Cluster of Excellence
The Politics of Inequality

Steinbeis Transfer Center
Linguistic Data Analysis

UNIVERSITÄT
PASSAU

Information Retrieval & Natural Language Processing
Week 10: Personalized search

Annette Hautli-Janisz, Prof. Dr.

16 January 2025

Today

Part 1:

Assessing relevance

Part 2:

Personalized search, or: types of features used beyond core ranking

Assessing relevance

Given information needs and documents, you need to collect relevance assessments by humans.

Standard approach: Pooling, i.e., relevance is assessed over a subset of the collection that is formed from the top k documents returned by a number of different IR systems.

Humans and their relevance judgements are quite idiosyncratic and variable → need to measure how much agreement between judges there is.

Kappa statistic

Kappa = $\frac{P(A)-P(E)}{1-P(E)}$

P(A) is the observed agreement.
P(E) is the expected agreement.

Kappa = 1 if two judges always agree.
Kappa = 0 if two judges agree at the rate given by chance.
Kappa < 0 if two judges agree worse than at random.

In a two-class decision, P(E) = 0.5. But normally, class distribution is skewed, therefore we use *marginal* statistics to calculate P(E).

Kappa statistic

Go through the example in the IRR book, p. 151.

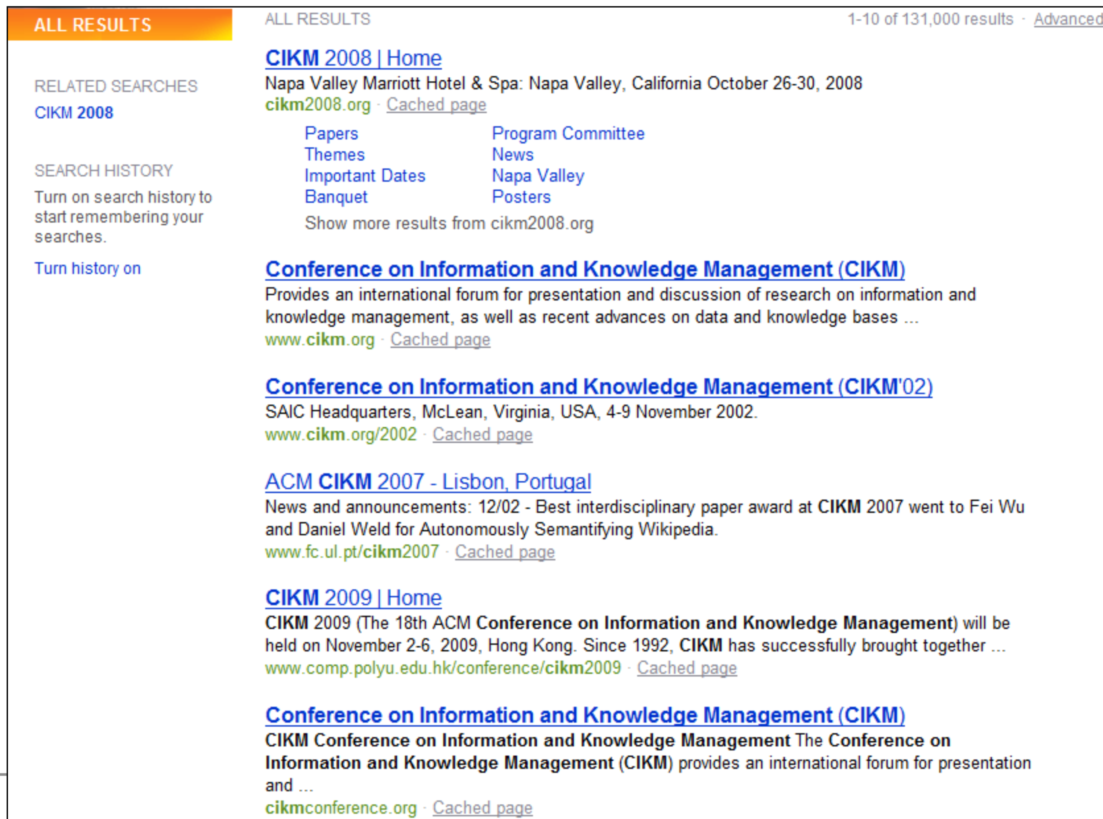
		Judge 2 Relevance			
		Yes		No	Total
Judge 1	Yes	300		20	320
Relevance	No	10		70	80
	Total	310		90	400

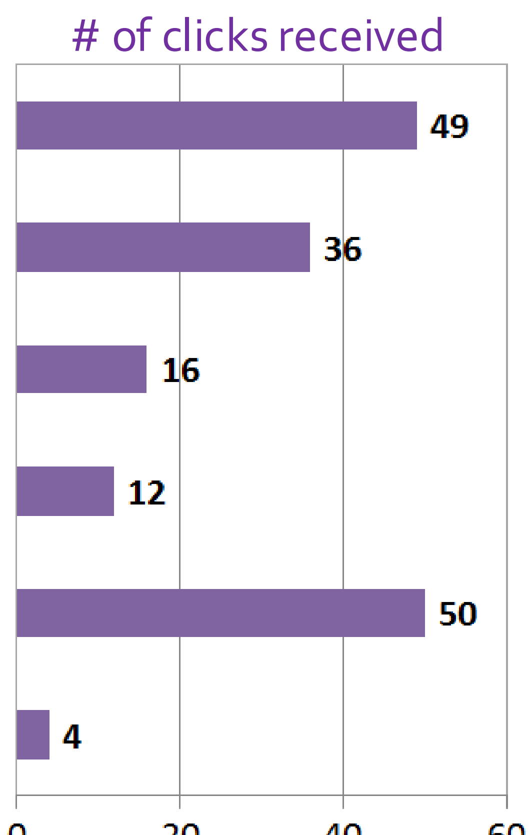
Calculate P(A), P(relevant), P(non-relevant), P(E) and kappa.

User behavior

Search results for *CIKM* (in 2009)

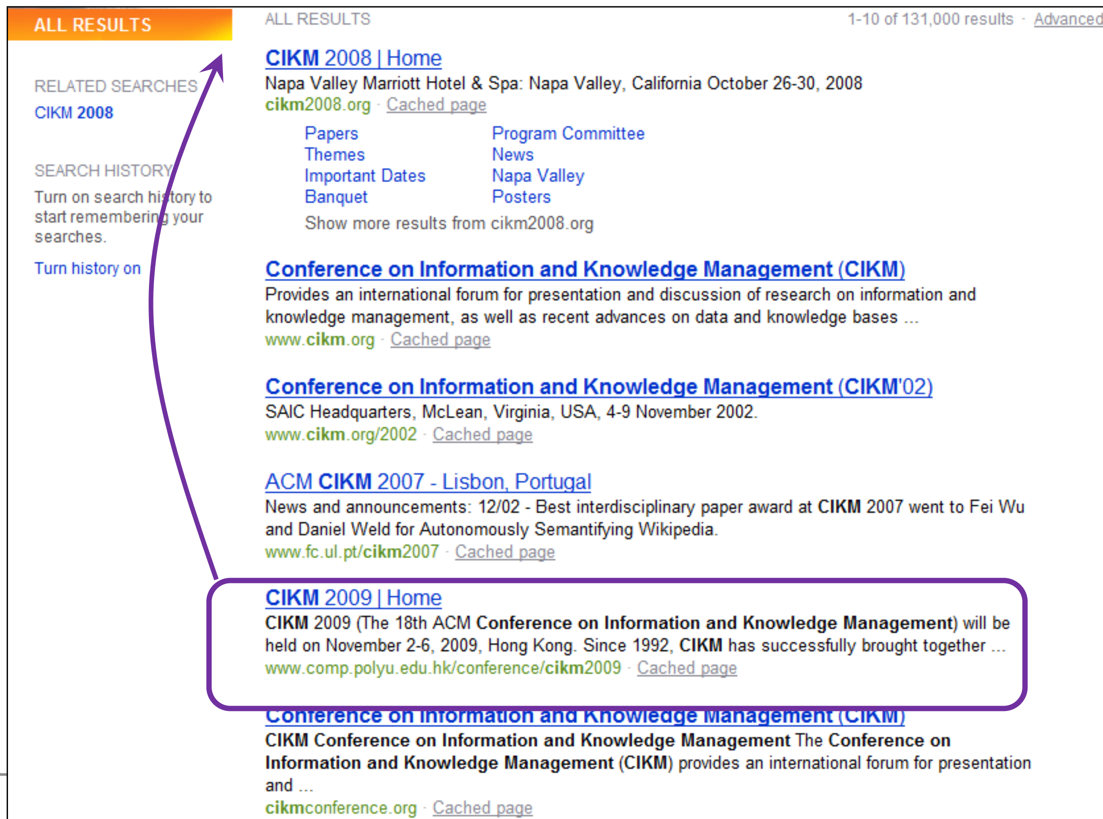
See Fan Guo and Chao Liu's 2009/2010 CIKM tutorial "Statistical models for web search: Click log analysis".

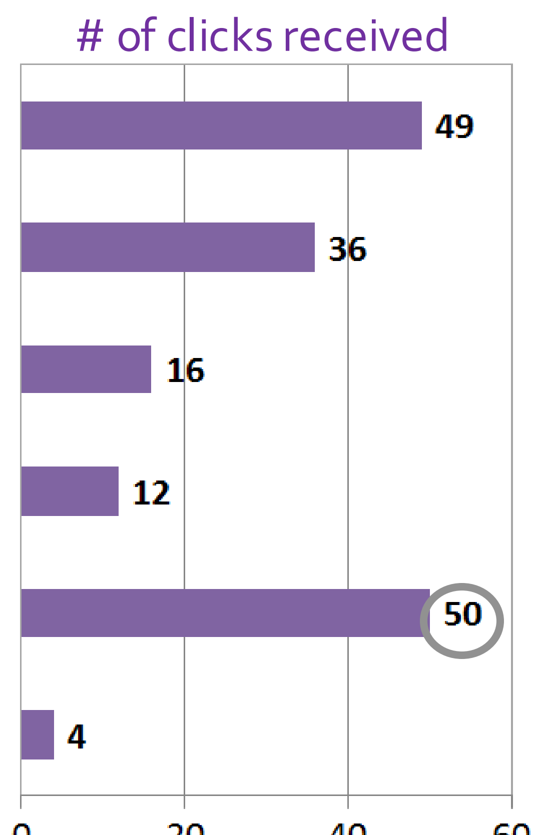




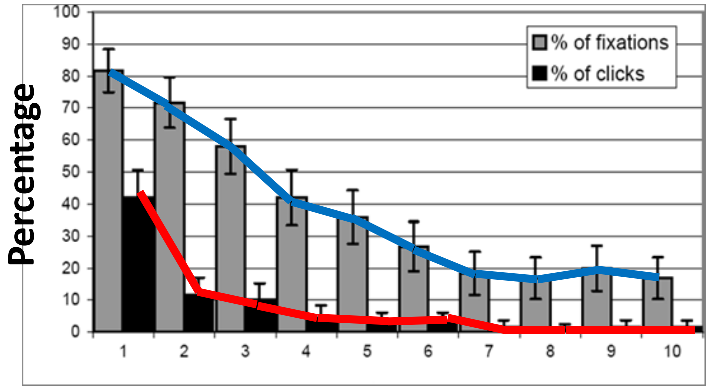
User behavior

Adapt ranking to user clicks? But there is a strong position bias, so absolute click rates are unreliable.

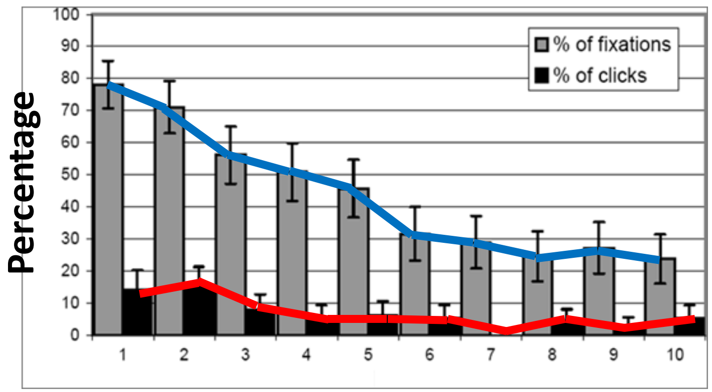




Click Position bias



Normal Position



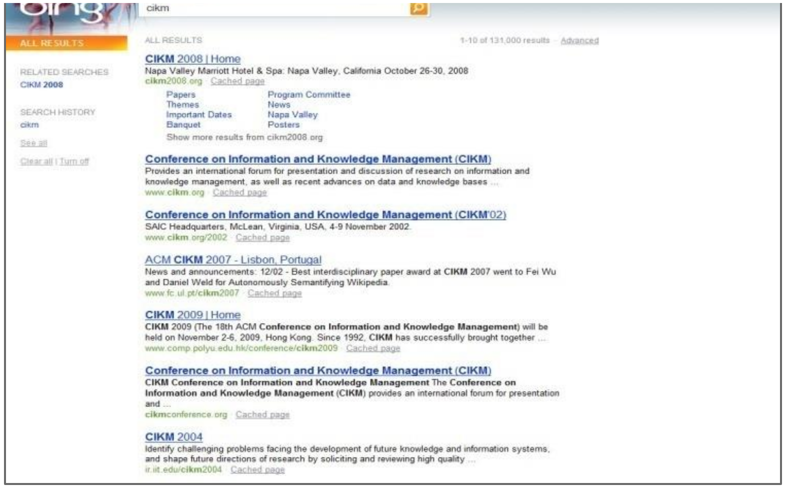
Reversed Impression

Higher positions receive more user attention (eye fixation) and clicks than lower positions.

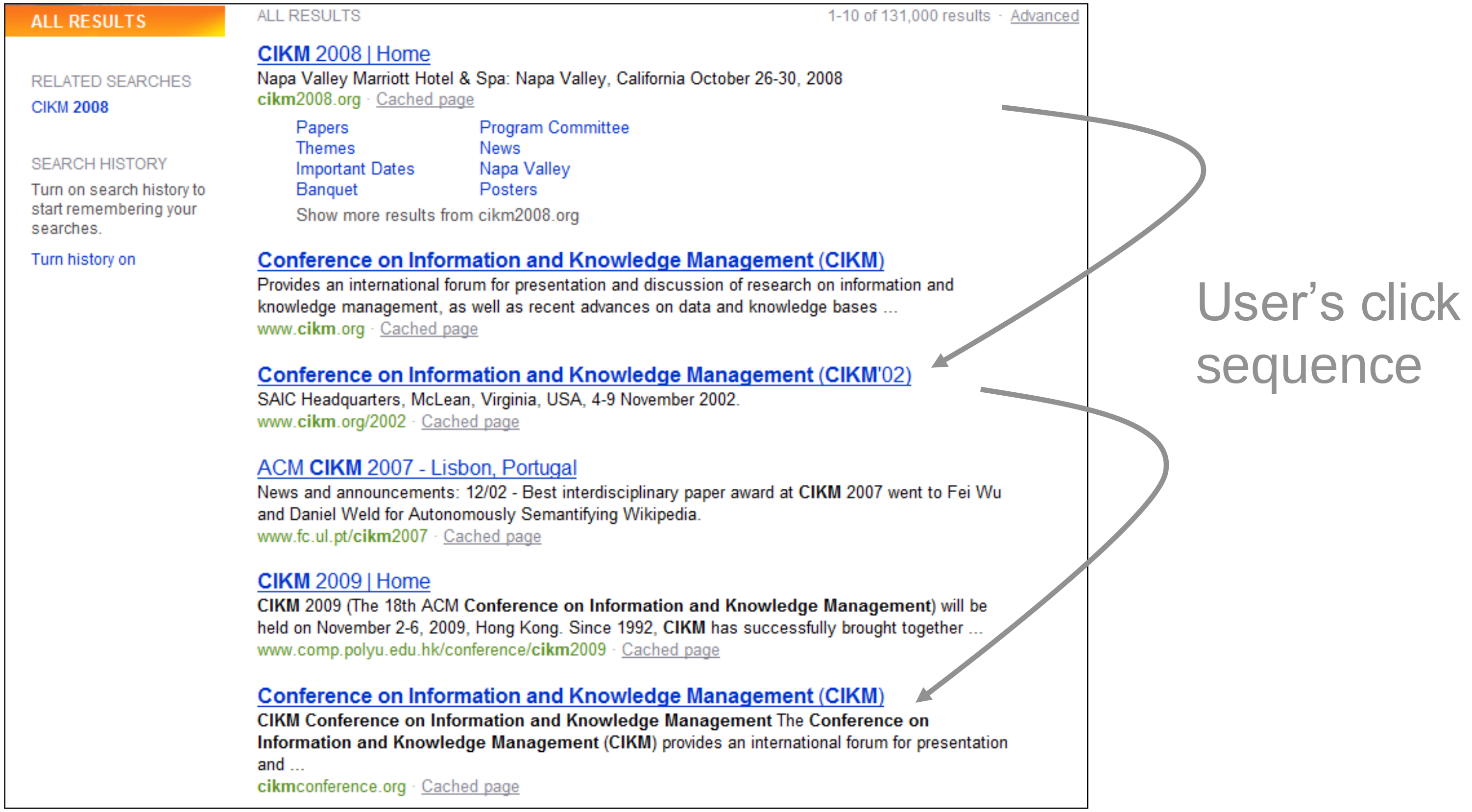
This is true even in the extreme setting where the order of positions is reversed.

“Clicks are informative but biased”. (Joachims 2007)

Eye-tracking user study



Relative versus absolute ratings



Hard to conclude: Result1 > Result3
Probably can conclude Result3 > Result2

A/B test

Common practice to test modern search engine systems.

Two-sample hypothesis testing:

- Two versions (A and B) of a system are compared, which are identical except for one variation that might affect a user's behavior, e.g., BM25 with different parameter settings

- Randomized experiment
  - Separate the population into equal size groups – 10% random users for system A and 10% random users for system B
  - Null hypothesis: no difference between system A and B

A/B test

Behavior-based metrics:

- *Abandonment Rate*: fraction of queries for which no results are clicked on
- *Reformulation Rate*: fraction of queries that are followed by another query during the same session
- *Queries per Session*: mean number of queries issued by a user during a session
- *Clicks per Query*: mean number of results that are clicked for each query
- *Time to First/Last Click*: mean time from query being issued until last click on any result

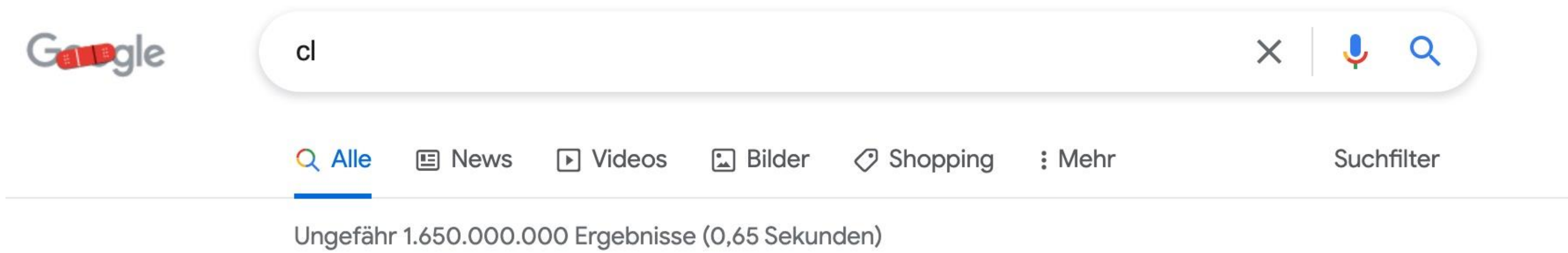
A/B test

How do the metrics change as the ranking gets worse?

- *Abandonment Rate*
- *Reformulation Rate*
- *Queries per Session*
- *Clicks per Query*
- *Time to First/Last Click*

Search in context

Queries are difficult to interpret in isolation.

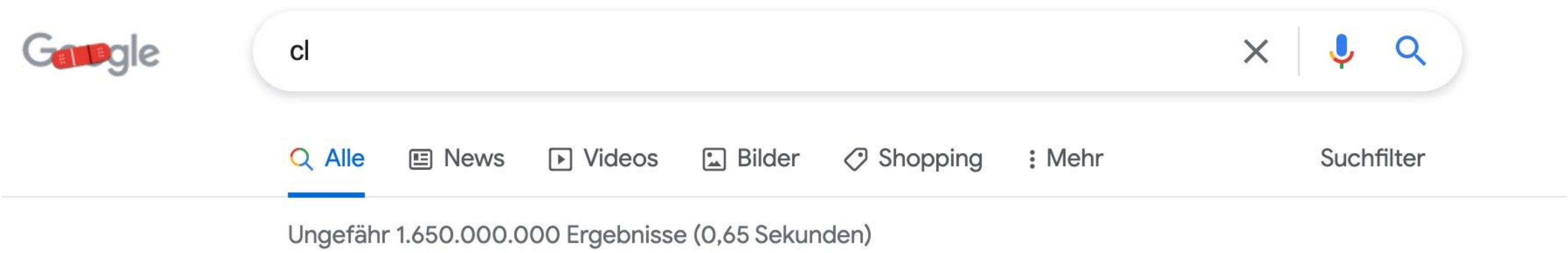


Easier if we model: **who** is asking, **what** have they done in the past, **where** are they, **what time** is it, etc.

Association for
Computational Linguistics

Search in context

Queries are difficult to interpret in insolation.

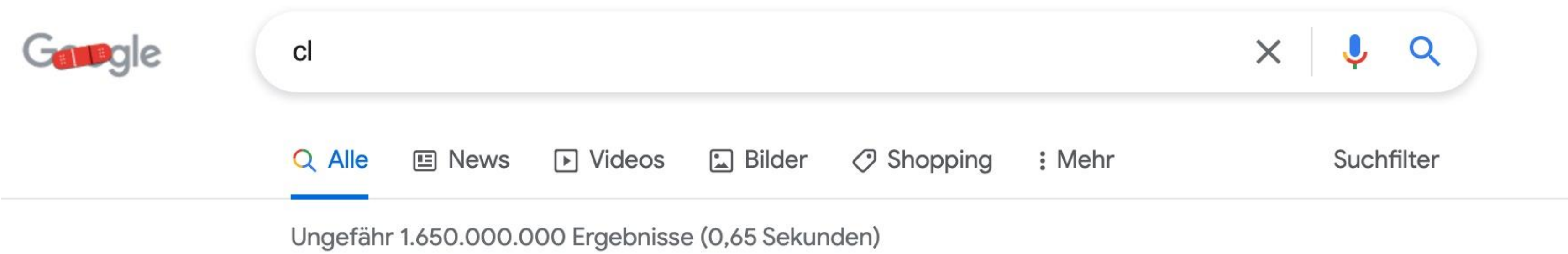


Easier if we model: **who** is asking, **what** have they done in the past, **where** are they, **when** is it, etc.

Searcher:
(CL | world’s best soccer player 2022) versus (CL | computational linguist)

Search in context

Queries are difficult to interpret in insolation.

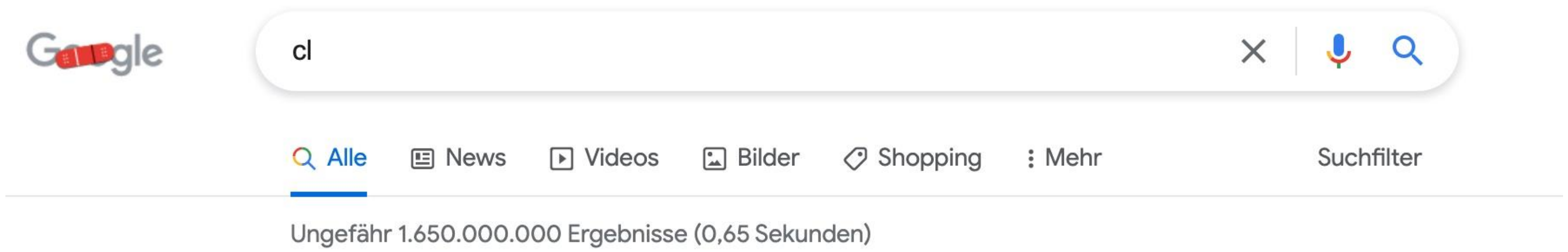


Easier if we model: **who** is asking, **what** have they done in the past, **where** are they, **when** is it, etc.

Previous actions:
(CL | Champions League) versus (CL | computational linguists)

Search in context

Queries are difficult to interpret in insolation.

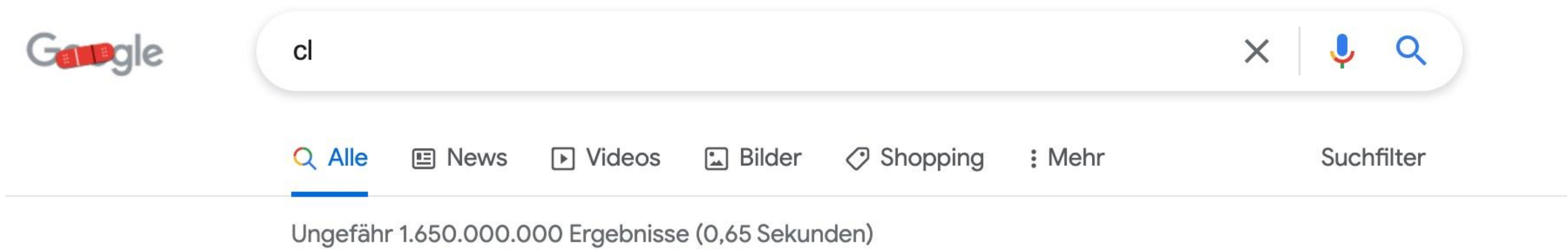


Easier if we model: **who** is asking, **what** have they done in the past, **where** are they, **when** is it, etc.

Location:
(CL | at Champions League final) versus. (CL | at ACL conference)

Search in context

Queries are difficult to interpret in insolation.



Easier if we model: **who** is asking, **what** have they done in the past, **where** are they, **when** is it, etc.

Time:
(CL | December submission) versus. (CL | August conference)

Personalization

Using a single ranking for everyone, in every context, at every point in time, limits how well a search engine can do.

→ Enhance the performance of the search engine by using

- core ranking
- personalization

Potential for personalization

Teevan, Dumais, Horvitz 2010:

Aim: Quantify the variation in relevance for the same query across different individuals.

Explicit judgements from different people:

- 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

Recap: Discounted cumulative gain

Popular measure for evaluating web search and related tasks.

Two assumptions:

1. Highly relevant documents are more useful than marginally relevant documents.
2. The lower the ranked position of a relevant document, the less useful it is for the user, since it’s less likely to be examined.

Focus on retrieving highly relevant documents.

Recap: Discounted cumulative gain

Designed for non-binary notions of relevance.

Uses graded relevance as a measure of usefulness, or gain, from examining a document.

Gain is accumulated starting at the top of the ranking and may be reduced, or discounted, at lower ranks.

Recap: Discounted cumulative gain

Summarize a ranking:

- Imagine the relevance judgements are on a scale of [0, r], with r > 2.
- Cumulative gain (CG) at rank *n*
  - The ratings of the n documents are $r_1, r_2, r_3, \ldots, r_n$
  - $CG = r_1 + r_2 + r_3 + \ldots + r_n$
- Discounted Cumulative Gain (DCG) at rank *n*
  - $DCG = r_1 + r_2 /\log_2 2 + r_3/\log_2 3 + \ldots + r_n/\log_2 i$

or $DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i}$

Recap: Discounted cumulative gain

Example:
There are 10 ranked documents judged on a 0-3 relevance scale:

3, 2, 3, 0, 0, 1, 2, 2, 3, 0

Compute the discounted gain and the DCG for all ranks (logarithm with base 2).

Normalized discounted cumulative gain

Normalize with DCG_{ideal}, the ideal ranking of the results.

- sort the results in decreasing order of relevance
- calculate DCG for that ranking

- $NDCG = DCG_n / DCG_{ideal}$

Original ranking: 3, 2, 3, 0, 0, 1, 2, 2, 3, 0

Ideal ranking: 3, 3, 3, 2, 2, 2, 1, 0, 0, 0

→ DCG_{ideal} = , NDCG =

Potential for personalization

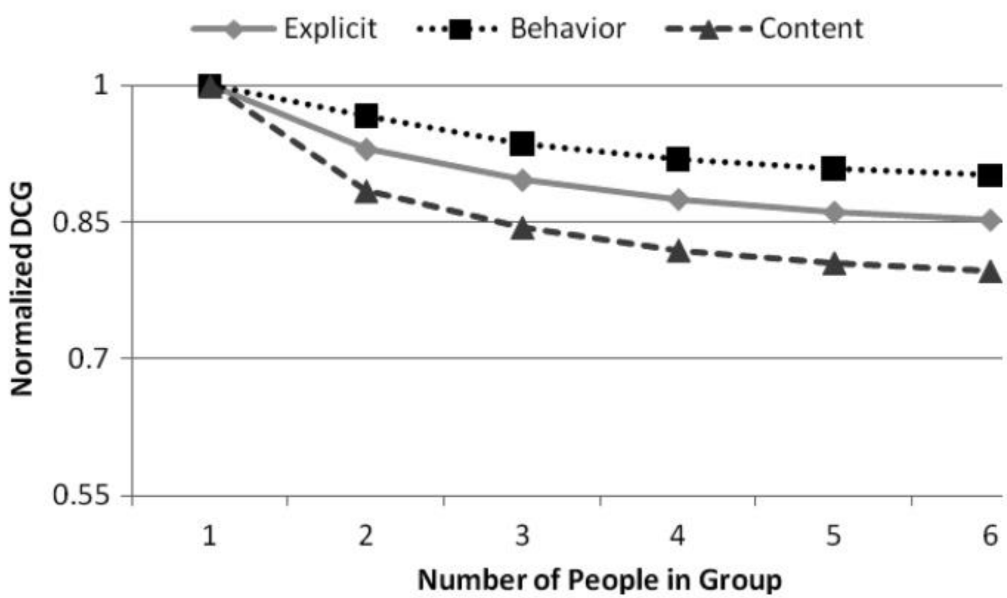


Fig. 5. The potential for personalization curves according to the three different measures of relevance. Explicit relevance judgments for the 17 unique queries that at least six people evaluated are compared with 24 queries for which there are at least six content-based implicit judgments and the 44,002 behavior-based queries for which there are behavior-based implicit judgments.

Teevan, Dumais, Horvitz 2010

Some literature on personalization

Liu et al. 2019. Personalization in text information retrieval: A survey. Journal of the Association for Information Science and Technology.

"*Personalization is aimed at tailoring search toward individual users and user groups by taking into account additional information about users besides their queries.*"

Started about 10-15 years ago, rich effort in industry and academia.

User models

Part A: Constructing user models

- sources of evidence:
  - content: queries, web pages, explicit profile, etc.
  - behavior: explicit feedback, implicit feedback, visited web pages etc.
  - context: location, date, time (of day/week/month), device etc.
- time frame: short-term, long-term
- who: individual, group

Part B: Using user models

- reside where: client, server
- how used: reranking, query expansion/suggestion
- when used: always, sometimes, context learned

User models

Part A: Constructing user models

- sources of evidence:
  - content: queries, web pages, explicit profile, etc.
  - behavior: explicit feedback, implicit feedback, visited web pages etc.
  - context: location, date, time (of day/week/month), device etc.
- time frame: short-term, long-term
- who: individual, group

Part B: Using user models

- reside where: client, server
- how used: **reranking, query expansion/suggestion**
- when used: always, sometimes, context learned

Personalizing search

Pitkow et al. 2002: Two general ways of personalizing search

Query expansion:

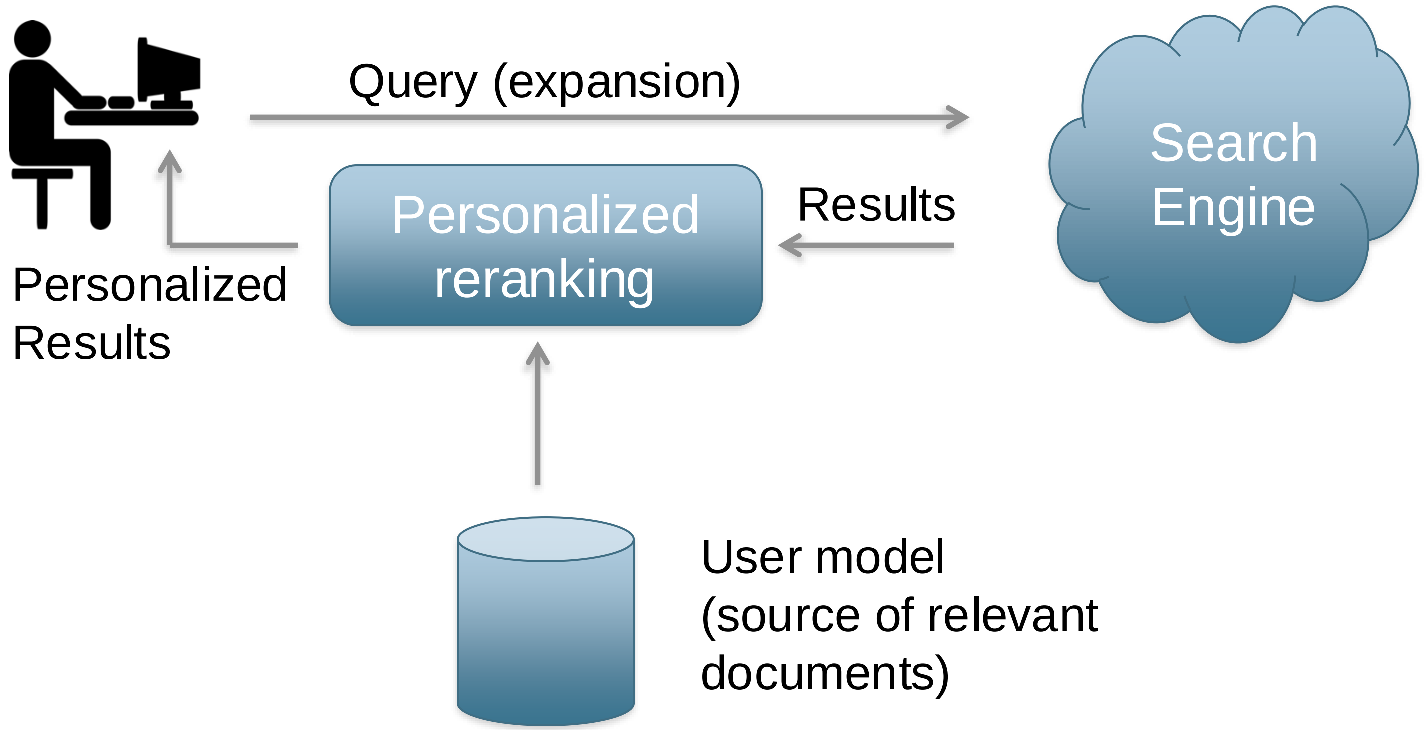
- modify or augment user query
- e.g., query term “IR” can be augmented with either “information retrieval” or “Ingersoll-Rand” depending on user interest
- ensures that there are enough personalized results

Reranking:

- issue the same query and fetch the same results …
- … but rerank the results based on a user profile
- allows both personalized and globally relevant results

Personalizing search

Teevan, Dumais and Horvitz 2005:



Personalization via location

User location is one of the most important features for personalization.

- country:
  - queries like ‘football’ and ‘biscuit’ in the UK versus the US
- state/metro/city:
  - queries like ‘zoo’, ‘craigslist’, ‘Ebay Kleinanzeigen’
- fine-grained location:
  - queries like ‘pizza’, ‘restaurant’, ‘coffee shop’

Personalization via location

Not all queries are location sensitive:

- ‘facebook’ is not asking for the closest Facebook office
- ‘national park’ is not necessarily asking for the closest national park

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

- University of Passau home page has address, but is not location sensitive.

Personalization via location

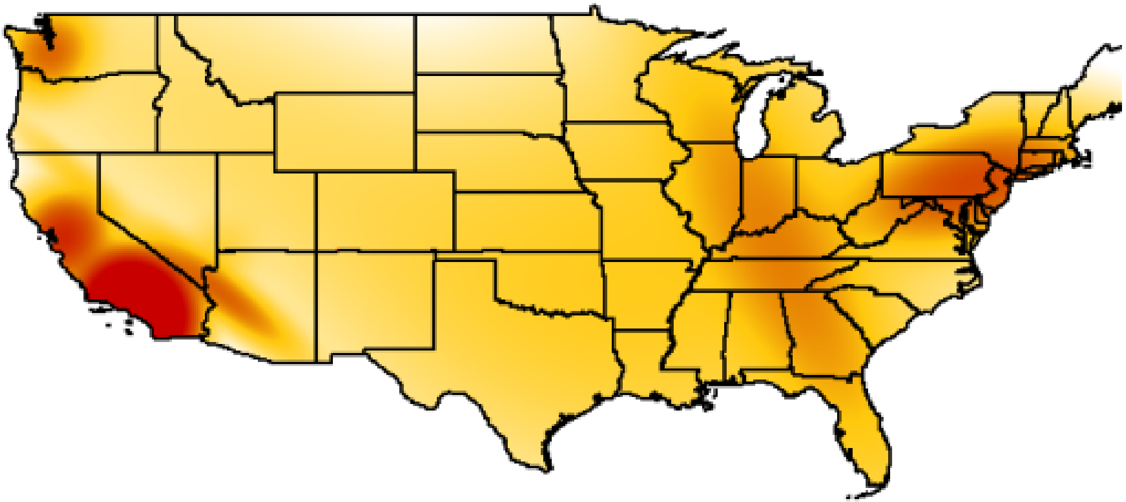
Key idea in Bennett et al. 2011.

Usage statistics, rather than locations mentioned in a document, best represent where it is relevant.
→ 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

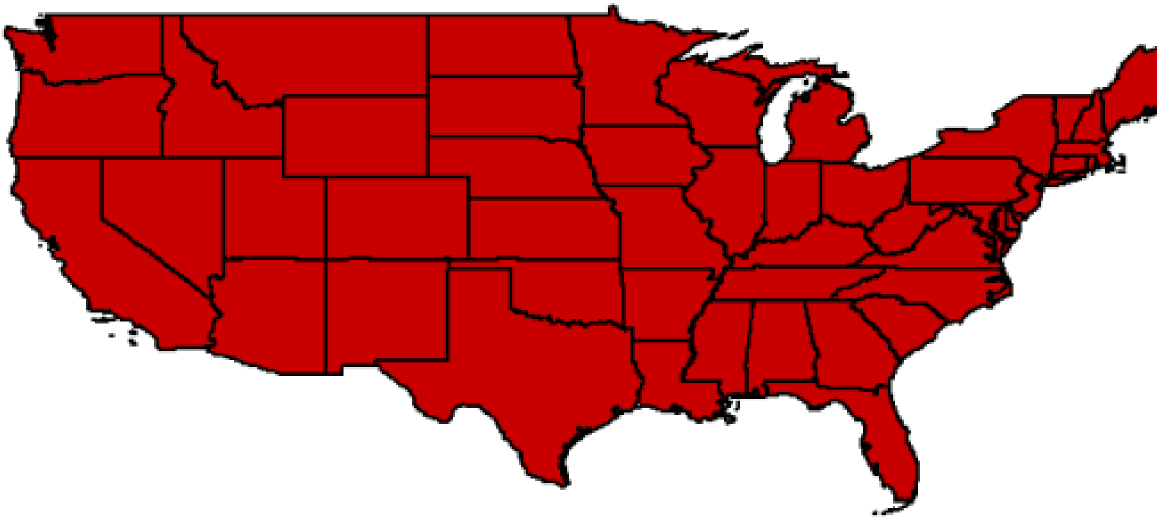
Use the logs data to estimate the probability of the location of the user given they viewed this URL: P(location = x | URL)
→ model of the locations in which a website is likely of interest.



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

Location interest model

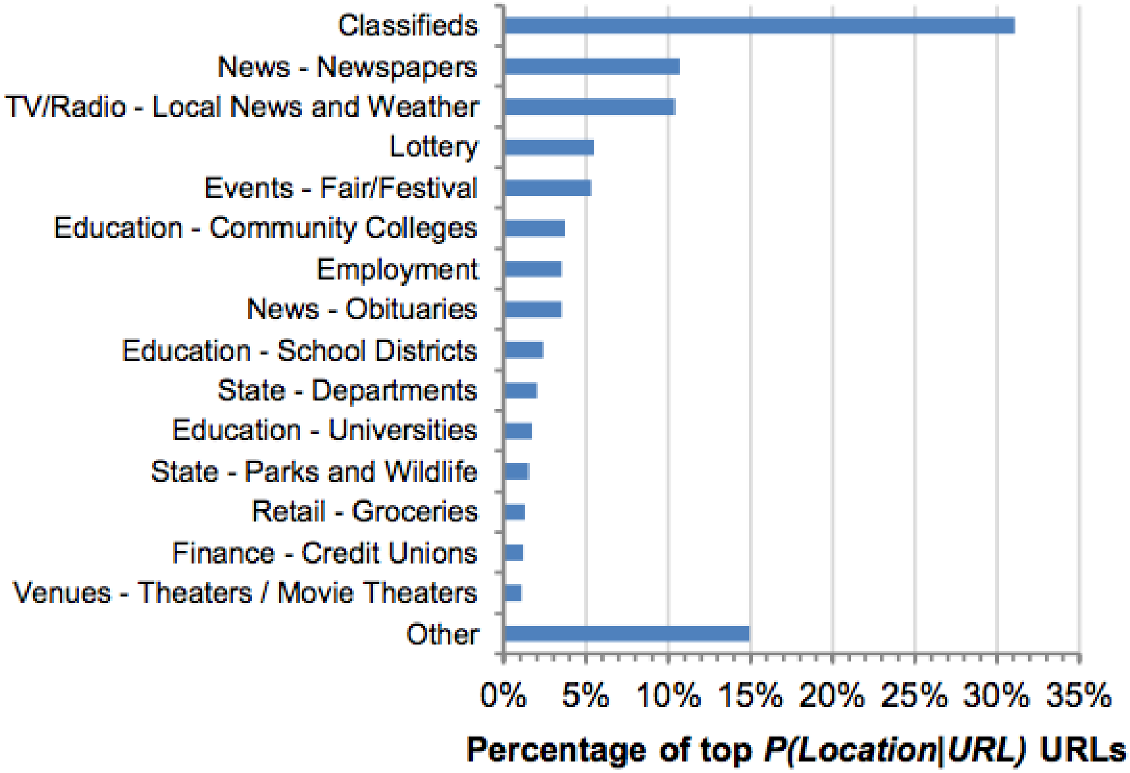
Use the logs data to estimate the probability of the location of the user given they viewed this URL: P(location = x | URL)
→ model of the locations in which a website is likely of interest.



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

Location interest model

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



Issues with personalization

Resistance to over-personalization. Creepy!

https://constructor.io/blog/when-personalization-goes-wrong-and-how-to-fix-it/

Concerns about personal data tracking.

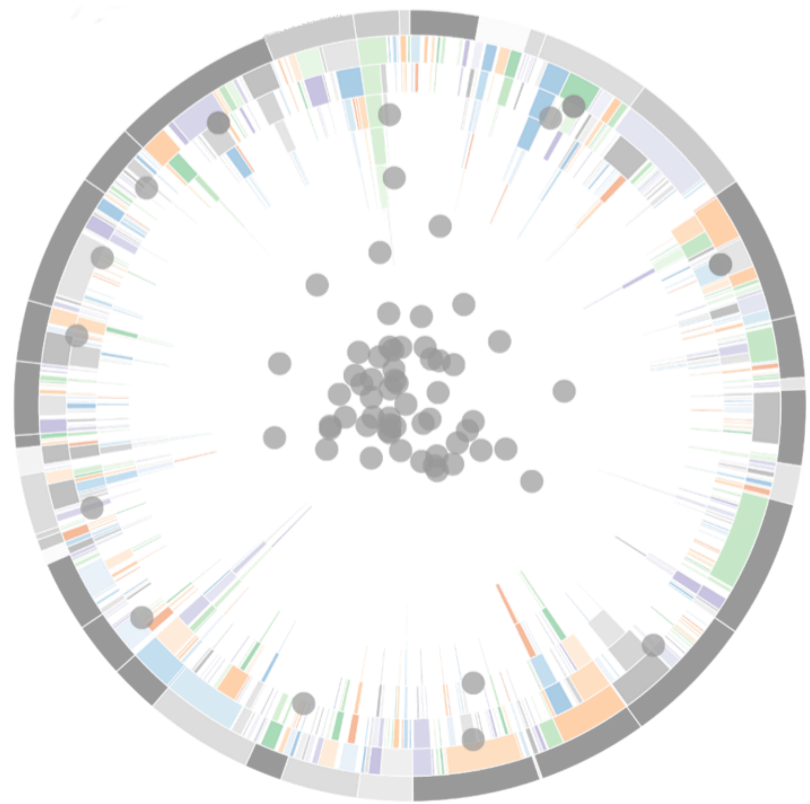
- intensive tracking of browser habits
- tracking of personal information
- storing that information

Cluster of Excellence
The Politics of Inequality

Steinbeis Transfer Center
Linguistic Data Analysis

UNIVERSITÄT
PASSAU

Thank you.
Questions?
Comments?



Annette Hautli-Janisz, Prof. Dr.
cornlp-teaching@uni-passau.de