Research-Insight

Providing Insight on Research by Publication Network Analysis

George Brova, Fangbo Tao

May 7th, 2013

(Joint work with Tobias Lei, Xiao Cheng, Rucha, etc..)

Motivation

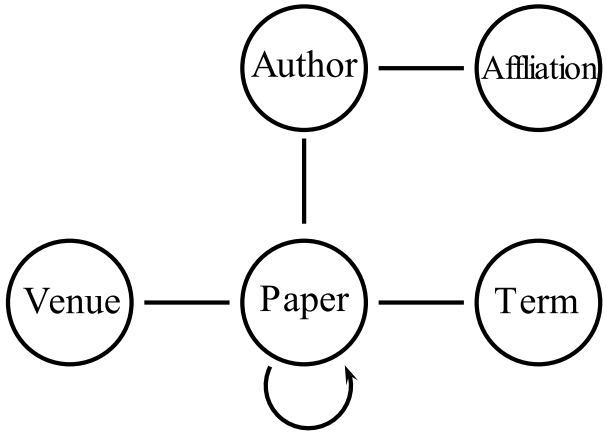
- When doing research, what's the confusing part that an "information system" may help?
  - What's **your** next research "big thing"?
  - Who is the guy **you** should collaborate with?
  - Which papers **you** need to read? Latest one? Related ones?
- Global insight? Personalized insight!
  - Previous works, common affiliation, social connections, paper already read, etc…

What should be correct?

- Lexical Similarity
- Citation #
- Freshness
- Related to your work?
- Social Similarity
- Authoritative Author/ Conferences

Data Source: CSR-Net

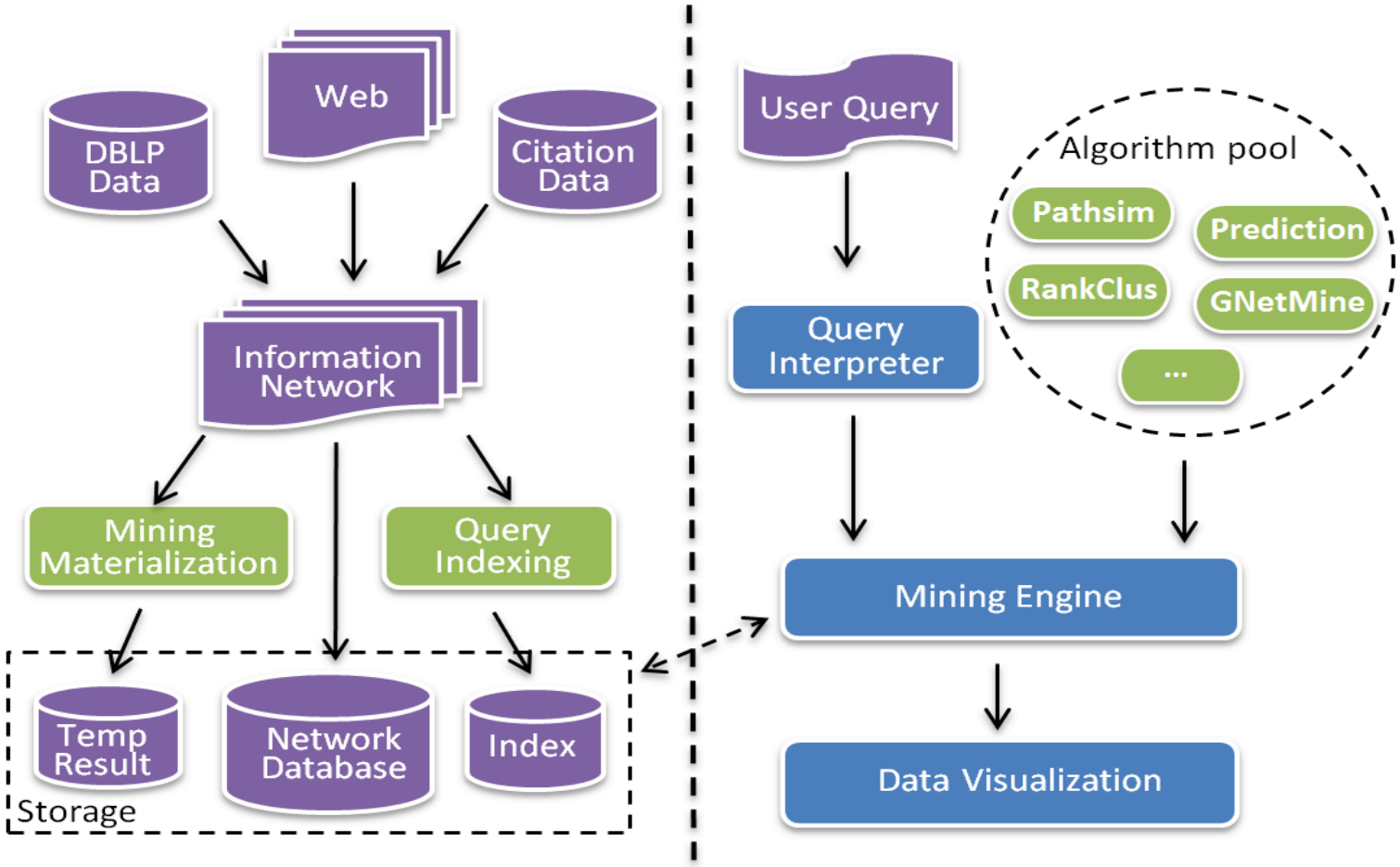
- DBLP Dataset
  - An information-rich CS publication network
- Crawling Web Information
  - Partial Hierarchical Affiliation Info: University-Department-Research Group
- Citation Data
  - ArnetMiner & Citeseer



Functions We Support

- Intelligent Literature Search
- Collaboration Prediction
- Similarity Search
- Ranking-based Clustering
- Academic Profile Generation
  - Historical Affiliations Prediction
  - Advisor/Advisee Finding

System Architecture



Literature Recommendation

- Traditional keyword-based search system (G-Scholar)
  - Measure the document similarity between query and paper
- Combine Network Structural Similarity & Document Similarity
  - Document Similarity
  - Authority of the paper
  - Closeness of the personalized information

Document Similarity: Lucene

- Easy and fast to implement
- Close to state-of-the-art
- Can already search multiple fields
  - Title has more weights than abstract

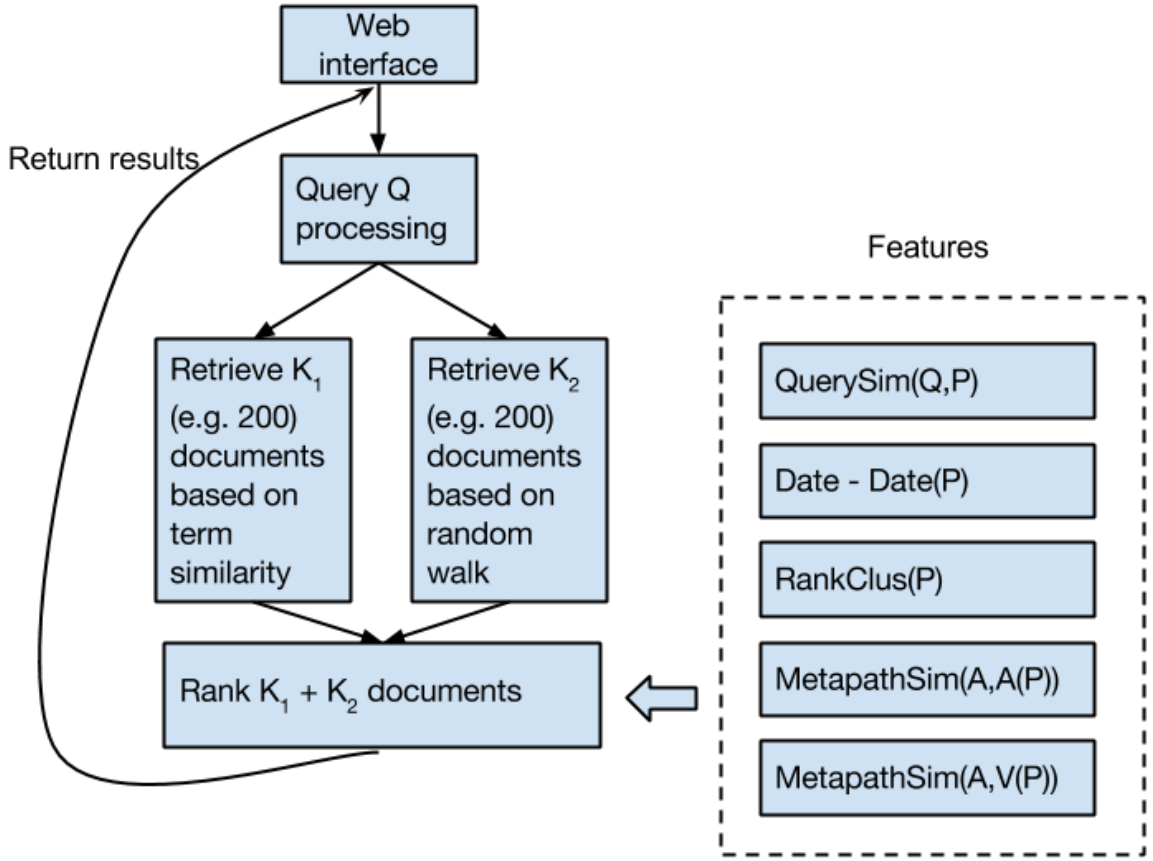
Authority of the paper: NetClus!

- Citation data, yes!
- Not enough
  - Paper may be too fresh
  - Authoritative papers != paper with high citation
- Using NetClus to leverage structural information
  - Authority of Author
  - Authority of Conference
- The ranking is cluster-based, not global.
  - Example: "Yizhou Sun" should rank higher for topic "Heterogeneous information network"
  - We locate the cluster first, using local ranking as features

Personalized results: PathSim

- Meta-path:
  - APA: Author-Paper-Author
  - APVPA: Author-Paper-Paper-Venue-Author
  - AFA: Author-Affiliation-Author
- PathSim: Similarity of nodes along meta-paths
- Different metapaths indicate different "types" of similarity. For example,
  - APA = frequent co-authorship,
  - APVPA = similar publication distribution
  - AFA = social similarity

A hybrid model



Evaluation of paper search

- Each component function has been evaluated [Usually has a corresponding paper]
- How to evaluate? No explicit ground truth!
- We're thinking about it a lot, two directions:
  - User study: training weights of all the features to fit user's ranking criteria best.
  - Feedback from online query: Implicit & Explicit.

Collaboration Prediction

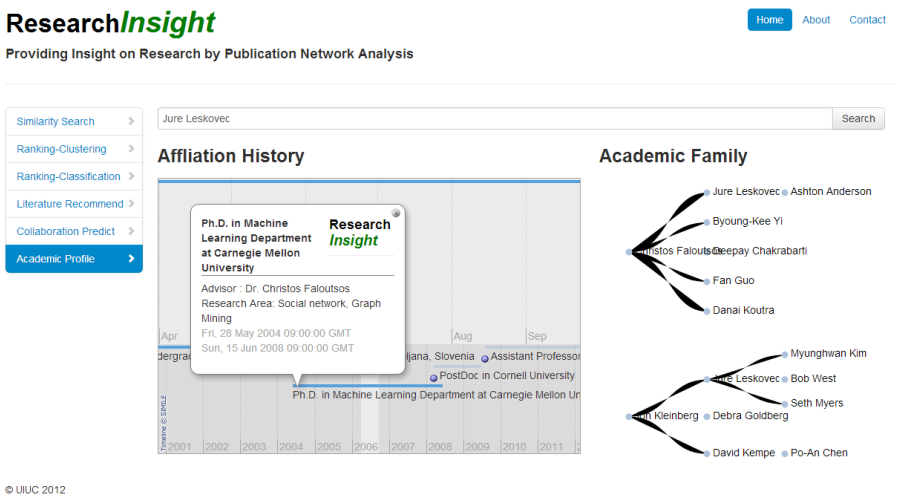
- For each researcher with his publication history and affiliation, one may get
  - Advisor and group mates → Social Similarity
  - Other professor/student in the same institution in a related discipline → Social Similarity
  - Researchers in same field but different affiliations → Topic Similarity
  - Authorities for your research topics → Authority
- A meta-path based framework
  - APA: Author-Paper-Author
  - APVPA: Author-Paper-Paper-Venue-Author
  - APTPA: Author-Paper-Term-Venue-Author
  - AFA: Author-Affiliation-Author
  - APPPA: Cite similar paper

Advanced Collaboration Prediction

- Specify a topic!
  - "Jiawei Han" + "Support Vector Machine" = potential collaborators
    - Authority of "Support Vector Machine"
    - Expert of "SVM" from same university/Group
    - Previous collaborators who have related knowledge of SVM.
  - For query-based collaborator prediction
    - Count "lexical similarity" as our feature
    - Adjust the weights of meta-paths.

Advisor/Advisee, Affiliation History

- Advisor/Advisee finder. [Chi, et.]
- Partial Affiliation Data from ACM-DL
- Advisor/Advisee←→Affiliation History



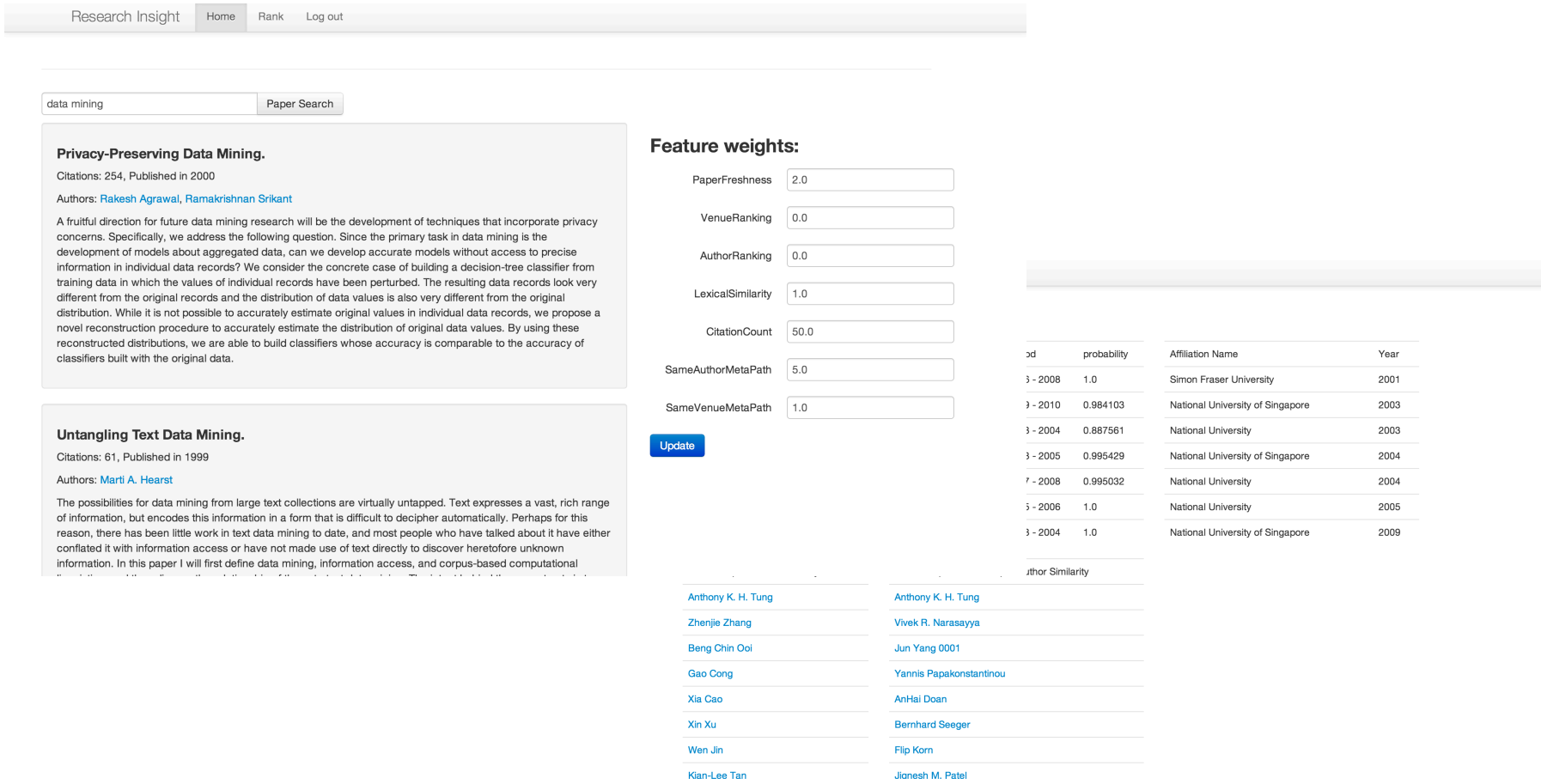
Other works

- Similarity Search
- Ranking based Clustering
- In the future:
  - Make it a testbed for heterogeneous information network.
  - Digest rich-text better, building cubes based on text.

Challenges

- Extendible design
- Large dataset
  - millions of papers, authors, citation
  - Efficiency/Space tradeoff
  - materialize paths on-line, with some minimal precomputation
- Balance research and engineering
  - Real system needs efficient solution
  - Some research is too complicated
  - Researching and engineering together.

Screen shots



Thanks!

- Question