# Comparing Collaborative and Content-based Filtering for Recommendation on Social Bookmarking Websites

Toine Bogers and Antal van den Bosch

ILK / TiCC

Tilburg University

# Overview

- Recommendation task + data sets
- What information sources do we have?
  - Usage patterns
  - Tags
  - Metadata

- What is it?
- What did we do with it?
- What did we find?

- Recommendations for recommendation

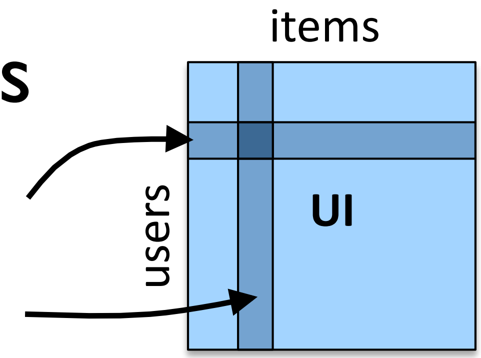
# Recommendation task & data sets

- Focused on Top-N item recommendation for social bookmarking websites
- Four data sets
  - del.icio.us (bookmarks)
  - BibSonomy (bookmarks)
  - citeulike (scientific articles)
  - BibSonomy (scientific articles)
- Evaluated using Mean Average Precision (MAP)

# Usage patterns

## What is it?

- Represent the items that users have added to their profiles
- Profile vectors
  - User profiles
  - Item profiles



- No explicit ratings available
  - Only binary information (1 or 0)
  - Or rather: unary!

# Usage patterns

## What did we do with it?

- Baseline: standard *k*-NN algorithm
  - – User-based CF vs. item-based CF
  - – Cosine similarity
  - – Unweighted vs. IDF-weighted profile vectors

# Usage patterns

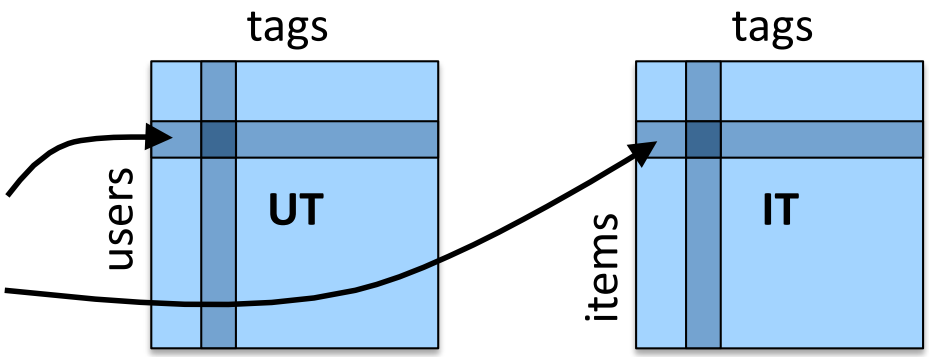
## What did we find?

- User-based vs. item-based
  - User-based CF slightly better on three data sets
  - Not statistically significant
  - Item-based CF significantly better on CiteULike
- Bookmarks vs. scientific articles
  - Recommending bookmarks is more difficult
  - More open domain and greater topical diversity
- IDF-weighting had no effect

# Tags

## What is it?

- Tags are keywords assigned to an item by a user
- Profile vectors
  - User tag profiles
  - Item tag profiles



- Values are tag occurrence counts

# Tags

## What did we do with it?

- Tag overlap between users/items as similarity
  - User-based vs. item-based filtering
  - Similarity metrics
    - Jaccard overlap
    - Dice's coefficient
    - Cosine similarity
  - Unweighted vs. IDF-weighted profiles (for cosine)

# Tags

## What did we find?

- CF with tag overlap
  - User-based CF performs significantly worse
  - Item-based CF performs much better
    - Often statistically significant improvements
  - Except on CiteULike: CF without tags better
- Similarity metric relatively unimportant
  - Cosine similarity slightly better
- IDF-weighting again had no effect

# Metadata

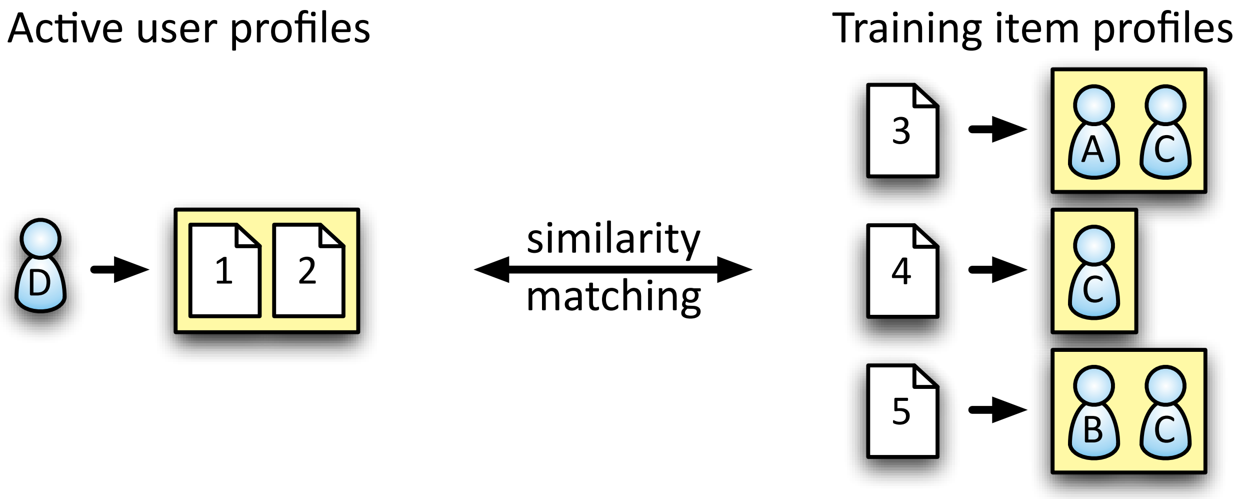
## What is it?

- Textual description of different aspects of an item
- Examples
  - Bookmarks: `<TITLE>`, `<URL>`, `<DESCRIPTION>`, …
  - Scientific articles: `<JOURNAL>`, `<YEAR>`, `<ABSTRACT>`, …
- Two types of metadata
  - Intrinsic, i.e., directly relating to the content
    - E.g., `<TITLE>`, `<DESCRIPTION>`, `<JOURNAL>`, `<AUTHOR>`, …
  - Extrinsic, i.e., administrative information
    - E.g., `<PAGES>`, `<MONTH>`, `<EDITION>`, …

# Metadata

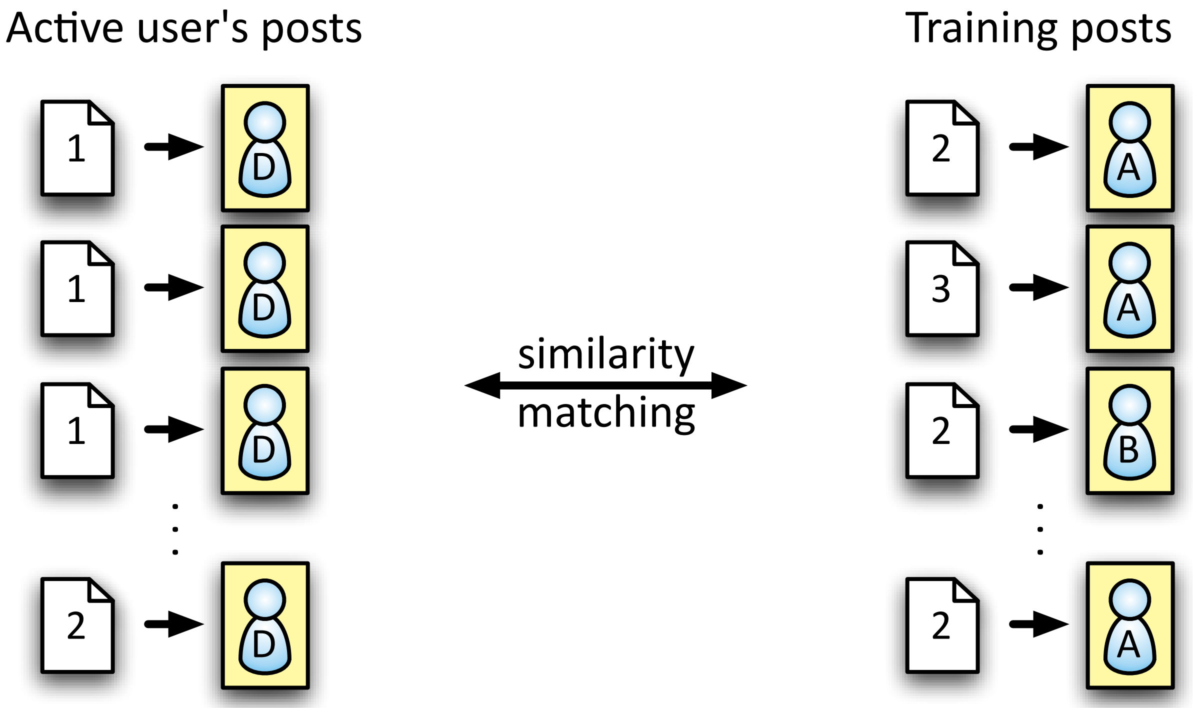
## What did we do with it?

- Content-based filtering
  - Profile-centric matching
    - Collate all of user’s metadata into a user profile
    - All metadata assigned to an item → item profile
    - Match and rank item profiles to user profiles



# Metadata

## What did we do with it?



– Post-centric matching
  - Construct metadata representations of each post
  - Match each of the user’s posts against all other posts
  - Match, rank, and aggregate all retrieved posts

# Metadata

## What did we do with it?

- Hybrid filtering
  - Combine CF with metadata-based approach
  - User-based CF with metadata-based similarities
    - Textual similarity between user profiles
  - Item-based CF with metadata-based similarities
    - Textual similarity between item profiles

# Metadata

## What did we find?

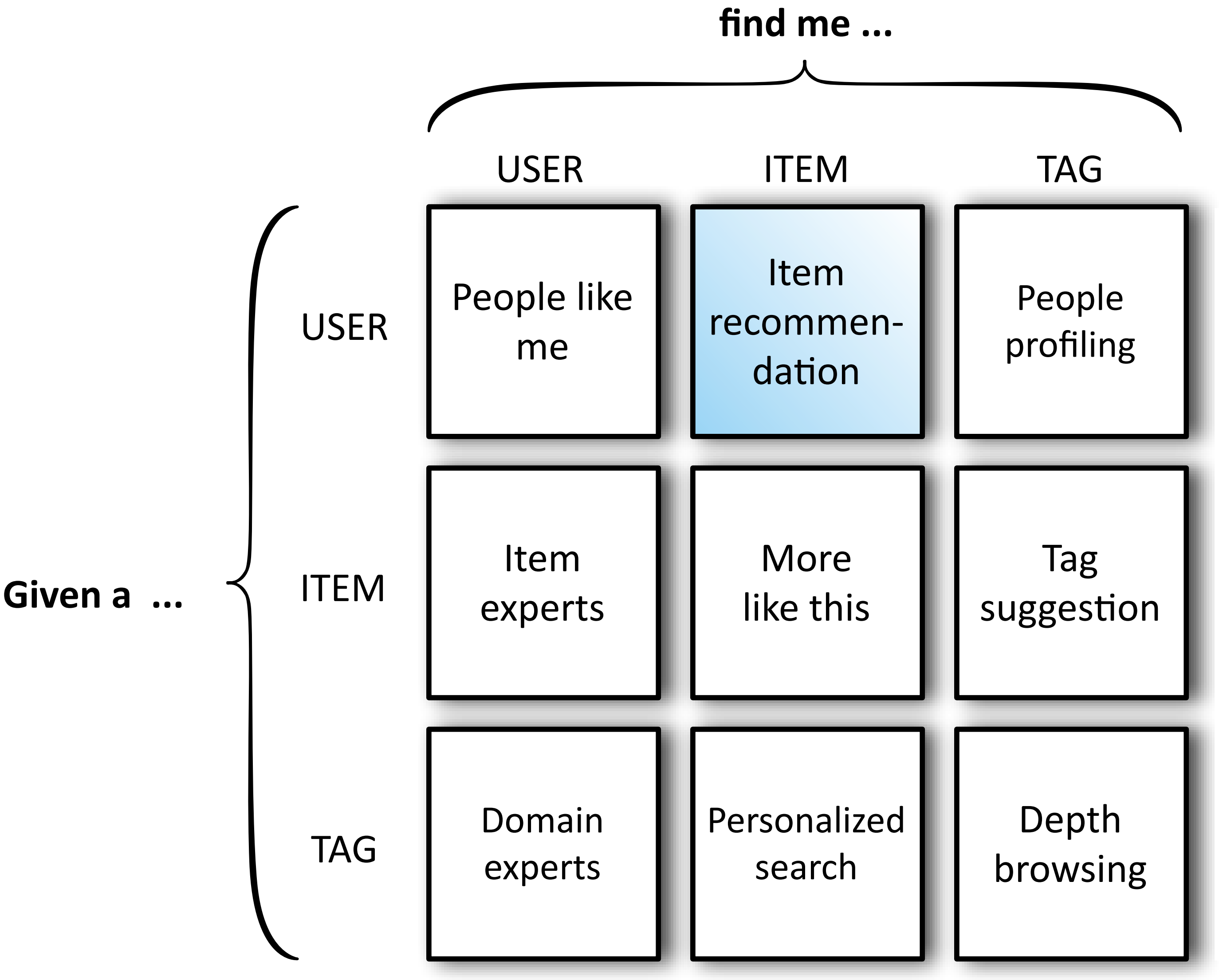
- Content-based filtering
  - Profile-level matching better than post-level
- Hybrid filtering
  - Item-based CF with metadata similarities works best
- No clear winner over all data sets
- Metadata
  - All intrinsic metadata combined works best
  - Best fields: `<TAGS>`, `<TITLE>`, `<AUTHOR>`, `<URL>`, `<ABSTRACT>`
  - Extrinsic metadata contributes little

# Recommendations for recommendation

- Using tag overlap in item-based CF works well
  - Easy to implement/adapt
- Metadata-based recommendation often better than CF
  - Not significantly
  - No clear winning algorithm
  - Easiest to implement using existing search engine
- Recommender fusion is promising
  - Investigate different combination techniques

Questions? Comments? Recommendations?

# Recommendation task



| Given a … / find me … | USER | ITEM | TAG |
|---|---|---|---|
| USER | People like me | Item recommendation | People profiling |
| ITEM | Item experts | More like this | Tag suggestion |
| TAG | Domain experts | Personalized search | Depth browsing |

# Data sets

| | Bookmarks | | Scientific articles | |
|---|---|---|---|---|
| | Delicious | BibSonomy | CiteULike | BibSonomy |
| # users | 1,243 | 192 | 1,322 | 167 |
| # items | 152,698 | 11,165 | 38,419 | 12,982 |
| # tags | 42,820 | 13,233 | 28,312 | 5,165 |
| # posts | 238,070 | 29,096 | 84,637 | 29,720 |

- Evaluated using Mean Average Precision (MAP)