2014 Fall Summary

Haoyan Cai
12/04/2014
Research Projects

• Author Name Disambiguation with heterogeneous information network and textual similarities (Sept.)

• NewsTwitter: A demo to study news events and related tweets (Oct. & Nov.)
Author Name Disambiguation

• Ambiguity is a main obstacle to information matching and linking.

• SimVec = [TextualSim(r1,r2), GraphSim(r1,r2)]
  - TextualSim(r1,r2) = [title, year, affiliations, phrase], Jaccard similarity, difference in year
  - GraphSim(r1,r2): meta path like similarities, co-author, citation, co-pub venue and etc.

• Learning weights by SVM. (ongoing)

• Clustering: Hierarchical Clustering, DBSCAN and maybe more. (future)
TwitterNews

- Overview: We just finished data collection, interface and initial clustering of news.
- My Work: Data collection, db management, near-duplicate tweets removal.
- Future: Stream clustering, co-clustering and other interesting topics like mining the patterns of social media (twitter) reactions to news events (clusters of news), sentiment changes as news events evolve.
- Please expect more detailed descriptions in Jingjing's presentation.
Other Activities

- **Courses**
  - CS 447 Natural Language Processing
  - CS 598 Advanced Information Retrieval
  - CS 591 Advanced Seminar on Data Mining
  - CS 591 TA Seminar

- **TA**
  - CS 105 Intro to Computing
Thanks!