Network of Epidemiology Digital Objects

Kui Xu, Naren Sundar

Client: Sandeep Gupta, S. M. Shamimul Hasan
Problem Statement

• To build a connected network of
  – Papers
  – Wiki pages
  – Websites
  – Videos
  – Other digital objects

• Represent using RDF for future graph analysis purposes
Starting with a simple example

- DBLP page of Madhav Marathe
- Crawl
  - Papers
  - Authors, co-authors
  - Publishers
  - Date published
- Build a crawling framework to produce a robust and consistent RDF network
Crawling
A snippet of produced RDF

What to crawl next?

- A single crawl does not suffice
- How do we decide what to crawl next?
- Use RDF queries to compute some statistics
- Derive the next seed
Author co-occurrence

• Current seed author r
• For each author s
  – P ← query RDF for papers by r
  – Q ← query RDF for papers by s
  – (P \land Q) / (P \lor Q)
• Call this A(s \mid r)
Publisher co-occurrence

• Current seed author r
• For each author s
  – P ← query RDF for publishers for r
  – Q ← query RDF for publishers for s
  – (P ∧ Q) / (P ∨ Q)
• Call this P(s | r)
Final ranking and seed selection

• Rank each author $s$ by
  
  $w_1 \times A(s \mid r) + w_2 \times P(s \mid r)$
  
  $w_1 + w_2 = 1$

• Higher weight $w_1$
  
  Prefer author co-occurrence

• Higher weight $w_2$
  
  Prefer common publishers for authors but not necessarily as co-authors
Example probabilities

| Author                        | A(s | r)                  |
|-------------------------------|------------------------|
| Madhav V. Marathe             | 1.0                    |
| S. S. Ravi                    | 0.33333333333333333    |
| Christopher L. Barrett        | 0.23529411764705882    |
| V. S. Anil Kumar              | 0.20212765957446807    |
| Harry B. Hunt III             | 0.18478260869565216    |
| Daniel J. Rosenkrantz         | 0.15217391304347827    |
| Keith R. Bisset               | 0.14673913043478262    |
| Richard Edwin Stearns         | 0.11956521739130435    |
| Aravind Srinivasan            | 0.10869565217391304    |
| Sven Oliver Krumke            | 0.09239130434782608    |
Repeat

Seed

Select seed

Crawl

Rank seeds

Update RDF
Future Work

• Make this feedback loop official as a reinforcement learning problem
• Other measures from RDF queries
• Increase crawled sources
• Other digital objects
• Move up from this example to network of epidemiology digital objects