Learning to Query: Focused Web Page Harvesting for Entity Aspects

Yuan Fang[1] Vincent Zheng[2] Kevin Chang[23]

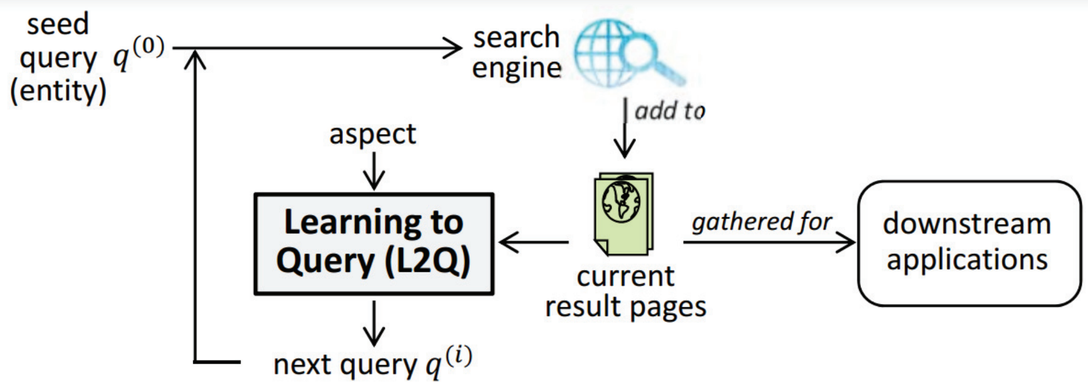
[1] Institute for Infocomm Research, Singapore
[2] Advanced Digital Sciences Center, Singapore
[3] University of Illinois at Urbana-Champaign, USA

Problem: *Learning to Query (L2Q)*

Entity type	Common aspects
celebrity	spouse, age, net worth
car	safety, cost, interior
business	address, hours, phone

abundant but scattered

Overall Workflow: *Iterative Querying*



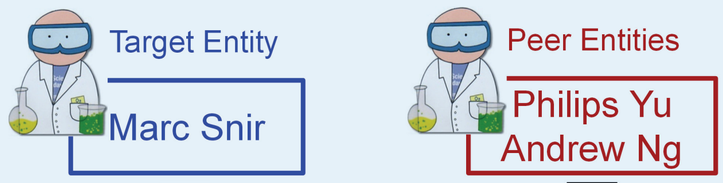
Seed query: Keywords (uniquely) identifying the entity

Target aspect: A pre-trained classifier Y for the target aspect

Utility: (precision/recall) In each iteration, $q^* = \arg\max_q \mathcal{U}^{(Y)}(q)$

Subproblem #1: *Domain-aware L2Q*

a) Domain Pages



entity pages P_E **+** ***domain pages*** P_D

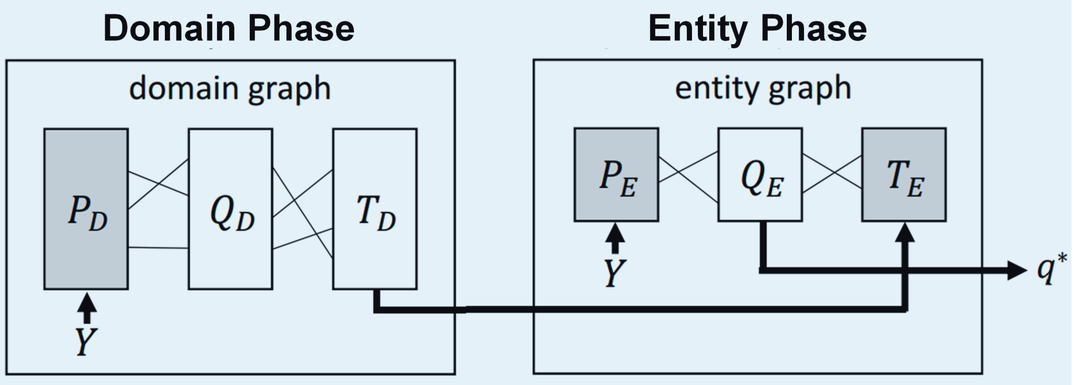
$$q^* = \arg\max_q \mathcal{U}^{(Y)}(q|P_E, P_D)$$

b) Template Abstraction

Entity	Example page content	Example query
Marc Snir	...many HPC papers in IJHPCA ...	`hpc ijhpca`
Philip Yu	...his data mining papers in TKDE ...	`data mining tkde`
Andrew Ng	...his recent AI paper in JMLR ...	`ai jmlr`

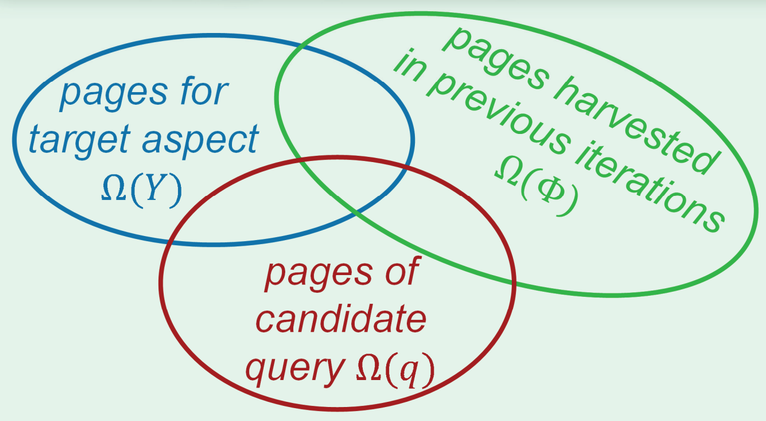
<topic> <journal>

c) Bridging Domain & Entity



Utility regularization (i.e. supervision on target aspect)

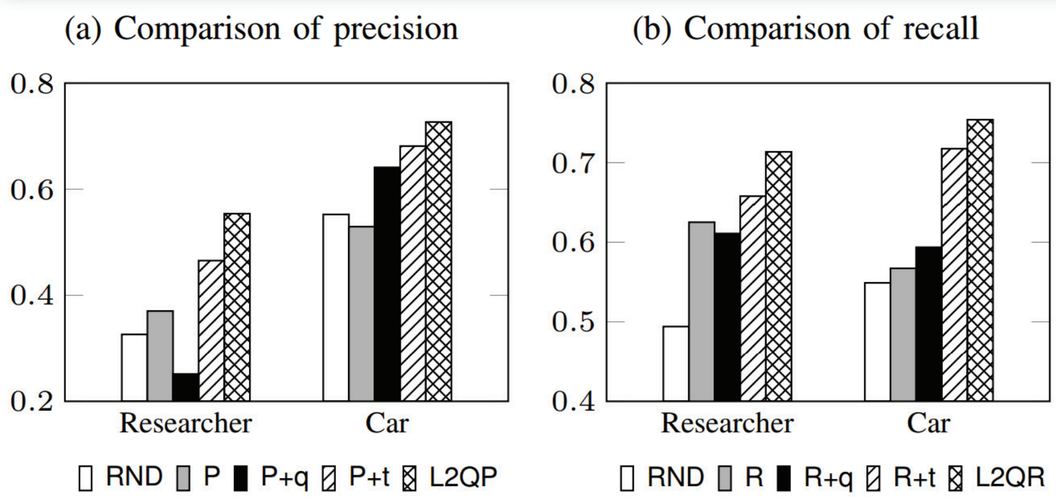
Subproblem #2: *Context-aware L2Q*



Collective precision
$$\frac{|(\Omega(q)\cup\Omega(\Phi))\cap\Omega(Y)|}{|\Omega(q)\cup\Omega(\Phi)|}$$

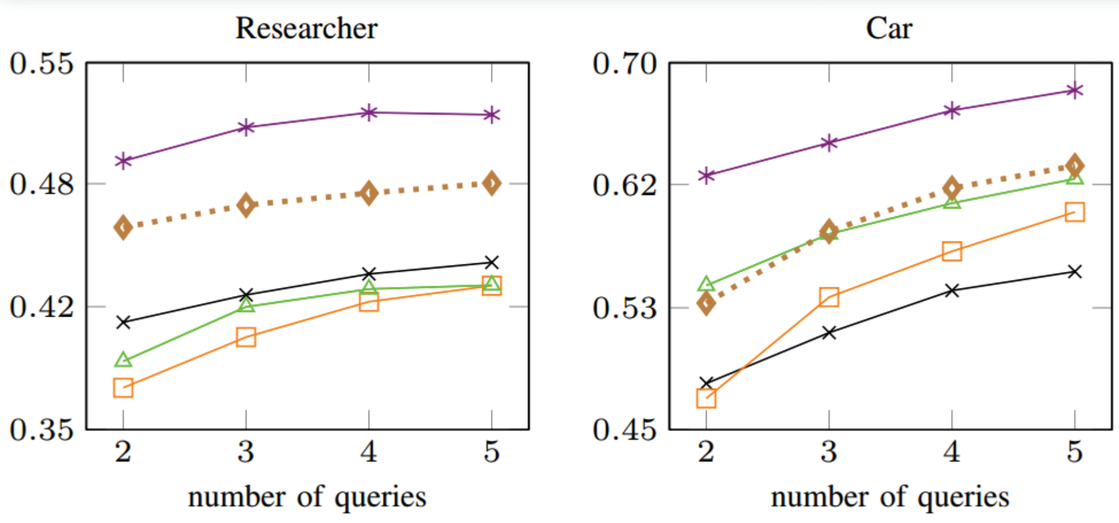
Collective recall
$$\frac{|(\Omega(q)\cup\Omega(\Phi))\cap\Omega(Y)|}{|\Omega(Y)|}$$

Result A: *Effect of Domain+Context*



RND: select query randomly
P/R: optimize precision/recall without domain and context
P/R+q: with domain pages but no templates, and without context
P/R+t: with domain pages and templates, without context
L2QP/L2QR: full approaches optimizing precision/recall

Result B: *Compare with Indep. Baselines*



L2QBAL: optimize for F-score, balancing L2QP & L2QR
LM: language feedback model
AQ: adaptive querying for text databases
HR: harvest rate for hidden structured databases
MQ: manually designed queries