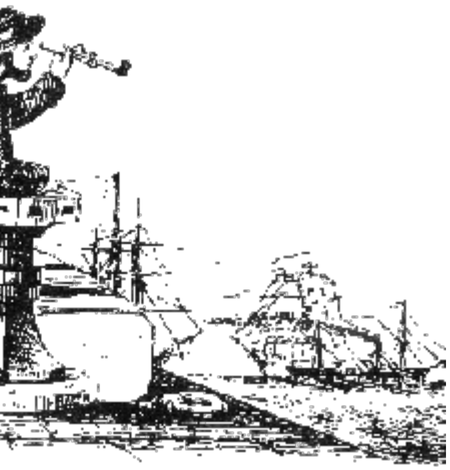


A Task-based Framework for User Behavior Modeling and Search Personalization*

Hongning Wang
Department of Computer Science
University of Virginia
hw5x@virginia.edu

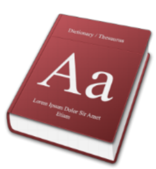
*work is done when visiting Microsof

Search Logs Provide Rich Context for Understanding Users' Search Tasks

85% information maintenance tasks a
52% information ga
tasks will span multi
queries [Agichtein et al. SI

5/29/2012	
5/29/2012 14:06:04	coney island Cincinnati
5/29/2012 14:11:49	sas
5/29/2012 14:12:01	sas shoes
5/30/2012	
5/30/2012 12:12:04	exit #72 and 275 lodging
5/30/2012 12:25:19	6pm.com
5/30/2012 12:49:21	coupon for 6pm
5/31/2012	
5/31/2012 19:40:38	motel 6 locations
5/31/2012 19:45:04	Cincinnati hotels near coney island

a good chance to customize the results!



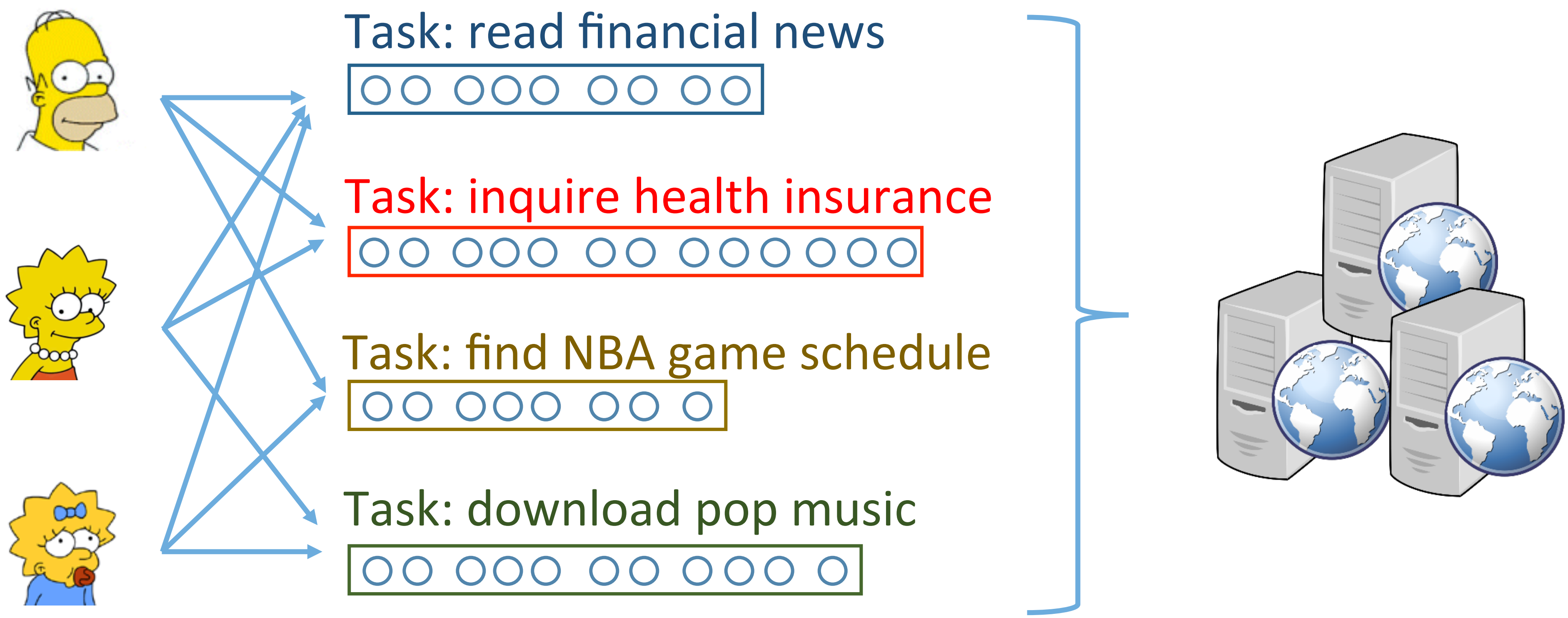
***Task**: an atomic information need that may result in one or more queries* [Jones et al. CIKM'08]

Query-based Analysis: An Isolated View

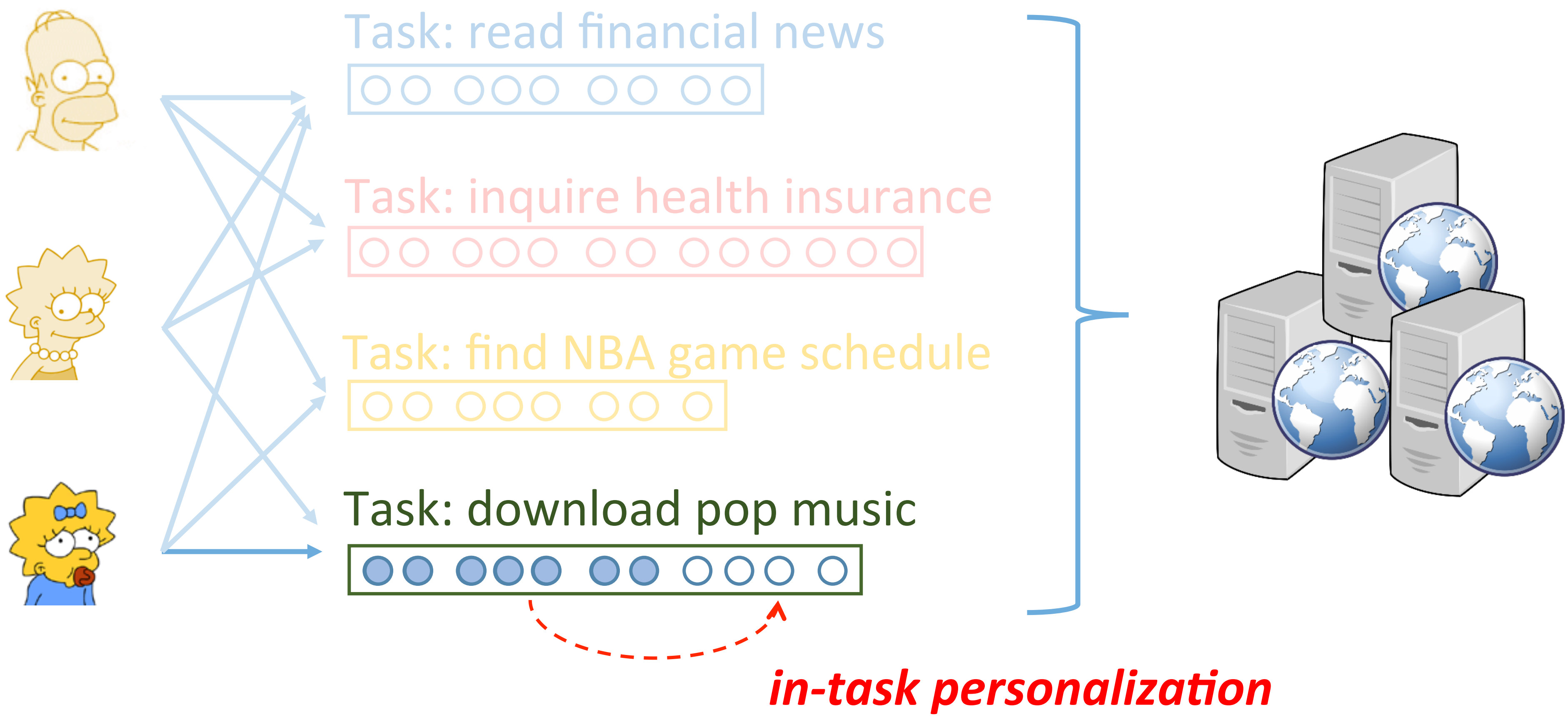
Search log mining approaches:

- **Query categories** *[Jansen et al. IPM 2000]*
- **Temporal query dynamics** *[Kulkarni et al.*
- ***Survey: [Silvestri 2010]***

Task-based Analysis: A Comprehensive View



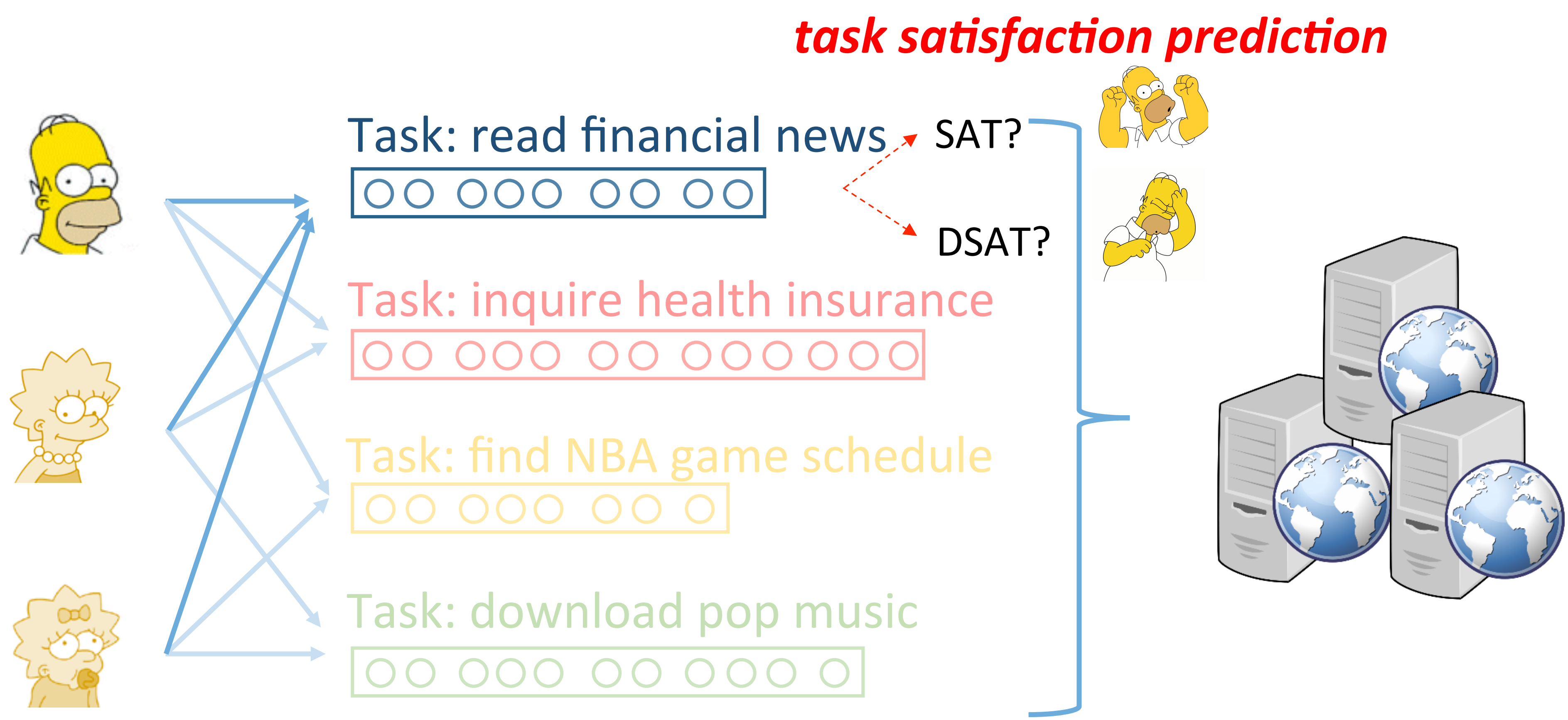
Task-based Analysis: A Comprehensive View



Task-based Analysis: A Comprehensive View

Cross-user: collaborative ranking

Task-based Analysis: A Comprehensive View



Research Questions

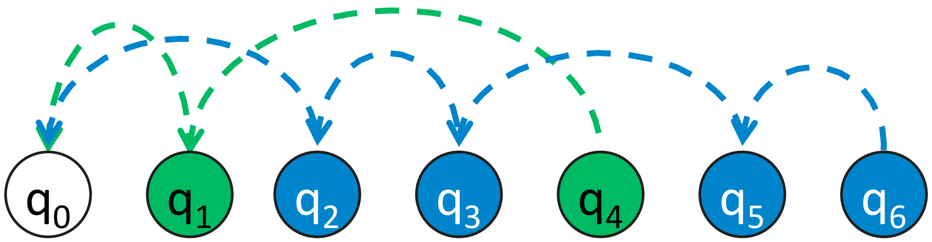
1. How to effectively extract search tasks from search logs?
2. How to represent and organize search tasks?
3. How to model users' in-task search behaviors?
4. How to optimize search services based on the identified search tasks?
5. How to interactively assist users to perform search tasks?
6. ....

Research Questions

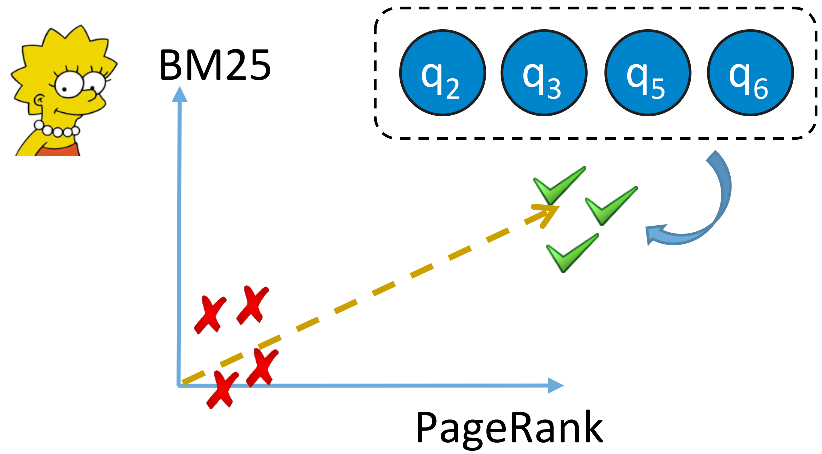
1. How to effectively extract search tasks from search logs?
2. How to represent and organize search tasks?
3. How to model users' in-task search behaviors?
4. How to optimize search services based on the identified search tasks?
5. How to interactively assist users to perform search tasks?
6. ....

A Task-based Framework

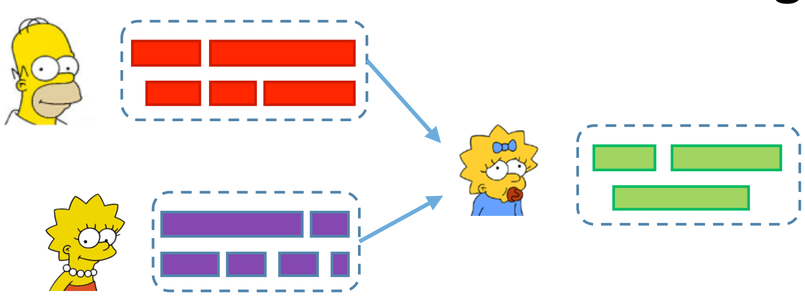
- Long-term task extraction



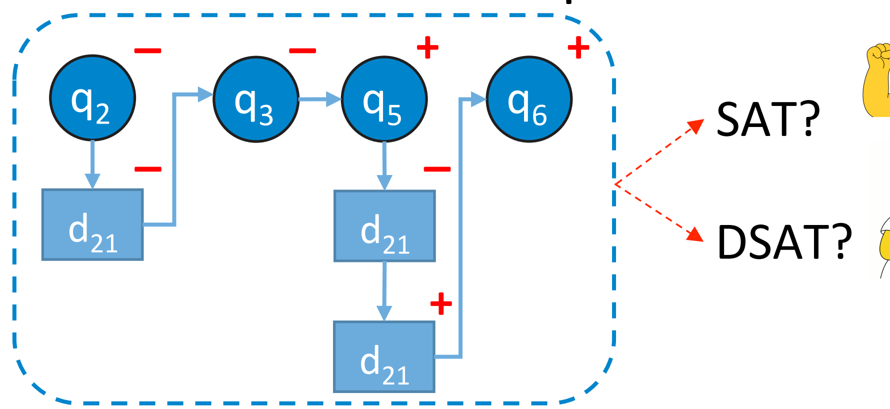
- In-task personalization



- Cross-user collaborative ranking

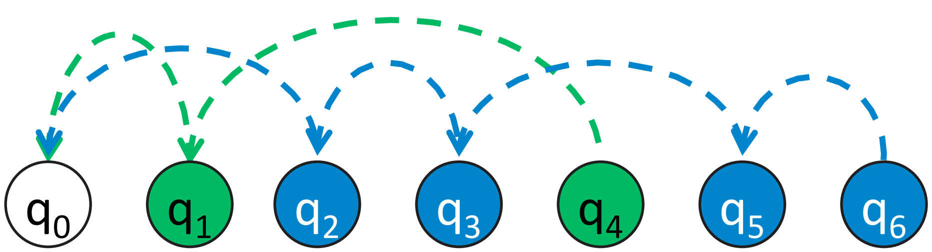


- Search-task satisfaction prediction

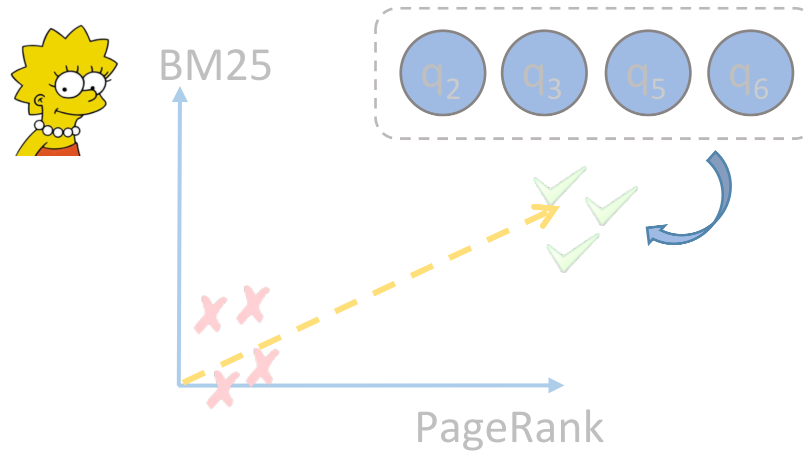


A Task-based Framework

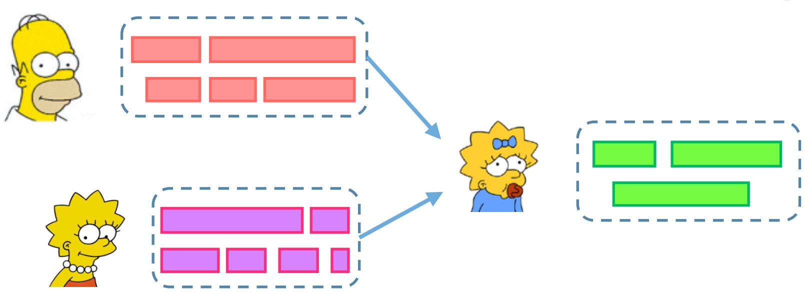
- Long-term task extraction



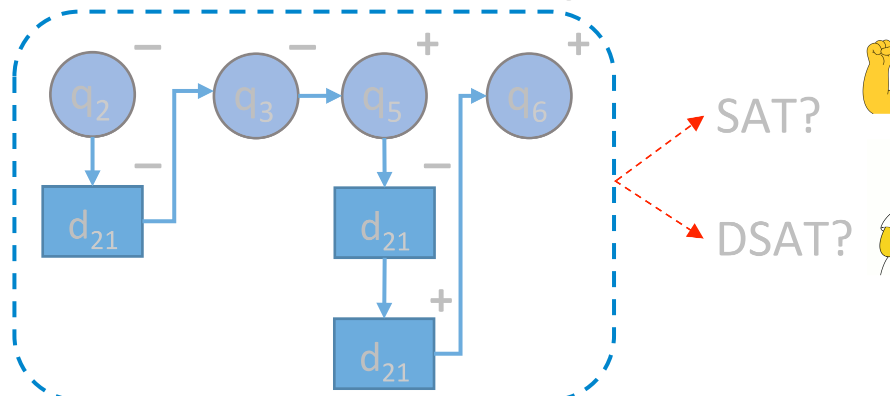
- In-task personalization



- Cross-user collaborative ranking



- Search-task satisfaction prediction



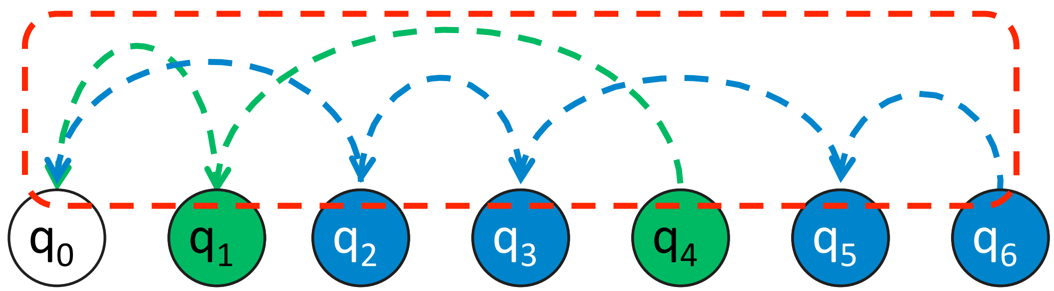
How to extract tasks? [WWW'13]

New perspective: best-link as task structure

ructure is lost

q_1 = "coney island Cincinnati" q_2 = "sas"
q_3 = "sas shoes" q_4 = "exit #72 and 275 lodging"
q_5 = "6pm.com" q_6 = "coupon for 6pm"

Latent!



$\mathcal{T}_1 = \{q_1, q_4\}$ $\mathcal{T}_2 = \{q_2, q_3, q_5, q_6\}$

red learning solution:

$$= \arg\max_{(y,h)\in\mathcal{Y}\times\mathcal{H}} w^{\mathsf{T}}\Phi(\mathcal{Q}, y, h)$$

Query features (9)
URL features (14)
Session features (3)

Existing solutions:
Binary classificatio
[Jones et al. CIKM'
Lucchese et al. WS
Kotov et al. SIGIR'
Liao et al. WWW'1

Experimental Results

Query Log Dataset

- Bing search log: May 27, 2012 – May 31, 2012
- Human annotation
  - 3 editors (inter-agreement: 0.68, 0.73, 0.77)

# User	# Session	# Query
7628	37547	114723
Query/User	Session/User	Query/Session
15.1 ± 17.2	4.9 ± 3.5	3.1 ± 1.2

7.2 ± 10.1	6.6 ± 8.2
Session/Task*	Task duration (mins)*
2.8 ± 2.6	491.1 ± 933.5

*count only in multi-query tasks

Example of Identified Search Tasks

- User
  - www.dailyastorian.info@5/28/2012 1:8:34 PM
    - www.dailyastorian.com@5/30/2012 11:45:15 AM
      - scott somers reedsport@5/30/2012 1:6:57 PM
        - scott somers reedsport@5/30/2012 1:8:0 PM
      - www.dailyastorian.com@5/31/2012 11:8:1 AM
  - plantar fasciitis symptoms@5/29/2012 2:35:53 PM
    - plantar fasciitis pictures@5/29/2012 2:36:22 PM
      - toe pain@5/29/2012 2:39:31 PM
        - toe pain@5/29/2012 2:40:55 PM
          - toe pain symptoms@5/29/2012 2:43:21 PM
            - foot pain@5/29/2012 2:43:57 PM
              - foot pain@5/29/2012 2:45:47 PM
            - hammer toe@5/29/2012 2:45:55 PM
              - hammer toe@5/29/2012 2:47:57 PM
    - chagas disease@5/31/2012 4:57:3 PM
  - clatsop community college columbia address@5/29/2012 2:31:47 PM
    - astoria safeway address@5/29/2012 4:37:11 PM

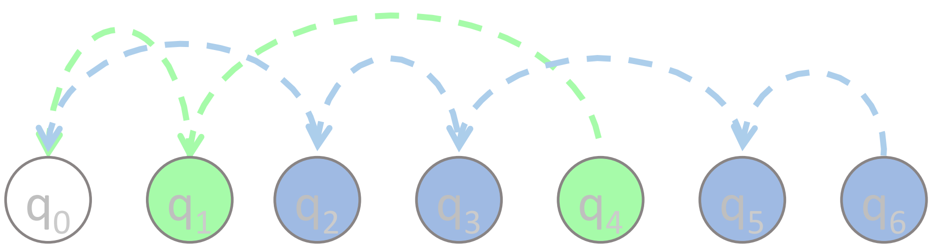
Search Task Extraction Performance

		p_{pair}	r_{pair}	$f1_{\text{CEAF}}$	NMI
o structures, ifferent post-processing	Q_wcc	0.8653	**0.9833***	0.4826	0.4058
	Q_htc	0.9213	0.8607	0.5461	0.5636
	AdaptClu	0.9059	0.9046	0.5583	0.5466
ent structural assumption	cluster-svm	0.9232	0.7908	0.5363	0.5602
	bestlink SVM	**0.9330***	0.9273	**0.5895***	**0.6046***

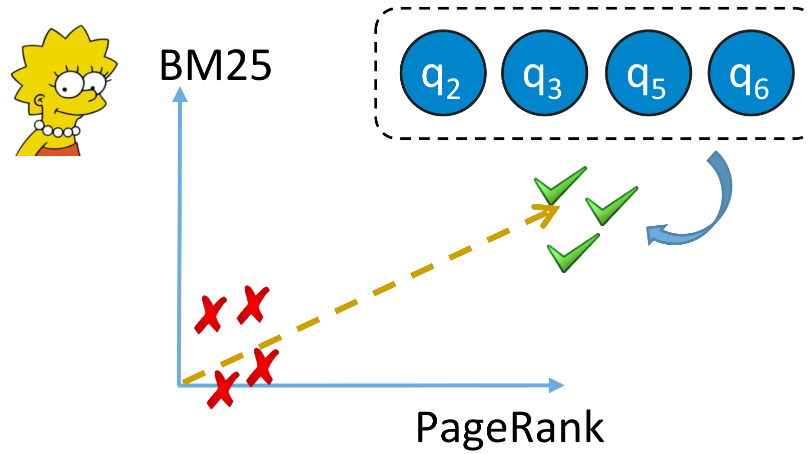
* indicates p-$value < 0.01$

A Task-based Framework

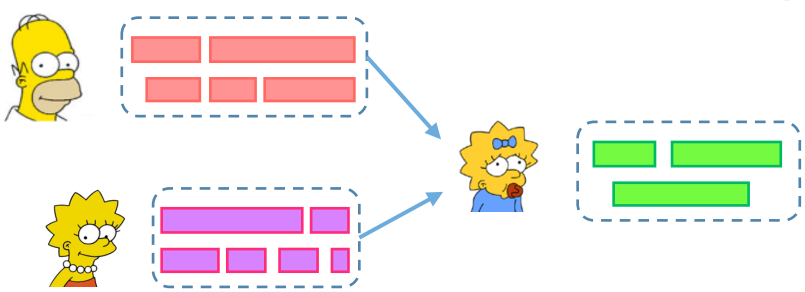
- Long-term task extraction



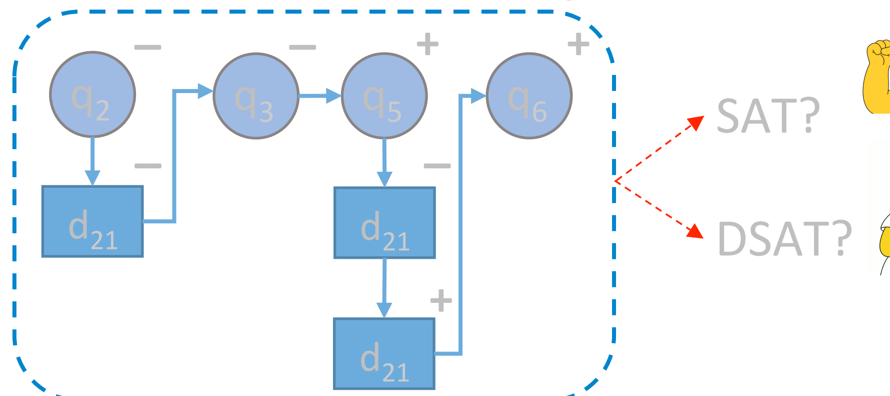
- In-task personalization



- Cross-user collaborative ranking



- Search-task satisfaction prediction

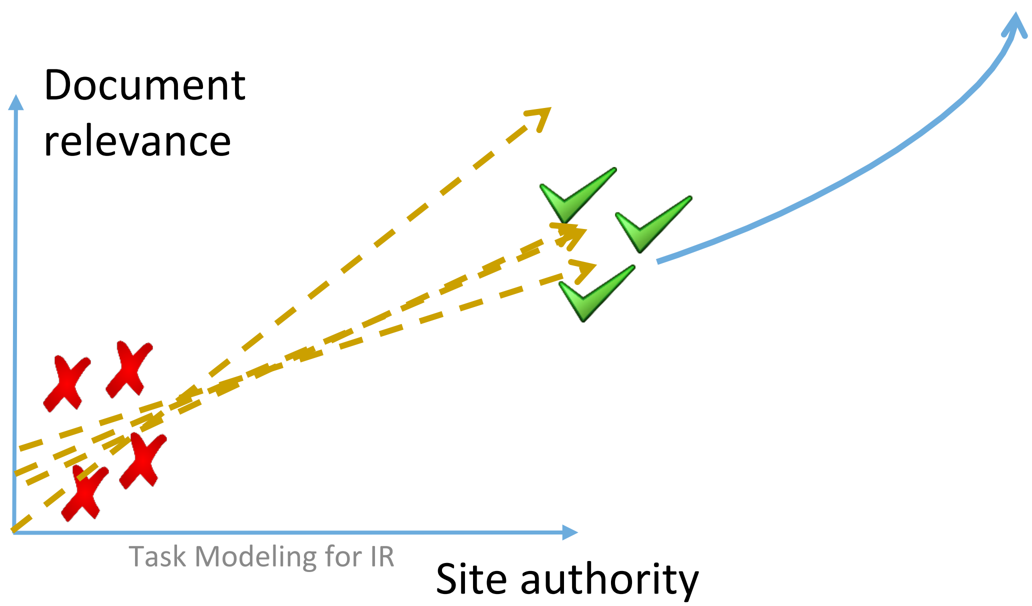


n-task Search Personalization

Search log:

Timestamp	Query	Clicks

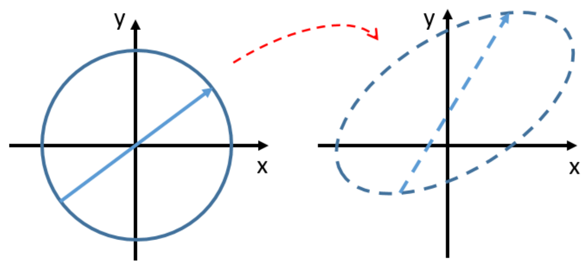
sting solutions:
- Extracting user-centric features [Teevan et al. SIGIR'05]
- Memorizing user clicks [White and Drucker WWW'07]



New perspective: personalized ranking model adaptation [SIGIR'13]

Adjust the generic ranking model's parameters with respect to each individual user's in-task ranking preferences

$$f(x) = w^{\mathsf{T}} x$$



mestamp	Query	Clicks
2012 14:06:04	coney island Cincinnati	✓ ✗ ✗ ✗
2012 12:12:04	drive direction to coney island	✗ ✗ ✓ ✗
2012 19:40:38	motel 6 locations	✗ ✓ ✗ ✗
2012 19:45:04	Cincinnati hotels near coney island	? ? ? ?

$O(V^2)$

$$f^u(x) = (A^u \tilde{w}^s)^{\mathsf{T}} x$$

$O(V)$

$$\mathbf{A}^u = \begin{pmatrix} a^u_{g(1)} & 0 & \cdots & b^u_{g(1)} \\ 0 & a^u_{g(2)} & \cdots & b^u_{g(2)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & a^u_{g(V)} & b^u_{g(V)} \end{pmatrix}$$

Linear Regression Based Model Adaptation

- Adapting global ranking model for each individual user

$$\min_{A^u} \quad L_{\text{adapt}}(A^u) = L(Q^u; f^u) + \lambda R(A^u)$$

$$\text{where} \quad f^u(x) = (A^u \tilde{w}^s)^{\mathsf{T}} x \text{ and } \tilde{w}^s = (w^s, 1)$$

Loss function from any linear learning-to-rank algorithm, e.g., RankNet, LambdaRank, RankSVM

Complexity of adaptation

n-task Personalization Evaluation

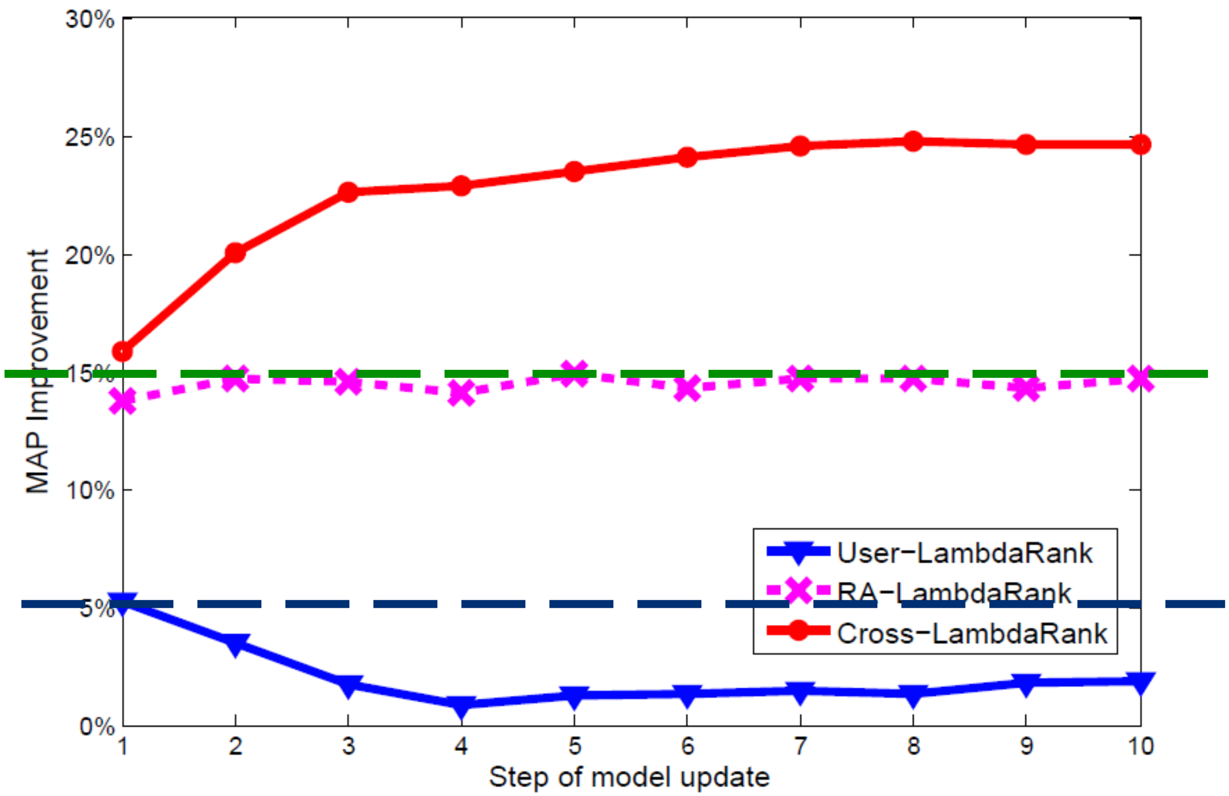
Dataset

- Bing query log: May 27, 2012 – May 31, 2012
- 1830 ranking features
    - BM25, PageRank, tf*idf and etc.

	# Users	# Queries	# Documents
Annotation Set	-	49,782	2,320,711
User Set	34,827	187,484	1,744,969

Adaptation Efficiency

- Against global model



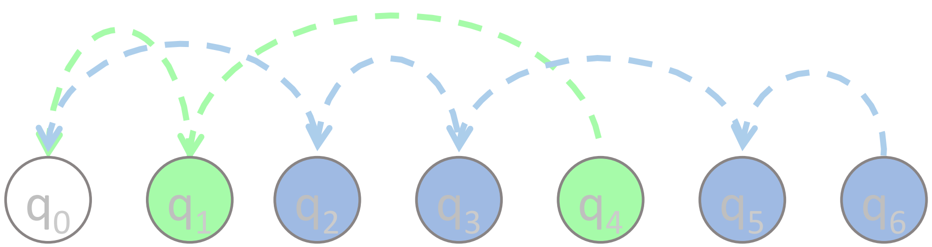
Adapting from global m
and sharing transforma

Cannot deal with sparsity in limited data

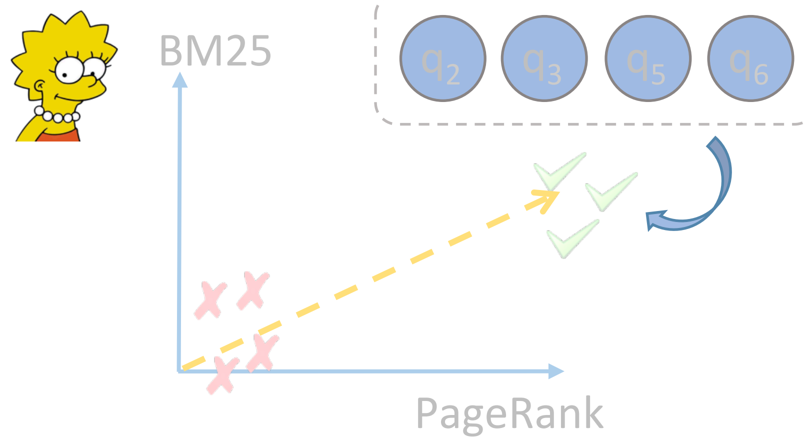
Cannot deal with variance in user clicks

A Task-based Framework

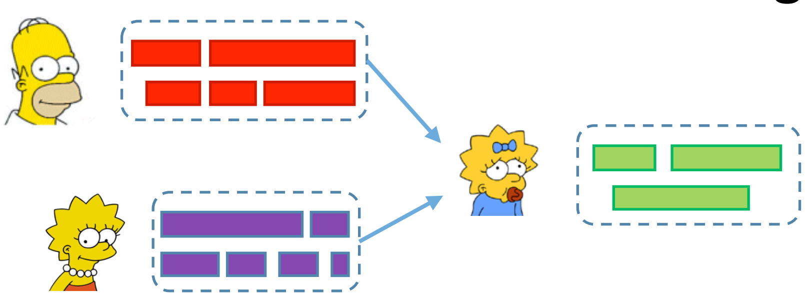
- Long-term task extraction



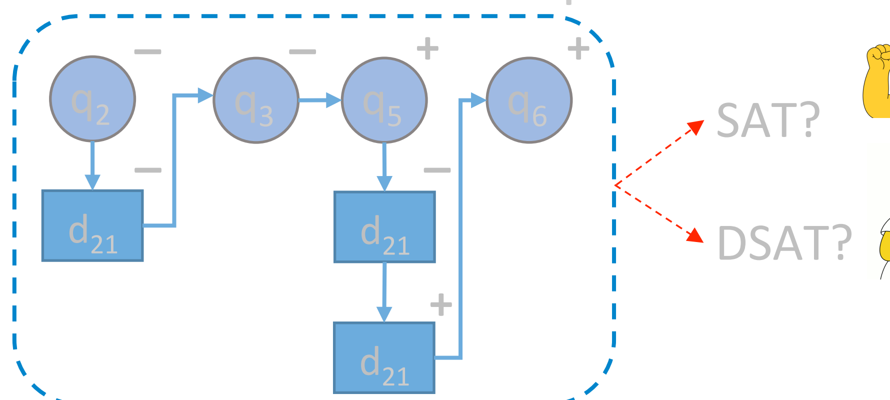
- In-task Personalization



- Cross-user collaborative ranking

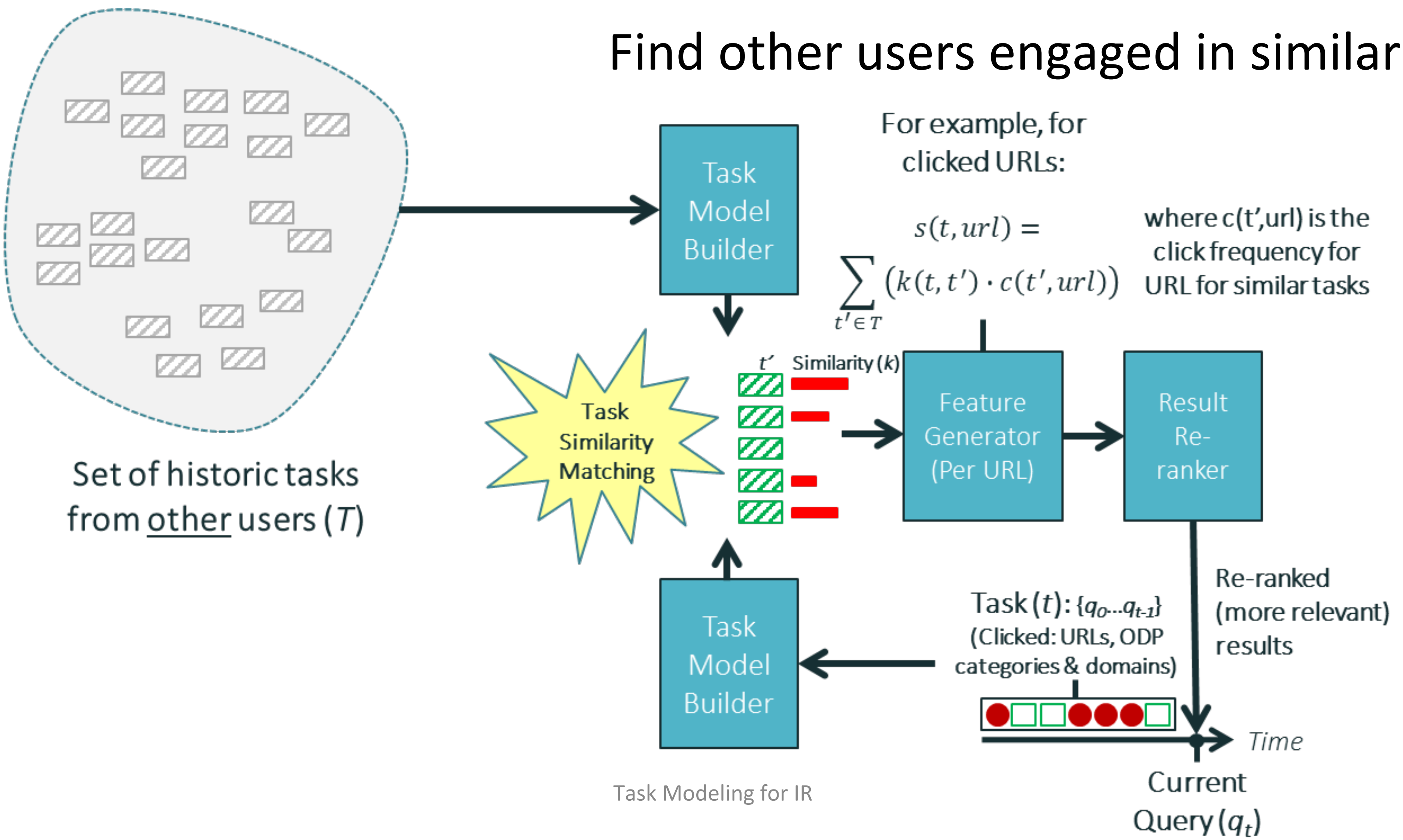


- Search-task satisfaction prediction



Task-Based Groupization [WWW2013b]

Find other users engaged in similar task



Task Match vs. Query Match

QG: same query, all users
QI: same query, same user
QGI: QG + QI

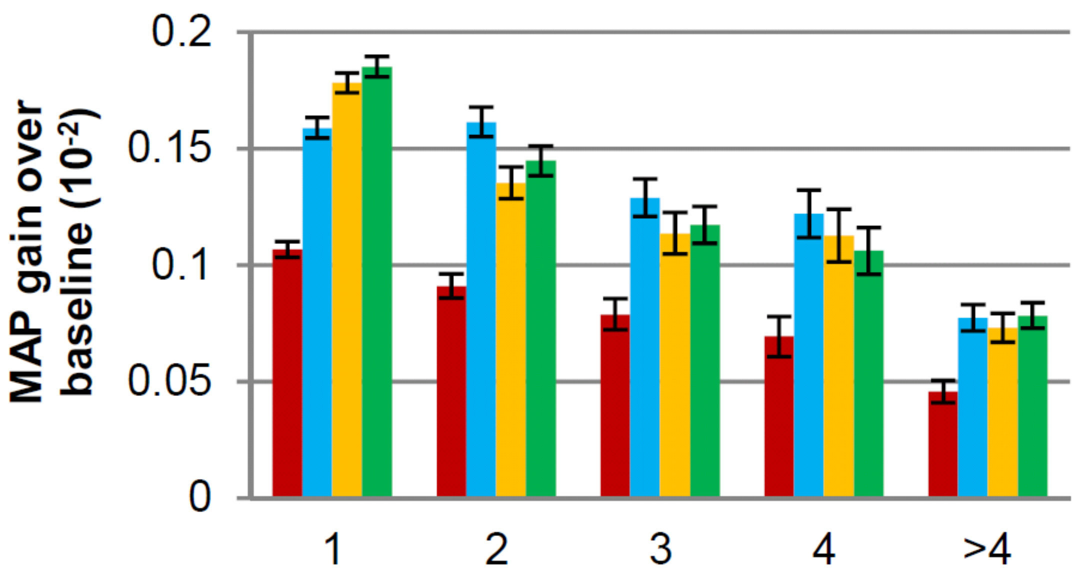
MAP/MRR gains on the test data, production ranker is the baseline.

Model	Δ MAP(10^{-2})	Δ MRR(10^{-2})	Rerank@1	Coverage	Win	Loss	Cost Rate
QG	**0.0888**±0.0023	**0.1076**±0.0024	0.46%	19.10%	28009	27507	98.21%
QI	**0.1425**±0.0028	**0.1431**±0.0029	0.70%	17.87%	26966	23214	86.09%
QGI	**0.1448**±0.0028	**0.1455**±0.0029	0.71%	19.10%	29259	25097	85.78%
TG	**0.1408**±0.0029	**0.1440**±0.0029	0.88%	67.37%	45866	37668	82.13%
TI	**0.1485**±0.0028	**0.1490**±0.0029	0.71%	19.44%	30932	26586	85.95%
TGI	**0.2292**±0.0035	**0.2318**±0.0036	1.22%	67.37%	32753	22292	68.06%

Some key findings:

- *Both query and task match get gains over baseline*
- *Task match better, especially when both feature groups used (TGI)*
- *Task match better coverage (> 3x) – re-rank@1 ~2x results as query*

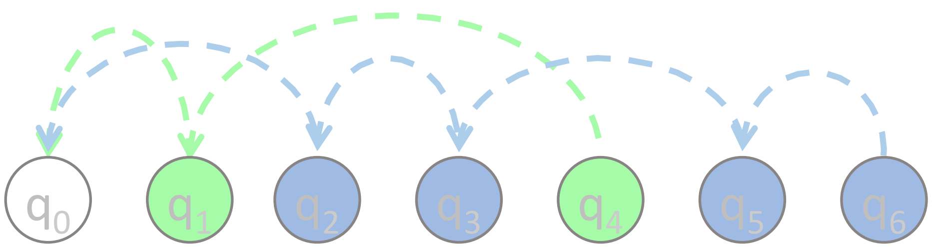
Effect of Query Sequence in Task



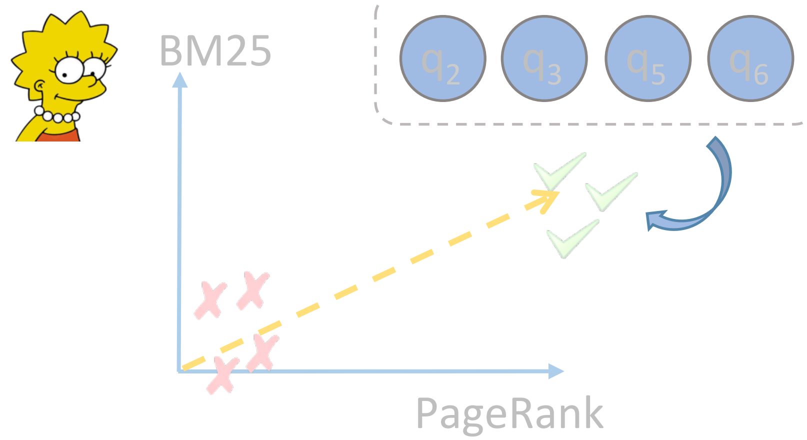
QG: Query-based Global Features
TG: Task-based Global Features
QI: Query-based Individual Features
TI: Task-based Individual Features

A Task-based Framework

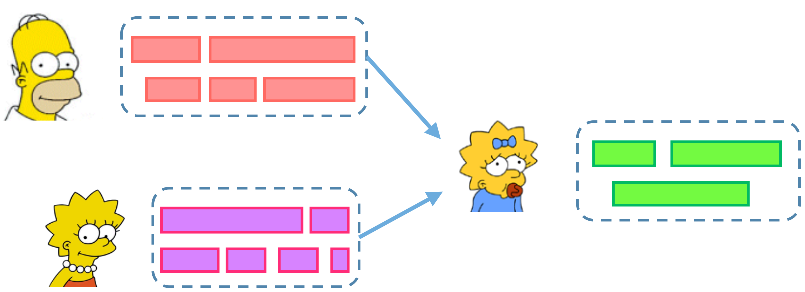
- Long-term task extraction



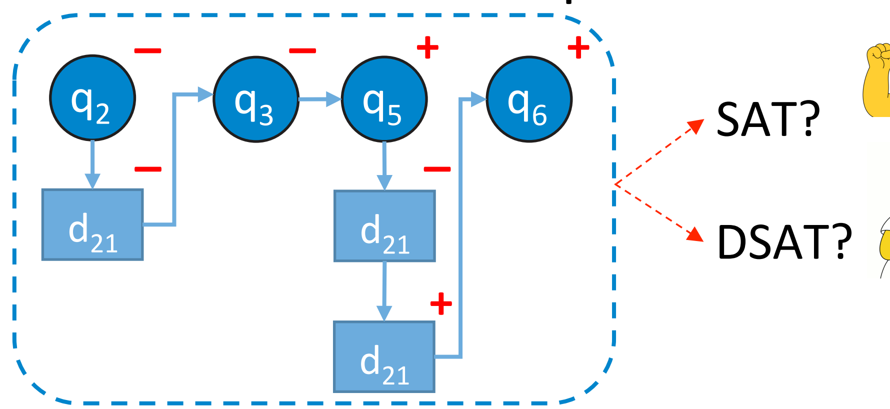
- In-task personalization



- Cross-user collaborative ranking



- Search-task satisfaction prediction



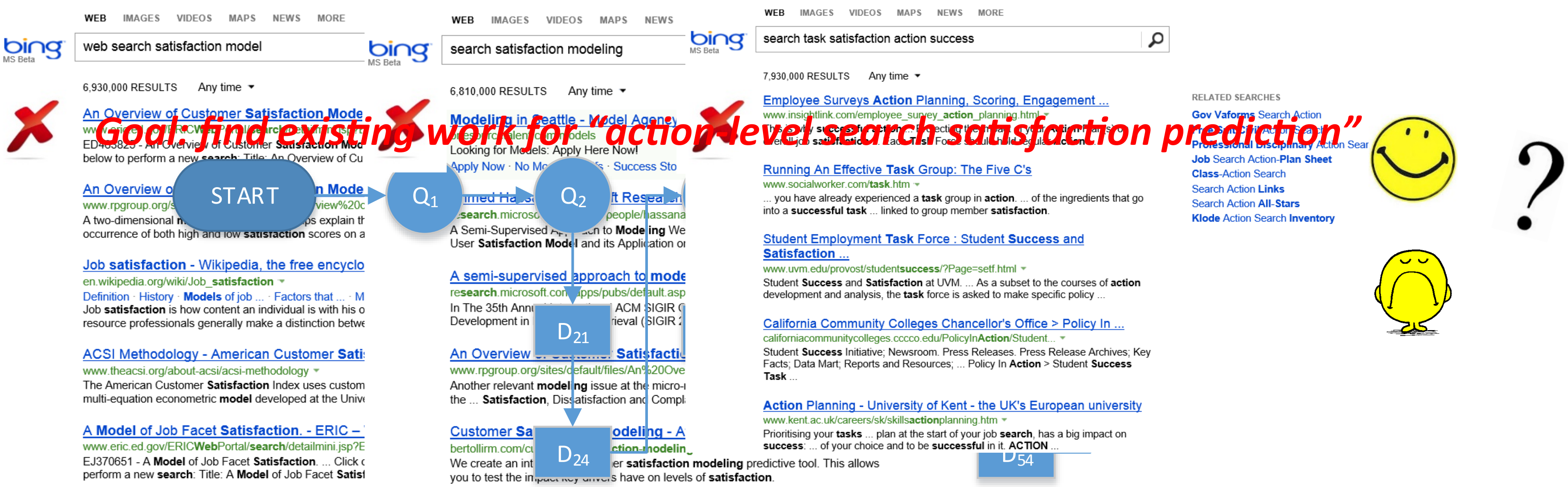
How to assess search result quality?

Query-level relevance evaluation [Ricardo et al. 1999]

- Metrics: MAP, NDCG, MRR

Task-level satisfaction evaluation [Hassan et al. WSDM'10]

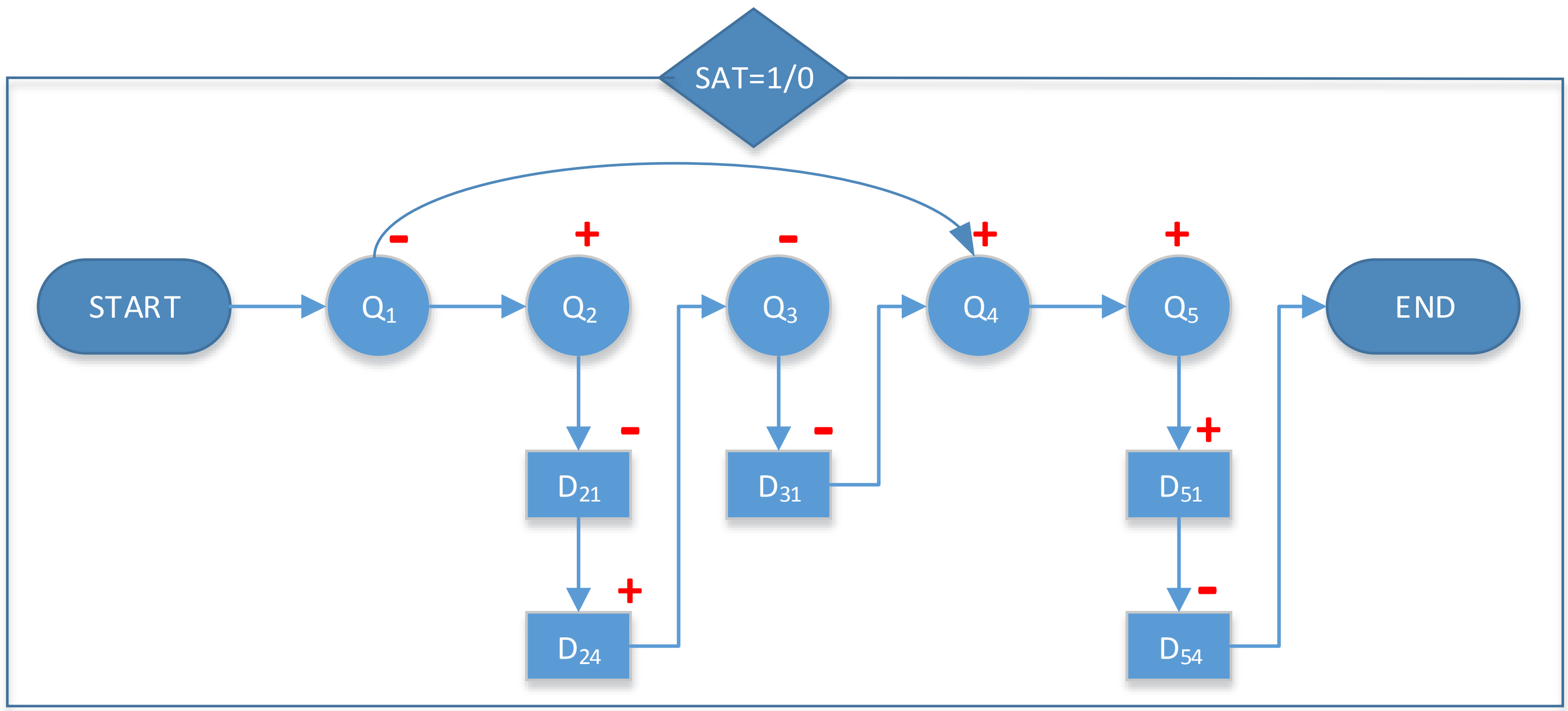
- Us



Modeling Latent Action Satisfaction for Search-task Satisfaction Prediction [SIGIR'14]

Hypothesis: *The desire for satisfaction drives users' interaction with search engines and that the satisfaction attained during the search-task contributes to the overall satisfaction*

alized as a latent tructural learning problem

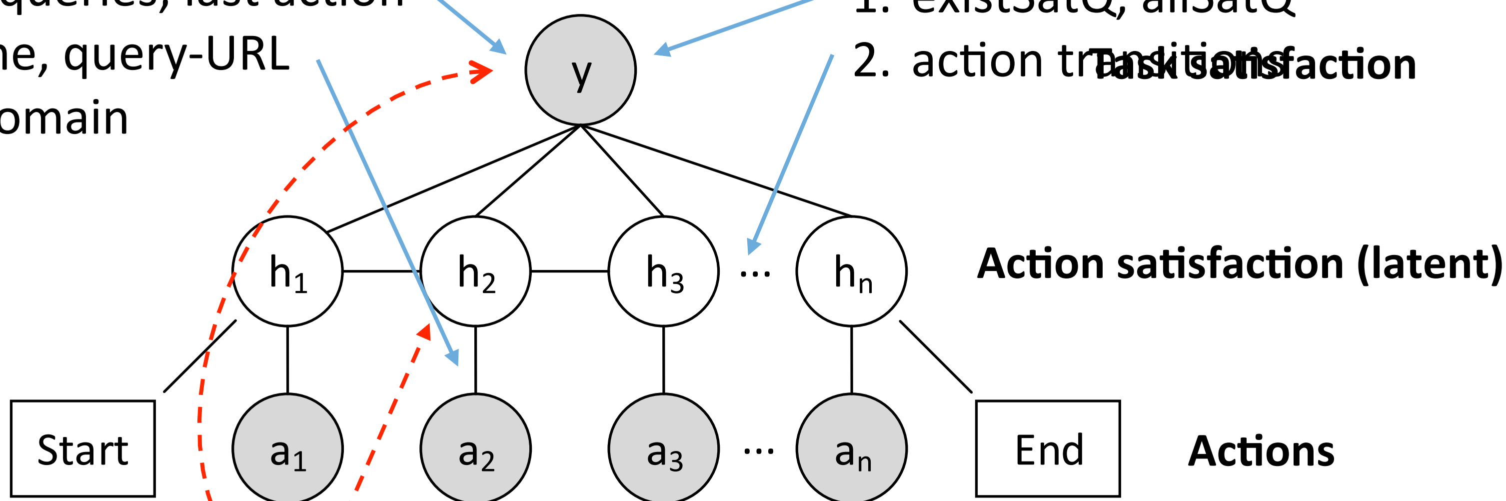


Modeling Latent Action Satisfaction for Search-task Satisfaction Prediction [SIGIR'14]

- Action-aware Task Satisfaction model (AcTS)
  - A latent structured learning approach

Short-range features:
1. #clicks, #queries, last action
2. Dwell time, query-URL match, domain

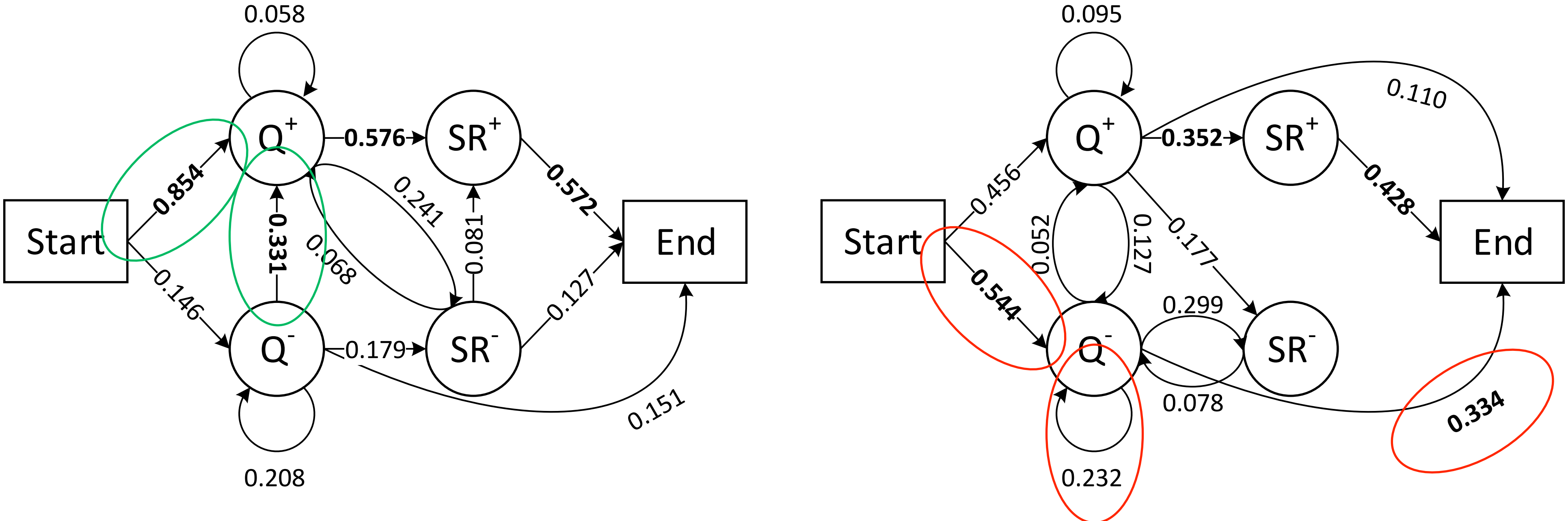
Long-range features:
1. existSatQ, allSatQ
2. action transitions



$$(\hat{y}, \hat{H}) = \underset{(y,H)\in\mathcal{Y}\times\mathcal{H}}{\arg\max}\; w^{\mathsf{T}}\underline{\Phi(A, H, y)}$$

Analysis of Search Behavior Patterns

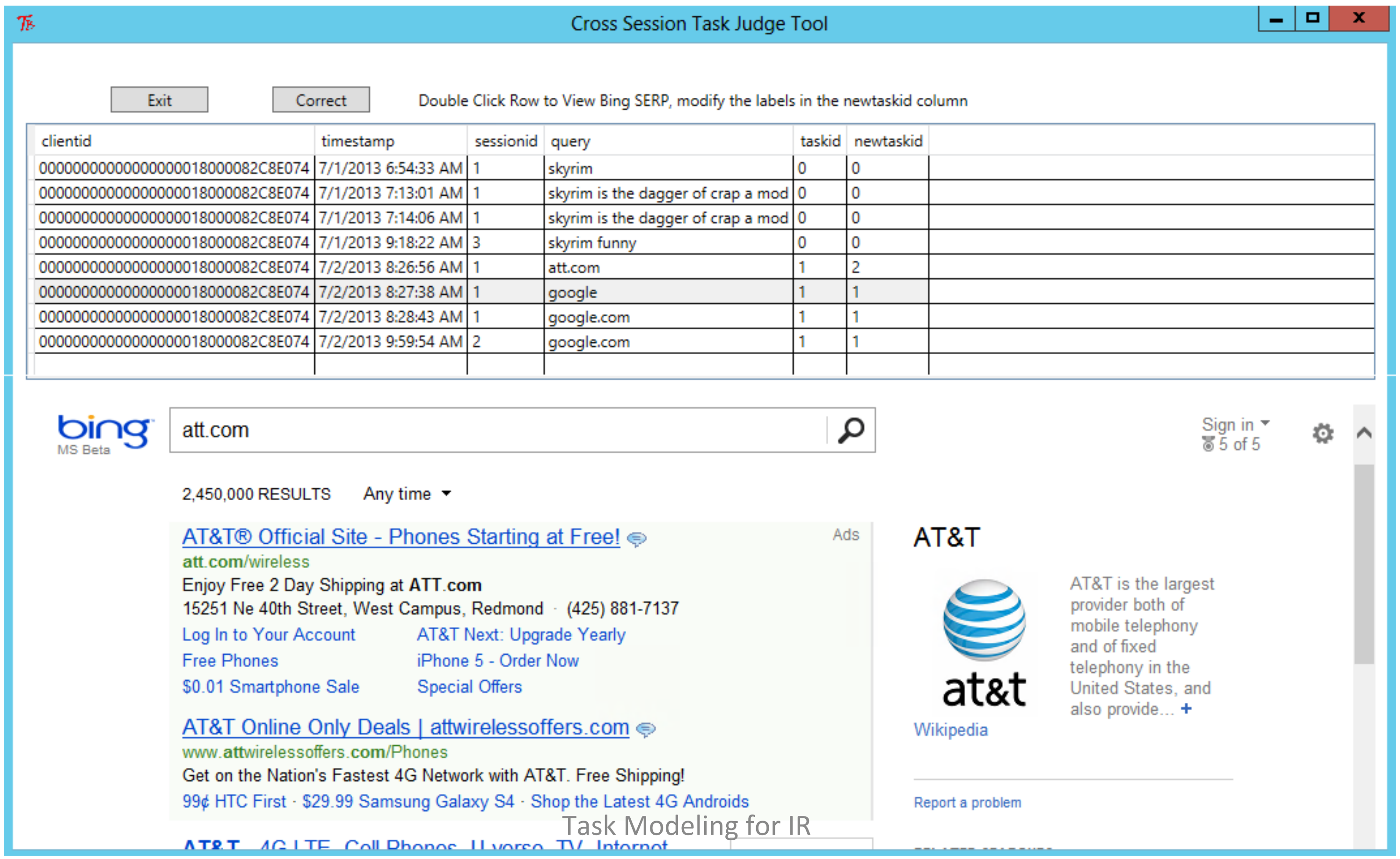
First-order transition probabilities between different actions within satisfying and unsatisfying tasks



(a) Satisfying tasks (b) Unsatisfying tasks

Deployed in Bing's Internal Tool Chain

It takes only 37 mins to process 14 days of Bing search query logs



Recent work in search-task mining

- Task-aware query recommendation *[Feild, H.* & Allan, J. *, SIGIR'13]*
  - Study query reformulation in tasks
- Click modeling in search tasks *[Zhang, Y. et al., KDD'11]*
  - Model users' click behaviors in tasks
- Query intent classification *[Cao H. et al., SIGIR'09]*
  - Explore rich search context for query classification

Conclusions

A task-based framework for user behavior modeling and search personalization

- Bestlink: an appropriate structure for search-task identification
- In-task personalization: exploiting users' in-task behaviors
- Cross-user collaborative ranking: leveraging search behaviors among differen users
- Search-task satisfaction prediction: modeling detailed action-level satisfactio

Future Directions

- Explore rich information about users for search-task identification
  - In-search, out-search behaviors
- From query-based search engine optimization to task-based
  - Optimize a user's long-term search utility
- Game-theoretic models for interacting with users
  - Machine and user collaborate to finish a task

References I

Hongning Wang, Yang Song, Ming-Wei Chang, Xiaodong He, Ryen White and Wei Chu. *Learning to Extract Cross-Session Search Tasks*. The 23rd International World-Wide Web Conference (WWW'2013), p1353-1364, 2013.

Hongning Wang, Xiaodong He, Ming-Wei Chang, Yang Song, Ryen White and Wei Chu. *Personalized Ranking Model Adaptation for Web Search*. The 36th Annual ACM SIGIR Conference (SIGIR'2013), p323-332, 2013.

Ryen White, Wei Chu, Ahmed Hassan, Xiaodong He, Yang Song and **Hongning Wang**. *Enhancing Personalized Search by Mining and Modeling Task Behavior*. The 23rd International World-Wide Web Conference (WWW'2013), p1411-1420, 2013.

Hongning Wang, Yang Song, Ming-Wei Chang, Xiaodong He, Ahmed Hassan and Ryen White. *Modeling Action-level Satisfaction for Search Task Satisfaction Prediction*. The 37th Annual ACM SIGIR Conference (SIGIR'2014), p123-132, 2014.

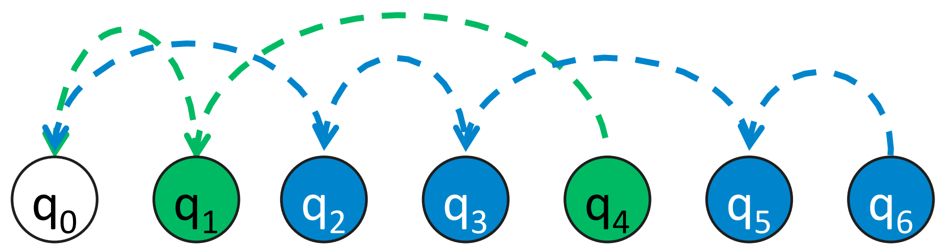
References II

- R. Jones and K. L. Klinkner. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. CIKM'08, pages 699–708. ACM.
- C. Lucchese, S. Orlando, R. Perego, F. Silvestri, and G. Tolomei. Identifying task-based sessions in search engine query log WSDM'11, pages 277–286. ACM.
- A. Kotov, P. N. Bennett, R. W. White, S. T. Dumais, and J. Teevan. Modeling and analysis of cross-session search tasks. SIGIR2011, pages 5–14, ACM.
- Z. Liao, Y. Song, L.-w. He, and Y. Huang. Evaluating the effectiveness of search task trails. WWW'12, pages 489–498. ACN
- Teevan, Jaime, Susan T. Dumais, and Eric Horvitz. "Personalizing search via automated analysis of interests and activities." *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2005.
- White, Ryen W., and Steven M. Drucker. "Investigating behavioral variability in web search." *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007.
- Shen, Xuehua, Bin Tan, and ChengXiang Zhai. "Context-sensitive information retrieval using implicit feedback." *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2005.
- Zhang, Y., Chen, W., Wang, D., & Yang, Q. User-click modeling for understanding and predicting search-behavior. In SIGKDD'11, (pp. 1388-1396). ACM.
- Feild, H., & Allan, J. Task-aware query recommendation. In SIGIR'13, (pp. 83-92). ACM.
- Cao, H., Hu, D. H., Shen, D., Jiang, D., Sun, J. T., Chen, E., & Yang, Q. Context-aware query classification. In SIGIR'09, (pp. 3-10). ACM.

Acknowledgements

- ChengXiang Zhai and team members in TIMAN group at UIUC
- Yang Song, Xiaodong He, Ming-Wei Chang, Ryen W. White and Kuansan Wang from Microsoft Research

Task: _a new perspective for us to understand users' search intent_



Thank you!

&A

Explore Domain Knowledge for Automating Model Learning

Same-query

5/30/2012	
5/30/2012 9:12:14	airline tickets
5/30/2012 9:20:19	rattlers
5/30/2012 9:42:21	charlize theron snow white
5/30/2012 21:13:34	charlize theron
5/30/2012 21:13:54	charlize theron movie opening
5/31/2012	
5/31/2012 8:56:39	sulphur springs school district
5/31/2012 9:10:01	airline tickets

Sub-query
Sub-query

A generalized margin

$$\tilde{\Delta}(\tilde{y}_n, \hat{y}, \hat{h}) = |\mathcal{Q}_n| - |\mathcal{C}_n| - \sum_{i,j} h(i,j)\tilde{\sigma}(y,(i,j))$$

queries

connected components

$$\tilde{\sigma}(y,(i,j)) = \begin{cases} \lambda^+ & \text{if } \tilde{y}(i) = \tilde{y}(\\ -\lambda^- & \text{if } \tilde{y}(i) \neq \tilde{y}(\\ 0 & \text{otherwise} \end{cases}$$

User-level improvement analysis

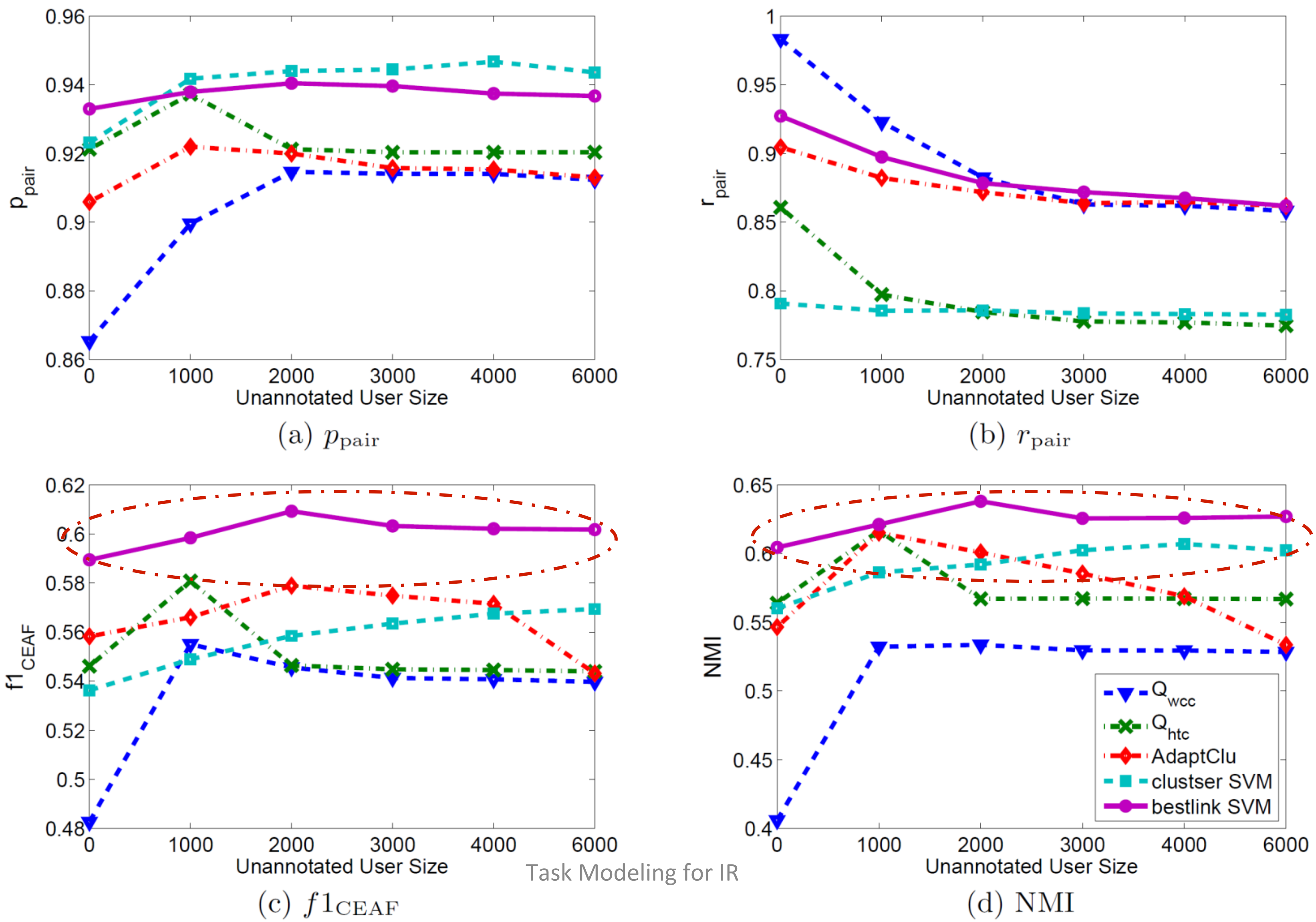
Adapted-LambdaRank against global LambdaRank model

per-user basis adaptation baseline

	User Class	Method	ΔMAP	ΔP@1	ΔP@3	ΔMRR
[10, ∞) queries	**Heavy**	RA	0.1843	0.3309	0.0120	0.1832
		Cross	0.1998	0.3523	0.0182	0.1994
[5, 10) queries	**Medium**	RA	0.1102	0.2129	0.0025	0.1103
		Cross	0.1494	0.2561	0.0208	0.1500
(0, 5) queries	**Light**	RA	0.0042	0.0575	-0.0221	0.0041
		Cross	0.0403	0.0894	-0.0021	0.0406

Use cross-training to determine feature grouping

Automating Model Learning with Domain Knowledge



(a) p_{pair} (b) r_{pair}

(c) $f1_{\mathrm{CEAF}}$ (d) NMI

Search Task Extraction Methods

- Baselines
  - QC_wcc/QC_htc *[Lucchese et al. WSDM' 11]*
    - Post-processing for binary same-task classification
  - Adaptive-clustering *[Cohen et al. KDD'02]*
    - Binary classification + single-link agglomerative clustering
  - Cluster-SVM *[Finley et al. ICML'05]*
    - All-link structural SVM

Binary classification based solution

Structured learning solution

Experimental Results

Four weeks of Bing query-click logs

- Logs collected from an A/B test with no other personalization

Week 1: Feature generation

- Compute $s\downarrow k$ for clicked URLs

Weeks 2-3: Learn re-ranking model (LambdaMART)

Week 4: Evaluation

- *Re-rank top-10 for each query*
- *Compute MAP and MRR for re-ranked lists (and coverage stats)*

	Training	Validation	Evaluation
Tasks	1,165,083	1,126,452	1,135,320
Queries/Task	1.678	1.676	1.666

Learn from Related Tasks

- For each URL u in top 10 for current query, compute score s_k

$$s_k(t,u) = \sum_{t' \in T} \left(k(t,t') \cdot w(t',u)\right)$$

  - $k(t,t')$: relatedness between t, related task t', computed in different ways
  - $w(t',u)$: importance of URL in related task (we use click frequency)

- Generate s_k for a range of different $k(t,t')$

Syntactic similarity,
URL similarity,
topical similarity,
etc.

Action-level Satisfaction Modeling

As additional features for document relevance estimation

- 4-month Bing search log
- Ranker: LambdaMART
- 398 standard ranking features, e.g., BM25, language model score and PageRank

igid assumption: task-
atisfaction equals to
ction-satisfaction

tasks are satisfactory

%	P@1	MAP	NDCG@5	MRR
MML	+4.926	+3.482	+3.573	+2.650
LogiReg	+5.110	+3.352	+3.783	+2.776
session-CRF	+4.752	+3.402	+3.896	+2.616
SUM	+5.101	+3.405	+3.946	+2.807
AcTS*	**+5.366**	**+3.819**	**+4.278**	**+2.955**

* Indicates *p*-value<0.01

ask-level satisfaction prediction performance

Toolbar data set [Hassan et al. CIKM'11]

- 7306 tasks from 153 users
- In-situ task satisfaction annotations from the actual users

Assumption:
action satisfaction
= task satisfaction

	Avg-f_1	T^+-f_1	T^--f_1	Accuracy
MML	0.707	0.897	0.518	0.830
LogiReg	0.740	0.918	0.563	0.861
Session-CRF	0.728	0.910	0.545	0.850
AcTS	**0.761***	**0.938***	**0.584***	**0.893***

* Indicates *p*-value<0.01