

# Social Search Networks of People and Search Engines

CS6200  
Information Retrieval

# Social Search

- Social search
  - ✦ Communities of users actively participating in the search process
  - ✦ Goes beyond classical search tasks
- Key differences
  - ✦ Users interact with the system
  - ✦ Users interact with other users either implicitly or explicitly

# Web 2.0

- Social search includes, but is not limited to, the so-called social media sites
  - ❖ Collectively referred to as “Web 2.0” as opposed to the classical notion of the Web (“Web 1.0”)
- Social media sites
  - ❖ User generated content
  - ❖ Users can tag their own and other’s content
  - ❖ Users can share favorites, tags, etc., with others
- Examples (from the last 10 years):
  - ❖ Digg, Twitter, Flickr, YouTube, Del.icio.us, CiteULike, MySpace, Facebook, and LinkedIn

# Social Search

- User tagging (i.e., manual indexing)
- Searching within communities
- Filtering and recommender systems
- Distributed search
  - \* Peer-to-peer (P2P, not covered here)
  - \* Metasearch (if there's time)

# User Tags and Manual Indexing

- Then: Library card catalogs
  - ✦ Indexing terms chosen with search in mind
  - ✦ Experts generate indexing terms
  - ✦ Terms are very high quality
  - ✦ Terms chosen from controlled vocabulary
- Now: Social media tagging
  - ✦ Tags not always chosen with search in mind
  - ✦ Users generate tags
  - ✦ Tags can be noisy or even incorrect
  - ✦ Tags chosen from *folksonomies*

# Expert Cataloguing

[CLICK HERE FOR ALL CALL NUMBERS AND ITEM AVAILABILITY](#)







Title	<u><a href="#">Legal ontology engineering [electronic resource] : methodologies, modelling trends, and the ontology of professional judicial knowledge / Nuria Casellas.</a></u>
Author	<u><a href="#">Casellas, Nuria.</a></u>
Publisher	Dordrecht ; New York : Springer, c2011.
Format	E-Book
Physical Descrip.	1 online resource (xxii, 297 p.)
Series	( <u><a href="#">Law, governance and technology series ; v.3</a></u> )
Bibliography	Includes bibliographical references and index.
Subject(s)	<u><a href="#">Legal research -- Data processing.</a></u> <u><a href="#">Ontologies (Information retrieval)</a></u> <u><a href="#">Semantic computing.</a></u>
Genre	<u><a href="#">Electronic books.</a></u>
Other Author(s)	<u><a href="#">SpringerLink (Online service)</a></u>
Related Entry	<u><a href="#">Springer e-books</a></u>
Series Added Entry	( <u><a href="#">Law, governance and technology series ; v.3.</a></u> )
ISBN	9789400714977 (electronic bk.) 9400714971 (electronic bk.)
Location(s)	<u><a href="#">SC Neilson Library / SC Internet / Online Resource</a></u>


[CLICK HERE FOR ALL CALL NUMBERS AND ITEM AVAILABILITY](#)

# Social Tagging

**flickr**<sup>®</sup> from **YAHOO!**  
Home The Tour Sign Up Explore | Upload

You aren't signed in [Sign In](#)

Favorite Actions    Share   




By [Honeybuy2012](#)  
No real name given + Add Contact

This photo was taken on April 17, 2012.

6 views

This photo belongs to

[Honeybuy2012's photostream](#) (439)



Tags

cherry • blossom

License

© All Rights Reserved

Privacy

This photo is visible to everyone

cherry blossom 16  
cherry blossom

# Social Tagging

flickr® from YAHOO!

Home The Tour Sign Up Explore Upload

You aren't signed in Sign In Help

Search

Favorite Actions Share

Newer Older



## cherry blossom obama

President Obama visits the Cherry Blossom Festival? Well, not really -- his motorcade passed by the Tidal Basin en route to Ft. McNair for his big Libya speech. A few lucky tourists got a chance to see it whizz by.

## Comments and faves

★ sarahmulca added this photo to their favorites. (13 months ago)

By philliefan99  
No real name given + Add Contact

This photo was taken on March 28, 2011 in Washington, District of Columbia, US, using a Pentax K-7.



346 views 1 comment 3 favorites

## This photo belongs to

philliefan99's photostream (9,048)



## This photo also appears in

- Obamarama (set)
- Barack Obama Photos (group)
- dcProper (DC only; please no... (group)
- We <3 DC (group)
- Washington DC Metro Area Sights (group)
- Washington DC/Metro Area (group)

## Tags

washington dc • district of columbia • 15th street sw • tidal basin • 2011 cherry blossom festival • bureau of printing and engraving • department of the treasury • president barack obama • motorcade • presidential limousine • aka the beast • dcist • POTUS • cadillac



# Social Tagging

 Pinboard

0.654 s

[recent](#) · [popular](#) · [tour](#) · [howto](#) · [log in](#)

**cshalizi** 7199

RSS

[« earlier](#)

[Language Log » Ask a baboon](#)

Oh, dear.

[linguistics](#) [primates](#) [experimental\\_psychology](#) [machine\\_learning](#) [perceptron](#) [logistic\\_regression](#)

[classifiers](#) [evisceration](#) [methodological\\_EPIC\\_FAIL](#) [liberman.mark](#)

[why\\_oh\\_why\\_cant\\_we\\_have\\_a\\_better\\_academic\\_publishing\\_system](#)

5 hours ago [copy to mine](#)

Search All

See this user's [network](#)

top tags    all tags · 2 · 5 · 10 · 20

[academia](#) [afghanistan](#) [agent-based\\_models](#)

[american\\_hegemony](#) [american\\_history](#)

[anthropology](#) [architecture](#) [art](#)

# Some Categories of Tags

- Content-based: cherry blossoms, car
- Context-based: Washington, DC
- Attributes: Nikon, B&W
- Subjective: delicious, awesome
- Organizational: to\_read

# Searching Tags

- Searching user tags is challenging
  - ✦ Most items have only a few tags
  - ✦ Tags are very short
- Boolean, probabilistic, vector space, and language modeling will fail if use naively
- Must overcome the vocabulary mismatch problem between the query and tags

# Tag Expansion

- Can overcome vocabulary mismatch problem by expanding tag representation with external knowledge
- Possible external sources
  - ❖ Thesaurus
  - ❖ Web search results
  - ❖ Query logs
- After tags have been expanded, can use standard retrieval models

# Tag Expansion

Age of Aquariums - Tropical Fish

Huge educational aquarium site for **tropical fish** hobbyists, promoting responsible **fish** keeping internationally since 1997.

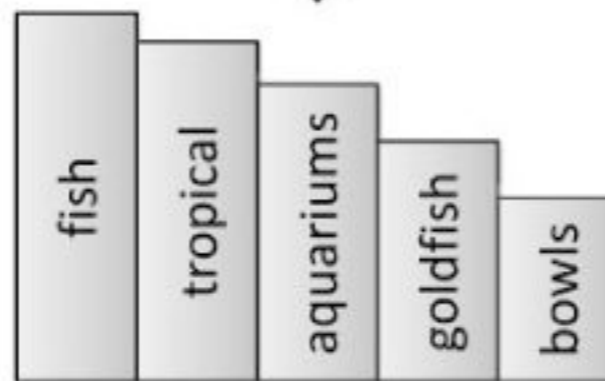
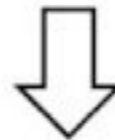
The Krib (Aquaria and Tropical Fish)

This site contains information about **tropical fish** aquariums, including archived usenet postings and e-mail discussions, along with new ...

...

Keeping Tropical Fish and Goldfish in Aquariums, Fish Bowls, and ...

Keeping **Tropical Fish** and Goldfish in Aquariums, **Fish** Bowls, and Ponds at AquariumFish.net.



$P(w \mid \text{"tropical fish"})$

# Searching Tags

- Even with tag expansion, searching tags is challenging
- Tags are inherently noisy and incorrect
- Many items may not even be tagged!
- Typically easier to find popular items with many tags than less popular items with few/no tags

# Inferring Missing Tags

- How can we automatically tag items with few or no tags?
- Uses of inferred tags
  - \* Improved tag search
  - \* Automatic tag suggestion

# Inferring Tags

- TF.IDF
  - ✦ Suggest tags that have a high TF.IDF weight in the item
  - ✦ Only works for textual items
- Classification
  - ✦ Train binary classifier for each tag
  - ✦ Performs well for popular tags, but not as well for rare tags
- Maximal marginal relevance
  - ✦ Finds tags that are relevant to the item and novel with respect to existing tags

$$MMR(t; T_i) = \left( \lambda Sim_{item}(t, i) - (1 - \lambda) \max_{t \in T_i} Sim_{tag}(t_i, t) \right)$$



# Browsing & Tag Clouds

- Search is useful for finding items of interest
- Browsing is more useful for exploring collections of tagged items
- Various ways to visualize collections of tags
  - ✦ Tag lists
  - ✦ Tag clouds
  - ✦ Alphabetical order
  - ✦ Grouped by category
  - ✦ Formatted/sorted according to popularity

# Tag Clouds

animals architecture **art** australia autumn baby band barcelona **beach** berlin  
**birthday** black blackandwhite blue **california** cameraphone canada canon  
car cat chicago china christmas church city clouds color concert day dog  
england europe **family** festival film florida flower flowers food  
france **friends** fun garden germany girl graffiti green halloween hawaii  
holiday home house india ireland **italy** japan july kids lake landscape light live  
**london** macro me mexico music nature new newyork night  
nikon nyc ocean paris park **party** people portrait red river rock  
sanfrancisco scotland sea seattle show sky snow spain spring street  
summer sunset taiwan texas thailand tokyo toronto **travel**  
tree trees **trip** uk usa vacation washington water wedding

# Searching within Communities

# Conversational Structure

The image shows a screenshot of a forum thread with four posts. The first post is by user 'clig' on Sep 9, 2010, at 6:39 PM, titled 'bluetooth update'. The post content is: 'having updated to 4.1 .apple have still not sorted bluetooth problem why do they not listen to us' and 'dell inspiron1 laptop Windows 7'. The second post is by 'dbx2spc' on Sep 9, 2010, at 7:10 PM, titled 'Re: bluetooth update', replying to 'clig'. The content is: 'Same here, I can pair it with my Pioneer DEH-P7000BT but once I turn off the car then restart only the works. I have to manually pair it again to get the bluetooth audio to work every time. APPLE PLEASE F YOUR ISSUE NOT PIONEERS OR ANY OTHERS.' and 'DEH-P7000BT iOS 4'. The third post is by 'jacksTLOS' on Sep 10, 2010, at 1:01 AM, titled 'Re: bluetooth update', replying to 'dbx2spc'. The content is: 'Reset your settings:', 'Settings- general- reset- reset all settings', 'This should fix your problems', and 'Vaio Windows 7'. The fourth post is by 'dbx2spc' on Sep 10, 2010, at 10:08 AM, titled 'Re: bluetooth update', replying to 'jacksTLOS'. The content is: 'Sorry but resetting does not do a \*\*\*\* bit of good.' and 'iOS 4'.

# Conversational Structure



**Mahj**  
Boston  
posts: 4,288

## 5. Re: Chihuly exhibit at MFA - simply fantastic!

29 June 2011, 4:57



Been twice so far, may go a third time. Quite a boon for the MFA, but do make sure you wander through the other exhibits as well.

[Reply](#)

[Report inappropriate content](#)

[Save this Post](#)



**bellman3**  
Massachusetts  
posts: 1,185  
reviews: 5

## 6. Re: Chihuly exhibit at MFA - simply fantastic!

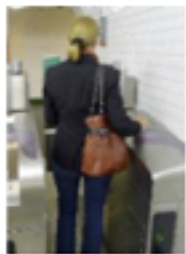
29 June 2011, 6:34

It really is very wonderful. We didn't know that photography was permitted, so I definitely want to go again to take pics. But even if I don't have photos, I don't think my memories of the pieces will fade anytime soon.

[Reply](#)

[Report inappropriate content](#)

[Save this Post](#)



**sojoh**  
Paris, France

## 7. Re: Chihuly exhibit at MFA - simply fantastic!

29 June 2011, 8:43

Mahj, I agree that the other exhibits at the MFA are well worth a visit too. It was a real pleasure to see the Impressionist paintings and to see a bit of the new Art of Americas wing. I'm already planning to return to the MFA during my next visit to Boston.

The screenshot shows a forum thread with three posts. The first post is by user 'clig' (stafford, registered Aug 19, 2010) titled 'bluetooth update' (posted Sep 9, 2010 6:39 PM), mentioning a 'dell inspiron1 laptop' and 'Windows 7'. The second post is by 'dbx2spc' (registered Sep 9, 2010) titled 'Re: bluetooth update' (posted Sep 9, 2010 7:10 PM), mentioning 'Pioneer DEH-P7000BT' and 'iOS 4'. The third post is by 'jacksTLOS' (Mesa, AZ, registered Sep 2, 2010) titled 'Re: bluetooth update' (posted Sep 10, 2010 1:01 AM), providing instructions to 'Reset your settings' and 'Settings- general- reset- reset all settings' on a 'Vaio Windows 7' device. The thread concludes with a post by 'dbx2spc' (registered Sep 9, 2010) titled 'Re: bluetooth update' (posted Sep 10, 2010 10:08 AM) stating 'Sorry but resetting does not do a \*\*\*\* bit of good.' on 'iOS 4'.

# Conversational Structure



Mahj  
Boston  
posts: 4,288

## 5. Re: Chihuly exhibit at MFA - simply fantastic!

29 June 2011, 4:57



Been twice so far, may go a third time. Quite a boon for the MFA, but do make sure you wander through the other exhibits as well.

Reply

Report inappropriate content

Save this Post



bellman3  
Massachusetts  
posts: 1,185  
reviews: 5

## 6. Re: Chihuly exhibit at MFA - simply fantastic!

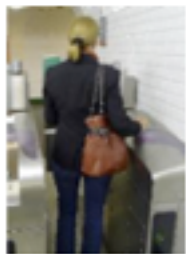
29 June 2011, 6:34

It really is very wonderful. We didn't know that photography was permitted, so I definitely want to go again to take pics. But even if I don't have photos, I don't think my memories of the pieces will fade anytime soon.

Reply

Report inappropriate content

Save this Post



sojoh  
Paris, France

## 7. Re: Chihuly exhibit at MFA - simply fantastic!

29 June 2011, 8:43

Mahj, I agree that the other exhibits at the MFA are well worth a visit too. It was a real pleasure to see the Impressionist paintings and to see a bit of the new Art of Americas wing. I'm already planning to return to the MFA during my next visit to Boston.

clig  
New! Posted: Sep 9, 2010 6:39 PM  
Posts: 3  
From: stafford  
Registered:  
Aug 19, 2010  
having updated to 4.1 .apple have still not sorted bluetooth problem why do they not listen to us  
dell inspiron1 laptop Windows 7

Re: bluetooth update  
New! Posted: Sep 9, 2010 7:10 PM in response to: clig  
Posts: 3  
Registered:  
Sep 9, 2010  
Same here, I can pair it with my Pioneer DEH-P7000BT but once I turn off the car then restart only the works. I have to manually pair it again to get the bluetooth audio to work every time. APPLE PLEASE FIX YOUR ISSUE NOT PIONEERS OR ANY OTHERS.  
DEH-P7000BT iOS 4

Re: bluetooth update  
New! Posted: Sep 10, 2010 1:01 AM in response to: dbx2spc  
Posts: 10  
From: Mesa, AZ  
Registered:  
Sep 2, 2010  
Reset your settings:  
Settings- general- reset- reset all settings  
This should fix your problems  
Vaio Windows 7

Re: bluetooth update  
New! Posted: Sep 10, 2010 10:08 AM in response to: jacksTLOS  
Posts: 3  
Registered:  
Sep 9, 2010  
Sorry but resetting does not do a \*\*\*\* bit of good.  
iOS 4

# Conversational Structure



**Mahj**  
Boston  
posts: 4,288

**5. Re: Chihuly exhibit at MFA - simply fantastic!**  
29 June 2011, 4:57

**Destination Expert ?**  
for Boston

Been twice so far, may go a third time. Quite a boon for the MFA, but do make sure you wander through the other exhibits as well.

Reply

Report inappropriate content

Save this Post



**bellman3**  
Massachusetts  
posts: 1,185  
reviews: 5

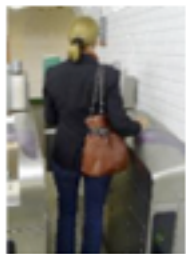
**6. Re: Chihuly exhibit at MFA - simply fantastic!**  
29 June 2011, 6:34

It really is very wonderful. We didn't know that photography was permitted, so I definitely want to go again to take pics. But even if I don't have photos, I don't think my memories of the pieces will fade anytime soon.

Reply

Report inappropriate content

Save this Post



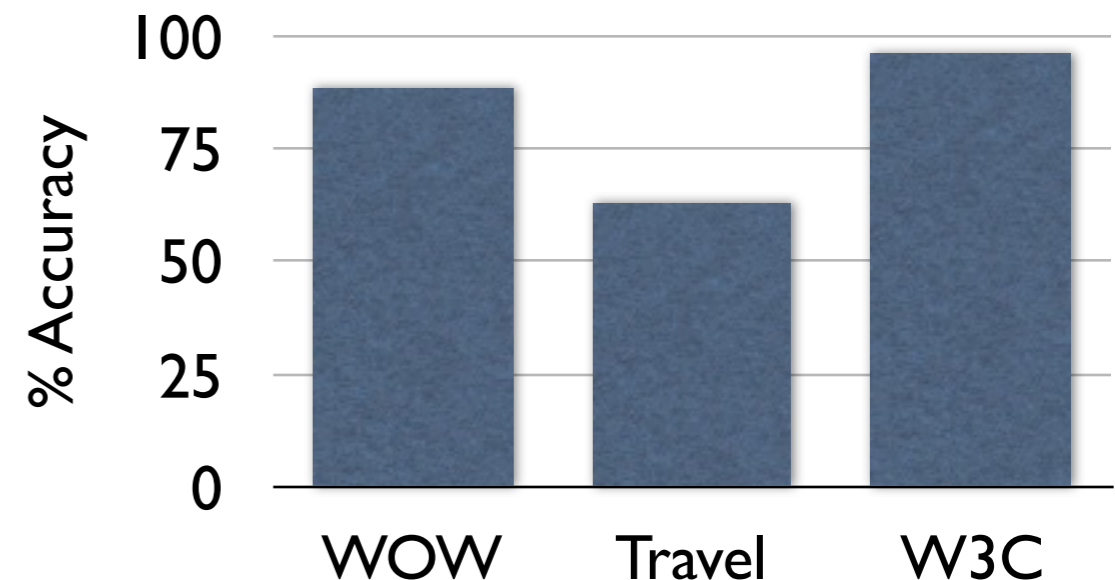
**sojoh**  
Paris, France

**7. Re: Chihuly exhibit at MFA - simply fantastic!**  
29 June 2011, 8:43

**Mahj**, I agree that the other exhibits at the MFA are well worth a visit too. It was a real pleasure to see the Impressionist paintings and to see a bit of the new Art of Americas wing. I'm already planning to return to the MFA during my next visit to Boston.

The screenshot shows a forum thread with three posts:

- Post 1:** User **clig** (stafford, registered Aug 19, 2010) posts "having updated to 4.1 .apple have still not sorted bluetooth problem why do they not listen to us" on a "dell inspiron1 laptop Windows 7".
- Post 2:** User **dbx2spc** (registered Sep 9, 2010) replies: "Same here, I can pair it with my Pioneer DEH-P7000BT but once I turn off the car then restart only the works. I have to manually pair it again to get the bluetooth audio to work every time. APPLE PLEASE FIX YOUR ISSUE NOT PIONEERS OR ANY OTHERS." on a "DEH-P7000BT IOS 4".
- Post 3:** User **jacksTLOS** (Mesa, AZ, registered Sep 2, 2010) replies: "Reset your settings: Settings- general- reset- reset all settings. This should fix your problems" on a "Vaio Windows 7".
- Post 4:** User **dbx2spc** replies: "Sorry but resetting does not do a \*\*\*\* bit of good." on "IOS 4".



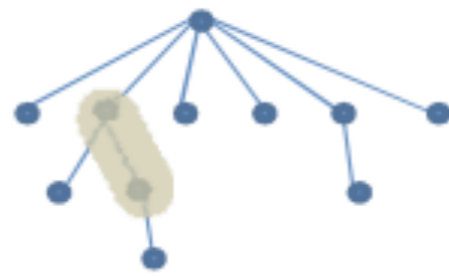
# Retrieving Social Data



Thread



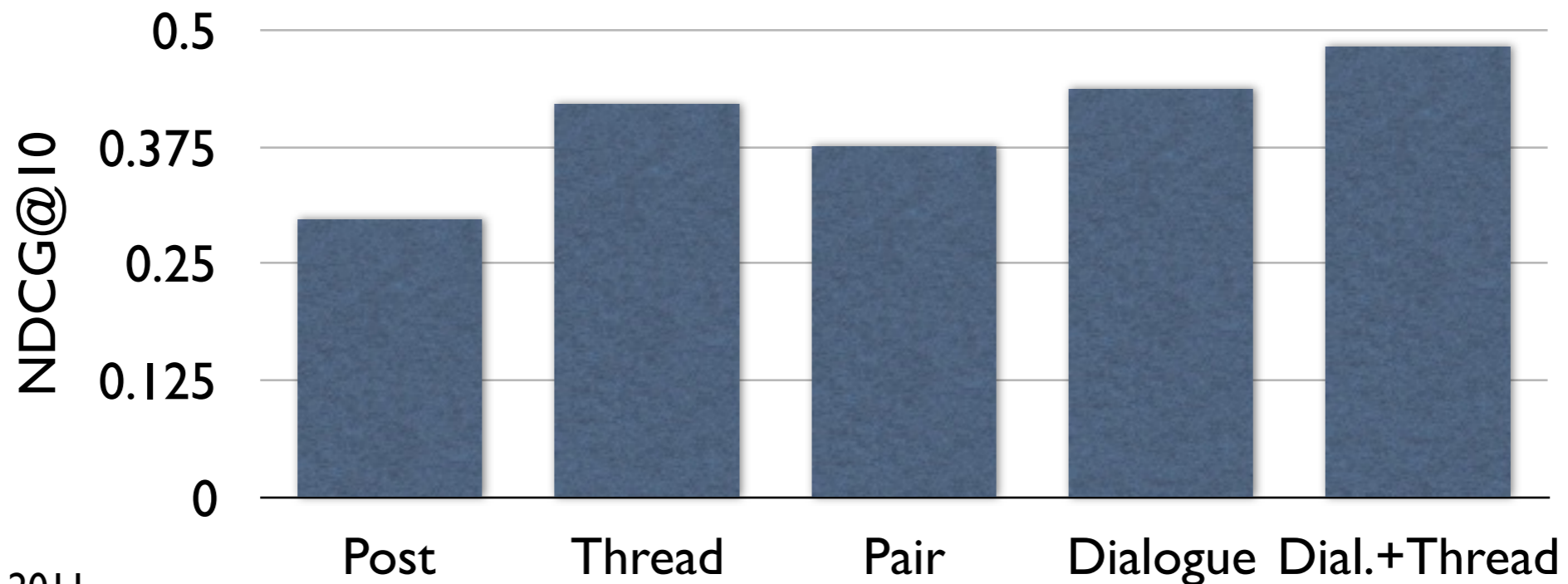
Posting



Pair



Dialogue





# Finding Communities

- How can we find users with common interests?
- How can we find documents on a common topic?
- Graph clustering
- Hypertext Induced Topic Search (HITS)
  - ❖ Nodes may be *Hubs* or *Authorities*
  - ❖ Iterative solution

# HITS Algorithm

---

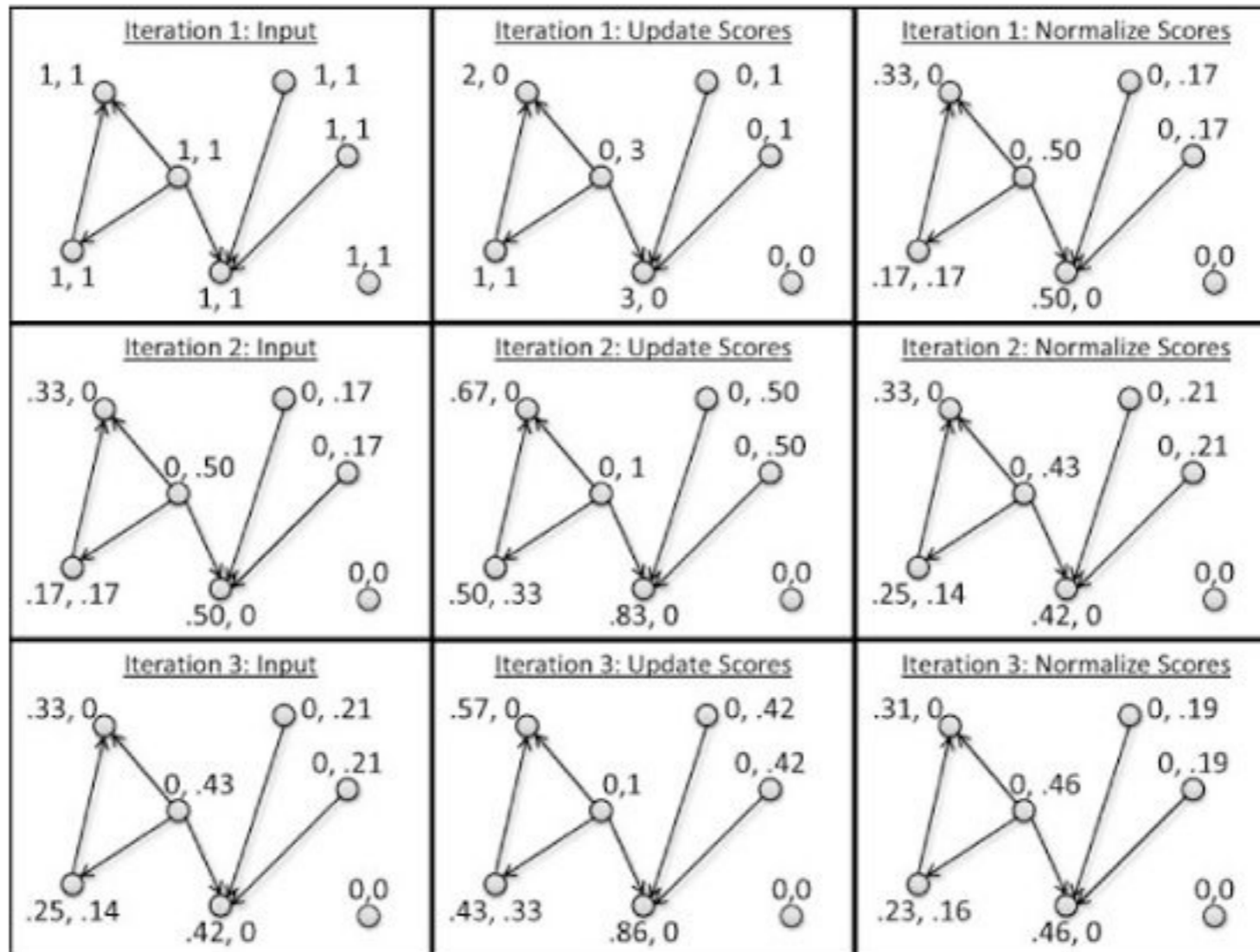
## Algorithm 3 HITS

---

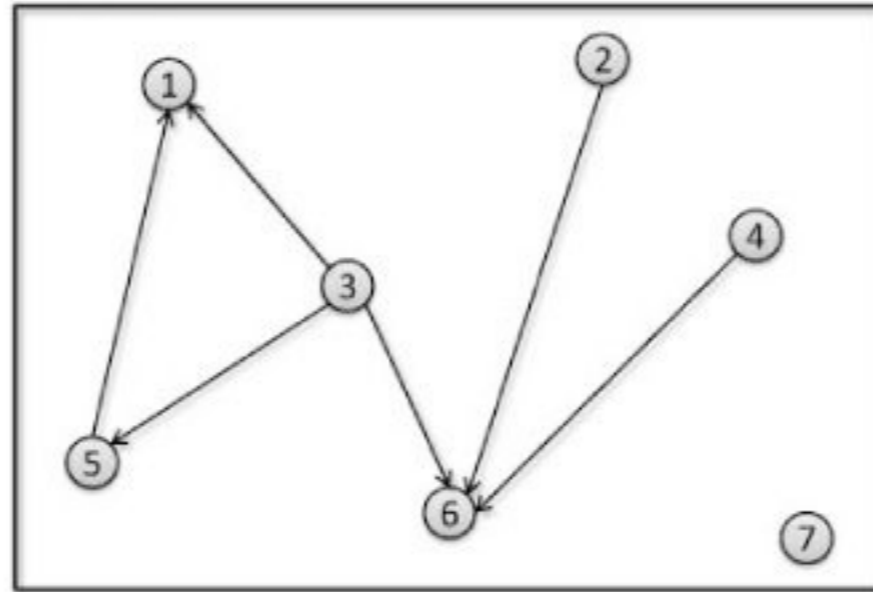
```
1: procedure HITS( $G = (V, E), K$ )
2:    $A_0(p) \leftarrow 1 \forall p \in V$ 
3:    $H_0(p) \leftarrow 1 \forall p \in V$ 
4:   for  $i = 1$  to  $K$  do
5:      $A_i(p) \leftarrow 0 \forall p \in V$ 
6:      $H_i(p) \leftarrow 0 \forall p \in V$ 
7:      $Z_A \leftarrow 0$ 
8:      $Z_H \leftarrow 0$ 
9:     for  $p \in V$  do
10:      for  $q \in V$  do
11:        if  $(p, q) \in E$  then
12:           $H_i(p) \leftarrow H_i(p) + A_{i-1}(q)$ 
13:           $Z_H \leftarrow Z_H + A_{i-1}(q)$ 
14:        end if
15:        if  $(q, p) \in E$  then
16:           $A_i(p) \leftarrow A_i(p) + H_{i-1}(q)$ 
17:           $Z_A \leftarrow Z_A + H_{i-1}(q)$ 
18:        end if
19:      end for
20:    end for
21:    for  $p \in V$  do
22:       $A_i(p) \leftarrow \frac{A_i(p)}{Z_A}$ 
23:       $H_i(p) \leftarrow \frac{H_i(p)}{Z_H}$ 
24:    end for
25:  end for
26:  return  $A_K, H_K$ 
27: end procedure
```

---

# HITS Iteration



# Clustering Nodes



Node:

1 2 3 4 5 6 7

Vector:

$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$
---	---	---	---	---	---	---

# **EM Clustering (for $k$ clusters)**

# EM Clustering (for k clusters)

- EM algorithm
  - Viterbi version – called “k-means clustering”
  - Full EM version – called “Gaussian mixtures”

# EM Clustering (for k clusters)

- EM algorithm
  - Viterbi version – called “k-means clustering”
  - Full EM version – called “Gaussian mixtures”

# EM Clustering (for k clusters)

- EM algorithm
  - Viterbi version – called “k-means clustering”
  - Full EM version – called “Gaussian mixtures”
- **Expectation step:** Use current parameters (and observations) to reconstruct hidden structure



# EM Clustering (for k clusters)

- EM algorithm
  - Viterbi version – called “k-means clustering”
  - Full EM version – called “Gaussian mixtures”
- **Expectation step:** Use current parameters (and observations) to reconstruct hidden structure
- **Maximization step:** Use that hidden structure (and observations) to reestimate parameters

# EM Clustering (for k clusters)

- EM algorithm
  - Viterbi version – called “k-means clustering”
  - Full EM version – called “Gaussian mixtures”
- **Expectation step:** Use current parameters (and observations) to reconstruct hidden structure
- **Maximization step:** Use that hidden structure (and observations) to reestimate parameters

# EM Clustering (for k clusters)

- EM algorithm
  - Viterbi version – called “k-means clustering”
  - Full EM version – called “Gaussian mixtures”
- **Expectation step:** Use current parameters (and observations) to reconstruct hidden structure
- **Maximization step:** Use that hidden structure (and observations) to reestimate parameters
- **Parameters:** k points representing cluster centers

# EM Clustering (for k clusters)

- EM algorithm
  - Viterbi version – called “k-means clustering”
  - Full EM version – called “Gaussian mixtures”
- **Expectation step:** Use current parameters (and observations) to reconstruct hidden structure
- **Maximization step:** Use that hidden structure (and observations) to reestimate parameters
- **Parameters:** k points representing cluster centers
- **Hidden structure:** for each data point (word type), which center generated it?

# Searching with Communities

- Identify communities
  - ❖ Graph clustering
  - ❖ Hypertext induced topic search
- Exploiting community knowledge
  - ❖ Authorities in HITS graph
  - ❖ Explicit question answering and feedback
  - ❖ Bonus: Training data for retrieval models

# Community QA



stackoverflow

Questions Tags Users Badges Unanswered

Search Results **relevance** newest votes active

PageRank search

Want better search results? [See our search tips!](#)

---

**3** votes  
**3** answers  
116 views

**What are some pagerank alternatives?**  
... (for example, pagerank knows who's linking(edges) to whom)

algorithm machine-learning graph-theory graph-algorithm

asked Jan 8 at 21:44  
 **Lostsoul**  
1,273 ●6●20

---

**3** votes  
**1** answer  
537 views

**getting pagerank**  
... how can I retrieve pagerank of any page indexed by google ? Has google any api or page for this ??

java c++ google pagerank search-engines

asked Mar 21 '10 at 20:46  
 **oneat**  
873 ●2●18

---

**2** votes  
**3** answers  
2k views

**pagerank implementation in java**  
... I am looking for a java implementation of the pagerank algorithm.

java jung

asked Mar 1 '10 at 4:40  
 **Larry**  
11 ●2

# Community QA



## What are some pagerank alternatives?

**CAREERS 2.0** by stackoverflow + Have projects on BitBucket? Import them easily to your profile

▲ 3 This is strictly related to the graph algorithm(not SEO or anything). I'm interested in knowing if there are other algorithms out there that solely use the structure of a graph(not content like keywords, etc) to make inferences?

▼ ☆ 2 So for example, if your given a large graph full of nodes how can you make inferences assuming you have no idea what the values within the nodes actually mean(for example, pagerank knows who's linking(edges) to whom and doesn't know anything about the content itself)?

👤 This is not exclusive to web searching, anything that uses graph structure to make inferences.

📌 [algorithm](#) [machine-learning](#) [graph-theory](#) [graph-algorithm](#)

🔗 [link](#) | [improve this question](#)

🕒 edited [Jan 8 at 22:51](#)

asked [Jan 8 at 21:44](#)  
 [Lostsoul](#)  
1,273 ● 6 ● 20  
99% accept rate

[feedback](#)

## 3 Answers

[active](#) [oldest](#) **[votes](#)**

▲ 3 As well as HITS [as suggested by [@larsmans](#)], there is also [SALSA](#), which is considered more "stable" from HITS [and thus is less vulnerable to be affected by spammers].

▼ You are also encourage to have a look at this [survey or ranking algorithms](#)

👍 [link](#) | [improve this answer](#)

answered [Jan 8 at 22:21](#)  
 [amit](#)  
23.4k ● 1 ● 11 ● 40

Twitter use SALSA in their user recommendation technique. – [steve](#) [Jan 10 at 22:15](#)

# Community QA

- Pros

- ❖ Answers to complex information needs
- ❖ Compare multiple opinions
- ❖ Feedback, interaction with others

- Cons

- ❖ Latency
- ❖ All the drawbacks of human interaction



# Community QA

---

- What part of Mexico gets the most tropical storms?  
How do you pronounce the french words, coeur and miel?  
GED test?  
Why do I have to pay this fine?  
What is Schrödinger's cat?  
What's this song?  
Hi...can u ppl tell me sumthing abt death dreams??  
What are the engagement and wedding traditions in Egypt?  
Fun things to do in LA?  
What lessons from the Tao Te Ching do you apply to your everyday life?  
Foci of a hyperbola?  
What should I do today?  
Why was iTunes deleted from my computer?  
Heather Locklear?  
Do people in the Australian Defense Force (RAAF) pay less tax than civilians?  
Whats a psp xmb?  
If  $C(-3, y)$  and  $D(1, 7)$  lie upon a line whose slope is 2, find the value of  $y$ ?  
Why does love make us so irrational?  
Am I in love?  
What are some technologies that are revolutionizing business?
-

# Community QA

---

What part of Mexico gets the most tropical storms?  
How do you pronounce the french words, coeur and miel?  
GED test?

Why do I have to pay this fine?  
What is Schrödinger's cat?  
What's this song?

Hi...can u ppl tell me sumthing abt death dreams??  
What are the engagement and wedding traditions in Egypt?  
Fun things to do in LA?

What lessons from the Tao Te Ching do you apply to your everyday life?  
Foci of a hyperbola?  
What should I do today?

Why was iTunes deleted from my computer?  
Heather Locklear?

Do people in the Australian Defense Force (RAAF) pay less tax than civilians?  
Whats a psp xmb?

If  $C(-3, y)$  and  $D(1, 7)$  lie upon a line whose slope is 2, find the value of  $y$ ?  
Why does love make us so irrational?

Am I in love?

What are some technologies that are revolutionizing business?

---



**GIGO?**

# IR in Communities

- How to match queries to community QA (or forums, or other social media)?
  - ❖ Match query to questions, answers, both?
  - ❖ Generally more effective to match questions
- Questions, and other posts, are short
- More problems with *vocabulary mismatch*

# Retrieval as Translation

# Retrieval as Translation

Naive translation model

$$P(Q|A) = \prod_{w \in Q} \sum_{t \in \mathcal{V}} P(w|t)P(t|A)$$

# Retrieval as Translation

Naive translation model

$$P(Q|A) = \prod_{w \in Q} \sum_{t \in \mathcal{V}} P(w|t)P(t|A)$$

Interpolation with  
language model

$$P(Q|A) = \prod_{w \in Q} \frac{(1 - \beta) f_{w,A} + \beta \sum_{t \in \mathcal{V}} P(w|t) f_{t,A} + \mu \frac{c_w}{|C|}}{|A| + \mu}$$

# Retrieval as Translation

Naive translation model

$$P(Q|A) = \prod_{w \in Q} \sum_{t \in \mathcal{V}} P(w|t)P(t|A)$$

Example “translations”

everest	xp	search
everest	xp	search
mountain	window	google
tallest	install	information
29,035	drive	internet
highest	computer	website
mt	version	web
ft	click	list
measure	pc	free
feet	program	info
mount	microsoft	page

Interpolation with language model

$$P(Q|A) = \prod_{w \in Q} \frac{(1 - \beta) f_{w,A} + \beta \sum_{t \in \mathcal{V}} P(w|t) f_{t,A} + \mu \frac{c_w}{|C|}}{|A| + \mu}$$

# Retrieval as Translation

Naive translation model

$$P(Q|A) = \prod_{w \in Q} \sum_{t \in \mathcal{V}} P(w|t)P(t|A)$$

Interpolation with language model

$$P(Q|A) = \prod_{w \in Q} \frac{(1 - \beta) f_{w,A} + \beta \sum_{t \in \mathcal{V}} P(w|t) f_{t,A} + \mu \frac{c_w}{|C|}}{|A| + \mu}$$

Example “translations”

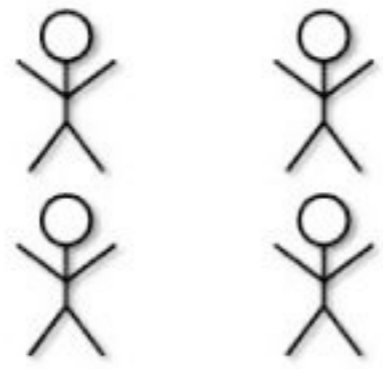
everest	xp	search
everest	xp	search
mountain	window	google
tallest	install	information
29,035	drive	internet
highest	computer	website
mt	version	web
ft	click	list
measure	pc	free
feet	program	info
mount	microsoft	page

How to estimate  $P(w|t)$ ?  
*Q/A pairs!*

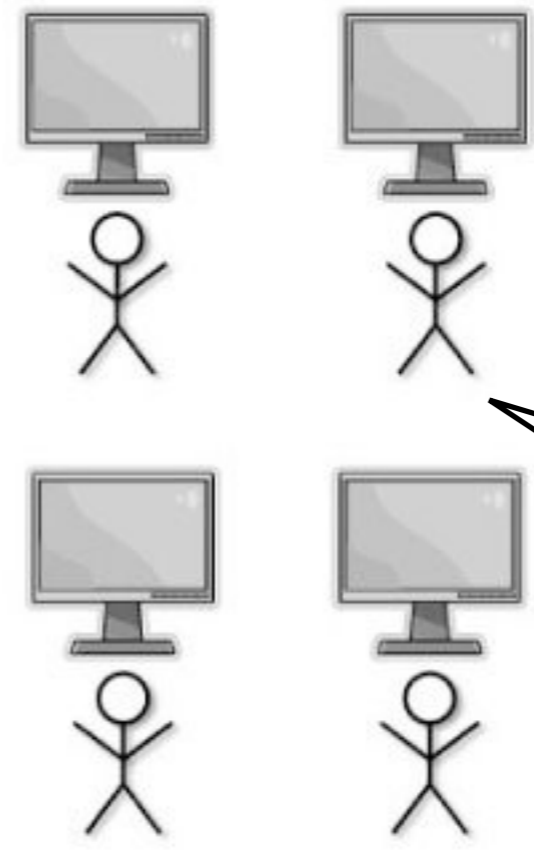


# Collaborative Search

The community searches together



Co-located Collaborative Searching



Shared queries, results, relevance judgments, etc.

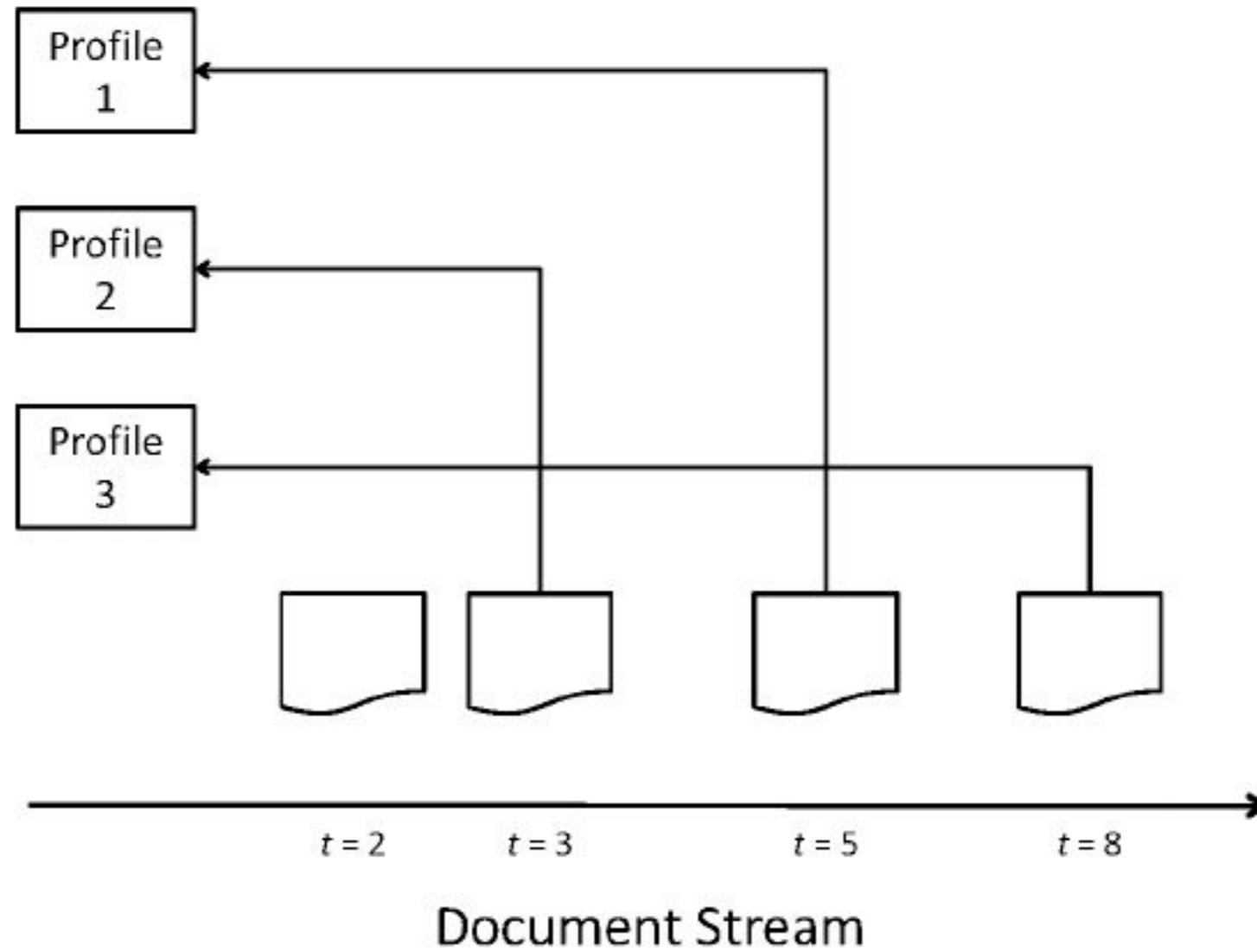
Remote Collaborative Searching

# Filtering and Recommending

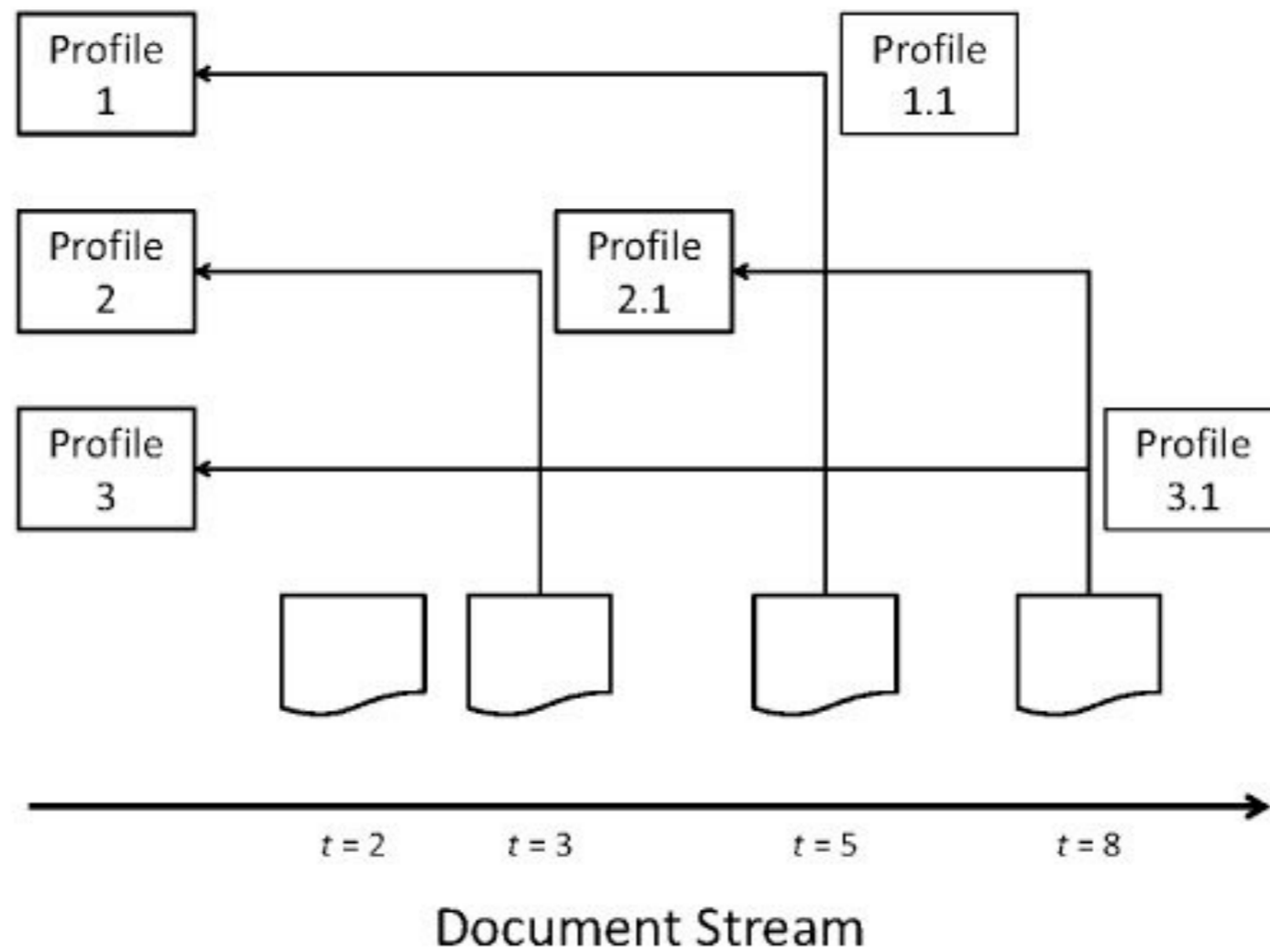
# Filtering and Recommending

- **User profiles** the fundamental data structure
  - \* Profiles can be static or dynamic
  - \* Query features + user/social features
- **Filtering**
  - \* Canonical case: query an endless stream
- **Recommending** (a.k.a. “collaborative filtering”)
  - \* Jointly infer relevance from lots of profiles

# Static Filtering



# Adaptive Filtering



# Filtering Models

# Filtering Models

## Profile model

$$P(w|P) = \frac{(1-\lambda)}{\sum_{i=1}^K \alpha_i} \sum_{i=1}^K \alpha_i \frac{f_{w,T_i}}{|T_i|} + \lambda \frac{c_w}{|C|}$$

# Filtering Models

## Profile model

$$P(w|P) = \frac{(1 - \lambda)}{\sum_{i=1}^K \alpha_i} \sum_{i=1}^K \alpha_i \frac{f_{w,T_i}}{|T_i|} + \lambda \frac{c_w}{|C|}$$

## Document model

$$P(w|D) = (1 - \lambda) \frac{f_{w,D}}{|D|} + \lambda \frac{c_w}{|C|}$$



# Filtering Models

## Profile model

$$P(w|P) = \frac{(1 - \lambda)}{\sum_{i=1}^K \alpha_i} \sum_{i=1}^K \alpha_i \frac{f_{w,T_i}}{|T_i|} + \lambda \frac{c_w}{|C|}$$

## Document model

$$P(w|D) = (1 - \lambda) \frac{f_{w,D}}{|D|} + \lambda \frac{c_w}{|C|}$$

## Relevance (adaptive) model

$$P(w|P) = \frac{1}{|Rel|} \sum_{D_i \in Rel} \sum_{D \in \mathcal{C}} P(w|D) P(D_i|D) \\ \approx \frac{1}{|Rel|} \sum_{D_i \in Rel} P(w|D_i)$$

# Filtering Models

## Profile model

$$P(w|P) = \frac{(1-\lambda)}{\sum_{i=1}^K \alpha_i} \sum_{i=1}^K \alpha_i \frac{f_{w,T_i}}{|T_i|} + \lambda \frac{c_w}{|C|}$$

## Document model

$$P(w|D) = (1-\lambda) \frac{f_{w,D}}{|D|} + \lambda \frac{c_w}{|C|}$$

## Relevance (adaptive) model

$$P(w|P) = \frac{1}{|Rel|} \sum_{D_i \in Rel} \sum_{D \in \mathcal{C}} P(w|D) P(D_i|D) \\ \approx \frac{1}{|Rel|} \sum_{D_i \in Rel} P(w|D_i)$$

## Kullback-Leibler divergence

$$-KL(P||D) = \sum_{w \in V} P(w|P) \log P(w|D) - \sum_{w \in V} P(w|P) \log P(w|P)$$

# Filtering Models

## Profile model

$$P(w|P) = \frac{(1-\lambda)}{\sum_{i=1}^K \alpha_i} \sum_{i=1}^K \alpha_i \frac{f_{w,T_i}}{|T_i|} + \lambda \frac{c_w}{|C|}$$

## Document model

$$P(w|D) = (1-\lambda) \frac{f_{w,D}}{|D|} + \lambda \frac{c_w}{|C|}$$

## Relevance (adaptive) model

$$P(w|P) = \frac{1}{|Rel|} \sum_{D_i \in Rel} \sum_{D \in C} P(w|D) P(D_i|D) \\ \approx \frac{1}{|Rel|} \sum_{D_i \in Rel} P(w|D_i)$$

## Kullback-Leibler divergence

$$-KL(P||D) = \sum_{w \in V} P(w|P) \log P(w|D) - \sum_{w \in V} P(w|P) \log P(w|P)$$

<i>Model</i>	<i>Profile Representation</i>	<i>Profile Updating</i>
Boolean	Boolean Expression	N/A
Vector Space	Vector	Rocchio
Language Modeling	Probability Distribution	Relevance Modeling
Classification	Model Parameters	Online Learning

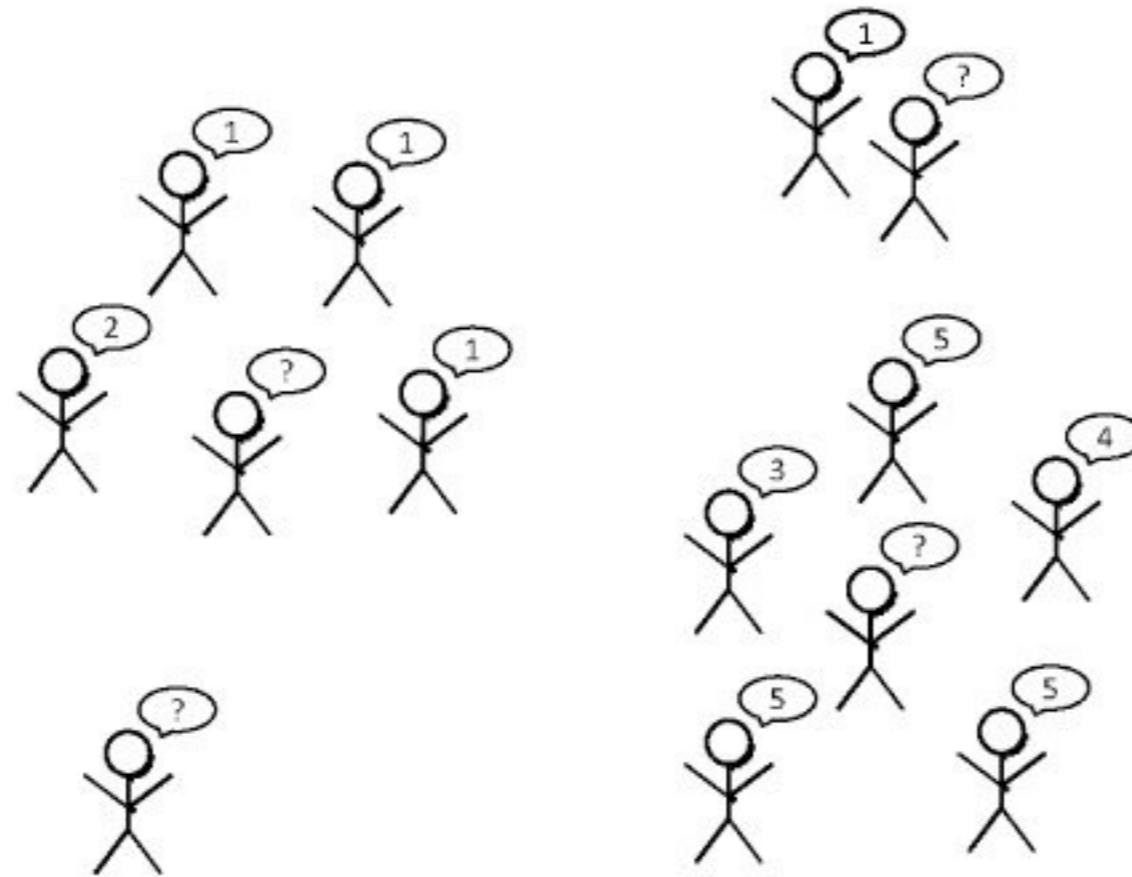
# Scaling and Evaluation

- Treat profiles as “documents” and index them
- Treat incoming documents as queries
- Tradeoffs vary by application

	Relevant	Non-relevant
Retrieved	TP	FP
Not retrieved	FN	TN

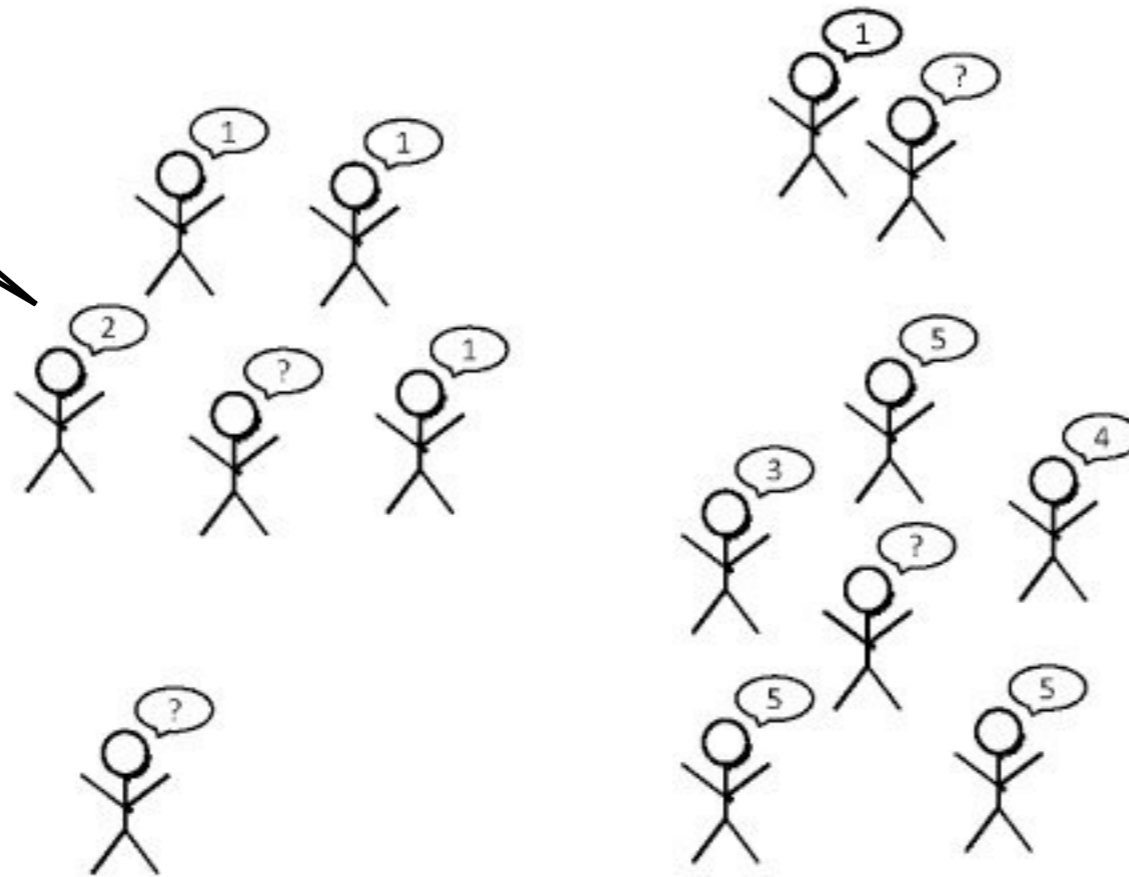
$$U = \alpha \cdot TP + \beta \cdot TN + \delta \cdot FP + \gamma \cdot FN$$

# Recommending



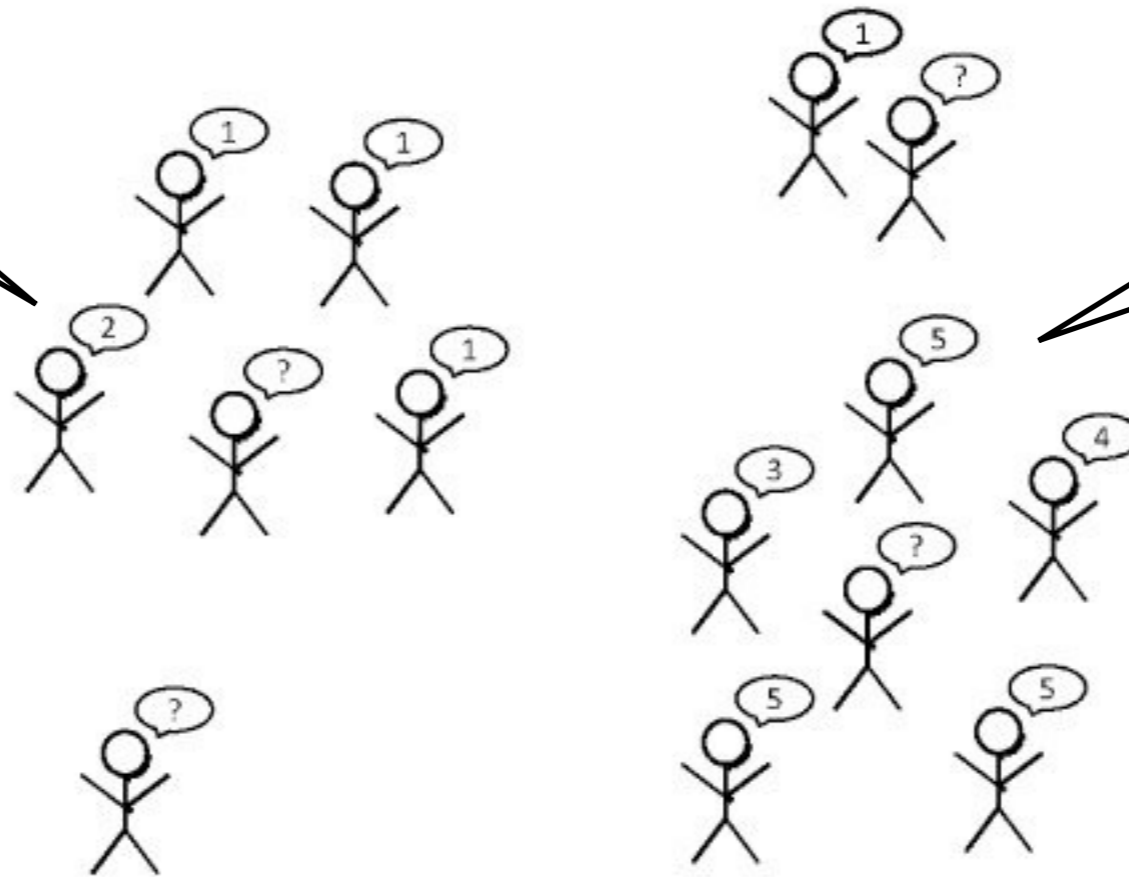
# Recommending

Imagine there's only one document/product



# Recommending

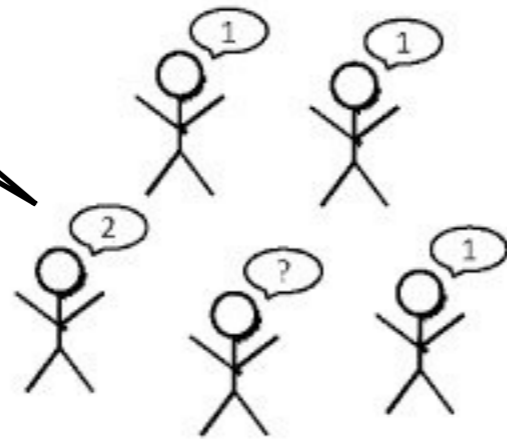
Imagine there's only one document/product



Similar users probably make similar judgments and have similar tastes

# Recommending

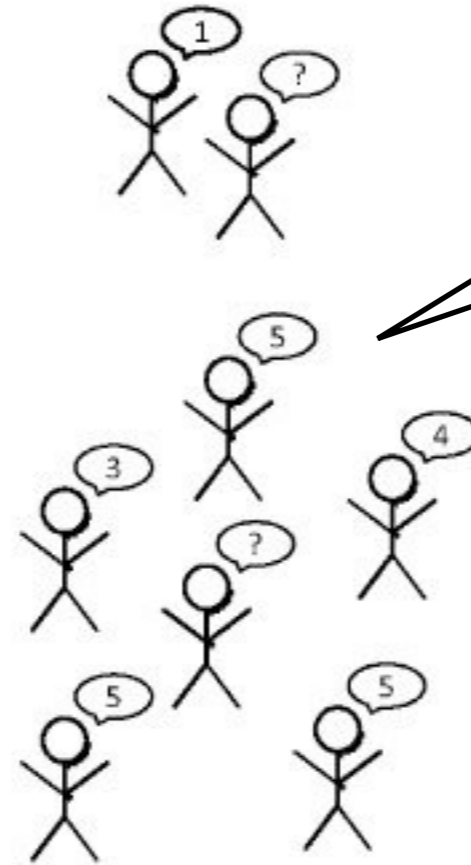
Imagine there's only one document/product



User without mainstream tastes?



Similar users probably make similar judgments and have similar tastes





# Rating by Distance

- Cluster users based on overall rating similarity (or other features)

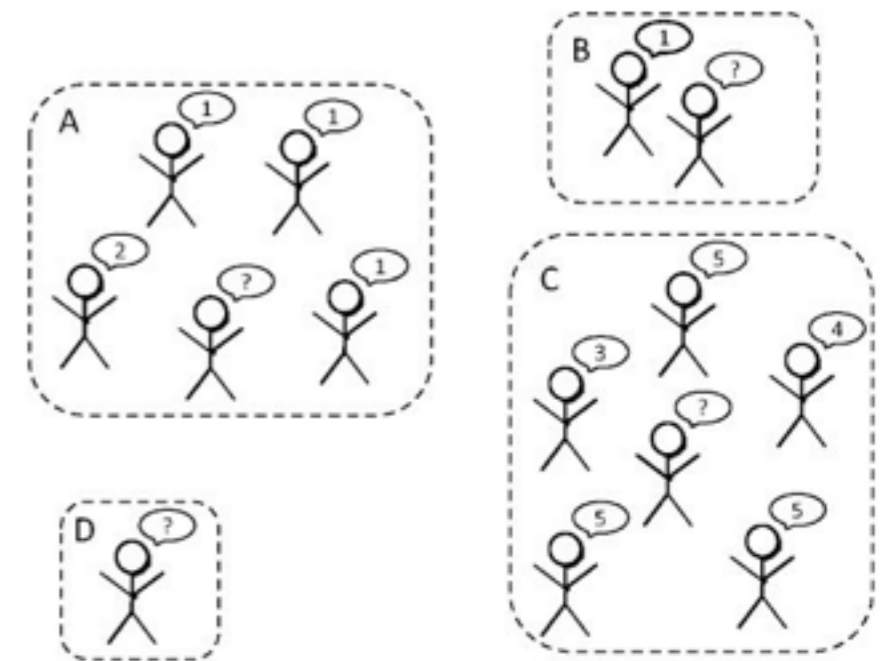
$$\frac{\sum_{i \in I_u \cap I_{u'}} (r_u(i) - \hat{r}_u) \cdot (r_{u'}(i) - \hat{r}_{u'})}{\sqrt{\sum_{i \in I_u \cap I_{u'}} (r_u(i) - \hat{r}_u)^2 \sum_{i \in I_u \cap I_{u'}} (r_{u'}(i) - \hat{r}_{u'})^2}}$$

- ✦ Assign unknown ratings the cluster average

$$\hat{r}_u(i) = \frac{1}{|Cluster(u)|} \sum_{u' \in Cluster(u)} r_{u'}(i)$$

- Average nearest neighbors

$$\hat{r}_u(i) = \bar{r}_u + \frac{1}{\sum_{u' \in \mathcal{N}(u)} sim(u, u')} \sum_{u' \in \mathcal{N}(u)} sim(u, u') (r_{u'}(i) - \bar{r}_{u'})$$



# Evaluating Recommenders

- Exact-match accuracy (usually too harsh)

- Absolute error

$$ABS = \frac{1}{|\mathcal{U}||\mathcal{I}|} \sum_{u \in \mathcal{U}} \sum_{i \in \mathcal{I}} |\hat{r}_u(i) - r_u(i)|$$

- Mean squared error

$$MSE = \frac{1}{|\mathcal{U}||\mathcal{I}|} \sum_{u \in \mathcal{U}} \sum_{i \in \mathcal{I}} (\hat{r}_u(i) - r_u(i))^2$$

- Other task-dependent measures