Social Tag-Based Recommendation Services

Jordan Bentley

Advisor: Elke Rundensteiner

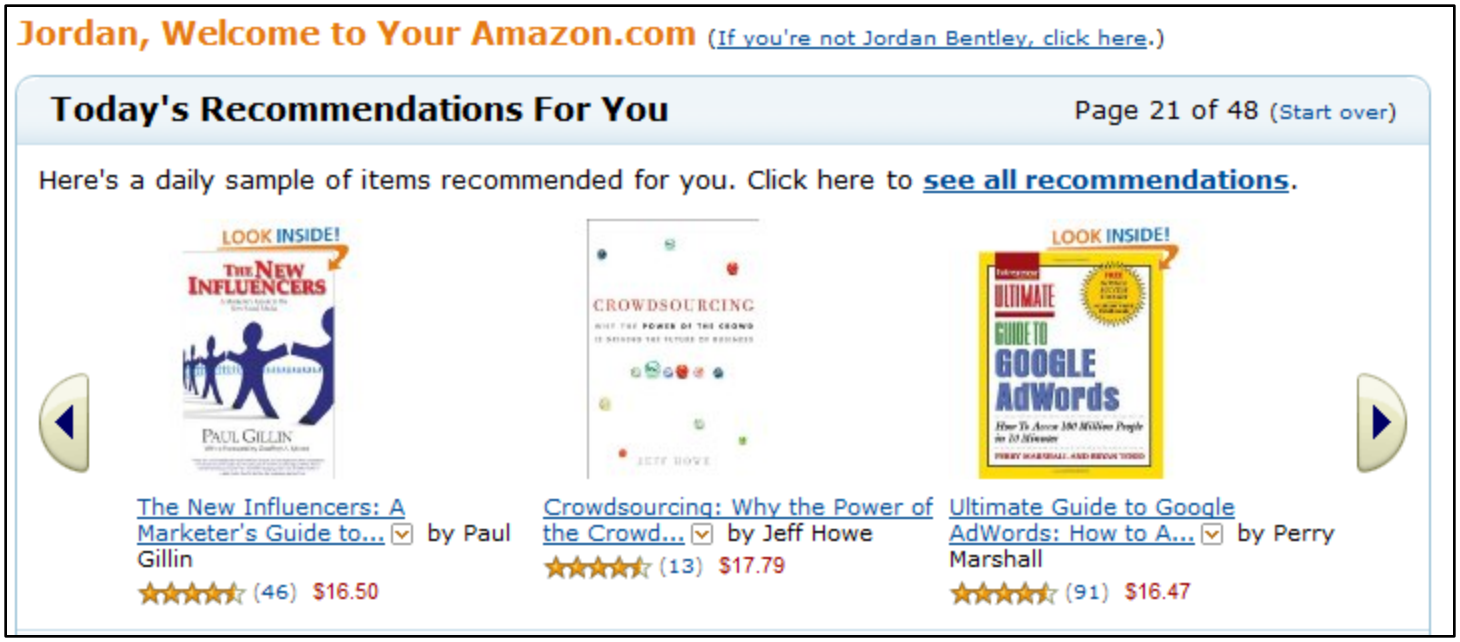
Collaborator: Dr. K. Claypool, MIT Lincoln Lab

Outline

- Introduction to Collaborative Filtering
- Tag based solutions
- Problems
  - Tag Unification
  - Recommendation Algorithm

Collaborative Filtering

- Predicts user interest in a given resource
  - i.e., Netflix, Amazon



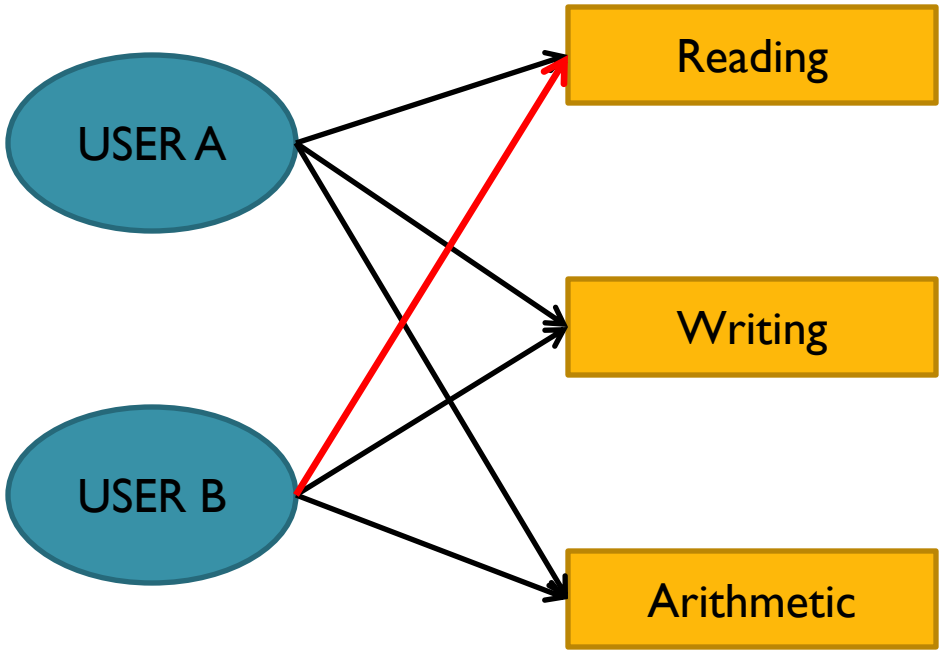
"Amazon.com: Recommended for You" Amazon.com. 7 April 2009. < http://www.amazon.com/gp/yourstore/ >.

Collaborative Filtering

- User-User model
  - The "Gold Standard" of collaborative filtering models[1]
  - Finds users with similar interests

[1] McNee, S. M., Kapoor, N., and Konstan, J. A. 2006. Don't look stupid: avoiding pitfalls when recommending research papers. In *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work* (Banff, Alberta, Canada, November 04 - 08, 2006). CSCW '06. ACM, New York, NY, 171-180. DOI= http://doi.acm.org/10.1145/1180875.1180903

User-User Model



Collaborative Filtering

- Ratings Based
  - This is the most common form
  - Takes into consideration how much a user likes an item but not why
  - Needs CONTEXT[1]

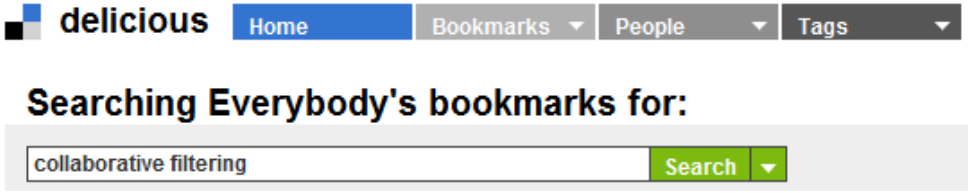
[1] Nakamoto, R.Y., Nakajima, S., Miyazaki, J., Uemura, S., Kato, H., and Inagaki, Y. 2008. Reasonable tag-based collaborative filtering for social tagging systems. In *Proceeding of the 2nd ACM Workshop on information Credibility on the Web* (Napa Valley, California, USA, October 30 - 30, 2008). WICOW '08. ACM, New York, NY, 11-18. DOI= http://doi.acm.org/10.1145/1458527.1458533

Solution – TAGS

Author: Luca Cremonini Source:
http://www.railsonwave.it/railsonwave/2007/12/web-2-0-map

Impact

- Tag based social bookmarking sites already exist without recommendation engines
  - Del.icio.us
  - Citeulike
  - Flickr
- Only tag searching

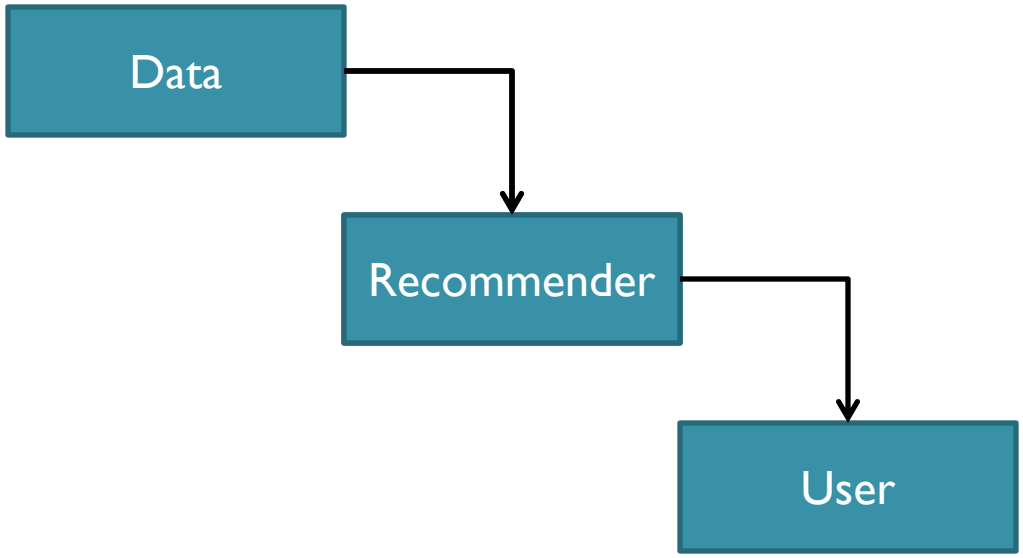


"Delicious" delicious.com. 7 April 2009. <http://delicious.com/>.

Objective

- Create a collaborative filtering algorithm that uses tags to take context in to account
  - Targeting academic papers for researchers

The Process

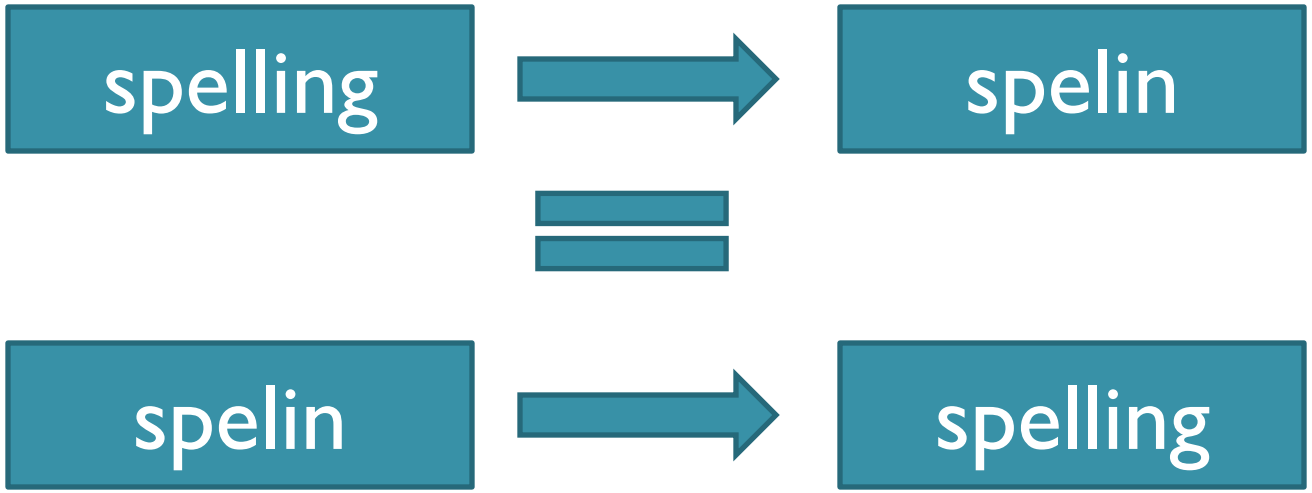


Tag Unification

- Problems in the data:
  - Misspellings
  - Form
    - Walk -> Walking, Walked, Walker, Walks, etc
- Easy solution:
  - Use a dictionary and an existing spell checking and stemming algorithm
- But what if the words do not appear in any dictionary?

Tag Unification

- Tag changes are bi-directional
- Focus is on consistency
  - Not correct spelling

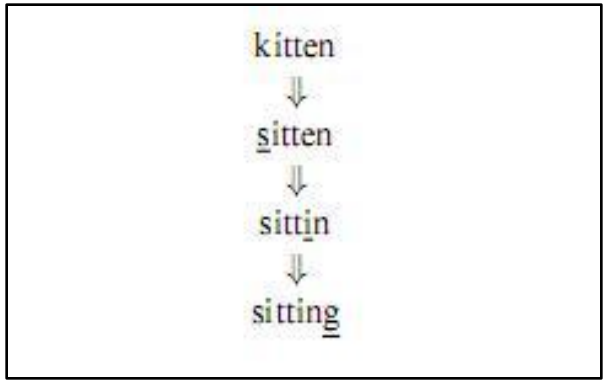


Tag Unification

- Proposed heuristic method
  - Starts with no predefined dictionary
  - Uses multiple measures to find good changes
    - Edit Distance
    - Co- Occurrence
    - Length

Edit Distance

- Traditional method of difference in spelling



- This example has an edit distance of 3

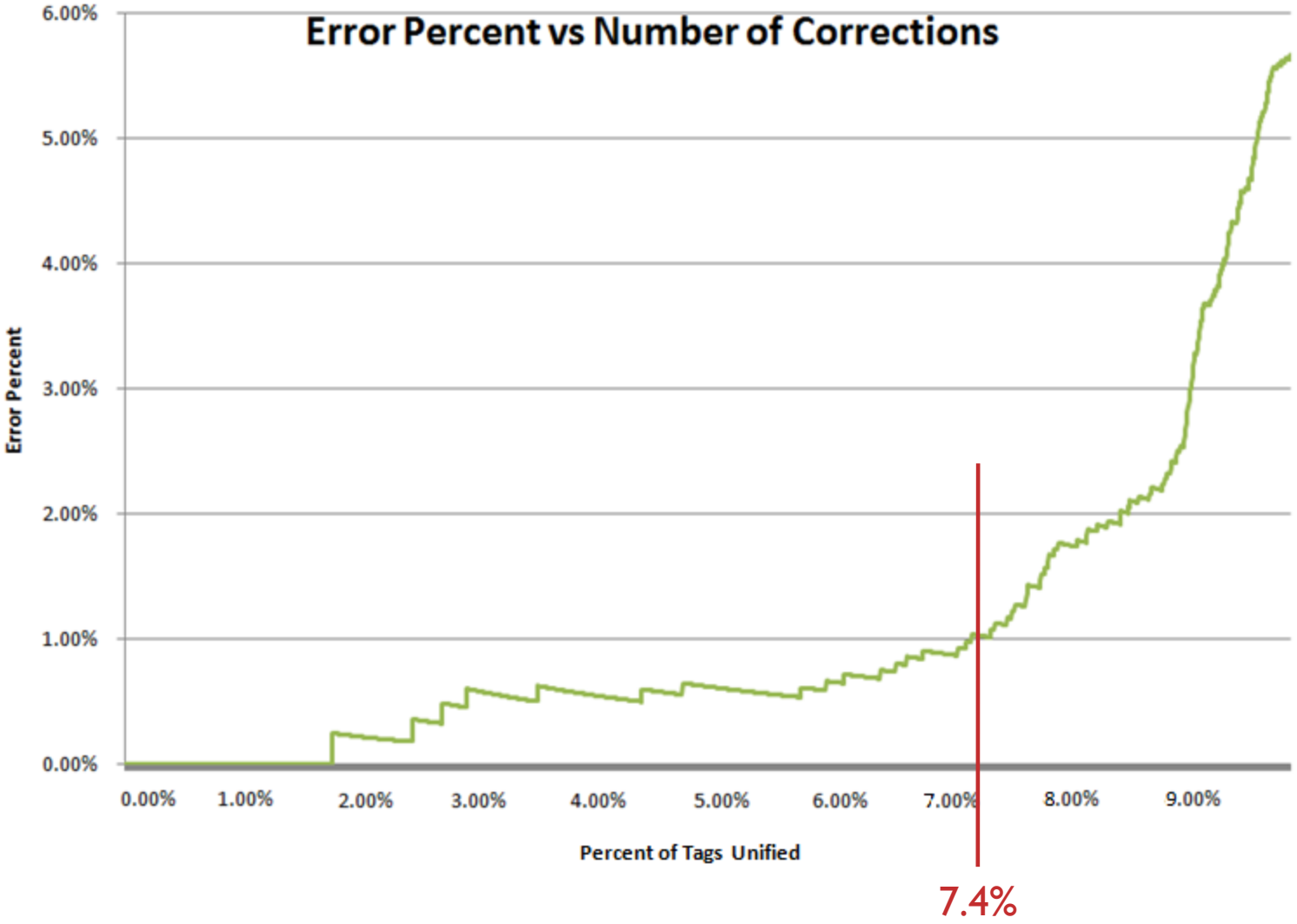
Tag Unification

$$\text{distance}_{t,p} = EditDistance_{t,p} * \frac{Resources \in (t \cap p)}{Resources \in (t \cup p)} * \sqrt{LEN(t) - 5}$$

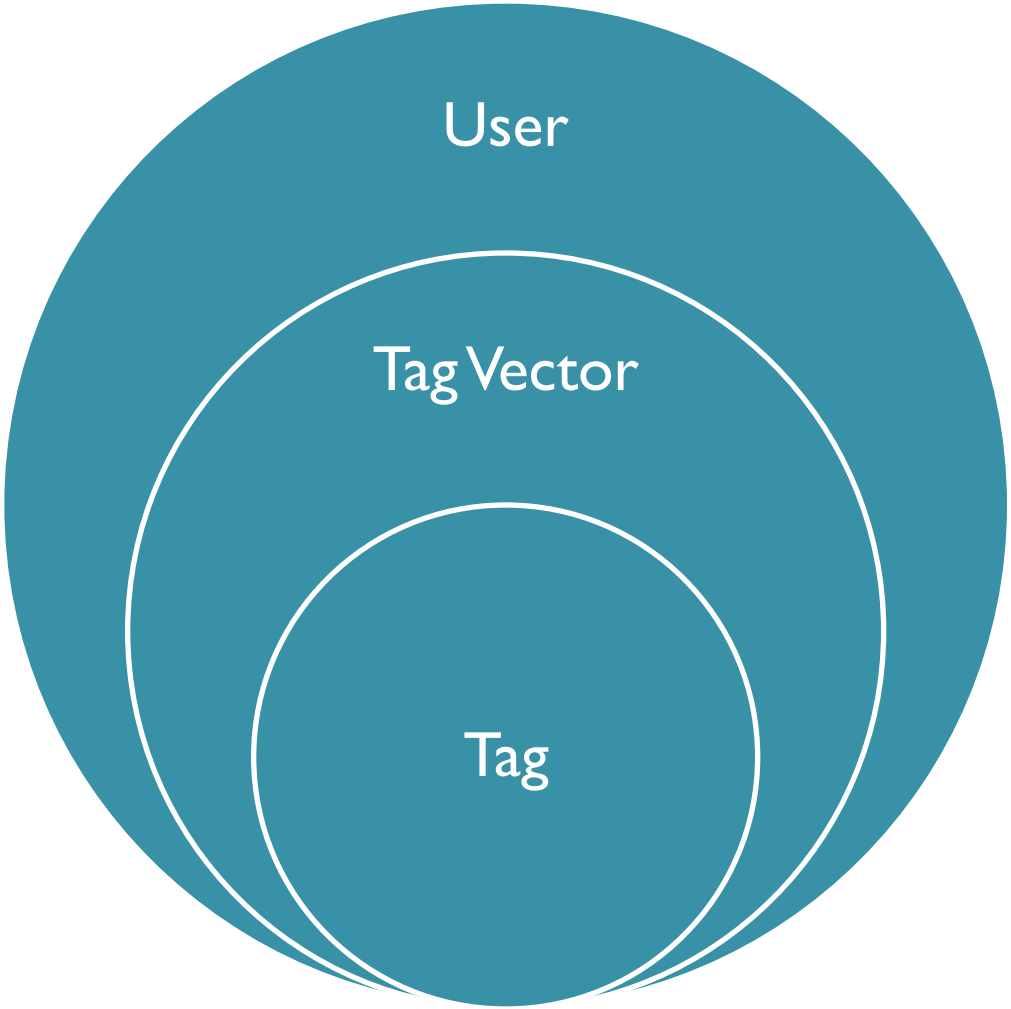
Tag Unification

- Evaluation had to be done by hand
  - A ‘good’ match was any change that preserved meaning
    - Even if the change was to an *incorrect* spelling, it was a good change

Tag Unification



Similarity Layers

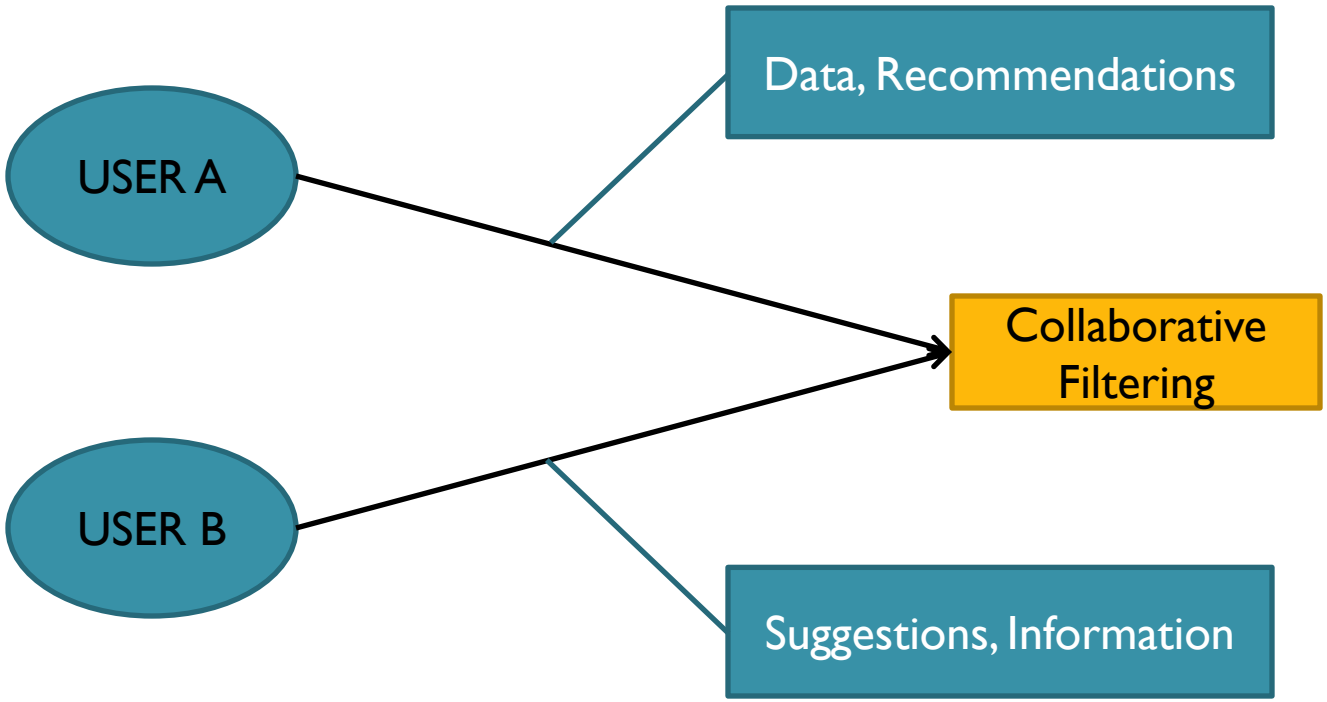


Tag Similarity

- Uses the co-occurrence from tag unification

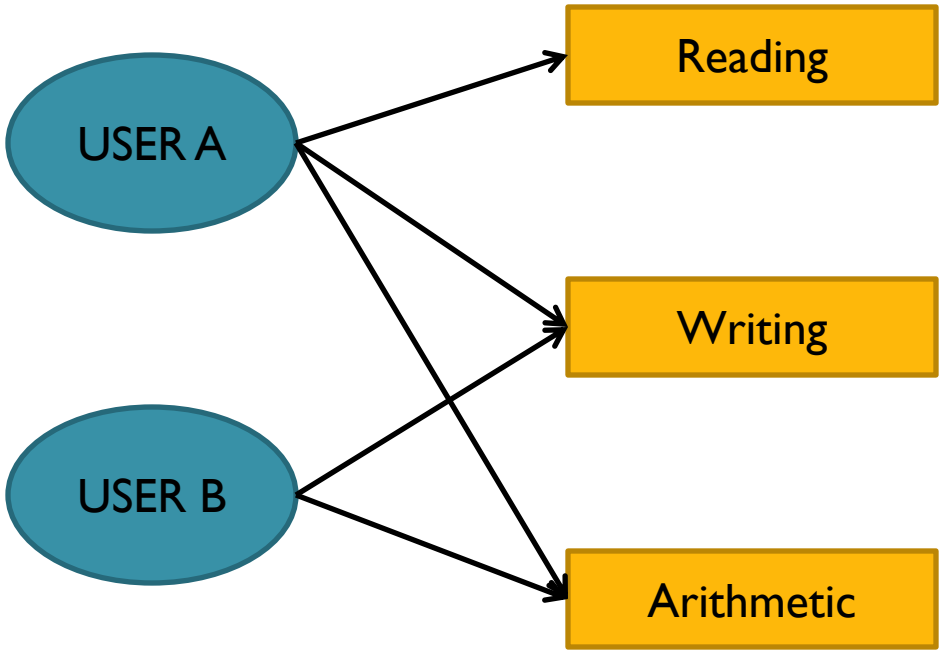
$$\mathrm{sim}_{a,b} = \frac{Resources \in (a \cap b)}{Resources \in (a \cup b)}$$

Tag-Vector Similarity



Data	x Suggestions	=	Low Match
Data	x Information	=	High Match
Recommendations	x Suggestions	=	High Match
Recommendations	x Information	=	Low Match

User Similarity



$$Sim(A_{writing}, B_{writing}) + Sim(A_{arithmetic}, B_{arithmetic})$$

Evaluation

1. Remove random resource from a random user
2. Generate recommendations for that user
3. Evaluate how often the user is recommended the removed resource

Evaluation

- Dataset
  - Obtained from citeulike [1]
  - 115,548 unique resources
  - 23,133 unique tags
  - 3,567 users

[1] "CiteULike: Everyone's library" CiteULike.com. 7 April 2009. < http://www.citeulike.com >.

Evaluation

- 24% of removed resources were in the top 50 recommendations
- 50% were in the top 270
- 95% were in the top 2000

Future work

- Compare Tag based methods to ratings based methods
  - Survey in progress
  - Will be able to run both methods on the same data
  - Data will be reusable for future studies

Thank you

Questions?