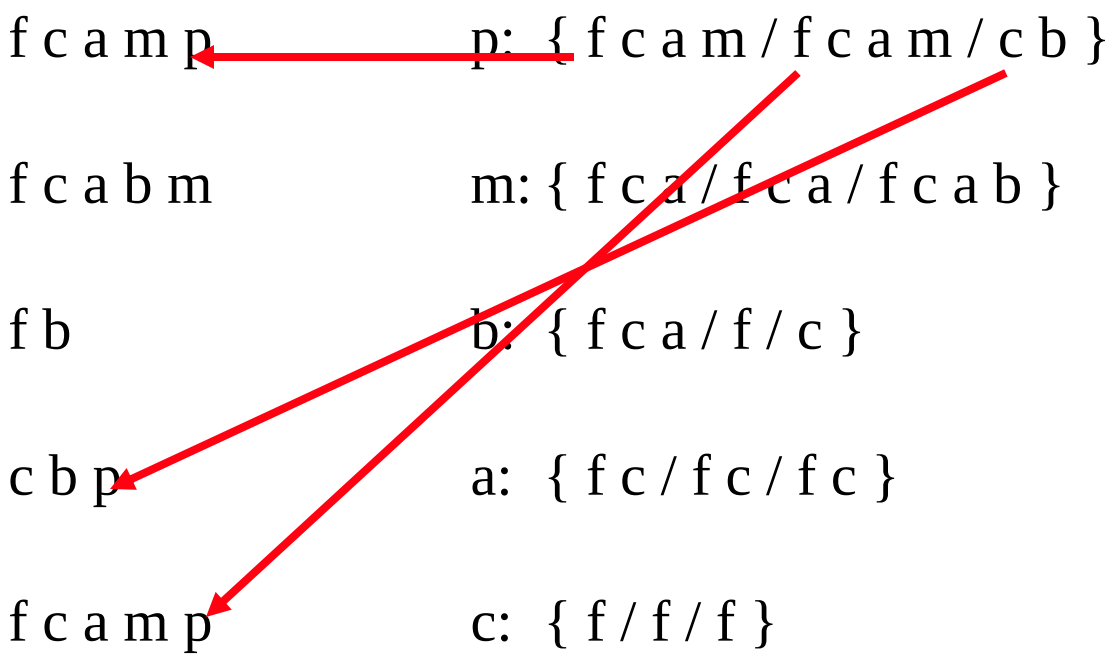
	f: 4 c: 4 a: 3 b: 3 m: 3 p: 3	
f a c d g i m p		f c a m p
a b c f l m o		f c a b m
b f h j o	o: 2 d: 1 e: 1 g: 1 h: 1 i: 1 k: 1 l : 1 n: 1	f b
b c k s p		c b p
a f c e l p m n		f c a m p

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

Parallel Projection

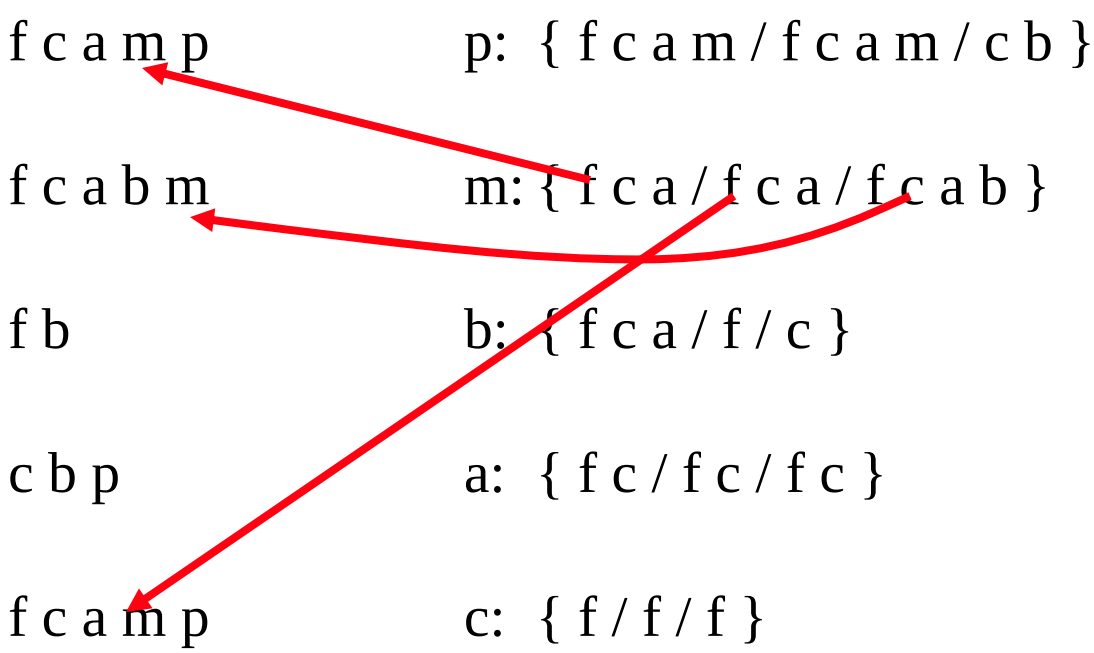
- According to Observation 2, we construct CDB of item A; then from this CDB, we find those patterns containing *A*
- How to construct the CDB of *A*?
  - If a transaction contains *A*, this transaction should appear in the CDB of *A*
  - Given a transaction {*B, A, C*}, it should appear in the CDB of *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



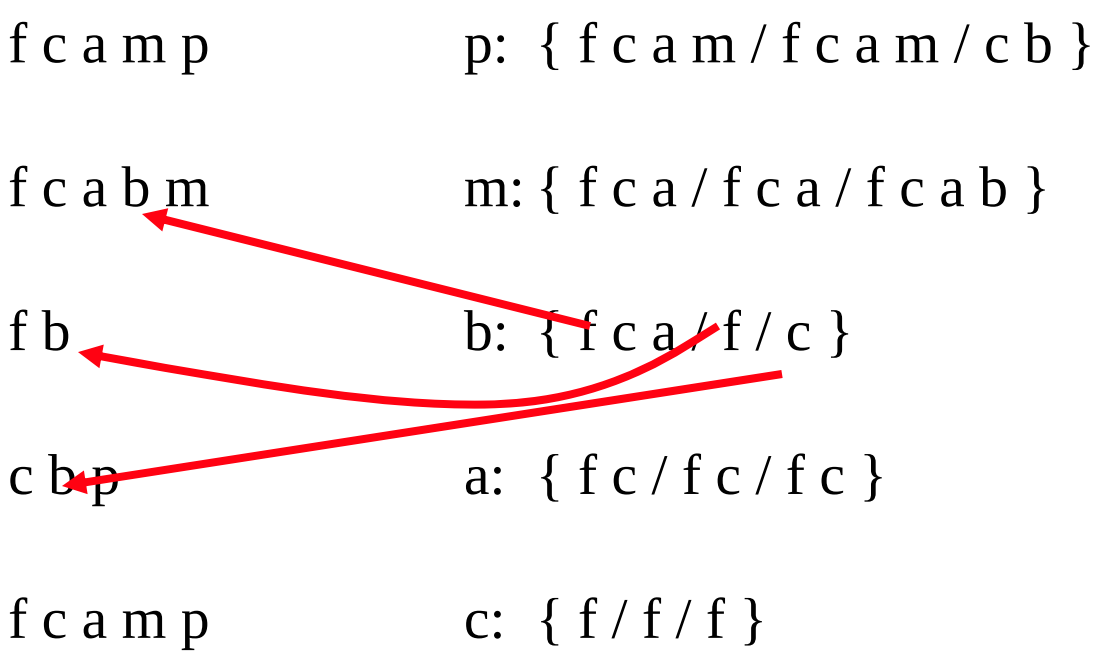
Example of Projection of a database into CDBs.
Left: sorted transactions in order of *f*, *c*, *a*, *b*, *m*, *p*
Right: conditional databases of frequent items

Example of Projection



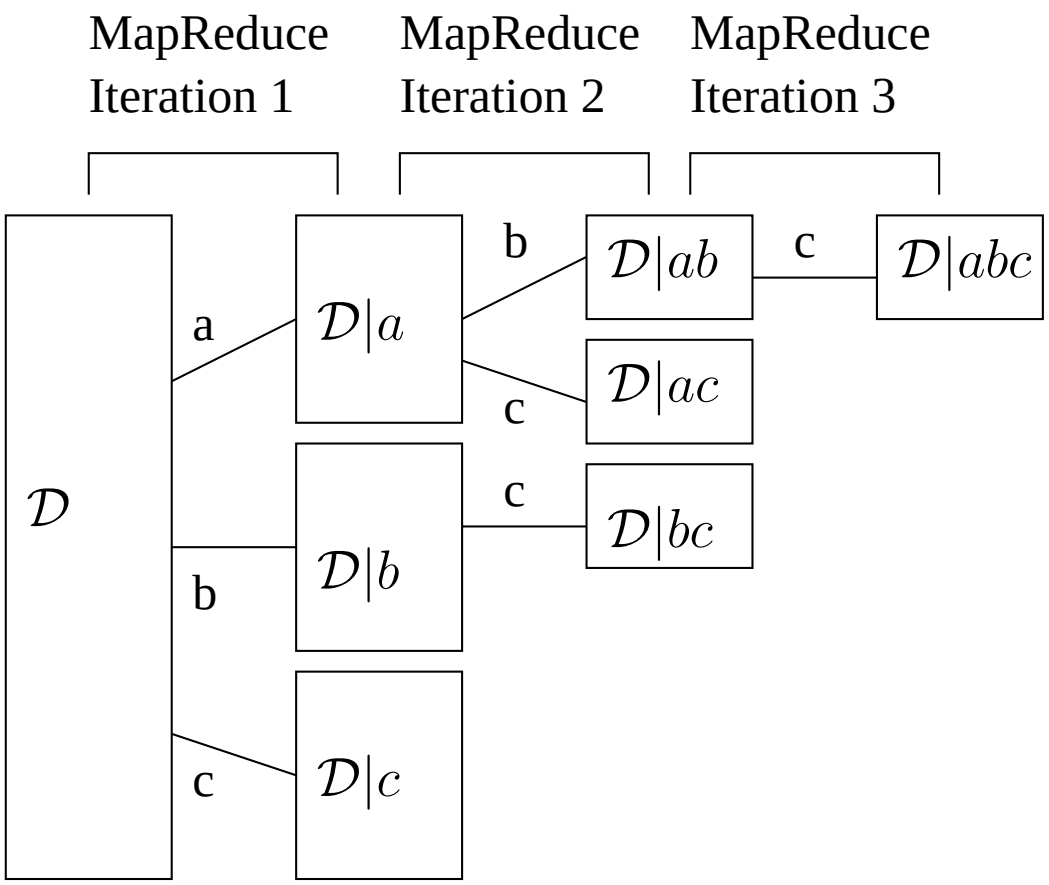
Example of Projection of a database into CDBs.
Left: sorted transactions;
Right: conditional databases of frequent items

Example of Projection



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]



- 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 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

p:{fcam/fcam/cb} p:3, pc:3

Reduce inputs (conditional databases) key: value	Reduce outputs (patterns and supports) key: value
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

Outline

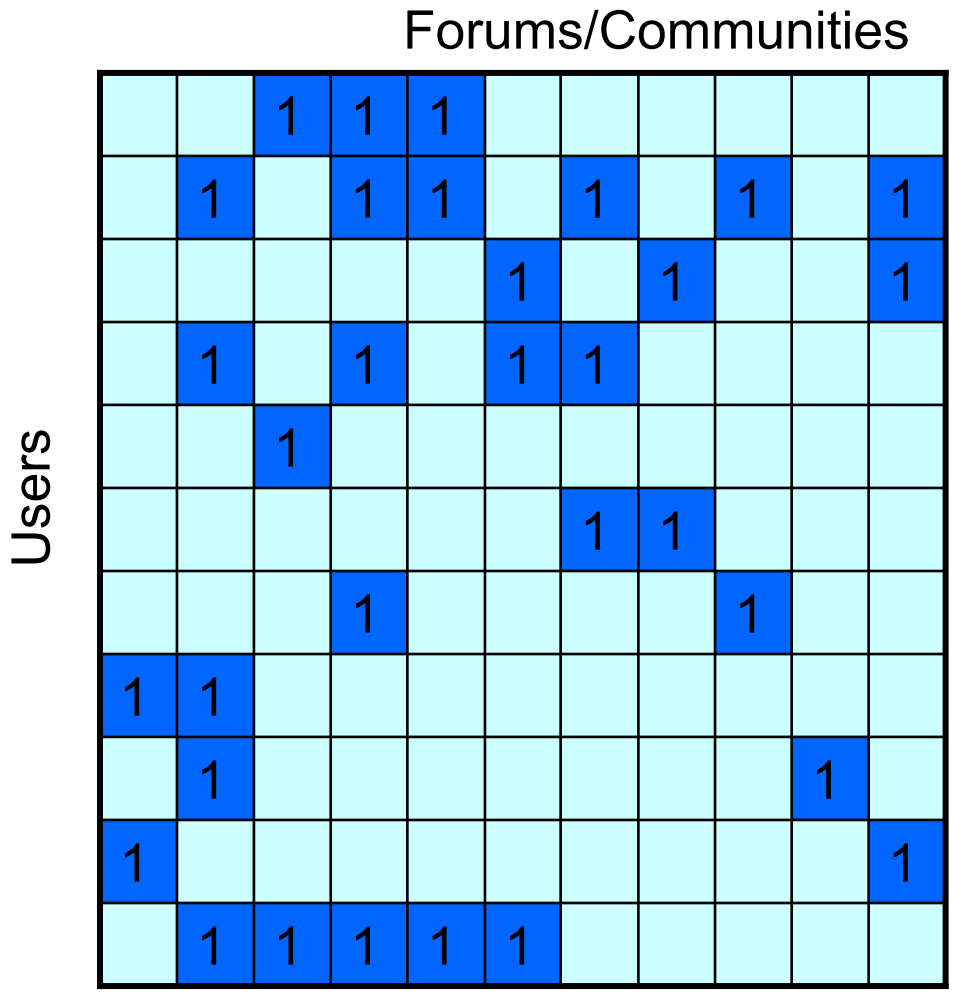
- Applications
  - Confucius
  - OpenSocial
- Key Subroutines
  - Clustering
  - Frequent Itemset Mining (FIM)
  - Combinational Collaborative Filtering
    - with PLSA
    - with LDA
  - Support Vector Machines

Collaborative Filtering

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

↓

Predict further *membership*

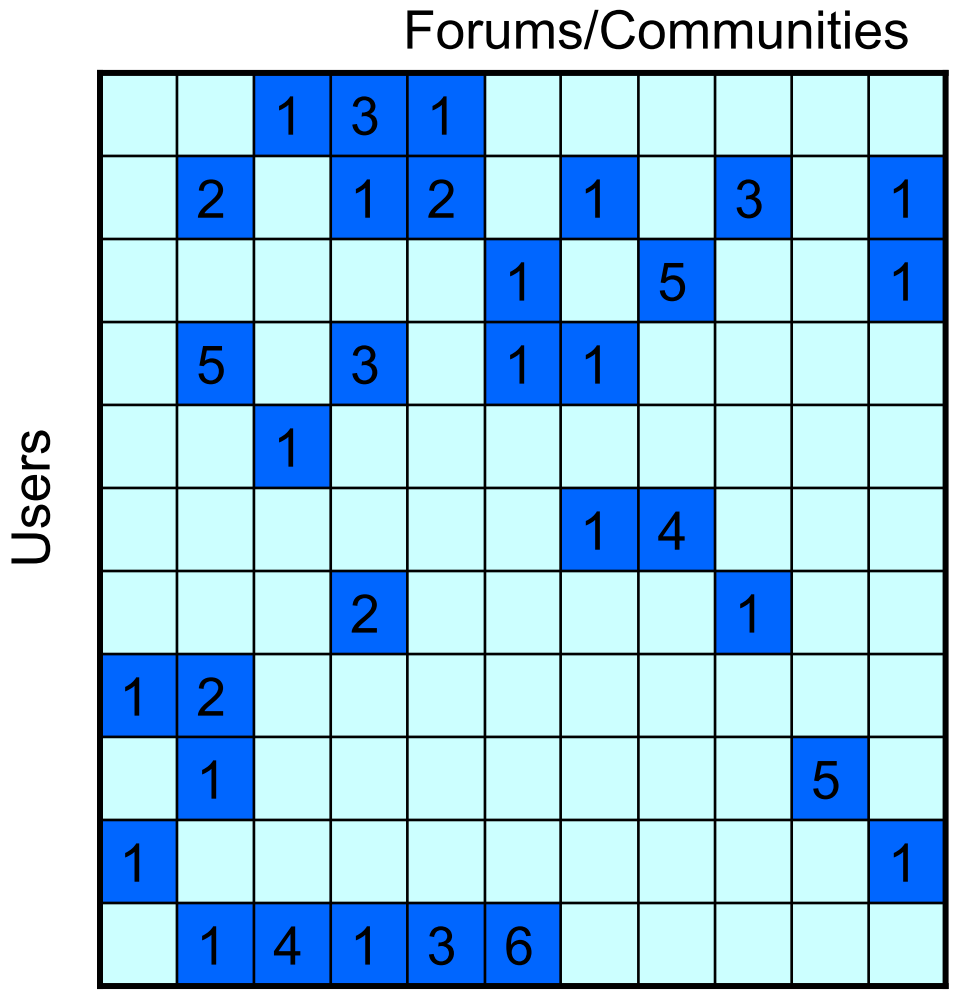


Collaborative Filtering

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

↓

Predict further *membership*

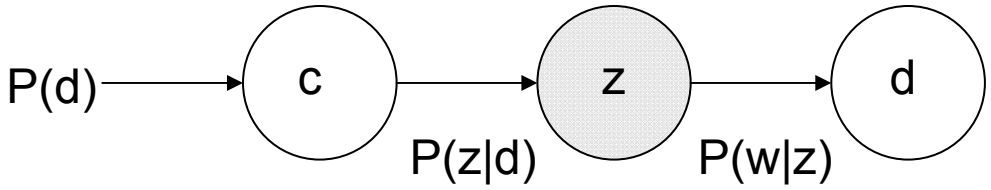


Notations

- Given a collection of co-occurrence data
  - Community: $C = \{c_1, c_2, \ldots, c_N\}$
  - User: $U = \{u_1, u_2, \ldots, u_M\}$
  - Description: $D = \{d_1, d_2, \ldots, d_V\}$
  - Latent aspect: $Z = \{z_1, z_2, \ldots, z_K\}$
- Models
  - Baseline models
    - Community-User (C-U) model
    - Community-Description (C-D) model
  - CCF: Combinational Collaborative Filtering
    - *Combines* both baseline models

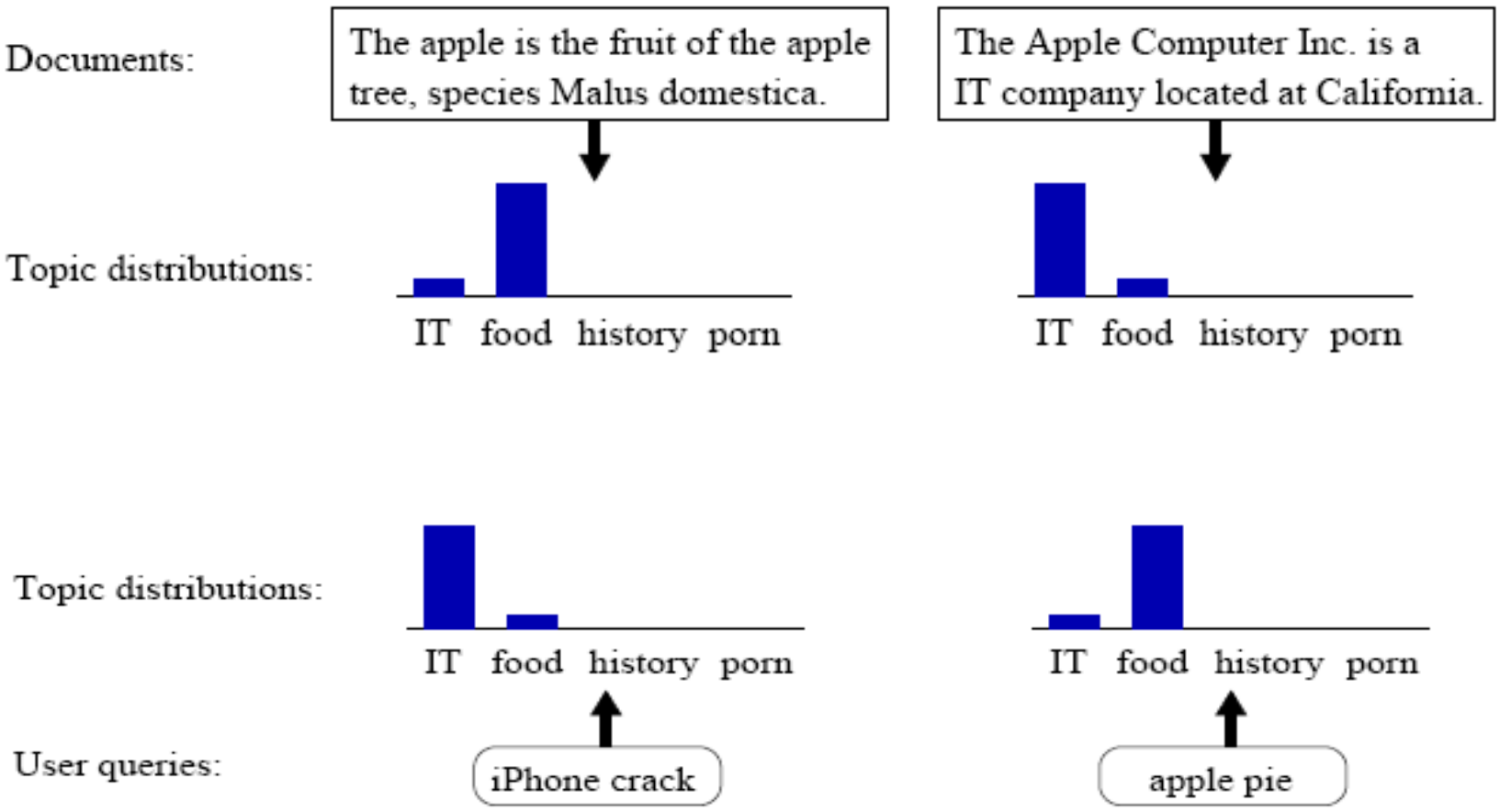
Probabilistic Latent Semantic Analysis (PLSA) [Hoffman 1999; Hoffman 2004]

- Document is viewed as a bag of words
- A *latent semantic layer* is constructed in between documents and words
- $P(d, c) = P(d|c)\,P(c) = P(c)\sum_z P(d|z)P(z|c)$



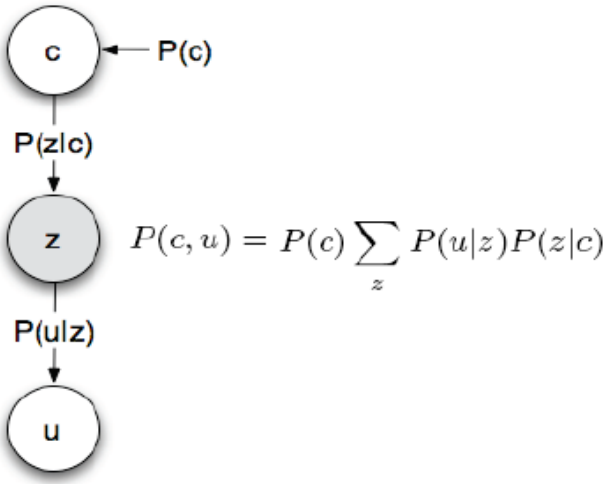
- Probability delivers explicit meaning
  - P(c|c), P(d|d), P(d, c)
- Model learning via EM or Gibbs sampling

Example of Latent Analysis



Baseline Models

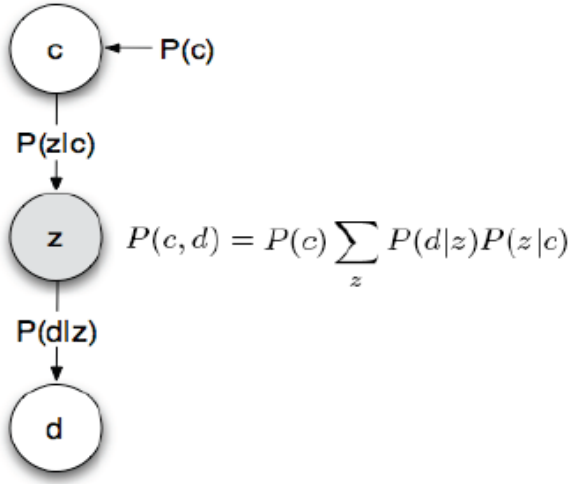
Community-User (C-U) model



$$P(c,u) = P(c)\sum_z P(u|z)P(z|c)$$

- Community is viewed as a bag of users
- ***c*** and ***u*** are rendered conditionally independent by introducing ***z***
- Generative process, for each user *u*
  1. A community *c* is chosen uniformly
  2. A topic *z* is selected from P(z|c)
  3. A user *u* is generated from P(u|z)

Community-Description (C-D) model

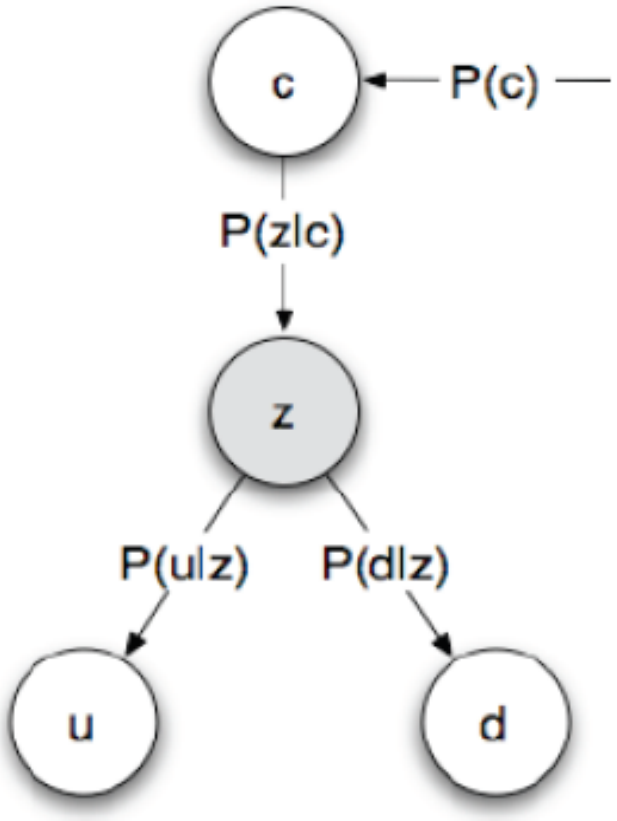


$$P(c,d) = P(c)\sum_z P(d|z)P(z|c)$$

- Community is viewed as a bag of words
- ***c*** and ***d*** are rendered conditionally independent by introducing ***z***
- Generative process, for each word *d*
  1. A community *c* is chosen uniformly
  2. A topic *z* is selected from P(z|c)
  3. A word *d* is generated from P(d|z)

CCF Model [Chen, et. al. KDD 08]

Combinational Collaborative Filtering (CCF) model



- CCF *combines* both baseline models
- A community is viewed as
  - a bag of users AND a bag of words
- By adding C-U, CCF can perform personalized recommendation which C-D alone *cannot*
- By adding C-D, CCF can perform better recommendation than C-U alone, which may suffer from sparsity
- CCF can do that C-U and C-D cannot
  - P(d|u), relate user to word
  - Useful for user targeting ads

Empirical Study

- Orkut Dataset
  - Collected in July, 2007
  - Two types of data were extracted
    - Community-user, community-description
  - 312,385 users
  - 109,987 communities
- Machine farm
  - Up to 200 machines in Google datacenters
  - Each machine is configured with:
    - A CPU faster than 2GHz
    - Memory larger than 4GBytes
- Evaluations
  - Community recommendation
  - Speedup

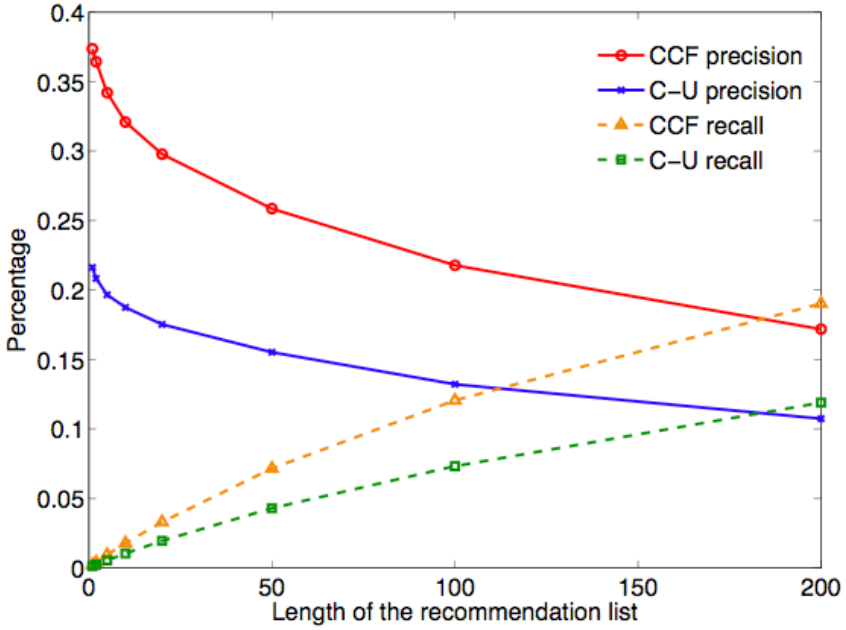
Community Recommendation

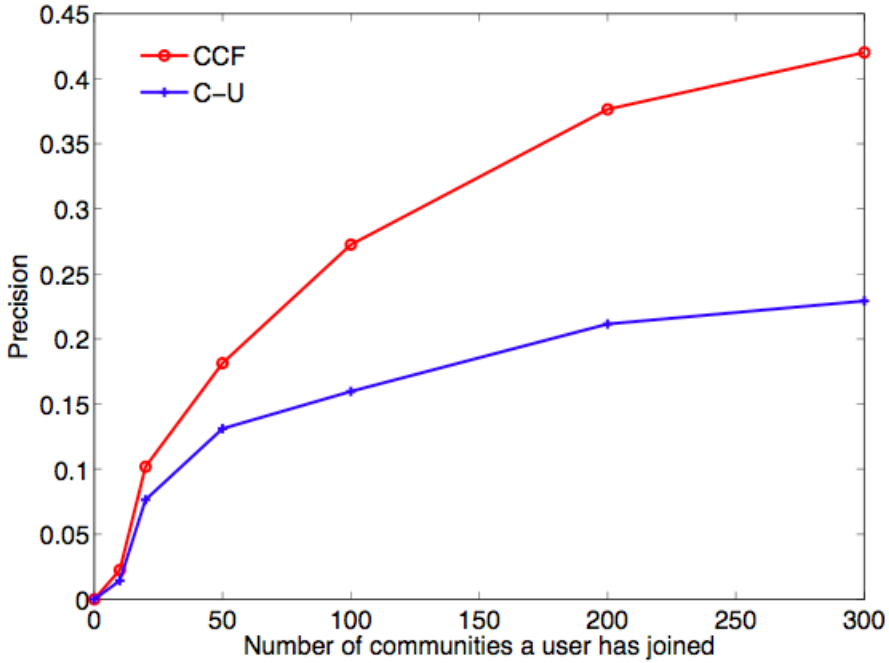
- Evaluation Method
  - Leave-one-out: randomly delete one community for each user
  - Check if a removed community can be recovered?
- Evaluation metric
  - Precision and Recall





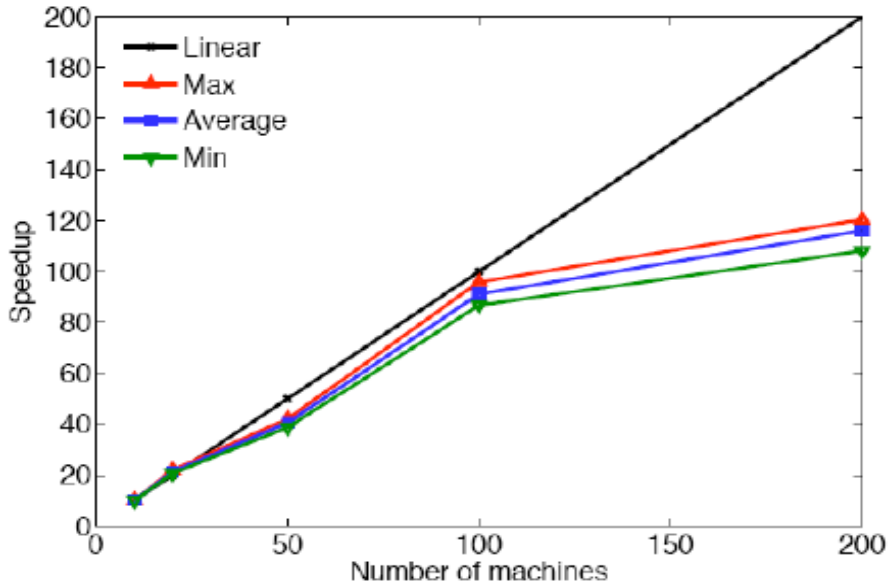
Results





- CCF outperforms C-U
- The more information, the higher accuracy

Gibbs Sampling MapRedue Speedup

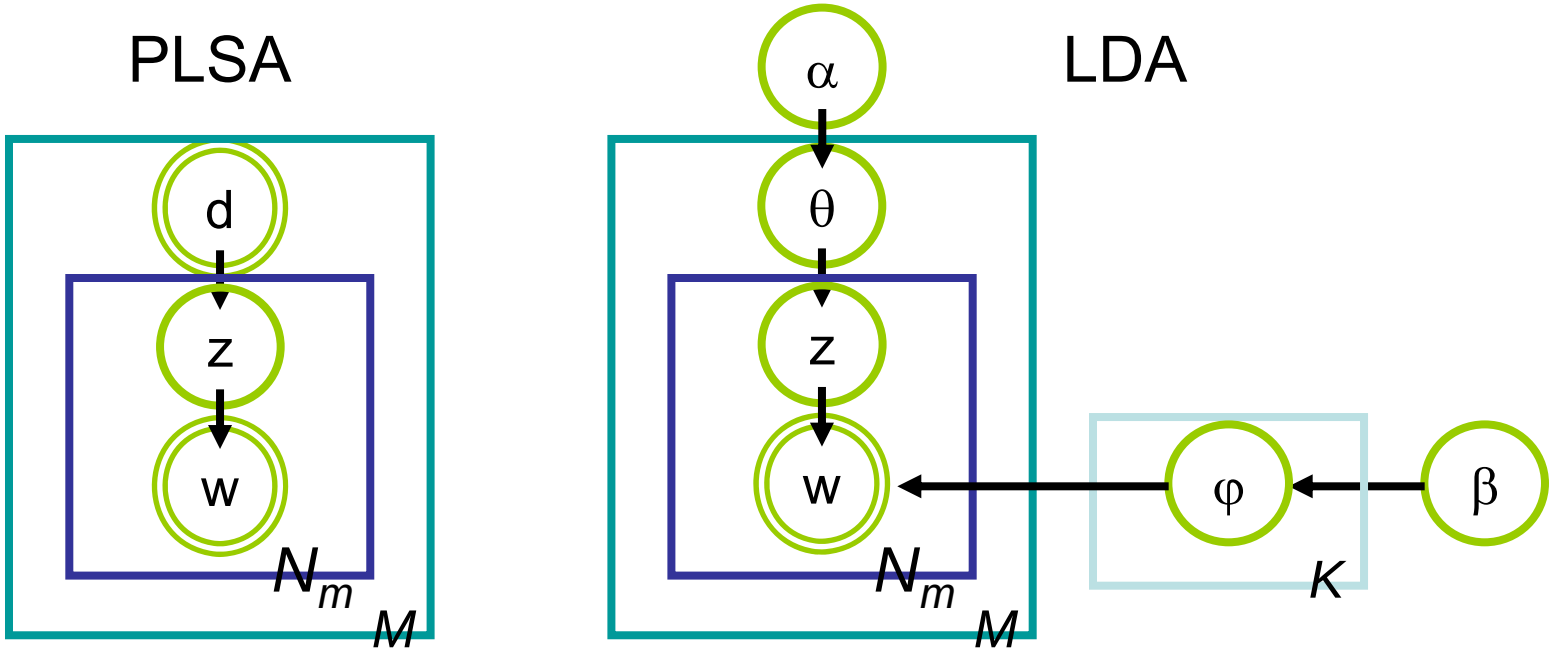


Machines	Time (sec.)	Speedup
10	9,233	10
20	4,326	21.3
50	2,280	40.5
100	1,014	91.1
200	796	116

- The Orkut dataset enjoys a linear speedup when the number of machines is up to 100
- Reduces the training time from one day to less than 14 minutes

Extensions

- Expand CCF to incorporate more types of information
- Replace PLSA with LDA



…Extensions

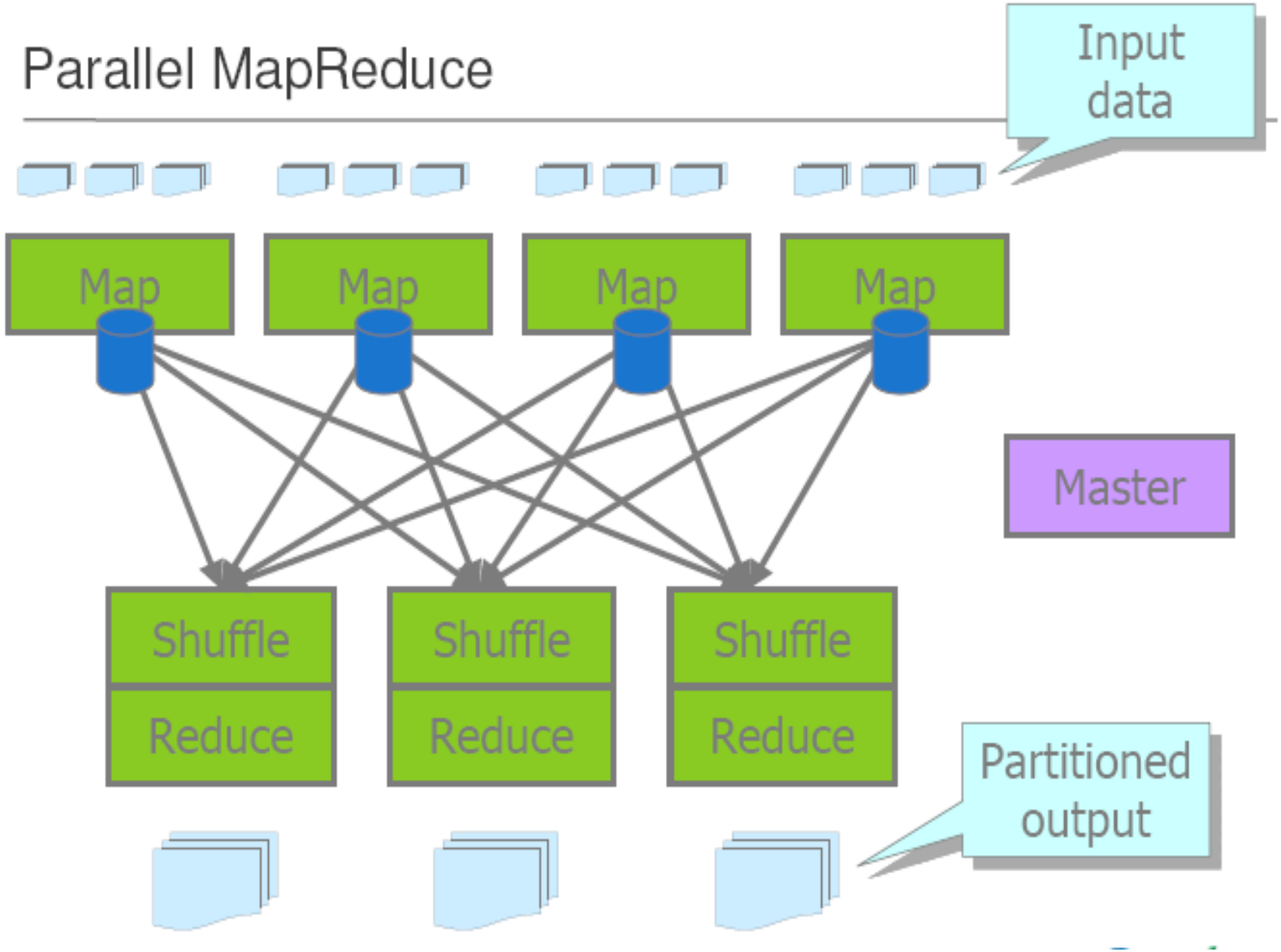
- Consider time dimension
- Perform incremental learning
- Construct topic hierarchy
- etc…

Outline

- Applications
  - Confucius
  - OpenSocial
- Key Subroutines
  - Clustering
  - Frequent Itemset Mining (FIM)
  - Combinational Collaborative Filtering
    - with PLSA
    - with LDA
  - Support Vector Machines [NIPS 07]
- Distributed Computing Perspectives

Distributed Computing Perspectives

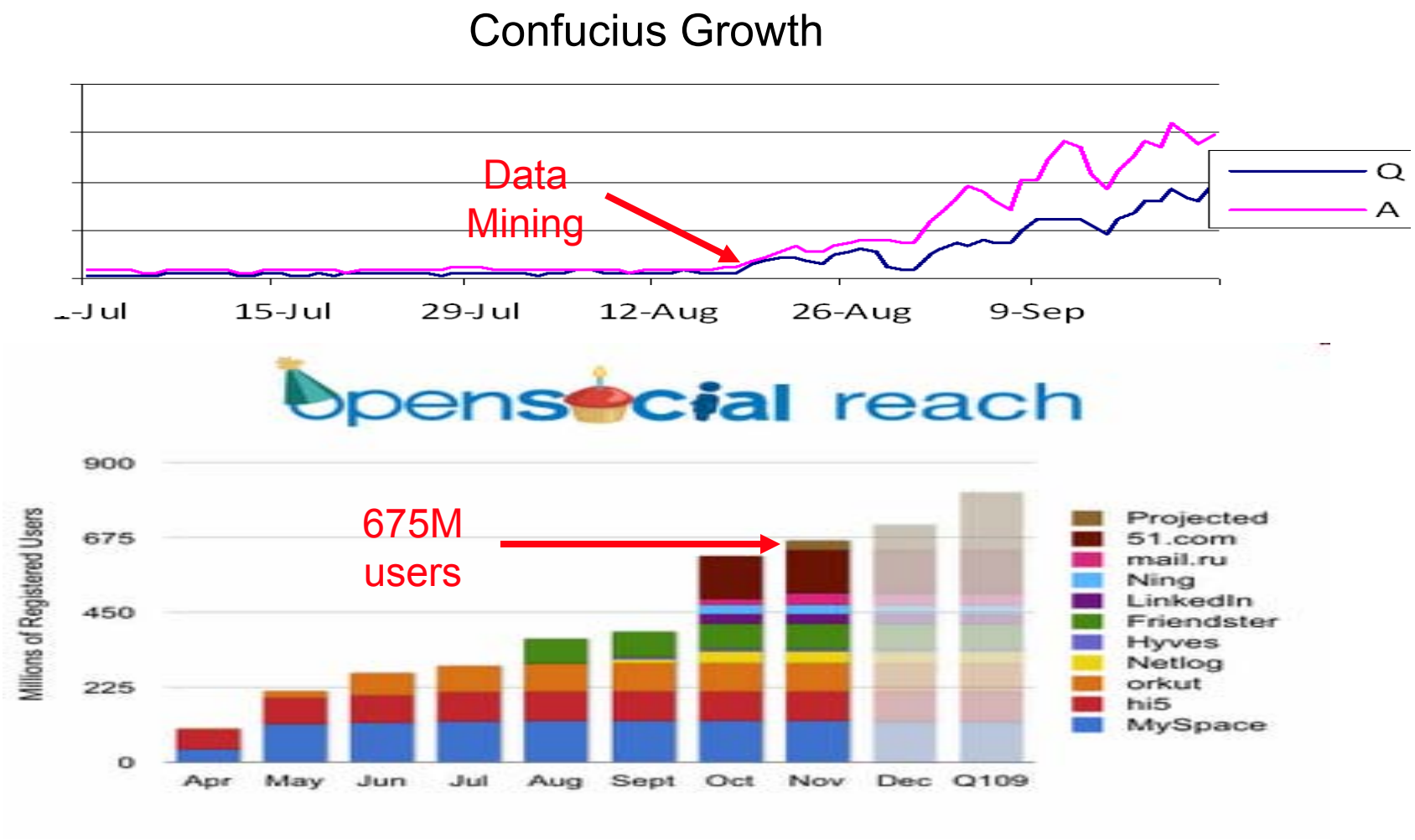
- Iterative
  - Most algorithms do a series of iterations
  - Data dependency: Iteration *t+1* depends on *t*
- Parallelize each iteration
  - In computation
  - In storage
- Auto Fault Recovery
  - Critical for large-scale tasks


Parallel MapReduce
Input data
Map
Map
Map
Map
Master
Shuffle
Shuffle
Shuffle
Reduce
Reduce
Reduce
Partitioned output

Comparison between Parallel Computing Frameworks

	MapReduce	Project Doe	MPI
GFS/IO and task rescheduling overhead between iterations	Yes	No +1	No +1
Flexibility of computation model	AllReduce only +0.5	 +0.7	Flexible +1
Efficient AllReduce	Yes +1	Yes +1	Yes +1
Recover from faults between iterations	Yes +1	Yes +1	No
Recover from faults within each iteration	Yes +1	Yes +1	No
Final Score for scalable machine learning	**3.5**	**4.7**	**3**

Conclusions…



…Conclusions

- Seven ML subroutines (disciples) of Confucius
- Recommendation is the push model of search
- Recommendation systems demand scalability
- ML algorithms demand “better” distributed computing models than MapReduce

…Conclusions

- Have parallelized key subroutines for mining massive data sets
  - Spectral Clustering [ECML 08]
  - Frequent Itemset Mining [ACM RS 08]
  - Combinational Collaborative Filtering [KDD 08]
    - with PLSA
    - with LDA
  - Support Vector Machines [NIPS 07]
- Relevant papers
  - http://infolab.stanford.edu/~echang/
- Open Source PSVM
  - http://code.google.com/p/psvm/

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.