Learning to Query:
Focused Web Page Harvesting for Entity Aspects

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

ICDE 2016 @ Helsinki

[1] Institute for Infocomm Research, Singapore

[2] Advanced Digital Sciences Center, Singapore

[3] University of Illinois at Urbana-Champaign, USA

In this talk

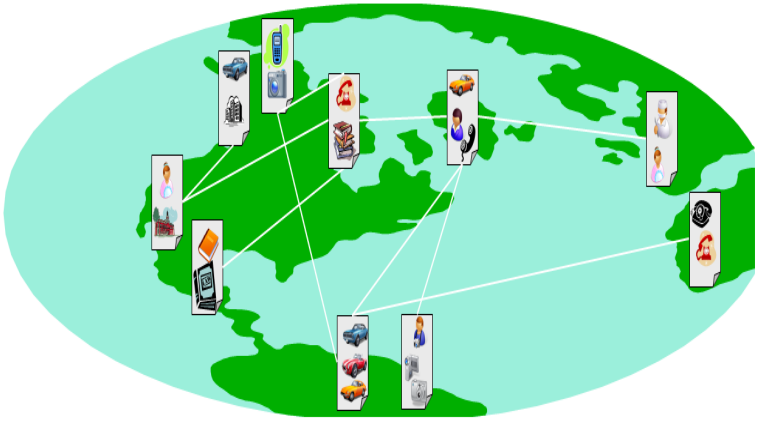
- **Problem: L2Q**
- Challenges and solution
  - Domain-awareness
  - Context-awareness
- Experimental Study
- Conclusion

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

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

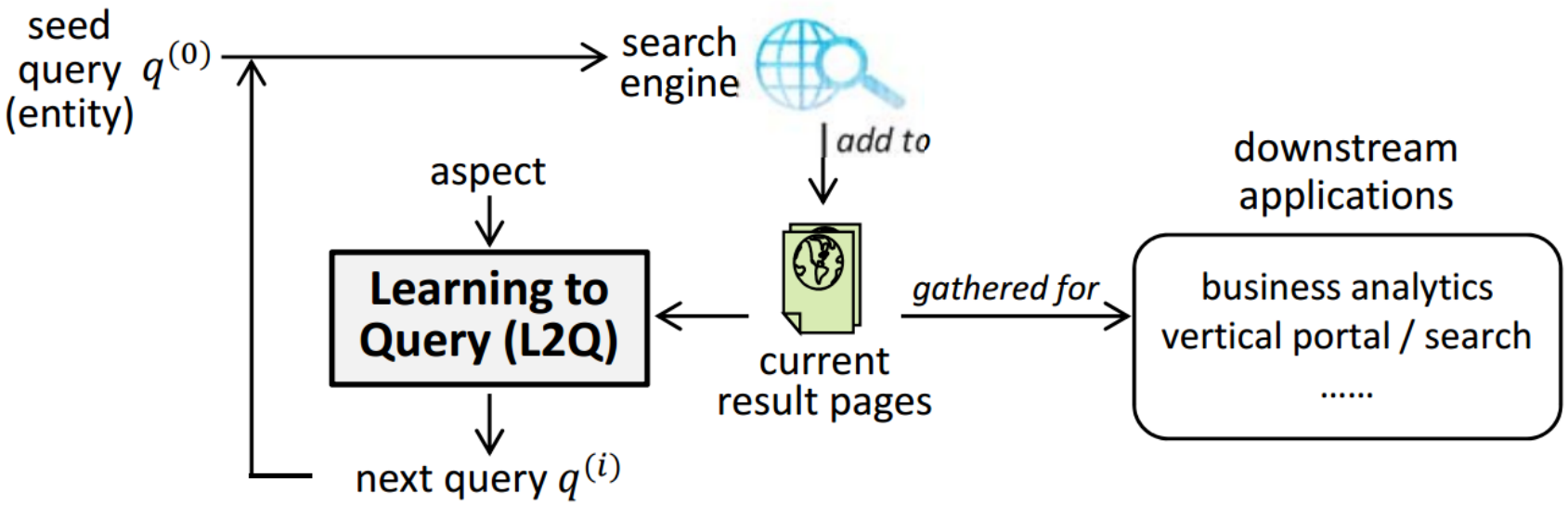
Motivation:
Focused Web Page Harvesting for Entity Aspects



Business Analytics

Vertical portal or search

High level problem: Learning to query (L2Q)



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

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

Subproblem #1: Domain-aware L2Q

Target Entity

Marc Snir

Peer Entities

Philips Yu
Andrew Ng
……

✓ Gradually harvested in earlier iterations

entity pages P_E

✓ Pre-collected / previously harvested pages

domain pages P_D

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

Subproblem #1: Vocabulary variations

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`

topics

hpc
data mining
ai
…

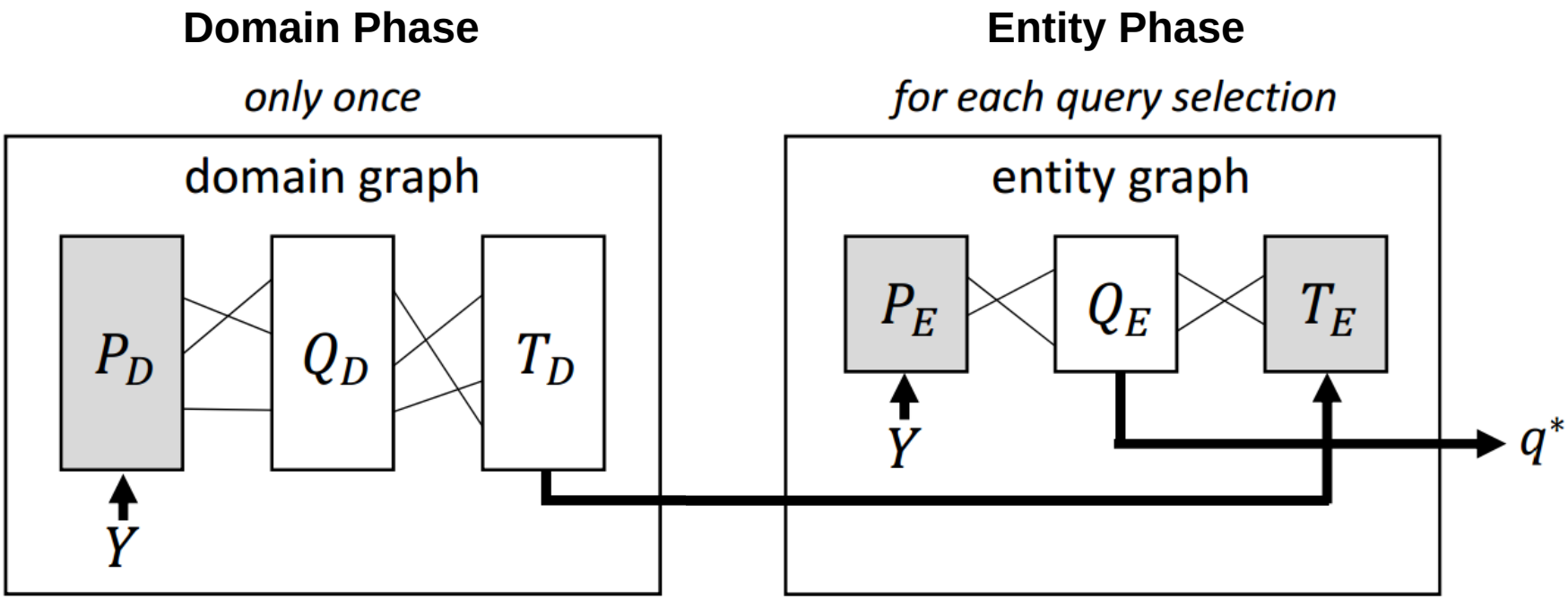
journals

ijhpca
tkde
jmlr
…

<topic> <journal>

(template)

Subproblem #1: Bridging domain and entity phases



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

In this talk

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

Subproblem #2: Context-aware L2Q

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

Marc Snir HPC

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 Murgo 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
inside**hpc**.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/**snirmarc**
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 Parallel

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 ...

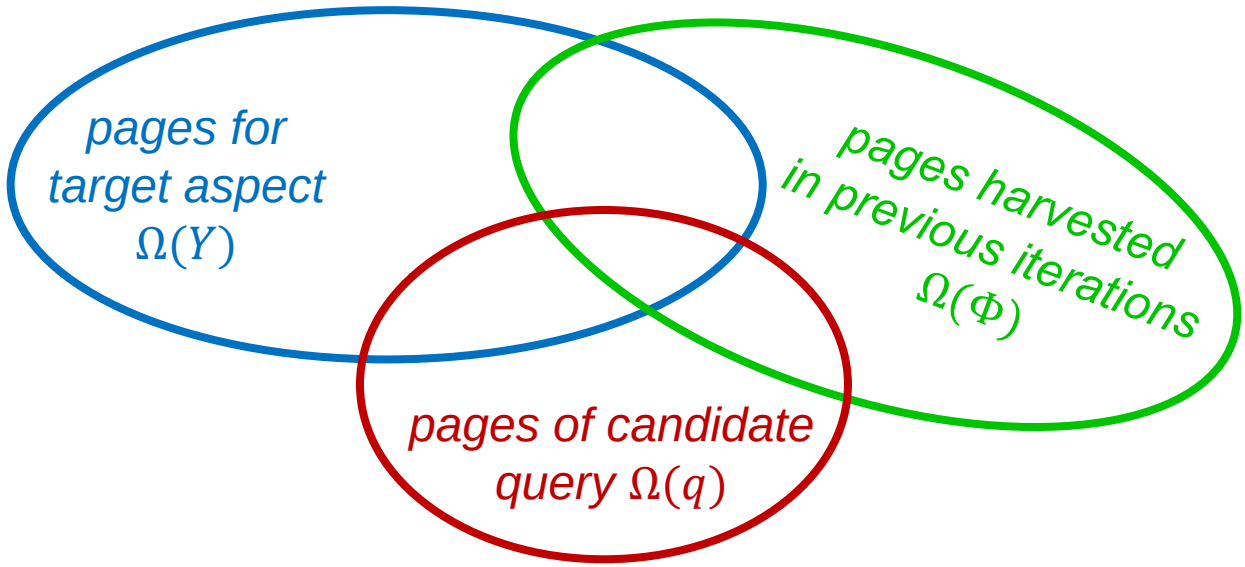
Marc Snir - Department of Computer Science at Illinois
https://cs.illinois.edu/directory/profile/**snir**
Marc Snir. Michael Faiman and Saburo Muroga Professor. (217) 244-6568 ...
Architecture, Compilers, and **Parallel** Computing - **Parallel** Computing ...

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=HaI6LesAAAAJ
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 ...

Subproblem #2: Accounting for redundancy



Collective Utilities

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

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

In this talk

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

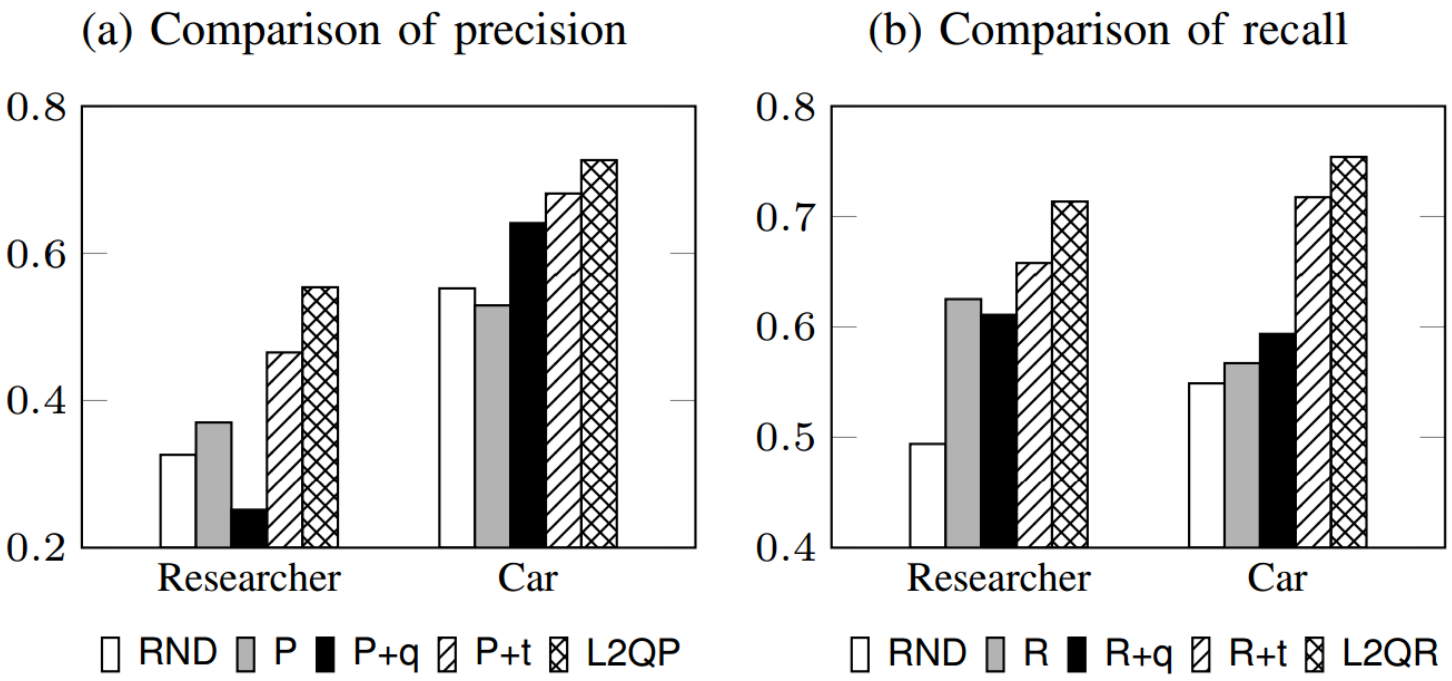
Experiment setup

- **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

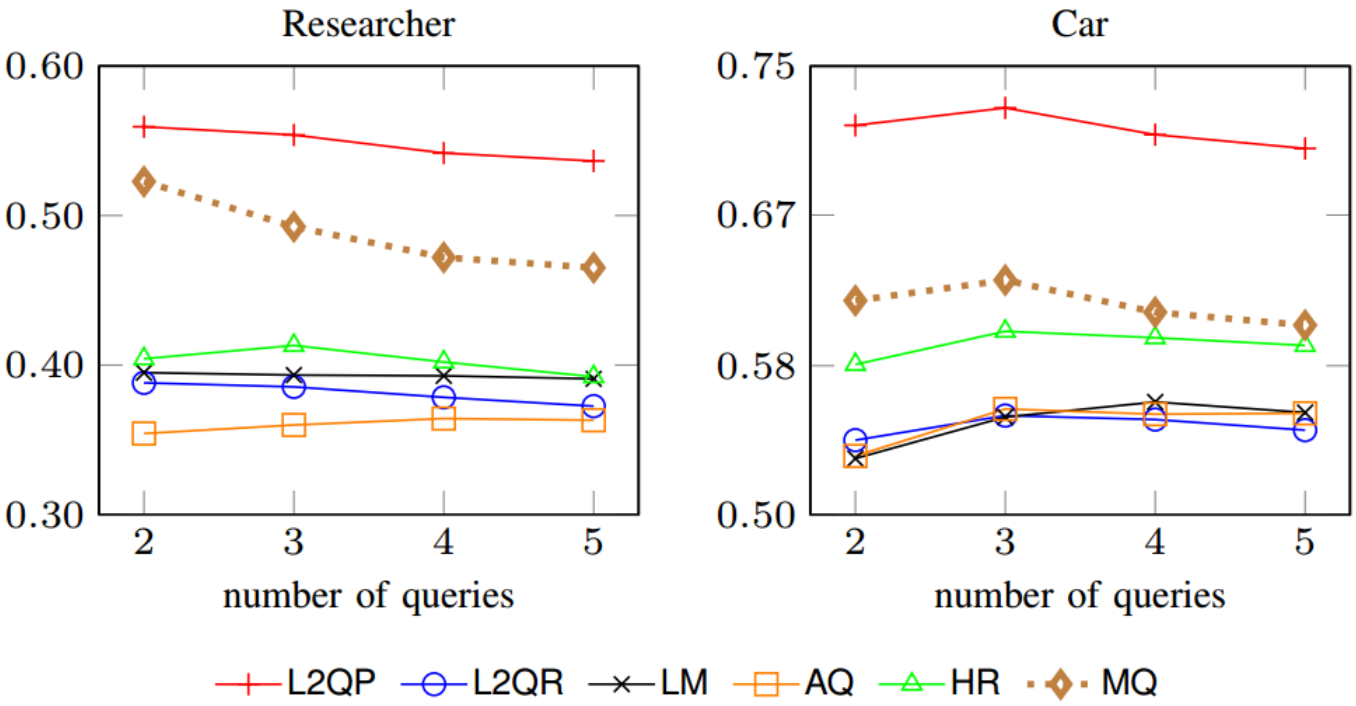
- **Utilities** of two forms: precision & recall
- **Evaluation metrics**
  - Precision, recall
  - Combined F-score
- Metrics reported are **normalized**
  - Against ideal precision/recall
  - Ideal metrics computed by "peeking" at un-retrieved pages

Finding #1:
Effect of domain and context-awareness



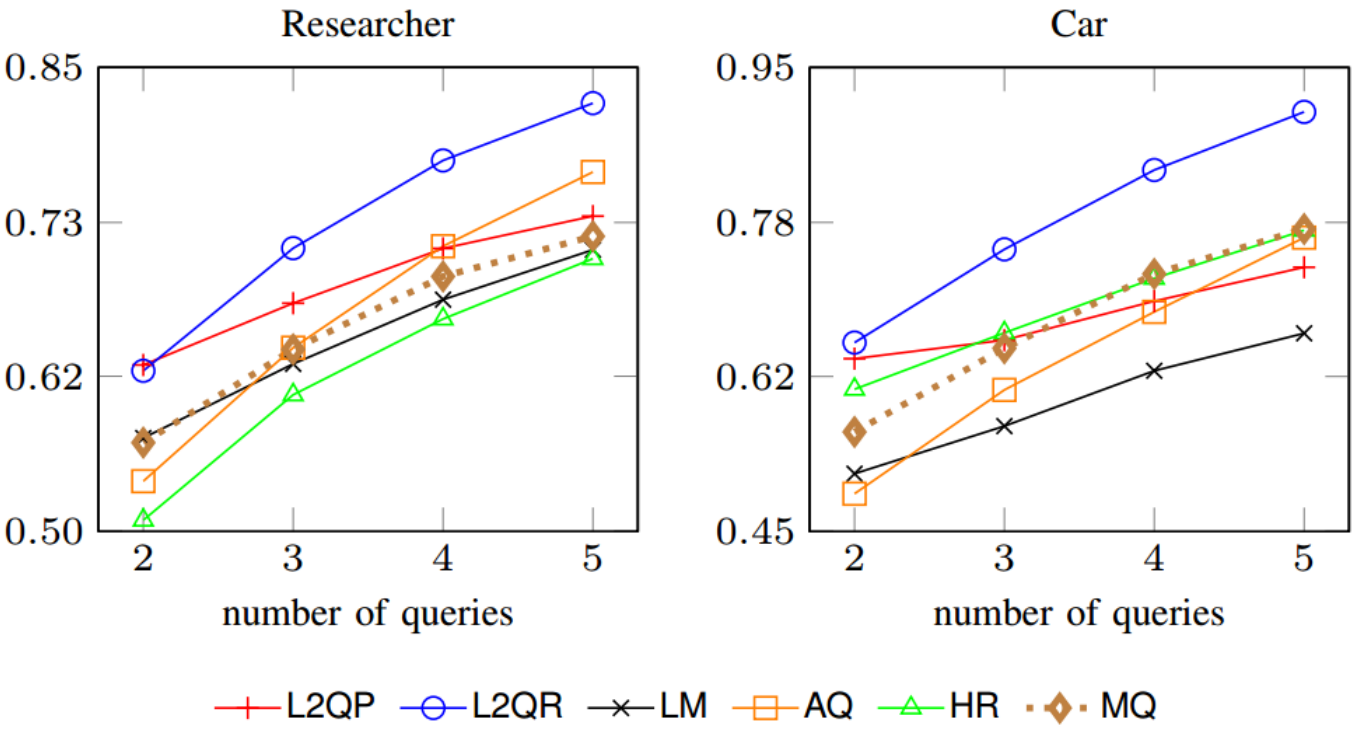
- **RND**: select query randomly
- **P/R**: optimize precision/recall without domain and context-awareness
- **P/R+q**: with domain pages, but do not employ templates, and without context
- **P/R+t**: with domain pages and templates, without context
- **L2QP/L2QR**: full approaches optimizing precision/recall

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



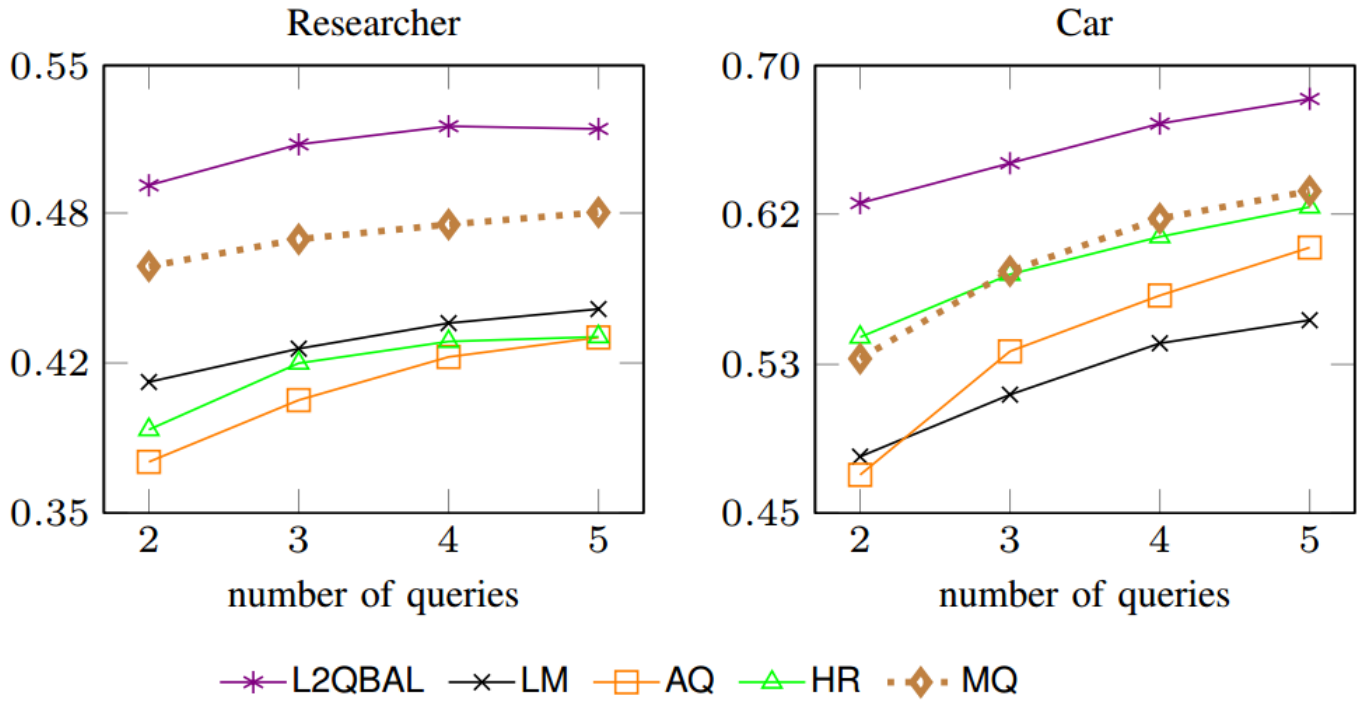
- **LM**: language feedback model
- **AQ**: adaptive querying for text databases
- **HR**: harvest rate for hidden structured databases
- **MQ**: manually designed queries

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



- **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



- **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

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

Conclusion

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