Network of Epidemiology Digital Objects

Kui Xu, Naren Sundar

Client: Sandeep Gupta, S. M. Shamimul Hasan

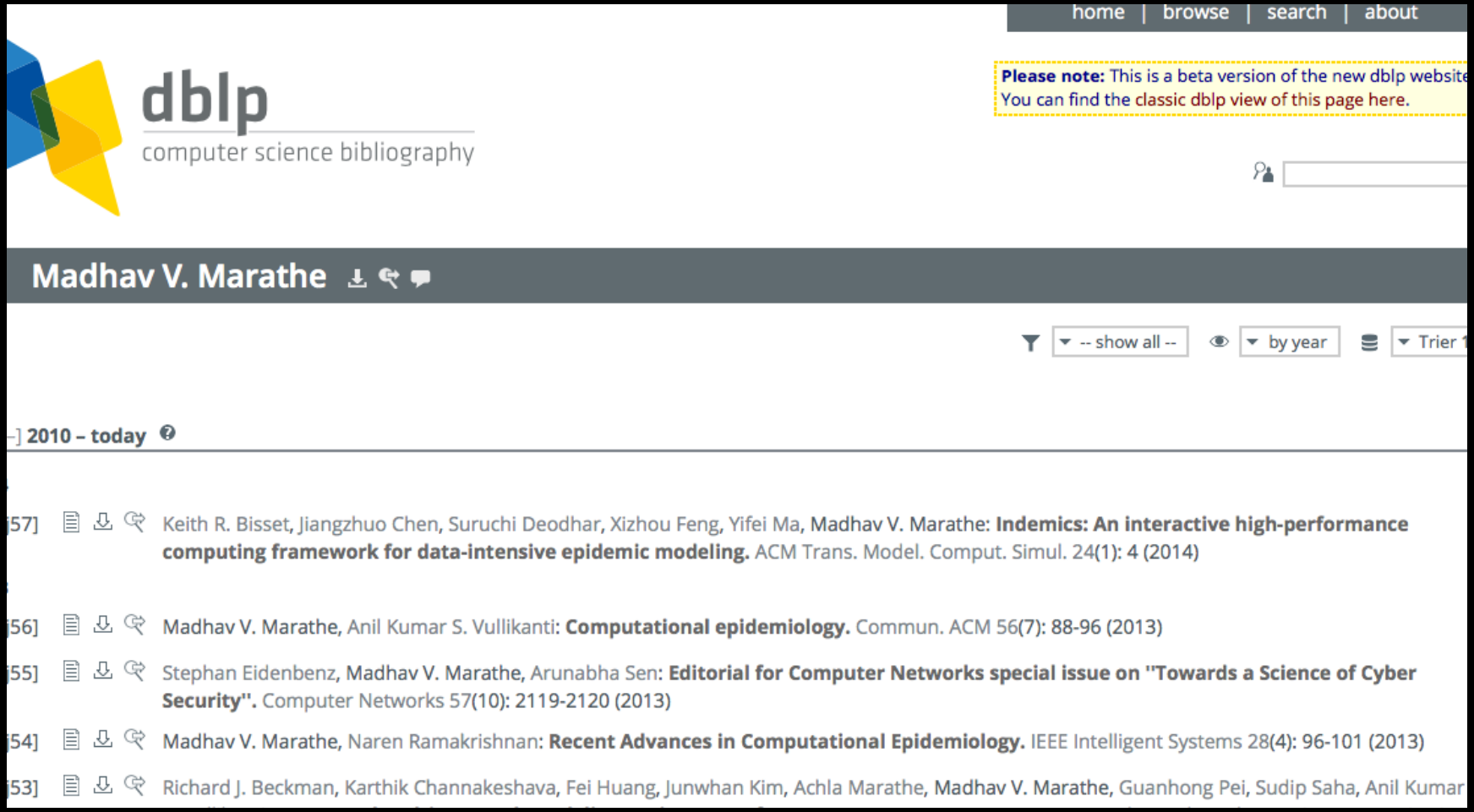
Problem Statement

- 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


home | browse | search | about
dblp
computer science bibliography
Please note: This is a beta version of the new dblp website
You can find the classic dblp view of this page here.
Madhav V. Marathe
-- show all --
by year
Trier
2010 – today
[57] Keith R. Bisset, Jiangzhuo Chen, Suruchi Deodhar, Xizhou Feng, Yifei Ma, Madhav V. Marathe: Indemics: An interactive high-performance computing framework for data-intensive epidemic modeling. ACM Trans. Model. Comput. Simul. 24(1): 4 (2014)
[56] Madhav V. Marathe, Anil Kumar S. Vullikanti: Computational epidemiology. Commun. ACM 56(7): 88-96 (2013)
[55] Stephan Eidenbenz, Madhav V. Marathe, Arunabha Sen: Editorial for Computer Networks special issue on "Towards a Science of Cyber Security". Computer Networks 57(10): 2119-2120 (2013)
[54] Madhav V. Marathe, Naren Ramakrishnan: Recent Advances in Computational Epidemiology. IEEE Intelligent Systems 28(4): 96-101 (2013)
[53] Richard J. Beckman, Karthik Channakeshava, Fei Huang, Junwhan Kim, Achla Marathe, Madhav V. Marathe, Guanhong Pei, Sudip Saha, Anil Kumar

A snippet of produced RDF

<http://doi.acm.org/10.1145/1869459.1869495> <http://papers.org/title> "Parallel inclusion-based points-to analysis." .

<http://doi.acm.org/10.1145/1869459.1869495> <http://papers.org/pubDate> "2010-10-27" .

<http://doi.acm.org/10.1145/1869459.1869495> <http://papers.org/publisher> "OOPSLA" .

<http://doi.acm.org/10.1145/1869459.1869495> <http://papers.org/author1> "Mario Méndez-Lojo" .

<http://doi.acm.org/10.1145/1869459.1869495> <http://papers.org/author2> "Augustine Mathew" .

<http://doi.acm.org/10.1145/1869459.1869495> <http://papers.org/author3> "Keshav Pingali" .

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 ∧ Q) / (P ∨ Q)
- Call this A(s | 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)

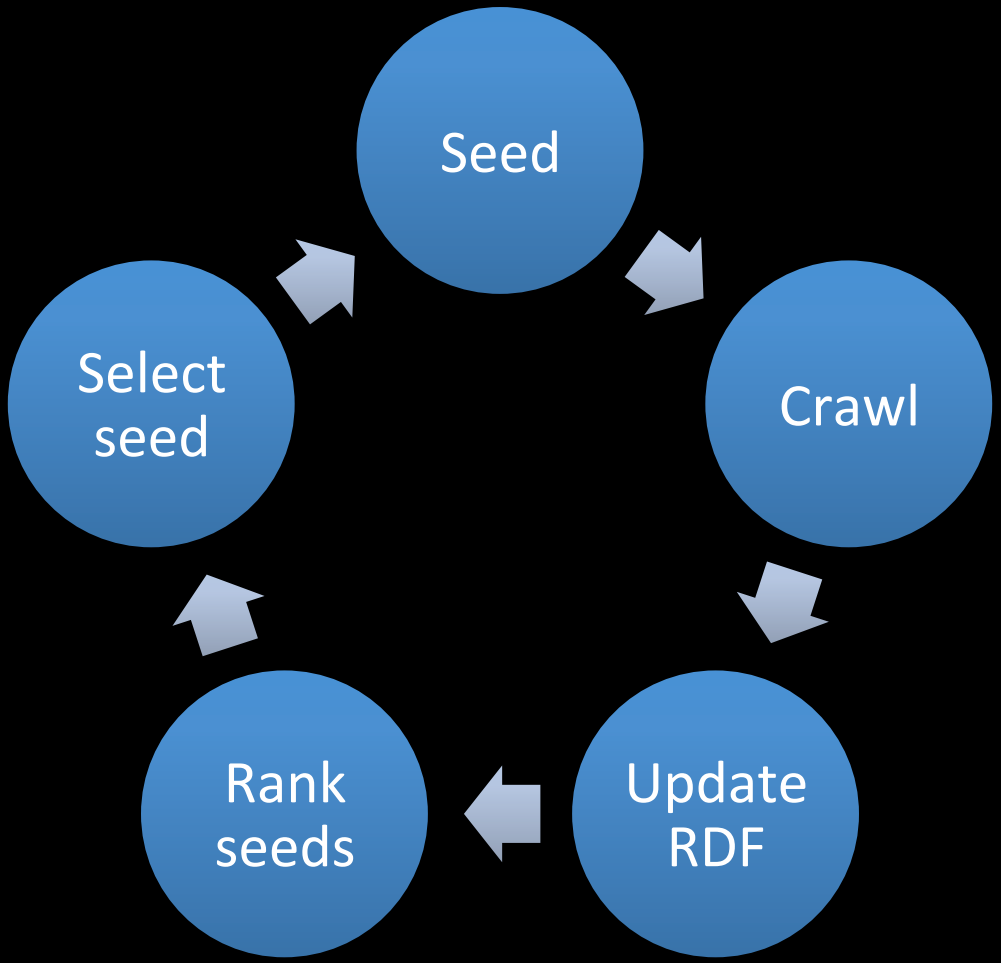
Final ranking and seed selection

- Rank each author s by
  - $w_1 * A(s \mid r) + w_2 * P(s \mid r)$
  - $w_1 + w_2 = 1$
- 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	0.3333333333333333
Christopher L. Barrett	0.23529411764705882
V. S. Anil Kumar	0.2012765957446807
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



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