11-442 / 11-642 / 11-742
Search Engines

Personalization

Jamie Callan
Carnegie Mellon University
callan@cs.cmu.edu

Lecture Outline

Today, three approaches to personalization

- Topic-based personalization
- Long-term vs. short-term personalization
- Personalization for typical vs. atypical information needs

This lecture is based on work done at Microsoft Research

Personalization is an active area of research

Our goals

- Get a sense of what is being done, and how it is being done

Personalization #1

Web search engines are tuned to satisfy a user population

- How can they be tuned to satisfy individuals?

Solution components

- **Representation:** Summarizing a person's interests / preferences
- **Learning:** Obtaining interests / preferences from data
- **Ranking:** Use interests / preferences in a retrieval algorithm
  - – Not our focus today

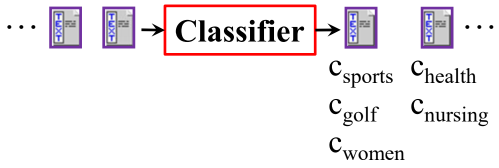
I have simplified this discussion to make it easier to understand (i.e., it isn't exactly what Sontag, et al. proposed)

(Sontag, et al., 2012)

Personalization #1

<u>Before</u> indexing, a classifier assigns each document to [0..n] categories

- E.g., categories from the top layers of the Open Directory
- *Controlled vocabulary indexing*

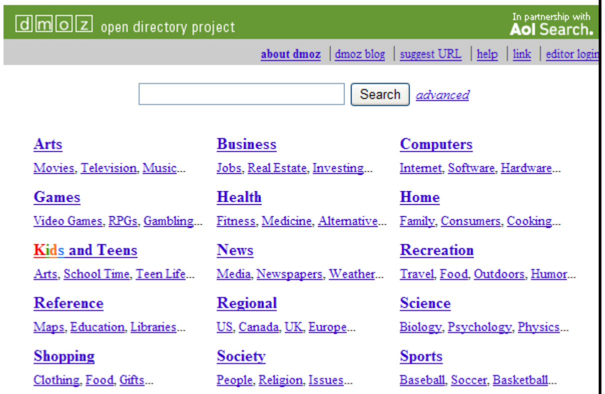


… → Classifier → …

c_{sports} c_{health}

c_{golf} $c_{nursing}$

c_{women}



dmoz open directory project

In partnership with Aol Search.

about dmoz | dmoz blog | suggest URL | help | link | editor login

Search advanced

Arts Movies, Television, Music...

Business Jobs, Real Estate, Investing...

Computers Internet, Software, Hardware...

Games Video Games, RPGs, Gambling...

Health Fitness, Medicine, Alternative...

Home Family, Consumers, Cooking...

Kids and Teens Arts, School Time, Teen Life...

News Media, Newspapers, Weather...

Recreation Travel, Food, Outdoors, Humor...

Reference Maps, Education, Libraries...

Regional US, Canada, UK, Europe...

Science Biology, Psychology, Physics...

Shopping Clothing, Food, Gifts...

Society People, Religion, Issues...

Sports Baseball, Soccer, Basketball...

(Closed March 17, 2017)

(Sontag, et al., 2012)

Personalization #1

User representation

- Model a person's interest in different topic categories
  - A probabilistic distribution over categories

	movies	tv	music	…	golf	football
Bob	0.01%	2.33%	0.92%	…	2.00%	3.21%
Mary	2.73%	1.88%	2.12%	…	0.08%	0.00%
:	:	:	:	…	:	:

- Train a model for each person (e.g., Bob)
  - Use Bob's queries and clicks from the Bing search log

$$p(category) = \frac{1}{|\text{clicked}|} \sum_{d \in \text{clicked}} p(category|d)$$

(Sontag, et al., 2012)

Personalization #1

Architecture

- Use a highly-tuned ranker to get an initial ranking
  - E.g., Bing
- Rerank the top *n* documents using a combination of the initial ranking score and how well document *d* categories match categories for user *u*
  - E.g., $\beta\, p_{Relevance}(d|q) + (1 - \beta)\, p_{CategoryMatch}(d|q, u)$
    $\beta = 0.3$ in their experiments
  - CategoryMatch can be implemented in different ways
    - » $\sum_{c \in d} p(c|d)\, p(c|u)$
    - » …

(Sontag, et al., 2012)

Personalization #1:
Data

25 days of search history

- **Train:** Search history from Sep 1-20, 2010
    - – Users must have at least 100 satisfied result clicks
- **Test:** Search history from Sep 21-25, 2010
    - – **Queries:** 1 word long, non-navigational
        - » Ambiguous, but not rare
    - – **Relevance:** The last satisfied result click in a session
        - » Thus, just 1 relevant document per query

102,417 queries from 54,581 users

(Sontag, et al., 2012)

Personalization #1:
Experimental Results

Bing vs. Personalized Bing

- **Metric:** Mean reciprocal rank (MRR)
- **ODP classifier accuracy:** 60% Micro-averaged F_1, 86% coverage
- **Effect of personalization**
    - – 1-2% improvement in overall MRR
    - – 17-18% improvement in MRR for results that change position
- **Effect of personalization on acronyms**
    - – 5% improvement in overall MRR
    - – 17-22% improvement in MRR for results that change position

Good results, because the search engine is highly tuned

(Sontag, et al., 2012)

Personalization #1

Key ideas

- A person's long-term interest in different high-level topics can be inferred from training data
- Documents can be automatically assigned to those categories
- A personalized search engine considers several types of evidence
  - How well the document matches the query
  - The query-independent value of the document
    - PageRank, spam score, popularity, …
  - Whether the document is on a topic the person is interested in
- This form of personalization seems to improve results
  - On ambiguous queries, anyway ☺

Lecture Outline

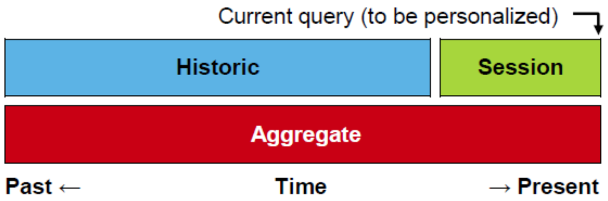
Three approaches to personalization

- Topic-based personalization
- Long-term vs. short-term personalization
- Personalization for typical vs. atypical information needs

Personalization #2

Personalization can be based on three types of information

- Information acquired over a long period of time ("historic")
- Information from the current search session ("session")
- A combination of historic and session information



Current query (to be personalized)

Historic

Session

Aggregate

Past ← Time → Present

Treat these as different views of a person's history

- Each view has the same features (calculated from different data)

(Bennett, et al., 2012)

Personalization #2:
Query-Document-User Features

Three views of a person's history: Historic, session, aggregate

Features per view

- Cosine between topic categories of document and a search history view
- Cosine between topic categories of document and matching queries (and subsets, and supersets)
  - 'deep neural networks' subset: 'neural networks'
  - 'deep neural networks' superset: 'deep neural networks toolkits'
- url click count
- url click counts for matching queries (and subsets and supersets)

(Bennett, et al., 2012)

Personalization #2:
Query Features

Query features

- **Ambiguity measures:** Click entropy, topic entropy
  - How much do people click on different pages or topics for this query?
  - Higher entropy means more disagreement among users
    - » E.g., people agree about "Kim Kardashian"
    - » E.g., people disagree about "healthy diets"
- **Difficulty measures:** Position in session, length, frequency
- **Document rank (not personalized)**

(Bennett, et al., 2012)

Personalization #2:
Query History Features

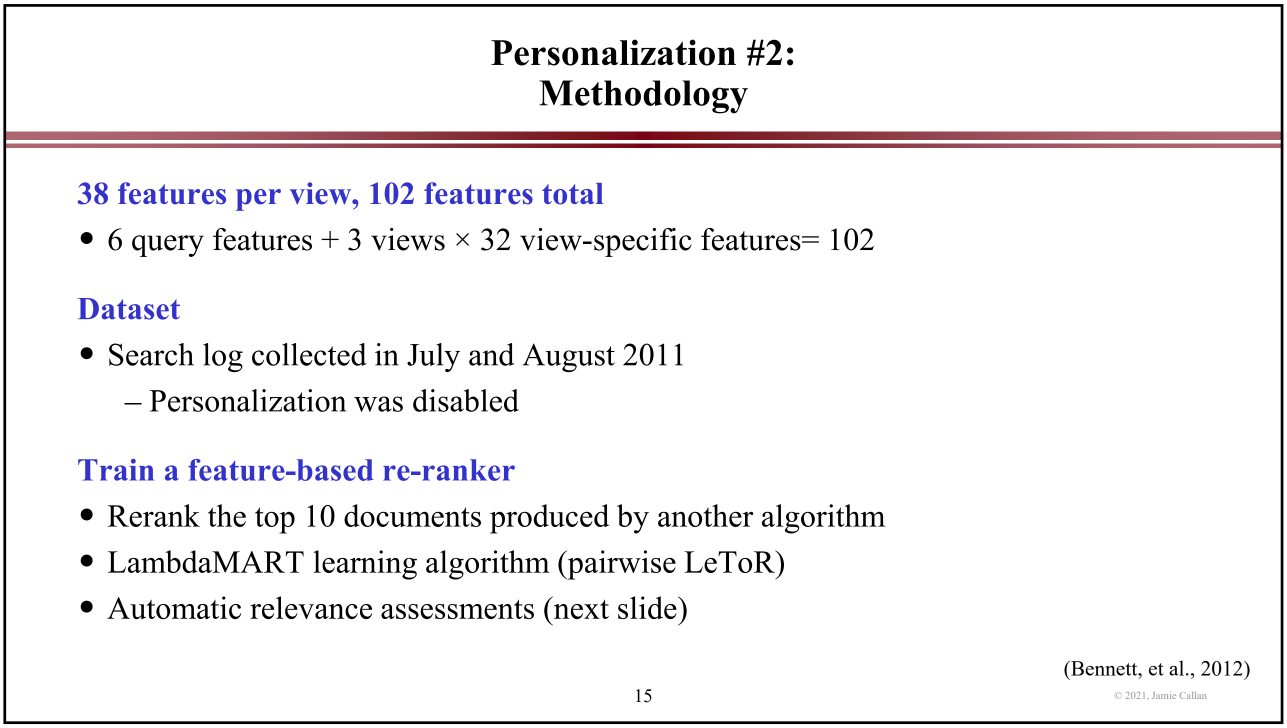
Features per view

- Number of queries
- Number of sessions with this query
- Number of subset queries
- Number of superset queries

Focus of user profile

- User topic entropy
- User query (and subset and superset) entropy
- User position entropy, user query position entropy

(Bennett, et al., 2012)



Personalization #2: Methodology

38 features per view, 102 features total

- 6 query features + 3 views × 32 view-specific features= 102

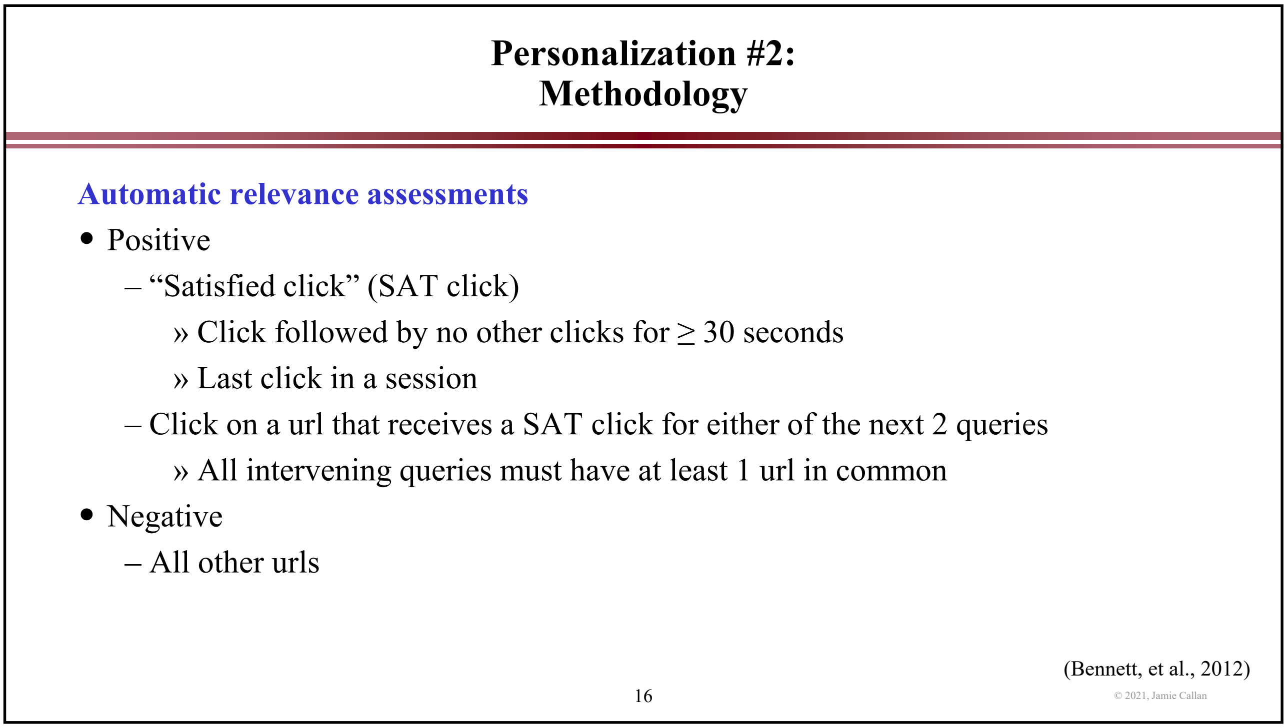
Dataset

- Search log collected in July and August 2011
  - Personalization was disabled

Train a feature-based re-ranker

- Rerank the top 10 documents produced by another algorithm
- LambdaMART learning algorithm (pairwise LeToR)
- Automatic relevance assessments (next slide)

(Bennett, et al., 2012)



Personalization #2: Methodology

Automatic relevance assessments

- Positive
  - "Satisfied click" (SAT click)
    - » Click followed by no other clicks for ≥ 30 seconds
    - » Last click in a session
  - Click on a url that receives a SAT click for either of the next 2 queries
    - » All intervening queries must have at least 1 url in common
- Negative
  - All other urls

(Bennett, et al., 2012)

Personalization #2: Methodology

Conditions

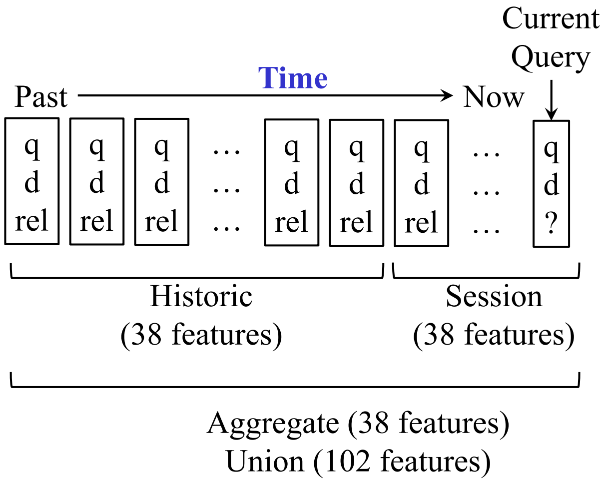
- **Session:** Current session only (6 + 32 = 38 features)
- **Historic:** Everything except the current session (6 + 32 = 38 features)
- **Aggregate:** Everything prior to the current query (6 + 32 = 38 features)
- **Union:** Session ∪ Historic ∪ Aggregate (6 + 32 + 32 + 32 = 102 features)

Consider only queries where MAP@10 changes ($\delta_{MAP@10} \neq 0$)

- They considered all queries, which dampens the effect predictably (so I'm not showing those results)

(Bennett, et al., 2012)

Personalization #2: Training Data



Historic (38 features)

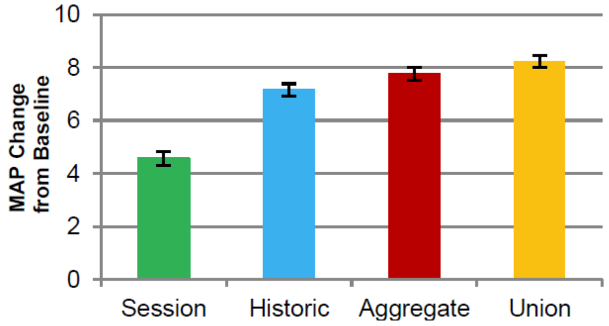
Session (38 features)

Aggregate (38 features)

Union (102 features)

Personalization #2:
The Value of Each View

What is the value of each view?

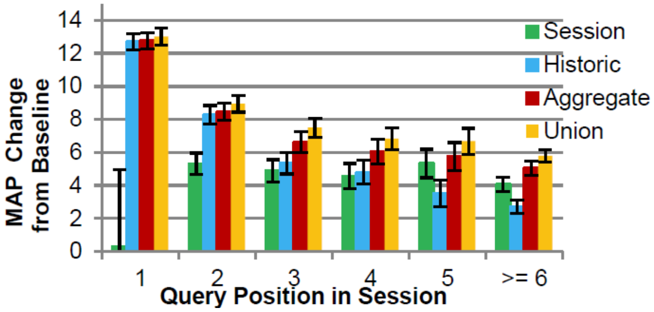


Best case: When ranker can do differential weighting of views

(Bennett, et al., 2012)

Personalization #2:
The Value of Each View Evolves

What is the value of each view throughout the session?

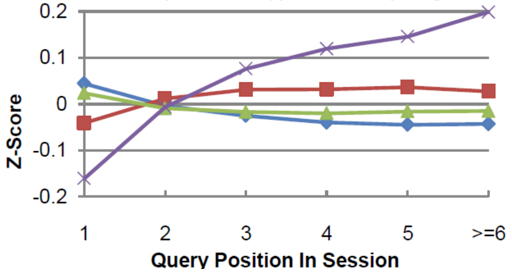


- Historic information provides the most gain early in the session
- Session information provides most of the gain late in the session
- Personalization has less of an effect late in a session
  - – User queries are more well-developed

(Bennett, et al., 2012)

Personalization #2:
How Queries Change in a Session

How do queries change throughout the session?

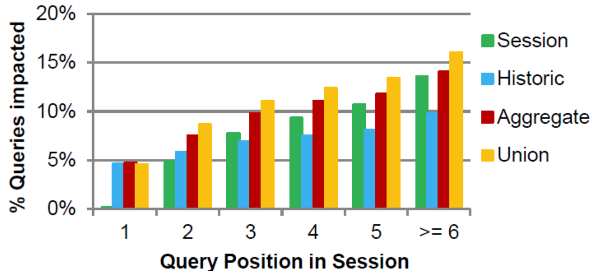


- Initial queries are short and ambiguous
  - Click entropy at position 1 is biased by navigational queries
- Later queries are longer and more specific

(Bennett, et al., 2012)

Personalization #2:
The Effects of Personalization in a Session

What is the effect of personalization throughout the session?



- Personalization affects <u>more queries</u> later in the session
  - Even for the historic method
  - Perhaps longer sessions pertain to this user's <u>typical</u> interests?

(Bennett, et al., 2012)

Lecture Outline

Three approaches to personalization

- Topic-based personalization
- Long-term vs. short-term personalization
- Personalization for typical vs. atypical information needs

Personalization #3

Most personalization techniques assume that a person is represented by a user profile that changes slowly

- E.g., the method just discussed

These techniques may not work well when a person searches for atypical information

- **Atypical:** Not typical (for this individual)
- E.g., a sudden medical problem, a gift, a vacation, …

Is this a serious problem?

(Eickhoff, et al., 2013)

Personalization #3

Dataset

- 4 months of Bing English query log data
- 200 active users (not a huge population)
  - – 380K queries in 44K sessions
  - – An average of 8.4 queries/session
  - – An average of 1.8 sessions/day
  - – These were not all their queries – just the queries used
- 30-minute session limit
- Discard navigational sessions (proprietary classifier)

(Eickhoff, et al., 2013)

Personalization #3:
Is the Session Atypical?

Crowd-sourced typicality labels were created for Month 4 sessions

- 5 point Likert scale
- Average results of 5 workers per session

(Eickhoff, et al., 2013)

Personalization #3:
Is the Session Atypical?

Crowd workers saw a user profile based on Month 3

- The most common ODP categories (e.g., 4 in this example)
- In each category, the 3 most frequent queries and their difficulty

55% Sports/Baseball:
"ncaa baseball", **"ectb baseball"**, **"pg baseball"**
14% Society/Religion and Spirituality:
"pope benedict bio", **"shamanistic travel"**, **"sacred heart newton"**
5% Reference/Education
"matlab student version", **"umass email"**, **"my math lab"**
5% Sports/Hockey
"elmira pioneers", **"umass lax"**, **"necbl"**

(color indicates query difficulty: **easy**, **medium**, **hard**)

(Eickhoff, et al., 2013)

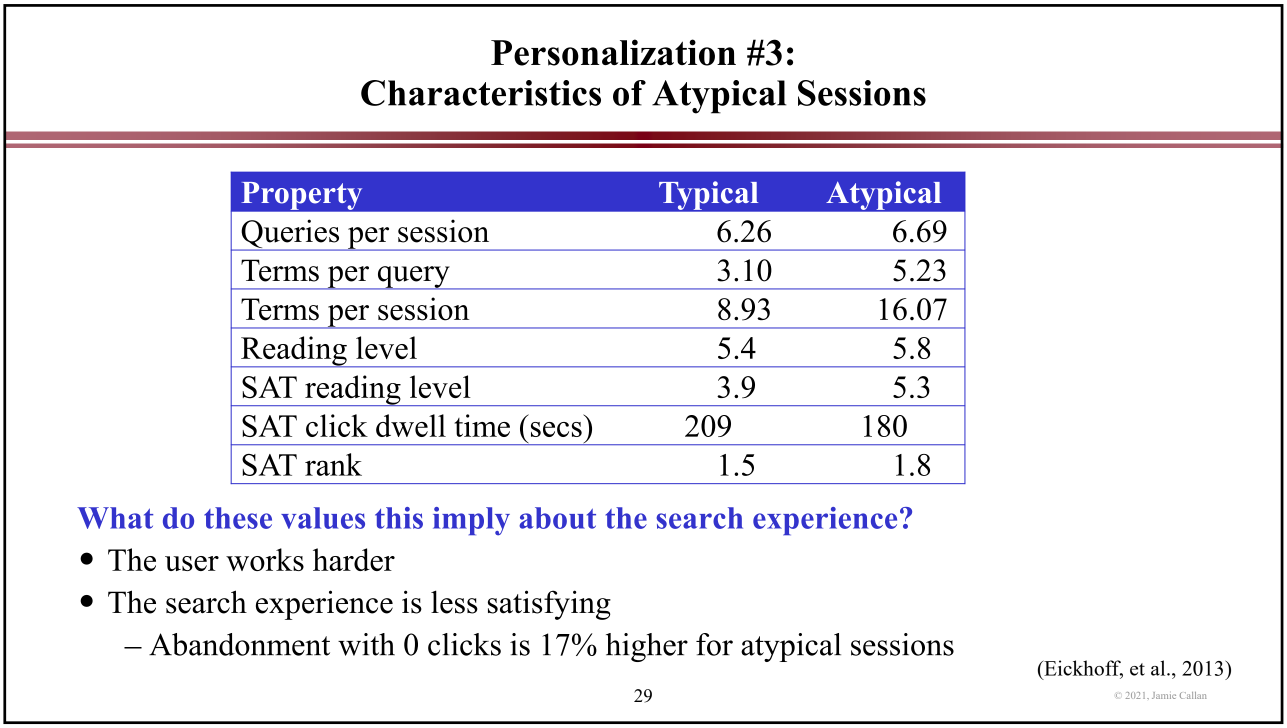
Personalization #3:
Is the Session Atypical?

6% of sessions were labeled atypical

- 74% of users had at least one atypical session in Month 4
- 7.5% of a person's monthly queries were from atypical sessions
  - – But, significant variation across users

Atypical sessions are not typical, but also not rare

(Eickhoff, et al., 2013)



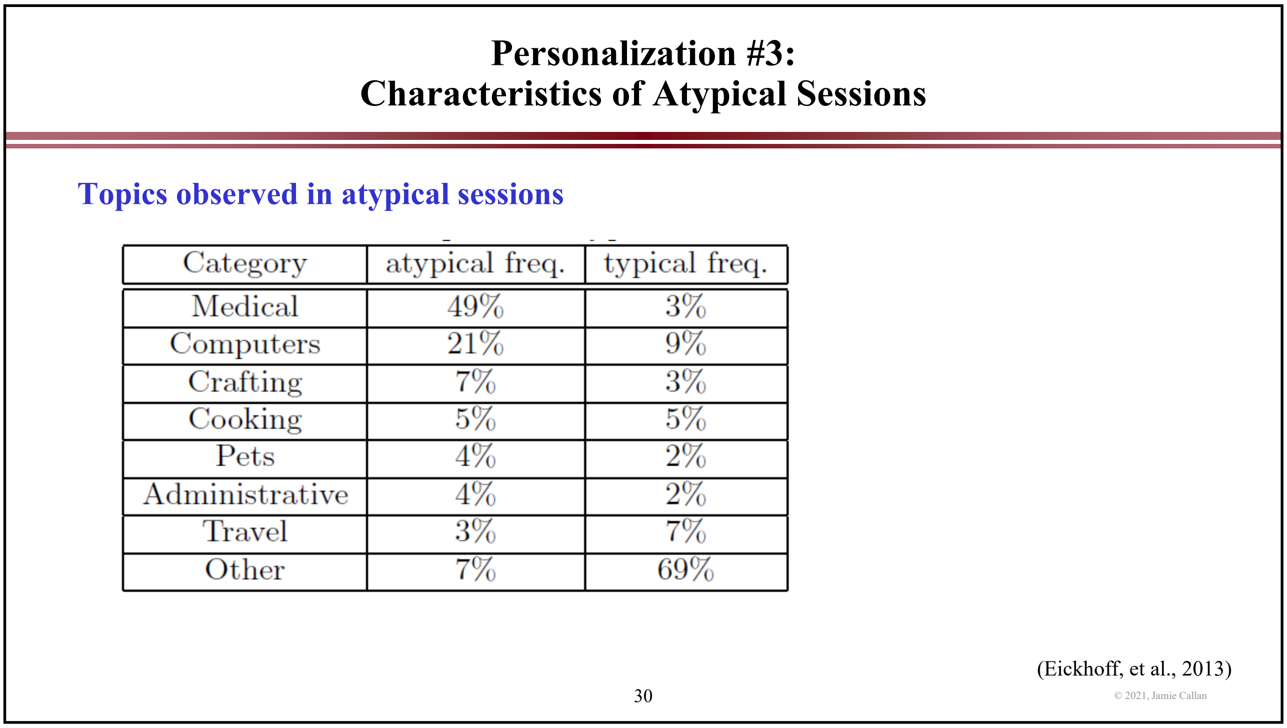
Personalization #3: Characteristics of Atypical Sessions

Property	Typical	Atypical
Queries per session	6.26	6.69
Terms per query	3.10	5.23
Terms per session	8.93	16.07
Reading level	5.4	5.8
SAT reading level	3.9	5.3
SAT click dwell time (secs)	209	180
SAT rank	1.5	1.8

What do these values this imply about the search experience?

- The user works harder
- The search experience is less satisfying
  - Abandonment with 0 clicks is 17% higher for atypical sessions

(Eickhoff, et al., 2013)



Personalization #3: Characteristics of Atypical Sessions

Topics observed in atypical sessions

Category	atypical freq.	typical freq.
Medical	49%	3%
Computers	21%	9%
Crafting	7%	3%
Cooking	5%	5%
Pets	4%	2%
Administrative	4%	2%
Travel	3%	7%
Other	7%	69%

(Eickhoff, et al., 2013)

Personalization #3:
Detecting Atypical Sessions

Detection of atypical sessions in a search log

- Build long-term profiles for each user
- Measure divergence between a person's long-term profile and the current session

(Eickhoff, et al., 2013)

Personalization #3:
Session features

Features calculated across all queries in a session

- Session length, avg query length, unique terms/session
- Ratio of queries that appear to contain a question word
- Advanced operator ratio, position of longest query
- Query part-of-speech (POS) ratios
- Clicks/query, SAT clicks/query, SAT click ratio, median SAT click rank, SAT click dwell time
- Avg reading level, avg SAT clicked reading level
- Indicators for the 7 topic categories shown earlier (medical, …)
- Unique topics (in clicked documents) per session

(Eickhoff, et al., 2013)

Personalization #3:
Session features

Query log observations

- Many atypical sessions contain natural language questions
  - – Measure % of queries per session that contain ‘who’, ‘what’, ‘where’, ‘when’, ‘why’, and ‘how’
  - – Measure relative frequencies of nouns, verbs, adjectives, misc
- People struggling are more likely to use AND, OR, NOT, and “ ”
- Success is more likely if the last query in the session is the longest
- Exploratory sessions tend to be more diverse (cover more topics)

(Eickhoff, et al., 2013)

Personalization #3:
Divergence

Divergence is measured in several ways

- Divergence of each session <u>feature</u> from this user’s historical norms
- Cosine distance between session and historical <u>vocabularies</u>
- Cosine distance between session and historical <u>topic categories</u>

(Eickhoff, et al., 2013)

Personalization #3:
Session features

Most informative 10 features (out of 34 features total)

Feature	Rank by IG	Rank by χ^2
query length divergence	1	1
query length	2	2
question ratio	3	4
verb ratio divergence	4	3
topic divergence	5	5
longest query position	6	8
SAT RL	7	6
SAT RL divergence	8	7
adjective ratio divergence	9	9
noun ratio	10	10

RL: Reading Level

- **7 features use only query information (1-5, 9-10)**
- **3 features use interaction with documents (6-8)**

(Eickhoff, et al., 2013)

Personalization #3:
Detecting Atypical Sessions

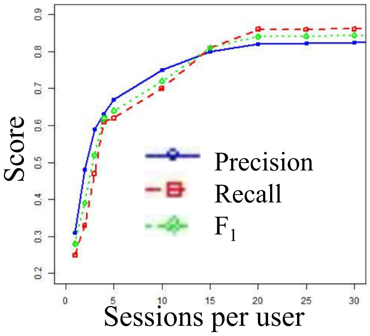
Classifier: Logistic regression

Accuracy on unseen data: P=0.80, R=0.68, F_1=0.74

- Comparable to human assessors matching a majority vote label

How much training data is required?

- About 20 sessions per user
  - About 14 days for most users
- More data didn't help
- **Caveat:** This is a small-scale study



(Eickhoff, et al., 2013)

Personalization #3:
Typical vs. Atypical Personalization

Prior research showed that personalization is most effective when using session and historic information ("aggregate")

Is this true for atypical sessions?

(Eickhoff, et al., 2013)

Personalization #3:
Detecting Atypical Sessions

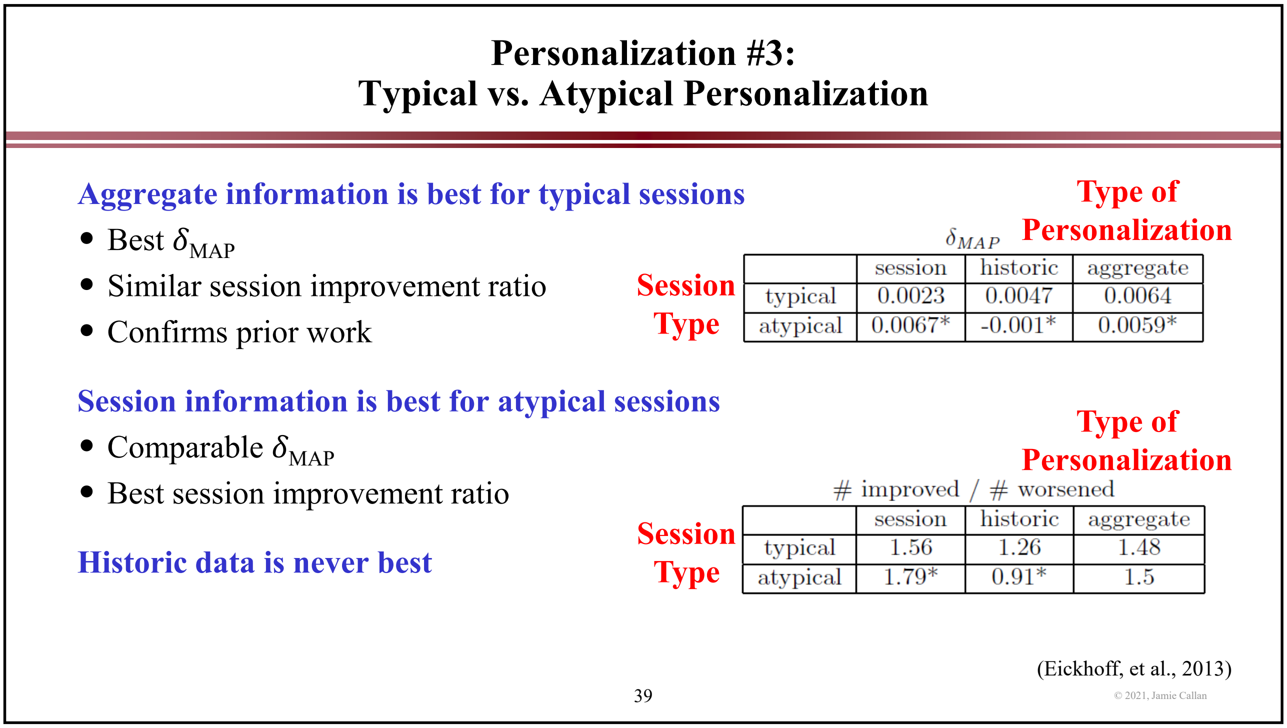
Dataset

- Search log collected in July and August 2011 (same dataset as Personalization #2)
- 155,000 unique users
- 10.4 million sessions
- An average of 174 queries / user

Train a feature-based re-ranker

- Rerank the top 10 documents produced by another algorithm
- LambdaMART learning algorithm
- Features mentioned earlier
- SAT clicked documents were treated as relevant

(Eickhoff, et al., 2013)



Personalization #3:
Typical vs. Atypical Personalization

Aggregate information is best for typical sessions

- Best δ_{MAP}
- Similar session improvement ratio
- Confirms prior work

Type of Personalization

δ_{MAP}

Session Type

	session	historic	aggregate
typical	0.0023	0.0047	0.0064
atypical	0.0067*	-0.001*	0.0059*

Session information is best for atypical sessions

- Comparable δ_{MAP}
- Best session improvement ratio

Historic data is never best

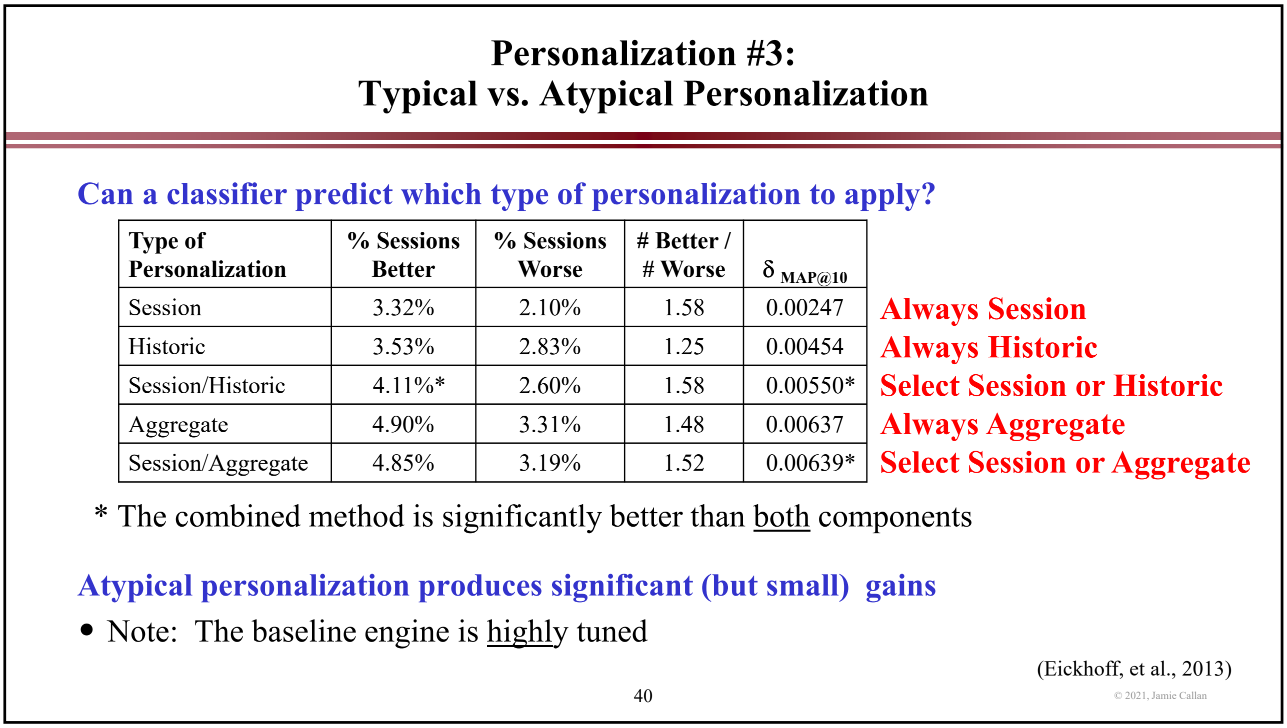
Type of Personalization

improved / # worsened

Session Type

	session	historic	aggregate
typical	1.56	1.26	1.48
atypical	1.79*	0.91*	1.5

(Eickhoff, et al., 2013)



Personalization #3:
Typical vs. Atypical Personalization

Can a classifier predict which type of personalization to apply?

Type of Personalization	**% Sessions Better**	**% Sessions Worse**	**# Better / # Worse**	**$\delta_{\text{MAP@10}}$**	
Session	3.32%	2.10%	1.58	0.00247	**Always Session**
Historic	3.53%	2.83%	1.25	0.00454	**Always Historic**
Session/Historic	4.11%*	2.60%	1.58	0.00550*	**Select Session or Historic**
Aggregate	4.90%	3.31%	1.48	0.00637	**Always Aggregate**
Session/Aggregate	4.85%	3.19%	1.52	0.00639*	**Select Session or Aggregate**

* The combined method is significantly better than both components

Atypical personalization produces significant (but small) gains

- Note: The baseline engine is highly tuned

(Eickhoff, et al., 2013)

Lecture Outline

Three approaches to personalization

- Topic-based personalization
- Long-term vs. short-term personalization
- Personalization for typical vs. atypical information needs

All of this work was done at Microsoft Research

Personalization is an active area of research

Our goals

- Get a sense of what is being done, and how it is being done

For More Information

- P.N. Bennett, R.W. White, W. Chu, S.T. Dumais, P. Bailey, F. Borisyuk, and X. Cui. Modeling the impact of short-and long-term behavior on search personalization. In Proceedings of SIGIR 2012. 2012.
- C. Eickhoff, K. Collins-Thompson, P. N. Bennett, and S. Dumais. Personalizing atypical web search sessions. In Proceedings of WSDM '13. 2013.
- D. Songtag, K. Collins-Thompson, P.N. Bennett, R.W. White, S. Dumais, and B. Billerbeck. "Probabilistic models for personalizing web search." WSDM. 2012.