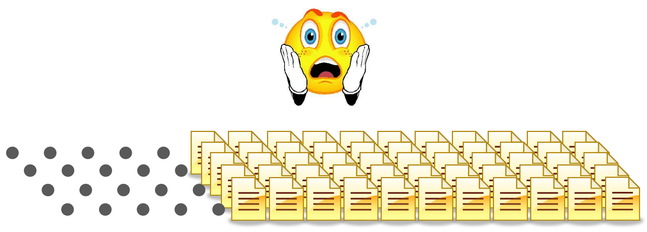
Personalized Reading Recommendations for *Saccharomyces* Genome Database

Motivation

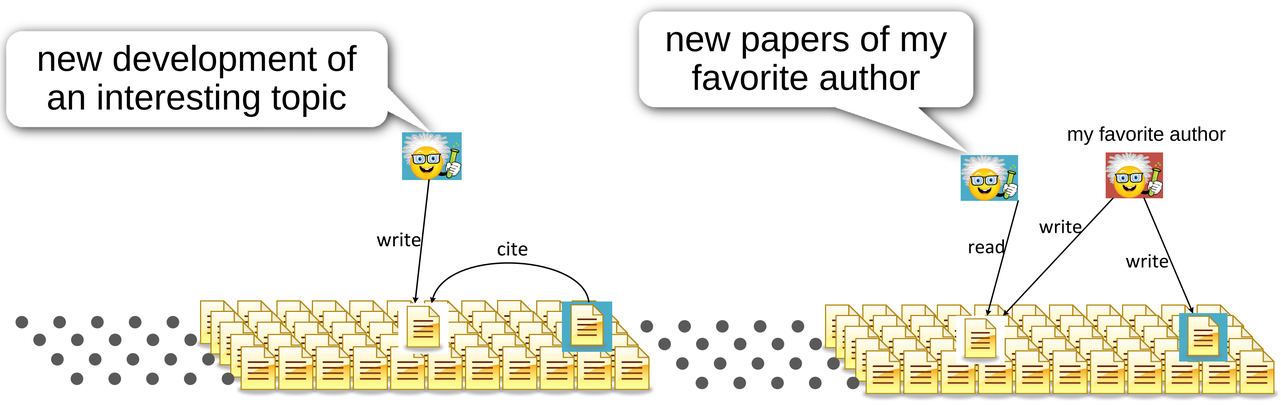
Information Overload for Scientists

The rapid growth of research in biology, and the increasing degree to which different subareas of biology are connected, make it difficult to monitor the published literature effectively.



a paper river

Recommendation strategies with rich meta-data



Path Ranking Algorithm (PRA)

Combines logic, random walks, and statistical learning (Lao & Cohen, ECML 2010)

Link Prediction Task

Given
- a directed edge-labeled graph
- a relation type r
- a source node s (also called a query)

Find
- the set of nodes G, s.t. r(s,t) for each t in G

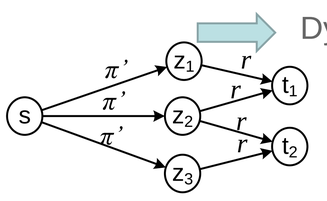
Ranking

$$score(s,t) = \sum_{\pi \in B} P(s \to t;\pi)\theta_\pi$$

e.g. π=<*Read*,*Write*$^{-1}$,*Write*>

expressive | robust

Path-Constrained Random Walks



$$P(s \to t;\pi) = \sum_z P(s \to z;\pi')P(z \to t;r)$$

scalable — Sampling can make it x100 more efficient (Lao & Cohen, KDD'10)

Feature Selection (B) with Labeled Data

given training query set $\{(s_i, G_i)\}$

$$hits(f) = \sum_i I\left[P(s_i \to G_j \mid \pi)\right] \geq h$$

$$accuracy(\pi) = \frac{1}{N}\sum_i P(s_i \to G_j \mid \pi) \geq a$$

I(): the indicator function

N: total number of queries

Parameter Estimation (θ)

For a recommendation relation r
generate positive and negative node pairs $\{(s_i, t_i)\}$

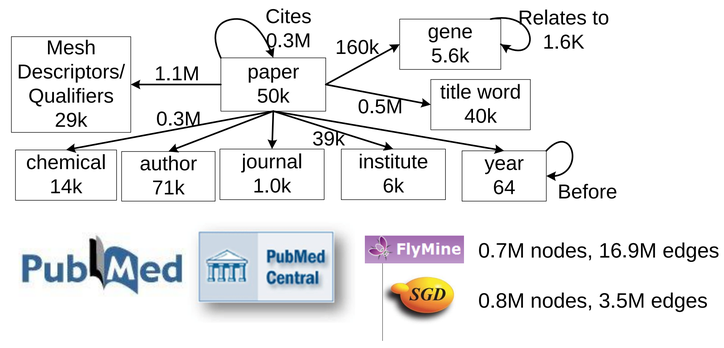
For each (s_i, t_i) generate (x_i, y_i)
x_i is a vector of RW features of different paths π
y_i is a binary label $r(s_i, t_i)$

Estimate θ by elastic-net logistic regression

$$\theta = \arg\max_\theta \left[\sum_i l_i(\theta, x_i, y_i) - \lambda_1 \|\theta\|_1 - \lambda_2 \|\theta\|_2^2\right]$$

Literature Recommendation

Dataset

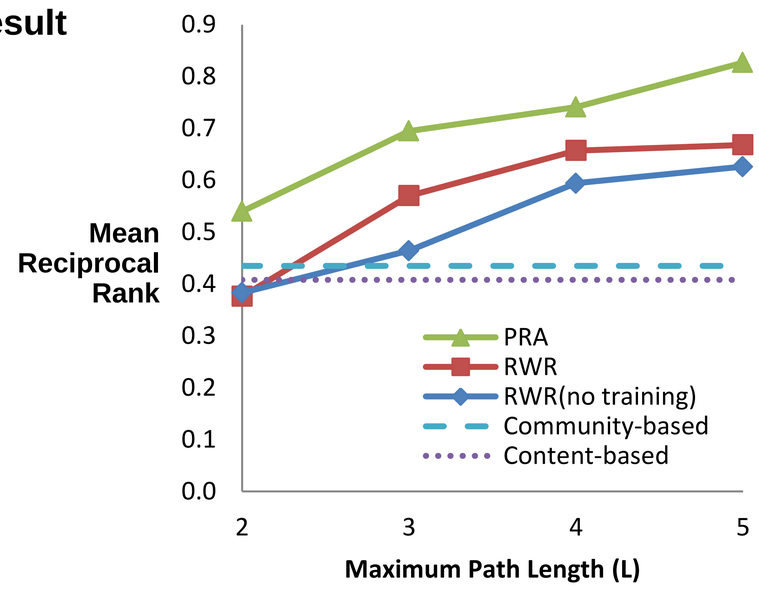


Task

Given a user predict papers this user is going to read

Over 20 years' data collected from Dr. John Woolford's computer

Result



ID	Path	Comments
ℓ=2		
1	author $\xrightarrow{\text{Read}}$ paper $\xrightarrow{\text{Cite}^{-1}}$ paper	follow up papers to what I read
ℓ=3		
2	author $\xrightarrow{\text{Read}}$ paper $\xrightarrow{\text{Write}^{-1}}$ author $\xrightarrow{\text{Write}}$ paper	papers from my favorite authors
3	author $\xrightarrow{\text{Read}}$ paper $\xrightarrow{\text{Read}^{-1}}$ author $\xrightarrow{\text{Write}}$ paper	papers of scientist who read the same papers as I do
4	author $\xrightarrow{\text{Read}}$ paper $\xrightarrow{\text{HasMajorMQ}}$ topic $\xrightarrow{\text{HasMajorMQ}^{-1}}$ paper	papers about my favorite topics
5	author $\xrightarrow{\text{Read}}$ paper $\xrightarrow{\text{HasTitle}}$ word $\xrightarrow{\text{HasTitle}^{-1}}$ paper	papers with similar titles to what I read before
6	author $\xrightarrow{\text{Read}}$ paper $\xrightarrow{\text{Cite}}$ paper $\xrightarrow{\text{Cite}^{-1}}$ paper	papers which cite the same papers as what I read
ℓ=4		
7	year $\xrightarrow{\text{After}}$ year $\xrightarrow{\text{Read}}$ paper $\xrightarrow{\text{Cite}}$ paper $\xrightarrow{\text{Cite}^{-1}}$ paper	papers which share citations with what I read last year
8	author $\xrightarrow{\text{Read}}$ paper $\xrightarrow{\text{Write}^{-1}}$ author $\xrightarrow{\text{Write}}$ paper	papers from my favorite authors
9	author $\xrightarrow{\text{Read}}$ paper $\xrightarrow{\text{HasMajorMQ}}$ topic $\xrightarrow{\text{HasMajorMQ}^{-1}}$ paper	papers about my favorite topics
10	year $\xrightarrow{\text{After}}$ year $\xrightarrow{\text{Read}}$ paper $\xrightarrow{\text{HasMD}}$ topic $\xrightarrow{\text{HasMD}^{-1}}$ paper	papers involving MeSH descriptors I read about last year
11	author $\xrightarrow{\text{Read}}$ paper $\xrightarrow{\text{Read}^{-1}}$ author $\xrightarrow{\text{Write}}$ paper	papers of scientist who read the same papers as I do
ℓ=5		
12	year $\xrightarrow{\text{After}}$ year $\xrightarrow{\text{After}}$ year $\xrightarrow{\text{Read}}$ paper $\xrightarrow{\text{HasMajorMQ}}$ topic $\xrightarrow{\text{HasMajorMQ}^{-1}}$ paper	papers which share MeSH descriptors with what I read 2 years back
13	year $\xrightarrow{\text{After}}$ year $\xrightarrow{\text{Read}}$ paper $\xrightarrow{\text{Cite}}$ paper $\xrightarrow{\text{Read}^{-1}}$ author $\xrightarrow{\text{Write}}$ paper	papers by users who read what I cited last year
14	year $\xrightarrow{\text{After}}$ year $\xrightarrow{\text{Read}}$ paper $\xrightarrow{\text{HasTitle}}$ word $\xrightarrow{\text{HasTitle}^{-1}}$ paper	papers which share title words with what I read last year
15	author $\xrightarrow{\text{Write}}$ paper $\xrightarrow{\text{HasMajorMQ}}$ topic $\xrightarrow{\text{HasMQ}^{-1}}$ paper	papers which share MeSH qualifier with what published
16	year $\xrightarrow{\text{After}}$ year $\xrightarrow{\text{After}}$ year $\xrightarrow{\text{Read}}$ paper $\xrightarrow{\text{HasTitle}}$ word $\xrightarrow{\text{HasTitle}^{-1}}$ paper	papers involving title words I read about 2 years back

Carnegie Mellon

Ni Lao, William W. Cohen {nlao, wcohen}@cs.cmu.edu

This work is supported by NSF grant IIS-0811562 and NIH grant R01GM081293

lti

ML MACHINE LEARNING DEPARTMENT