Name Entity Cube in History Domain
Software design

Wenbin Li

The Saarbrücken Graduate School of Computer Science

Please note: some details will be changed within the actual implementation.

content

- Data Preparation
- Preprocessing
- Relationship Analysis
- Visualization
- Software Engineering
- Schedule
- Motivation
- Demo

Data preparation

- About Civil War
- Sources:
  - University of Virginia: HIUS 403: "Digital History and the American Civil War." (Newspaper)
  - Operation:
    - Retrieve text-format newspaper from the website
  - Tools:
    - Websphinx http://www.cs.cmu.edu/~rcm/websphinx/

preprocessing

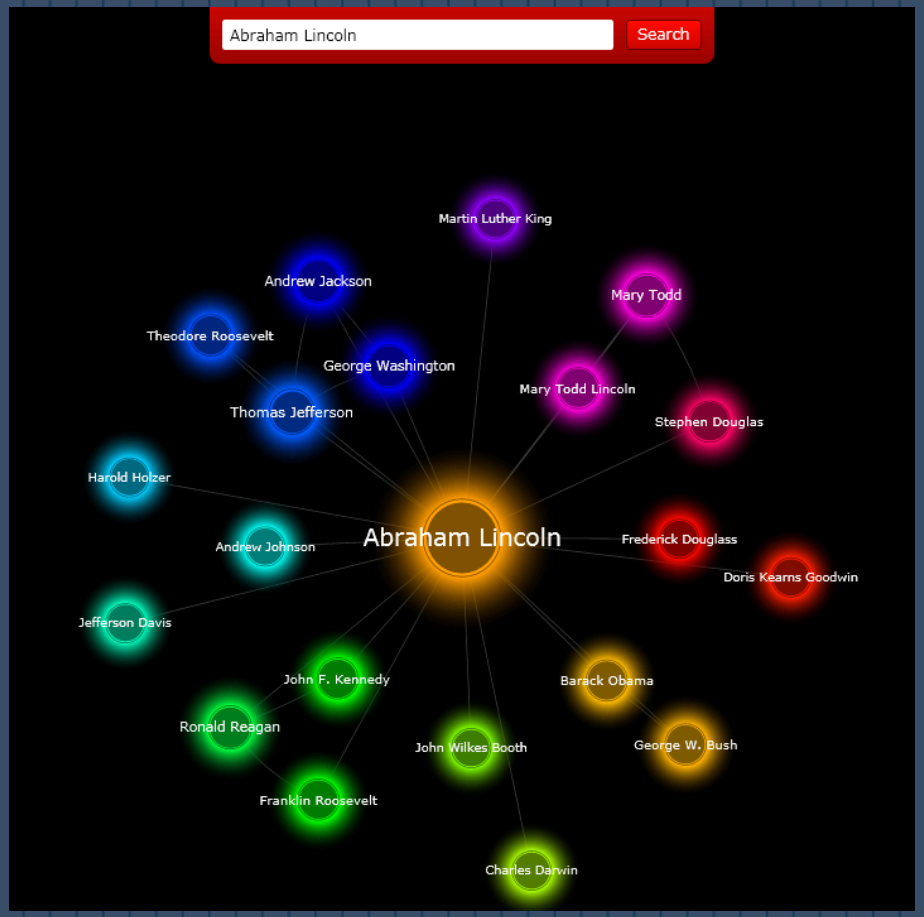
- POS tagging
- NER tagging
- Tools: Stanford NLP Tools http://nlp.stanford.edu/software/index.shtml

Relationship analysis

- Record:
  - { Pairs (Name_A, Name_B), weight }
- Data Structure:
  - Undirected Graph
- Tools:
  - **JUNG** Java Universal Network/Graph Framework

visualization

- Relationship Web Search



Software engineering

- GUI Design
- Extensibility
  - Customize Database for specific research topics, e.g. Civil War → World War II
- Modularity
  - Data Retrieval → Preprocessing → Relationship Analysis → Visualization
- Usability
  - Balance between Precision and Recall
  - (ensure low-occurrence name appear in the map might conflict with specific threshold in the RA algorithm – to be settled)

schedule

- Prototype (1 week)
- Report (2-3 days)

Motivation

- General idea
  - Object-level Search
- Specific usefulness
  - Comparative Study in History Character

Object-level search

- Traditional search
  - For Internet: Page-Level Search
  - More Generally: input entry → return related entries
- Object-Level Search (Vertical Search)
  - More precisely meet users' information need
  - More stereo about the query

Object-level search

- Example:
- Comparison:

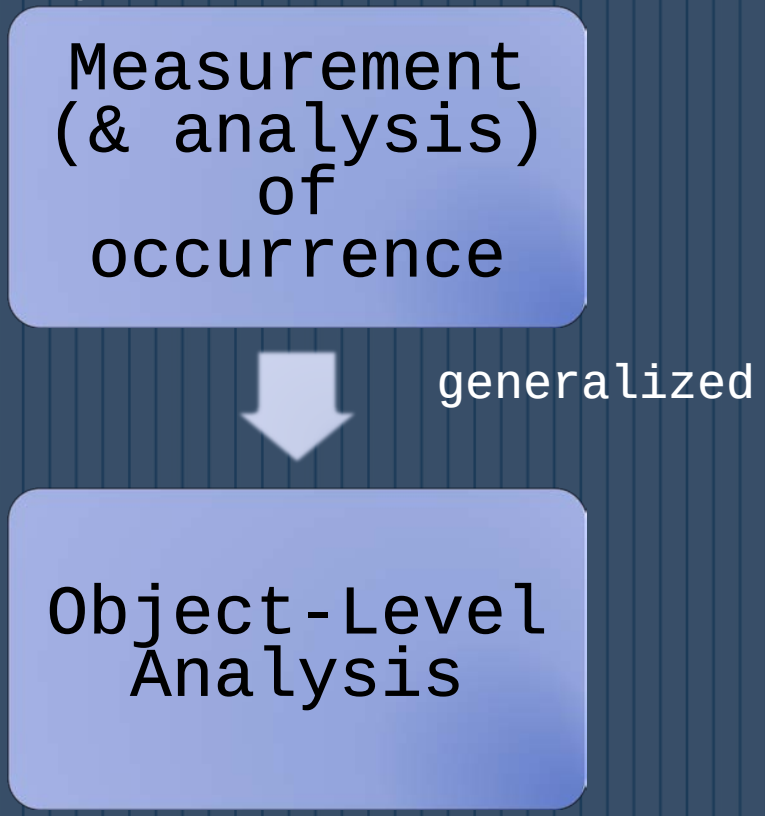
	Traditional Search	Object-Level Search
Tech.	IR	DB ML
Pros.	Ease of author Ease of use	Powerful Query Capability; Aggregate Answer
Cons.	Limited Query Capability	Where & How to get the Objects?

Relative Search

- Display Related history characters (clustering analysis)
- Comparative analysis (need further interface for access to every entry)
- More delicate work can be done by semantic analysis of different name entities (edge representation)

Relationship btw OLS & RS

- RS can be seen as a simple case for OLS.
- Technically speaking



Demo

Discussion

- Further analysis of data
- Possible measurement of relationship
- Correspondent GUI

End

- Thanks!