Yelp Data Challenge

`https://github.com/apptsunami/yelpdatachallenge`

Phase 2

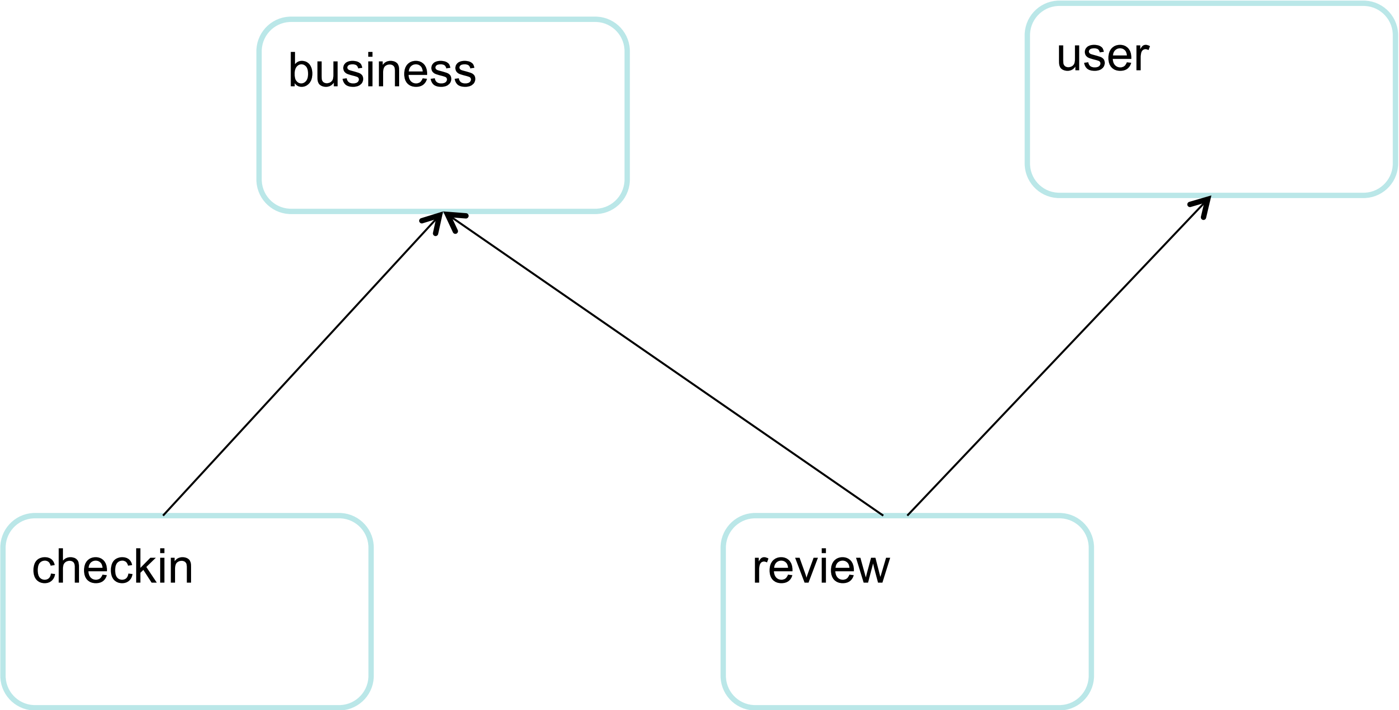
6/3/2013

schan@apptsunami.com

Analysis Phase

- Basic Collaborative Filtering
  - Only use “similar” users’ rating
  - Pearson correlation coefficient
    - Calculated by similarity in ranking of same businesses
- Enhancements
  - Use additional attributes of “similar” users in the prediction formula
  - Corner cases of Pearson correlation coefficient

Yelp Data Model



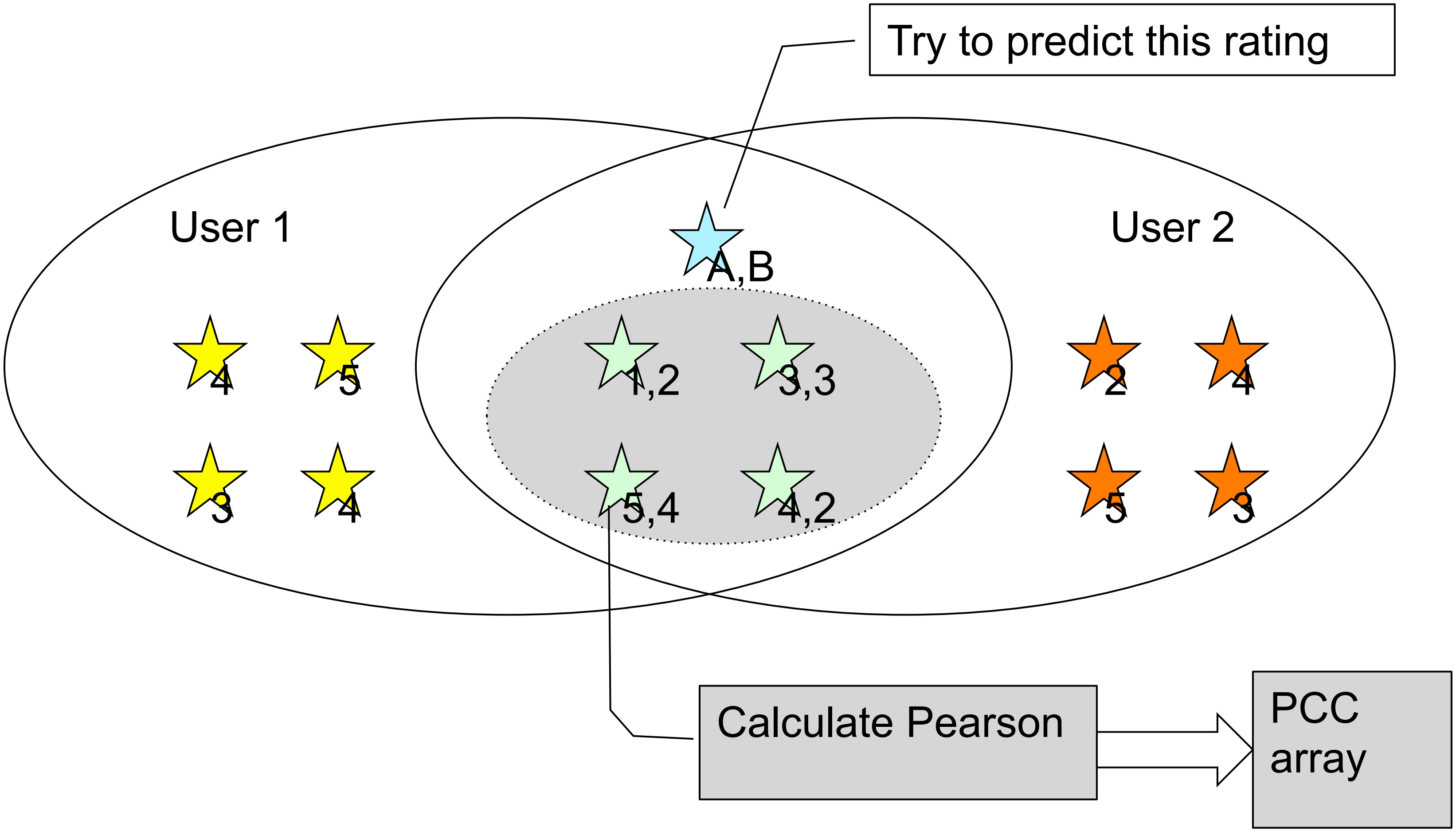
Sample *review* record

```
{"votes": {"funny": 0, "useful": 5, "cool":
2}, "user_id": "rLtl8ZkDX5vH5nAx9C3q5Q",
"review_id": "fWKvX83p0-ka4JS3dc6E5A",
"stars": 5, "date": "2011-01-26", ”text":
"My wife took me here on my birthday for
breakfast and it was excellent. I can't wait
to go back!", "type”: "review",
"business_id": "9yKzy9PApeiPPOUJEtnvkg"}
```

Basic Collaborative Filtering

- For each review (`userId1, businessId, stars`)
  - □ Gather all reviews by the same user
  - □ For each user who reviews the same business (`userId2, businessId`)
    - Gather all reviews by the same user
    - Compute the Pearson correlation coefficient based on businesses ranked by both users
  - □ Calculate a predicted stars
  - □ Calculated the error (predicted stars – stars)
- Calculate the RMS of all errors

Compute “Similarity”

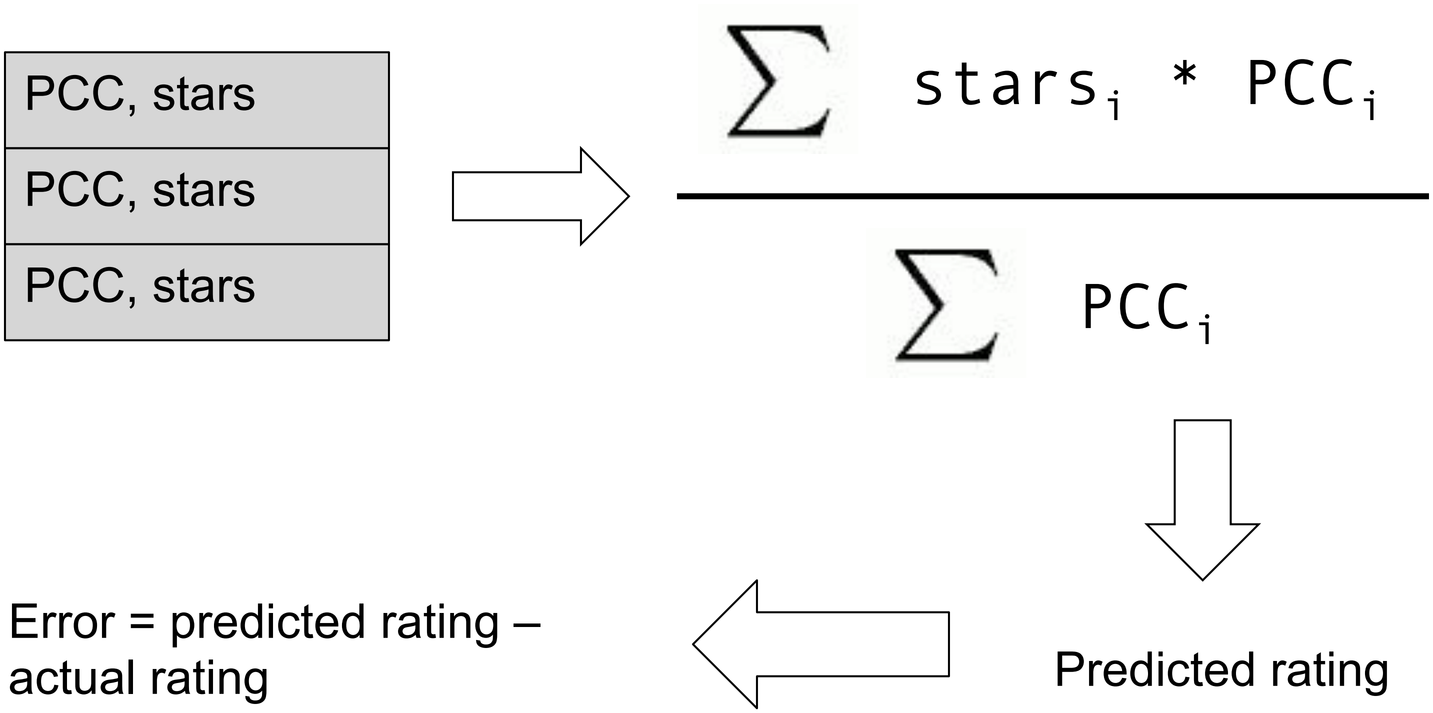


Try to predict this rating

Calculate Pearson

PCC array

Predict Rating



PCC, stars

PCC, stars

PCC, stars

$$\frac{\sum \mathrm{stars}_i * \mathrm{PCC}_i}{\sum \mathrm{PCC}_i}$$

Predicted rating

Error = predicted rating – actual rating

Lessons Learned

- RMS better than random guesses
- Some businesses don’t have many reviews
- Two users often do not have many businesses voted in common
  - Pearson requires at least two data points
- With a small number of businesses ranked often they have the same stars value
  - Pearson does not compute when variance is zero
- Execution speed slows down innovation

Enhancements

- Same gender?
  - Add weight to similarity if both users are of the same gender
- Minimum PCC?
  - Eliminate noise (users with low PCC)
- Corner case of Pearson?
  - Users who rank all businesses the same within the sample set

More Enhancements

- Content-Based filtering
- Business categories
- Usefulness of reviews
- “Cold start” for some businesses and some users
  - Too few ranking

Questions?

- schan@apptsunami.com