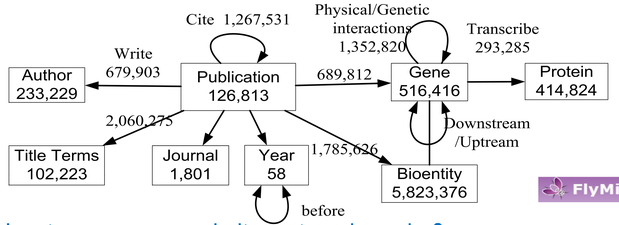
Relational Retrieval Using a Combination of Path-Constrained Random Walks

Relational Retrieval and Proximity Measures

Retrieval with Rich Meta-Data

Gene recommendation: author, year→gene
Reference recommendation: title words, year→paper
Expert-finding: title words, genes→author
Venue recommendation: title words, genes→venue



How to measure proximity on typed graphs?

The Limitation of RWR proximity measure:

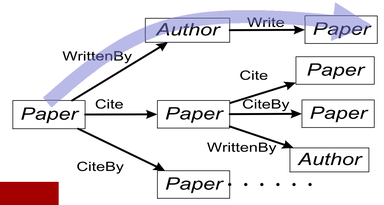
Random Walks with Restart (RWR) is a commonly used similarity measure on labeled graphs. It can be improved by supervised learning of edge weights. However, its one-parameter-per-edge label is limited because the context of an edge label appears is ignored

Path	Comments
$author \xrightarrow{Read} paper \xrightarrow{Contain} gene \xrightarrow{Contain^{-1}} paper$	Don't read about **genes** which I have already read
$author \xrightarrow{Read} paper \xrightarrow{Write^{-1}} author \xrightarrow{Write} paper$	Read about my favorite **authors**
Path	**Comments**
$author \xrightarrow{Write} paper \xrightarrow{Contain} gene \xrightarrow{Contain^{-1}} paper$	Read about the **genes** that I am working on
$author \xrightarrow{Write} paper \xrightarrow{publish^{-1}} institute \xrightarrow{publish} paper$	Don't need to read paper from **my own lab**

Path Constrained Random Walk

Given a query $q=(\mathbf{E}_q, T_q)$, recursively define a distribution for each path

$$h_{E_q,P}(e) = \sum_{e' \in range(P')} h_{E_q,P'}(e') \cdot \frac{R_l(e',e)}{|R_l(e')|}$$



Path Ranking Algorithm (PRA)

Retrieval model

A retrieval model can rank target entities by linearly combine the distributions of different paths

$$score(e;\theta,L) = \sum_{P \in \mathbf{P}(q,L)} h_P(e)\theta_P$$

in matrix form s=Aθ

Parameter Estimation

Given a set of training data

D={($q^{(m)}$, $A^{(m)}$, $y^{(m)}$)}, m=1…M, $y^{(m)}(e)$=1/0

define a regularized objective function

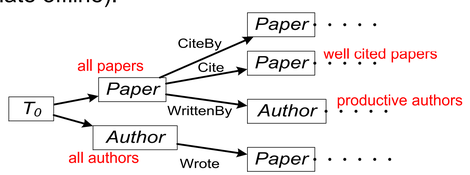
$$O(\theta) = \sum_{m=1..M} o_m(\theta) - \lambda_1 |\theta|_1 - \lambda_2 |\theta|_2 / 2$$

$$o_m(\theta) = |P_m|^{-1} \sum_{i \in P_m} \ln p_i^{(m)} + |N_m|^{-1} \sum_{i \in N_m} \ln(1 - p_i^{(m)})$$

$$p_i^{(m)} = p(y_i^{(m)} = 1 | q^{(m)};\theta) = \frac{\exp(\theta^T A_i^{(m)})}{1 + \exp(\theta^T A_i^{(m)})}$$

Ext.1: Query Independent Paths

Generalize PageRank to multiple entity and relation type setting (can be calculate offline).



Ext.2: Popular Entity Biases

There are entity specific characteristics which cannot be captured by a general model

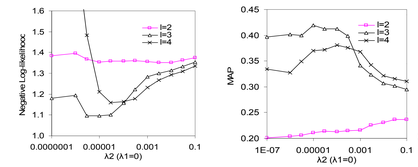
$$s(e;\theta) = \sum_{P.T_{last}=T_q} h_P^T(e)\theta_P + \theta_e + \sum_{e' \in \mathcal{E}_q} \theta_{e',e},$$

in matrix form $s = A\theta + \theta^{(b)} + \Theta q$

For efficiency we add to the model top J parameters (measured by $|O(\theta)/\theta_e|$) at each LBFGS iteration.

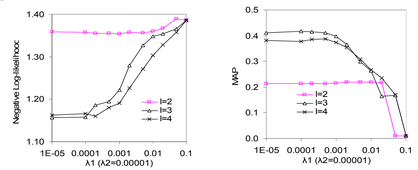
Results

L2 Regularization improves the retrieval quality



Reference recommendation task on Yeast data

L1 Regularization reduces the number of features



Reference recommendation task on Yeast data

Example Features

ID	Weight	Feature	
1	272.4	$word \to paper \xrightarrow{Cite^{-1}} paper \xrightarrow{Cite} paper$	1) papers co-cited with the on-topic papers
2	156.7	$word \to paper \xrightarrow{Cite} paper$	2) Aggregated citations of the on-topic papers
3	100.5	$gene \to paper \xrightarrow{Cite^{-1}} paper \xrightarrow{Cite} paper$	
4	83.7	$word \to paper \xrightarrow{Cite^{-1}} paper$	
5	50.2	$gene \to paper \xrightarrow{Cite} paper$	
6	41.4	$word \to paper$	6) resembles an ad-hoc retrieval system
7	29.3	$year \to paper \xrightarrow{Cite} paper$	7,8) papers cited during the past two years
8	13.0	$year \xrightarrow{Before^{-1}} year \to paper \xrightarrow{Cite} paper$	
	...		
9	3.7	$T^* \to paper \xrightarrow{Cite} paper$	9) well cited papers
10	2.9	GAL4>Nature. 1988. GAL4-VP16 is an unusually potent transcriptional activator.	
11	2.1	CYC1>Cell. 1979. Sequence of the gene for iso-1-cytochrome c in Saccharomyces cerevisiae.	
	...		10,11) (important) early papers about specific query terms (genes)
12	-5.4	$year \xrightarrow{Before^{-1}} year \to paper$	
13	-39.1	$year \to paper$	12,13) general papers published during the past two years
14	-49.0	$T^* \to year \to paper$	14) old papers

Main Result

A PRA+qip+pop model trained for the reference recommendation task on the yeast data

Corpus	Task	RWR	PRA			
		Trained	Trained	+qip	+pop	+qip +pop
Yeast	Ven	44.2	45.7 (+3.4)	46.4 (+5.0)	48.7 (+10.2)	49.3 (+11.5)
Yeast	Ref	16.0	16.9 (+5.6)	18.3 (+14.4)	19.1 (+19.4)	19.8 (+23.8)
Yeast	Exp	11.1	11.9 (+7.2)	12.4 (+11.7)	12.5 (+12.6)	12.9 (+16.2)
Yeast	Gen	14.4	14.9 (+3.5)	15.1 (+4.9)	15.1 (+4.9)	15.3 (+6.3)
Fly	Ven	48.3	50.4 (+4.3)	51.1 (+5.8)	50.7 (+5.0)	51.7 (+7.0)
Fly	Ref	20.5	20.8 (+1.5)†	21.0 (+2.4)	21.6 (+5.4)	21.7 (+5.9)
Fly	Exp	7.2	7.6 (+5.6)†	8.3 (+15.3)	7.9 (+9.7)	8.5 (+18.1)
Fly	Gen	19.2	20.7 (+7.8)	21.1 (+9.9)	21.1 (+9.9)	21.0 (+9.4)

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

--- in Machine Learning 2010 ---