Personalization in Search

SI650 / EECS540 Information Retrieval

October 29, 2025

Imagine the search “michigan”…

Imagine the search “michigan”

https://www.michigan.org

Pure Michigan | Official Travel & Tourism Website for Michigan

You are on the brink of planning a vacation so unique, it can only be classified as Pure **Michigan**. **Michigan's** Upper Peninsula Named #1 Destination for Fall ...

Travel Guide

The 2022 Pure Michigan Spring/Summer Travel Guide highlights ...

Things to Do

All Attractions - 10 Amazing Hidden Gems - Trip Ideas - **...**

First-Time Visitors

For first time Michigan visitors, find inspiration for your visit to the ...

Trip Ideas

Let our articles take some of the guess work out of travel ...

More results from michigan.org »

https://www.tripadvisor.com › United States › Michigan

Michigan 2022: Best Places to Visit - TripAdvisor

Michigan Tourism: Tripadvisor has 1890105 reviews of Michigan Hotels, Attractions, and Restaurants making it your best Michigan resource.

https://www.travel-mi.com › Michigan-Travel-Guide

2022 Michigan Travel Guide+ Map Over 43 Cities, ...

Feb 19, 2022 — Get more inspiration, photos, landmarks and ideas from these popular **tourist** cities. Includes northern **Michigan/Michigan's** Upper Peninsula.

Imagine the search “michigan”

https://www.michigan.gov › som

State of Michigan: SOM

Use the **Michigan** Voter Information Center to find your polling place, view your sample ballot, learn about voting equipment and more.

https://en.wikipedia.org › wiki › Michigan

Michigan - Wikipedia

Michigan is a state in the Great Lakes region of the upper Midwestern United States. With a population of nearly 10.12 million and an area of nearly 97,000 ...

U.S. House delegation: 7 Democrats; 7 R…
Most of state: UTC−05:00 (Eastern)
Largest city: Detroit
Lowest elevation (Lake Erie): 571 ft (17…

History · Government · Geography

https://www.michigan.org

Pure Michigan | Official Travel & Tourism Website for ...

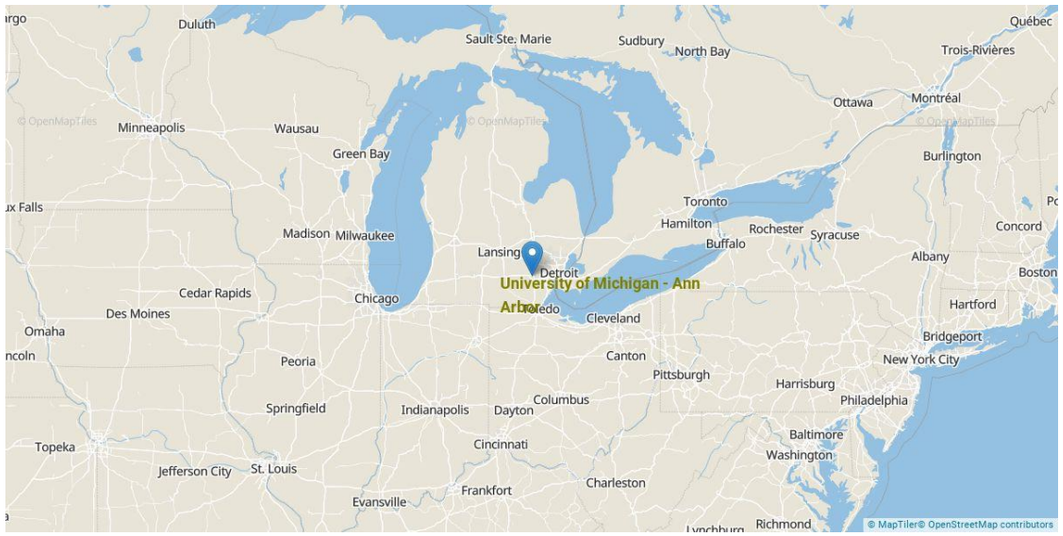
Find inspiration for your future getaway with **Michigan's** webcams and fall color updates. From scenic routes on the open road to rugged trails, ...

https://umich.edu

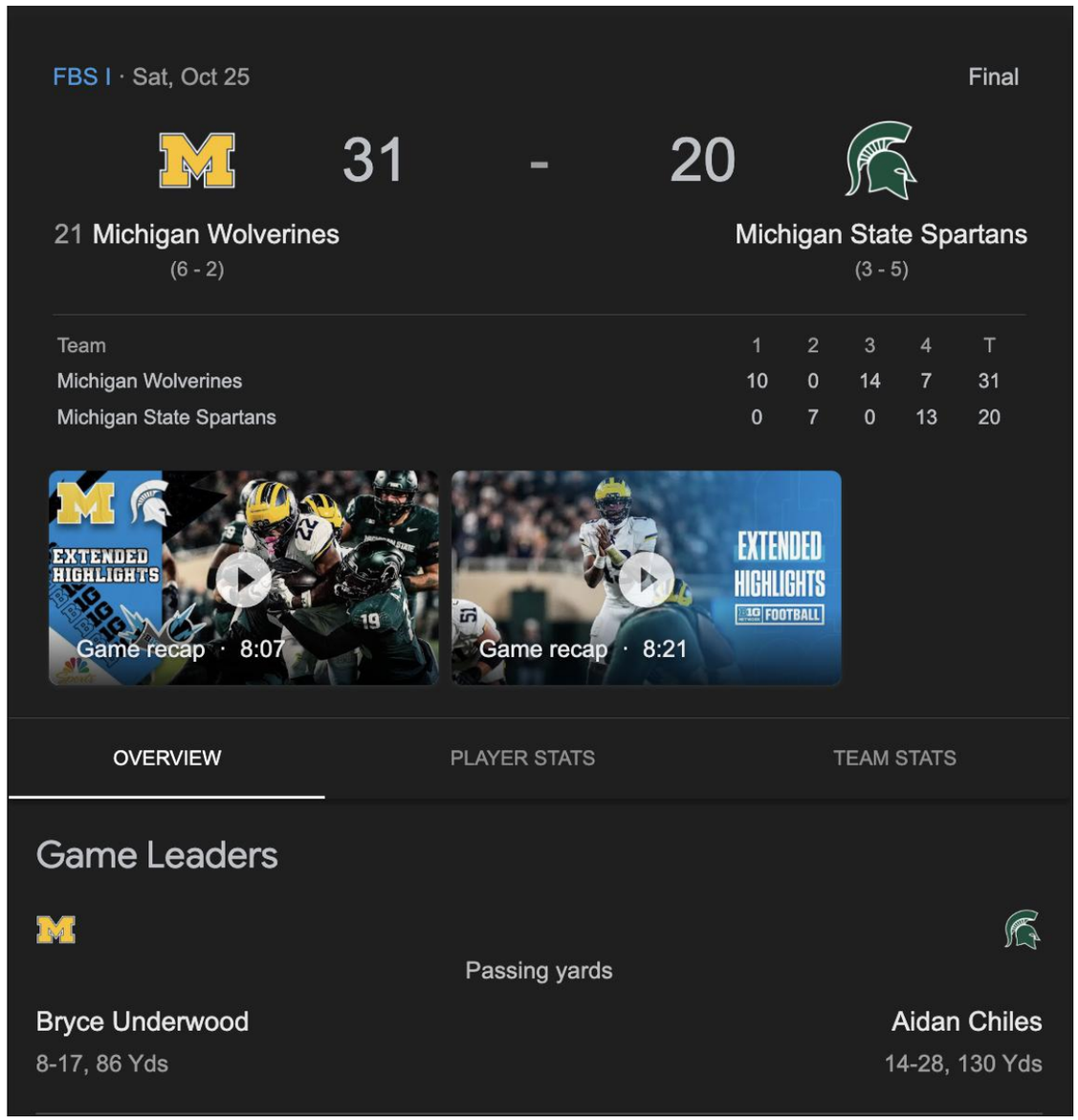
University of Michigan

A top-ranked public university, the University of **Michigan** has a tradition of excellence in research, learning and teaching, sports and the arts, and more.

Imagine the search “michigan”



on Saturday, Oct 25



The same query can have different intents

- Queries are ambiguous—and often short
- We need to use extra information to fulfill the user's information needs
- Personalization can help make use of these signals!
- Today's Lecture:
  - What do users do that we can learn from?
  - How can we personalize?
  - How to move personalization beyond the person?

michigan tourism

Travel Guide

The 2022 Pure Michigan Spring/Summer Travel Guide highlights ...

Things to Do

All Attractions - 10 Amazing Hidden Gems - Trip Ideas - ...

First-Time Visitors

For first time Michigan visitors, find inspiration for your visit to the ...

Trip Ideas

Let our articles take some of the guess work out of travel ...

More results from michigan.org »

Popular destinations in Michigan

About these results

These destinations are ranked mainly by popularity, and the cost and convenience of travel from your location. Factors include frequency of mentions across the web, destination search queries, travel time, number of stops, and airport changes during layovers.

Detroit

Michigan city famed for cars & Motown

45 min

Grand Rapids

Frederik Meijer Gardens & breweries

$258 49m

Cherry Festival & City Opera House

$238 1h

More destinations in Michigan

User behavior on the web

User Behaviors on Web

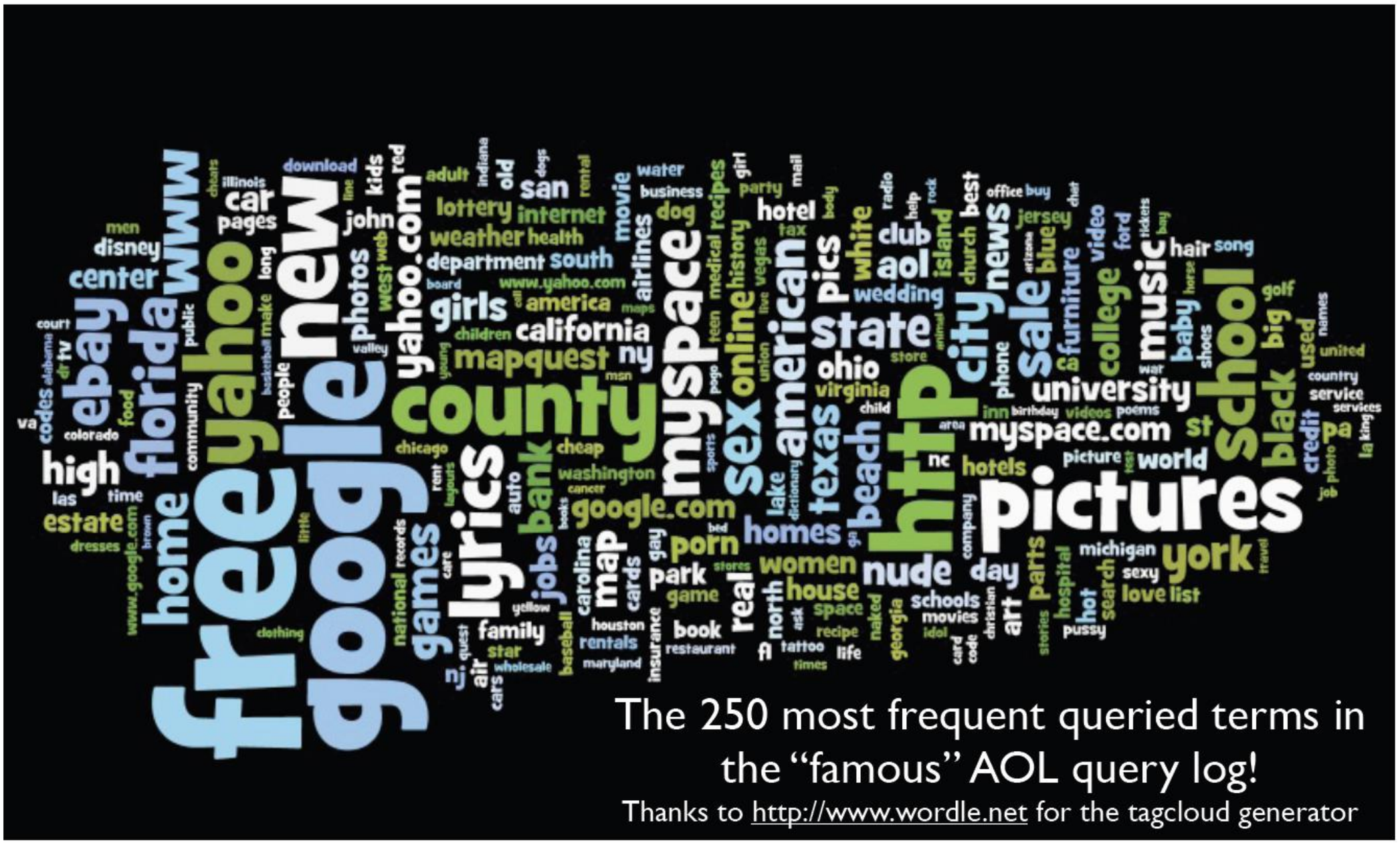
- The better you understand me, the better you can serve me
- Basic task: infer the information need, intents, interests of users from their past behaviors, and then predict their future behavior (e.g., click given a query)
- Sample problems:
  - Identifying sessions in query logs
  - Predicting accesses to a given page (e.g., for caching)
  - Recognizing human vs. automated queries
  - Recommending alternative queries, landing pages, …

Query Log analysis

- Main idea: log the user behaviors/actions in web search
- Analyze the log to better understand the users

AnonID	Query	QueryTime	ItemRank	ClickURL
100218	tennessee department of transportation	2006-03-01 11:08:30	1	http://www.tdot.state.tn.us
100218	tennessee federal court	2006-03-01 11:53:44	1	http://www.constructionweblinks.com
100218	state of tennessee emergency communications board	2006-03-01 12:56:18	1	http://www.tennessee.gov
100218	state of tennessee emergency communications board	2006-03-01 12:56:18	1	http://www.tennessee.gov
100218	state of tennessee emergency communications board	2006-03-01 12:56:18	2	http://www.tennessee.gov
100218	state of tennessee emergency communications board	2006-03-01 12:56:18	1	http://www.tennessee.gov
100218	dixie youth softball	2006-03-02 10:36:48	2	http://www.dixie.org
100218	cdwg	2006-03-03 14:29:07	1	http://www.cdwg.com
100218	cdwg scam cdwge	2006-03-03 14:30:11		
100218	escambia county sheriff's department	2006-03-07 09:26:51	1	http://www.escambiaso.com
100218	escambia county sheriff's department	2006-03-07 09:26:51	2	http://www.escambiaso.com
100218	escambia county sheriff's department	2006-03-07 09:26:51	1	http://www.escambiaso.com
100218	escambia county sheriff's department	2006-03-07 09:26:51	1	http://www.escambiaso.com
100218	pensacola police department	2006-03-07 09:34:28	1	http://www.pensacolapolice.com
100218	memphis pd	2006-03-07 09:42:33	1	http://www.memphispolice.org
100218	nashville metro pd	2006-03-07 09:44:43	1	http://www.police.nashville.org
100218	florida highway patrol	2006-03-07 09:48:35	1	http://www.fhp.state.fl.us
100218	tennessee highway patrol	2006-03-07 09:49:52	1	http://www.state.tn.us
100218	florida bureau of investigations	2006-03-07 09:51:08	2	http://www.flsbi.com
100218	florida bureau of investigations	2006-03-07 09:51:08	1	http://www.fhp.state.fl.us
100218	government finance officers asssociation	2006-03-07 21:16:11		
100218	state of tennessee controllers manual	2006-03-07 21:17:12		
100218	state of tennessee audit controllers manual	2006-03-07 21:17:40	3	http://www.comptroller.state.tn.us
100218	state of tennessee audit controllers manual	2006-03-07 21:17:40	4	http://www.tbr.state.tn.us
100218	state of tennessee audit controllers manual	2006-03-07 21:17:40	9	http://audit.tennessee.edu
100218	internal controls for municipalities under 10 000	2006-03-07 21:38:04	1	http://www.nysscpa.org
100218	internal controls for municipalities under 10 000	2006-03-07 21:38:04	4	http://www.massdor.com
100218	municipality fraud detection techniques	2006-03-07 21:41:40		
100218	municipal fraud audit detection internal controls	2006-03-07 21:43:15		
100218	internal fraud controls for municipalities cities towns local government	2006-03-07 21:45:13	1	http://www.whitehouse.gov
100218	internal fraud controls for municipalities cities towns local government	2006-03-07 21:45:13	4	http://www.nhlgc.org
100218	internal fraud controls for municipalities cities towns local government	2006-03-07 21:45:13	7	http://www.sao.state.ut.us
100218	evaluating internal controls a local government managers guide	2006-03-07 21:51:18	5	http://www.allbusiness.com

Query Log analysis



– Slide from Ricardo Baeza-Yates

Query Log Analysis in Literature

- Enhance ranking – retrieval, advertisement
- Query suggestion; refinement; expansion; substitution, …
- Spelling check
- Other tasks …

Query Log Analysis in Literature

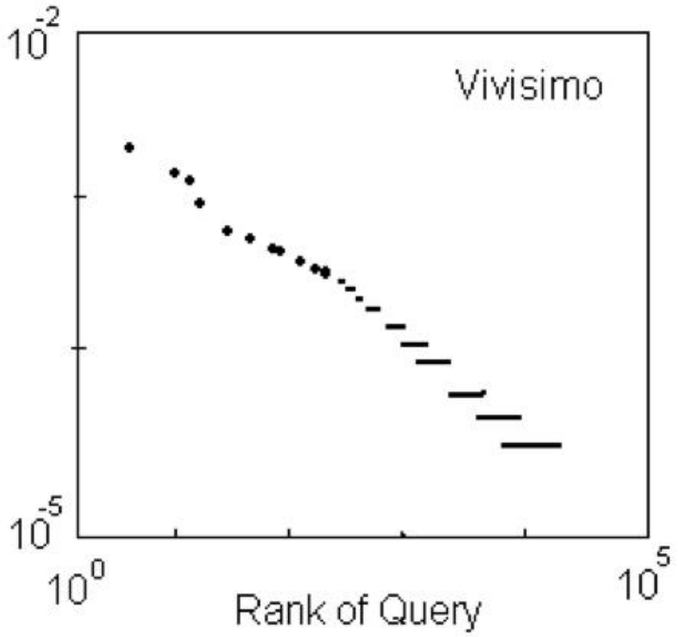
Query log name	Public	Period	# Queries	# Sessions	# Users
Excite '97	Y	Sep '97	1,025,908	211,063	~ 410,360
Excite '97 (small)	Y	Sep '97	51,473	N.D.	~ 18,113
Altavista	N	Aug 2^{nd} - Sep 13^{th} '98	993,208,159	285,474,117	N.D.
Excite '99	Y	Dec '99	1,025,910	325,711	~ 540,000
Excite '01	Y	May '01	1,025,910	262,025	~ 446,000
Altavista (public)	Y	Sep '01	7,175,648	N.D.	N.D.
Tiscali	N	Apr '02	3,278,211	N.D.	N.D.
TodoBR	Y	Jan - Oct '03	22,589,568	N.D.	N.D.
TodoCL	N	May – Nov '03	N.D.	N.D.	N.D.
AOL (big)	N	Dec 26^{th} '03 – Jan 1^{st} '04	~ 100,000,000	N.D.	~ 50,000,000
Yahoo!	N	Nov '05 – Nov '06	N.D.	N.D.	N.D.
AOL (small)	Y	Mar 1^{st} - May 31^{st} '06	36,389,567	N.D.	N.D.

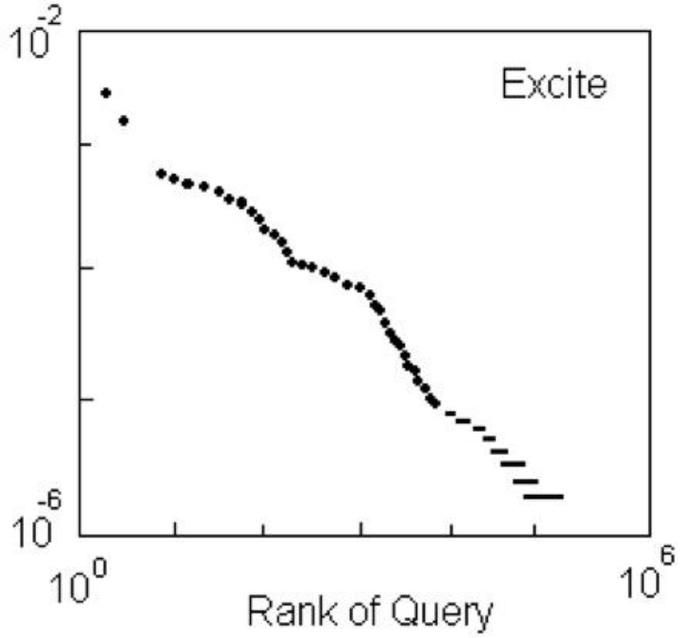
- Mei and Church (2008): MSN Search – 18 months,
  637 million unique queries,
  585 million unique urls,
  193 million unique IP addresses

Main Results of Query Log Analysis

- Average number of terms in a query is ranging from a low of 2.2 to a high of 2.6
- The most common number of terms in a query is 2
- 45% (2001) of queries are about Commerce, Travel, Economy, People (was 20%1997)
  - The queries about adult content or entertainment decreased from 20% (1997) to around 7% (2001)
- The majority of users don't refine their query
  - The number of users who viewed only a single page increase 29% (1997) to 51% (2001) (Excite)
  - 85% of users viewed only first page of search results (AltaVista)

Power-law Characteristics in Query Frequency





Power-Law in log-log space

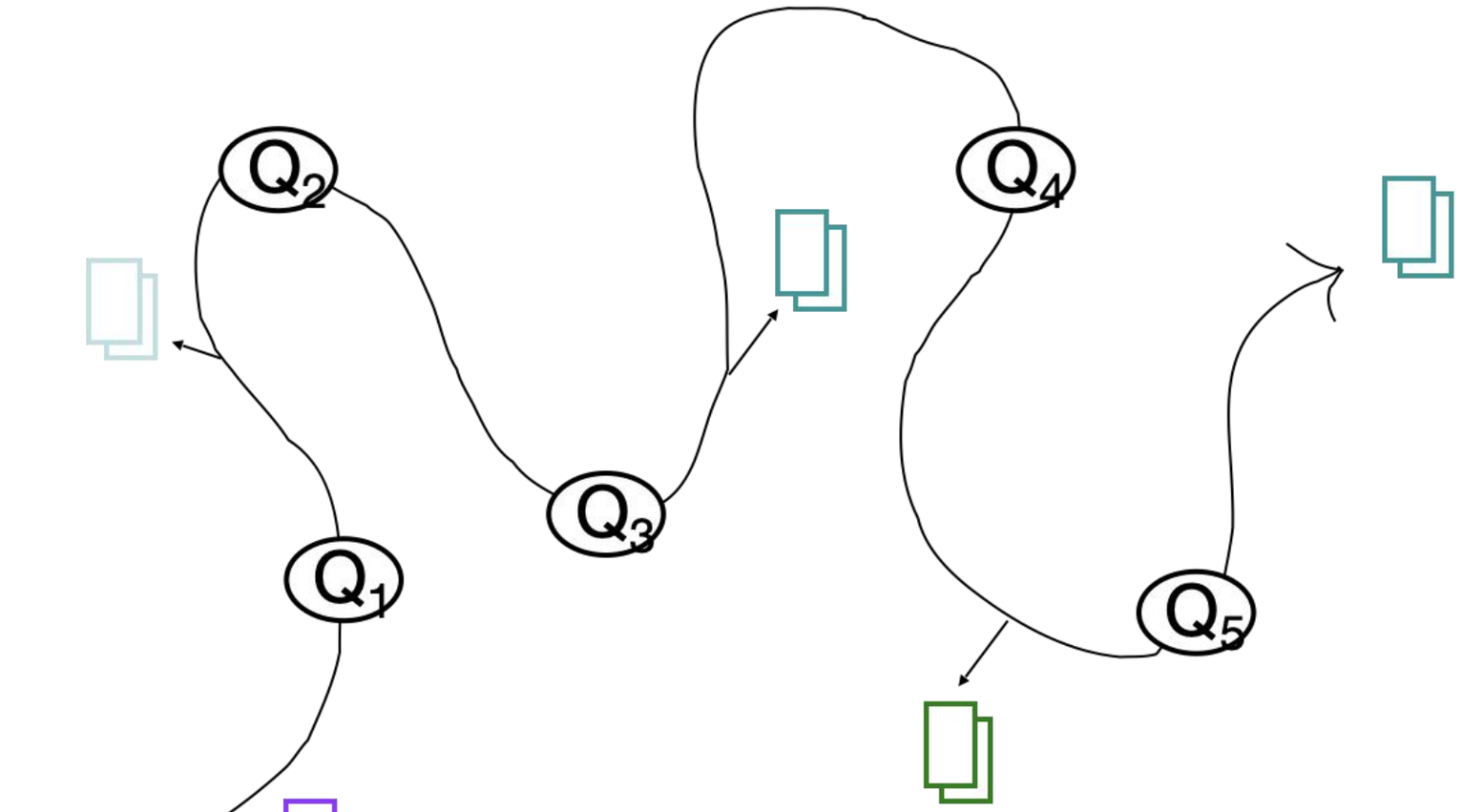
- Frequency f(r) of Queries with Rank r
  - 110000 queries from Vivisimo
  - 1.9 Million queries from Excite
- There are strong regularities in terms of patterns of behavior in how we search the Web

Entropy of Search Logs - How Hard is Search? With Personalization? With Backoff?

- Traditional Search
  - H(URL | Query)
  - H(URL,Query)-H(Query)
  - 2.8 (= 23.9 – 21.1)
  - In other words: How much uncertainty remains about the URL once we know the query
- Personalized Search
  - H(URL | Query, IP)
  - H(All)−H(All But URL)
  - 1.2 (= 27.2 – 26.0)

	Entropy (H)
Query	21.1
URL	22.1
IP	22.1
All But IP	23.9
All But URL	26.0
All But Query	27.1
All Three	27.2

A sketch of a searcher… “moving through many actions towards a general goal of satisfactory completion of research related to an information need.”



(Slide from “Marti Hearst, UCB SIMS, Fall 98)

Why Personalized Search?

Personalized Search

- Ambiguous query: MSR
  - Microsoft Research
  - Mountain Safety Research

- Disambiguate based on user's prior clicks
- If you know who I am, you should give me what I want
- Research issues:
  - What if we don't have enough history?
  - History v.s. new information needs
  - Privacy, privacy, privacy!

Ambiguity

- Unlikely that a short query can unambiguously describe a user's information need
- For example, the query [chi] can mean
  - Calamos Convertible Opportunities & Income Fund quote
  - The city of Chicago
  - Balancing one's natural energy (or ch'i)
  - Computer-human interactions

Personalization

- Ambiguity means that a single ranking is unlikely to be optimal for all users
- Personalized ranking is the only way to bridge the gap
- Personalization can use
  - Long term behavior to identify user interests, e.g., a long term interest in user interface research
  - Short term session to identify current task, e.g., checking on a series of stock tickers
  - User location, e.g., MTA in New York vs Baltimore
  - Social network
  - …

Potential for Personalization

- How much can personalization improve ranking? How can we measure this? [Teevan, Dumais, Horvitz 2010]
- Ask raters to explicitly rate a set of queries
  - But rather than asking them to guess what a user's information need might be …
  - ... ask which results *they would personally consider relevant*
  - Use self-generated and pre-generated queries

Computing potential for personalization

- For each query q
  - Compute average rating for each result
  - Let R_q be the optimal ranking according to the average rating
  - Compute the NDCG value of ranking R_q for the ratings of each rater i
    - Let Avg_q be the average of the NDCG values for each rater
- Let Avg be the average Avg_q over all queries
- Potential for personalization is (1 – Avg)

Example: NDCG values for a query

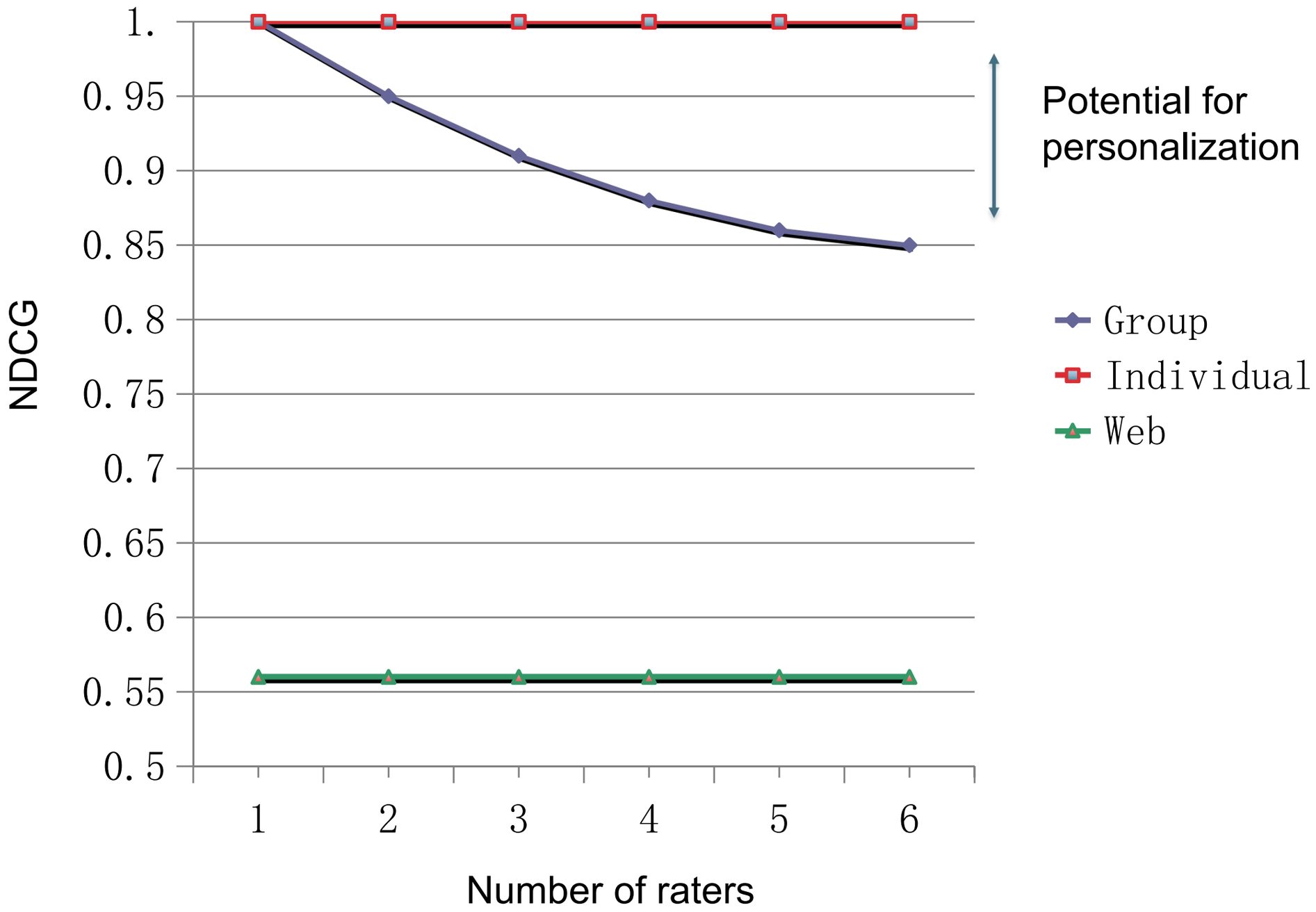
Table V.

The best ranking of the results for "slr Digital Camera" for user A and for user B. The rightmost section shows the best possible ranking if the same list must be returned to user A and user B. The normalized DCG for the best ranking when only one person is taken into account is 1. When more than one person must be accounted for, the normalized DCG drops

Best Ranking for User A		Best Ranking for User B		Best Ranking for Group (A + B)			
	Gain		Gain		Gain		
Web Result	A	Web Result	B	Web Result	A	B	A+B
usa.canon.com/consu...	1	..wikipedia.org/DSLR	2	..wikipedia.org/DSLR	1	2	3
..about.com/professio...	1	..about.com/professio...	1	..about.com/professio...	1	1	2
..wikipedia.org/Digital...	1	..about.com/..reviews...	1	usa.canon.com/consu...	1	0	1
..wikipedia.org/DSLR	1	usa.canon.com/consu...	0	..about.com/..reviews...	0	1	1
..about.com/..reviews...	0	amazon.com/..-Rebel-...	0	..wikipedia.org/Digital...	1	0	1
amazon.com/..-Rebel-...	0	amazon.com/Canon-4...	0	amazon.com/..-Rebel-...	0	0	0
amazon.com/Canon-4...	0	..wikipedia.org/Digital...	0	amazon.com/Canon-4...	0	0	0
olympusamerica.com/...	0	olympusamerica.com/...	0	olympusamerica.com/...	0	0	0
olympusamerica..body...	0	olympusamerica..body...	0	olympusamerica..body...	0	0	0
astore.amazon.com/p...	0	astore.amazon.com/p...	0	astore.amazon.com/p...	0	0	0
	A		B		A	B	Avg
Normalized DCG	1.00	Normalized DCG	1.00	Normalized DCG	0.97	0.96	0.97

Potential for personalization =0.03

Potential for personalization graph



The World Wide Web

A Looooooong Ago

Content

2,700 websites (14% .com)

Tools

Mosaic only 1 year old

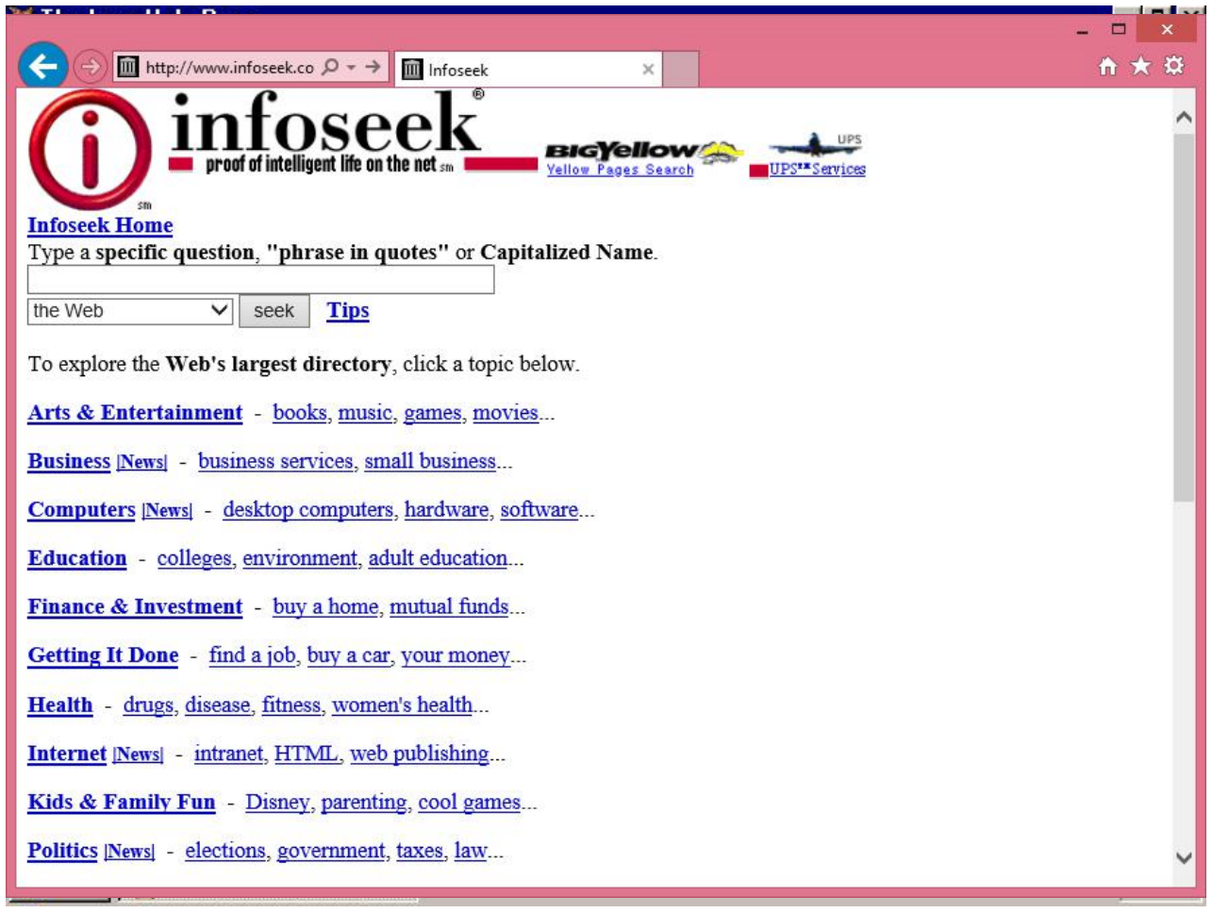
Pre-Netscape, IE, Chrome

4 years pre-Google

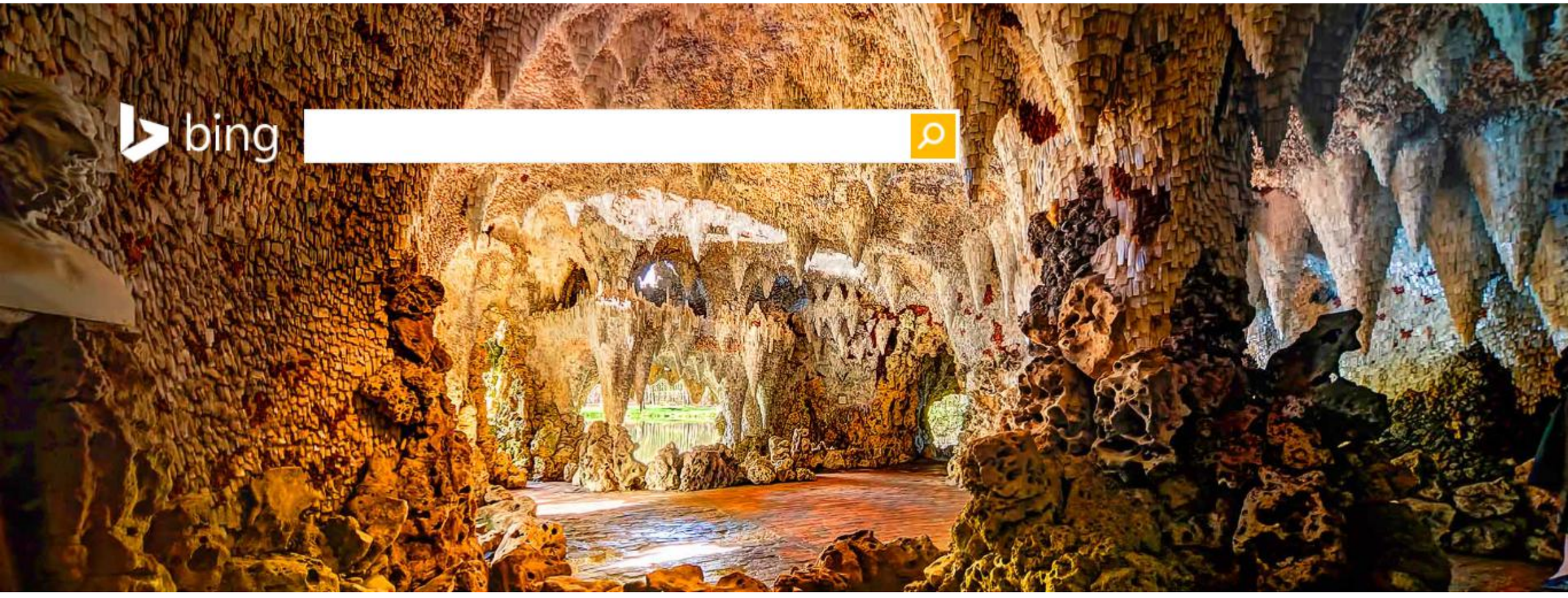
Search Engines

54,000 pages indexed by Lycos

1,500 queries per day

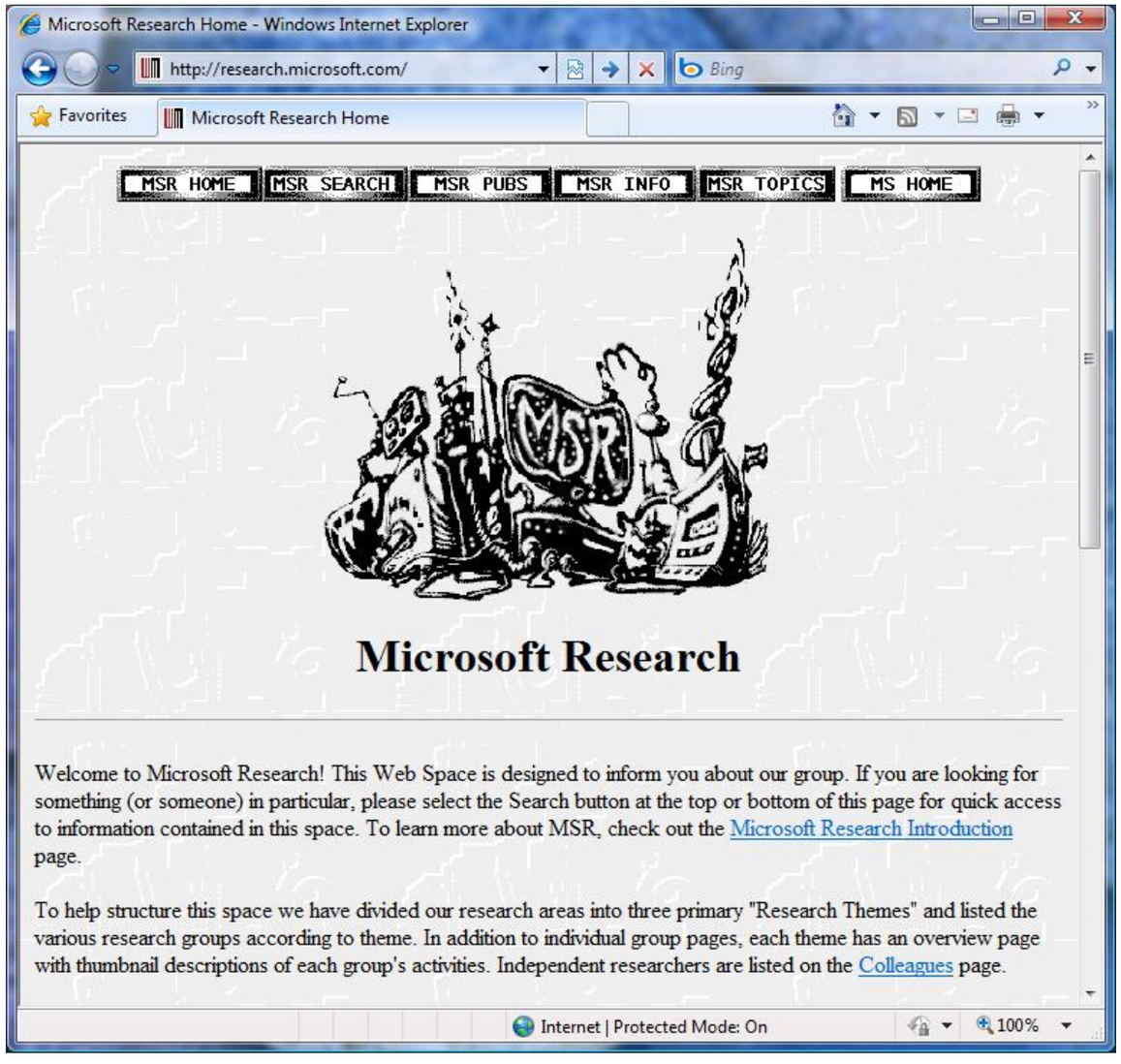


The World Wide Web Today



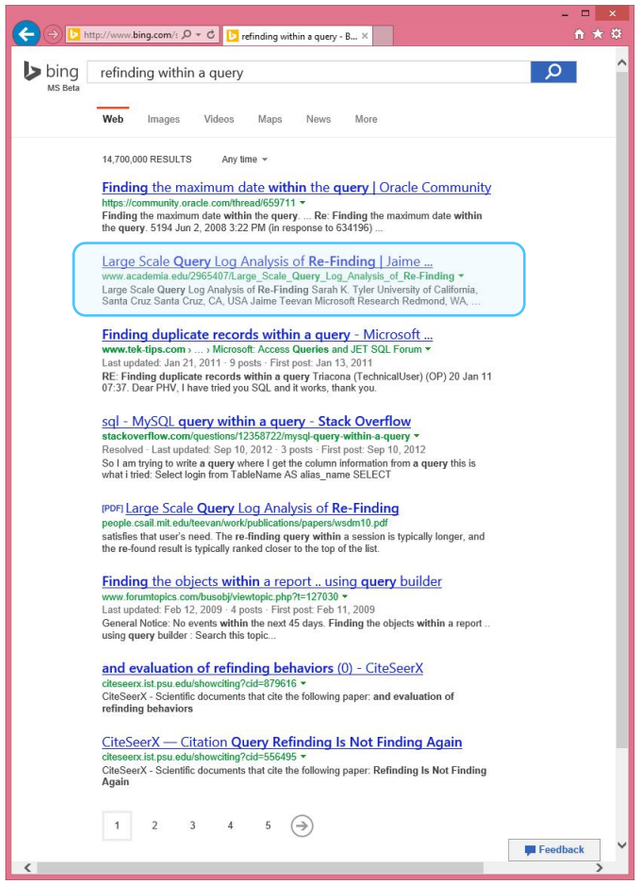
- Trillions of pages indexed.
- Billions of queries per day.

1996



- We assume information is static.
- But web content changes!

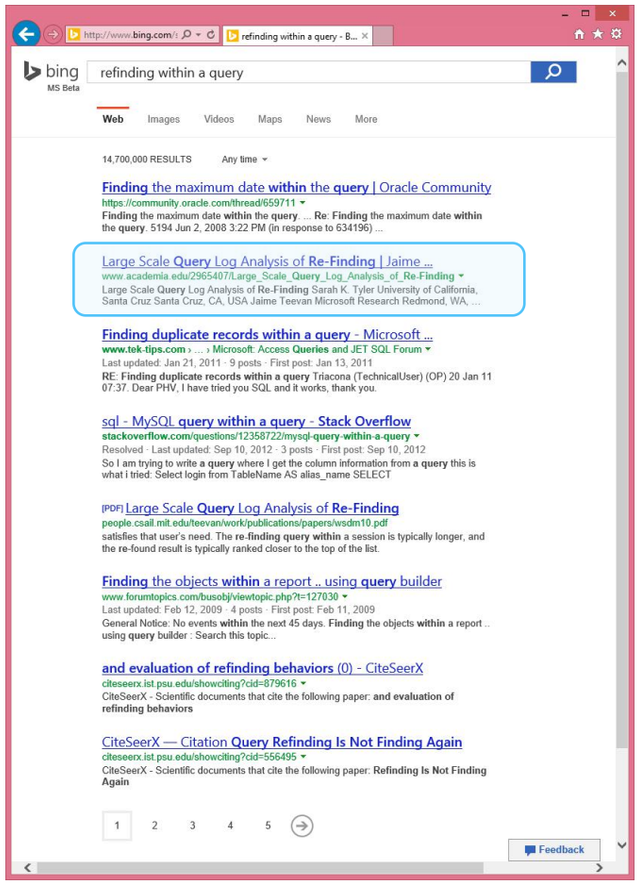
Search Results Change

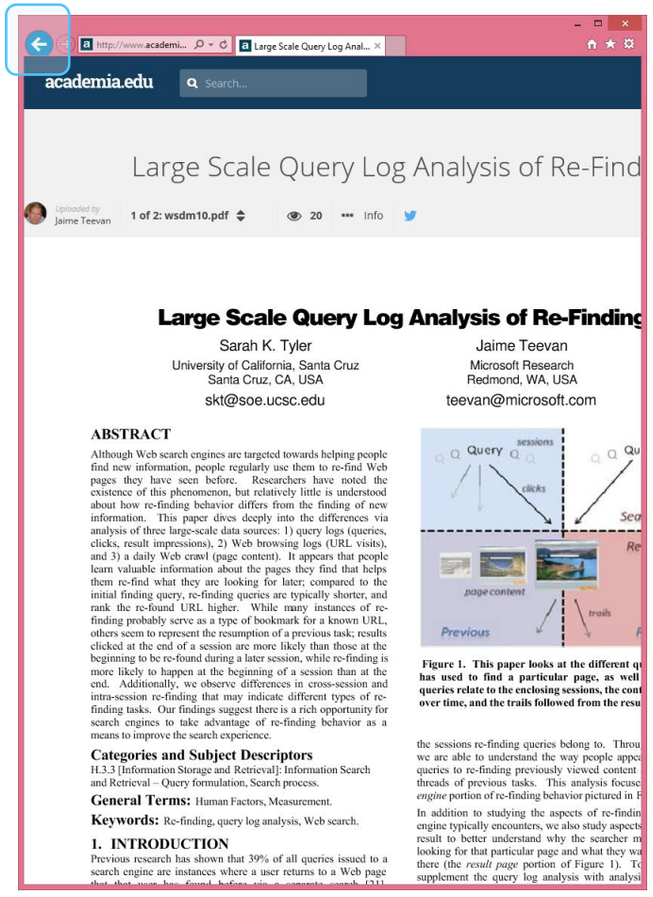


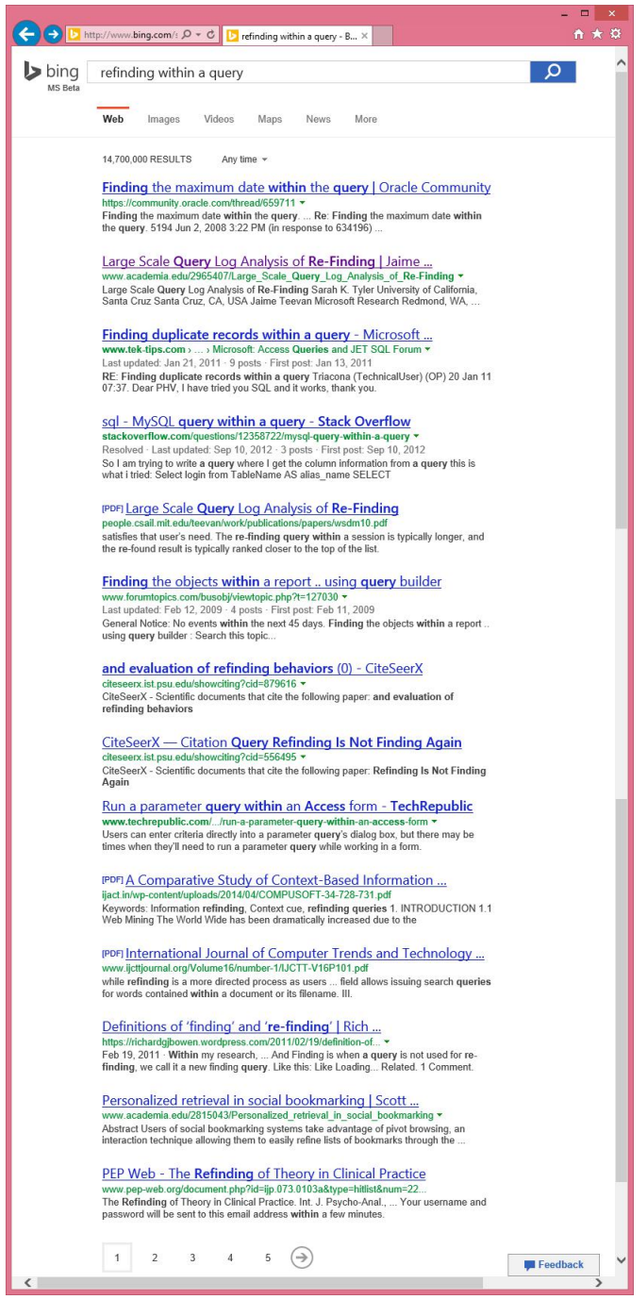
- New, relevant content
- Improved ranking
- Personalization
- General instability

Can change during a query!

Search Results Change







Behavioral Data

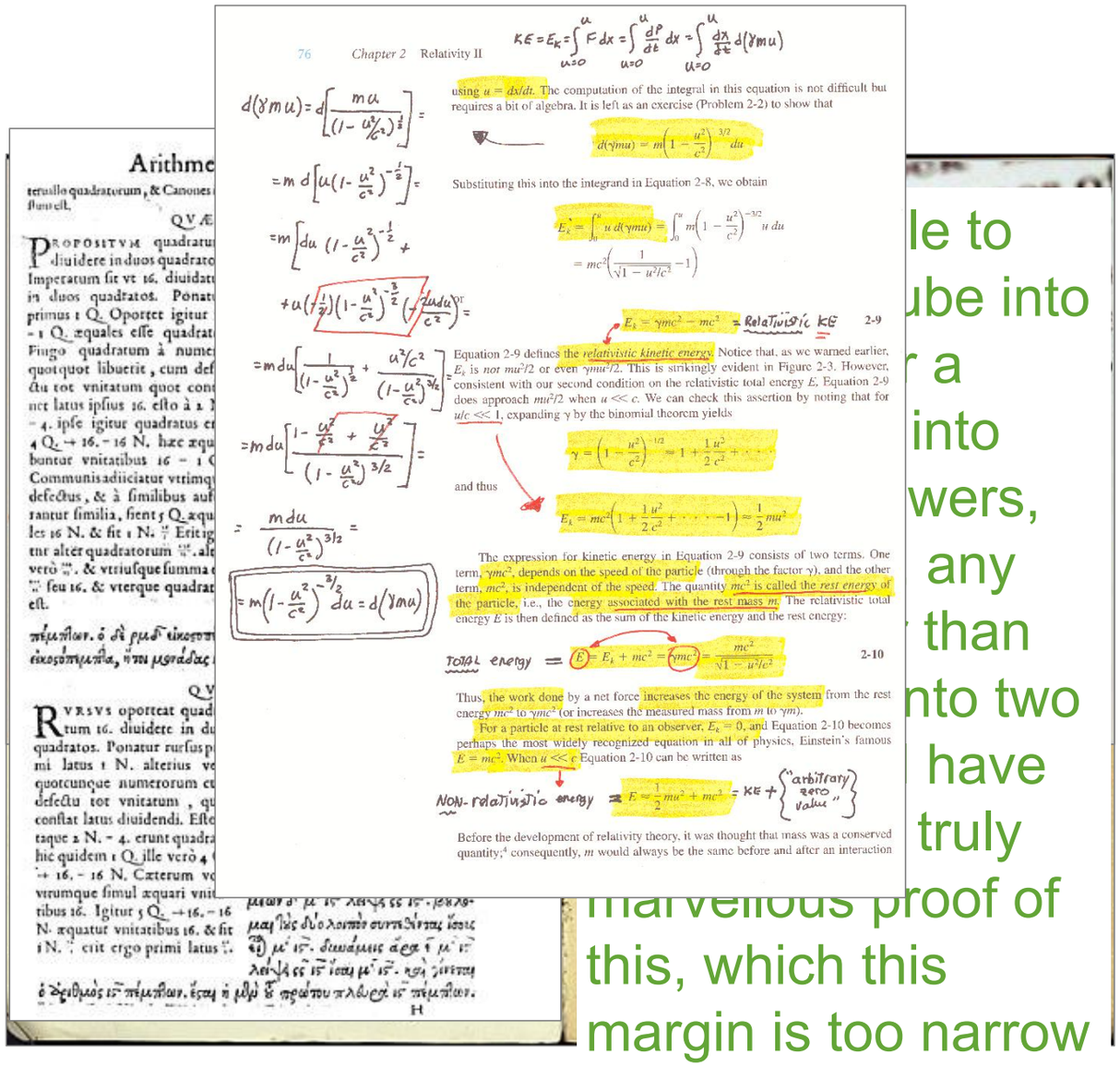
Many Years AGo

Marginalia adds value to books

Students prefer annotated texts

Do we lose marginalia when we move to digital documents?

No! Scale makes it possible to look at experiences in the aggregate, and to tailor and personalize



Arithme

teruallo quadratorum, & Canones
ſtum eſt.

QVÆ

PROPOSITVM quadratu
diuidere in duos quadrato
Imperatum ſit vt 16. diuidat
in duos quadratos. Ponat
primus 1 Q. Oportet igitur
– 1 Q. æquales eſſe quadrat
Fingo quadratum à nume
quotquot libuerit, cum def
ctu tot vnitatum quot cont
net latus ipſius 16. eſto à 2
– 4. ipſe igitur quadratus e
4 Q. + 16. – 16 N. hæc æqu
buntur vnitatibus 16 – 1
Communis adiiciatur vtrimq
defectus, & à ſimilibus auf
rantur ſimilia, fient 5 Q. æqu
les 16 N. & fit 1 N. $\frac{3}{5}$ Erit ig
tur alter quadratorum $\frac{144}{25}$. al
verò $\frac{81}{25}$. & vtriuſque ſumma
$\frac{225}{25}$ ſeu 16. & vterque quadrat
eſt.

πέμπλων. ὁ δὲ ρμδ̄ εἰκοσ
εἰκοσάπεμπτα, ἥτις μονάδας

QV

RVRSVS oportet quad
rum 16. diuidere in du
quadratos. Ponatur rurſus p
mi latus 1 N. alterius ve
quotcunque numerorum cu
defectu tot vnitatum, qu
conflat latus diuidendi. Eſto
taque 2 N. – 4. erunt quadra
hic quidem 1 Q. ille verò 4
+ 16. – 16 N. Cæterum v
virumque ſimul æquari vni
tibus 16. Igitur 5 Q. + 16. – 16
N. æquatur vnitatibus 16. & fit
1 N. $\frac{4}{5}$. erit ergo primi latus $\frac{16}{25}$.

ὁ δὲ ἑξῆς ις̄ πέμπλων. ἔσται ἡ μὲν β̄ τοῦ πλευρὰ ις̄ πέμπλων.

H

76 Chapter 2 Relativity II

$$KE = E_k = \int_{u=0}^{u} F\,dx = \int_{u=0}^{u} \frac{dp}{dt}\,dx = \int_{u=0}^{u} \frac{dx}{dt}\,d(\gamma m u)$$

$$d(\gamma m u) = d\left[\frac{mu}{(1-u^2/c^2)^{\frac{1}{2}}}\right] =$$

$$= m\,d\left[u\left(1-\frac{u^2}{c^2}\right)^{-\frac{1}{2}}\right] =$$

$$= m\left[du\,(1-\frac{u^2}{c^2})^{-\frac{1}{2}} + u\left(-\frac{1}{2}\right)\left(1-\frac{u^2}{c^2}\right)^{-\frac{3}{2}}\left(-\frac{2u\,du}{c^2}\right)\right] =$$

$$= m\,du\left[\frac{1}{(1-\frac{u^2}{c^2})^{\frac{1}{2}}} + \frac{u^2/c^2}{(1-\frac{u^2}{c^2})^{3/2}}\right] =$$

$$= m\,du\left[\frac{1-\frac{u^2}{c^2}+\frac{u^2}{c^2}}{(1-\frac{u^2}{c^2})^{3/2}}\right] =$$

$$= \frac{m\,du}{(1-\frac{u^2}{c^2})^{3/2}} =$$

$$= m\left(1-\frac{u^2}{c^2}\right)^{-3/2} du = d(\gamma m u)$$

using $u = dx/dt$. The computation of the integral in this equation is not difficult but requires a bit of algebra. It is left as an exercise (Problem 2-2) to show that

$$d(\gamma m u) = m\left(1-\frac{u^2}{c^2}\right)^{-3/2} du$$

Substituting this into the integrand in Equation 2-8, we obtain

$$E_k = \int_0^u u\,d(\gamma m u) = \int_0^u m\left(1-\frac{u^2}{c^2}\right)^{-3/2} u\,du$$
$$= mc^2\left(\frac{1}{\sqrt{1-u^2/c^2}} - 1\right)$$

$$E_k = \gamma mc^2 - mc^2 \quad \text{Relativistic KE} \qquad 2\text{-}9$$

Equation 2-9 defines the *relativistic kinetic energy*. Notice that, as we warned earlier, E_k is *not* $mu^2/2$ or even $\gamma mu^2/2$. This is strikingly evident in Figure 2-3. However, consistent with our second condition on the relativistic total energy E, Equation 2-9 does approach $mu^2/2$ when $u \ll c$. We can check this assertion by noting that for $u/c \ll 1$, expanding γ by the binomial theorem yields

$$\gamma = \left(1-\frac{u^2}{c^2}\right)^{-1/2} \approx 1 + \frac{1}{2}\frac{u^2}{c^2} + \cdots$$

and thus

$$E_k = mc^2\left(1 + \frac{1}{2}\frac{u^2}{c^2} + \cdots - 1\right) \approx \frac{1}{2}mu^2$$

The expression for kinetic energy in Equation 2-9 consists of two terms. One term, γmc^2, depends on the speed of the particle (through the factor γ), and the other term, mc^2, is independent of the speed. The quantity mc^2 is called the *rest energy* of the particle, i.e., the energy associated with the rest mass m. The relativistic total energy E is then defined as the sum of the kinetic energy and the rest energy:

$$\text{TOTAL energy} = E = E_k + mc^2 = \gamma mc^2 = \frac{mc^2}{\sqrt{1-u^2/c^2}} \qquad 2\text{-}10$$

Thus, the work done by a net force increases the energy of the system from the rest energy mc^2 to γmc^2 (or increases the measured mass from m to γm).

For a particle at rest relative to an observer, $E_k = 0$, and Equation 2-10 becomes perhaps the most widely recognized equation in all of physics, Einstein's famous $E = mc^2$. When $u \ll c$ Equation 2-10 can be written as

$$\text{NON-relativistic energy} \approx E = \frac{1}{2}mu^2 + mc^2 = \text{KE} + \{\text{"arbitrary zero value"}\}$$

Before the development of relativity theory, it was thought that mass was a conserved quantity;[4] consequently, m would always be the same before and after an interaction

le to
ube into
r a
into
wers,
any
than
nto two
have
truly
marvelous proof of this, which this margin is too narrow to contain.

Past Surprises About Web Search

- Early log analysis
  - –Excite logs from 1997, 1999
  - –Silverstein et al. 1999; Jansen et al. 2000; Broder 2002
- Queries are not 7 or 8 words long
- Advanced operators not used or “misused”
- Nobody used relevance feedback
- Lots of people search for sex
- Navigational behavior common
- *Prior experience was with library search*

Flaxman, Seth, Sharad Goel, and Justin M. Rao. "Filter bubbles, echo chambers, and online news consumption." *Public opinion quarterly* 80.S1 (2016): 298-320.

Search is Complex, Multi-Stepped process

- Typical query involves more than one click
  - 59% of people return to search page after their first click
  - Clicked results often not the endpoint
    - People orienteer from results using context as a guide
    - Not all information needs can be expressed with current tools
    - Recognition is easier than recall
- Typical search session involves more than one query
  - 40% of sessions contain multiple queries
  - Half of all search time spent in sessions of 30+ minutes
- Search tasks often involves more than one session
  - 25% of queries are from multi-session tasks

Which query has less variation in which link a user clicks ?

- campbells soup recipes v. vegetable soup recipe
- tiffany's v. tiffany
- nytimes v. connecticut newspapers
- www.usajobs.gov v. federal government jobs
- singaporepools.com v. singapore pools

Less variation = users more likely to click the same links for a query

Navigational Queries with Low Variation

- Use everyone's clicks to identify queries with low click entropy
  - 12% of the query volume
  - Only works for popular queries
- Clicks predicted only 72% of the time
  - Double the accuracy for the average query
  - But what is going on the other 28% of the time?
- Many typical navigational queries are not identified
  - People visit interior pages
    - craigslist – 3% visit http://geo.craigslist.org/iso/us/ca
  - People visit related pages
    - weather.com – 17% visit http://weather.yahoo.com

Key questions in personalization

- What kinds of queries need personalized?
- When to use personalization?
- How to model a user's behavior?
- How to personalize?

Personalized Search

Introduction

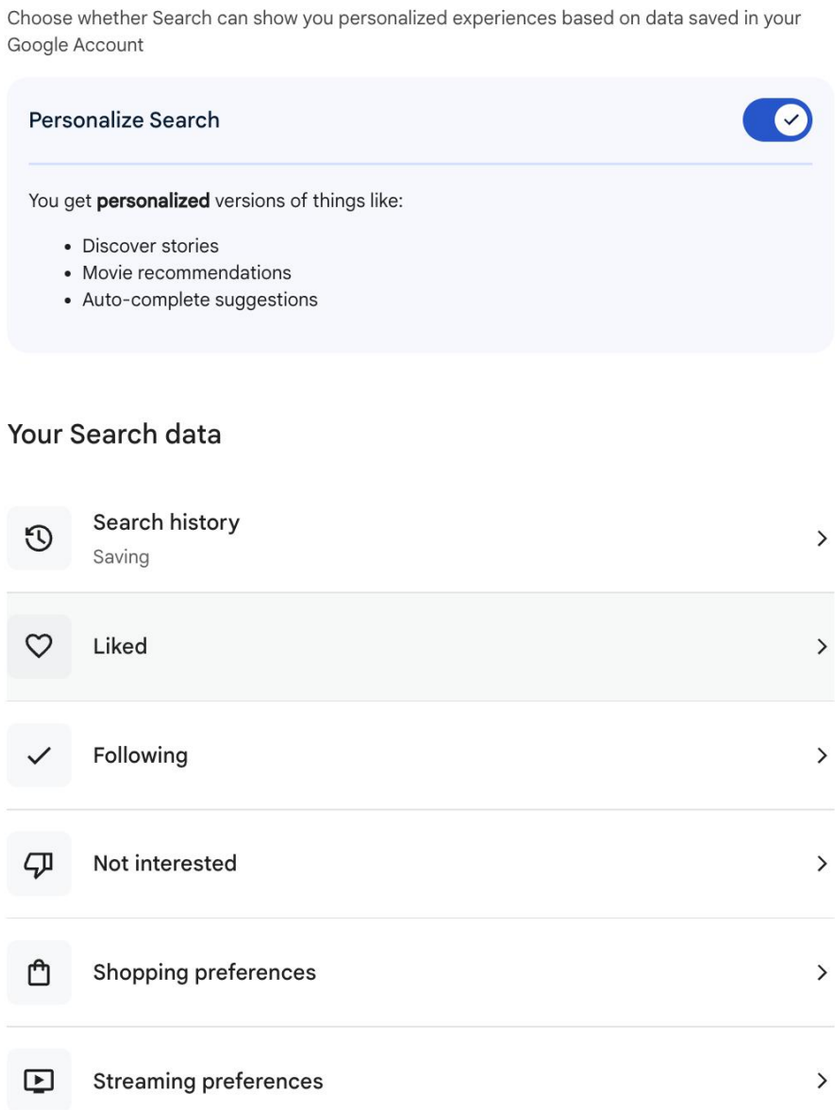
- The vast majority of search queries are short and ambiguous. Different users consider the same query to mean different things.
  - Ex. “ajax” Is the cleaning product ajax? Or Dutch football team Ajax Amsterdam? Or the greek mythological hero Ajax?
- Personalized search is a potential solution to all these problems.
- Matthijs and Radlinski (2011) describe one way Bing personalized results
  - Building an improved user profile for personalizing web search results.

Introduction

- To successfully personalize search results, it is essential to be able to identify what types of results are relevant to users.
  - (1) Ask users to label documents
  - (2) Infer personal relevance automatically.
- Identify relevant parts by…
  - parsing web page structure,
  - using term extraction and extracting noun phrases
  - model users from extract information
- Using document judgments obtained from a small number of users for 72 queries to assess potential approaches, then select three methods for complete online evaluation.

Personalization Strategies

- A user could be represented by :
  - a list of terms
  - weights associated with those terms
  - a list of visited URLs
  - the number of visits to each
  - a list of past search queries
  - pages clicked for these search queries
  - …



Personalization Strategies

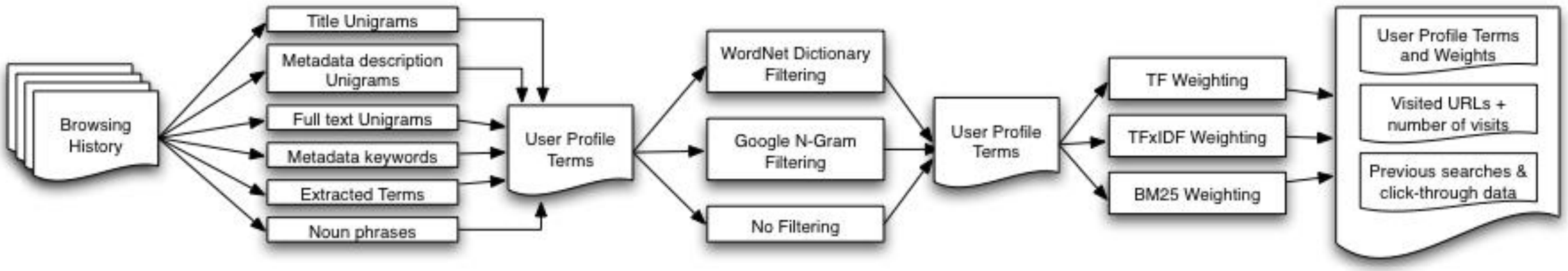


Figure 1: User Profile Generation Steps and Workflow

- User Profile Generation – use to rerank
  - Data Capture
  - Data Extraction
  - Term List Filtering
  - Term Weighting
- Re-ranking Strategies
  - Scoring Methods
  - Rank and Visit Scoring

Personalization Strategies -User Profile Generation

- Data Capture

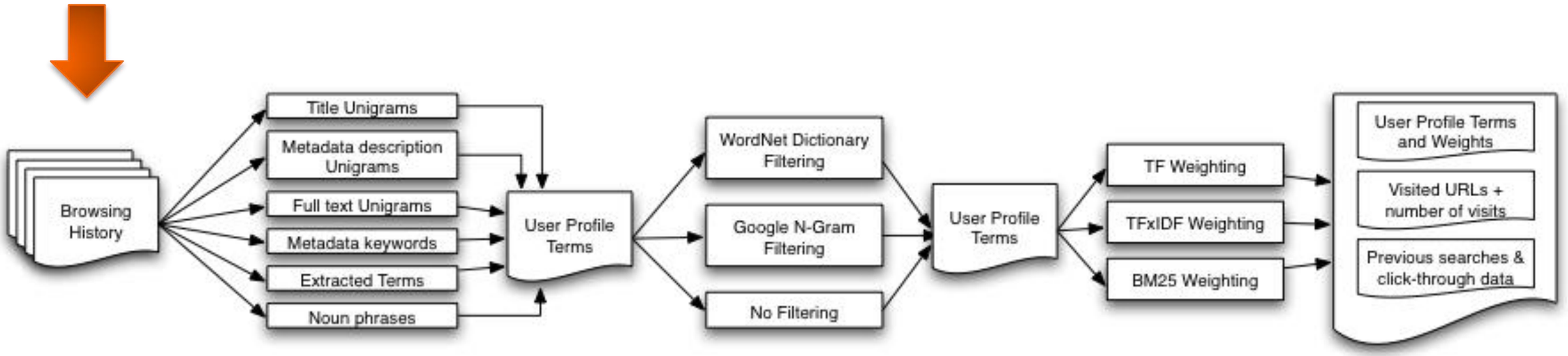


Figure 1: User Profile Generation Steps and Workflow

Where do user profiles come from?

- Signing into search accounts—e.g., linking google search your your google account
- Early days: Browser extension—e.g., “Altergo” (A FireFox add-on) to get user browsing records.

Table 1: Captured Data Statistics

Metric	Total	Min	Max	Mean
Page Visits	530,334	51	53,459	10,607
Unique Page Visits	218,228	36	26,756	4,365
Google Searches	39,838	0	4,203	797
Bing Searches	186	0	53	4
Yahoo Searches	87	0	29	2
Wikipedia Pages	1,728	0	235	35

Preprocessing the history to get a browsing profile

- Data Extraction

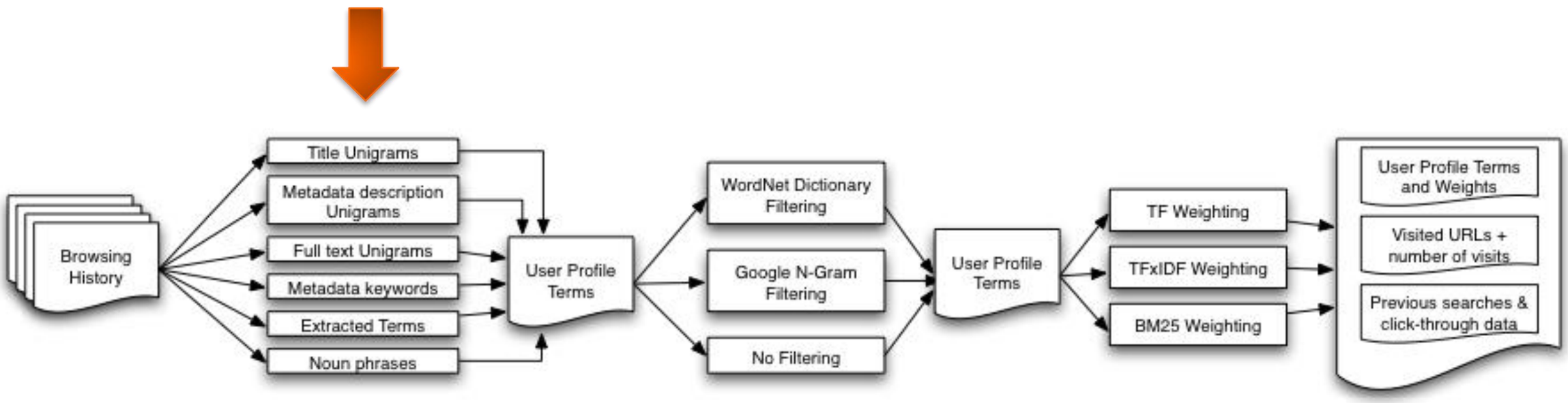


Figure 1: User Profile Generation Steps and Workflow

Personalization Strategies: User Profile Generation

- Full Text Unigrams
  - The body text of each web page.
- Title Unigrams
  - The words inside any <title> tag on the html pages.
- Metadata Description Unigrams
  - The content inside any <meta name="description"> tag.
- Metadata keywords Unigrams
  - The content inside any <meta name="keywords"> tag.
- Extracted Terms
  - Term Extraction algorithm on the full text of each visited web page. It attempts to summarize the web page's text into a set of important keywords.
- Noun Phrases
  - Taking the text from each web pages and splitting it into sentence using OpenNLP Tool, then using the Clark & Curran Statistical Language Parser which assigns a constituent tree to the sentence and part of speech tags to each word.

Reducing noise in user profiles

- Term List Filtering to reduce the number of noisy terms in user representation
  - E.g., removing infrequent words or words not in WordNet
    - When could this go wrong?

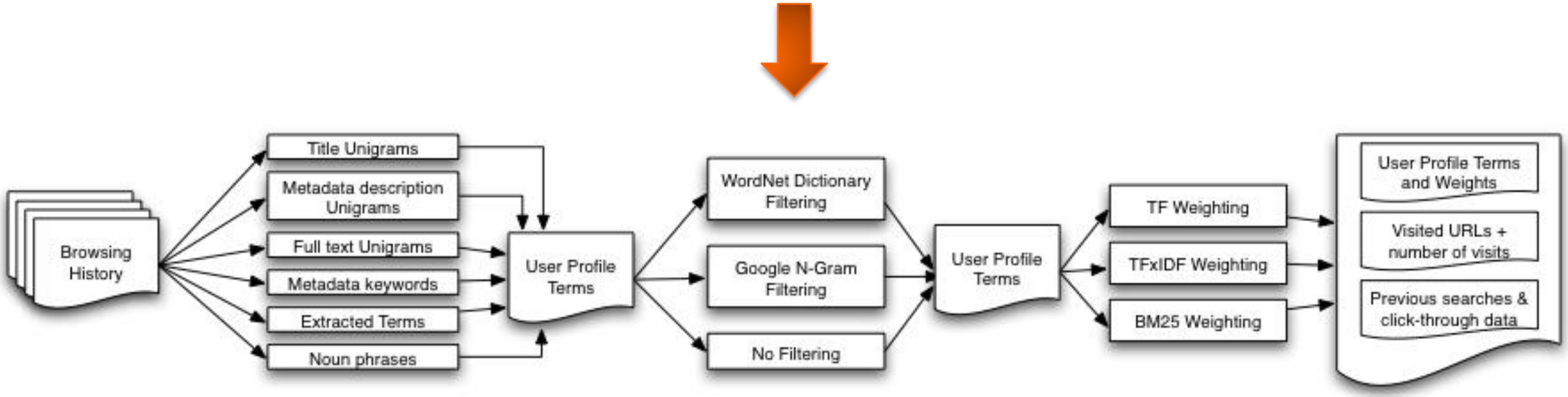


Figure 1: User Profile Generation Steps and Workflow

Personalization Strategies -User Profile Generation

- Term Weighting

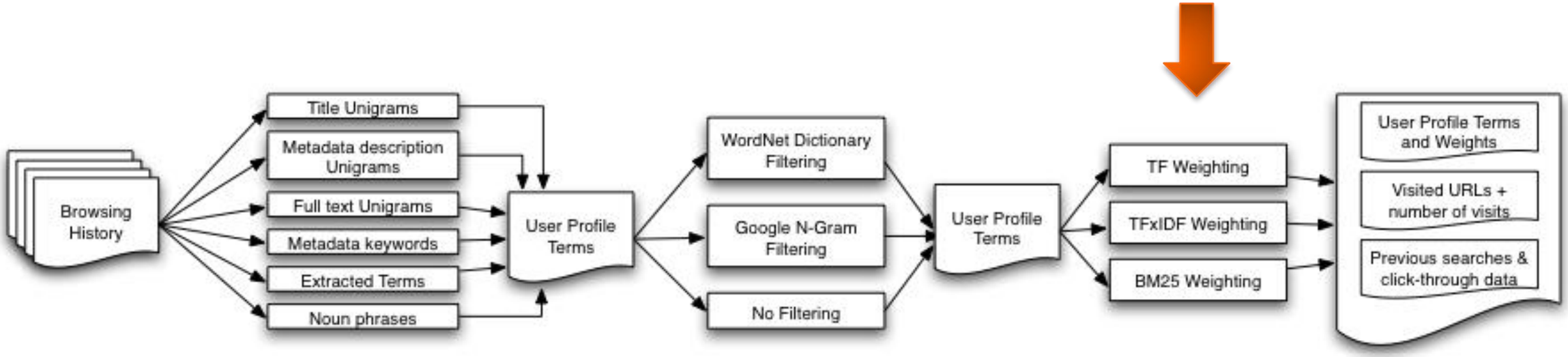


Figure 1: User Profile Generation Steps and Workflow

How to incorporate profile information into ranking?

Personalization Strategies -User Profile Generation

- TF Weighting

$$\vec{F}_{t_i} = \begin{bmatrix} f_{title_{t_i}} \\ f_{mdesc_{t_i}} \\ f_{text_{t_i}} \\ f_{mkeywt_{t_i}} \\ f_{terms_{t_i}} \\ f_{nphrases_{t_i}} \end{bmatrix}$$

- 0 : ignore particular field
- 1 : including particular field
- 1/ N_i :N_i is the total number of terms in
- TF-IDF Weighting

$$w_{TF}(t_i) = \vec{F}_{t_i} \cdot \vec{\alpha}$$

Weight vector

$$w_{TFIDF}(t_i) = \frac{1}{log(DF_{t_i})} \times w_{TF}(t_i)$$

What about updating BM25?

$$S(Q,D)=\sum_{t\in Q\cap D}\ln\frac{N-df(t)+0.5}{df(t)+0.5}\cdot\frac{(k1+1)\cdot c(t,D)}{k1(1-b+b\frac{|D|}{avdl})+c(t,D)}\cdot\frac{(k3+1)\cdot c(t,Q)}{k3+c(t,Q)}$$

What about updating BM25?

$$S(Q,D) = \sum_{t \in Q \cap D} \ln \frac{N - df(t) + 0.5}{df(t) + 0.5} \cdot \frac{(k1+1)\cdot c(t,D)}{k1(1 - b + b\frac{|D|}{avdl}) + c(t,D)} \cdot \frac{(k3+1)\cdot c(t,Q)}{k3 + c(t,Q)}$$

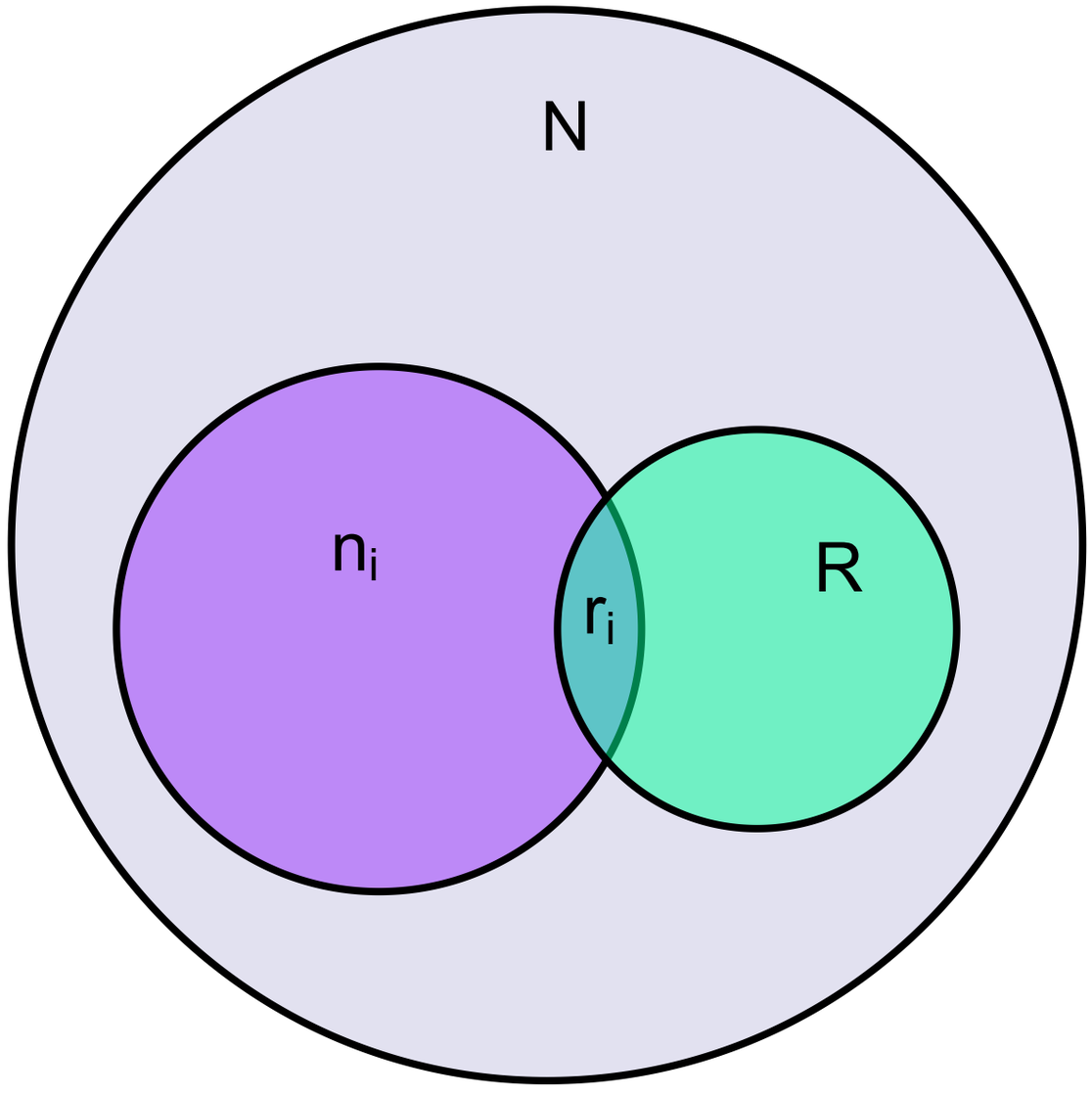
(1) (2) (3)

Which of the above components (1), (2), or (3) gives us a variant form of IDF?

Which of the above components (1), (2), or (3) gives us a variant form of QTF?

BM25 with Relevance Feedback

- N : the number of documents on the web
- n_i: the number of documents in the corpus that contain the term t
- R : the number of documents in the user's browsing history
- r_i is the number of Relevant documents that contains this term



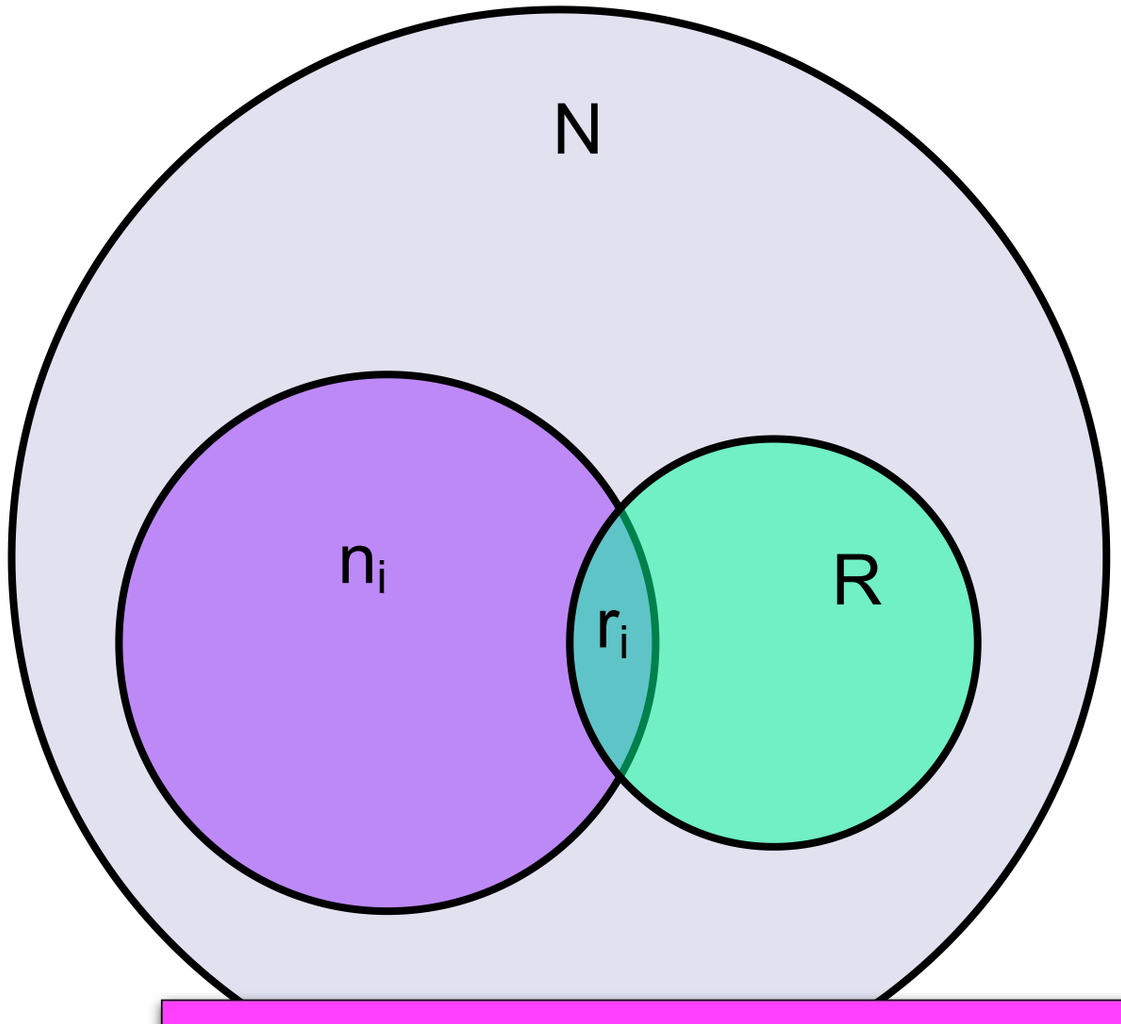
$$\text{Score} = \sum \text{tf}_i * w_i$$

$$w_i = log \frac{N - n_i + 0.5}{n_i + 0.5}$$

With no relevance, we weight words with the IDF

BM25 with Relevance Feedback

- N : the number of documents on the web
- n_i: the number of documents in the corpus that contain the term t
- R : the number of Relevant documents
- r_i is the number of Relevant documents that contains this term



$$\text{Score} = \sum tf_i * w_i$$

Our aim: computing the **odds ratio** that a term appears in *relevant* documents versus *non-relevant* documents.

$$w_i = log \frac{(r_i+0.5)(N-n_i-R+r_i+0.5)}{(n_i-r_i+0.5)(R-r_i+0.5)}$$

When we have relevance judgments, we can weight terms based on how frequent they are in relevance documents

BM25 with Relevance Feedback

A bit more on why the below formula captures the **odds ratio** that a term appears in *relevant* documents versus *non-relevant* documents.

$$w_i = log \frac{(r_i+0.5)(N-n_i-R+r_i+0.5)}{(n_i-r_i+0.5)(R-r_i+0.5)}$$

Odds that term appears in relevant docs = $\frac{r_i/R}{(R-r_i)/R} = \frac{r_i}{R-r_i}$

Odds that term appears in irrelevant docs = $\frac{(n_i-r_i)/(N-R)}{(N-n_i-R+r_i)/(N-R)} = \frac{n_i-r_i}{N-n_i-R+r_i}$

The **log ratio** of those two odds gives us: $\log \frac{r_i(N-n_i-R+r_i)}{(n_i-r_i)(R-r_i)}$

+0.5 for smoothing and avoiding divide by zero

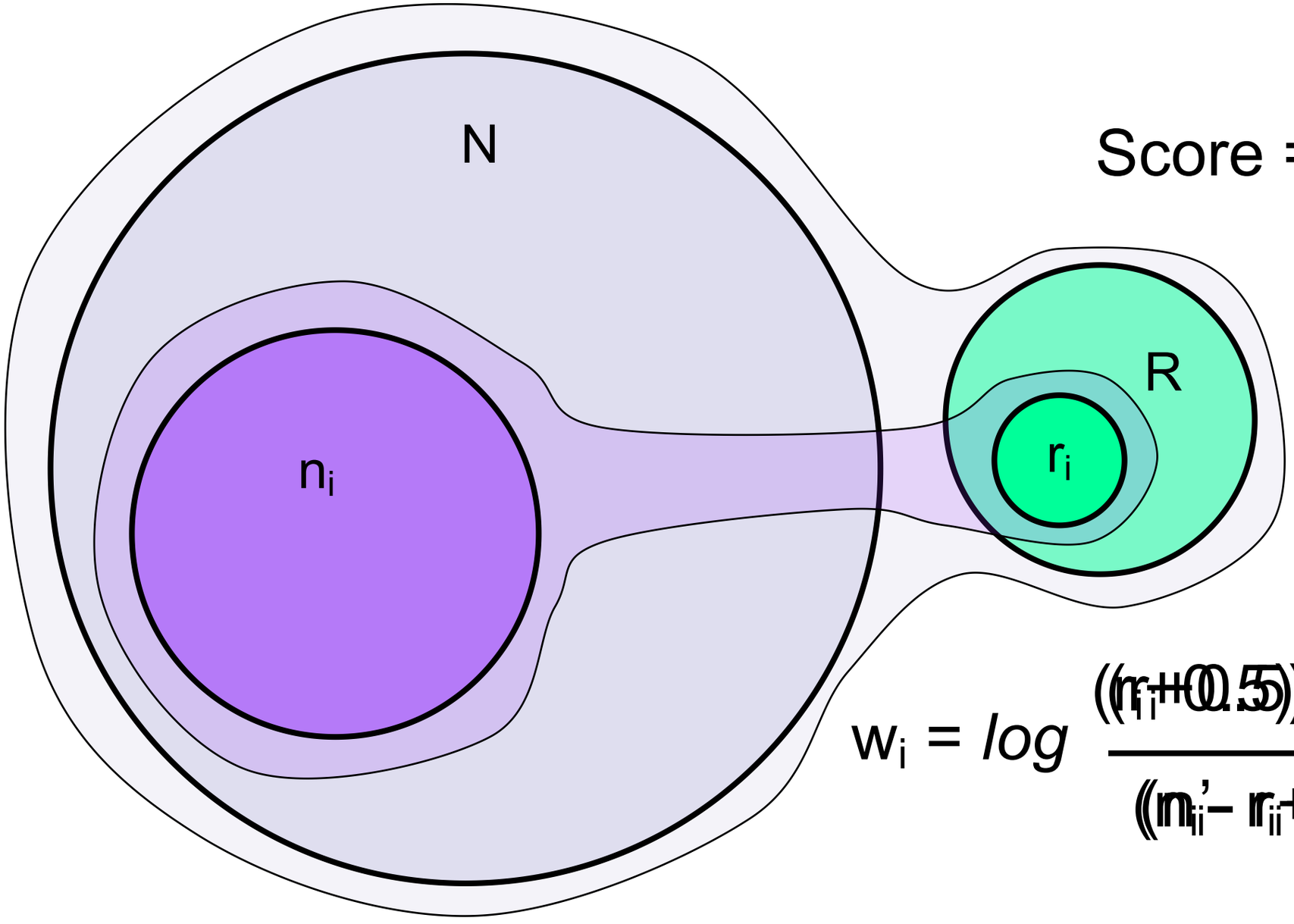
Personalization Strategies: Personalized BM25 Weighting

$$w_{pBM25}(t_i) = \log \frac{(r_i + 0.5)(N - n_{t_i} + 0.5)}{(n_{t_i} + 0.5)(R - r_{t_i} + 0.5)}$$

- N : the number of documents on the web
- n_{ti}: the number of documents in the corpus that contain the term t_i
- R : the number of documents in the user's browsing history
- r_{ti} is the number of documents in the browsing history that contains this term within the selected input data source

User Model as Personalized Relevance Feedback

- N : the number of documents on the web
- n_i: the number of documents in the corpus that contain the term t
- R : the number of documents in the user's browsing history
- r_i is the number of documents in the browsing history that contains this term within the selected input data source

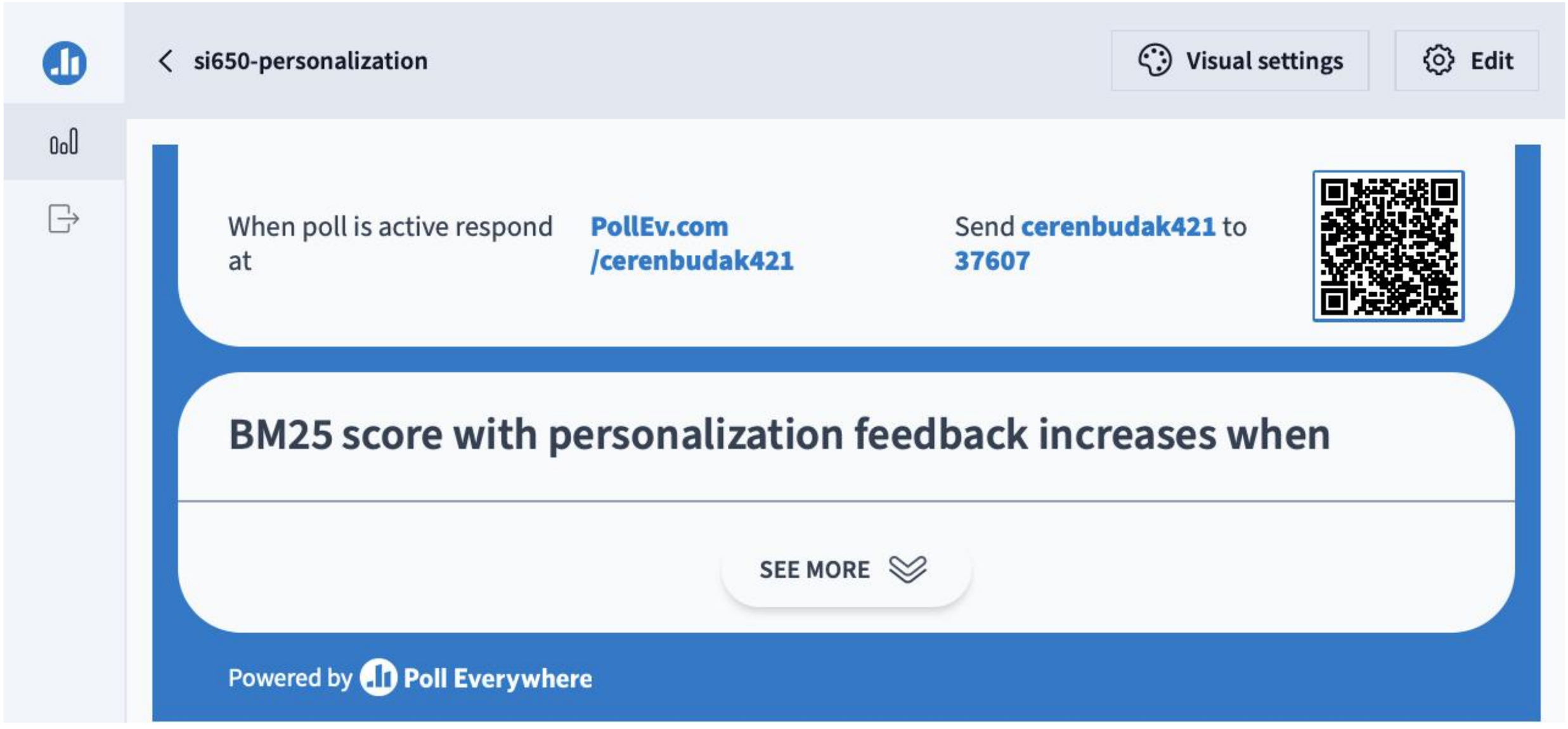


$$\text{Score} = \sum tf_i * w_i$$

$$N' = N+R$$

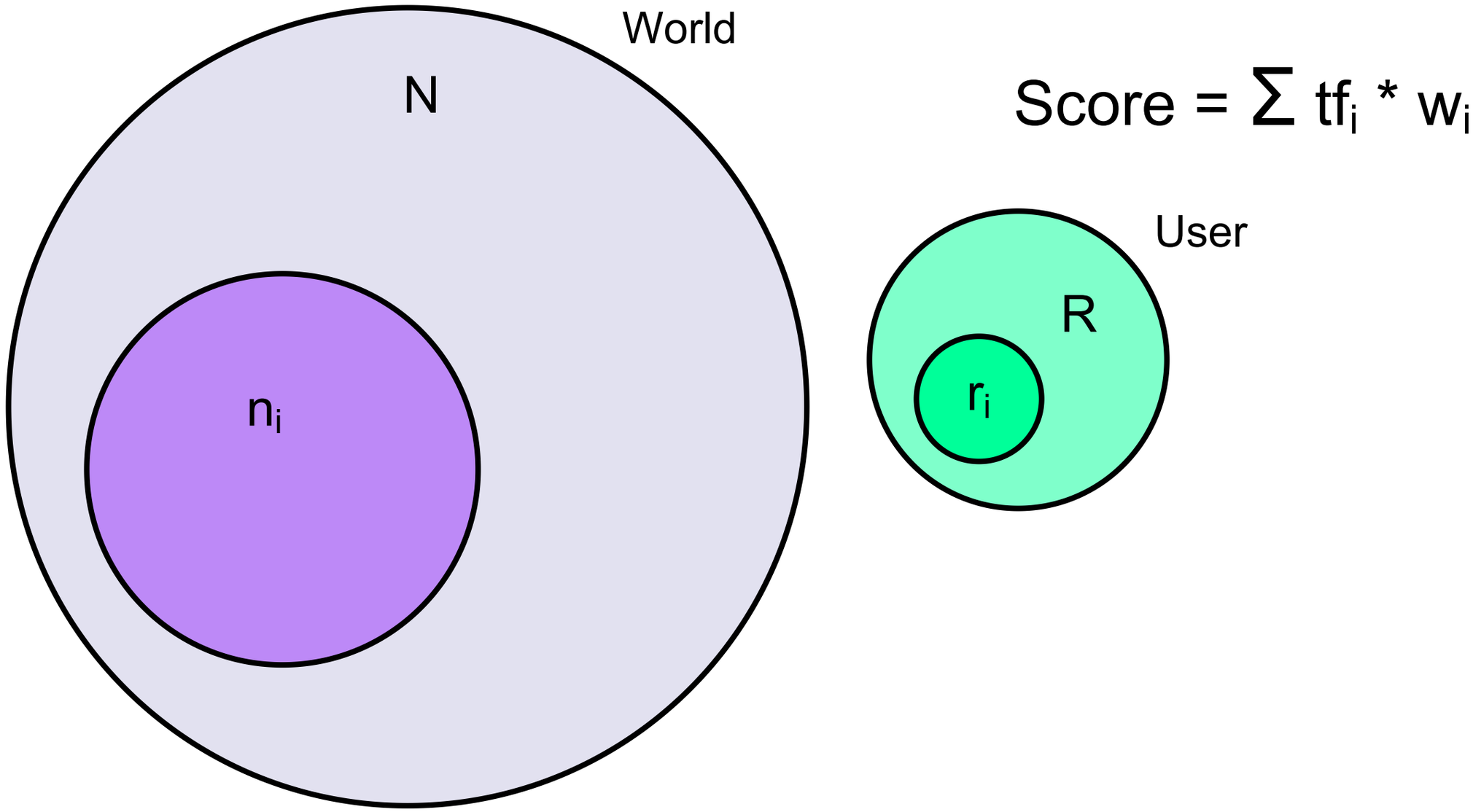
$$n_i' = n_i+ri$$

$$w_i = log \frac{((r_i+0.5)(N-n_i'-R+r_i+0.5))}{((n_i'-r_i+0.5)(R-r_i+0.5))}$$



User Model as Relevance Feedback

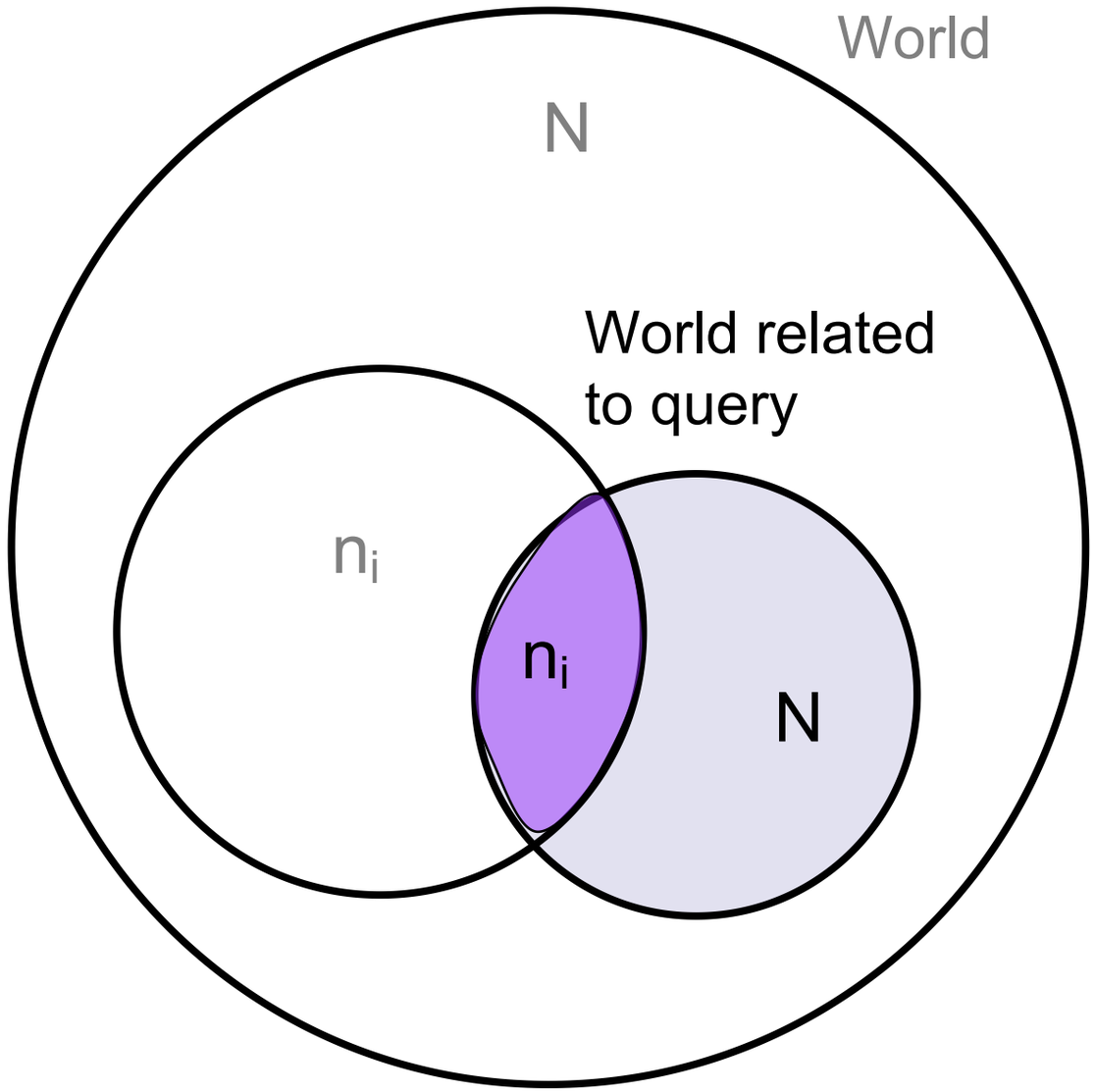
- N : the number of documents on the web
- n_i: the number of documents in the corpus that contain the term t
- R : the number of documents in the user's browsing history
- r_i is the number of documents in the browsing history that contains this term within the selected input data source



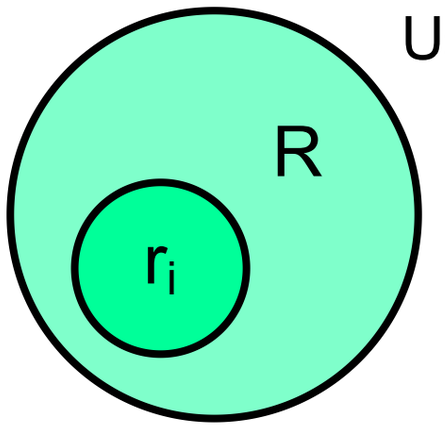
$$Score = \sum tf_i * w_i$$

User Model as Relevance Feedback

- N : the number of documents on the web
- n_i: the number of documents in the corpus that contain the term t
- R : the number of documents in the user's browsing history
- r_i is the number of documents in the browsing history that contains this term within the selected input data source

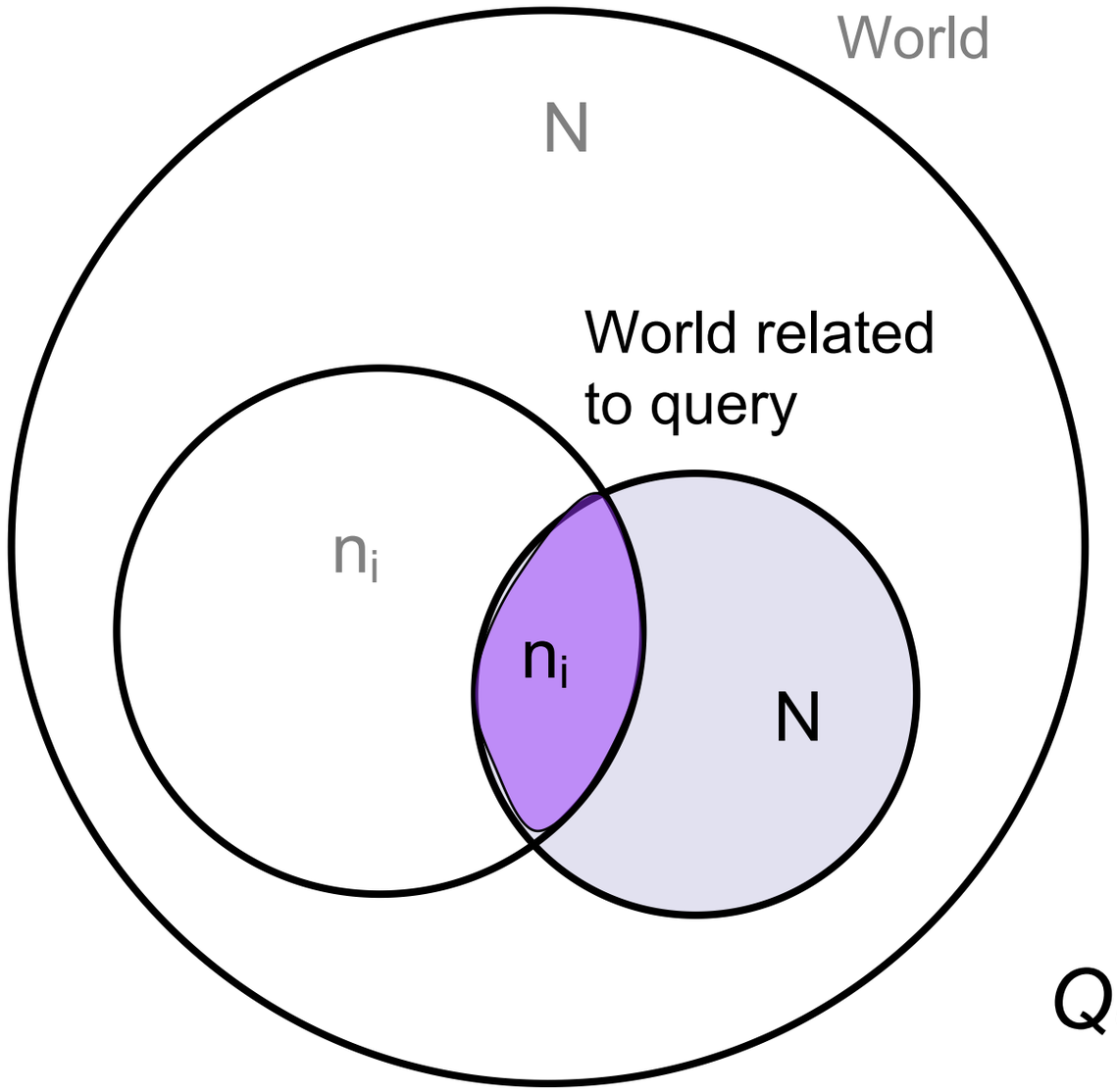


$$\text{Score} = \sum tf_i * w_i$$

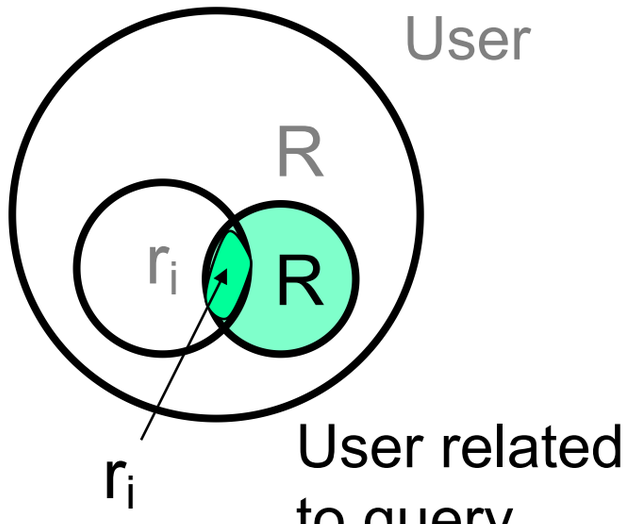


User Model as Relevance Feedback

- N : the number of documents on the web
- n_i: the number of documents in the corpus that contain the term t
- R : the number of documents in the user's browsing history
- r_i is the number of documents in the browsing history that contains this term within the selected input data source



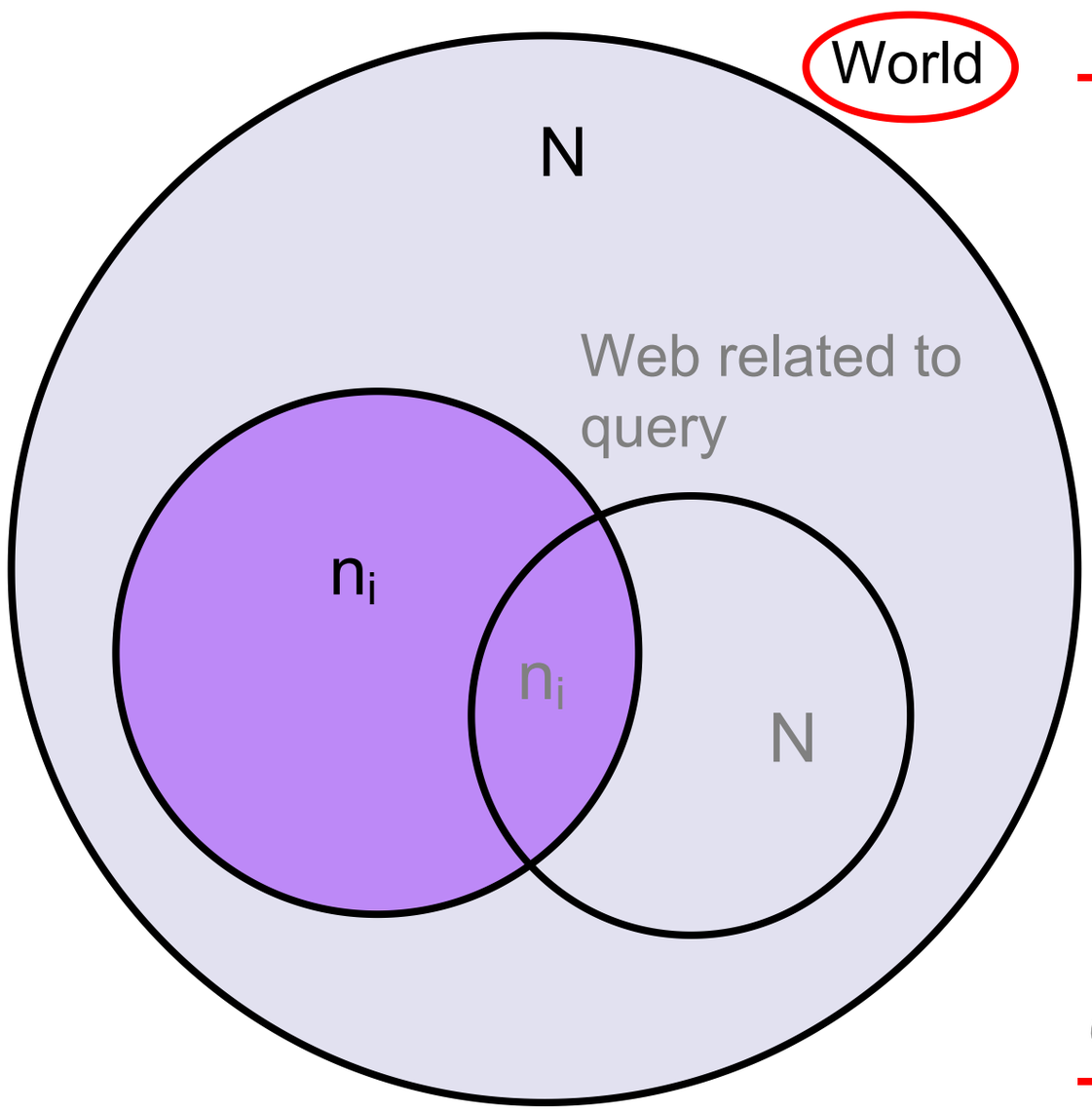
$$\text{Score} = \sum tf_i * w_i$$



Query Focused Matching

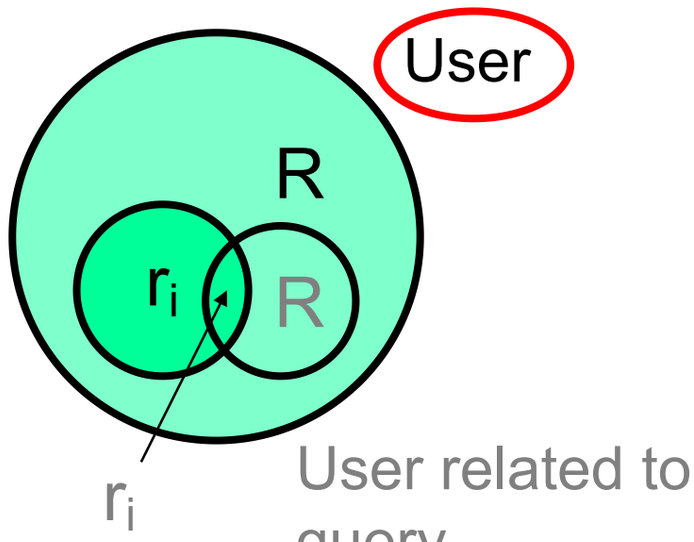
User Model as Relevance Feedback

- N : the number of documents on the web
- n_i: the number of documents in the corpus that contain the term t
- R : the number of documents in the user's browsing history
- r_i is the number of documents in the browsing history that contains this term within the selected input data source



World Focused Matching

$$Score = \sum tf_i * w_i$$



Query Focused Matching

What about personalized language models?

- Recap of language model
- Rank documents based on query likelihood

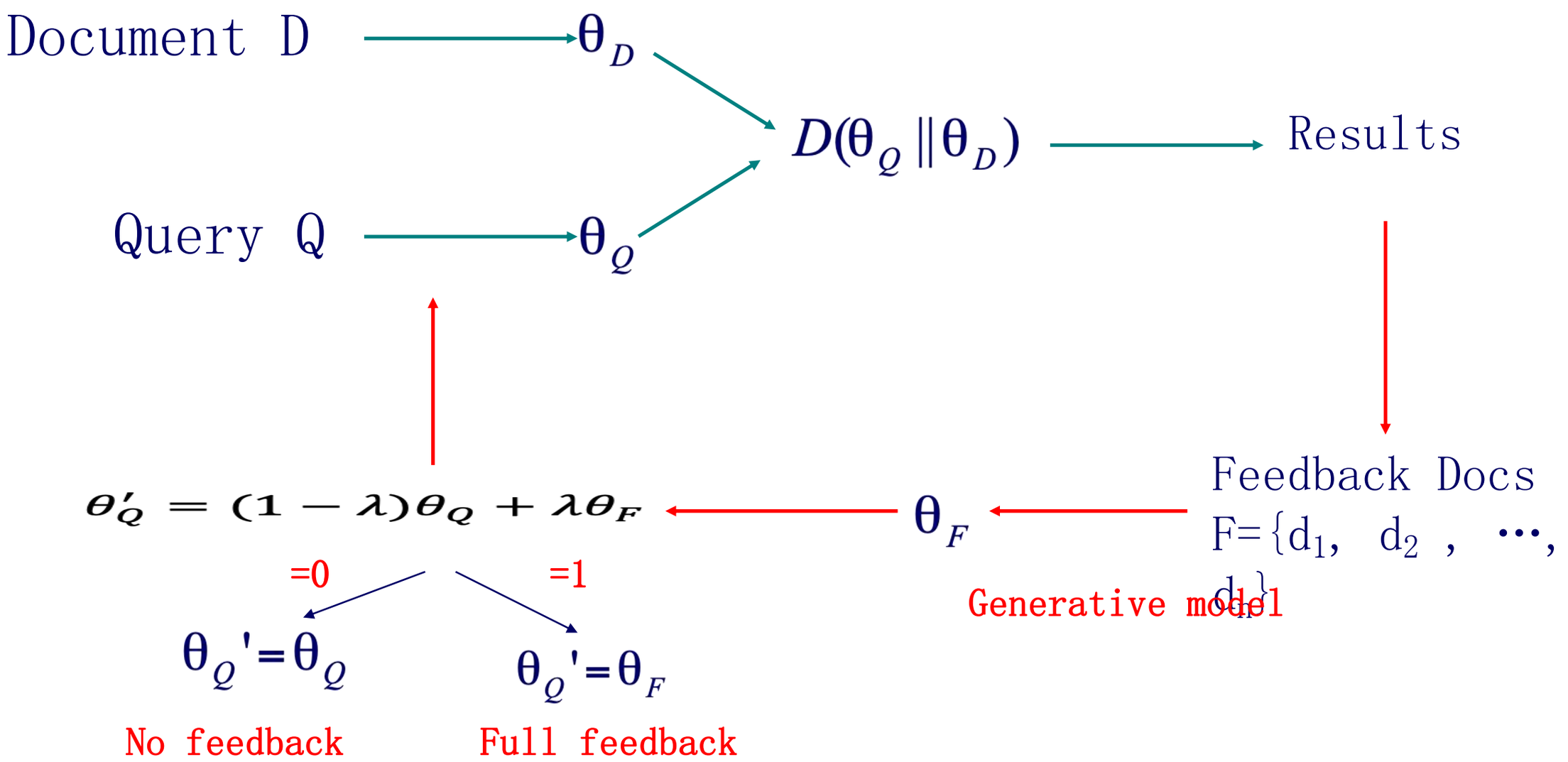
$$\log p(q \mid d) = \sum_{w_i \in q} \log p(w_i \mid d)$$

$$where,\ q = w_1 w_2 ... w_n$$

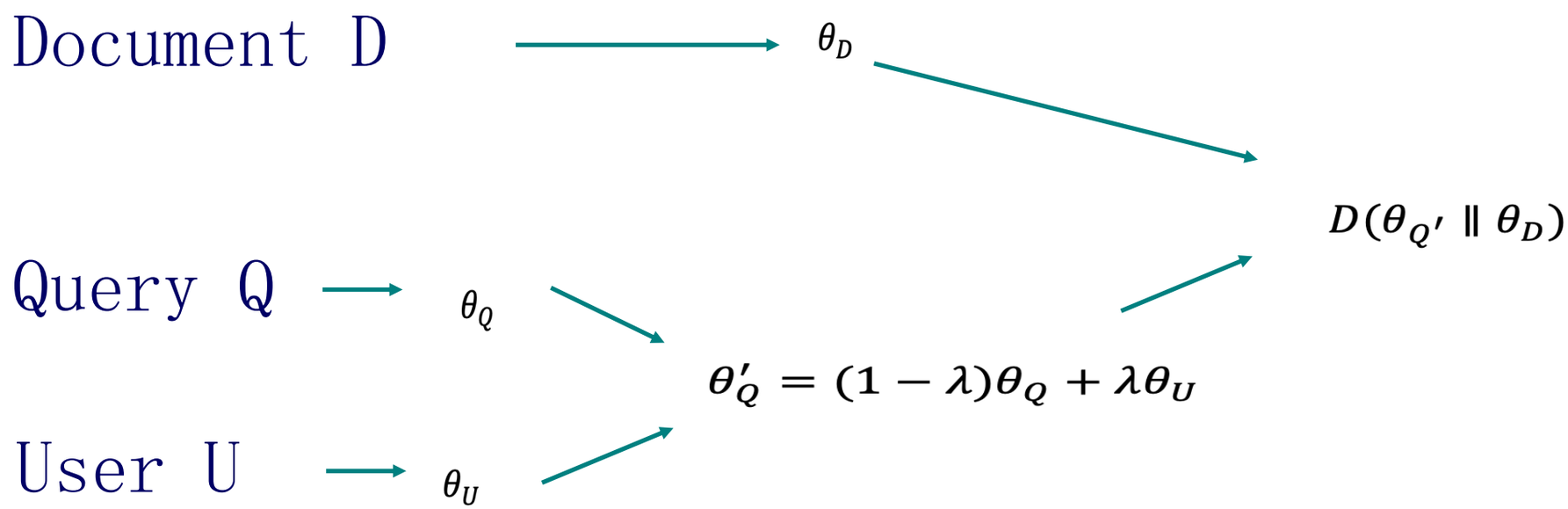
Document language model

- Difficulty: Documents are given, i.e., $p(w|d)$ is fixed
  - No way to update the language model of the query!

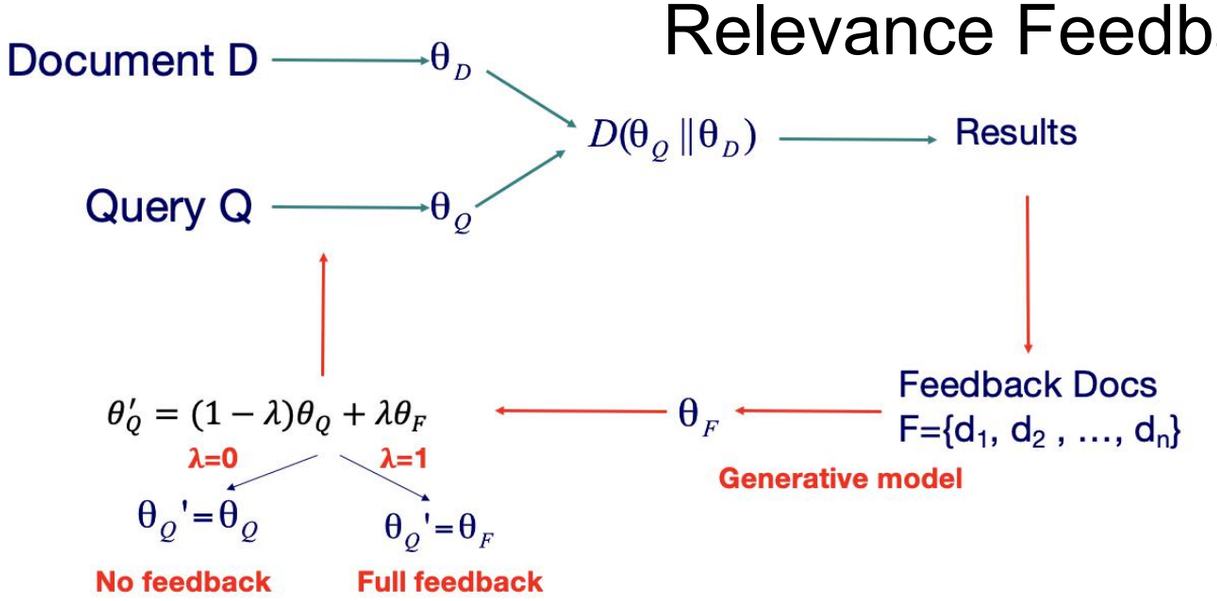
Recall how we incorporated feedback in language models…

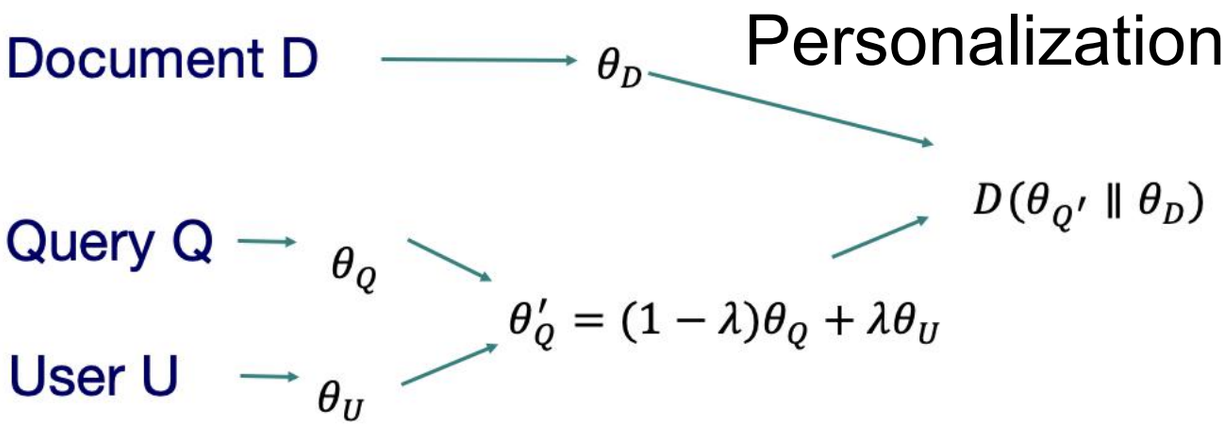


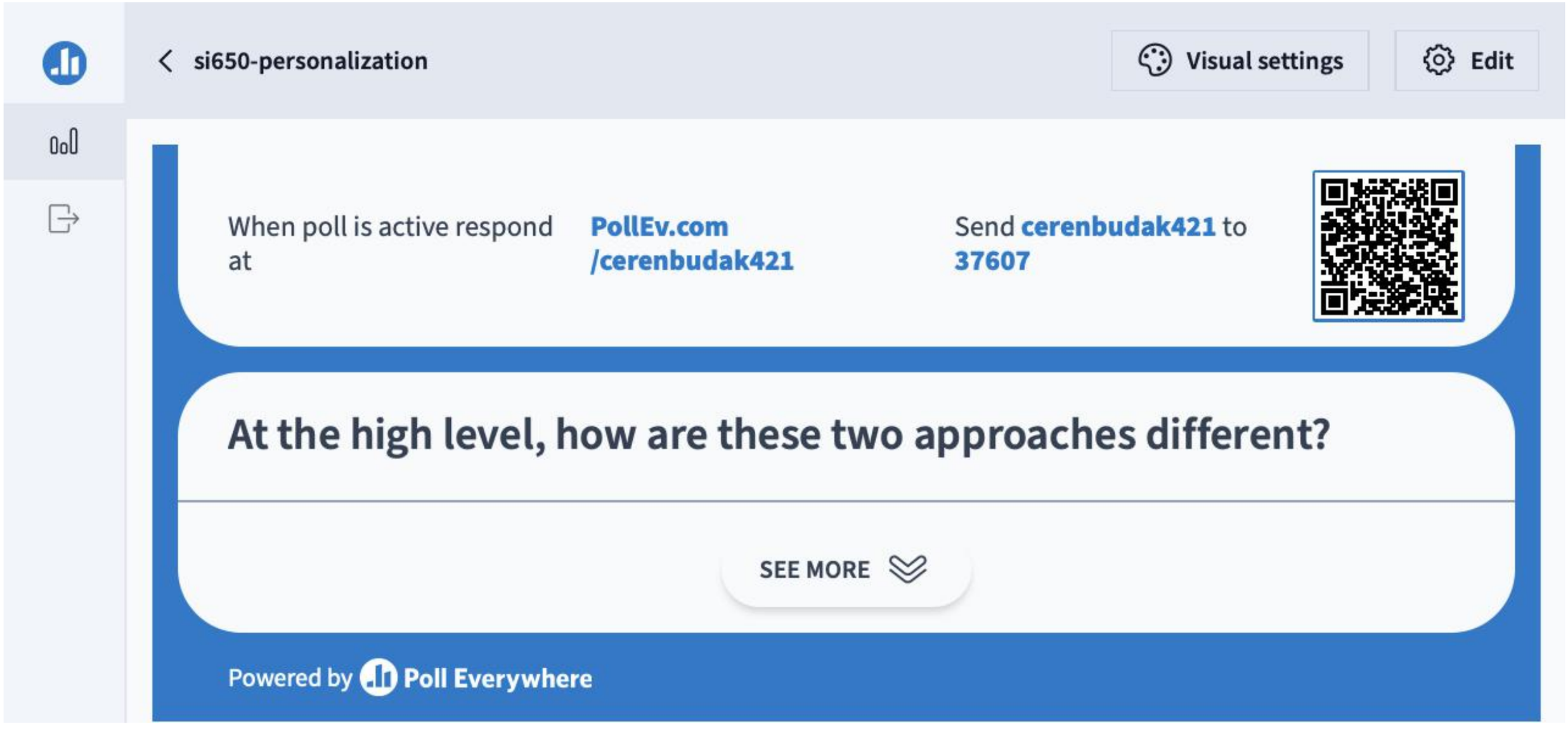
Personalization in language models



Spot the difference?







Reranking results after retrieval

- Chances are high that even for an ambiguous query the search engine will be quite successful in returning pages for the different meanings of the query.
- Idea: retrieve and re-rank the first 50 results retrieved for each query.

- Scoring Methods: $score_M(s_i) = \sum_{z=1}^{N_{s_i}} f_{t_z} \times w(t_z)$
  - To score a document represented by snippet s_i
    - For each unique word in the search snippet's title and summary (there are N_si of those)
    - f_{tz} : the number of occurrences of t_z within the snippet.
    - w_{tz} : the weight of t_i within user profile.

Personalization Strategies: Re-ranking Strategies

- Unique Matching

$$score_{UM}(s_i) = \sum_{z=1}^{N_{s_i}} w(t_z)$$

- Language Model
  - N_{si} is the total number of words in the snippet's title and summary.
  - w_{total} stands for the sum of all the weights within the user profile.

$$score_{LM}(s_i) = \sum_{z=0}^{N_{s_i}} \log\left(\frac{w(t_z)+1}{w_{total}}\right)$$

Personalization without text: PClick

- PClick assumes that for a query q submitted by a user u, the web pages frequently clicked by u in the past are more relevant to u.
- |Clicks(q, p, u)| is the number of clicks on web page p by user u for query q in the past.
- |Clicks(q,●,u)| is the total click number on query q by u
- β is a smoothing factor set to 0.5
- Note that PClick makes no use of the terms and weights associated to the user's profile and is only based on click-through data for a given query!

$$score_{PC}(s_i) = \frac{|Clicks(q, p, u)|}{|Clicks(q, \bullet, u)| + \beta}$$

Personalization Strategies: Re-ranking Strategies

- Rank and Visit Scoring
  - To adjust the snippet scores by two ways:
  - The snippet's original rank r_{si}

$$\text{finalScore}(s_i) = score(s_i) \times \frac{1}{1 + log(r_{s_i})}$$

  - This extends PClick in that it boosts all URLs that have previously been visited, while PClick only boosts URLs that have directly been clicked for the current search query.
  - The number of previous visits to that web page (n_i) times a factor v

$$\text{finalScore}(s_i) = score(s_i) * (1 + v \times n_i)$$

Evaluation Approach

- Step I.
  - Starting with an offline NDCG based evaluation to pick the optimal parameter configurations
- Step II.
  - Then we evaluate with the more realistic and harder online interleaved evaluation

How to choose how much to weight personalization?

- Consider the weighting: $\text{FinalScore}(s_i) = \alpha \cdot \text{BaseScore}(s_i) + (1 - \alpha) \cdot \text{PersonalScore}(s_i)$
- Reranking Performance
  - MaxNDCG:
    - which yielded the highest average NDCG score on the offine dataset;
  - MaxQuer
    - which improved the most queries;
  - MaxNoRank
    - the method with highest NDCG that does not take the original Google ranking into account
  - MaxBestPar
    - obtained by greedily selecting each parameter

Table 5: Selected personalization strategies. *Rel* indicates relative weighting, $v = 10$ indicates setting parameter v to 10 in Equation 10. For parameter descriptions, see Section 3.

Strategy	Profile Parameters							Ranking Parameters		
	Full Text	Title	Meta Keywords	Meta Descr.	Extracted Terms	Noun Phrases	Term Weights	Snippet Scoring	Google Rank	Urls Visited
MaxNDCG	–	Rel	Rel	–	–	Rel	TF-IDF	LM	1/log	v=10
MaxQuer	–	–	–	–	Rel	Rel	TF	LM	1/log	v=10
MaxNoRank	–	–	Rel	–	–	–	TF	LM	–	v=10
MaxBestPar	–	Rel	Rel	–	Rel	–	pBM25	LM	1/log	v=10

Offline Evaluation of performance

Table 6: Summary of offline evaluation results

Method	Average NDCG	+/=/− Queries
Google	0.502 ± 0.067	–
Teevan	0.518 ± 0.062	44/0/28
PClick	0.533 ± 0.057	13/58/1
MaxNDCG	$\mathbf{0.573 \pm 0.042}$	48/1/23
MaxQuer	0.567 ± 0.045	**52/2/18**
MaxNoRank	0.520 ± 0.060	13/52/7
MaxBestPar	0.566 ± 0.044	45/5/22

- Compare Average NDCG
  - MaxNDCG & MaxQuer are better than others.
  - Although MaxNoRank is better than Google,But this maybe a result of overfitting the parameters given the small offline dataset.
- Compare numbers of query
  - How about PClick ? Because PClick only works on repeated queries.

Evaluation Approach - Offline Evaluation

- Parameter Effect:
  - Profile parameter
  - Ranking parameter

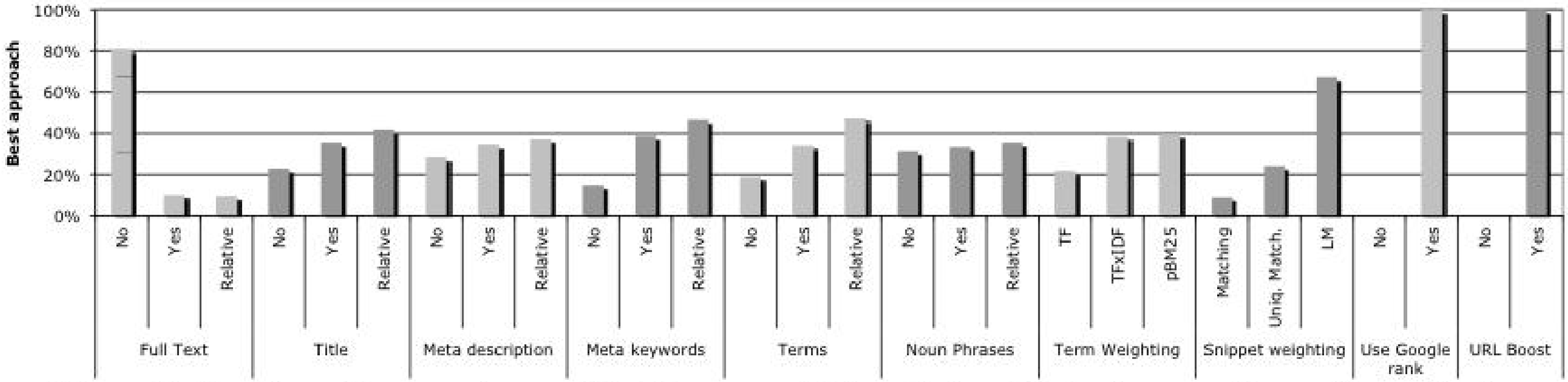


Figure 3: Fraction of parameter combinations on which each investigated parameter performs best.

Offline Evaluation

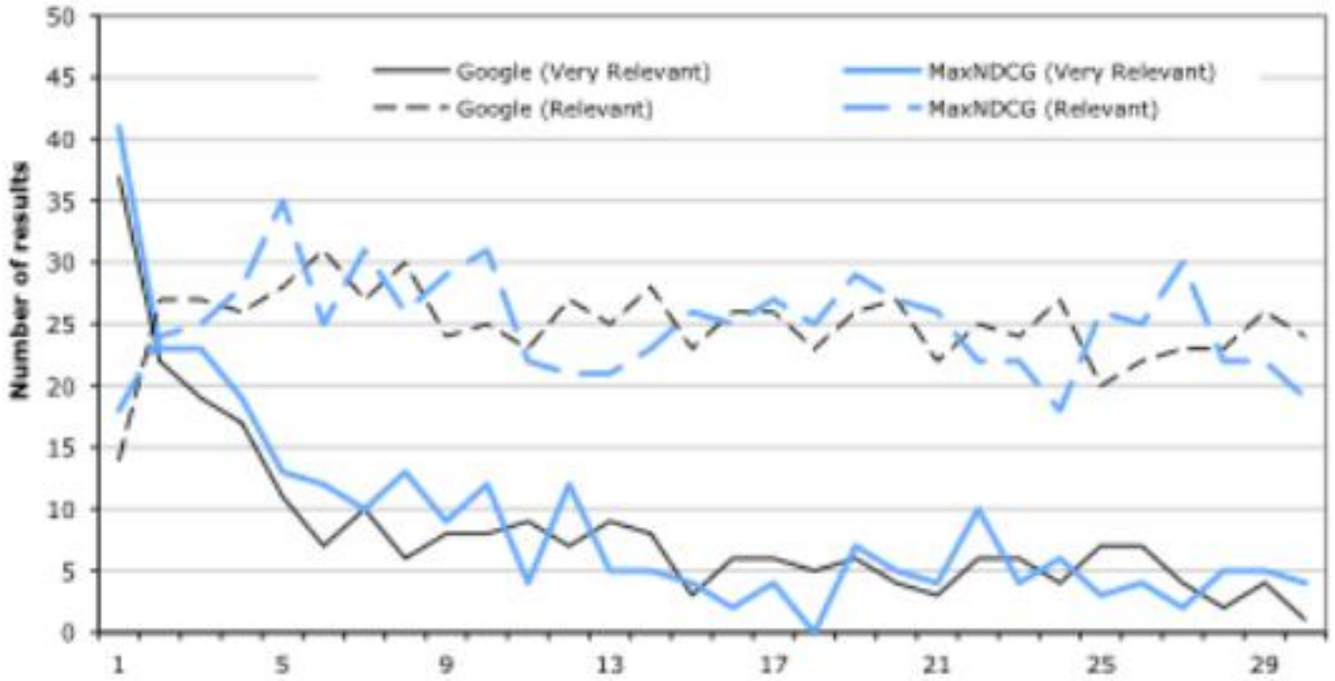


Figure 2: Distribution of relevance at rank for the Google and MaxNDCG rankings

- The personalization strategy considers the Google rank and is less aggressive at high ranks.

Online Evaluation with real users

- Interleaving Implementation
  - Is used to produce a combined ranking.
  - To avoid presenting slightly different rankings every time a search page is refreshed.

Algorithm 1 Team-Draft Interleaving [18]

Input: Rankings $A = (a_1, a_2, \dots)$ and $B = (b_1, b_2, \dots)$
Init: $I \leftarrow (); TeamA \leftarrow \emptyset; TeamB \leftarrow \emptyset;$
while $(\exists i : A[i] \notin I) \wedge (\exists j : B[j] \notin I)$ **do**
  if $(|TeamA| < |TeamB|) \vee$
    $((|TeamA| = |TeamB|) \wedge (RandBit() = 1))$ **then**
    $k \leftarrow \min_i\{i : A[i] \notin I\}$ ... *top result in A not yet in I*
    $I \leftarrow I + A[k];$ ..... *append it to I*
    $TeamA \leftarrow TeamA \cup \{A[k]\}$ ..... *clicks credited to A*
  else
    $k \leftarrow \min_i\{i : B[i] \notin I\}$ ... *top result in B not yet in I*
    $I \leftarrow I + B[k]$ ..... *append it to I*
    $TeamB \leftarrow TeamB \cup \{B[k]\}$ .... *clicks credited to B*
  end if
end while
Output: Interleaved ranking I, $TeamA$, $TeamB$

Online Evaluation with real users

Table 7: Results of online interleaving test

Method	Queries	Google Vote	Re-ranked Vote
MaxNDCG	2,090	624 (39.5%)	955 (60.5%)
MaxQuer	2,273	812 (47.3%)	905 (52.7%)
MaxBestPar	2,171	734 (44.8%)	906 (55.2%)

- MaxNDCG is best, matching the offline finding.

Online Evaluation with real users

Table 8: Queries impacted by search personalization

Method	Unchanged	Improved	Deteriorated
MaxNDCG	1,419 (67.9%)	500 (23.9%)	171 (8.2%)
MaxQuer	1,639 (72.1%)	423 (18.6%)	211 (9.3%)
MaxBestPar	1,485 (68.4%)	467 (21.5%)	219 (10.1%)

- The Unchanged column indicates the number of queries for which the clicked result was at the same rank for both the non-personalized and the personalized ranking.
- The Improved column shows how often the clicked result was brought up.
- The Deteriorated column shows the number of queries for which the clicked result was pushed down.

Online Evaluation with real users

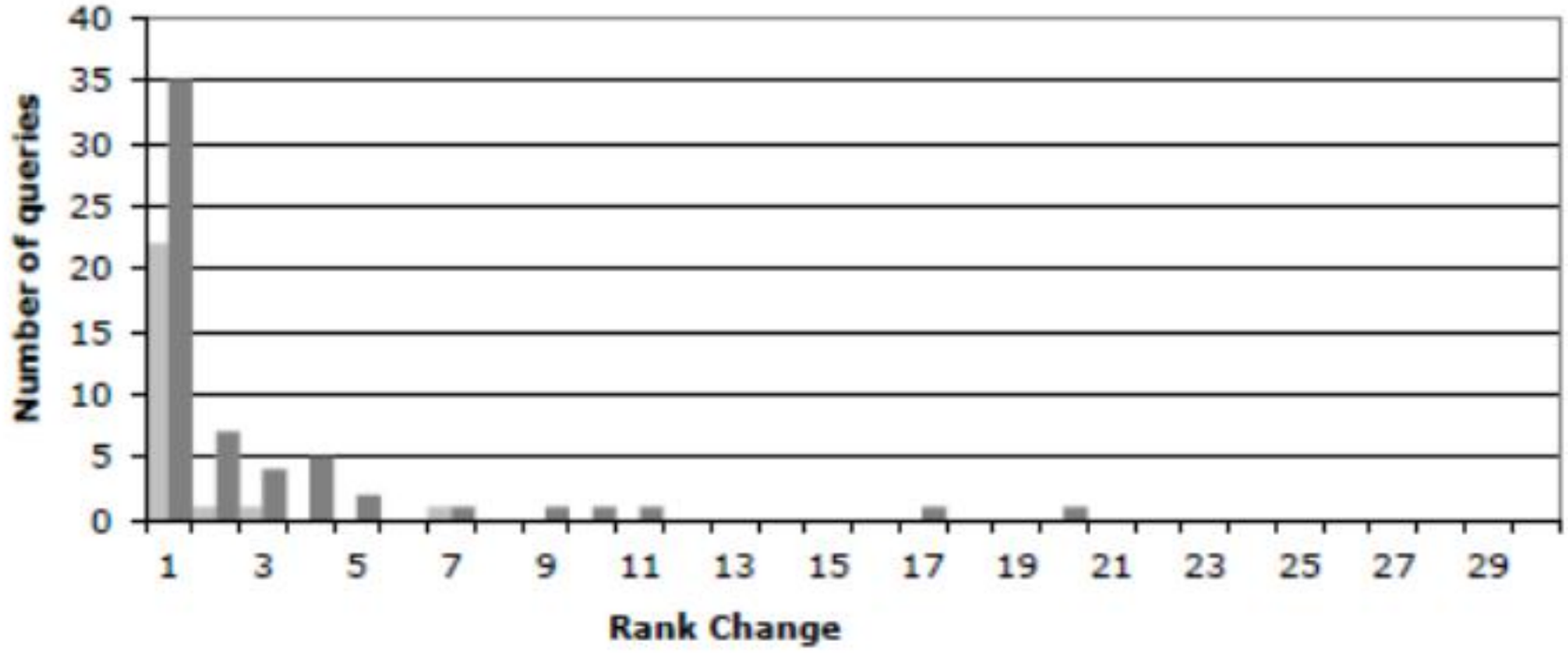


Figure 4: Rank differences for deteriorated (light grey) and improved queries (dark grey) for MaxNDCG

So which strategy to use to select method/parameters?

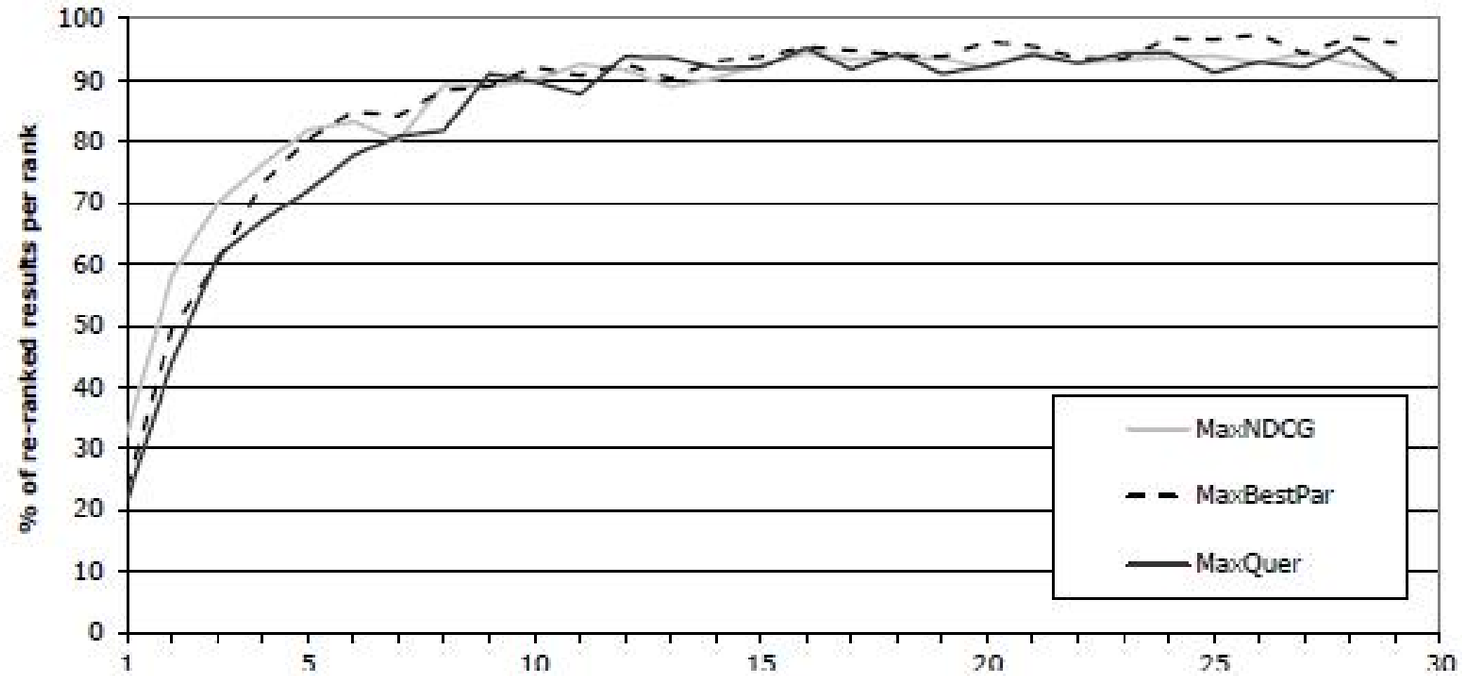


Figure 5: Degree of personalization per rank

- MaxNDCG is the most effective approach to select personalization parameters

Personalization via Location

User location

- User location is one of the most important features for personalization
  - Country
    - Query [football] in the US vs the UK
  - State/Metro/City
    - Queries like [zoo], [craigslist], [giants]
  - Fine-grained location
    - Queries like [pizza], [restaurants], [coffee shops]

Challenges

- Not all queries are location sensitive
  - [facebook] is not asking for the closest Facebook office
  - [seaworld] is not necessarily asking for the closest SeaWorld

- Different parts of a site may be more or less location sensitive
  - NYTimes home page vs NYTimes Local section

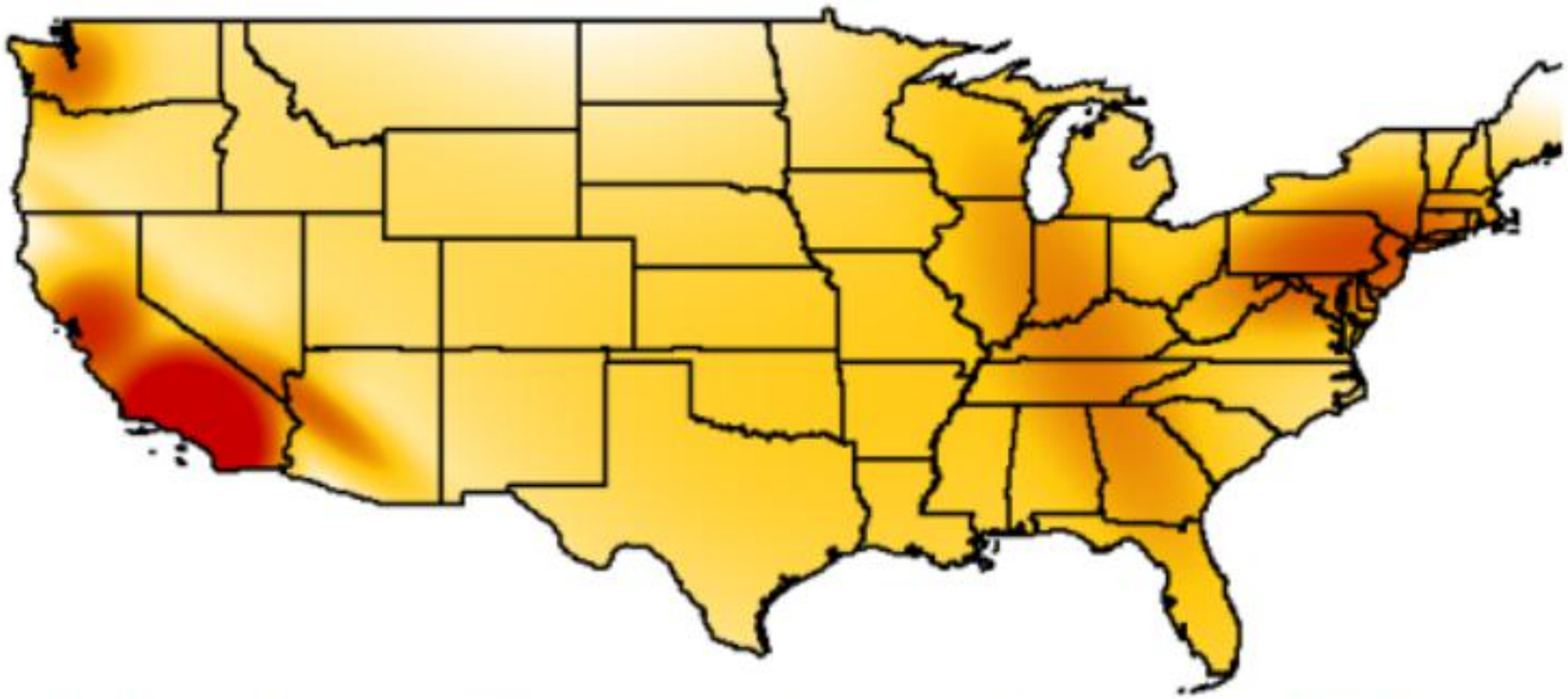
- Addresses on a page don’t always tell us how location sensitive the page is
  - UMich home page has address, but not location sensitive

Key idea

- Usage statistics, rather than locations mentioned in a document, best represent where it is relevant [Bennett et al. 2011]
  - I.e., if users in a location tend to click on that document, then it is relevant in that location
- User location data is acquired from anonymized logs (with user consent, e.g., from a widely distributed browser extension)
  - User IP addresses are resolved into geographic location information

Location interest model

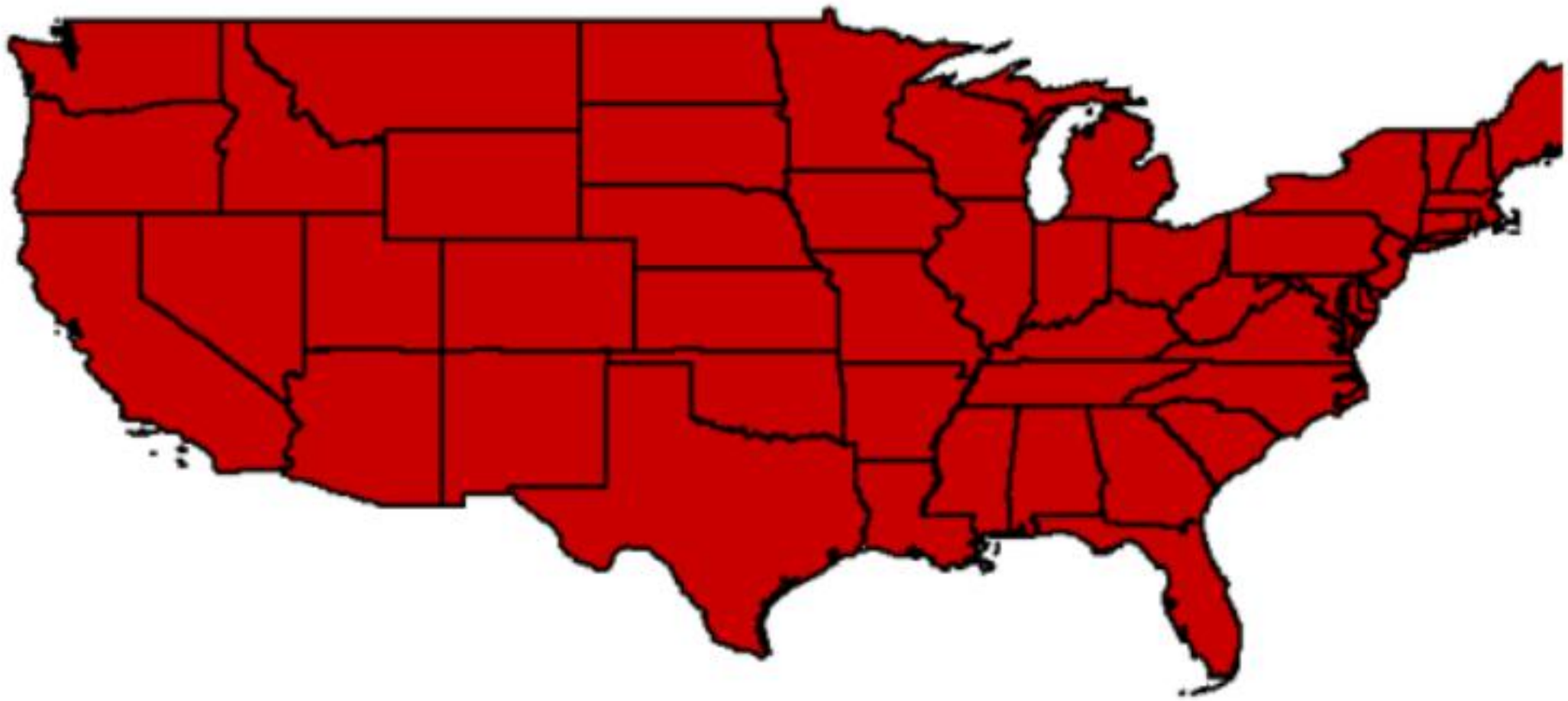
- Use the logs data to estimate the probability of the location of the user given they viewed this URL: $P(location=x \mid URL)$



(c) Los Angeles Times: Reviews and Recommendations
http://findlocal.latimes.com/

Location interest model

- Use the logs data to estimate the probability of the location of the user given they viewed this URL: $P(location=x \mid URL)$



(d) Los Angeles Times: Crossword Puzzles and Games
http://games.latimes.com/

Learning the location interest model

- For compactness, represent location interest model as a mixture of 5-25 2-d Gaussians (x is [lat, long])

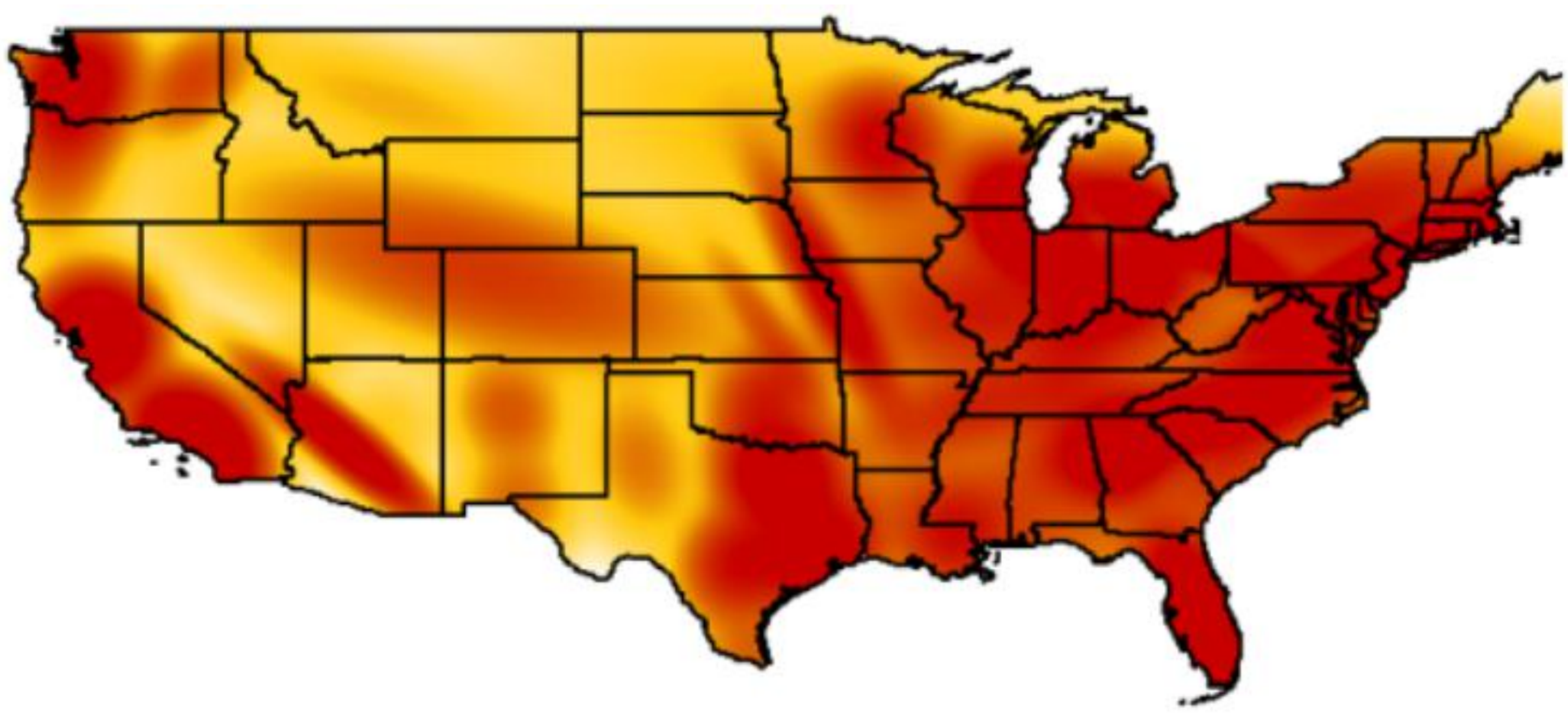
$$P(location = x|URL) = \sum_{i=1}^{n} w_i N(x; \mu_i, \Sigma_i)$$

- Learn Gaussian mixture model using EM
  - Expectation step: Estimate probability that each point belongs to each Gaussian
  - Maximization step: Estimate most likely mean, covariance, weight

- As a result, predict the **most likely location clusters** for any new user visiting a URL.

More location interest models

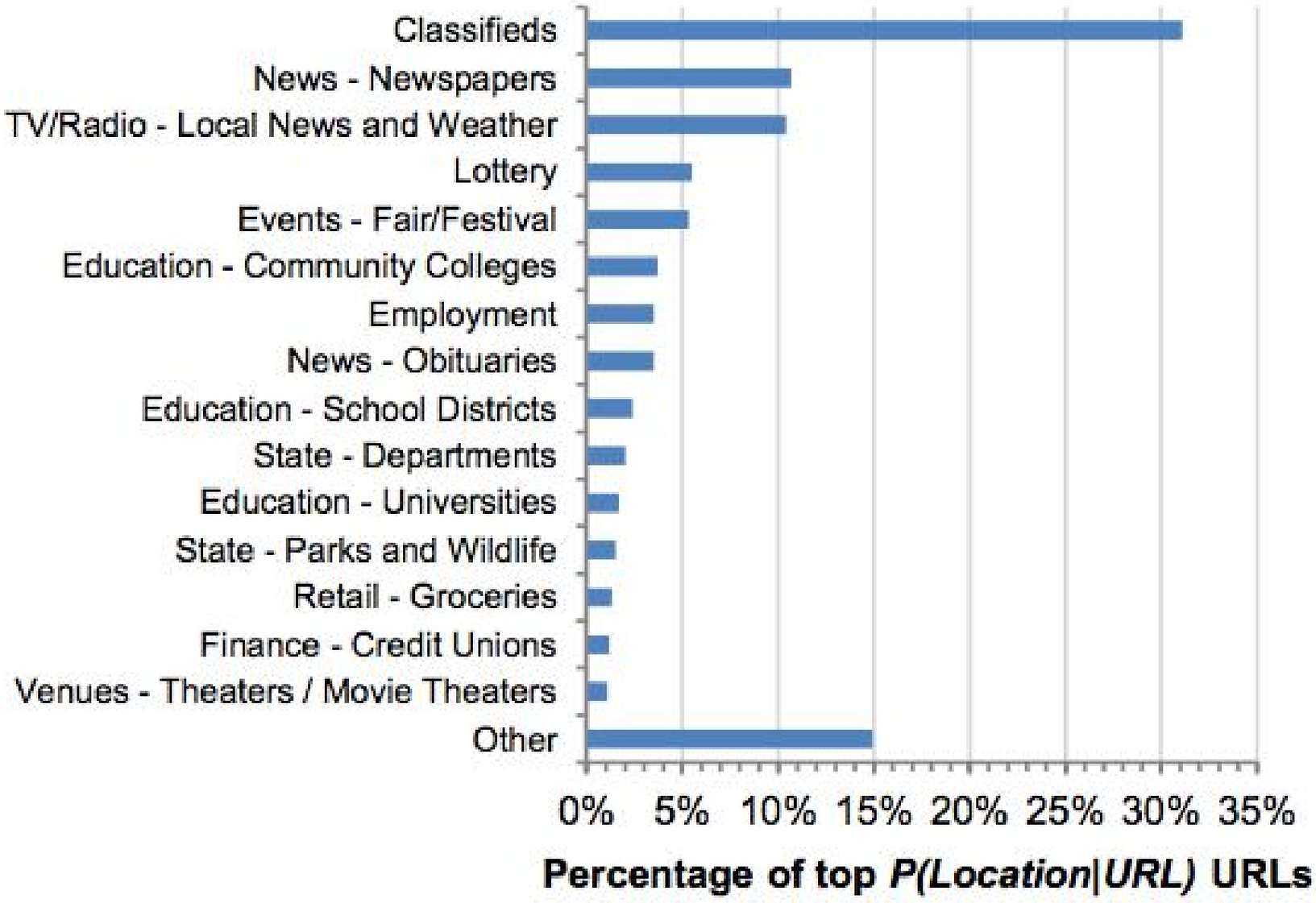
- Learn a location-interest model for queries
  - Using location of users who issued the query
- Learn a background model showing the overall density of users



(e) Background Model

Topics in URLs with high P(user location | URL)

What query topics would be most local for you?



Location sensitive features

- Non-contextual features (user-independent)
  - Is the query location sensitive? What about the URLs?
  - Feature: Entropy of the location distribution
    - Low entropy means distribution is peaked and location is important
  - Feature: KL-divergence between location model and background model
    - High KL-divergence suggests that it is location sensitive
  - Feature: KL-divergence between query and URL models
    - Low KL-divergence suggests URL is more likely to be relevant to users issuing the query

More location sensitive features

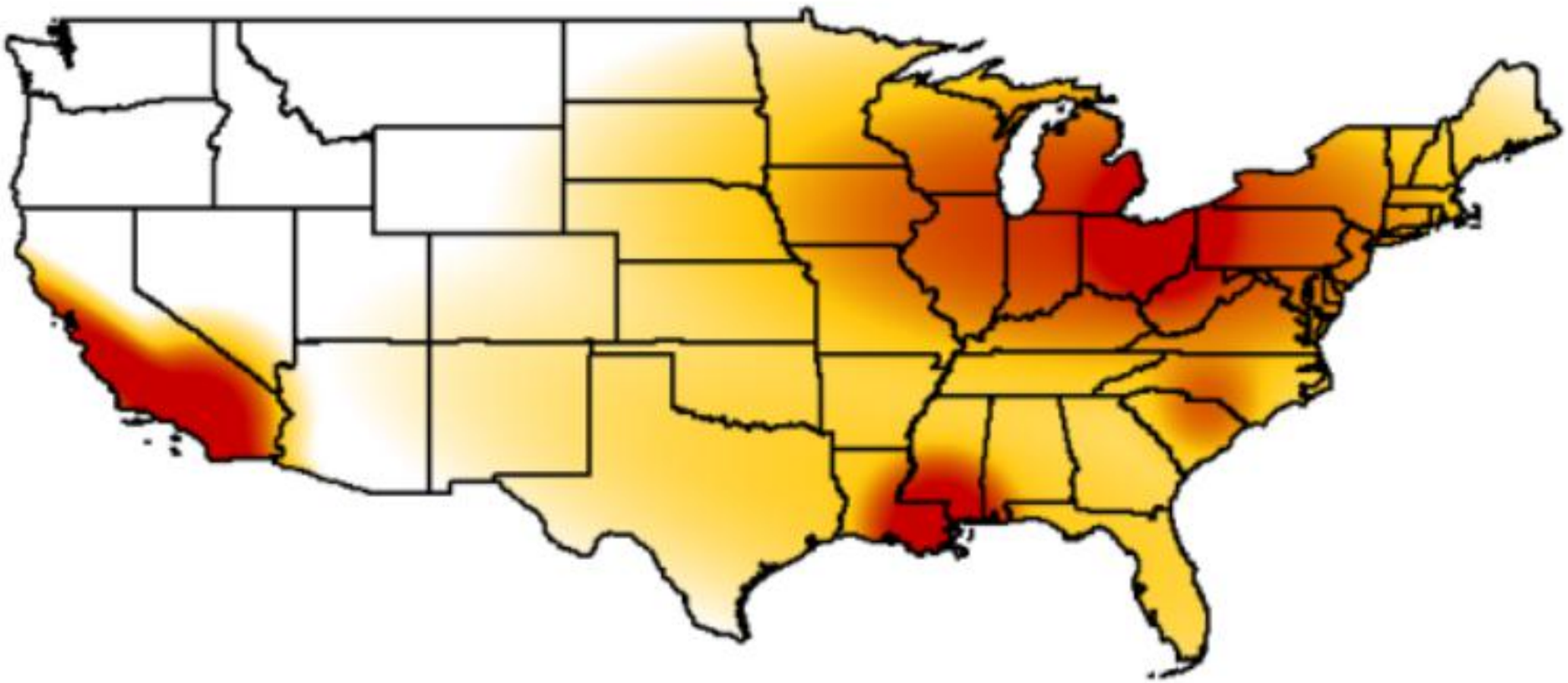
- Contextual features (user-dependent)
  - Feature: User's location (naturally!)
  - Feature: Probability of the user's location given the URL
    - Computed by evaluating URL's location model at user location
    - Feature is high when user is at a location where URL is popular
    - Downside: large population centers tend to higher probabilities for all URLs
  - Feature: Use Bayes rule to compute P(URL | user location)
  - Feature: Also create a normalized version of the above feature by normalizing with the background model
  - Features: Versions of the above with query instead of URL

Learning to rank

- Add location features (in addition to standard features) for machine learned ranking
  - Training data derived from logs
  - P(URL | user location) turns out to be an important feature
  - KL divergence of the URL model from the background model also plays an important role

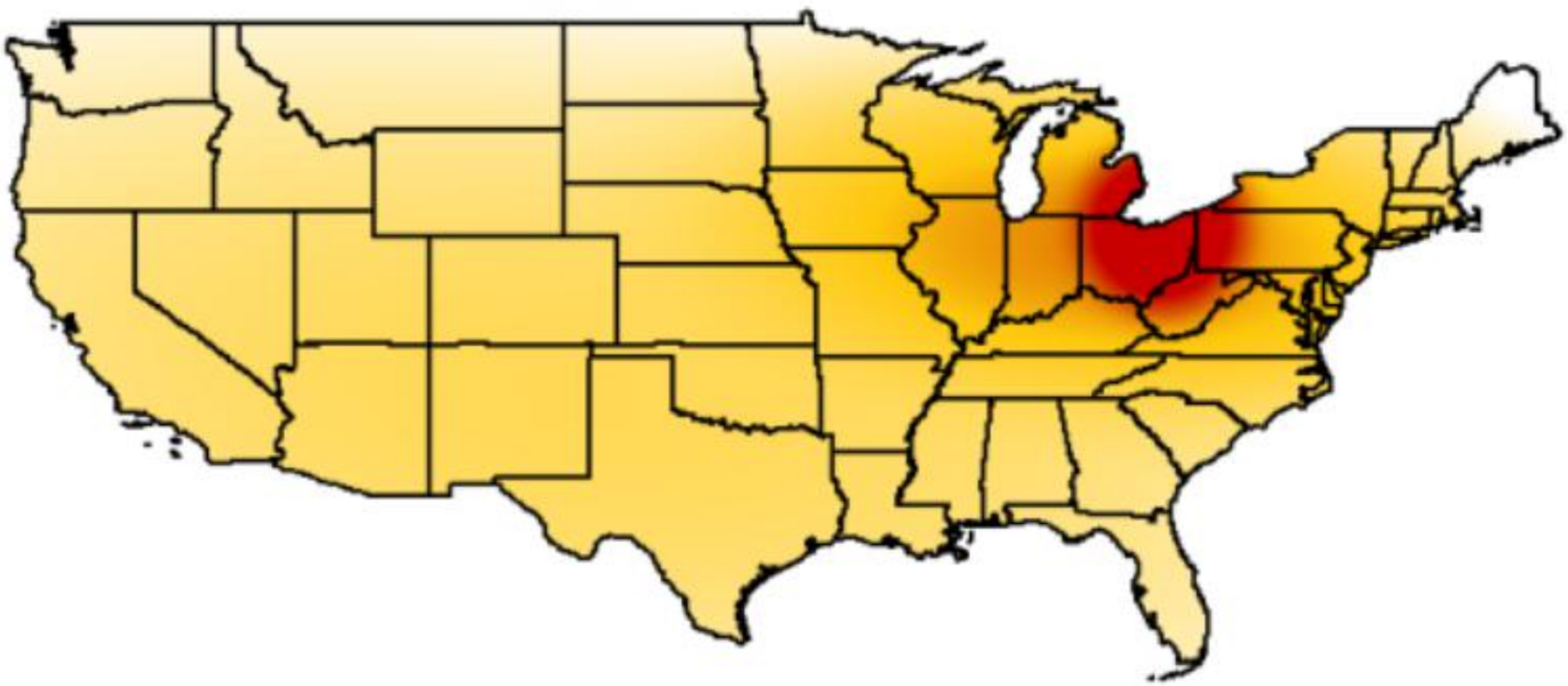
Query model for rta bus schedule

User in New Orleans



URL model for top original result

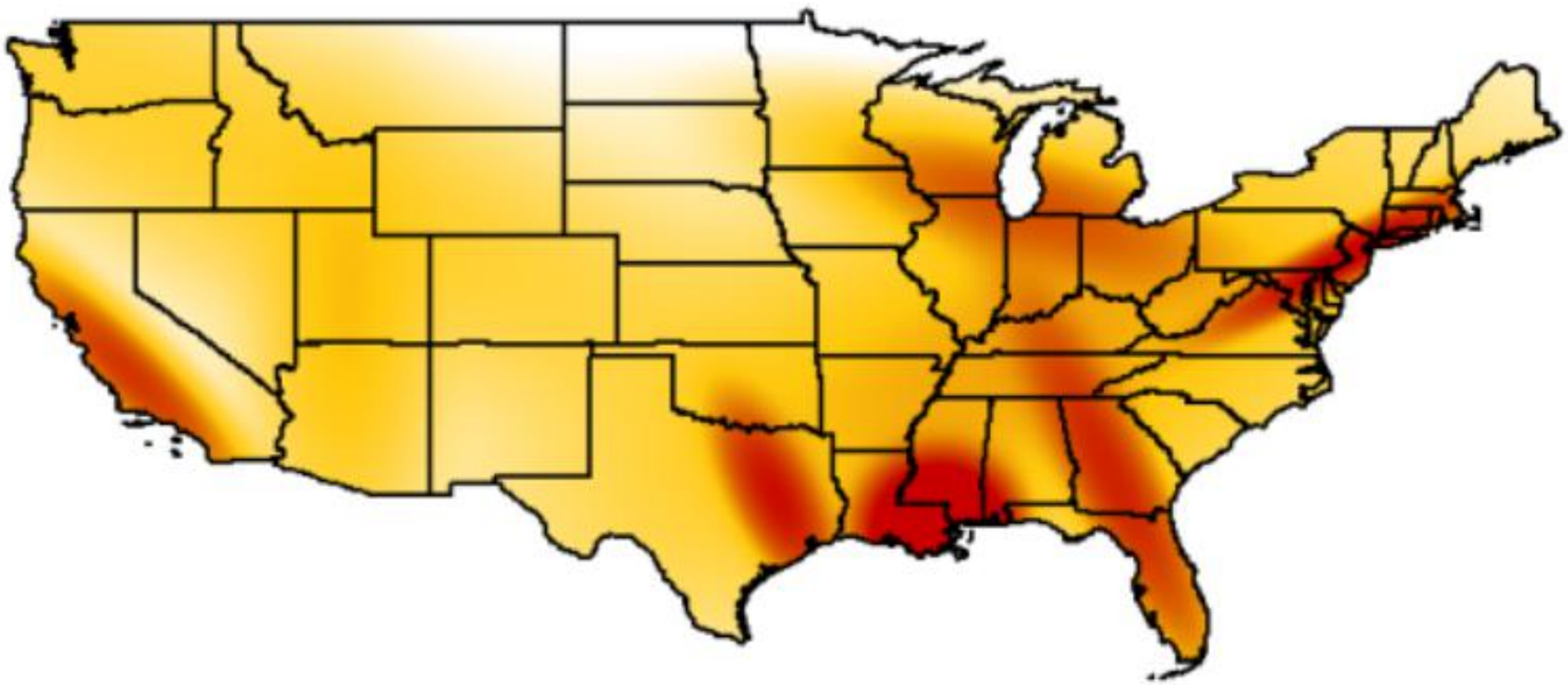
User in New Orleans



(a) http://www.riderta.com/maps-schedules.asp

URL model for promoted URL

User in New Orleans



(b) http://www.norta.com/

Group Analysis for Personalization

People Express Things Differently

- Differences can be a challenge for Web search
- Personalization
  - Closes the gap using more about the person
- *Group*ization
  - Closes the gap using more about the *group*

Identifying Groups

- Explicitly
  - Tasks: Tools for collaboration [Morris & Horvitz 2007]
  - Traits: Profiles
- Implicitly
  - Interests: Sites visited, queries
  - Tasks: Query
  - Location: IP address [Mei & Church 2008]
  - Gender: Queries [Jones et al. 2007]
  - Interesting area to explore: Social networks

How to Take Advantage of Groups?

- Who do we share interests with?
- Do we talk about things similarly?
- What algorithms should we use?

Related Work

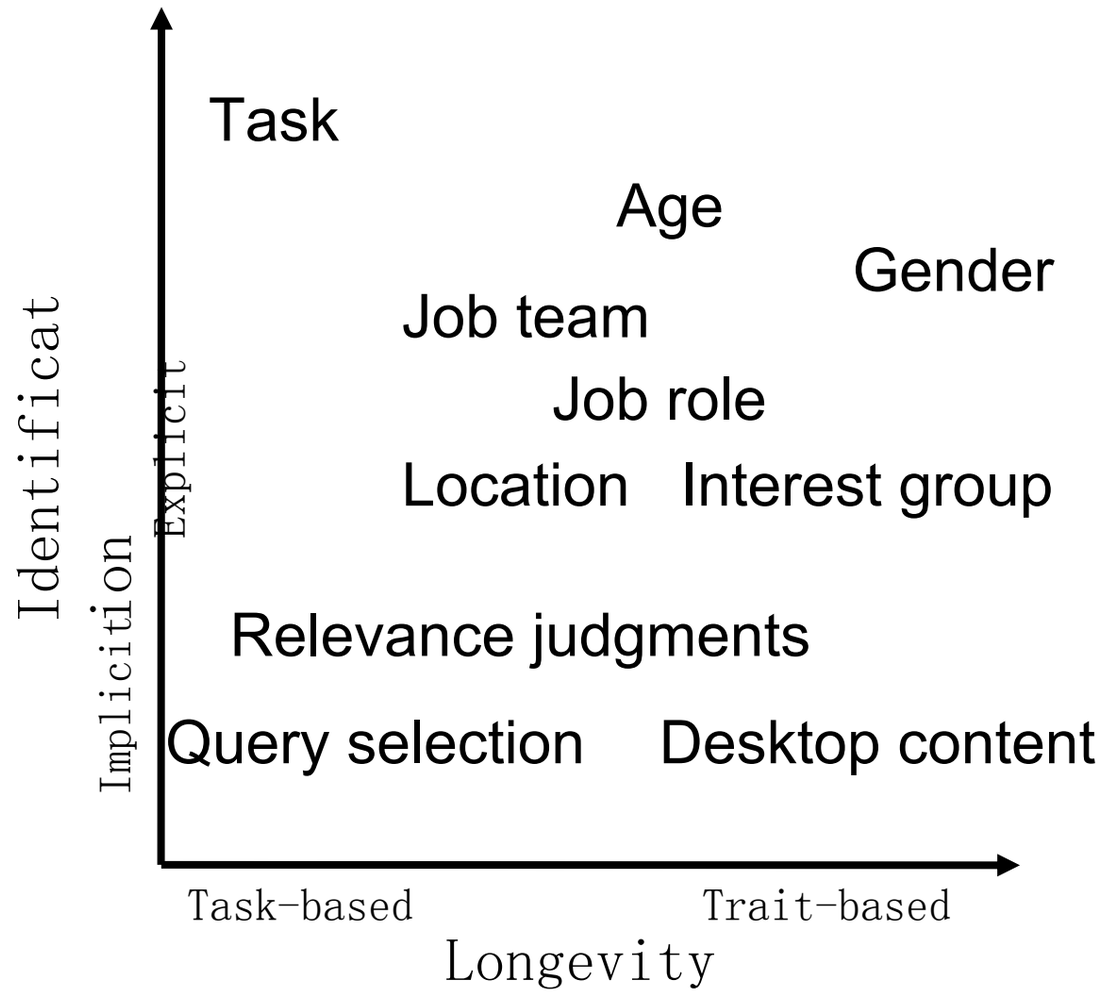
- Personalization
  - Implicit information valuable [Dou et al. 2007; Shen et al. 2005]
  - More data = better performance [Teevan et al. 2005]
- Collaborative filtering & recommender systems
  - Identify related groups
    - Browsed pages [Almeida & Almeida 2004; Sugiyama et al. 2005]
    - Queries [Freyne & Smyth 2006; Lee 2005]
    - Location [Mei & Church 2008], company [Smyth 2007], etc.
  - Use group data to fill in missing personal data
    - Typically data based on user behavior

How One Can Answer these Questions

- Who do we share interests with?
  - Similarity in query selection
  - Similarity in what is considered relevant
- Do we talk about things similarly?
  - Similarity in user profile
- What algorithms should we use?
  - Groupize results using groups of user profiles
  - Evaluate using groups' relevance judgments

There are many types of groups!

- Group longevity
  - Task-based
  - Trait-based
- Group identification
  - Explicit
  - Implicit



Teevan et al. (2009) studied different types of groups

Trait-based dataset

- 110 people
  - Work
  - Interests
  - Demographics
- Microsoft employees

Task-based dataset

- 10 groups x 3 (= 30)
- Know each other
- Have common task
  - “Find economic pros and cons of telecommuting”
  - “Search for information about companies offering learning services to corporate customers”

Queries Studied

Trait-based dataset

- Challenge
  - Overlapping queries
  - Natural motivation
- Queries picked from 12
  - Work

    c# delegates, live meeting
  - Interests

    bread recipes, toilet train dog

Task-based dataset

- Common task
  - Telecommuting v. office

    pros and cons of working in an office

    social comparison telecommuting versus office

    telecommuting

    working at home cost benefit

Data Collected

- Queries evaluated
- Explicit relevance judgments
  - 20 - 40 results
  - Personal relevance
    - *Highly relevant*
    - *Relevant*
    - *Not relevant*
- User profile: Desktop index

Who do we share interests with?

- Variation in query selection
  - Work groups selected similar work queries
  - Social groups selected similar social queries
- Variation in relevance judgments
  - Judgments varied greatly (κ =0.08)
  - Task-based groups most similar
  - Similar for one query ≠ similar for another

Do we talk about things similarly?

- Group profile similarity
  - Members more similar to each other than others
  - Most similar for aspects related to the group

	In task group	Not in group	Difference
All queries	0.42	0.31	34%
Group queries	0.77	0.35	120%

- Clustering profiles recreates groups
- Index similarity ≠ judgment similarity
  - Correlation coefficient of 0.09

What algorithms should we use?

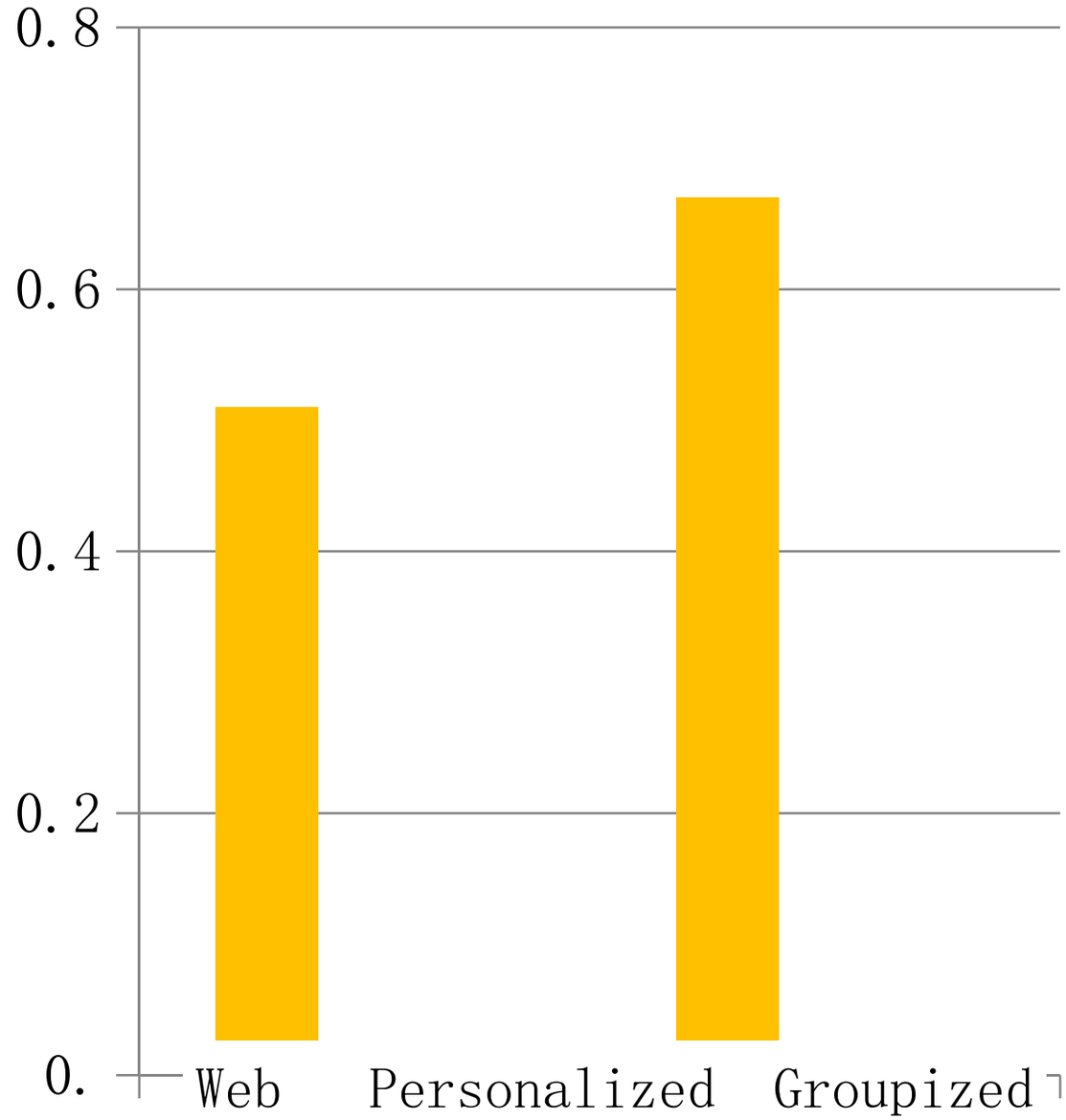
- Calculate personalized score for each member using Teevan et al. (2005)
  - Content: User profile as relevance feedback

$$\sum_{term_i} tf_i \log\left(\frac{(r_i + 0.5)(N - n_i - R + r_i + 0.5)}{(n_i - r_i + 0.5)(R - r_i + 0.5)}\right)$$

  - Behavior: Previously visited URLs and domains
- Sum personalized scores across group
- Produces same ranking for all members

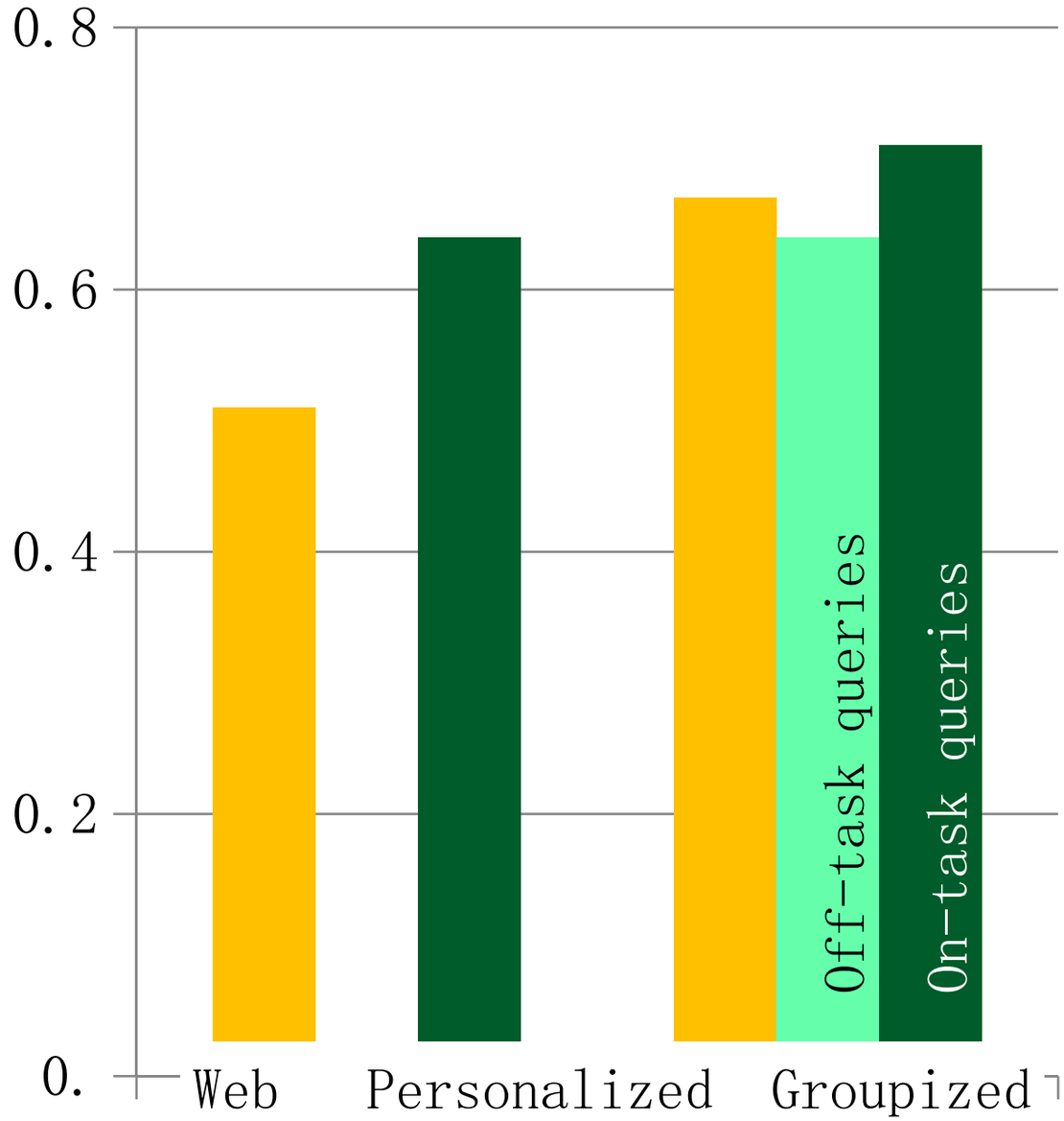
Performance: Task-Based Groups

- Personalization improves on Web
- Groupization gains +5%

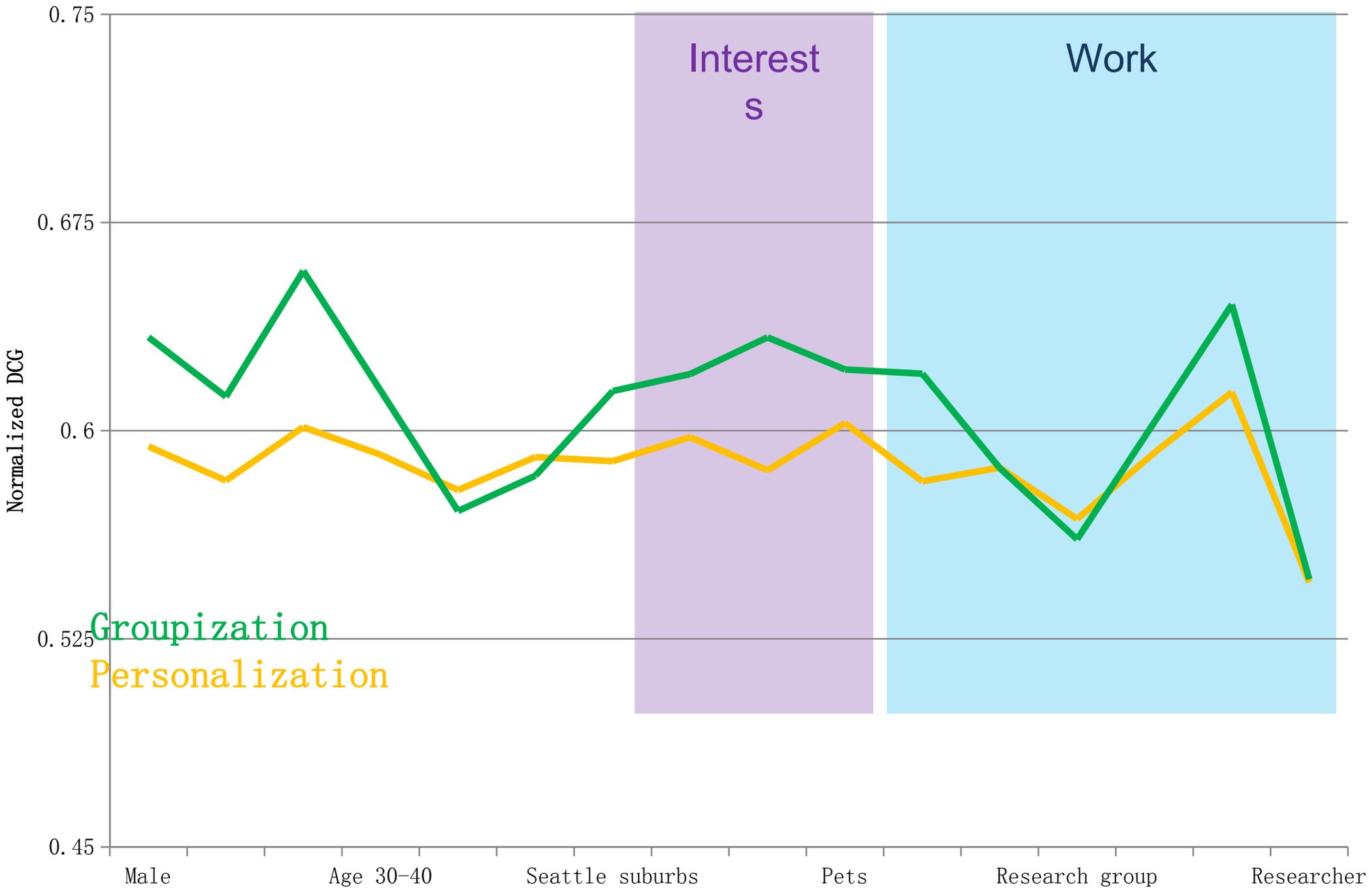


Performance: Task-Based Groups

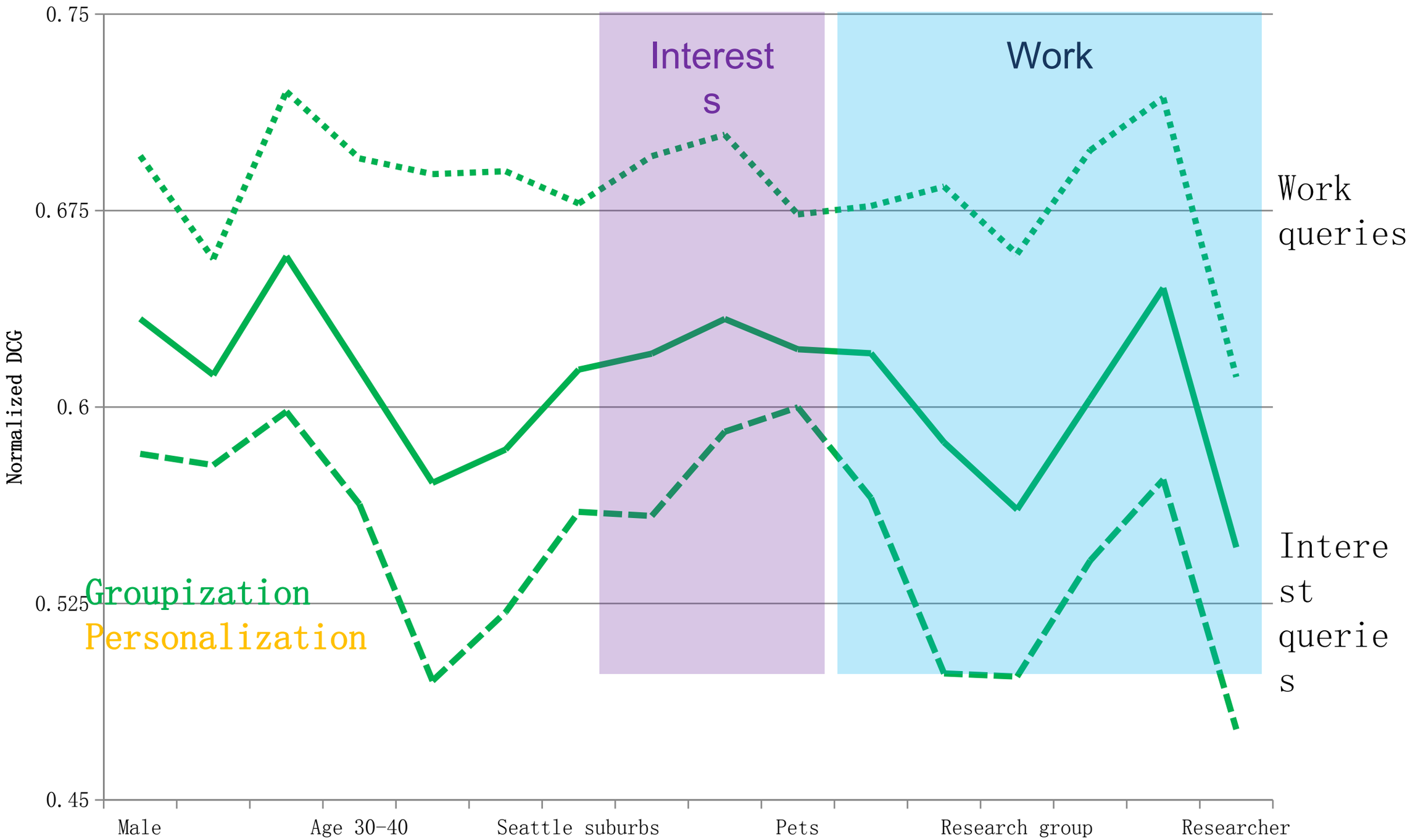
- Personalization improves on Web
- Groupization gains +5%
- Split by query type
  - On-task v. off-task
  - Groupization the same as personalization for off-task queries
  - 11% improvement for on-task queries



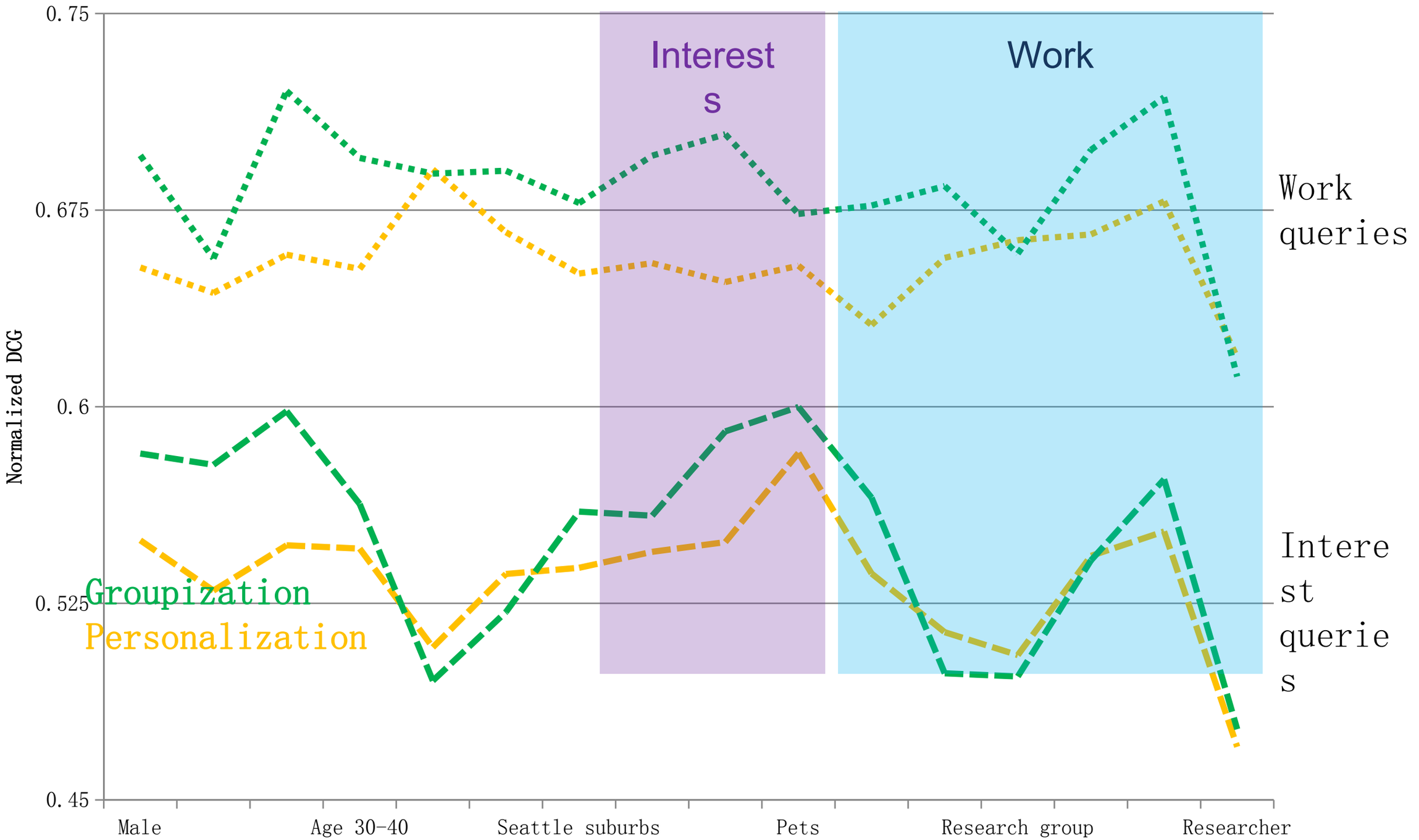
Performance: Trait-Based Groups



Performance: Trait-Based Groups



Performance: Trait-Based Groups



Groupization take aways

- Who do we share interests with?
  - Depends on the task
- Do we talk about things similarly?
  - Variation in profiles even with similar judgments
- What algorithms should we use?
  - Groupization can take advantage of variation for group-related tasks

What should you know about personalization?

- Queries are ambiguous but users are internally clear in what they want
- Personalization can help resolve ambiguity
- Can incorporate personalization directly into VSMs and Language Models through incorporating what users previously looked at
- Can personalized based on previous behavior too—no need to restrict to text
- Metadata like Location and Groups provide powerful signals
- Personalization features are very common components of Learning to Rank

Project Update

Things to know for the update

- Three big components:
  - You need to have all* your data in place
  - 20* query* annotations per person
    - 10 train/validation
    - 10 test
    - queries need to be *interesting*
- You should score a baseline system on the 10 test queries *and* a naive system (more later)
  - one common baseline is BM25

What is a baseline?

- A baseline needs to be some reasonable “first approach”
  - untuned BM25 for search
  - item-item model for recommender systems
  - …
- Low effort to get it running but needed to put performance in context

What is a naive system?

- Needed to set the lowest possible bar for performance
  - Search: retrieve results sorted by length/popularity/likes/etc.
    - Could also return documents randomly (using metadata is stronger though!)
  - Recommend systems: the mean rating
  - Classification: the most common class
  - …
- Your system should probably beat this “system” at least

You can use existing IR libraries!

- PyTerrier and Pyserini are popular choices
  - Efficient and fast implementation of all of the things we've learned in class
- Can use vector-based search engines too
  - Or deep learning models directly
- You do not need to use the homework code for anything

Why do annotation?

- Practical answer: most projects do not yet have any labeled to evaluate with
- Pedagogical answer: annotation exposes you to the specific IR task your system needs to solve
  - *Very* helpful in building intuition for which features to use

How do we do annotation?

- Not intended to be a huge pain!
- Approach it like most TREC challenges:
  - Collect “top” results from a few IR systems
  - Rate just those results for relevance
    - Assume everything else is irrelevant (lowest score)
- We recommend using some kind of scale for relevance: perhaps a 5 point scale (may differ based on your needs)

Search annotation guide

- Use an IR system with a few models to collect to most-relevant ~50 results for a query
  - BM25 should be one
- Merge all these results into a single list (keep unique documents)
  - Aim for 100-200 documents
- If you know something about your data, manually search for a few documents and include those too
- Put documents/data in a google sheet and add a column for relevance
  - make the tab name the query so you don't forget
  - randomize the order (remove bias!)
- Quickly work through the documents and score them
  - Take a break after each one if helpful!

Other tips/tricks

- More diverse models will give you a broader selection of documents in total
- Don’t feel the need to finish everything at once
  - Don’t wait until the last minute too!
- Feel free to use an IR library to get you started
- Discuss annotation scales with partner(s), if applicable

Any questions on the project update?