Scholarly Article Recommendation for Newbies

Samaneh Ebrahimi, Francis Gallego, Ari Kapusta, Kangqi Ni, and Payam Siyari

Data and Visual Analytics - CSE 6242
Spring 2016

INTRODUCTION

Motivation

- **Researchers spend much time searching for papers**
  - Notably when exploring a new area or doing interdisciplinary research
- **Literature Survey typically performed manually**
  - There is no commercially available and user-friendly software for such task
- **Solution** would ideally be a sub-system in Google Scholar, Microsoft Academic Research, Scopus, etc
  - Could incorporate network centrality as importance metric

Related Work

- **Graph-Sensemaking:** Representing information from graphs for user interpretation
- **Article Recommender Systems:** Mostly based on contents, not for newbies
- **Ranking Systems:** Mostly to do with search and retrieval systems
- **Centrality Algorithms:** A measure of nodes' importance

Dataset

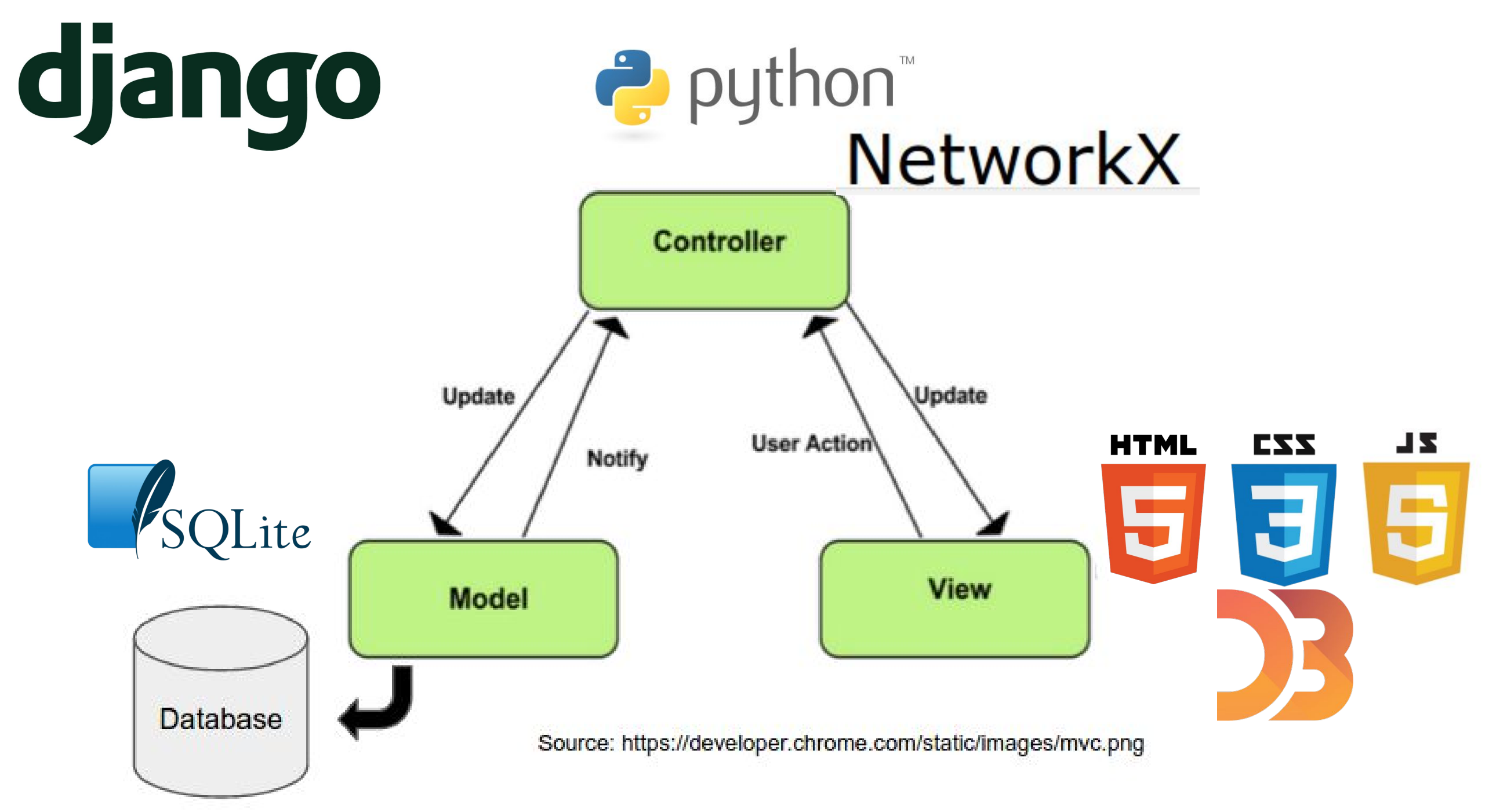
Our primary dataset is a SQL database containing a citation network graph scraped from **Google Scholar**. It contains **83,000 articles (nodes)** and **150,000 citation** relationships across a wide range of subject areas, making up approximately **187.5 MB [1]**. This is a subset of all Google Scholar data.

[1] Duen Horng Chau, et. al., "Apolo: making sense of large network data by combining rich user interaction and machine learning". CHI, 2011.

METHODOLOGY

Approach

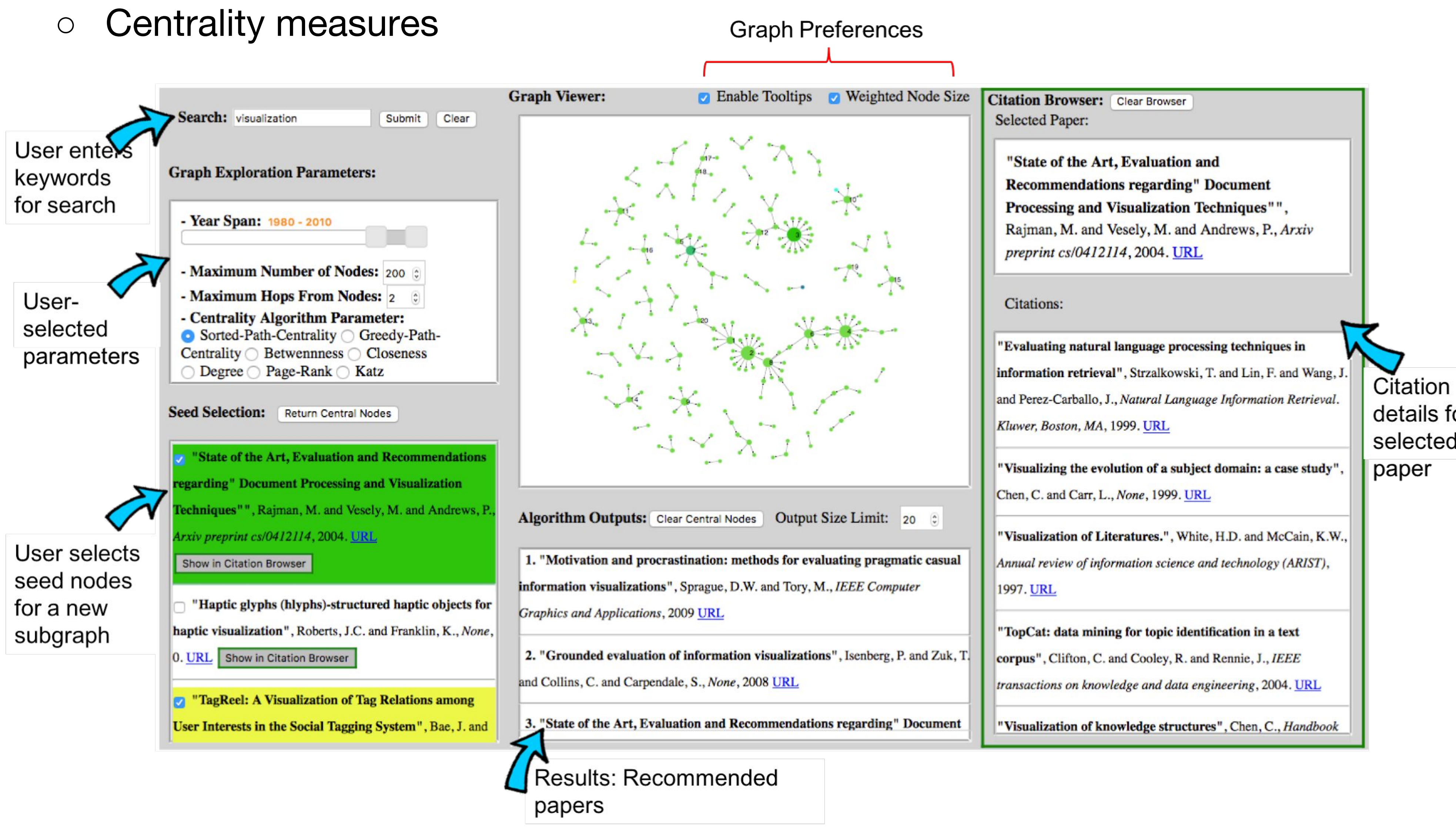
- A *Model-View-Controller* architecture
- **Django** code framework
  - Allows for fast and parallel development
- **SQLite** database in Model
- **Python** (with NetworkX) controller
- HTML, CSS, JS and **D3** View



Source: https://developer.chrome.com/static/images/mvc.png

Our App

- Web-based
- Search assistance with user-set parameters
- Active graph visualization (*better viewed in the demo*)
  - Filtering relevant nodes for further user's focus
  - Brief yet effective use of colors and sizings for better user engagement
- Recommendation system
  - Relevant citation network generation by keyword matching
  - Centrality measures



Graph Preferences

User enters keywords for search

User-selected parameters

User selects seed nodes for a new subgraph

Citation details for a selected paper

Results: Recommended papers

Search: visualization Submit Clear

Graph Exploration Parameters:

- **Year Span:** 1980 - 2010
- **Maximum Number of Nodes:** 200
- **Maximum Hops From Nodes:** 2
- **Centrality Algorithm Parameter:**
● Sorted-Path-Centrality ○ Greedy-Path-Centrality ○ Betwennness ○ Closeness ○ Degree ○ Page-Rank ○ Katz

Seed Selection: Return Central Nodes

☑ **"State of the Art, Evaluation and Recommendations regarding" Document Processing and Visualization Techniques""**, Rajman, M. and Vesely, M. and Andrews, P., *Arxiv preprint cs/0412114*, 2004. URL
Show in Citation Browser

□ **"Haptic glyphs (hlyphs)-structured haptic objects for haptic visualization"**, Roberts, J.C. and Franklin, K., *None*, 0. URL Show in Citation Browser

☑ **"TagReel: A Visualization of Tag Relations among User Interests in the Social Tagging System"**, Bae, J. and

Graph Viewer: ☑ Enable Tooltips ☑ Weighted Node Size

Algorithm Outputs: Clear Central Nodes Output Size Limit: 20

1. **"Motivation and procrastination: methods for evaluating pragmatic casual information visualizations"**, Sprague, D.W. and Tory, M., *IEEE Computer Graphics and Applications*, 2009 URL

2. **"Grounded evaluation of information visualizations"**, Isenberg, P. and Zuk, T. and Collins, C. and Carpendale, S., *None*, 2008 URL

3. **"State of the Art, Evaluation and Recommendations regarding" Document**

Citation Browser: Clear Browser
Selected Paper:

"State of the Art, Evaluation and Recommendations regarding" Document Processing and Visualization Techniques"", Rajman, M. and Vesely, M. and Andrews, P., *Arxiv preprint cs/0412114*, 2004. URL

Citations:

"Evaluating natural language processing techniques in information retrieval", Strzalkowski, T. and Lin, F. and Wang, J. and Perez-Carballo, J., *Natural Language Information Retrieval. Kluwer, Boston, MA*, 1999. URL

"Visualizing the evolution of a subject domain: a case study", Chen, C. and Carr, L., *None*, 1999. URL

"Visualization of Literatures.", White, H.D. and McCain, K.W., *Annual review of information science and technology (ARIST)*, 1997. URL

"TopCat: data mining for topic identification in a text corpus", Clifton, C. and Cooley, R. and Rennie, J., *IEEE transactions on knowledge and data engineering*, 2004. URL

"Visualization of knowledge structures", Chen, C., *Handbook*

EVALUATIONS

User Study

- Find the 5 papers with highest degree within 1 edge of a given seed paper
  - 5 subjects (the authors) in informal study
  - Using our App vs Google Scholar and web-searches

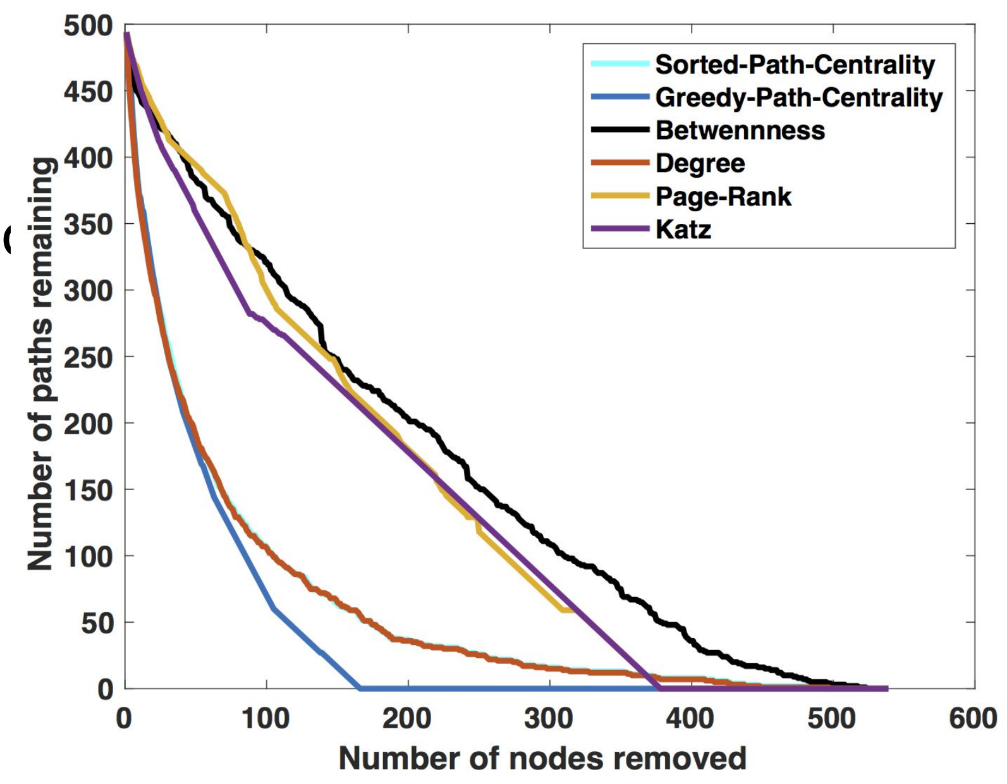
	Our App	Google Scholar
Mean (std) (seconds)	8.8 (1.9)	398 (24)

  - Greatly decreases time to perform search
    - Mean decrease of more than 6 minutes from manual search

Experimental Results

- Maximize the **coverage** of the citation pathways
  - A measure for *concise and parsimonious* results
  - **Path-Centrality:** Number of paths co sources to sinks
  - A *Supermodular Minimization Problem*
    - Greedy algorithm provides nice appx. bounds

$$C = argmin_s\ \mathcal{P}_G^-(s)$$
$$s.t.\ \ |s| = k$$



Summary

- **Much faster** to find important papers in a relevant network.
- **Active visualization** allows interaction with search results and the graph, and allows visualization of graph-relevant features.
- **Easy to visualize** relevant papers in one page, *no need to tediously cross reference* articles across multiple pages .