Relational Retrieval Using a Combination of Path-Constrained Random Walks

Ni Lao, William W. Cohen

Carnegie Mellon University

2010.9.22

Outline

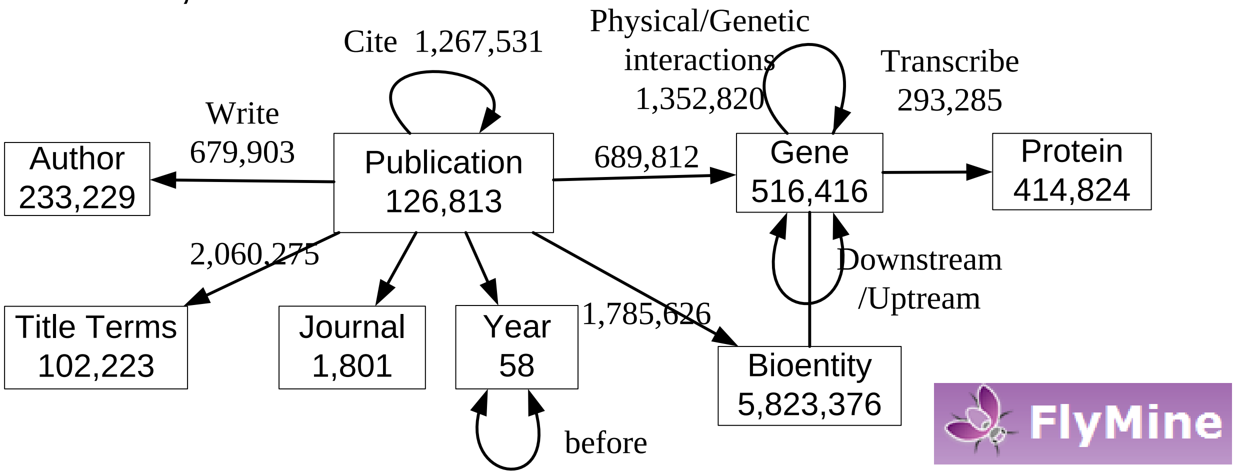
- Relational Retrieval Problems
  - Path-constrained random walks
  - The need for retrieval strategy mining
- Retrieval Models with PCRW
  - Path Ranking Algorithm (PRA)
  - Ext.1: query-independent experts (generalization of PageRank)
  - Ext.2: popular entity experts
- Experiment

Relational Retrieval Problems

- Data of many retrieval/recommendation tasks can be represented as labeled directed graphs
  - E.g. scientific literature
  - Typed nodes: documents, terms, metadata
  - Labeled edges: citation, authorOf, datePublished
- Can support a family of *typed proximity queries*
  - ad hoc retrieval: term nodes → documents
  - Reference (citation) recommendation: topic → paper
  - Expert finding: topic → user
  - Collaborator recommendation : scientist → scientist
- How to measure the proximity between query and target nodes?

Biology Literature Data

- Data of this study
  - Yeast: 0.2M nodes, 5.5M links
  - Fly: 0.8M nodes, 3.5M links

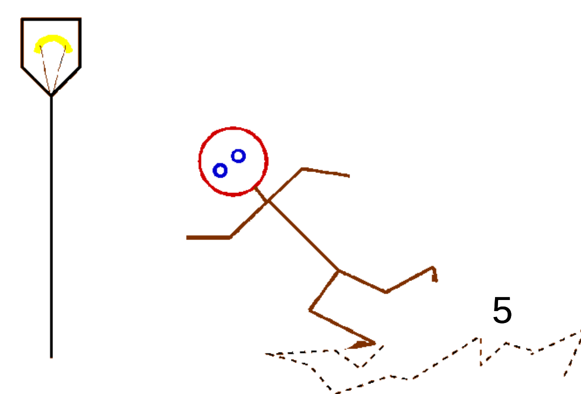


- Tasks
  - Gene recommendation: author, year→gene
  - Reference recommendation: title words,year→paper
  - Venue recommendation: genes, title words→journal
  - Expert-finding: title words, genes→author

Random Walks with Restart (RWR) as A Proximity Measured

- RWR is a commonly used similarity measure on Labeled Graphs
  - Topic-sensitive Pagerank (Haveliwala, 2002)
  - Personalized Pagerank (Jeh &. Widom, 2003)
  - ObjectRank (Balmin et al., 2004),
  - Personal information management (Minkov & Cohen, 2007)

- RWR can be improved by supervised learning of edge weights
  - quadratic programming (Tsoi et al., 2003),
  - simulated annealing (Nie et al., 2005),
  - back-propagation (Diligenti et al., 2005; Minkov & Cohen, 2007),
  - limit memory Newton method (Agarwal et al., 2006)



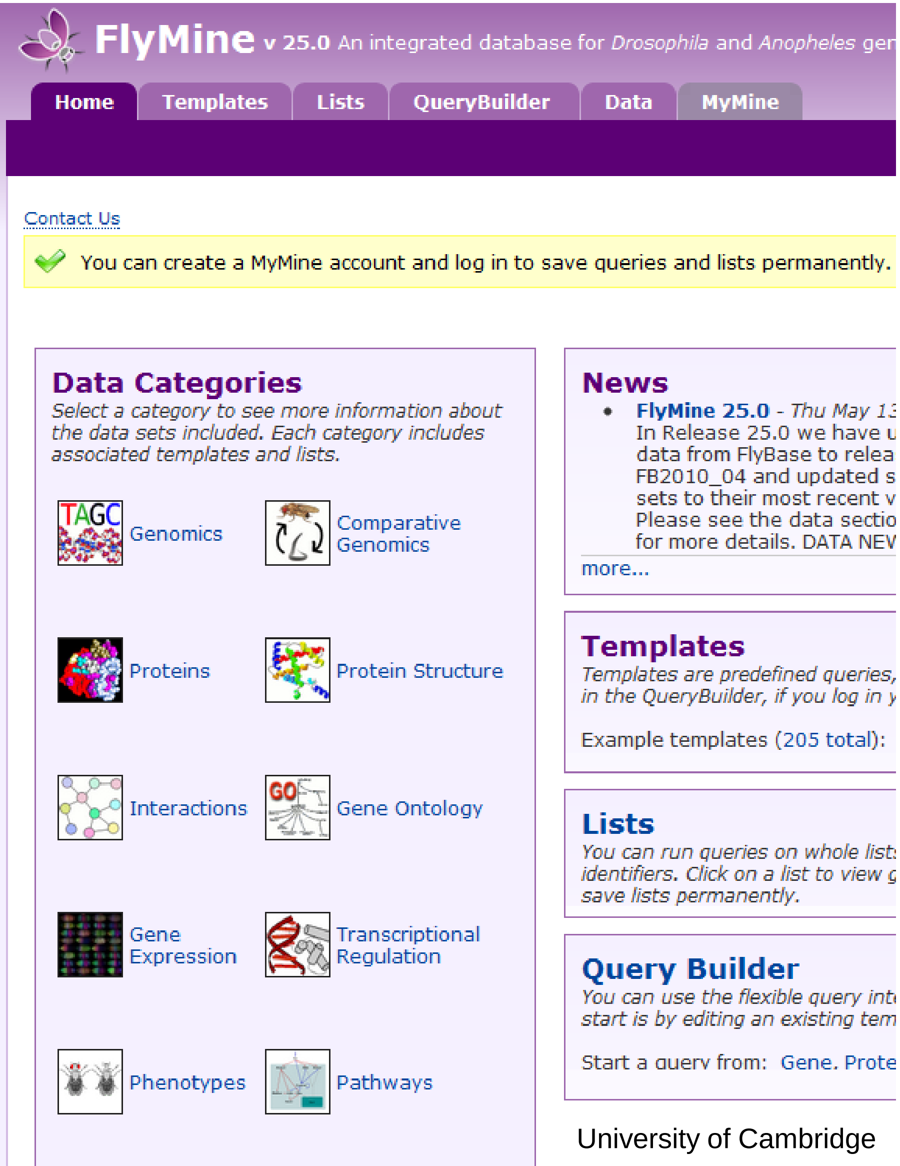
The Limitation of RWR

- One-parameter-per-edge label is limited because the context in which an edge label appears is ignored
  - E.g. (observed from real data)

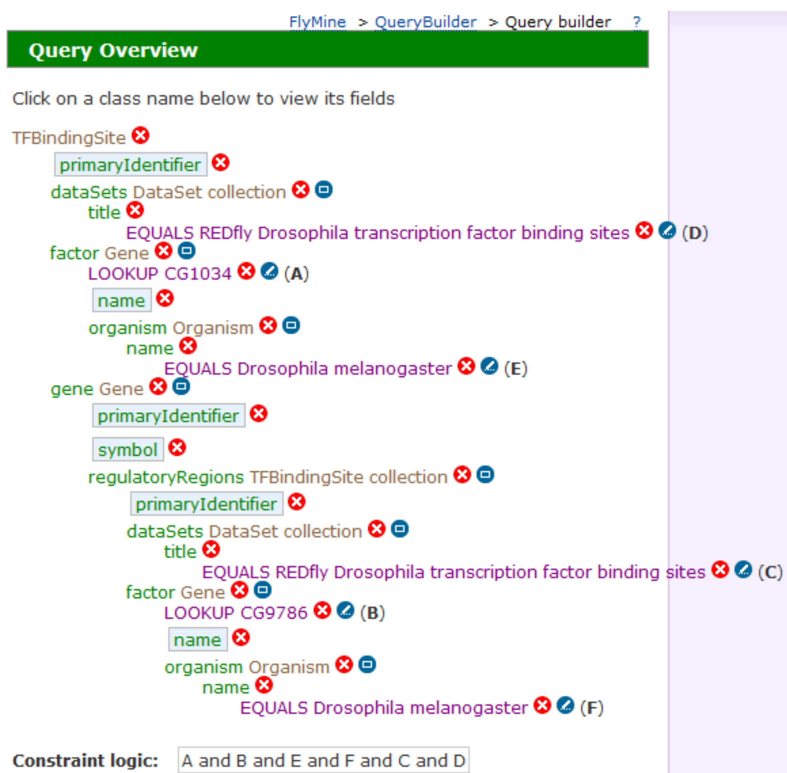
Path	Comments
$author \xrightarrow{Read} paper \xrightarrow{Contain} gene \xrightarrow{Contain^{-1}} paper$	Don't read about **genes** which I have already studied
$author \xrightarrow{Read} paper \xrightarrow{Write^{-1}} author \xrightarrow{Write} paper$	Read about my favorite **authors**

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**

The Need for Retrieval Strategy Mining



- A ramification of considering paths on heterogeneous graph with complex schema
- Consider non-expert users….
  - Who are willing to give some labels



University of Cambridge

The Proposed Approach

- Automatically generate, evaluate and combine different retrieval strategies (paths)
- An example -- reference recommendation
  - In the TREC-CHEM Prior Art Search Task, researchers found that it is more effective to first find patents about the topic, then aggregate their citations
  - Our proposed model can discover this kind of retrieval schemes and assign proper weights to combine them. E.g.

Weight Path

$$272.4 \; word \xrightarrow{HasTitle^{-1}} paper \xrightarrow{Cite^{-1}} paper \xrightarrow{Cite} paper$$

$$156.7 \; word \xrightarrow{HasTitle^{-1}} paper \xrightarrow{Cite} paper$$

$$41.4 \; word \xrightarrow{HasTitle^{-1}} paper$$

Outline

- Relational Retrieval Problems
  - Path-constrained random walks
  - The need for retrieval strategy mining
- Retrieval Models with PCRW
  - Path Ranking Algorithm (PRA)
  - Ext.1: query-independent experts (generalization of PageRank)
  - Ext.2: popular entity experts
- Experiment

Definitions

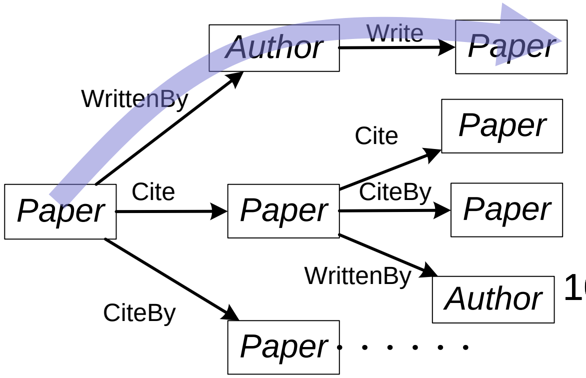
- An *Entity-Relation graph* G=(**T**,**E**,**R**), is
  - a set of *entities types* **T**={T}
  - a set of *entities* **E**={e}, Each entity is typed with $e.T \in$ **T**
  - a set of relations **R**={R}
- A *Relation path* $P=(R_1, \ldots, R_n)$ *is a sequence of relations*
  - E.g.

$$year \xrightarrow{PublishedIn^{-1}} paper$$

$$year \xrightarrow{PublishedIn^{-1}} paper \xrightarrow{Cite} paper$$

- *Path Constrained Random Walk*
  - Given a query $q=(\mathbf{E}_q, T_q)$
  - Recursively define a distribution for each path

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



Supervised PCRW Retrieval Model

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

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

  – or in matrix form $s=A\theta$

- This mode can be optimized by maximizing the probability of the observed relevance

$$p_e^{(m)}=p(y_e^{(m)}=1\,|\,q^{(m)};\theta)=\frac{\exp(\theta^T A_e^{(m)})}{1+\exp(\theta^T A_e^{(m)})}$$

  – Given a set of training data D={($q^{(m)}$, $A^{(m)}$, $y^{(m)}$)}, $y_e^{(m)}$=1/0

Parameter Estimation (Details)

- We can define a regularized objective function

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

- Use average log-likelihood as the objective $o_m(\theta)$

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

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

  - $P(m)$ the index set or relevant entities,
  - $N(m)$ the index set of irrelevant entities (sampled)

Parameter Estimation (Details)

- Selecting the negative entity set N_m
  - Few positive entities vs. thousands (or millions) of negative entities?
  - First sort all the negative entities with an initial model (uniform weight 1.0)
  - Then take negative entities at the k(k+1)/2-th position,

- The gradient

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

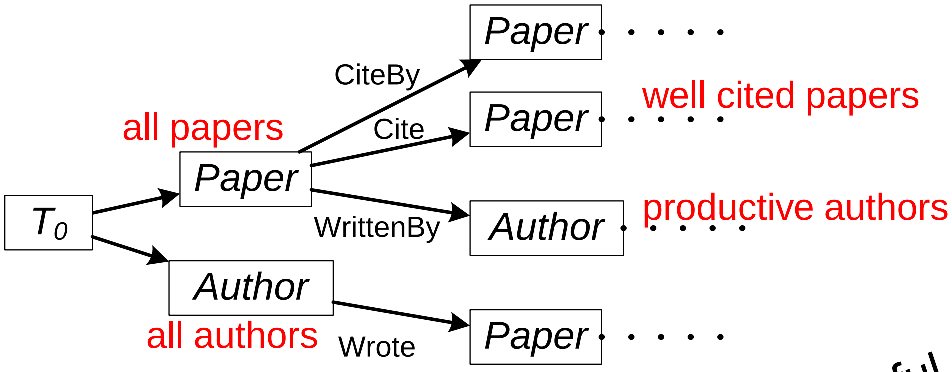
- Use orthant-wise L-BFGS (Andrew & Gao, 2007) to estimate θ
  - Efficient
  - Can deal with L1 regularization

Ext.1: Query Independent Paths

- PageRank
  - assign an importance score (query independent) to each web page
  - later combined with relevance score (query dependent)

- Generalize to heterogeneous graph
  - We include to each query a special entity e_0 of special type T_0
  - T_0 has relation to all other entity types, and e_0 has links to all entities
  - Therefore, we have a set of query independent relation paths
  - (distributions of which can be calculate offline)

- Example



Simple yet powerful idea

Ext.2: Popular Entity Biases

- There are entity specific characteristics which cannot be captured by a general model
  - E.g. log mining
  - Some document with lower rank to a query may be interesting to the users because of features not captured in the data
  - E.g. personalization
  - Different users may have completely different information needs and goals under the same query
  - The identity of entity matters

Ext.2: Popular Entity Biases

- For a task with query type T_0, and target type T_q,
  - Introduce a bias θ_e for each entity e in $I_E(T_q)$
  - Introduce a bias $\theta_{e',e}$ for each entity pair (e',e) where e in $I_E(T_q)$ and e' in $I_E(T_0)$
- Then $$s(e;\theta) = \sum_{P.T_{last}=T_q} h_P^T(e)\theta_P + \theta_e + \sum_{e' \in \mathcal{E}_q} \theta_{e',e},$$
  - Or in matrix form $s = A\theta + \theta^{(b)} + \Theta q$
- Efficiency consideration
  - Only add to the model top J parameters (measured by $|O(\theta)/\theta_e|$) at each LBFGS iteration

Outline

- Relational Retrieval Problems
  - Path-constrained random walks
  - The need for retrieval strategy mining
- Retrieval Models with PCRW
  - Path Ranking Algorithm (PRA)
  - Ext.1: query-independent experts (generalization of PageRank)
  - Ext.2: popular entity experts
- Experiment

Setup

- Data sources for bio-informatics
  - PubMed on-line archive of over 18 million biological abstracts
  - PubMed Central (PMC) full-text copies of over 1 million of these papers
  - Saccharomyces Genome Database (SGD) a database for yeast
  - Flymine a database for fruit flies

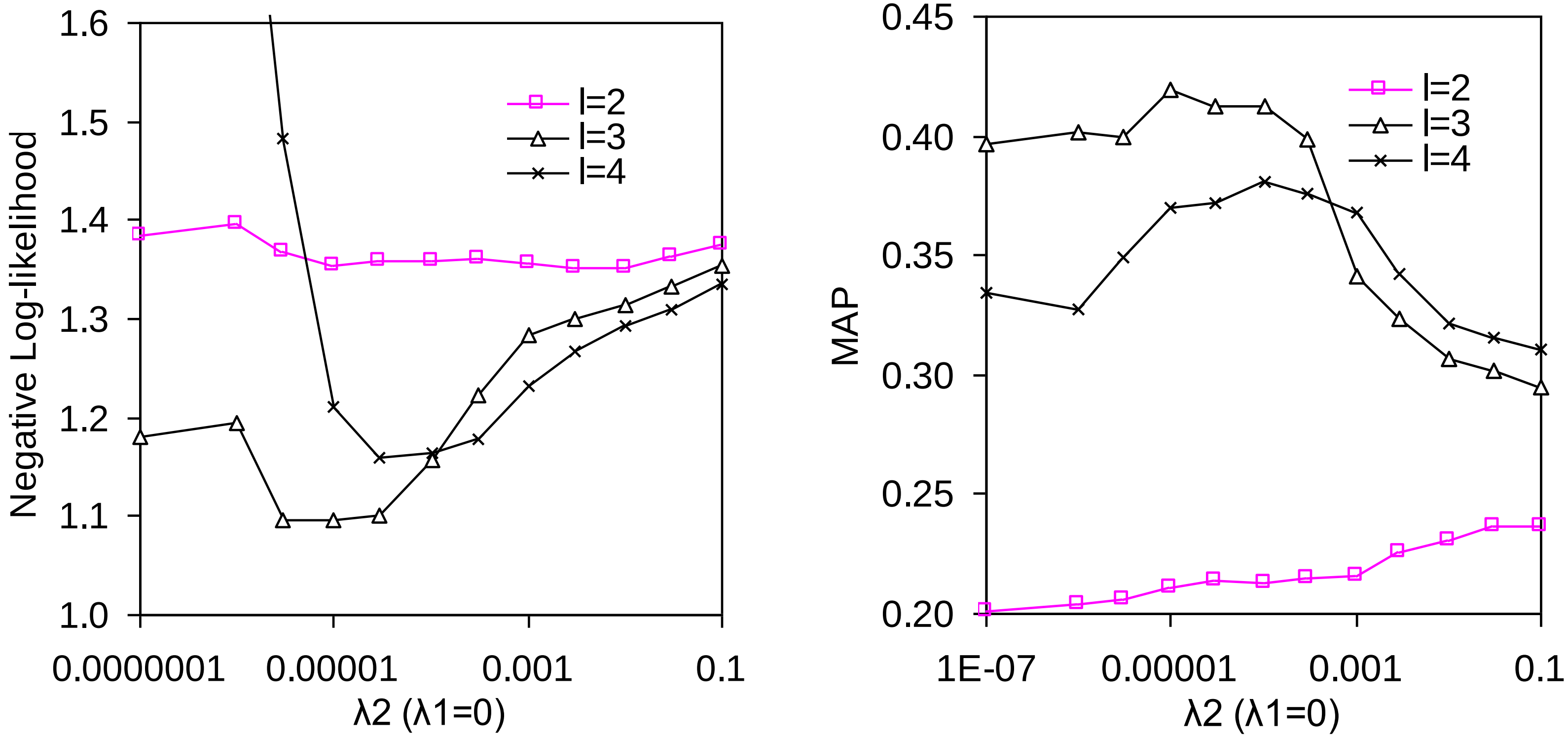
- Tasks
  - Gene recommendation: author, year→gene
  - Venue recommendation: genes, title words→journal
  - Reference recommendation: title words,year→paper
  - Expert-finding: title words, genes→author

- Data split
  - 2000 training, 2000 tuning, 2000 test

- Time variant graph (for training)
  - each edge is tagged with a time stamp (year)
  - only consider edges that are earlier than the query during random walk

L2 Regularization

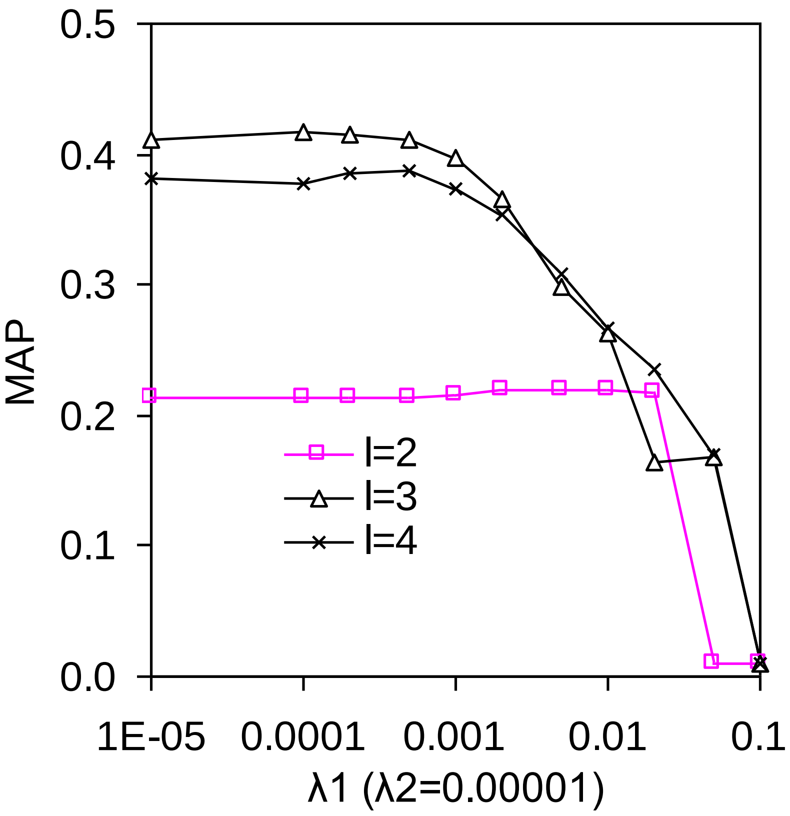
- Improves retrieval quality

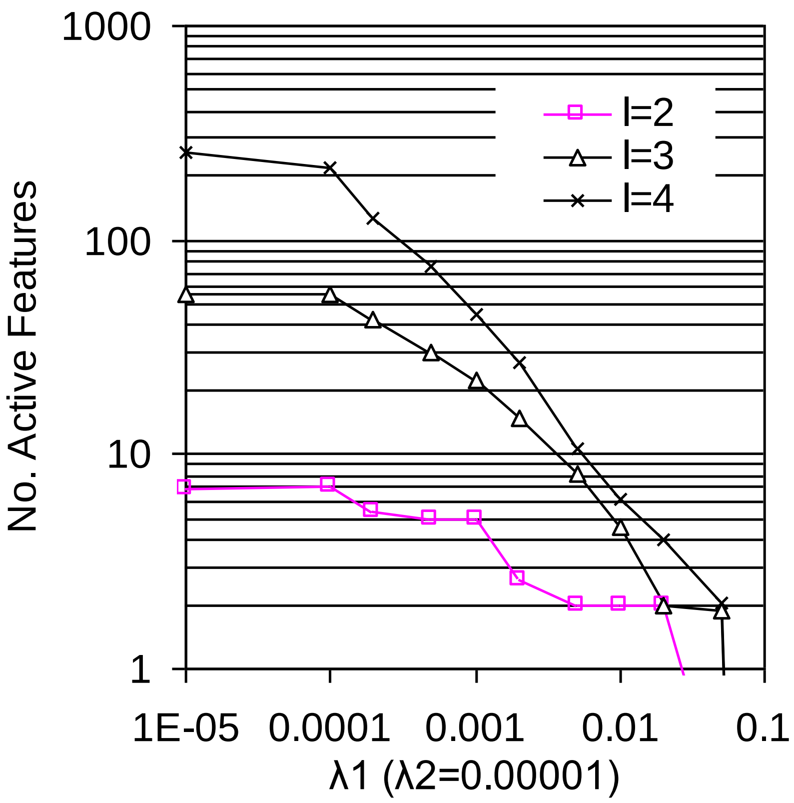


On the personal paper recommendation task

L1 Regularization

- Can help select features





On the personal paper recommendation task

Example Features

ID	Weight	Feature	
1	272.4	$word \rightarrow paper \xrightarrow{Cite^{-1}} paper \xrightarrow{Cite} paper$	1) Papers co-cited with the on-topic papers
2	156.7	$word \rightarrow paper \xrightarrow{Cite} paper$	2) Aggregated citations of the on-topic papers
3	100.5	$gene \rightarrow paper \xrightarrow{Cite^{-1}} paper \xrightarrow{Cite} paper$	
4	83.7	$word \rightarrow paper \xrightarrow{Cite^{-1}} paper$	
5	50.2	$gene \rightarrow paper \xrightarrow{Cite} paper$	
6	41.4	$word \rightarrow paper$	6) Resembles an ad-hoc retrieval system
7	29.3	$year \rightarrow paper \xrightarrow{Cite} paper$	
8	13.0	$year \xrightarrow{Before^{-1}} year \rightarrow paper \xrightarrow{Cite} paper$	7,8) Papers cited during the past two years
	...		
9	3.7	$T^* \rightarrow paper \xrightarrow{Cite} paper$	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	$year \xrightarrow{Before^{-1}} year \rightarrow paper$	
13	-39.1	$year \rightarrow paper$	12,13) General papers published during the past two years
14	-49.0	$T^* \rightarrow year \rightarrow paper$	14) old papers

A model trained for reference recommendation task on the yeast data

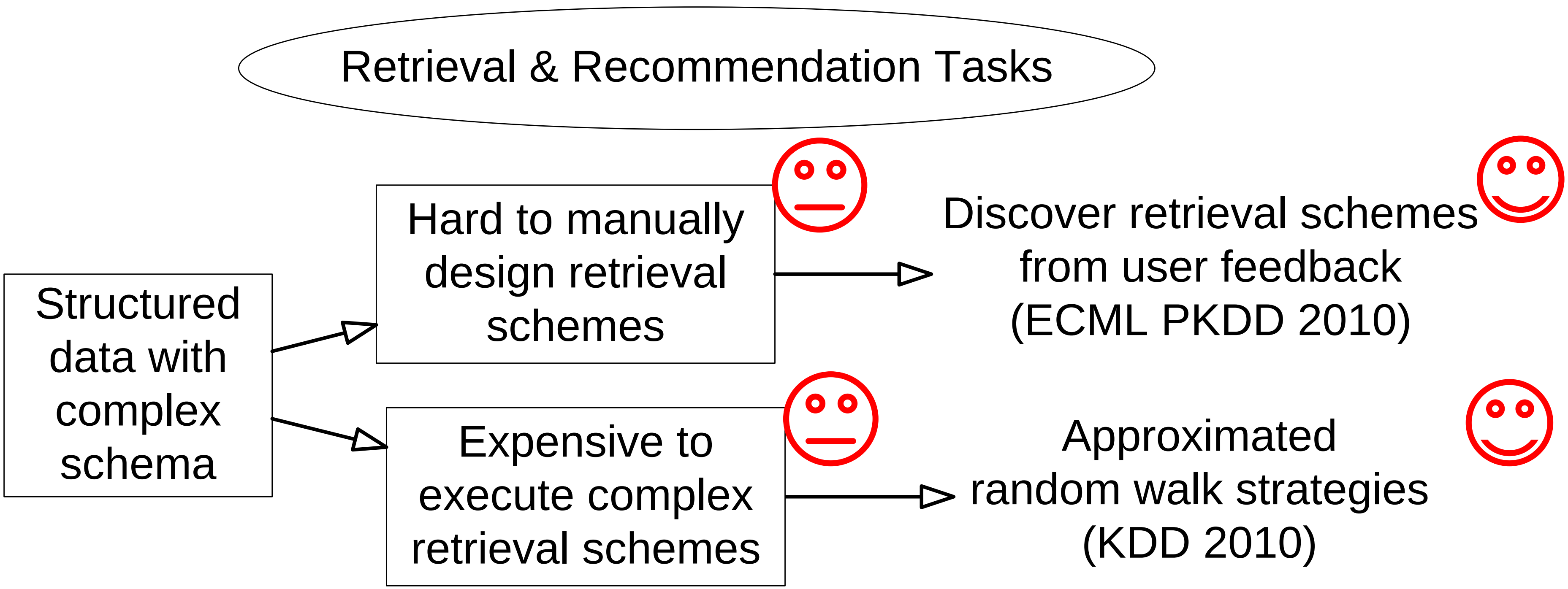
Evaluations

- Compare the MAP of PCRW to
  - Random walk with restart (RWR) model
  - Query independent paths (qip)
  - Popular entity biases (pop)

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)

Except these† , all improvements are statistically significant at $p<0.05$ using paired t-test

Summary


Retrieval & Recommendation Tasks
Structured data with complex schema
Hard to manually design retrieval schemes
Expensive to execute complex retrieval schemes
Discover retrieval schemes from user feedback (ECML PKDD 2010)
Approximated random walk strategies (KDD 2010)