4. Personalization

Outline

4.1. Objectives

4.2. Concerns

4.3. Potential

4.4. Link Analysis

4.5. Query Expansion

4.6. Retrieval Model

4.7. Re-Ranking

1. Objectives

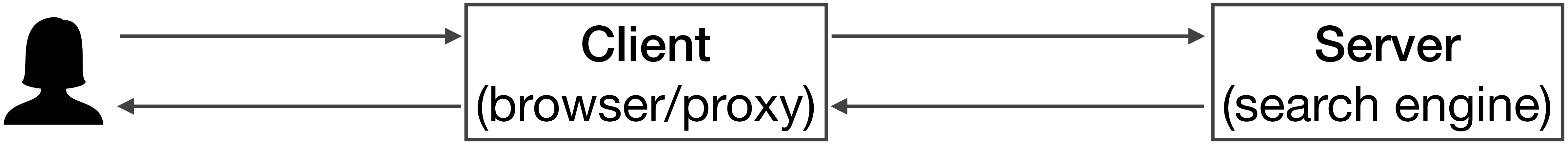
- Our focus will be on **web search**; personalization also affects other applications (e.g., recommender systems, advertising)

- Personalization can serve **different objectives** in web search
  - **disambiguate** the query based on user profile (e.g., jaguar)
  - **adapt** query results to the user profile or abilities (e.g., reading level)
  - **localize** results based on the user location (e.g., uds, coffee shop)

Data Sources

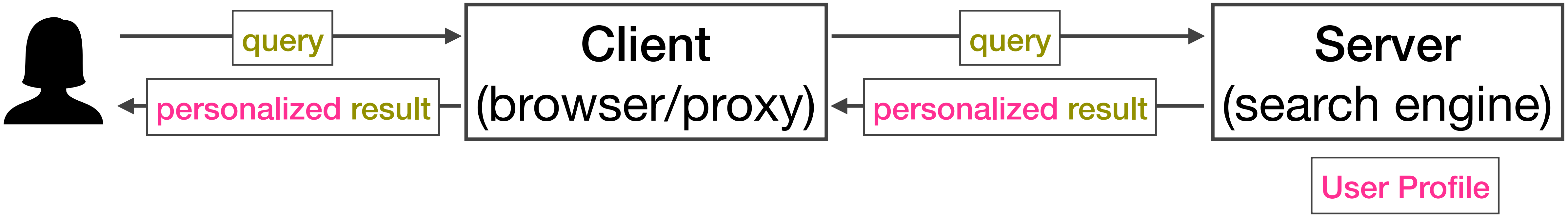
- Search results can be personalized using **different data sources**
  - **Feedback** (e.g., about relevance of search results)
  - **Traits** (e.g., age, gender, income level, education level, religion)
  - **Social profiles** (e.g., likes on facebook, tweets)
  - **Behavior** (e.g., short/long-time browsing, search, and click histories)
  - **Desktop** (e.g., office documents, e-mail)

Client vs. Server



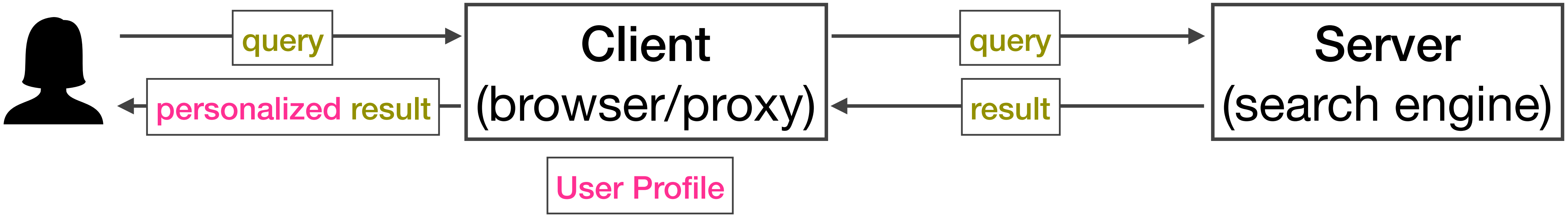
- Search results can be personalized in **different locations** [12]
  - **Server:** the search engine knows the user profile and personalizes the search result according to it
  - **Client:** only the client knows the user profile and personalizes the generic result from the search engine according to it
  - **Client-Server Cooperation:** the client knows the user profile and reveals parts of it to the search engine to personalize the result

Client vs. Server



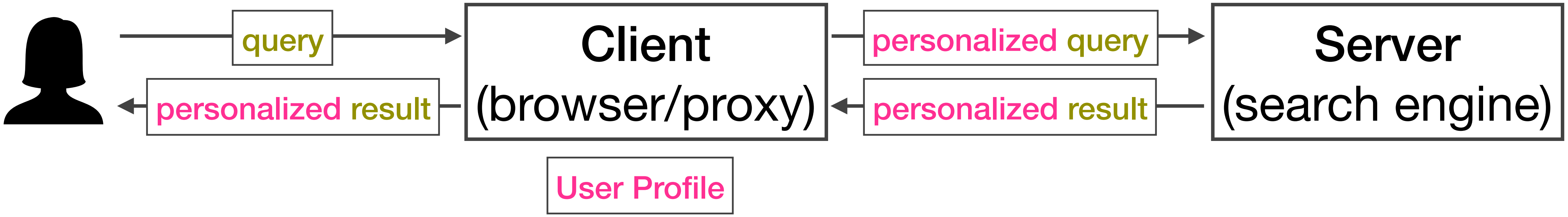
- Search results can be personalized in **different locations** [12]
  - **Server:** the search engine knows the user profile and personalizes the search result according to it
  - **Client:** only the client knows the user profile and personalizes the generic result from the search engine according to it
  - **Client-Server Cooperation:** the client knows the user profile and reveals parts of it to the search engine to personalize the result

Client vs. Server



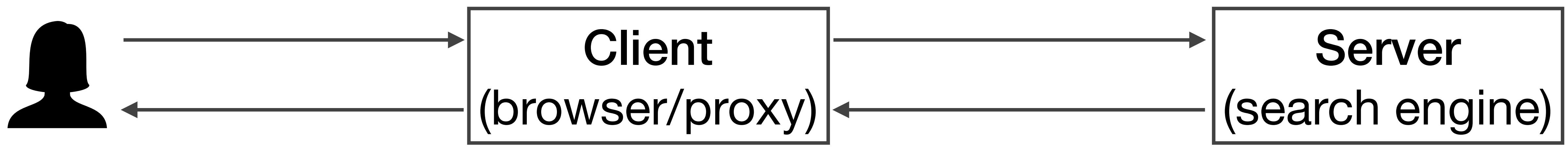
- Search results can be personalized in **different locations** [12]
  - **Server:** the search engine knows the user profile and personalizes the search result according to it
  - **Client:** only the client knows the user profile and personalizes the generic result from the search engine according to it
  - **Client-Server Cooperation:** the client knows the user profile and reveals parts of it to the search engine to personalize the result

Client vs. Server



- Search results can be personalized in **different locations** [12]
  - **Server:** the search engine knows the user profile and personalizes the search result according to it
  - **Client:** only the client knows the user profile and personalizes the generic result from the search engine according to it
  - **Client-Server Cooperation:** the client knows the user profile and reveals parts of it to the search engine to personalize the result

Client vs. Server



- Search results can be personalized in **different locations** [12]
  - **Server:** the search engine knows the user profile and personalizes the search result according to it
  - **Client:** only the client knows the user profile and personalizes the generic result from the search engine according to it
  - **Client-Server Cooperation:** the client knows the user profile and reveals parts of it to the search engine to personalize the result

Methods

- Search results can be personalized using **different methods**
  - **Link analysis:** by computing a **user-specific static score** for each web page, reflecting its importance relative to the user profile
  - **Query expansion:** by **augmenting the query** with terms from the user profile to disambiguate it and inform the search engine
  - **Retrieval model:** by **directly considering** the user profile when deciding which documents to return as results and how to order them
  - **Re-ranking:** by considering the **generic results** returned by the search engine and **re-ranking** them considering the user profile

2. Concerns

- Personalization of search results requires **data about the user**
  - **personal traits** (e.g., gender, age, income level)
  - **search, click, or browsing histories**
- **Privacy is a concern** in the post-Snowden era

- Personalization of search results **can affect users and society**
  - by not exposing users to **views different from their own**
  - by only showing results fitting the **user's interests, location, intellect**
- **Filter bubble** is a concern regarding the effects of personalization

Privacy

- Shen et al. [10] study the **tension** between **privacy preservation** and **personalization** and define four levels of privacy protection
  - **Level 1: Pseudo Identity**
    (user identity is replaced by an identifier in the search system)
  - **Level 2: Group Identity**
    (multiple users share a single user identifier in the search system)
  - **Level 3: No Identity**
    (search system does not know the user identity)
  - **Level 4: No Personal Information**
    (search system does not know any personal information)

How Much Do They Know?

"*You have zero privacy anyway. Get over it.*"

(Scott McNealy, former CEO of Sun Microsystems)

- Bi et al. [1] examine to what extent a **user's demographics** can be **inferred** purely based on the **search queries she issues**
- myPersonality.org data provides the **Facebook likes** of **millions of anonymous users** together with their **demographic profiles**
- **Open Directory Project** (DMOZ.org) as **common representation** for liked entities on Facebook and queries issued by users

How Much Do They Know?

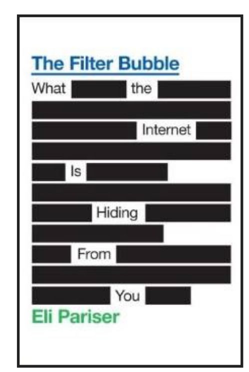
- Bing users as probability distributions over ODP topics
- Probability distributions over ODP topics for traits from Facebook

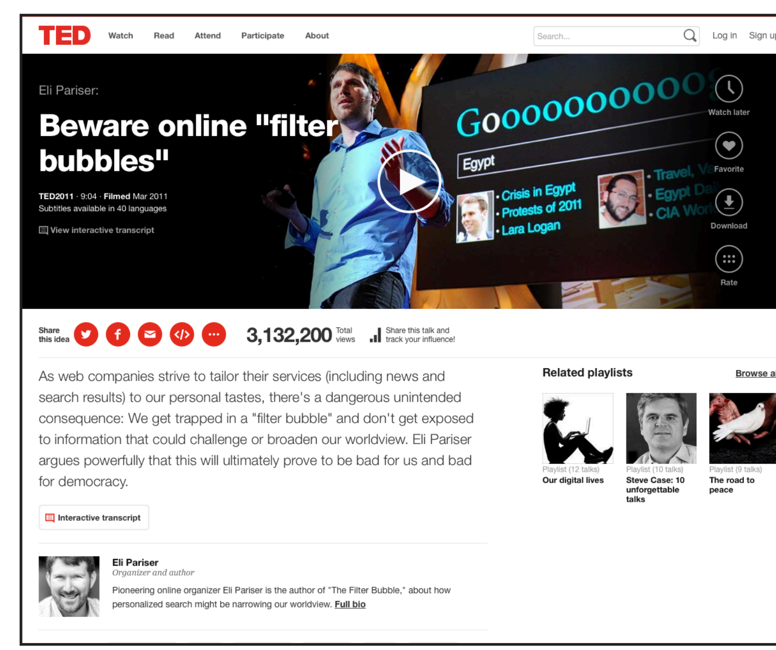
- Results: AUC (Area Under receiver operating characteristic Curve)
  - **0.803** for predicting **gender** based on queries issued
  - **0.735** for predicting **age** based on queries issued

Filter Bubble

- Eli Pariser [9] coined the notion "filter bubble", observing that personalization traps users by increasingly exposing them to content that is in line with what they know or believe

- <u>Examples</u>:
  - Query "egypt" brings up only tourism-related results, but none related to political situation
  - Query "bp" brings up stock-related results, but none related to oil spill





[TED talk]

Is the Filter Bubble Real?

- Hannak et al. [4] conducted a study with 200 Google users to measure the **degree of personalization** and identify **personal features** with an impact on search results
  - **120 queries** from Google Zeitgeist and WebMD (tech, news, etc.)
  - **200 users** from 43 different U.S. states recruited via **Mechanical Turk**
  - **scripted issuing of queries** through **HTTP proxy**

- Observations:
  - **extensive personalization** (at lower ranks)
  - **most personalized queries** related to **companies/stores** (localization)

Most Personalized	Least Personalized
gap	what is gout
hollister	dance with dragons
hgtv	what is lupus
boomerang	gila monster facts
home depot	what is gluten
greece	ipad 2
pottery barn	cheri daniels
human rights	psoriatic arthritis
h2o	keurig coffee maker
nike	maytag refrigerator

Is the Filter Bubble Real?

- To identify personal features that impact search results, Hannak et al. [4] **created different Google profiles** and compared results
  - **logged in / not logged in / cookies cleared** (little impact)
  - **browser user-agent** (no impact)
  - **geolocation from IP address** (big impact)
  - **gender** (no impact)
  - **search history** (no impact)
  - **click history** (no impact)
  - **browsing history** (no impact)

3. Potential

- <u>Question</u>: *How much can be gained, in terms of retrieval performance, by personalizing web search results?*

- Teevan et al. [11] estimate the **potential for personalization** (in terms of nDCG) using three kinds of data sources
  - **explicit relevance feedback** from 125 users on 699 queries (gain value {0, 1, 2} derived from graded relevance judgment)
  - **desktop data** of 59 users as **implicit feedback** on 822 queries (gain value [0, 1] based on cosine similarity to desktop)
  - **click logs** of 1.5 M users as **implicit feedback** on 2.4 M queries (gain value {0, 1} based on whether user clicked on result)

Potential for Personalization

- Given **feedback from an individual user**, we can determine the **optimal result** for her and **how much worse** the web result is

Potential for Personalization

- Given **feedback from an individual user**, we can determine the **optimal result** for her and **how much worse** the web result is

Result

d_2

d_1

d_4

d_3

d_5

Potential for Personalization

- Given **feedback from an individual user**, we can determine the **optimal result** for her and **how much worse** the web result is

Result	Feedback
d_2	0
d_1	2
d_4	0
d_3	1
d_5	1

Potential for Personalization

- Given **feedback from an individual user**, we can determine the **optimal result** for her and **how much worse** the web result is

Result	Feedback	Optimal Result
d_2	0	d_1
d_1	2	d_3
d_4	0	d_5
d_3	1	d_2
d_5	1	d_4

Potential for Personalization

- Given **feedback from an individual user**, we can determine the **optimal result** for her and **how much worse** the web result is

Result	Feedback	Optimal Result
d_2	0	d_1
d_1	2	d_3
d_4	0	d_5
d_3	1	d_2
d_5	1	d_4

nDCG: 1.0

Potential for Personalization

- Given **feedback from an individual user**, we can determine the **optimal result** for her and **how much worse** the web result is

Result	Feedback	Optimal Result
d_2	0	d_1
d_1	2	d_3
d_4	0	d_5
d_3	1	d_2
d_5	1	d_4

nDCG: 0.79

nDCG: 1.0

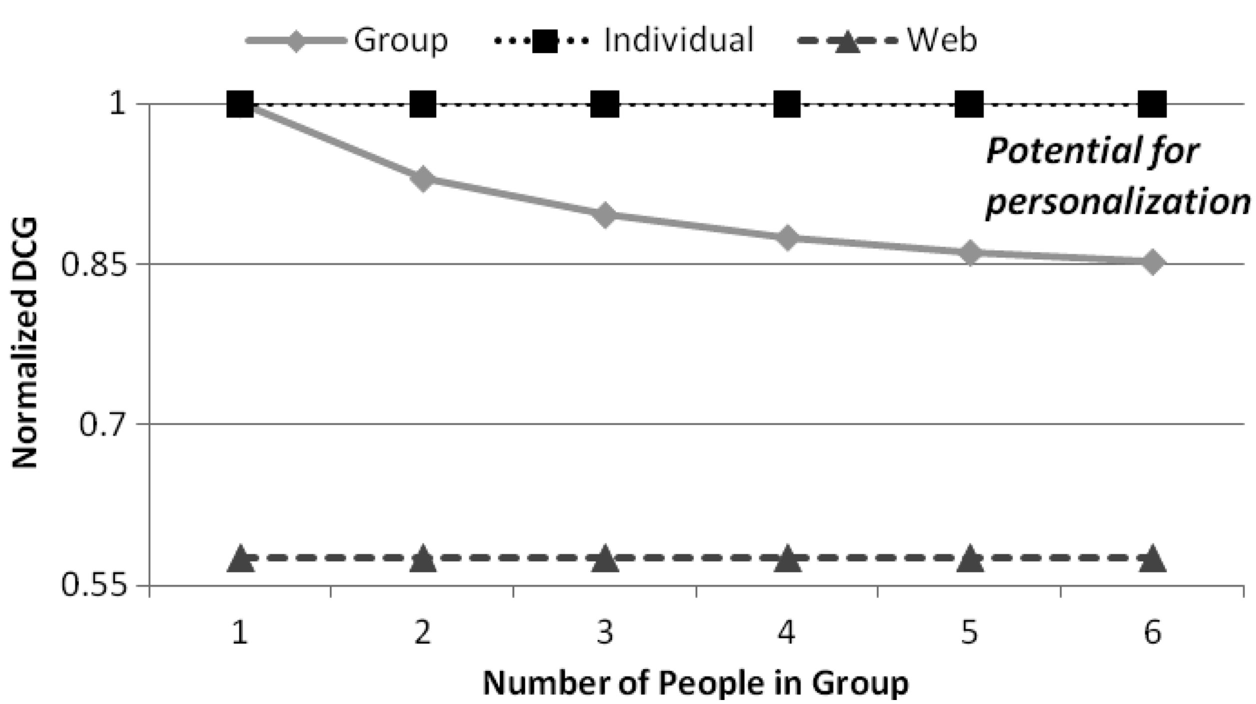
Potential for Personalization

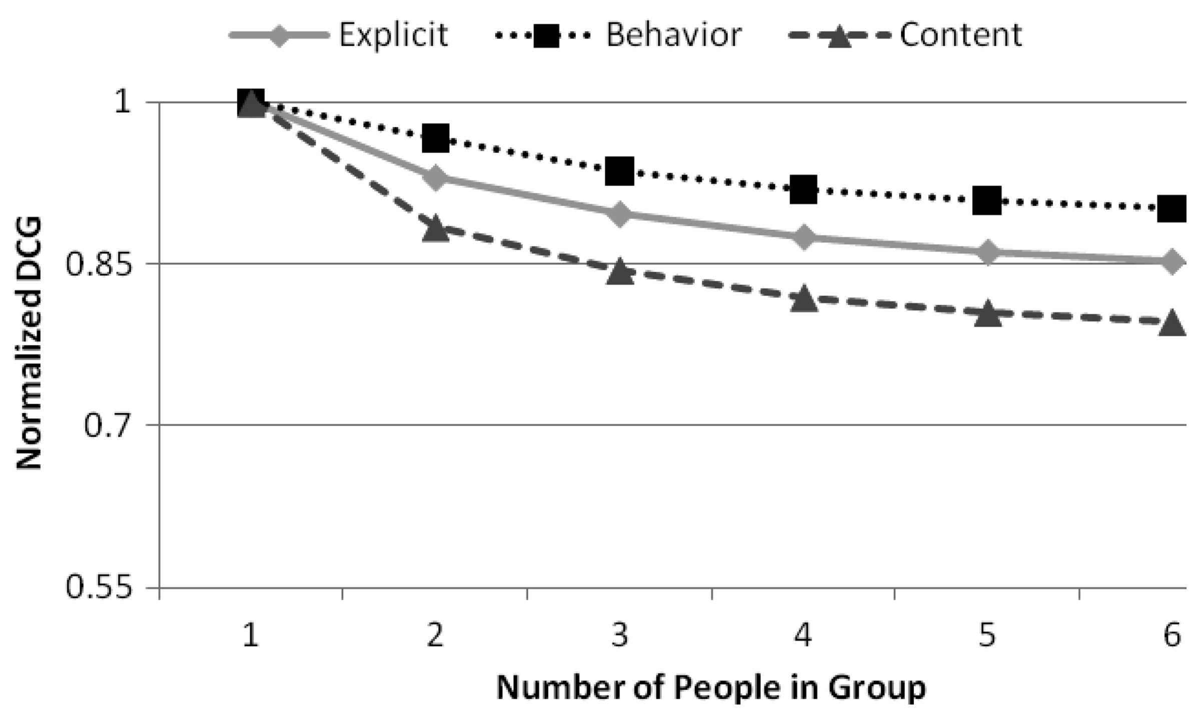
- Given **feedback from an individual user**, we can determine the **optimal result** for her and **how much worse** the web result is

Result	Feedback	Optimal Result
d_2	0	d_1
d_1	2	d_3
d_4	0	d_5
d_3	1	d_2
d_5	1	d_4

nDCG: 0.79 ⟵ Potential for Personalization ⟶ nDCG: 1.0

Potential for Personalization





- **Explicit relevance feedback**
    - Personalized result (nDCG 1.0)
    - Result for group of six (nDCG 0.85)
    - Web result (nDCG 0.58)

- **Potential for personalization**
    - smallest for click logs (behavior)
    - largest for desktop data (content)

Potential for Personalization

- Mei and Church [7] make use of **information theory** to estimate **how hard web search is** and **how much personalization helps**

Query (e.g., fb), **URL** (e.g., http://www.fb.com), **IP** (e.g., 139.19.54.9)

- Data: Click log from the Microsoft Live search engine (now: Bing)
  - **18 months** (until July 2007)
  - **193 M** unique **IP addresses** (users)
  - **637 M** unique **queries**
  - **585 M** unique **URLs**

Entropy

- **Entropy** measures the **degree of uncertainty** of a random variable X, thereby characterizing the size of the **search space**

$$H(X) = - \sum_x \mathrm{P}\left[\, x \,\right] \log \mathrm{P}\left[\, x \,\right]$$

- Example: Dice with six faces having uniform probability

$$H(D) \approx 2.58$$

Size of search space: 6

- Example: Dice with six faces; 1 has probability 0.8; others 0.04

$$H(D) \approx 1.19$$

Size of search space: 2.28

Conditional Entropy

- **Conditional entropy** measures the **remaining uncertainty** of a random variable **X** given the value of another random variable **Y**

$$H(X|Y) = H(X,Y) - H(Y)$$

- Example: Dice with six colored faces having uniform probability

1 2 3 4 5 6

Consider **N** = {even, odd} and **C** = {black, white}

$$H(N) = 1 \qquad H(C) \approx 0.92$$

$$H(N,C) \approx 1.46 \quad H(N|C) \approx 0.54$$

How Hard is Web Search?

- Given a click log, one can now estimate **how hard search** is as

$$H(URL|Query)$$

- Mei and Church [7] observe the following (conditional) entropies

$$H(URL|Query) \approx 3.5$$

$$H(URL, Query) \approx 26.41 \qquad H(Query) \approx 22.94$$

How Much does Personalization Help?

- Assuming that **IPs correspond to individuals**, we can estimate how **much easier search becomes** once the IP is known

$$H(URL|Query, IP) \approx 1.26$$

$$H(URL, Query, IP) \approx 31.67 \quad H(Query, IP) \approx 30.41$$

- Personalization reduces the size of the search space from about 11.31 to 2.39 (reflecting how many results users typically have to inspect)

4. Link Analysis

- Search results can be personalized by computing a **user-specific static score** for every web page that reflects its **importance relative to the user profile**

- Recap: PageRank (as part of the original Google search engine) operates on the **web graph** G(V, E) consisting of **web pages** (V) and **hyperlinks** (E)

$$r(v) = (1 - \epsilon) \sum_{(u,v) \in E} \frac{r(u)}{\text{out}(u)} + \frac{\epsilon}{|V|}$$

- PageRank models a **random surfer** who **follows random hyperlink** with probability (1 - ε) and **jumps to random web page** with probability ε

PageRank

- **PageRank scores** correspond to the **stationary state probabilities** of an **ergodic Markov chain** with **transition probability matrix** P

$$\mathbf{P} = (1-\epsilon)\,\mathbf{T} \; + \; \epsilon\,\mathbf{J}$$

with matrix **T** capturing **hyperlink following** as

$$\mathbf{T}_{ij} = \begin{cases} 1/\,\mathrm{out}(i) & : \quad (i,j) \in E \\ 0 & : \quad \text{otherwise} \end{cases}$$

and matrix **J** capturing **random jumps** as

$$\mathbf{J} = \begin{bmatrix} 1 & \ldots & 1 \end{bmatrix}^T \times \mathbf{j}$$

with **random jump vector j** as

$$\mathbf{j} = \begin{bmatrix} 1/|V| & \ldots & 1/|V| \end{bmatrix}$$

Power-Iteration Method

- **Power-iteration method** to compute PageRank vectors
  - **initialize** $$\pi^{(0)} = \begin{bmatrix} 1/|V| & \ldots & 1/|V| \end{bmatrix}$$
  - **repeat** $$\pi^{(i)} = \pi^{(i-1)} \times \mathbf{P}$$
  - **until convergence** $$\left| \pi^{(i)} - \pi^{(i-1)} \right| < \delta$$

Personalized PageRank

- Haveliwala [5] proposed a **topic-specific variant of PageRank** that performs **random jumps only to on-topic web pages**

- Let **C** ⊆ **V** be the web pages belonging to **topic C** (e.g., Sports), the random jump vector **j** is defined as

$$\mathbf{j}_i = \begin{cases} 1/|C| & : \quad i \in C \\ 0 & : \quad \text{otherwise} \end{cases}$$

- Web pages "closer" to on-topic web pages in **C** are favored

- **Personalized PageRank** considers a set of **user-specific favorite web pages** F as random jump targets

Personalized PageRank

- **Computing and storing personalized PageRank** scores for large numbers of users and/or web pages is **prohibitive**

- Jeh and Widom [6] discovered the **linearity of PageRank**
  - Let **j** and **j'** be **two random jump vectors** and **π** and **π'** be the **two corresponding PageRank vectors**, then

$$(\alpha\,\boldsymbol{\pi} + \beta\,\boldsymbol{\pi}') = (\alpha\,\boldsymbol{\pi} + \beta\,\boldsymbol{\pi}') \times \left((1-\epsilon)\,\mathbf{T} + \epsilon\begin{bmatrix}1 & \ldots & 1\end{bmatrix}^T \times (\alpha\,\mathbf{j} + \beta\,\mathbf{j}')\right)$$

- One can thus select a **small set of basis vectors**, compute the corresponding PageRank vectors, and obtain **user-specific PageRank scores as a linear combination** of them

5. Query Expansion

- Chirita et al. [2] personalize search results by augmenting the query with **terms selected from the user's desktop**

- **Local Desktop Analysis** issues the query locally against the user's desktop search engine and extracts terms from top-*k* pseudo-relevant documents, e.g., based on
  - **term frequency** (tf) or **document frequency** (df) (but not: tf.idf)
  - **dispersion analysis** (most frequent compounds: `adjective? noun+`)

Query Expansion

- **Global Desktop Analysis** precomputes term co-occurrence scores by analyzing documents from the user's desktop
  - **cosine similarity** $$\text{score}(a, b) = \frac{df(a \wedge b)}{\sqrt{df(a) \cdot df(b)}}$$
  - **mutual information** $$\text{score}(a, b) = \log \frac{|D| \cdot df(a \wedge b)}{df(a) \cdot df(b)}$$
- **Expansion terms** for a query **q** are then determined as those having the **highest aggregated score**

$$\text{agg\_score}(e) = \prod_{v \in \mathbf{q}} \text{score}(v, e)$$

- Experiments show significant improvement over baseline (Google) for ambiguous queries; but deterioration for clear queries

6. Retrieval Model

- Xue et al. [12] devise a **language modeling approach** to personalize results **based on what users have viewed**

- Let $V_{i,t}$ be documents that **user i** has viewed at **time t**, and let **nw** denote the current time period (e.g., day)

- **Short-term profile** for user **i** is estimated based on what the user has viewed within the last time period

$$P\left[\, v \,\middle|\, \theta_i^{st} \,\right] = \frac{\sum_{d \in V_{i,nw}} tf(v,d)}{\sum_{d \in V_{i,nw}} |d|}$$

User Model

- **Long-term profile** for user **i** is estimated based on what the user has viewed **within the last** h **time periods**

$$P\left[\, v \,\middle|\, \theta_i^{lt} \,\right] = \frac{\sum_{t=1}^{h} \sum_{d \in V_{i,nw-t}} tf(v,d) \cdot e^{-\rho\, t}}{\sum_{t=1}^{h} \sum_{d \in V_{i,nw-t}} |d| \cdot e^{-\rho\, t}}$$

  applying **exponential temporal decay** to give **lower weight** to what has been **viewed longer ago**

- **User language model** is then estimated as

$$P\left[\, v \,|\, \theta_i \,\right] = \beta\ P\left[\, w \,\middle|\, \theta_i^{st} \,\right] + (1-\beta)\ P\left[\, w \,\middle|\, \theta_i^{lt} \,\right]$$

Global Model

- **Global language model** for all users is obtained as

$$P[v \mid \theta_g] = \frac{1}{|U|} \sum_{i \in U} P[v \mid \theta_i]$$

with **U** as the set of **all users**

Group Model

- Users are **grouped into clusters** $c_1, \ldots, c_k$ based on the similarity of their user language models (e.g., using *k*-means with KLD)

- **Cluster language model** for cluster **c** is estimated as

$$P\left[\, v \mid \theta_c \,\right] = \frac{1}{|c|} \sum_{i \in c} P\left[\, v \mid \theta_i \,\right]$$

- For **query q** issued by **user i** identify a **single cluster c** as

$$\arg\min_{c} \left(\zeta \, KL(\theta_i \| \theta_c) + (1 - \zeta) \, KL(\theta_q \| \theta_c) \right)$$

and parameter ζ controlling **fit of cluster to user and/or query**

Combining the Models

- **Combined language model** to rank documents is estimated as

$$P[v \mid \theta] = \lambda \, P[v \mid \theta_q] + (1-\lambda)\left[\gamma \, P[v \mid \theta_i] + (1-\gamma)\left[\eta \, P[v \mid \theta_c] + (1-\eta) \, P[v \mid \theta_g]\right]\right]$$

  with **smoothing parameters** λ, γ, η controlling the influence of the **query, user, group, and global model**

- Experiments based on **click-through data** from 1,000 users of MSN search engine (now: Bing) and 50/50 split of queries

Model	NDCG1	NDCG5	NDCG10	NDCG20	NDCG30
q	0.422	0.434	0.441	0.416	0.384
$q+i$	0.664	0.655	0.613	0.535	0.467
$q+c$	0.724	0.674	0.635	0.515	0.438
$q+g$	0.672	0.667	0.626	0.546	0.497
$q+i+g$	0.707	0.674	0.641	0.556	0.474
$q+i+c$	0.712	0.675	0.64	**0.557**	0.474
$q+i+c+g$	**0.724**	**0.683**	**0.644**	0.555	**0.499**

7. Re-Ranking

- Matthijs and Radlinski [7] develop a **browser plug-in** that builds a **(local) user profile** which is then used to **re-rank Google search results** based on the information in their **snippets**

- **User profile** based on viewed web pages includes
  - unigrams from **full-text** (body) and **title**
  - unigrams from **meta-data fields** (description and keywords)
  - **extracted keywords** and **noun phrases**

- For each term v in the user profile, a **tf.idf weight** $w_{tf.idf}(v)$ is estimated with a document frequency from Google

Re-Ranking

- Given a query, the **search results returned by Google are re-ranked** taking into account the following factors
  - **matching score** between **search result title** and **user profile**

$$\mathrm{score_M}(r) = \prod_{v \,\in\, \mathtt{title}(r)} \log \frac{w_{tf.idf}(v) + 1}{\sum_{v'} w_{tf.idf}(v')}$$

  - **original rank** in Google result (logarithmically damped)

$$\mathrm{score_R}(r) = \frac{1}{1 + \log(\mathrm{rank}(r))}$$

  - **number of previous visits** to the URL

$$\mathrm{score_V}(r) = (1 + \alpha \cdot \mathrm{visits}(r))$$

  with tunable parameter α

Re-Ranking

- **Re-ranking Google top-50** results based on

$$\mathrm{score}_{\mathrm{M}}(r) \times \mathrm{score}_{\mathrm{R}}(r) \times \mathrm{score}_{\mathrm{V}}(r)$$

  improved nDCG from **0.502** to **0.573** (14%)
  in a user study with **six users** and **72 queries**

- While **relatively simple** the approach yields a **significant improvement** (p = 0.042) and can be **implemented locally** (i.e., without disclosing personal information)

Summary

- Search results are personalized to **resolve ambiguity**, **localize them**, or **adapt them** to the user's traits or interests
- Personalization can be achieved by leveraging **different data sources** including users traits, social media profiles, desktop
- **Privacy and filter bubble effects** are **serious concerns** regarding personalized search – with differing opinions
- **Potential impact of personalization** can be assessed through user studies or by observing their behavior at large scale
- Personalization of search results can be achieved using **different methods** including link analysis, retrieval models, and re-ranking

References

[1] **B. Bi, M. Shokouhi, M. Kosinki, T. Graepel:** *Inferring the Demographics of Search Users,* WWW 2013

[2] **P. A. Chirita, C. S. Firan, W. Nejdl:** *Personalized Query Expansion for the Web*, SIGIR 2007

[3] **M. R. Ghorab, D. Zhou, A. O'Connor, V. Wade:** *Personalised Information Retrieval: Survey and Classification*, UMUAI 23, 2012

[4] **A. Hannak, P. Sapiezynski, A. M. Kakhki, B. Krishnamurthy, D. Lazer, A. Mislove, C. Wilson:** *Measuring Personalization of Web Search*, WWW 2013

[5] **T. H. Haveliwala:** *Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search*, IEEE TKDE 15(4), 2003

[6] **G. Jeh and J. Widom:** *Scaling Personalized Web Search*, WWW 2003

[7] **N. Matthijs and F. Radlinski:** *Personalizing Web Search using Long Term Browsing History*, WSDM 2011

References

[8] **Q. Mei and K. Church:** *Entropy of Search Logs: How Hard is Search? With Personalization? With Backoff?*, WSDM 2008

[9] **E. Pariser:** *The Filter Bubble: What the Internet is Hiding from You,* Penguin Press, 2011

[10] **X. Shen, B. Tan, C. Zhai:** *Privacy Protection in Personalized Search,* SIGIR Forum 2007

[11] **J. Teevan, S. T. Dumais, E. Horvitz:** *Potential for Personalization*, ACM TOIS 17(1), 2010

[12] **G.-R. Xue, J. Han,Y. Yu:** *User Language Models for Collaborative Personalized Search*, ACM TOIS 27(2), 2009