Mining the Search Trails of Surfing Crowds: Identifying Relevant Websites from User Activity Data

Misha Bilenko and Ryen White

presented by Matt Richardson

Microsoft Research

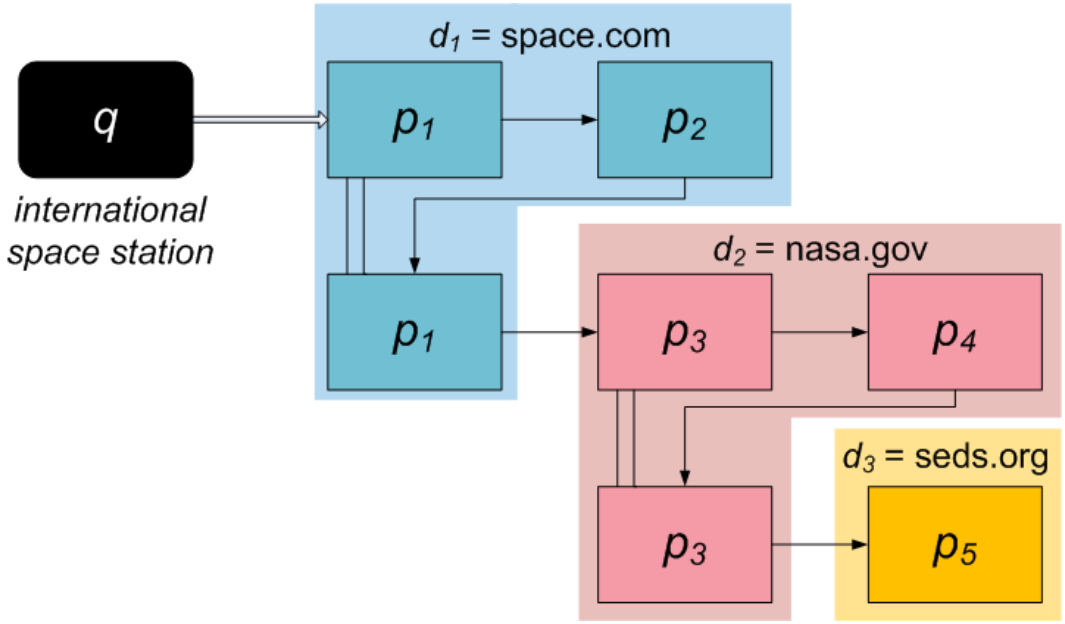
Search = Modeling User Behavior

- Retrieval functions estimate relevance from behavior of several user groups:
  - *Page authors* create page contents
    - TF-IDF/BM25, query-is-page-title, …
  - *Page authors* create links
    - PageRank/HITS, query-matches-anchor text, …
  - *Searchers* submit queries and click on results
    - Clickthrough, query reformulations
- Most user behavior occurs beyond search engines
  - Viewing results and browsing beyond them
  - What can we capture, and how can we use it?

Prior Work

- Clickthrough/implicit feedback methods
  - Learning ranking functions from clicks and query chains [Joachims ‘02, Xue *et al.* ‘04, Radlinski-Joachims ’05 ‘06 ‘07]
  - Combining clickthrough with traditional IR features [Richardson *et al.* ‘06, Agichtein *et al.* ‘06]
- Activity-based user models for personalization
  - [Shen *et al.* ‘05, Tan *et al.* ’06]
- Modeling browsing behavior
  - [Anderson *et al.* ‘01, Downey *et al.* ‘07, Pandit-Olston ’07]

Search Trails



- Trails start with a search engine query
- Continue until a terminating event
  - Another search
  - Visit to an unrelated site (social networks, webmail)
  - Timeout, browser homepage, browser closing

Trails vs. Click logs

- Trails capture dwell time
  - Both attention share and pageview counts are accounted
- Trails represent user activity across many websites
- Browsing sequences surface “under-ranked” pages

- Click logs are less noisy
- Position bias is easy to control

Predicting Relevance from Trails

- **Task: given a trails corpus $D = \{q_i \rightarrow (d_{i1}, \ldots, d_{ik})\}$ predict relevant websites for a new query *q***
- Trails give us the good pages for each query…

  …can't we just lookup the pages for new queries?
  - Not directly: 50+% of queries are unique
  - Page visits are also extremely sparse
- Solutions:
  - Query sparsity: term-based matching, language modeling
  - Pageview sparsity: smoothing (domain-level prediction)

Model 1: Heuristic

- Documents ≈ websites
- Contents ≈ queries preceding websites in trails
- Split queries into terms, compute frequencies
  - Terms include unigrams, bigrams, named entities
- Relevance is analogous to BM25 (TF-IDF)
  - Query-term frequency (QF) and inverse query frequency (IQF) terms incorporate corpus statistics and website popularity.

$$Rel(d_i, q) = \sum_{t_j \in q} QF(t_j, d_i) \times IQF(t_j)$$

Model 2: Probabilistic

- IR via language modeling [Zhai-Lafferty, Lavrenko]

$$Rel(d_i, q) = p(d_i|q) = \sum_{t_j \in q} p(t_j|q)\, p(d_i|\, t_j)$$

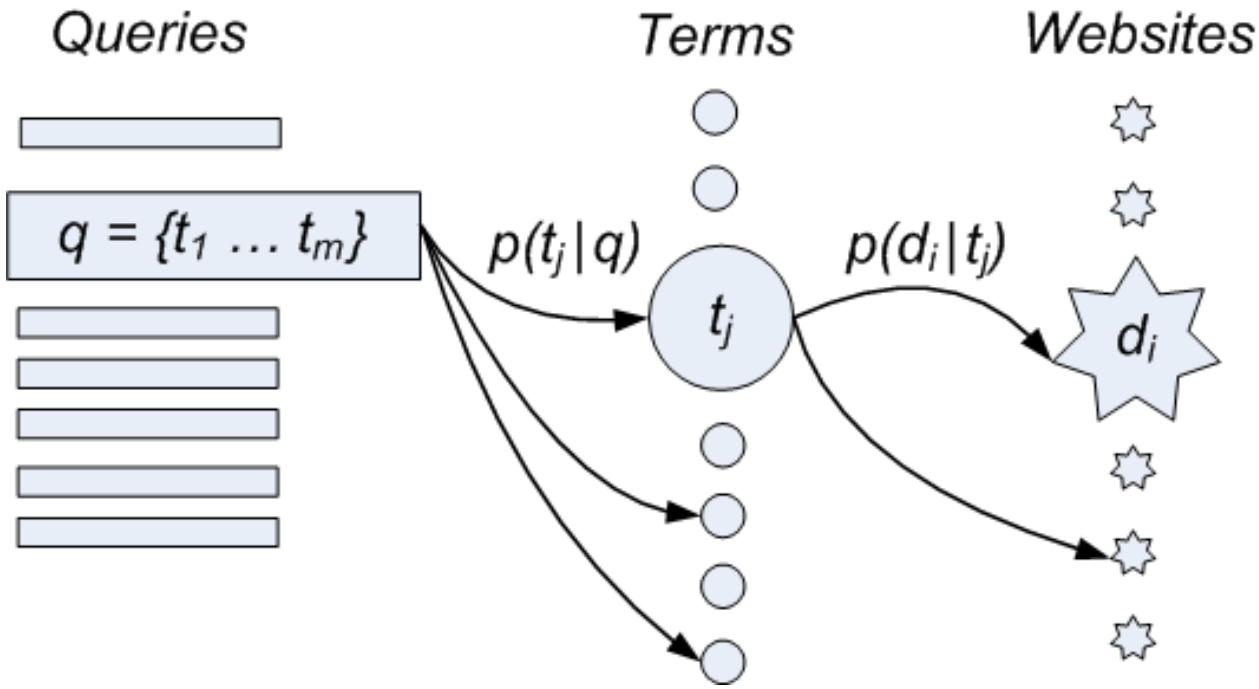
- Query-term distribution gives more mass to rare terms:

$$p(t_j|q) = \frac{\exp\left(-p(t_j)\right)}{\sum_{t_k \in q} \exp\left(-p(t_k)\right)}$$

- Term-website weights *combine dwell time and counts*

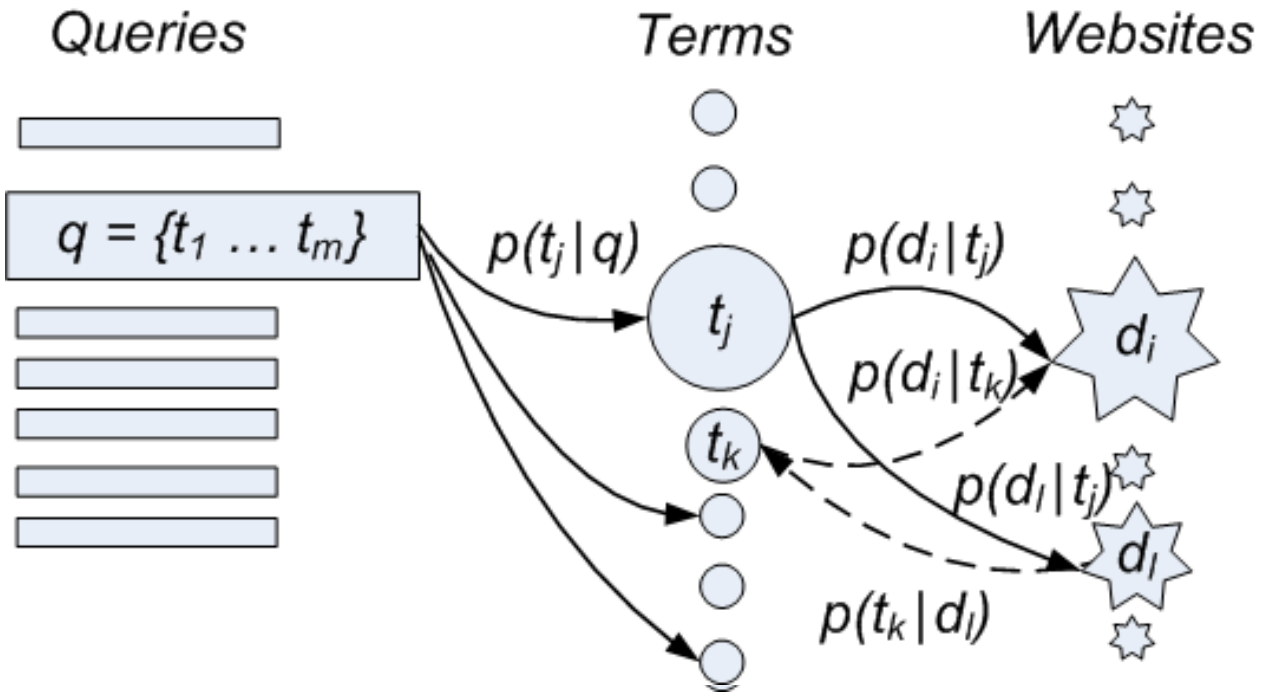
$$f(d_i, t_j) = \sum_{\forall q':\, t_j \in q';\; q' \to d_i,} \log(time(q', d_i)) \qquad p(d_i|t_j) = \frac{f(d_i, t_j)}{\sum_{d_k \in D} f(d_k, t_j)}$$

Model 2: Probabilistic (cont.)



- Basic probabilistic model is noisy
  - Misspellings, synonyms, sparseness

Model 3: Random Walks



- Basic probabilistic model is noisy
  - Misspellings, synonyms, sparseness
- Solution: random walk extension

Evaluation

- Train: 140+ million search trails (toolbar data)
- Test: human-labeled relevance set, 33K queries

  q =[*black diamond carabiners*]

URL	Rating
www.bdel.com/gear	Perfect
www.climbing.com/Reviews/biners/Black_Diamond.html	Excellent
www.climbinggear.com/products/listing/item7588.asp	Good
www.rei.com/product/471041	Good
www.nextag.com/BLACK-DIAMOND/	Fair
www.blackdiamondranch.com/	Bad

Evaluation (cont.)

- Metric: NDCG (Normalized Discounted Cumulative Gain)

$$NDCG(i) = \frac{DCG(i)}{DCG_{perfect}(i)} \qquad DCG(i) = \sum_i Gain\big(r(i)\big) \times Discount(i)$$

$$Gain(r(i)) = 2^{r(i)} - 1 \qquad Discount(i) = \log(1+i)$$

$$NDCG(i) = N_i \sum_i \frac{2^{r(i)} - 1}{\log(1+i)}$$

- Preferable to MAP, Kendall's Tau, Spearman's, etc.
  - Sensitive to top-ranked results
  - Handles variable number of results/target items
  - Well correlated with user satisfaction [Bompada *et al.* '07]

Evaluation (cont.)

- Metric: NDCG (Normalized Discounted Cumulative Gain)

$$NDCG(i) = N_i \sum_i \frac{2^{r(i)} - 1}{\log(1+i)}$$

Perfect ranking

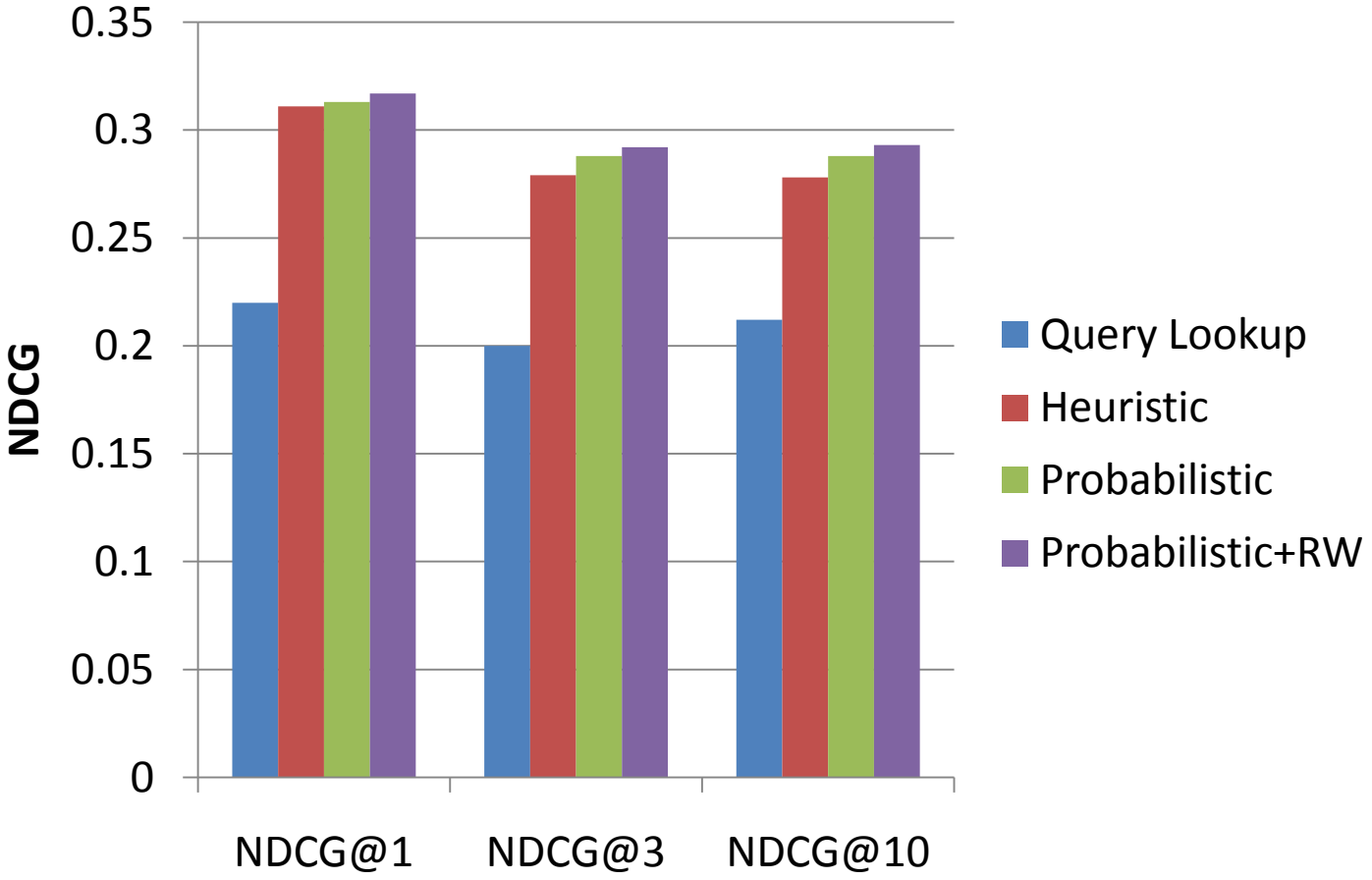
i	d	$r(i)$	$DCG_{perfect}(i)$
1	d_1	5	31
2	d_2	4	40.5
3	d_3	4	48.0
4	d_4	3	51.0
5	d_5	1	51.4

Obtained ranking

i	d	$r(i)$	$DCG(i)$	$NDCG(i)$
1	d_1	5	31	1
2	d_7	0	31	0.766
3	d_4	3	34.5	0.719
4	d_5	1	34.9	0.684
5	d_2	4	40.7	0.792

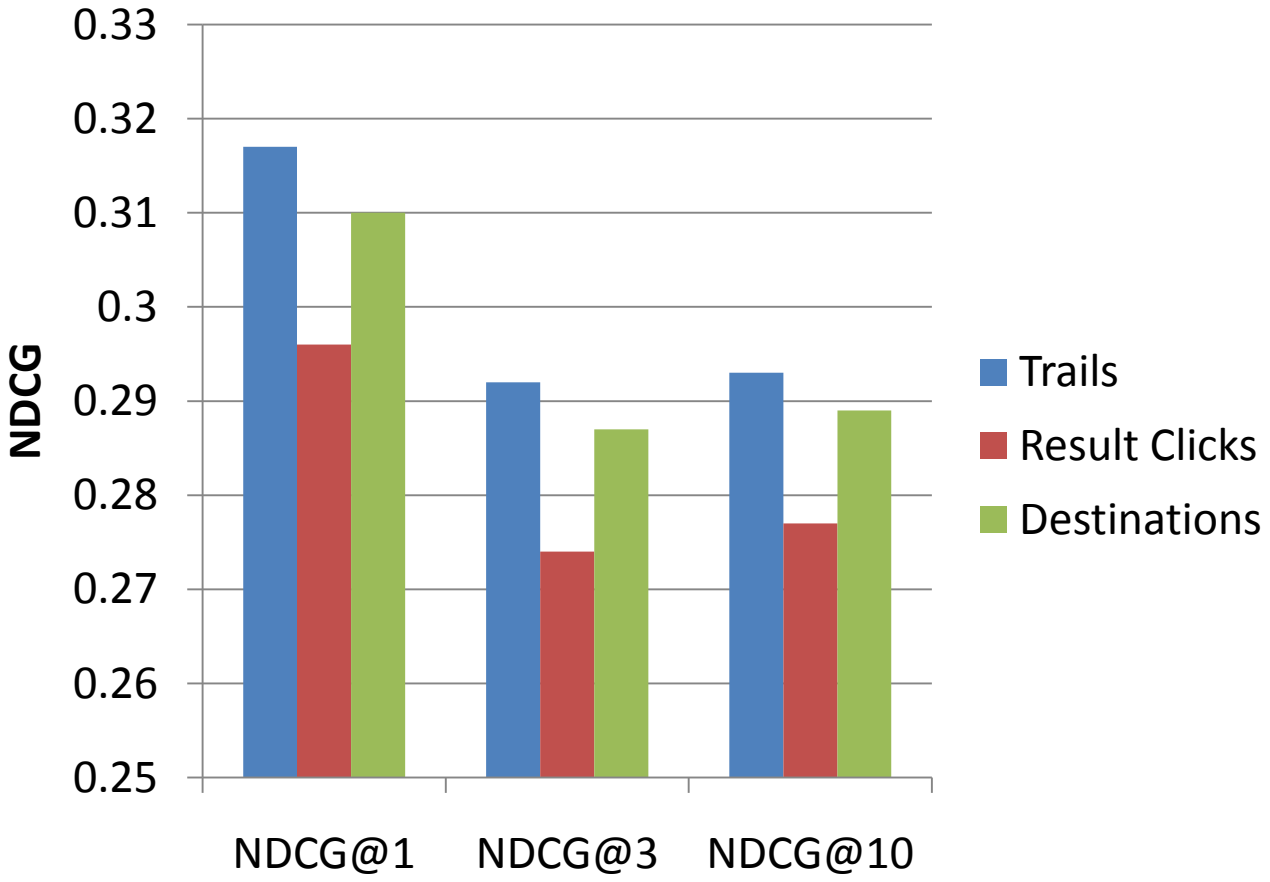
Results I: Domain ranking (cont.)

- Predicting correct ranking of domains for queries



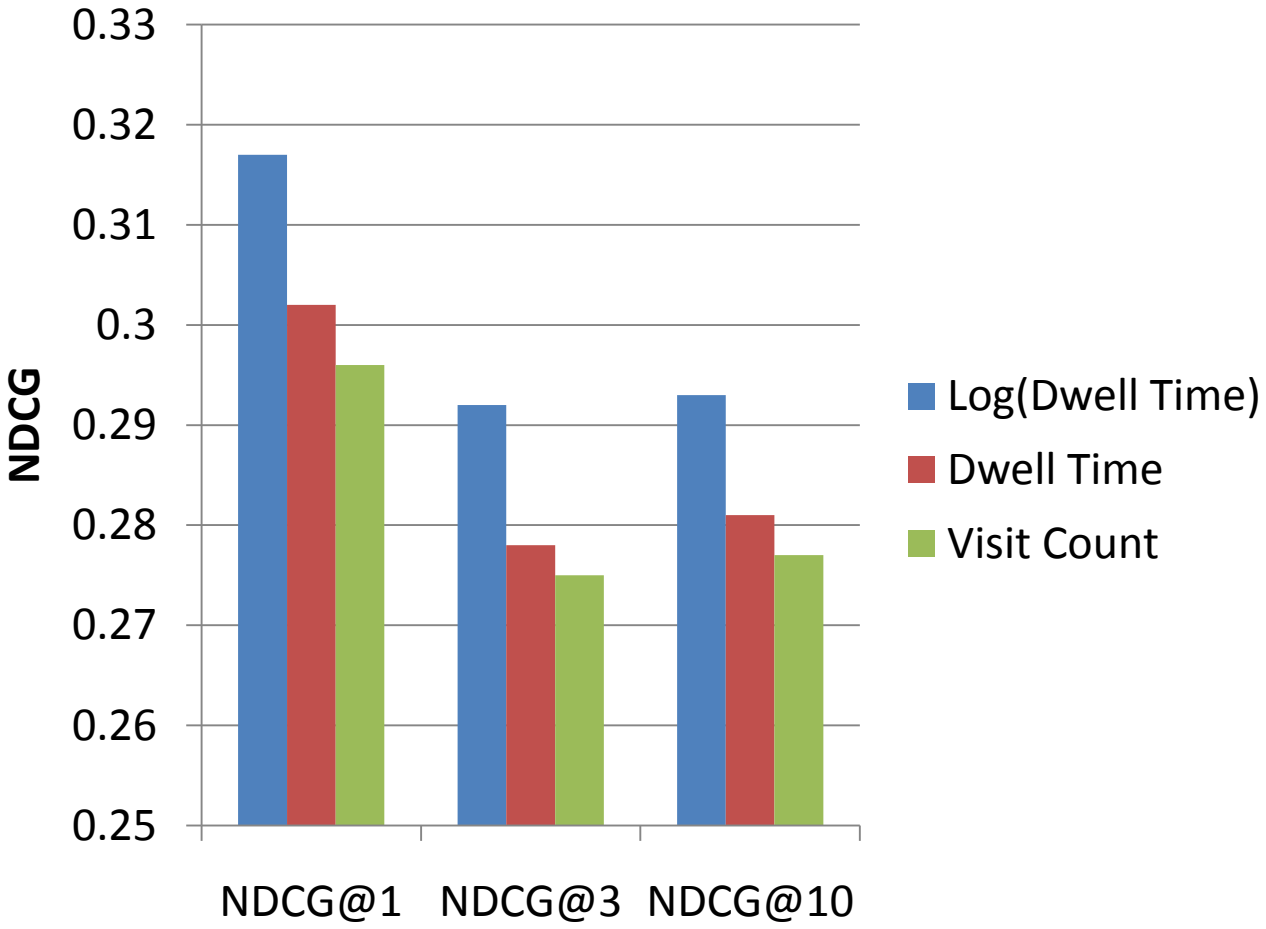
Results I: Domain ranking (cont.)

- Full trails vs. search result clicks vs. “destinations”



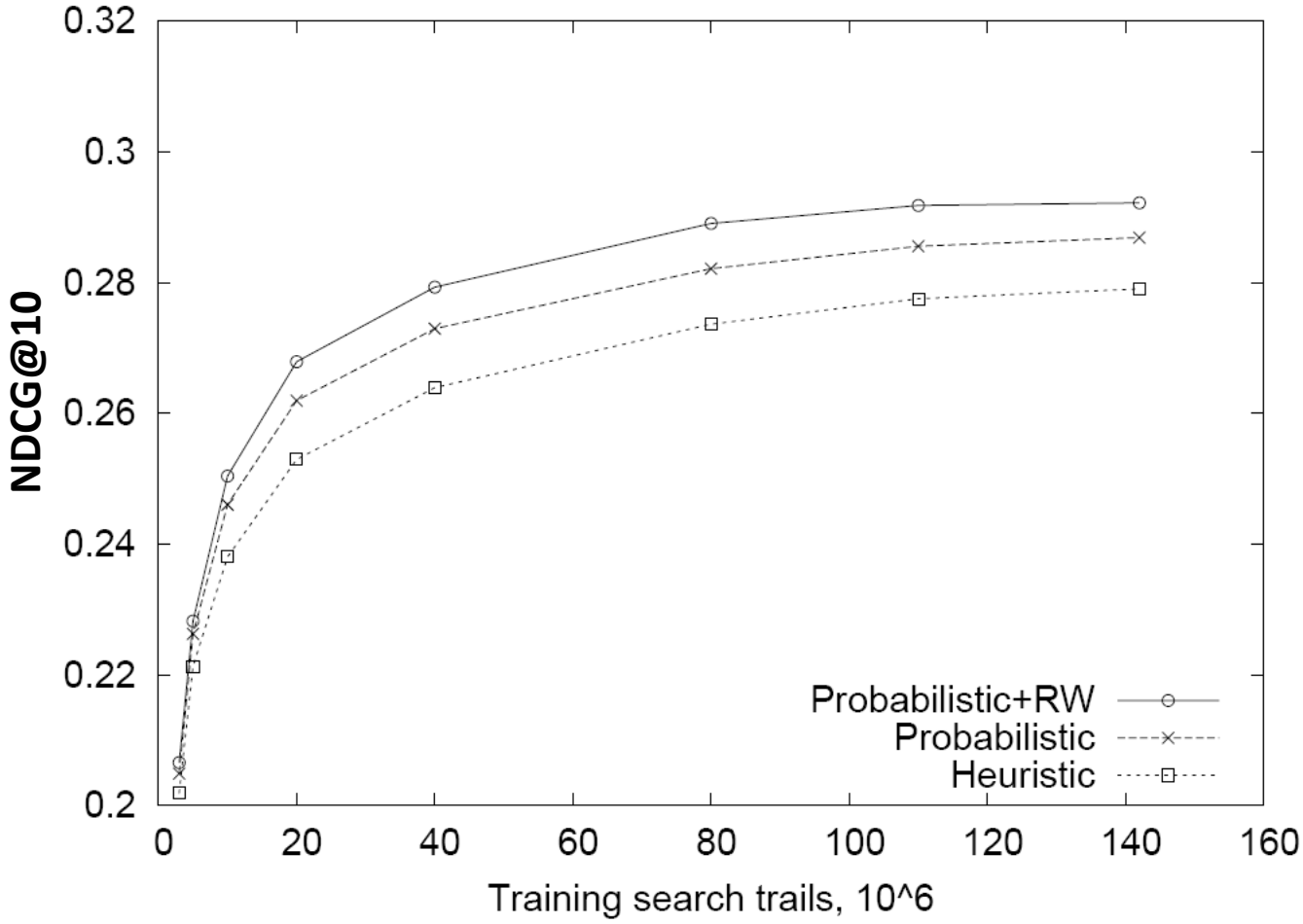
Results I: Domain ranking (cont.)

- Scoring based on dwell times vs. visitation counts



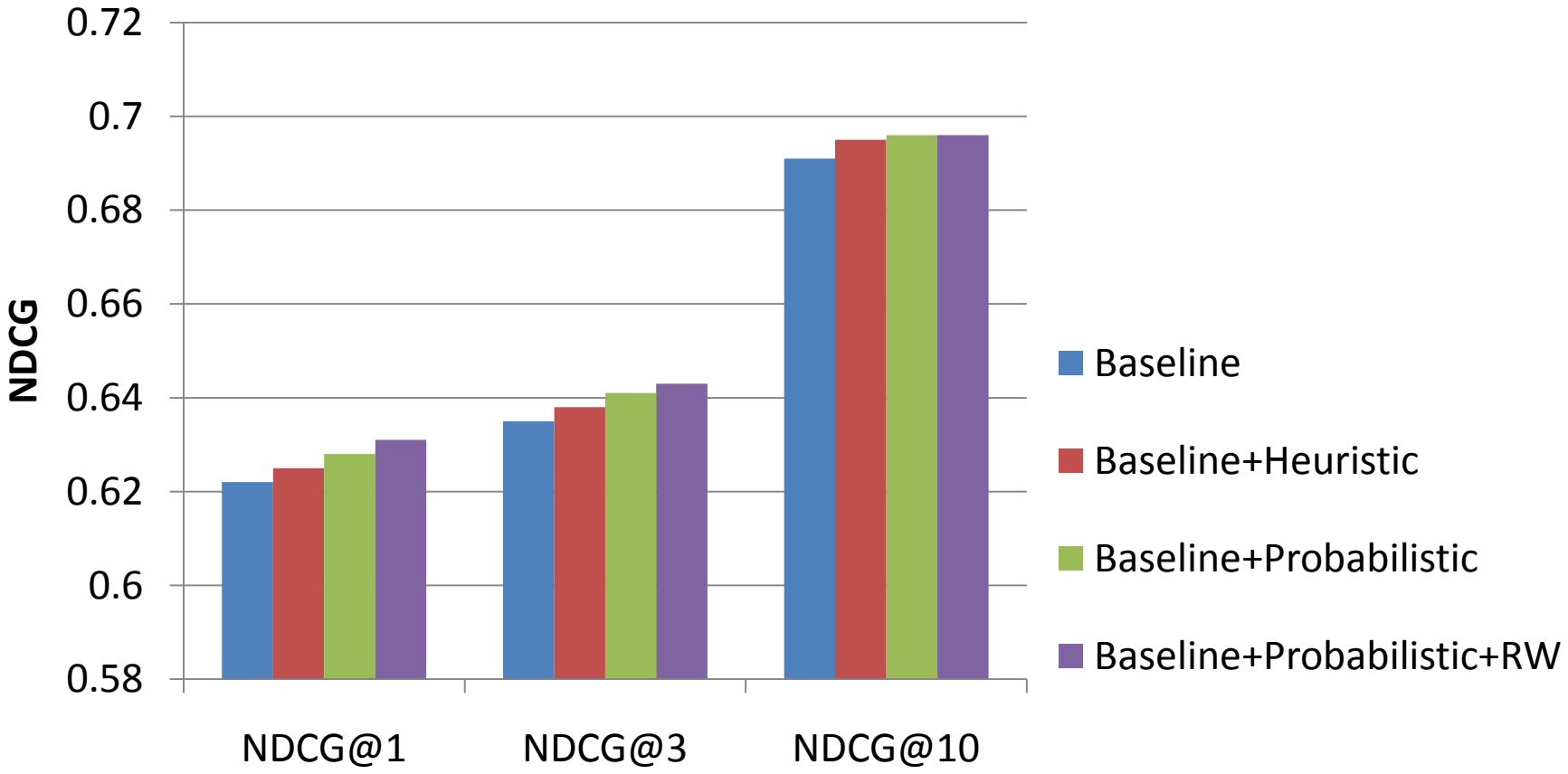
Results I: Domain ranking (cont.)

- What's better than data? LOTS OF DATA!



Results II: Learning to Rank

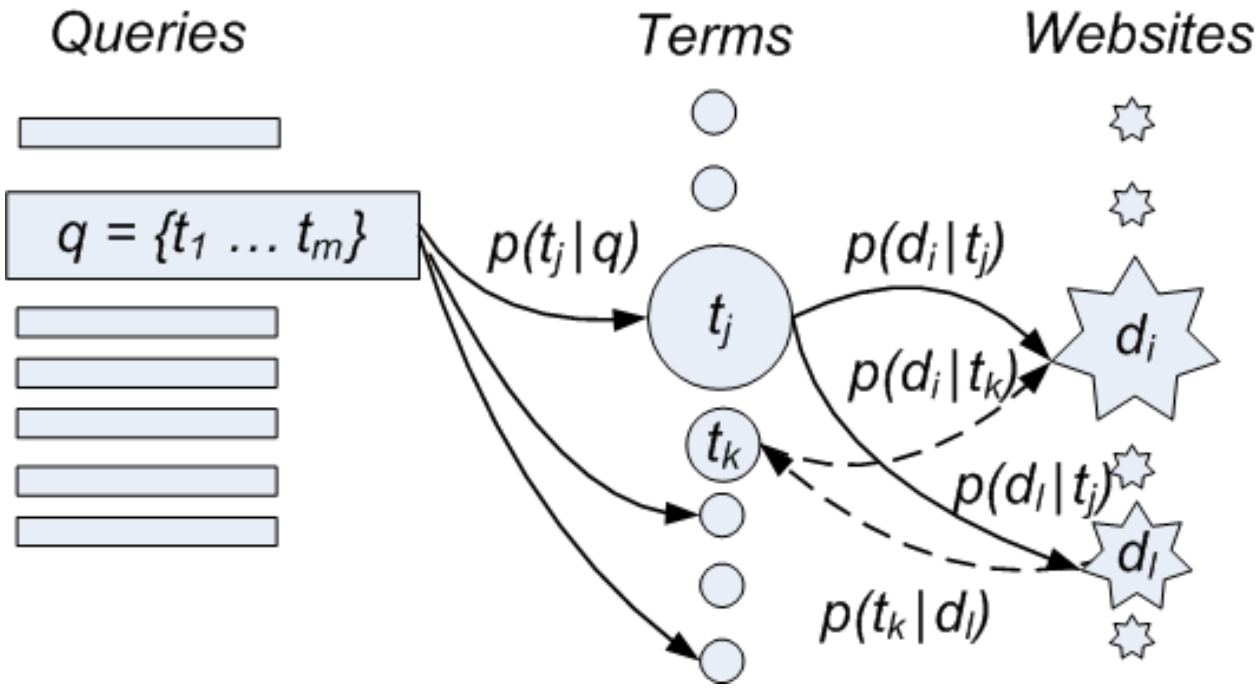
- Add *Rel*(*q*, d_i) as a feature to RankNet [Burges *et al.* '05]
  - Thousands of other features capture various content-, link- and clickthrough-based evidence



Conclusions

- Post-search browsing behavior (search trails) can be mined to extract users' implicit endorsement of relevant websites.
- Trail-based relevance prediction provides unique signal not captured by other (content, link, clickthrough) features.
- Using full trails outperforms using only search result clicks or search trail destinations.
- Probabilistic models incorporating random walks provide best accuracy by overcoming data sparsity and noise.

Model 3: Random Walks (cont.)



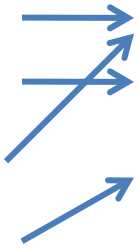
$$Rel_{RW}(d_i, q) = \sum_{t_j \in q} p(t_j|q)\,(\alpha p(d_i|t_j) + \\ +(1-\alpha)\sum_{d_l, t_k} p(d_l|t_j)p(t_k|d_l)p(d_i|t_k))$$

URLs vs. Websites

- Website ≈ domain
  - Sites: *spaces.live.com*, *news.yahoo.co.uk*
  - Not sites: *www2.hp.com*, *cx09hz.myspace.com*
- Scoring: $r(site) = \max_{page \in site} r(page)$

URL ranking

URL	Rating
www.bdel.com/gear	Perfect
www.rei.com/product/471041	Good
www.bdel.com/about	Fair
www.blackdiamondranch.com/	Bad



Website ranking

URL	Rating
bdel.com	Perfect
rei.com	Good
blackdiamondranch.com	Bad