

Learning to Query: Focused Web Page Harvesting for Entity Aspects

Yuan Fang¹, Vincent Zheng², Kevin Chang²³

ICDE 2016 @ Helsinki

¹ Institute for Infocomm Research, Singapore

² Advanced Digital Sciences Center, Singapore

³ University of Illinois at Urbana-Champaign, USA



Institute for
Infocomm Research



In this talk

2

- **Problem: L2Q**
- Challenges and solution
 - ▣ Domain-awareness
 - ▣ Context-awareness
- Experimental Study
- Conclusion

Entities and their aspects are abundant, but scattered, on the Web

3

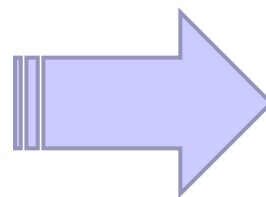
Entity type	Common aspects
celebrity	spouse, age, net worth, ...
car	safety, cost, interior, ...
business	address, opening hour, phone no., ...

People entities
alone: 10% of
Bing's search
volume



Motivation: Focused Web Page Harvesting for Entity Aspects

4

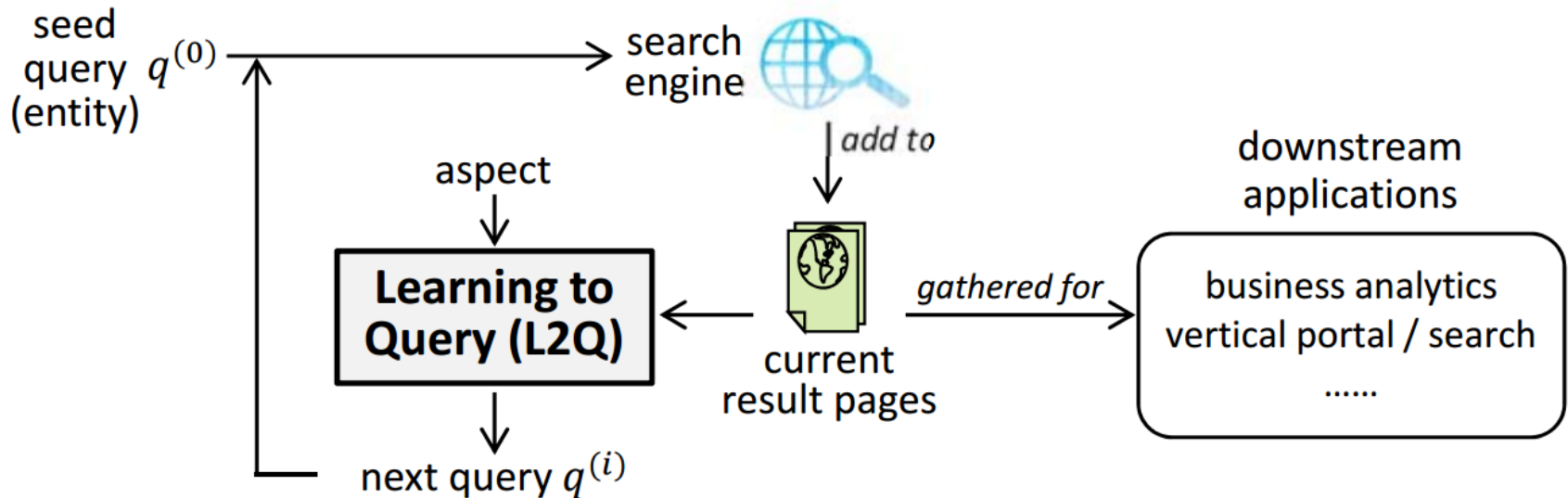


**Business
Analytics**

**Vertical portal
or search**

High level problem: Learning to query (L2Q)

5



Seed query

Keywords (uniquely) identifying the entity

Target aspect

A pre-trained classifier Y , mapping each page to “relevant” or “not relevant” to the target aspect

**Utility
(precision/recall)**

In each iteration, $q^* = \arg \max_q \mathcal{U}^{(Y)}(q)$

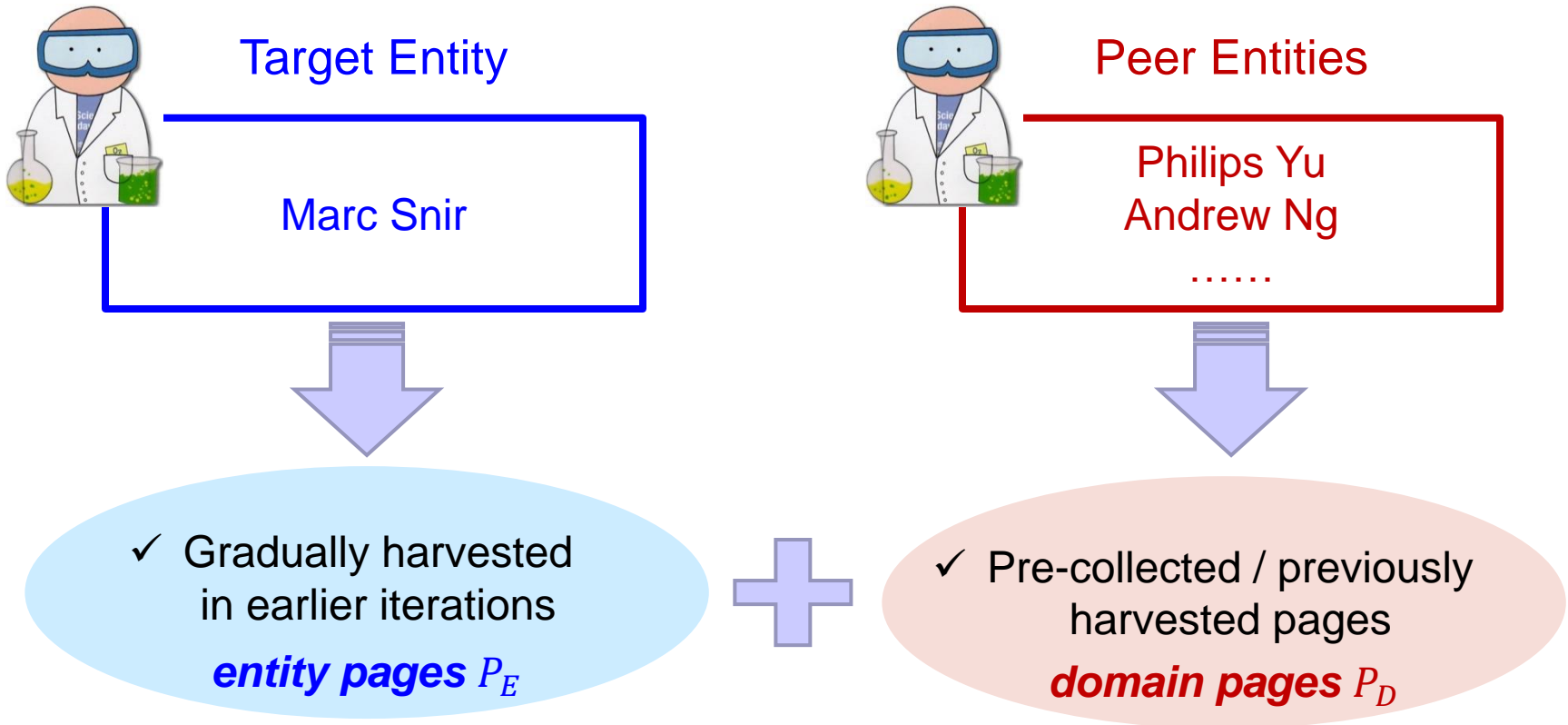
In this talk

6

- Problem: L2Q
- **Challenges and solution**
 - **Domain-awareness**
 - Context-awareness
- Experimental Study
- Conclusion

Subproblem #1: Domain-aware L2Q

7



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

Subproblem #1:

Vocabulary variations

8

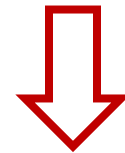
Entity	Example page content	Example query
Marc Snir Philip Yu Andrew Ng	...many HPC papers in IJHPCAhis data mining papers in TKDEhis recent AI paper in JMLR ...	hpc ijhpca data mining tkde ai jmlr

topics

hpc
data mining
ai
...

journals

ijhpca
tkde
jmlr
...

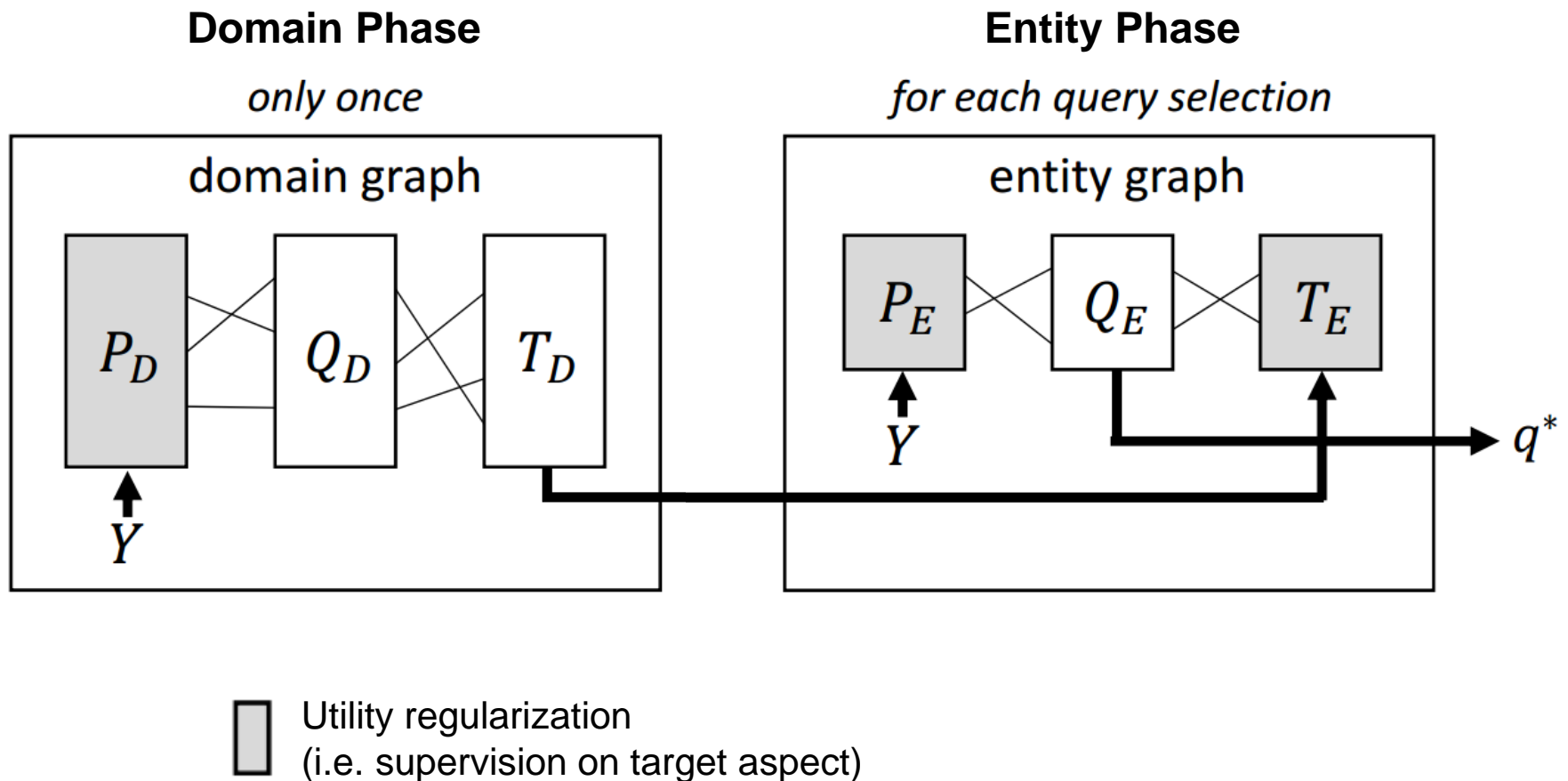


<topic> <journal>
(template)

Subproblem #1:

Bridging domain and entity phases

9



In this talk

10

- Problem: L2Q
- **Challenges and solution**
 - Domain-awareness
 - **Context-awareness**
- Experimental Study
- Conclusion

Subproblem #2:

Context-aware L2Q

11

- In iteration- i , a context of already fired queries $\Phi = \{q^{(0)}, q^{(1)}, \dots, q^{(i-1)}\}$.
- Queries can retrieve redundant pages

Marc Snir | Argonne National Laboratory

www.anl.gov/contributors/marc-snir ▼

Marc Snir is a parallel computing expert whose ongoing research and engagement in various supercomputing initiatives helps to advance the elite class of ...

Prof. Marc Snir Named "HPC Rock Star" | Department of Com...

<https://cs.illinois.edu/news/prof-marc-snir-named-hpc-rock-star> ▼

Jun 10, 2010 - Illinois computer science professor Marc Snir was named insideHPC.com's newest Rock Star of HPC. As the Faiman and Murgu Professor of ...

Marc Snir - Department of Computer Science at Illinois

<https://cs.illinois.edu/directory/profile/snir> ▼

Marc Snir. Michael Faiman and Saburo Muroga Professor. (217) 244-6568 NCSA receives NSF grant to develop Eclipse-based Workbench for HPC ...

Rock Stars of HPC: Marc Snir - insideHPC

insidehpc.com/2010/06/rock-stars-of-hpc-marc-snir/ ▼

Jun 10, 2010 - This month's HPC Rock Star is Marc Snir. During his time at IBM, Snir contributed to one of the most successful bespoke HPC architectures of ...

Marc Snir | LinkedIn

<https://www.linkedin.com/in/snimarc> ▼

Greater Chicago Area - Director, Mathematics and Computer Science Division at Argonne National Laboratory - Argonne National Laboratory
With exascale computing on the horizon, the performance variability of I/O systems represents a key challenge in sustaining high performance. In many HPC ...

Marc Snir - University of Illinois at Urbana-Champaign

snir.cs.illinois.edu/ ▼

Marc Snir is Director of the Mathematics and Computer Science Division at the Argonne National ... He currently pursues research in parallel computing. He was ...

Marc Snir - Department of Computer Science at Illinois

<https://cs.illinois.edu/directory/profile/snir> ▼

Marc Snir. Michael Faiman and Saburo Muroga Professor. (217) 244-6568 ... Architecture, Compilers, and Parallel Computing - Parallel Computing ...

Marc Snir | Argonne National Laboratory

www.anl.gov/contributors/marc-snir ▼

Marc Snir is a parallel computing expert whose ongoing research and engagement in various supercomputing initiatives helps to advance the elite class of ...

Marc Snir - Google Scholar Citations

scholar.google.com/citations?user=Hal6LesAAAAJ ▼

Argonne National Laboratory & University of Illinois at Urbana Champaign - mcs.anl.gov
The NYU Ultracomputer: Designing an MIMD Shared Memory Parallel Computer. A Gottlieb, R Grishman, CP Kruskal, KP McAuliffe, L Rudolph, M Snir.

Parallel computing pioneer Marc Snir to receive 2013 IEEE Se...

sc13.supercomputing.org > News and Media > Press Releases ▼

DENVER, CO – Dr. Marc Snir, a parallel computing pioneer whose innovative work has advanced the elite supercomputing systems that drive scientific ...

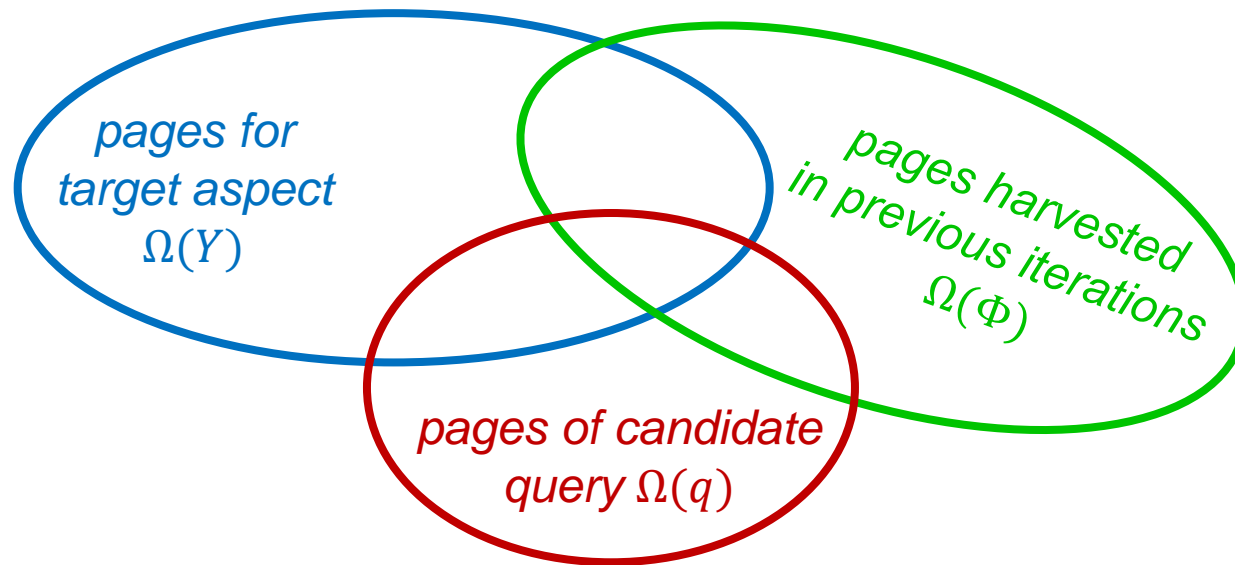
Marc Snir HPC

Marc Snir Parallel

Subproblem #2:

Accounting for redundancy

12



Collective Utilities

Collective precision :

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

Collective recall :

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

In this talk

13

- Problem: L2Q
- Challenges and solution
 - Domain-awareness
 - Context-awareness
- **Experimental Study**
- Conclusion

Experiment setup

14

- **Datasets:** two domains
 - 996 researchers & 143 car models
 - Pre-collected pages to simulate the corpus
- **Search engine:** language model
- **Dictionaries** for templates
 - Gathered from existing knowledge base
 - Manually compiled
- **Entity aspects**
 - 7 attributes for each domain
 - Pre-trained aspect classifier with high accuracy

Experiment methodology

15

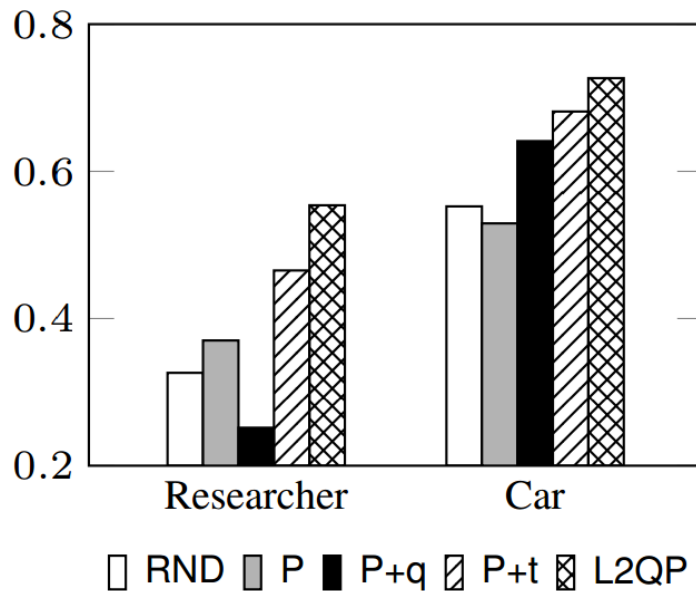
- **Utilities** of two forms: precision & recall
- **Evaluation metrics**
 - ▣ Precision, recall
 - ▣ Combined F-score
- Metrics reported are **normalized**
 - ▣ Against ideal precision/recall
 - ▣ Ideal metrics computed by “peeking” at unretrieved pages

Finding #1:

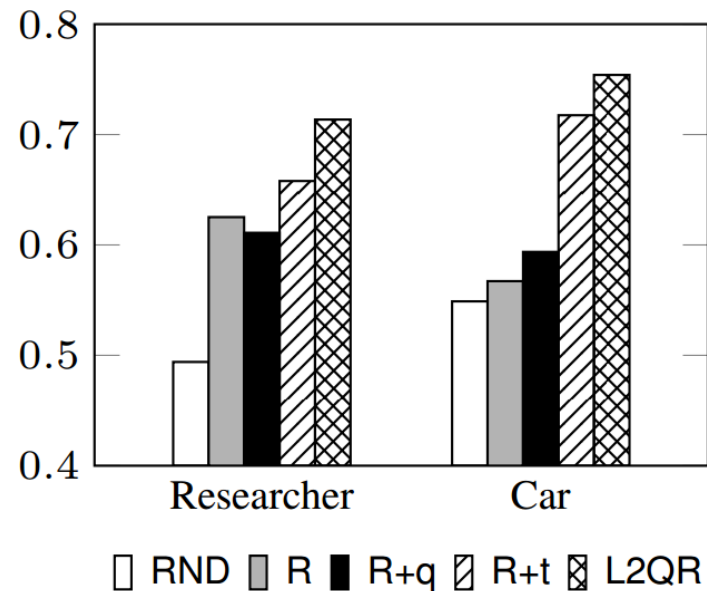
Effect of domain and context-awareness

16

(a) Comparison of precision



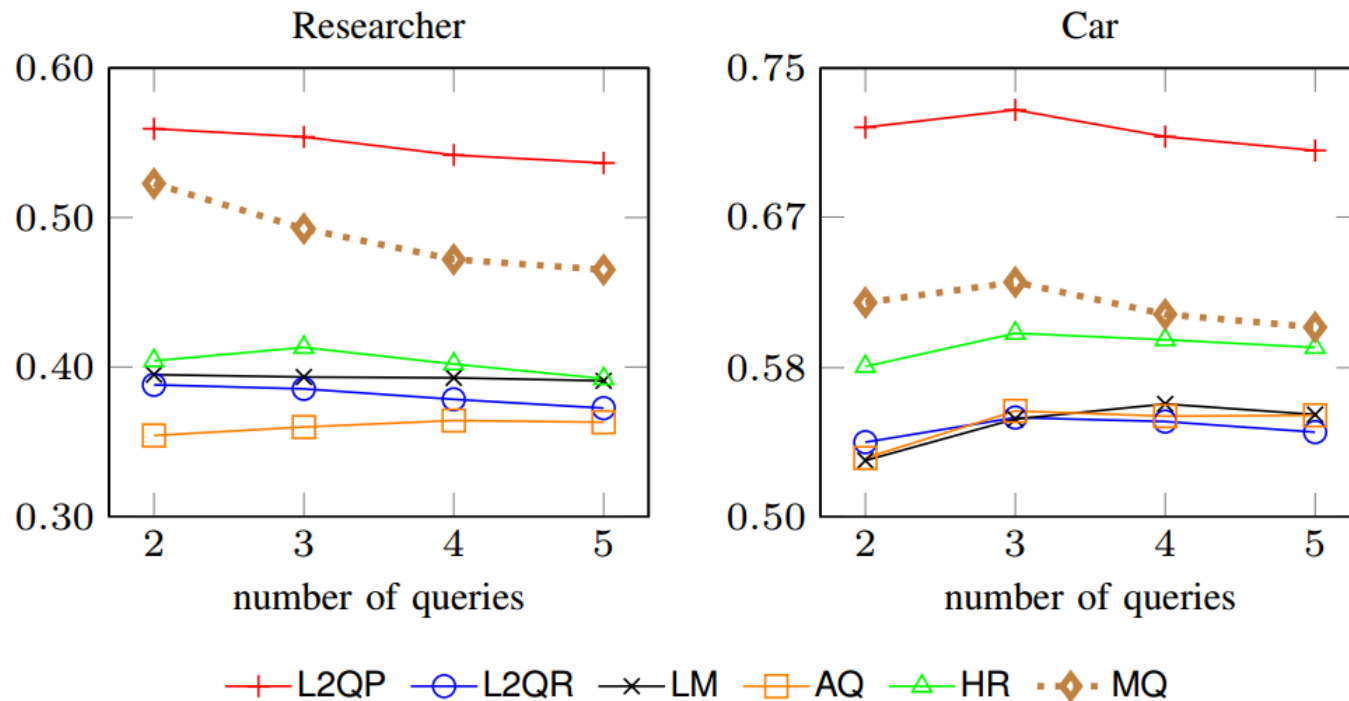
(b) Comparison of recall



- **RND**: select query randomly
- **P/R**: optimize precision/recall without domain and context-awareness
- **P/R+q**: with domain pages, but do not employ templates, and without context
- **P/R+t**: with domain pages and templates, without context
- **L2QP/L2QR**: full approaches optimizing precision/recall

Finding #2(a): Comparing precision with indep. baselines

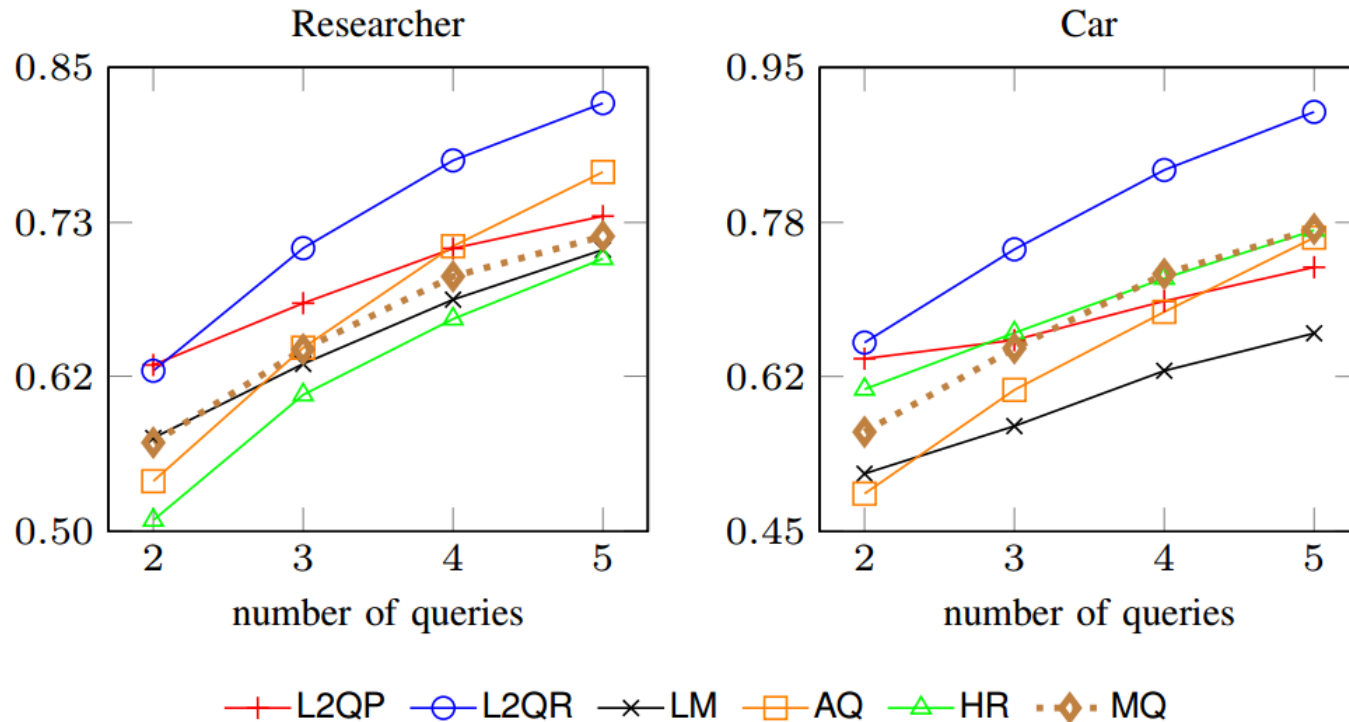
17



- **LM**: language feedback model
- **AQ**: adaptive querying for text databases
- **HR**: harvest rate for hidden structured databases
- **MQ**: manually designed queries

Finding #2(b): Comparing recall with indep. baselines

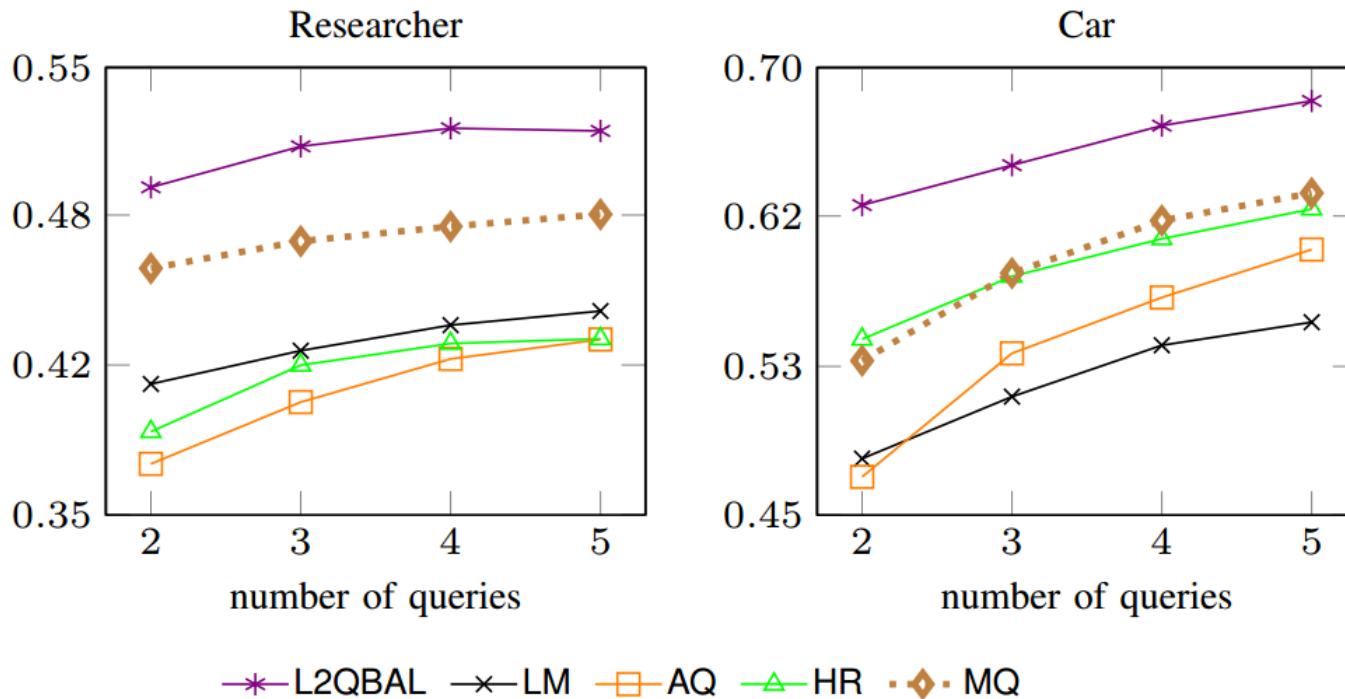
18



- **LM:** language feedback model
- **AQ:** adaptive querying for text databases
- **HR:** harvest rate for hidden structured databases
- **MQ:** manually designed queries

Finding #2(c): Comparing F-score with indep. baselines

19



- **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

In this talk

20

- Problem: L2Q
- Challenges and solution
 - ▣ Domain-awareness
 - ▣ Context-awareness
- Experimental Study
- **Conclusion**

Conclusion

21

- L2Q: a novel paradigm of crawling
- Domain-aware L2Q
 - ▣ Templates to handle vocabulary variations
- Context-aware L2Q
 - ▣ Collective utilities to account for page redundancy