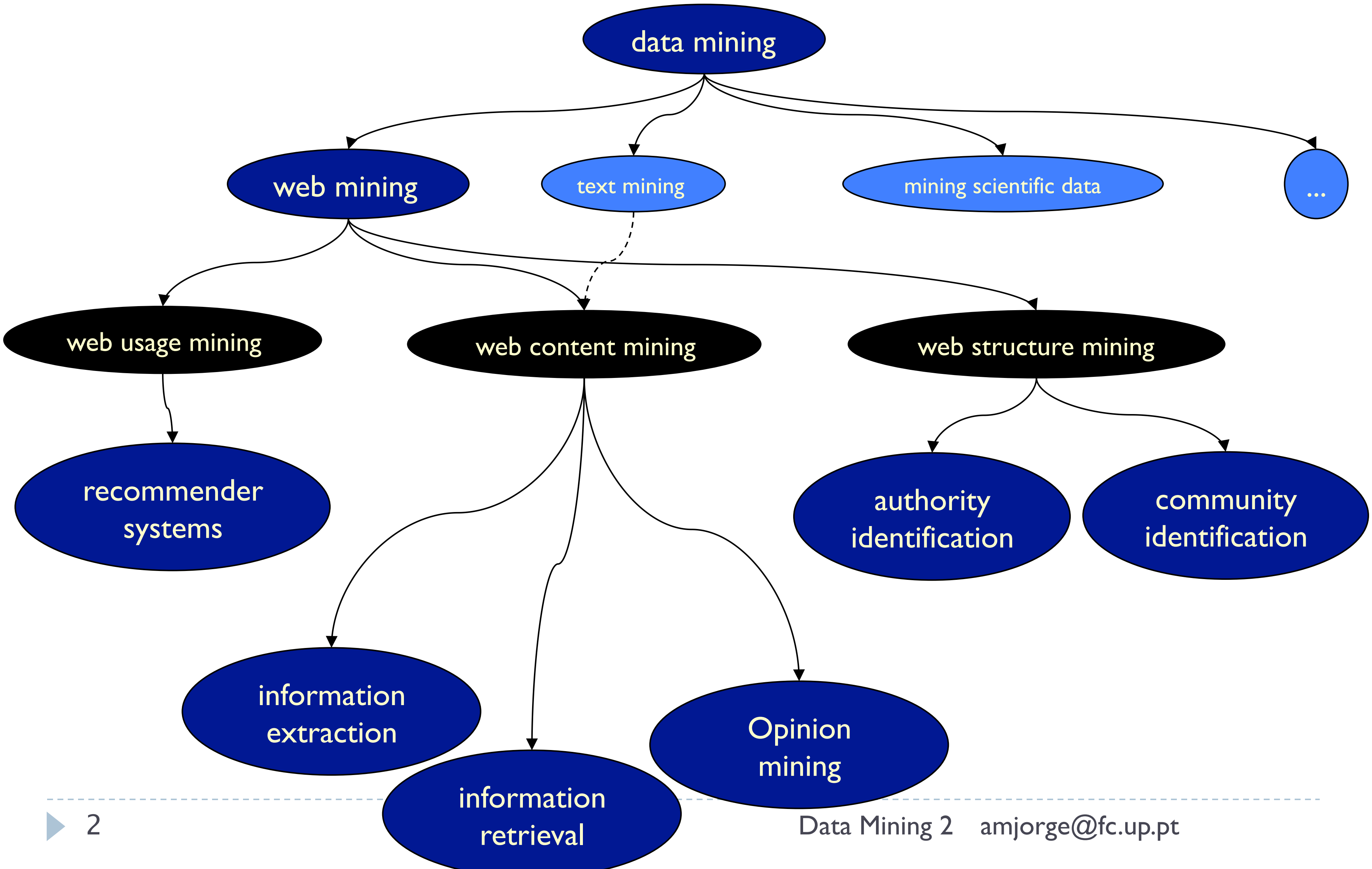
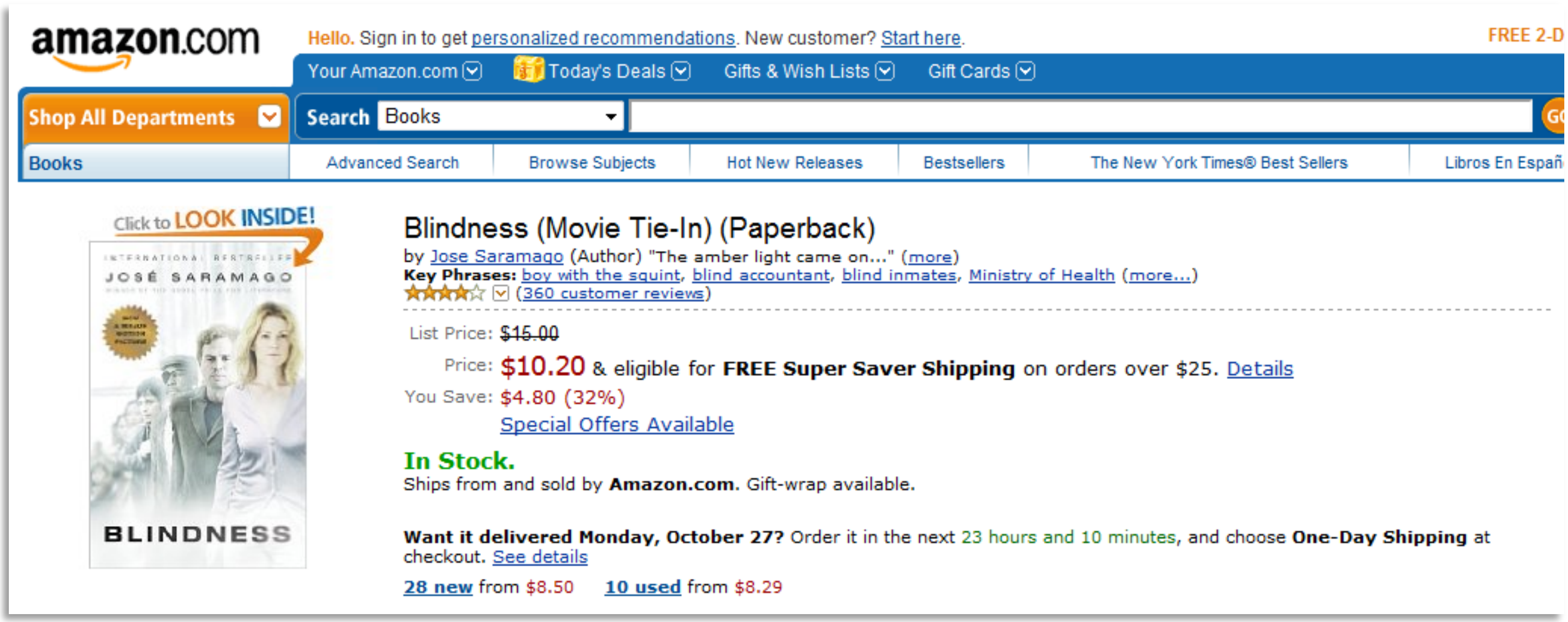
Web Mining: Recommender Systems: Collaborative Filtering: neighbours

Alípio Jorge, DCC-FC, Universidade do Porto
amjorge@fc.up.pt

Knowledge (sort of) tree



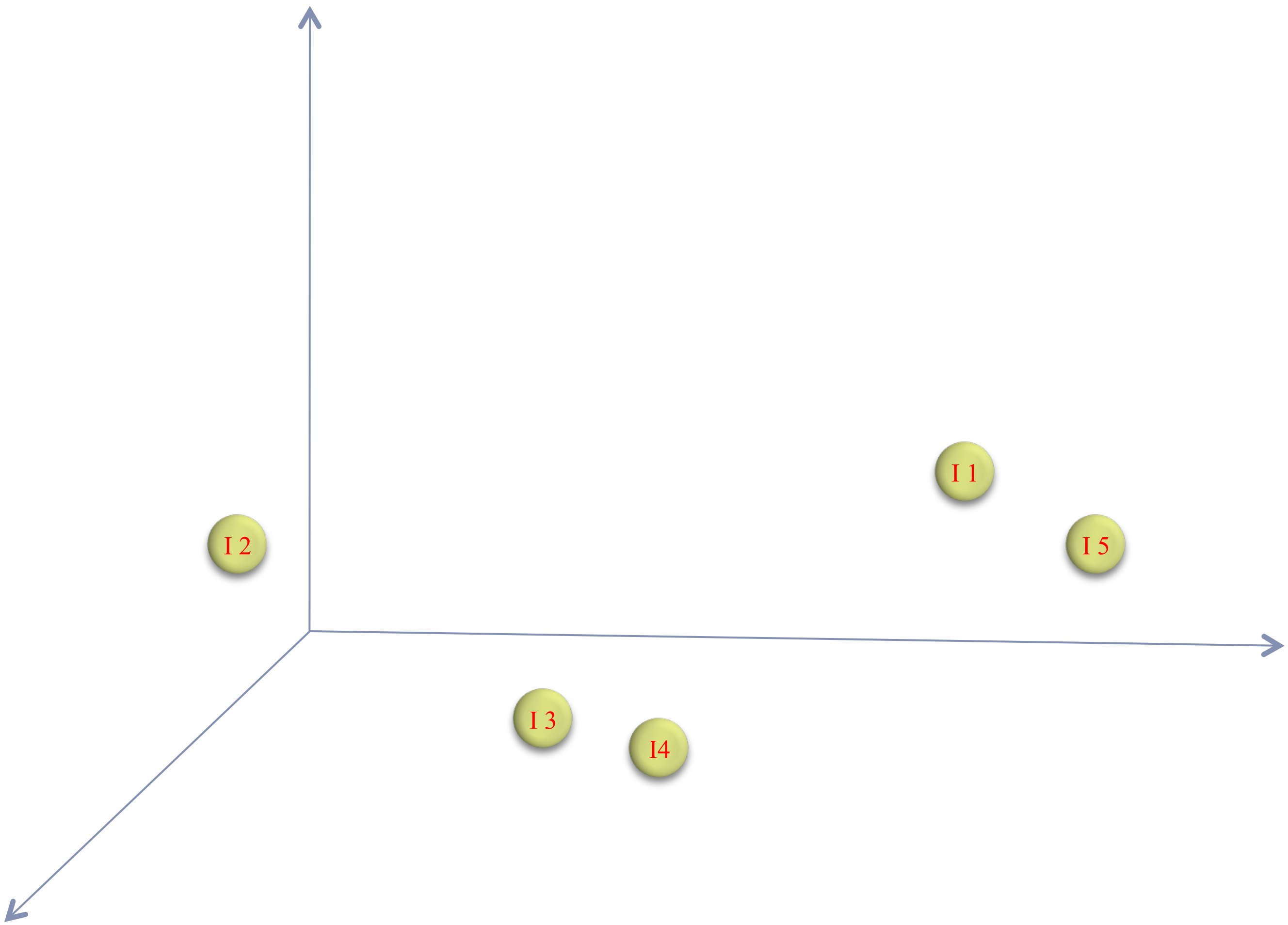

amazon.com
Hello. Sign in to get personalized recommendations. New customer? Start here.
FREE 2-D
Your Amazon.com
Today's Deals
Gifts & Wish Lists
Gift Cards
Shop All Departments
Search Books
Books
Advanced Search
Browse Subjects
Hot New Releases
Bestsellers
The New York Times® Best Sellers
Libros En Españ
Click to LOOK INSIDE!
JOSÉ SARAMAGO
BLINDNESS
Blindness (Movie Tie-In) (Paperback)
by Jose Saramago (Author) "The amber light came on..." (more)
Key Phrases: boy with the squint, blind accountant, blind inmates, Ministry of Health (more...)
(360 customer reviews)
List Price: $15.00
Price: $10.20 & eligible for FREE Super Saver Shipping on orders over $25. Details
You Save: $4.80 (32%)
Special Offers Available
In Stock.
Ships from and sold by Amazon.com. Gift-wrap available.
Want it delivered Monday, October 27? Order it in the next 23 hours and 10 minutes, and choose One-Day Shipping at checkout. See details
28 new from $8.50
10 used from $8.29


Customers Who Bought This Item Also Bought
Seeing by Jose Saramago (27)
The Gospel According to Jesus Christ by Jose Saramago (76) $10.20
All the Names by Jose Saramago (57) $11.20
The Cave by Jose Saramago (48)
Baltasar and Blimunda by Jose Saramago (42) $10.20

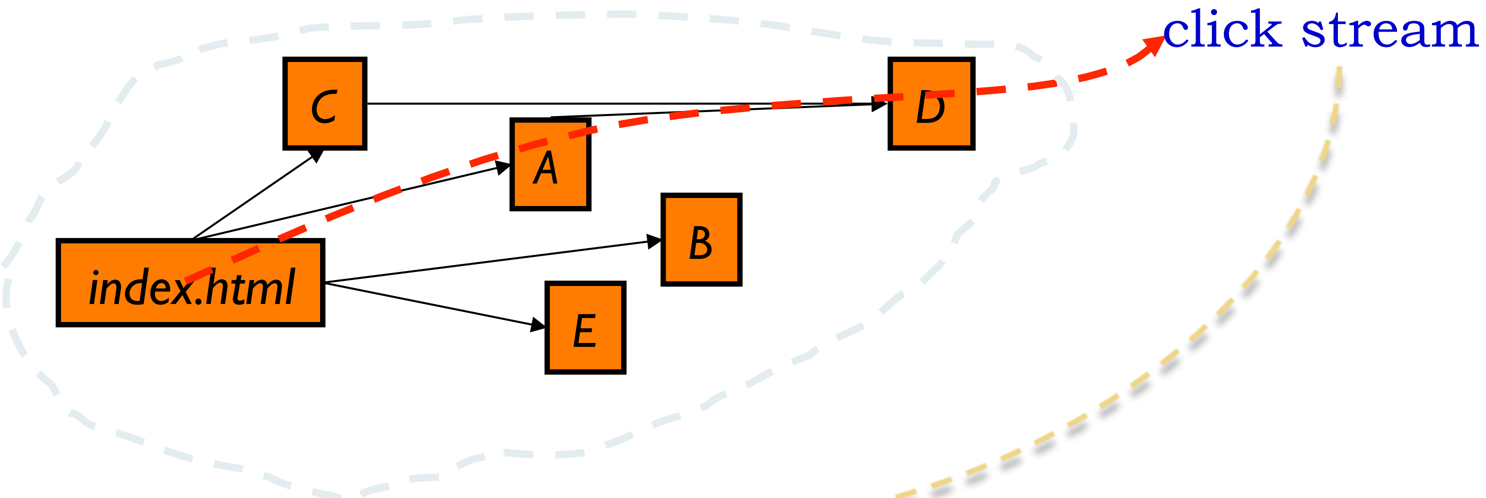
Collaborative Filtering

distance based methods

Collaborative Filtering (the idea: item based)



Collaborative Filtering (item based)



Obs.: A D

Sim. Matrix

	A	B	C	D	E	F
A	1	.1	0	.6	.3	.5
B		1	0	.2	.2	0
C			1	0	.7	.1
D				1	.5	.7
E					1	.3
F						1

Recommendations (top 2):

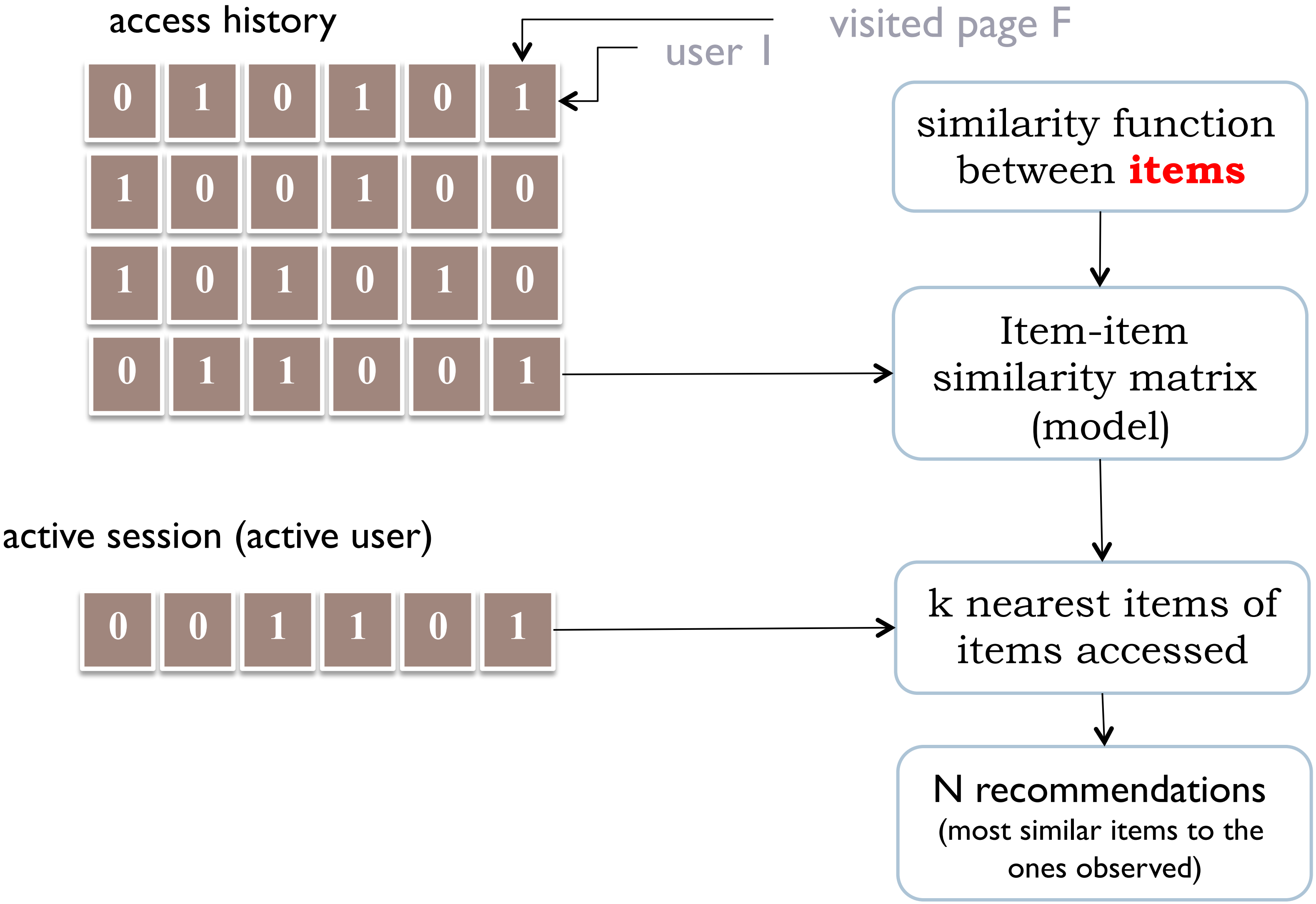
F (0.6)

E (0.4)

Collaborative Filtering Issues

- Binary…
  - web: accessed/didn't access
  - e-commerce: bought / didn't buy
- … vs. non-binary ratings
  - movies: five star system

Collaborative Filtering (item based)



Similarity measures

- Cosine

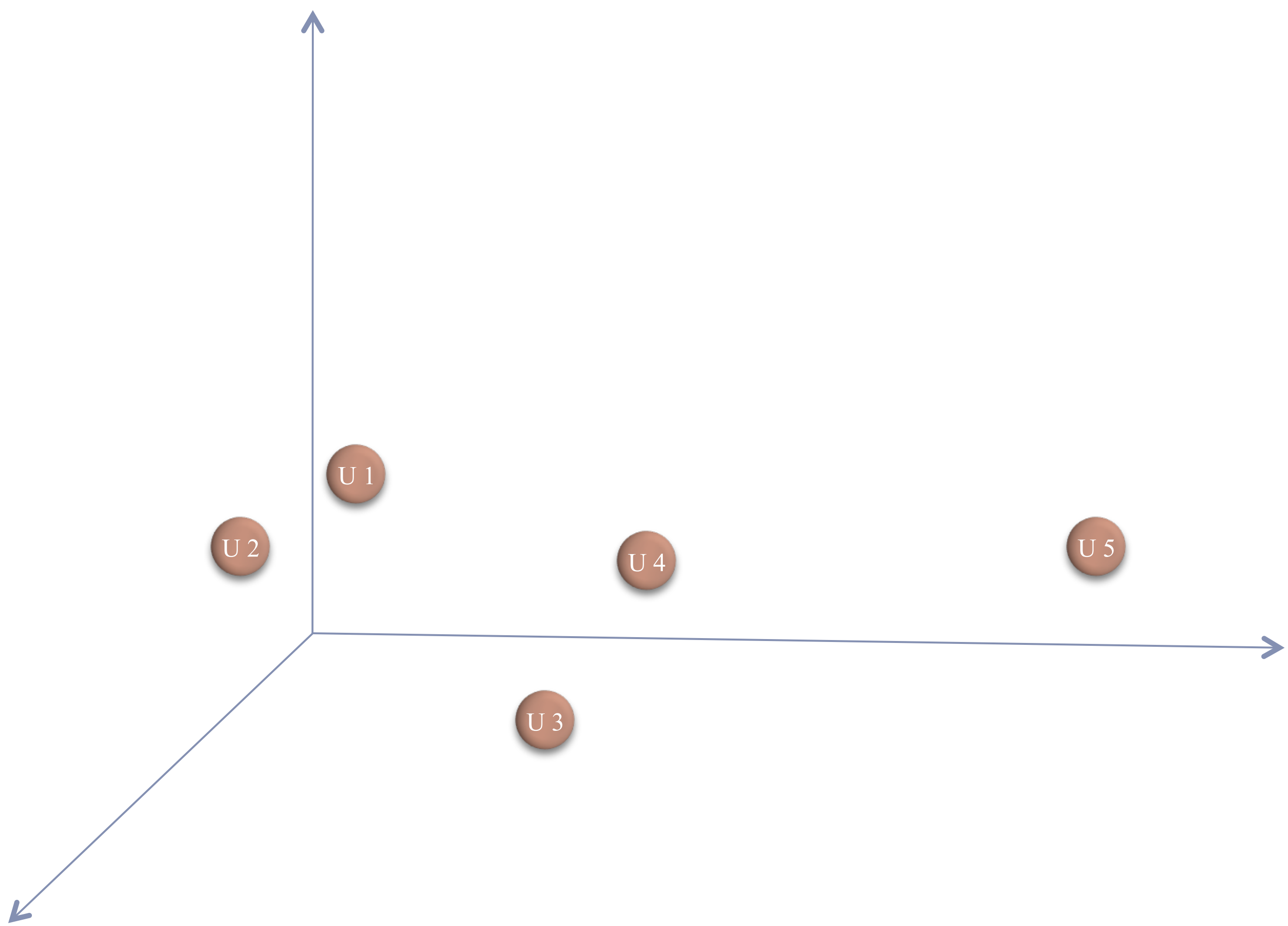
$$sim(i,j) = cos(\overrightarrow{i}, \overrightarrow{j}) = \frac{\#(I \cap J)}{\sqrt{\#I} \times \sqrt{\#J}}$$

- Pearson, Jaccard, ...

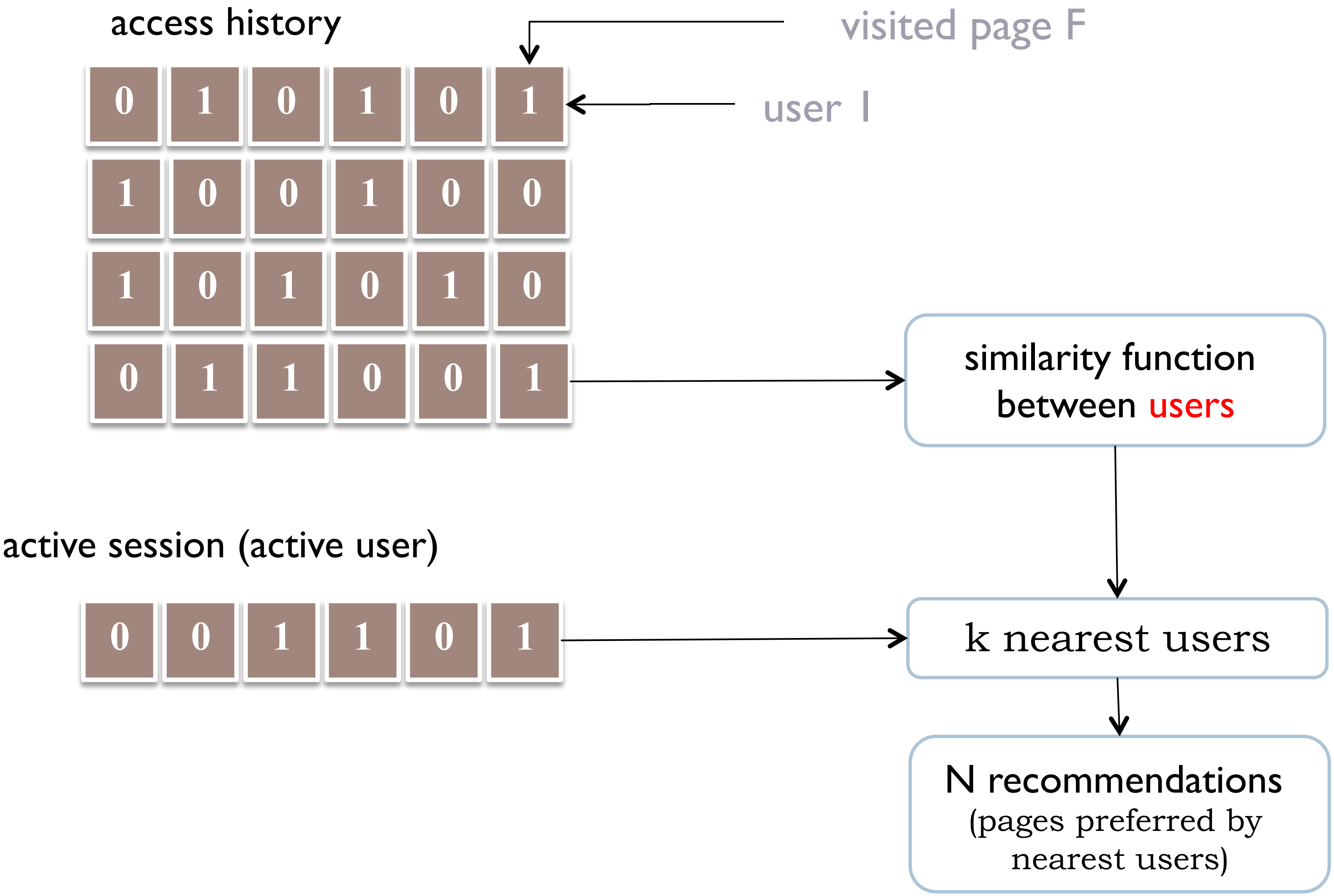
Collaborative Filtering Issues

- User based vs. Item based

Collaborative Filtering (the idea: user based)



Collaborative Filtering (user based)



Similarity measures

- Cosine - user based

$$sim(u,w) = cos(\overrightarrow{u}, \overrightarrow{w}) = \frac{\#(U \cap W)}{\sqrt{\#U} \times \sqrt{\#W}}$$

- Cosine - item based

$$sim(i,j) = cos(\overrightarrow{i}, \overrightarrow{j}) = \frac{\#(I \cap J)}{\sqrt{\#I} \times \sqrt{\#J}}$$

Producing recommendations

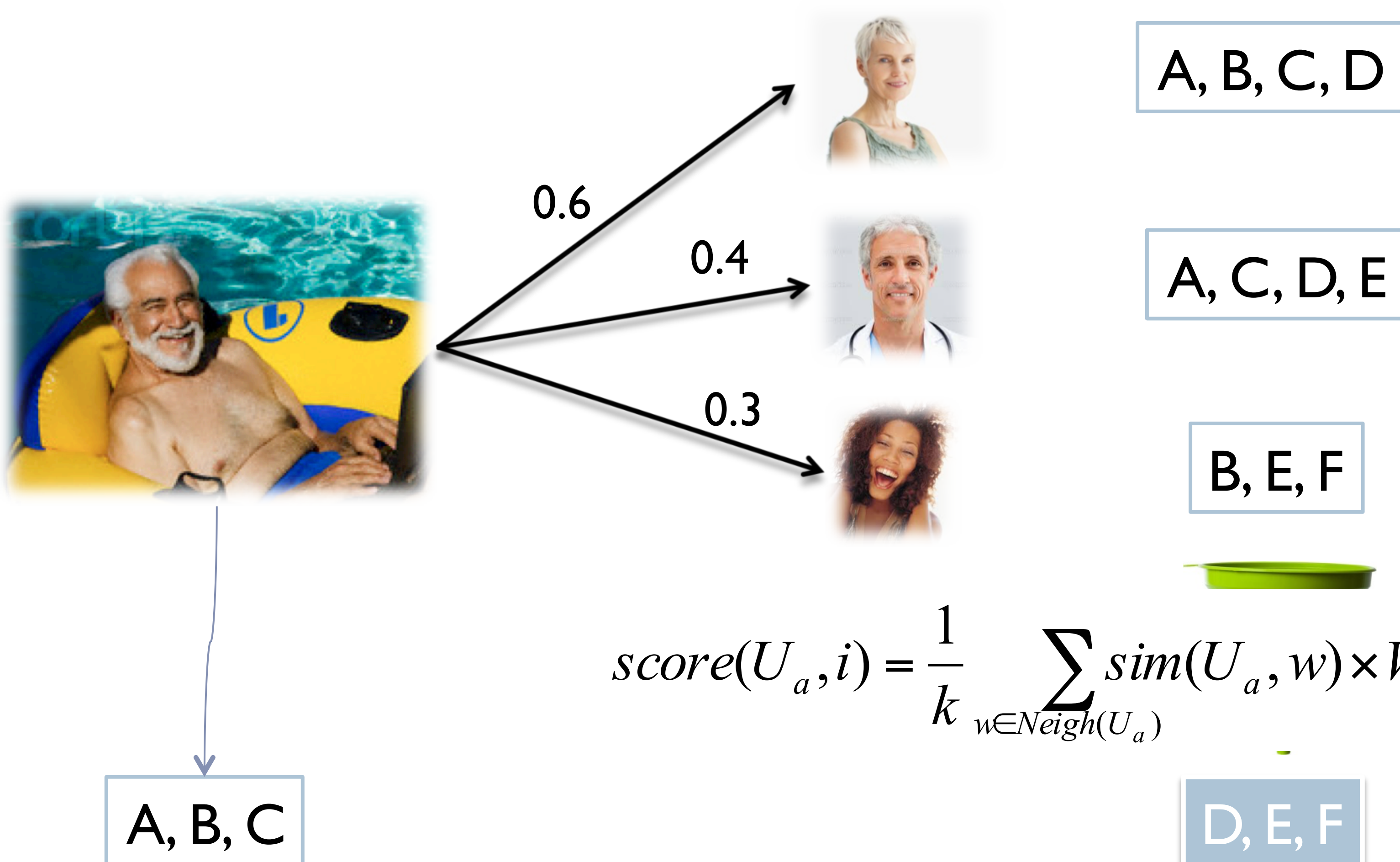
Producing recommendations

- User based
  - given active user U_a
  - find K nearest neighbors of U_a
  - compute the score of each item viewed by the neighbors
  - recommend items with highest score

$$score(U_a, i) = \frac{1}{k} \sum_{w \in Neigh(U_a)} sim(U_a, w) \times Viewed(w, i)$$

Collaborative Filtering - Algorithms

- k-nearest neighbor (user based, binary)



$$score(U_a, i) = \frac{1}{k} \sum_{w \in Neigh(U_a)} sim(U_a, w) \times Viewed(w, i)$$

Producing recommendations

- Item based
  - given the active session S_a
  - compute the score of each item i
    - find its k nearest neighbors
    - consider the intersection of S_a and the neighbors of i
  - recommend items with highest score

$$score(U_a, i) = \frac{\sum_{j \in S_a \cap Neigh(i)} sim(i,j)}{\sum_{j \in Neigh(i)} sim(i,j)}$$

Activity

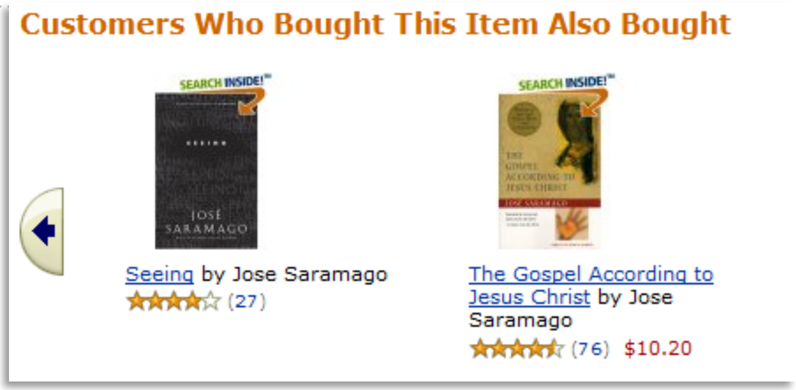
USER	PAGE
1	A
1	B
1	C
2	A
2	C
3	B
3	G
3	F
3	I
4	B
4	C
5	G
5	F
5	I
5	J
6	A
6	C

- Build the similarity (cos) matrix
  - for the user based approach
  - … item-based…
- Compute the recommendations for
  - session B,G
  - user 4

Ratings

Recommendation with ratings

- User give ratings to items
  - 5 star scale
  - or any numeric scale S



- Problem
  - predict the rating a user u in U
  - would give to an unseen item i in I

$$f : U \times I \rightarrow S$$

Recommendation with ratings

- How to recommend?
  - given an active user u_a
  - find items that maximize $f(u_a, i)$

$$top\,relevant\,item = \arg\max_{j \in I \setminus I_u} f(u_a, j)$$

Recommendation with ratings

USER	PAGE	RATING
1	A	1
1	B	3
1	C	2
2	A	4
2	C	2
3	B	4
3	G	5
3	F	3
3	I	4
4	B	5
4	C	4
5	G	3
5	F	4
5	I	5
5	J	3
6	A	5
6	C	3

- How would u2 rate B ?

Recommendation with ratings

- Methods
  - k-nearest neighbor
  - use knn who have rated the item

$$\hat{r}_{ui} = \frac{1}{|N_i(u)|} \sum_{v \in N_i(u)} r_{vi}$$

  - N_i(u) = neighbors of u who rated i
  - is this user-based or item-based?

Recommendation with ratings

USER	PAGE	RATING
1	A	1
1	B	3
1	C	2
2	A	4
2	C	2
3	B	4
3	G	5
3	F	3
3	I	4
4	B	5
4	C	4
5	G	3
5	F	4
5	I	5
5	J	3
6	A	5
6	C	3

- How would u2 rate B ?

```
> dm<-table(d$USER,d$PAGE)
> dm

  A B C F G I J
1 1 1 1 0 0 0 0
2 1 0 1 0 0 0 0
3 0 1 0 1 1 1 0
4 0 1 1 0 0 0 0
5 0 0 0 1 1 1 1
6 1 0 1 0 0 0 0
```

Recommendation with ratings

USER	PAGE	RATING
1	A	1
1	B	3
1	C	2
2	A	4
2	C	2
3	B	4
3	G	5
3	F	3
3	I	4
4	B	5
4	C	4
5	G	3
5	F	4
5	I	5
5	J	3
6	A	5
6	C	3

- How would u2 rate B (2 neighbors)?

```
> dism<-
  round(as.matrix(dist(dm)),2)
     1    2    3    4    5    6
1 0.00 1.00 2.24 1.00 2.65 1.00
2 1.00 0.00 2.45 1.41 2.45 0.00
3 2.24 2.45 0.00 2.00 1.41 2.45
4 1.00 1.41 2.00 0.00 2.45 1.41
5 2.65 2.45 1.41 2.45 0.00 2.45
6 1.00 0.00 2.45 1.41 2.45 0.00

> sort(dism[2,-2])
   6    1    4    3    5
0.00 1.00 1.41 2.45 2.45
```

Recommendation with ratings

USER	PAGE	RATING
1	A	1
1	B	**3**
1	C	2
2	A	4
2	C	2
3	B	4
3	G	5
3	F	3
3	I	4
4	B	**5**
4	C	4
5	G	3
5	F	4
5	I	5
5	J	3
6	A	5
6	C	3

- How would u2 rate B (2 neighbors)?

```
> (3+5)/2
[1] 4
```

Recommendation with ratings

USER	PAGE	RATING
1	A	1
1	B	**3**
1	C	2
2	A	4
2	C	2
3	B	4
3	G	5
3	F	3
3	I	4
4	B	**5**
4	C	4
5	G	3
5	F	4
5	I	5
5	J	3
6	A	5
6	C	3

- How would u2 rate B (2 neighbors)?
  - this method is not sensitive to distance
  - u1 is more similar to u2 than u4

```
> sort(dism[2,-2])
   6    1    4    3    5
0.00 1.00 1.41 2.45 2.45

> (1.41*5+1*3)/(1+1.41)
[1] 4.17
```

Recommendation with ratings

- Methods
  - k-nearest neighbor (weighed)
  - use knn who have rated the item

$$\hat{r}_{ui} = \frac{\sum_{v \in N_i(u)} w_{uv} r_{vi}}{\sum_{v \in N_i(u)} |w_{uv}|}$$

Recommendation with ratings

USER	PAGE	RATING
1	A	1
1	B	3
1	C	2
2	A	4
2	C	2
3	B	4
3	G	5
3	F	3
3	I	4
4	B	5
4	C	4
5	G	3
5	F	4
5	I	5
5	J	3
6	A	5
6	C	3

- How would u2 rate B ?
  - Using ratings to profile users

```
> for(i in 1:nrow(d))
      dm[d$USER[i],d$PAGE[i]]
      <-d$RATING[i]
> dm

   A B C F G I J
1  1 3 2 0 0 0 0
2  4 0 2 0 0 0 0
3  0 4 0 3 5 4 0
4  0 5 4 0 0 0 0
5  0 0 0 4 3 5 3
6  5 0 3 0 0 0 0
```

Recommendation with ratings

USER	PAGE	RATING
1	A	1
1	B	3
1	C	2
2	A	4
2	C	2
3	B	4
3	G	5
3	F	3
3	I	4
4	B	5
4	C	4
5	G	3
5	F	4
5	I	5
5	J	3
6	A	5
6	C	3

- How would u2 rate B (2 neighbors)?

```
> dism<-round(as.matrix(dist(dm)),2)
> dism
     1    2     3     4     5     6
1 0.00 4.24  7.48  3.00  8.54  5.10
2 4.24 0.00  9.27  6.71  8.89  1.41
3 7.48 9.27  0.00  8.19  5.57 10.00
4 3.00 6.71  8.19  0.00 10.00  7.14
5 8.54 8.89  5.57 10.00  0.00  9.64
6 5.10 1.41 10.00  7.14  9.64  0.00
```

Recommendation with ratings

USER	PAGE	RATING
1	A	1
1	B	3
1	C	2
2	A	4
2	C	2
3	B	4
3	G	5
3	F	3
3	I	4
4	B	5
4	C	4
5	G	3
5	F	4
5	I	5
5	J	3
6	A	5
6	C	3

- How would u2 rate B (2 neighbors)?

```
> simm<-max(dism)-dism
> simm
      1     2     3     4     5     6
1 10.00  5.76  2.52  7.00  1.46  4.90
2  5.76 10.00  0.73  3.29  1.11  8.59
3  2.52  0.73 10.00  1.81  4.43  0.00
4  7.00  3.29  1.81 10.00  0.00  2.86
5  1.46  1.11  4.43  0.00 10.00  0.36
6  4.90  8.59  0.00  2.86  0.36 10.00

> sort(simm[2,-2],decreasing=T)
   6    1    4    5    3
8.59 5.76 3.29 1.11 0.73
```

Recommendation with ratings

USER	PAGE	RATING
1	A	1
1	B	**3**
1	C	2
2	A	4
2	C	2
3	B	4
3	G	5
3	F	3
3	I	4
4	B	**5**
4	C	4
5	G	3
5	F	4
5	I	5
5	J	3
6	A	5
6	C	3

- How would u2 rate B (2 neighbors)?

```
> (3+5)/2
[1] 4
```

Recommendation with ratings

USER	PAGE	RATING
1	A	1
1	B	**3**
1	C	2
2	A	4
2	C	2
3	B	4
3	G	5
3	F	3
3	I	4
4	B	**5**
4	C	4
5	G	3
5	F	4
5	I	5
5	J	3
6	A	5
6	C	3

- How would u2 rate B (2 neighbors)?
  - this method is not sensitive to distance
  - u1 is more similar to u2 than u4

```
> sort(simm[2,-2],decreasing=T)
   6    1    4    5    3
8.59 5.76 3.29 1.11 0.73

> (5.76*3+3.29*5)/(5.76+3.29)
[1] 3.73
```

Recommendation with ratings

USER	PAGE	RATING
1	A	1
1	B	**3**
1	C	2
2	A	4
2	C	2
3	B	4
3	G	5
3	F	3
3	I	4
4	B	**5**
4	C	4
5	G	3
5	F	4
5	I	5
5	J	3
6	A	5
6	C	3

- How would u2 rate B (2 neighbors)?
  - what if users have **different perceptions**?
  - u4 is more generous than u1

```
> sort(dism[2,-2])
   6    1    4    3    5
0.00 1.00 1.41 2.45 2.45

> 3+(1.41*0.5+1*1)/(1+1.41)
[1] 3.70
```

Recommendation with ratings

- Methods
  - k-nearest neighbor (weighed and mean-centered)
  - clip if value is out of scale
  - mean-centering is a form of normalization. There are others,

$$\hat{r}_{ui} = \bar{r}_u + \frac{\sum_{v \in N_i(u)} w_{uv}(r_{vi} - \bar{r}_v)}{\sum_{v \in N_i(u)} |w_{uv}|}$$

Recommendation with ratings

- Methods
  - k-nearest neighbor
  - item-based : distances between items

$$\hat{r}_{ui} = \frac{1}{|N_u(i)|} \sum_{j \in N_u(i)} r_{uj}$$

  - consider only items j, neighbor to i, already rated by user
  - Activity: add weights and recenter

Activity

USER	PAGE	RATING
1	A	1
1	B	3
1	C	2
2	A	4
2	C	2
3	B	4
3	G	5
3	F	3
3	I	4
4	B	5
4	C	4
5	G	3
5	F	4
5	I	5
5	J	3
6	A	5
6	C	3

- How would u1 rate F (2 neighbors)?

```
> dism<-
  round(as.matrix(dist(dm)),2)
     1    2    3    4    5    6
1 0.00 1.00 2.24 1.00 2.65 1.00
2 1.00 0.00 2.45 1.41 2.45 0.00
3 2.24 2.45 0.00 2.00 1.41 2.45
4 1.00 1.41 2.00 0.00 2.45 1.41
5 2.65 2.45 1.41 2.45 0.00 2.45
6 1.00 0.00 2.45 1.41 2.45 0.00
```

Activity (mind the ratings)

USER	PAGE	RATING
1	A	1
1	B	3
1	C	2
2	A	4
2	C	2
3	B	4
3	G	5
3	F	3
3	I	4
4	B	5
4	C	4
5	G	3
5	F	4
5	I	5
5	J	3
6	A	5
6	C	3

- How would u1 rate F (2 neighbors)?

```
> simm
      1     2     3     4     5     6
1 10.00  5.76  2.52  7.00  1.46  4.90
2  5.76 10.00  0.73  3.29  1.11  8.59
3  2.52  0.73 10.00  1.81  4.43  0.00
4  7.00  3.29  1.81 10.00  0.00  2.86
5  1.46  1.11  4.43  0.00 10.00  0.36
6  4.90  8.59  0.00  2.86  0.36 10.00
```

Recommender lab

Activity - recommenderlab

```
library(recommenderlab)
help(Jester5k)
data(Jester5k)
d<-Jester5k
show(d)
nrow(d)
ncol(d)
as(d[1:3,], "list")
as(d[101:103,], "matrix")
d<-sample(d,1000)
```

Activity - recommenderlab

```
# look at distribution of ratings
hist(getRatings(d),breaks=100)
# and other data characteristics
mean(rowCounts(d))
hist(rowCounts(d),breaks=50)
hist(rowMeans(d),breaks=50)
hist(colMeans(d),breaks=50)
mean(colCounts(d))
```

Activity - recommenderlab

```
# model generation and use
model<-Recommender(d,method="IBCF")
# test with a user not in the training set
is.element("u4687",rownames(d))
recs <-predict(model,Jester5k["u4687",])
as(recs,"list")
# compare topNlist of IBCF and UBCF
model<-Recommender(d,method="UBCF")
recs <-predict(model,Jester5k["u4687",])
as(recs,"list")
```

Activity - recommenderlab

```
# check prediction of ratings
model<-Recommender(d,method="IBCF")
recs <-
  predict(model,Jester5k["u4687",],type="ratings")
as(recs,"list")
# manipulate one test example
as(Jester5k['u4687',],"list")
tst<-Jester5k["u4687",]
m<-as(tst,"matrix")
m[,c(5,7,27,69)]<-NA
tst<-as(m,"realRatingMatrix")
recs <-predict(model,tst,type="ratings")
as(recs,"list")
```

Activity - recommenderlab

```
# normalize
dn<-normalize(d)
hist(getRatings(dn))
# normalize 2
dn<-normalize(d,method="Z-score")
hist(getRatings(dn),breaks=50)
```

Activity - recommenderlab

```
# read csv in two column format
d<-read.csv(”tiny.bas",sep=" ",header=F)
d<-as(d,"realRatingMatrix")
d<-binarize(m,minRating=1)
as(d,"list")
plot(sort(rowCounts(d)),type="line")
# build model
model<-Recommender(d,"POPULAR")
show(as(model@model$topN,"list"))
```

Activity - recommenderlab

```
# ”make” user
au<-data.frame(V1="u",V2=c("C"))
au<-as(au,"realRatingMatrix")
au<-binarize(au,minRating=1)

# produce recommendations
as(predict(model,au,n=2),"matrix")
```

Activity - recommenderlab

- Observe on a few users the impact of normalization
- Find out which parameters are used by IBCF and UBCF and observe the impact of these.

- Questions:
  - Which score function should we use if we normalize ratings?
  - Apply IBCF to a different data set (even if a toy one).

Challenges

Challenges

- Scalability
- Sparsity
- Incrementality
- Cold start
- Considering context
- Background knowledge
- Combining content, structure and usage

Resources

- Articles
  - Breese, J.S., Heckerman, D., and Kadie, C. Empirical analysis of predictive algorithms for collaborative filtering. In Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence, pages 43--52, July 1998.
  - B. M. Sarwar, G. Karypis, J. A. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In Proc. of the 10th International World Wide Web Conference (WWW01), Hong Kong, May 2001.

Resources

- Articles
  - G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," IEEE Trans. Knowl. Data Eng., vol. 17, 2005, pp. 734-749.
  - G. Adomavicius, R. Sankaranarayanan, S. Sen, and A. Tuzhilin, "Incorporating contextual information in recommender systems using a multidimensional approach," ACM Transactions on Information Systems, vol. 23, 2005, pp. 103-145.
  - N. Good, J.B. Schafer, J.A. Konstan, A. Borchers, B. Sarwar, J. Herlocker, and J. Riedl, "Combining Collaborative Filtering with Personal Agents for Better Recommendations," Artificial Intelligence, 1999.
  - Z. Huang, H. Chen, and D. Zeng, "Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering," ACM Transactions on Information Systems, vol. 22, 2004, pp. 116-142.
  - C. Palmisano, A. Tuzhilin, and M. Gorgoglione, "Using Context to Improve Predictive Modeling of Customers in Personalization Applications," IEEE Transactions on Knowledge and Data Engineering, vol. 20, 2008, pp. 1535-1549.

Resources

- Book
  - Web Data Mining, Bing Liu
  - Recommender Systems Handbook, chapter 4, Springer (Ed. Ricci et al)