Locality Sensitive Hashing for Real time Recommender Systems and Entity Resolution

Shouheng Li
u4713006@anu.edu.au
The Australian National University
Outline

- Introduction
- Motivation
- Methods/Techniques
- Things to do
- Conclusion
- Question
Introduction

- Find similar items in databases
  - —a mission for data miners

Applications
- Recommender Systems
  - Amazon, Netflix, Google personalized news
- Entity Resolution
  - Record Linkage
  - data de-duplication
Motivation

- Real time similarity search needed
  - Quick response on user behaviours
    - Social media
    - Online credit card inquiry
    - National security database

- How to make it fast
  - Find approximate results
    - Grouping/blocking
  - Rank items by similarity
Locality Sensitive Hashing (LSH)

- Main idea: hashing
- Min-hashing
  - Hashes by smallest signature
Methods/Techniques

- Similarity Aware Indexing (SAI)
  - Fast record-matching
  - Pre-calculate similarities
Things to do

- Combine the two techniques (LSH and SAI)
  - Entity resolution
    - Based on bibliography dataset
  - Social recommender system
    - Based on Twitter dataset
Conclusion

- Motivation: find similar items
- How: approximate results and similarity rank
- Techniques: LSH and SAI
- Things to do
  - Combine two techniques
  - Build a social recommender system
  - Implement an entity resolution algorithm
Question

- Any questions?