


Query Recommendations for Interactive Database Exploration

Work by
Gloria Chatzopoulou*, UC Riverside
Magdalini Eirinaki, San Jose State Univ
Neoklis Polyzotis, UC Santa Cruz

Presented by
Wuzhou Zhang

Motivation

- Scientific community rely increasingly on *relational databases*
- Users, with *diverse* information needs, employ a web-based client to issue SQL queries for data analysis
- Users may find it hard to write *interesting* queries:
  - They are not SQL experts
  - They are not aware of all parts of the database

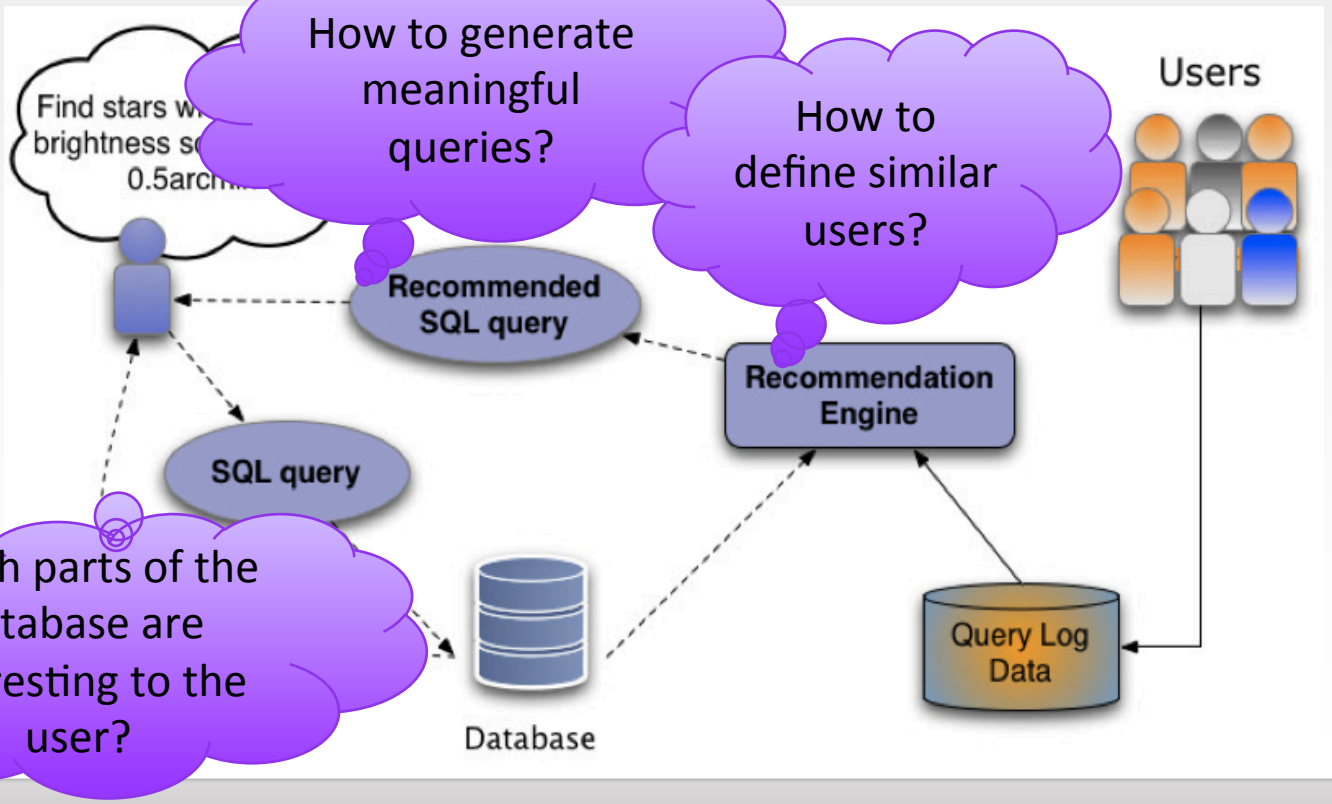
Goal: Assist users in finding useful information

Proposed Solution

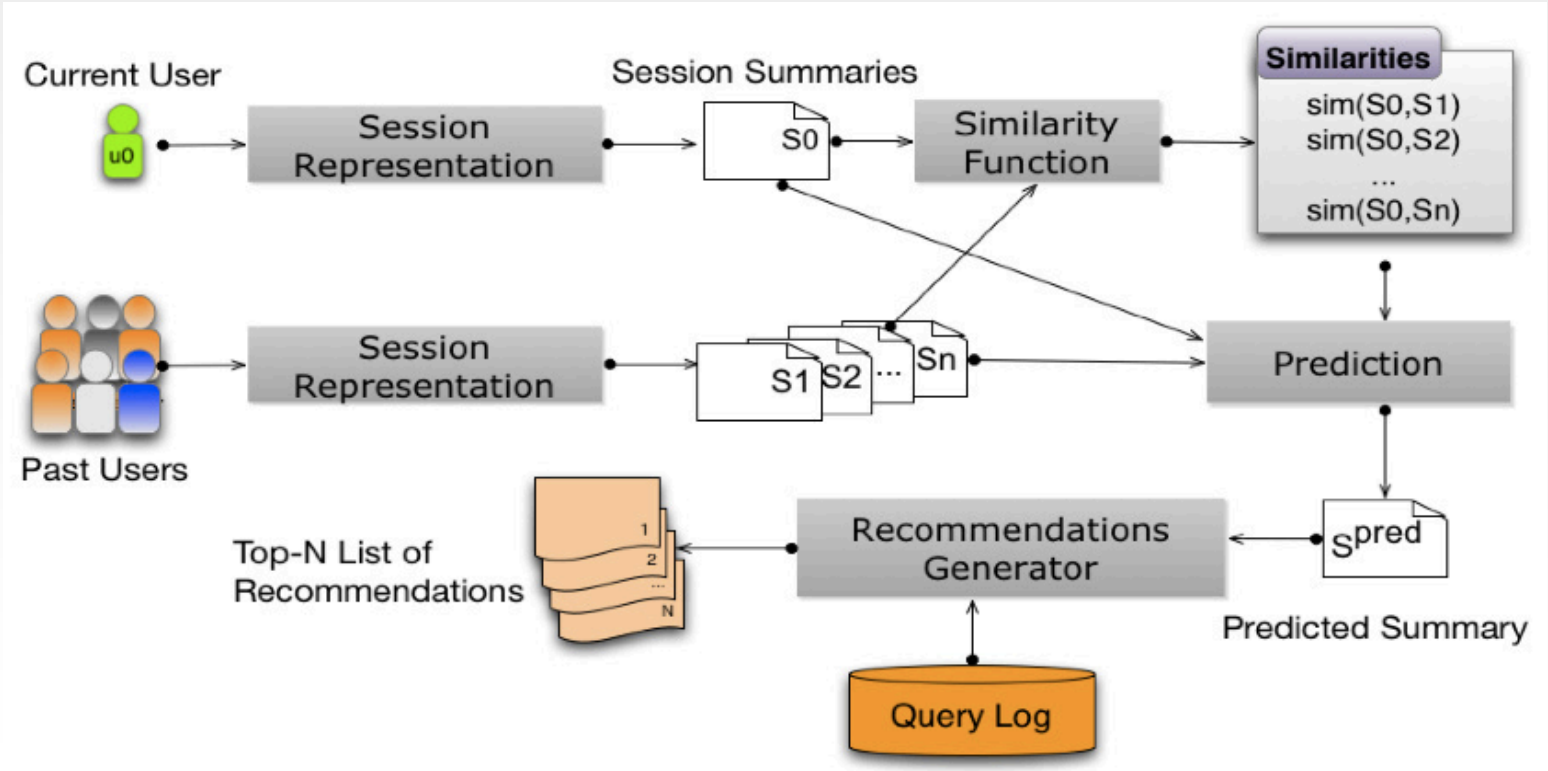
- Recommend queries to users based on the queries of other *similar* users
- Inspiration: *Collaborative Filtering*
- *Example:* Movie Recommendations

If Alice and Bob **both** like movie X and Alice likes movie Y
If Alice and Bob **both** query data X and Alice queries data Y
then
then
Bob is likely to be interested in seeing movie Y
Bob is likely to be interested in querying data Y

System Architecture



Conceptual Framework



Session Representation

R	a	b
	y	3
	s	4
	w	3
	r	2

L	a	c
	y	9
	s	3
	s	5
	t	8

q1: $R \bowtie_{R.a=L.a} L$

q2: $\sigma_{R.b=4}(F \bowtie_{R.a=L.a} L)$

Binary Weighting Scheme	Result Weighting Scheme
q1 = <1,1,0,0,1,1,1,0>	q1 = <0.33,0.33,0,0,0.33,0.33,0.33,0>
q2 = <0,1,0,0,0,1,1,0>	q2 = <0,0.50,0,0,0,0.50,0.50,0>
s0 = <1,2,0,0,1,2,2,0>	s0 = <0.33,0.83,0,0,0.33,0.83,0.83,0>

Similarity Function

- Vector-space similarity functions can be used
  - Cosine Similarity

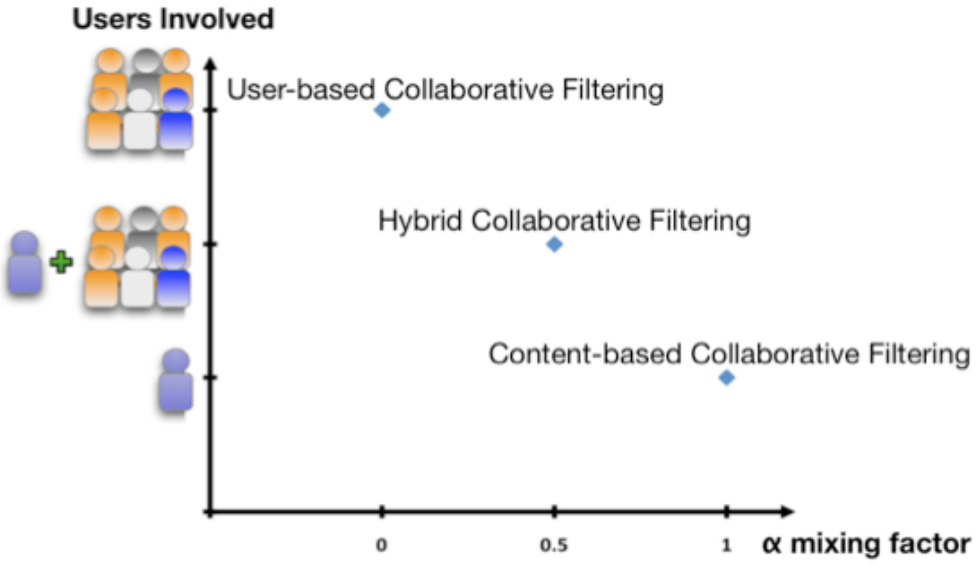
$$sim(S_i, S_j) = \frac{S_i S_j}{\|S_i\|_2 \|S_j\|_2}$$

- High similarity means that users are most likely interested in the same parts of the database

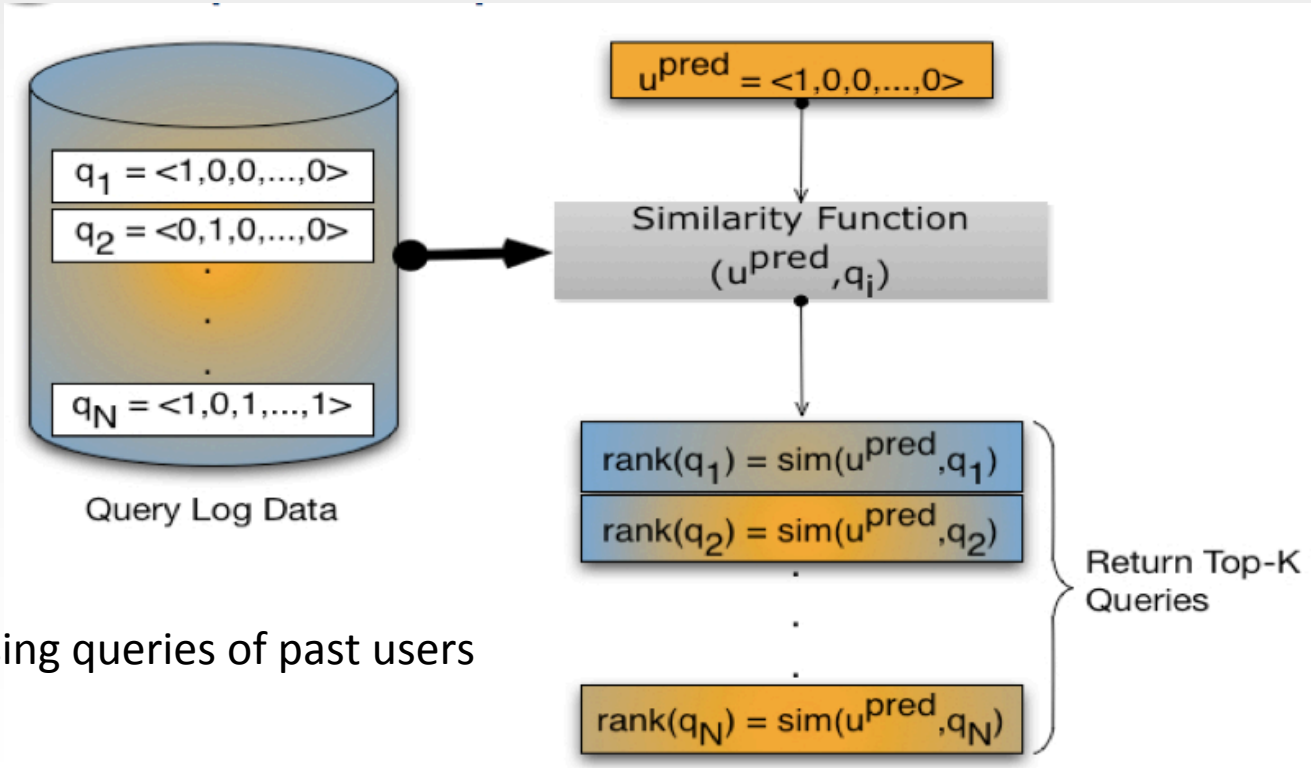
Predicted Summary

$$S_0^{\text{pred}} = \alpha * S_0 + (1 - \alpha) * \frac{\sum_{1 \le i \le h} sim(S_0, S_i) \cdot S_i}{\sum_{1 \le i \le h} sim(S_0, S_i)}$$

α: the mixing factor



Generating Recommendations



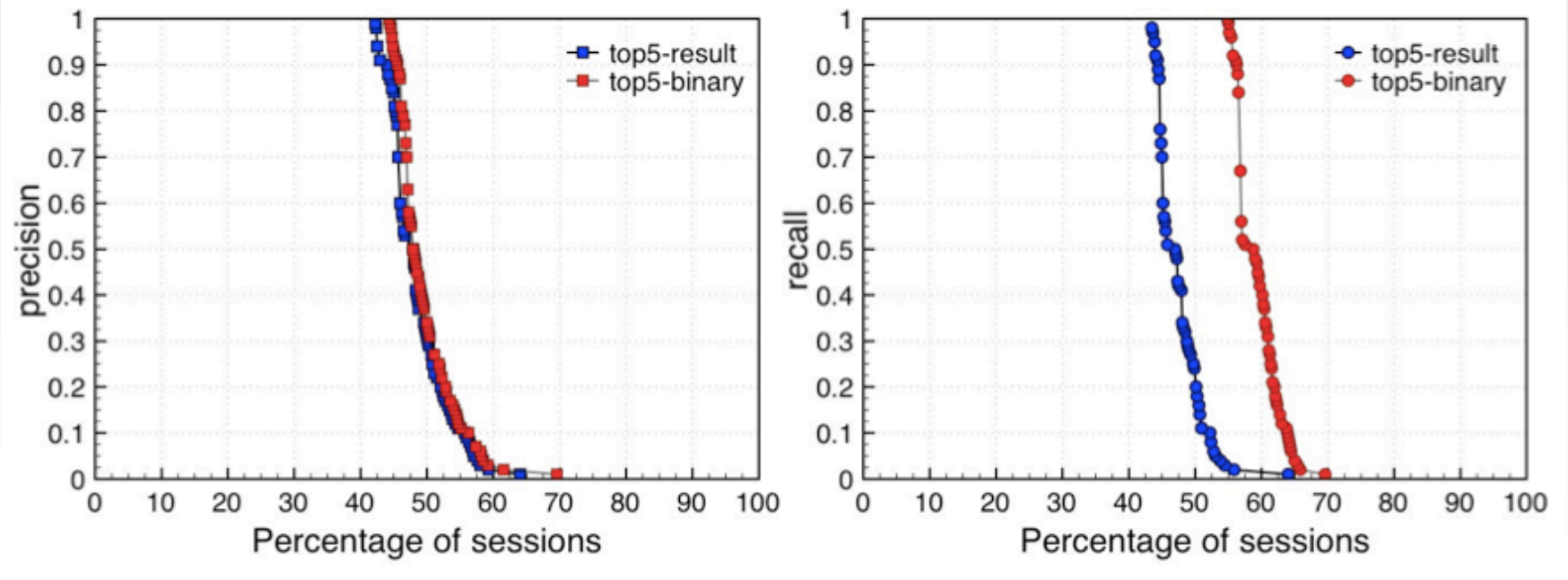
Using queries of past users

Experimental Setup

- SkyServer Dataset
- Evaluation Metrics: Precision and Recall
  - **High precision:** most witnesses of the recommended query are witnesses in the actual query.
  - **High Recall:** most witnesses of the actual query are witnesses in the recommended query.

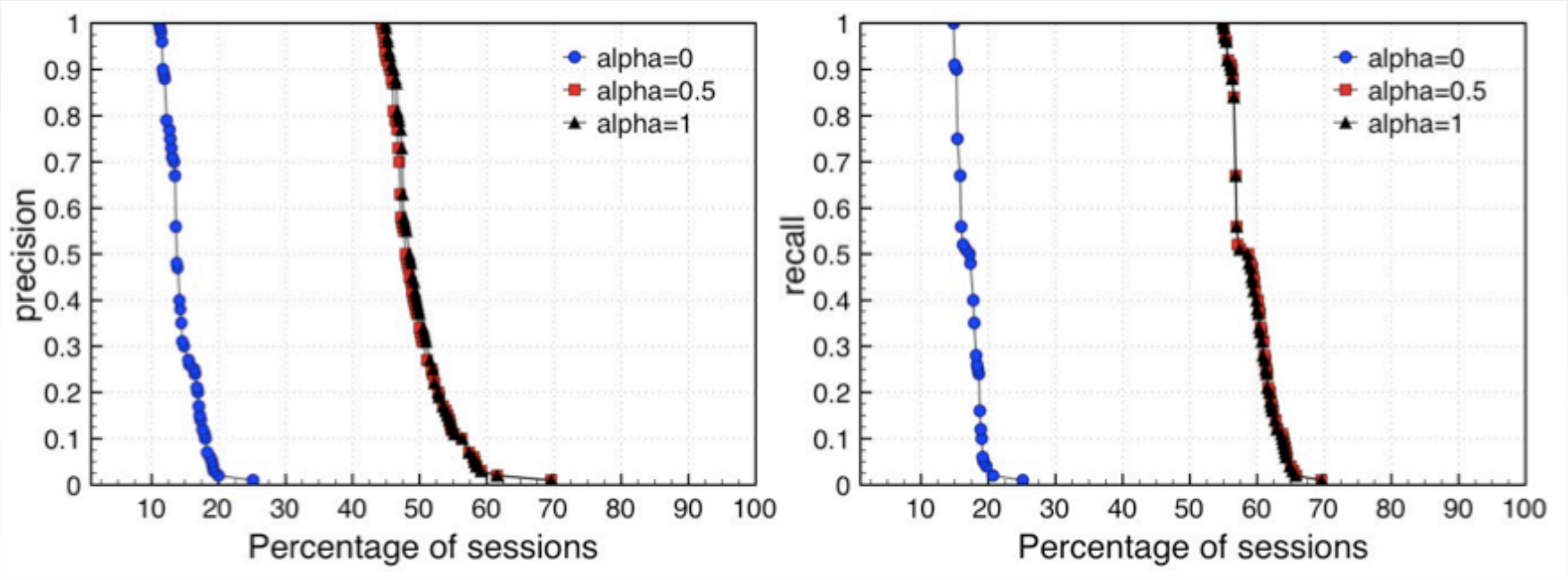
Database size	2.6TB
#Sessions	720
#Queries	6713
#Distinct queries	4037
#Distinct witnesses	13,602,430
Avg. number of queries per session	9.3
Min. number of queries per session	3

Binary vs Result Weighting Schemes



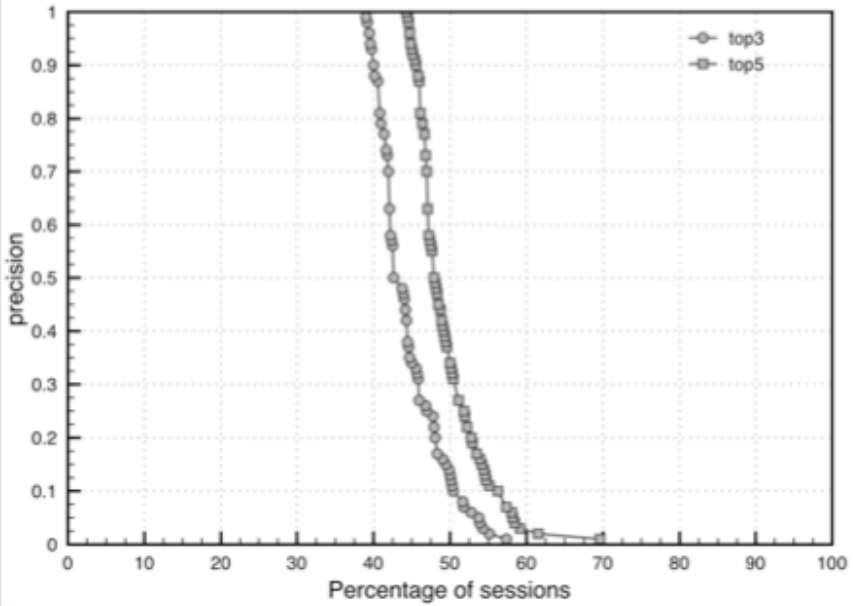
Binary outperforms Result Weighting Scheme

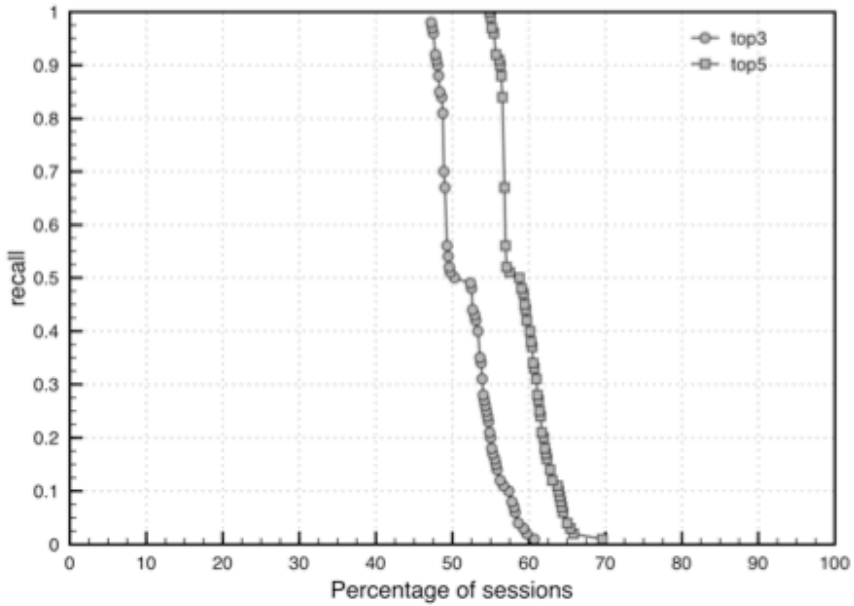
Effect of mixing factor α



Hybrid Collaborative Filtering yields better results

Top-3 vs Top-5 Binary Weights





The bigger recommendation set the higher accuracy

Discussion

- Performance improvement
- Though we can return the same tuple, queries might be different
- Query structure instead of tuples retrieved
- Correlation between sequence of queries (causality, incremental)
- Extension: automatically import other relations
- Relation to our project?

R	a	b
	y	3
	s	4
	w	3
	r	2

Recent work: fragment-based (attribute ref, tables ref, join and selection predicates)

Thanks!

- Reference
  - "Query Recommendations for Interactive Database Exploration." Chatzopoulou, Eirinaki, and Polyzotis. *SSDBM* 2009.
  - http://www.cs.washington.edu/education/courses/cse599c/10sp/lecture7/querie.pdf
  - http://ssdbm09.cs.uno.edu/papers/3b.pdf
  - http://www.engr.sjsu.edu/meirinaki/papers/CE+11-IEEEDebul.pdf