

IN4325



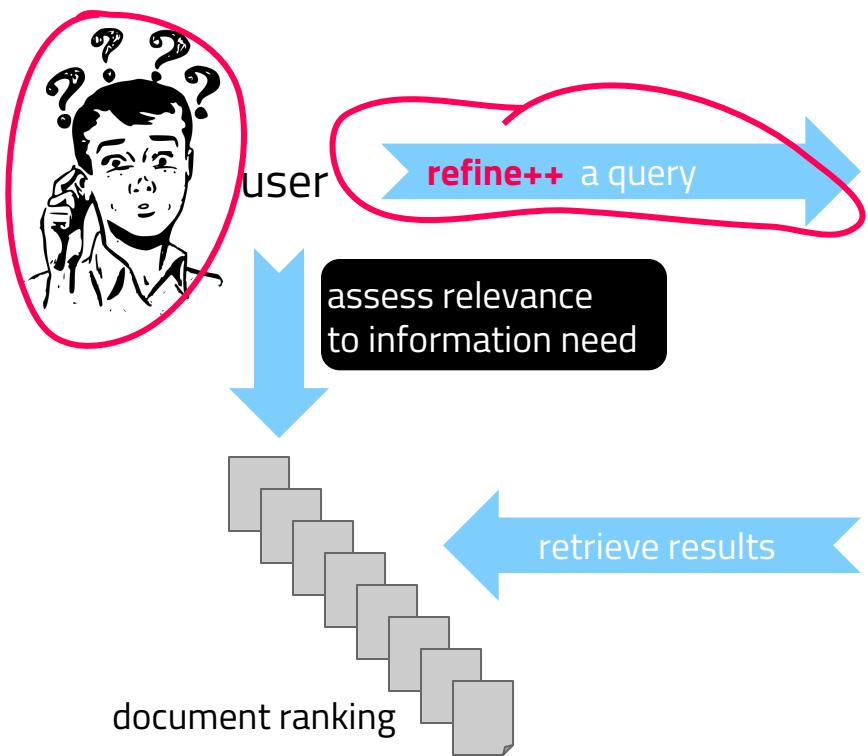
# Query autocomplete and Interactive IR

Claudia Hauff (WIS, TU Delft)

# The big picture

# The essence of IR

**Information need:** Looks like I need Eclipse for this job. Where can I download the latest beta version for macOS Sierra?



**Information need**  
Topic the user wants to know more about

**Query**  
Translation of need into an input for the search engine

**Relevance**  
A document is relevant if it (partially) provides answers to the information need

incomplete,  
underspecified  
& ambiguous

eclipse download osx

WWW, library records, medial reports, patents, ...

## Query refinement

Query expansion  
Pseudo-relevance feedback in LMs  
Spell checking

## Query autocomplete

## Interactive query expansion

Select the term(s) to augment your original query with.

## Query suggestions

Select the complete query to replace your original query with.

## Query completion

Select the complete query to replace your original query with whilst typing.

## Related queries

Select the complete query to replace your original query with.

# Overview

inf	information	information r	information r logged in
informatique	information	information ratio	information ratio
infomedics	information security officer	information retrieval	information retrieval
influenza	information technology	information radiators	information revolution
infinity	information bias	information risk theory	information risk
infographic	information ratio	information rights management	information rules
inflatie	information planet	information request	information radiators
inflatie 2017	information asset	information resources	information rights management
infinity war	information overload	information risk theory audience	information retrieval python
infacol	informationele positionering	information risk	information retrieval pdf
informatica actie	information icon	information retrieval vu	information retrieval techniques

Google Search

Suggestion of queries that (1) match the user's information needs and (2) yield a high-quality result ranking.

## Goals:

1. Reduce query entry time
2. Prepare results in advance of query submission
3. Help users formulate a more precise query

Requires the search system to infer the user's *intent*.

# CHIIR 2018: query priming study



diabetes cinnamon



diabetes cinnamon pills  
diabetes cinnamon rolls  
diabetes cinnamon and honey  
diabetes cinnamon dosage  
diabetes cinnamon comparison  
diabetes cinnamon survey  
diabetes cinnamon statistics  
diabetes cinnamon evidence

(1) QAC with query priming

diabetes cinnamon



cinnamon pills  
cinnamon rolls  
cinnamon and honey  
diabetes cinnamon dosage  
diabetes cinnamon tea  
diabetes cinnamon chromium picolinate  
diabetes cinnamon update  
diabetes cinnamon study

(2) Conventional QAC

## Findings:

1. With priming, users issue more queries
2. With priming, users (re)-visit the SERP more often
3. The priming effect varies relative to users' educational backgrounds (benefits highly educated users)

# Query-log based Query autocomplete

# Task

Given the current prefix  
(=query string the user  
has typed in so far),  
rank all possible  
candidates\* (=complete  
queries).

Display the top ranked  
candidates to the user.

\*assume for now that we have that list  
available



# Two strong baselines

Assumptions:

1. Access to a query log and document clicks
2. Access to a corpus
3. Access to a user's past queries

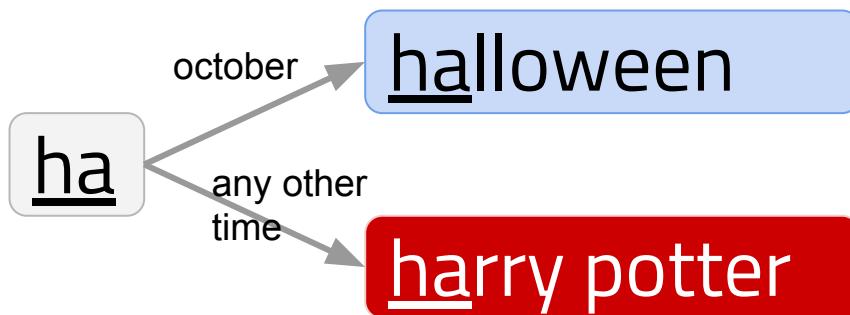
## Most popular ranker

Query candidates are ranked according to their past popularity

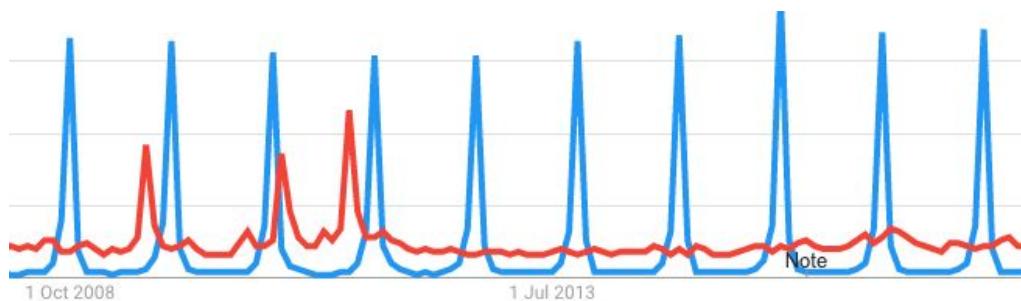
## Clicked documents ranker

Cosine similarity between a user's profile (previously clicked docs by that user) and the candidate query profile (previously clicked docs across all users for that query)

# Time-sensitive query autocomplete

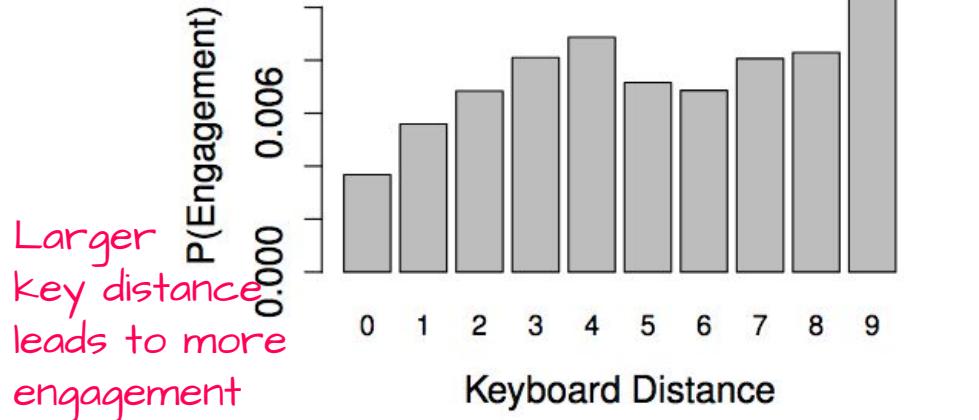
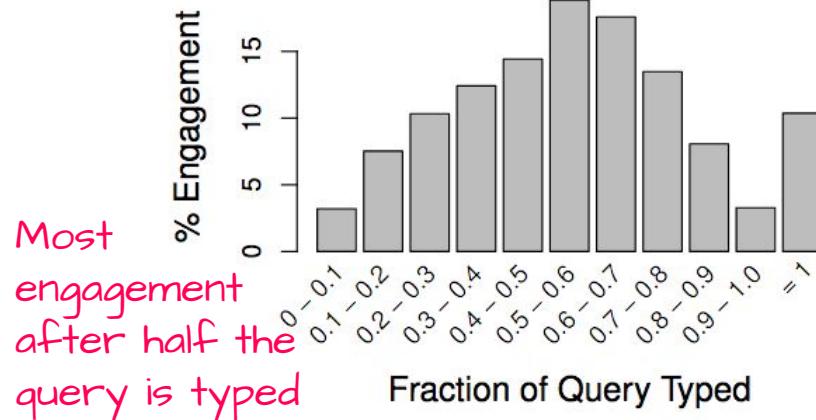
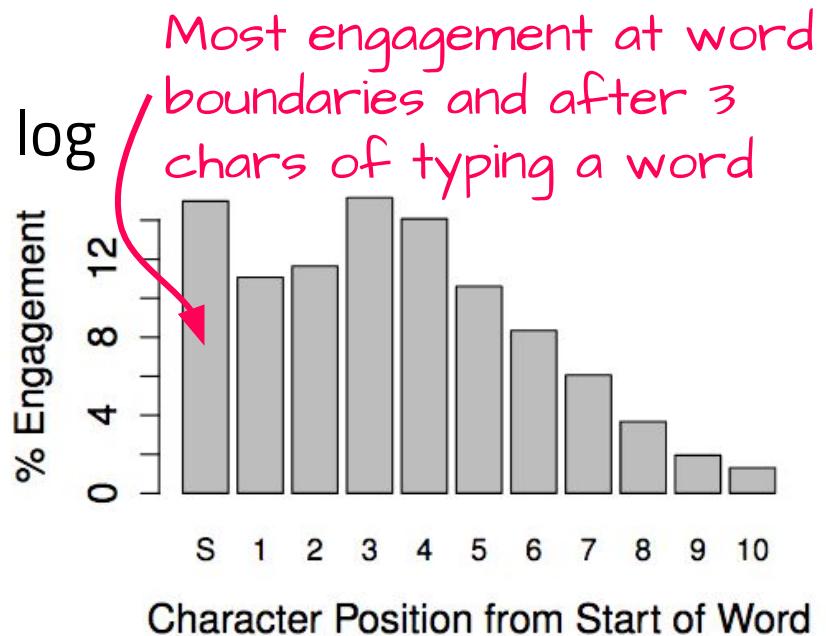
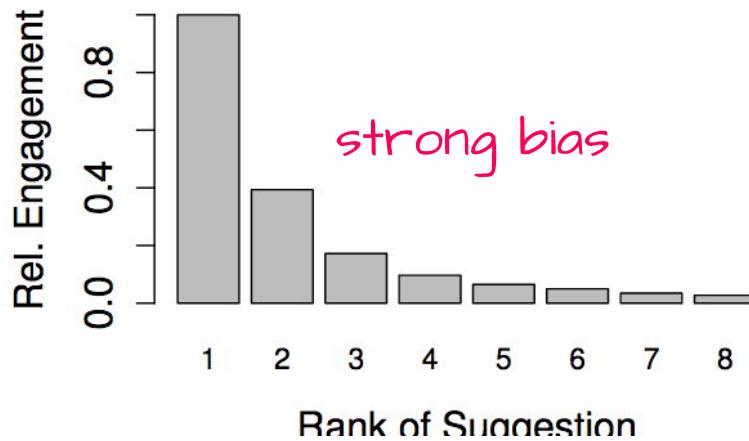


Approach: apply time-series modeling  
and rank candidates according to their  
forecasted frequencies



# User engagement with query completion

1.6M queries from Bing's search log



**Cross-lingual IR:** field of IR concerned with the retrieval of documents in a language different from the query language

**Cross-lingual query suggestions:** suggest queries in a different language from the original query



Web search engines are not everything ...

Large user base

Assumptions:

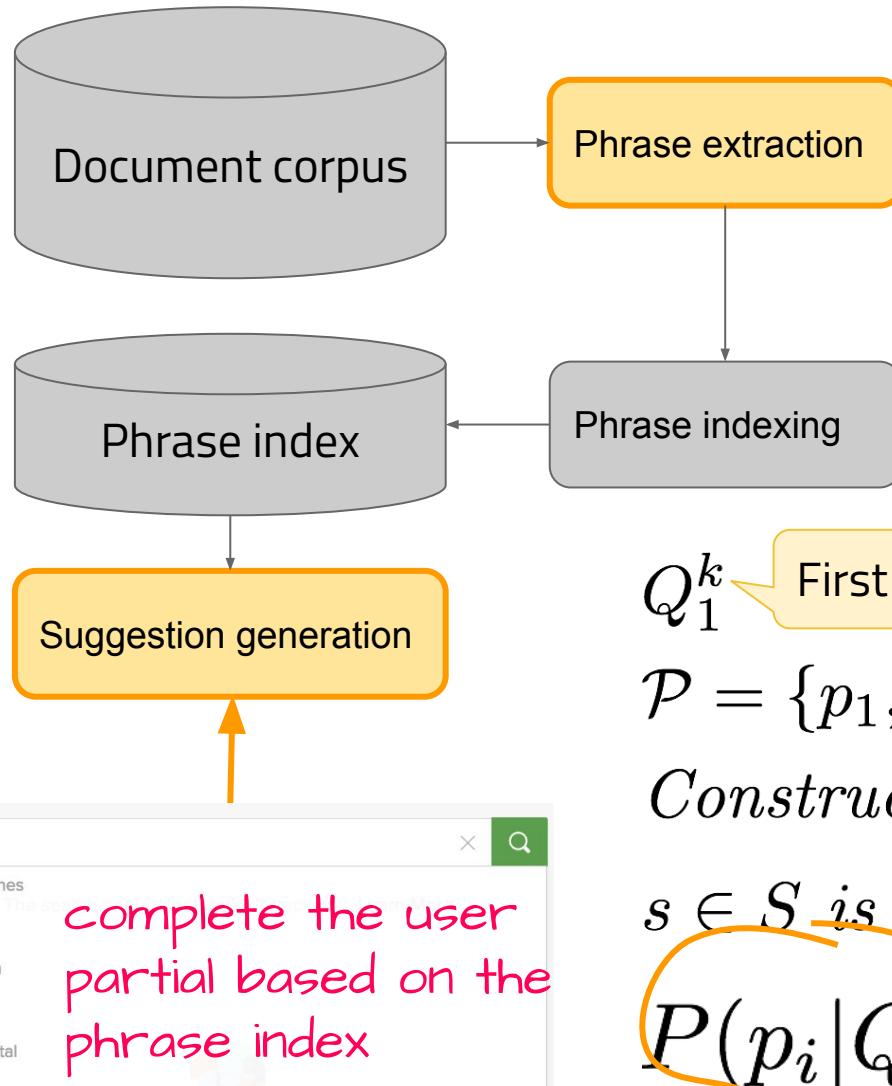
1. **Access to a query log and document clicks**
2. Access to a corpus *always possible*
3. **Access to a user's past queries**

What about search in specialized domains or personal search systems (PIM)?



# Corpus-based Query autocomplete

# Corpus-based query suggestions



- N-grams: unigrams, bigrams, trigrams
- Ignore N-grams starting with a stopword

$Q_1^k$  First  $k$  characters typed

$\mathcal{P} = \{p_1, \dots, p_n\}$  Set of all extracted phrases

Construct  $S \subset \mathcal{P}$ , such that each

$s \in S$  is a possible completion of  $Q_1^k$

$P(p_i | Q_1^k)$



# Corpus-based query suggestions

$P(p_i | Q_1^k)$  Probability that the user will type  $p_i$  given her first  $k$  typed characters

# Corpus-based query suggestions

$P(p_i | Q_1^k)$  Probability that the user will type  $p_i$  given her first  $k$  typed characters

$Q_1^k = \underline{Q_c} + \underline{Q_t}$  Completed word(s) plus word the user is currently typing

# Corpus-based query suggestions

$P(p_i | Q_1^k)$  Probability that the user will type  $p_i$  given her first  $k$  typed characters

$Q_1^k = \underline{Q_c} + \underline{Q_t}$  Completed word(s) plus word the user is currently typing

$P(p_i | Q_1^k) = \frac{P(p_i) \times P(Q_1^k | p_i)}{P(Q_1^k)}$  according to Bayes' theorem

# Corpus-based query suggestions

$P(p_i | Q_1^k)$  Probability that the user will type  $p_i$  given her first  $k$  typed characters

$Q_1^k = \underline{Q_c} + \underline{Q_t}$  Completed word(s) plus word the user is currently typing

$$\begin{aligned} P(p_i | Q_1^k) &= \frac{P(p_i) \times P(Q_1^k | p_i)}{P(Q_1^k)} \quad \text{according to Bayes' theorem} \\ &= \frac{P(p_i) \times P(Q_t | p_i) \times P(Q_c | p_i)}{P(Q_1^k)} \end{aligned}$$

Simplifying assumption:  
conditional independence

# Corpus-based query suggestions

$P(p_i | Q_1^k)$  Probability that the user will type  $p_i$  given her first  $k$  typed characters

$Q_1^k = \underline{Q_c} + \underline{Q_t}$  Completed word(s) plus word the user is currently typing

$$\begin{aligned} P(p_i | Q_1^k) &= \frac{P(p_i) \times P(Q_1^k | p_i)}{P(Q_1^k)} \quad \text{according to Bayes' theorem} \\ &= \frac{P(p_i) \times P(Q_t | p_i) \times P(Q_c | p_i)}{P(Q_1^k)} \\ &= \frac{P(Q_t) \times P(p_i | Q_t) \times P(Q_c | p_i)}{P(Q_1^k)} \end{aligned}$$

# Corpus-based query suggestions

$$P(p_i | Q_1^k)$$

Probability that the user will type  $p_i$  given her first  $k$  typed characters

$$Q_1^k = \underline{Q_c} + \underline{Q_t}$$

Completed word(s) plus word the user is currently typing

$$P(p_i | Q_1^k) = \frac{P(p_i) \times P(Q_1^k | p_i)}{P(Q_1^k)} \quad \text{according to Bayes' theorem}$$

$$\begin{aligned} & P(p_i)P(Q_t | p_i) \\ &= P(p_i, Q_t) \\ &= P(Q_t)P(p_i | Q_t) = \frac{P(Q_t) \times P(p_i | Q_t) \times P(Q_c | p_i)}{P(Q_1^k)} \end{aligned}$$

# Corpus-based query suggestions

$P(p_i | Q_1^k)$  Probability that the user will type  $p_i$  given her first  $k$  typed characters

$Q_1^k = \underline{Q_c} + \underline{Q_t}$  Completed word(s) plus word the user is currently typing

$$\begin{aligned} P(p_i | Q_1^k) &= \frac{P(p_i) \times P(Q_1^k | p_i)}{P(Q_1^k)} \quad \text{according to Bayes' theorem} \\ &= \frac{P(p_i) \times P(Q_t | p_i) \times P(Q_c | p_i)}{P(Q_1^k)} \\ &= \frac{\cancel{P(Q_t)} \times P(p_i | Q_t) \times P(Q_c | p_i)}{\cancel{P(Q_1^k)}} \end{aligned}$$

Remains static for all  $p_i$

$$\stackrel{rank}{=} P(p_i | Q_t) \times P(Q_c | p_i)$$

# Corpus-based query suggestions

$$P(p_i|Q_1^k) \stackrel{rank}{=} P(p_i|Q_t) \times P(Q_c|p_i)$$

Phrase that contains  
the completed word  $c_i$

Phrase selection  
probability

Phrase-query correlation  
bill gate\* vs. india gate\*  
**Context is needed!**

$$P(p_{ij}|Q_t) = P(c_i|Q_t) \times P(p_{ij}|c_i)$$

Term completion  
probability;  $c_i$  is a  
possible word  
completion

Term to phrase  
probability

$$P(Q_c|p_i) = \frac{P(Q_c, p_i)}{P(p_i)}$$

Assumption: phrases in the corpus that are  
more important have a higher chance of  
being used by the user for querying.  
Estimated based on corpus statistics.

Estimated based on corpus statistics; to  
avoid data sparseness, we simplify to the  
bag of words approach, i.e. search queries  
linux install firefox  
install firefox linux  
firefox install linux are treated  
in the same way.

# Corpus-based query suggestions

## Data sets

**TREC**: 200K news articles by the Financial Times published between 1991-1994

**Ubuntu**: 100K discussion threads crawled from [ubuntuforums.org](http://ubuntuforums.org)

Given a complete query, retain only the first keyword (Type-A) or the first keyword plus  $k > 2$  characters (Type-B)

## Baseline

**SimSearch**: search the phrase index for all phrases containing the partial user query; rank them in order of decreasing corpus frequency

Radioactive waste  
(TREC Topic 387)

Radioactive  
(Type-A)

Radioactive was  
(Type-B)





# Corpus-based query suggestions

## Data sets

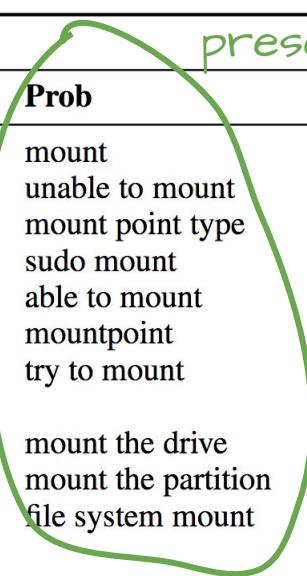
**TREC:** 200K news articles by the Financial Times published between 1991-1994

## Baseline

**SimSearch:** search the phrase index for all phrases containing the partial user query; rank them in order of decreasing corpus frequency

Query = mount			Query = falkland		
SimSearch	CompSearch	Prob	SimSearch	CompSearch	Prob
mount	mount	mount	falklands	falklands	falklands
mounted	mounted	unable to mount	falkland	falkland	falklands war
mounting	mounting	mount point type	falkland islands	falklanders	falkland islands
mounts	mounts	sudo mount	falklands war		falklands conflict
sudo mount	mountpoint	able to mount	falklands conflict		1982 falklands
unable to mount	mountcifs	mountpoint	1982 falklands		1982 falklands conflict
system mount	mountable	try to mount	falkland islands govern-		falkland islands govern-
file system mount	mounter	mount the drive	ment		ment
mount point type	mountunmount	mount the partition	1982 falklands conflict		falklands war in 1982
system mount point type	mountpoints	file system mount	falkland arms		1982 falklands war
			falklanders		invasion of the falklands

presented approach



What are other options besides the generic document corpus frequencies?



# Corpus-based query suggestions

<b>Ubuntu</b>			
	<b>SimSearch</b>	<b>CompSearch</b>	<b>Probabilistic</b>
<b>Type-A</b>	1.00	1.00	1.00
<b>Type-B</b>	0.75	1.00 <sup>s</sup>	1.00 <sup>s</sup>
<b>Overall</b>	0.875	1.00 <sup>s</sup>	1.00 <sup>s</sup>

<b>TREC</b>			
	<b>SimSearch</b>	<b>CompSearch</b>	<b>Probabilistic</b>
<b>Type-A</b>	1.00	1.00	1.00
<b>Type-B</b>	0.15	0.95 <sup>S</sup>	1.00 <sup>S</sup>
<b>Overall</b>	0.575	0.975 <sup>S</sup>	1.00 <sup>S</sup>

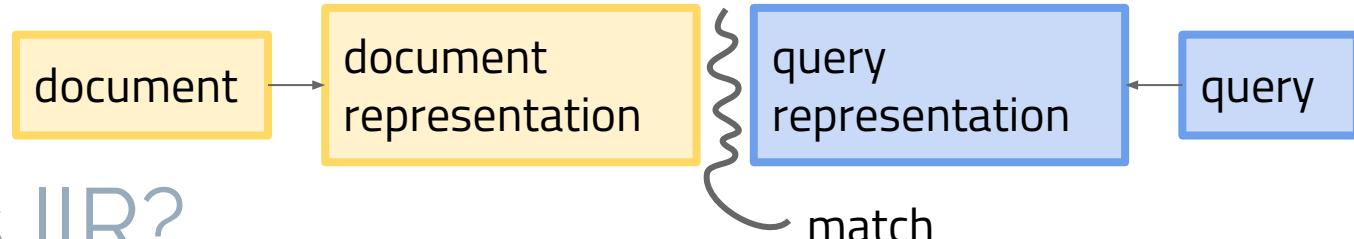
presented approach

Success rate: at least one meaningful suggestion for the partial query

**Table 4:** Success Rate of different query suggestion methods for the two datasets. Superscripts s and S indicate statistically significant improvements over SimSearch with  $p < 0.05$  and  $p < 0.01$ , respectively (one-tailed t-test).

# Interactive Information Retrieval

"classic" IR model



## What is IIR?

*"The area of interactive information retrieval covers research related to **studying** and **assisting** these diverse end users of information access and retrieval systems."*

(Ian Ruthven)

*"In interactive information retrieval, **users are typically studied** along with their interactions with systems and information."*

(Diane Kelly)

*"... the interactive approach to IR has led to a **focus on the user-oriented activities** of query formulation and reformulation, and inspection and judgement of retrieved items ..."* (Nick Belkin)

Many (many!) models have been proposed over the years. This is only a small selection.

# From past to present

Conceptual, observational and empirical work

- Observe users
- Propose a model that *describes* the observations well and has intuitive appeal

Bates' berrypicking

Kuhlthau's ISP

Fuhr's IPRP

approximately equivalent

Search Economic Theory

Mathematical models of information seeking and search

- Narrow down the 'search space' of **testable hypotheses**
- Pick the most promising hypotheses
- Design & execute user studies to (in)validate the hypotheses

Most often in IR  
when we talk  
about models we  
mean retrieval  
models.

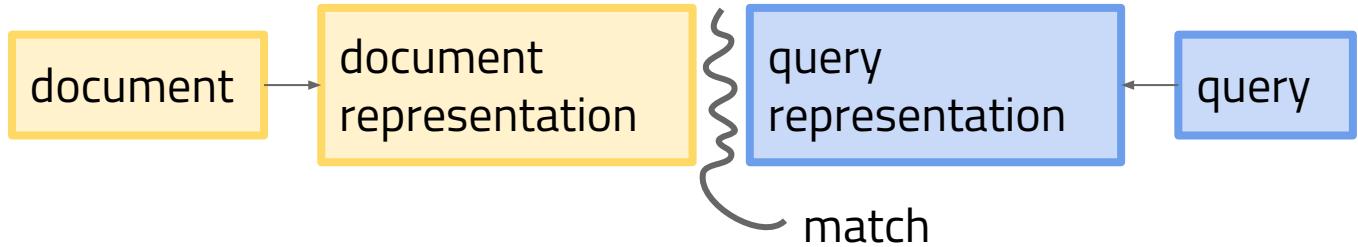
Not now though!

Now: **models**  
for interactive  
information  
seeking and  
retrieval

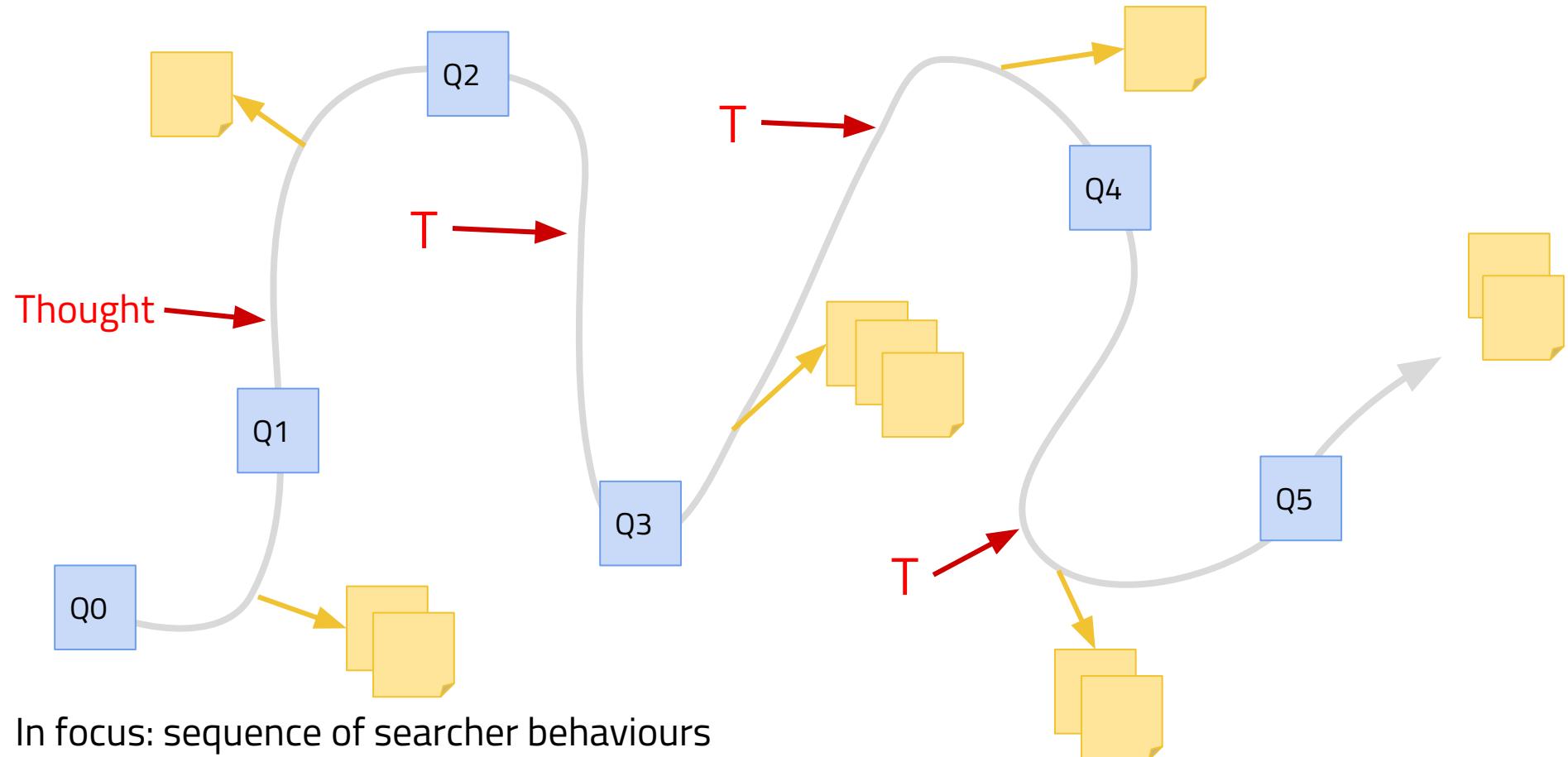


# Two early models of IIR

*"classic" IR model*



## Bates' berrypicking model (1989)



In focus: sequence of searcher behaviours  
Based on intuitions, informal observations

# Bates' berrypicking model (1989)

- Information needs **evolve over time**, they are not static throughout the search
- Users frequently start their search with just one sub-topic of a broader topic
- Each found piece of information can result in new ideas and search directions
- A query is not satisfied by a final retrieved set of documents, but by a **series of selections** of bits of information at each stage of the evolving search

**bit-at-a-time retrieval = berrypicking**



# Kuhlthau's Information Search Process model (1988)

Model designed based on **observations** of **high school students**' application of library skills (i.e. qualitative research)

Motivation: "*Findings are needed that define the **experience** of people in an information search from their **own perspective**.*"

Systematic development of theory

Goal: **grounded theory** of the library search process



# Kuhlthau's Information Search Process model (1988)

Exploratory study based on:

- Observations in the natural setting (school library)
- Interviews (45 minutes)
- Journals (diaries)
- Search logs
- Time lines
- Flow charts
- Assessed writing probes



*Describe how you felt when the teacher announced the research assignment.*

*Describe how and why you chose your topic.*

*How did you know when your search was completed?*

*What did you find most difficult about your search?*

Participants: 26 college-bound high school seniors

Assignment: write a paper

# Kuhlthau's Information Search Process model (1988)

Six stages

Task initiation

Topic selection

Prefocus exploration

Focus formulation

Information collection

Search closure

Feelings (affective)

uncertainty

optimism

confusion,  
frustration,  
doubt

clarity

sense of  
direction,  
confidence

relief

Thoughts (cognitive)

ambiguity

specificity

increased interest

Actions (physical)

seeking  
relevant  
information

seeking  
pertinent  
information

# Kuhlthau's Information Search Process model (1988)

Six stages

Task initiation

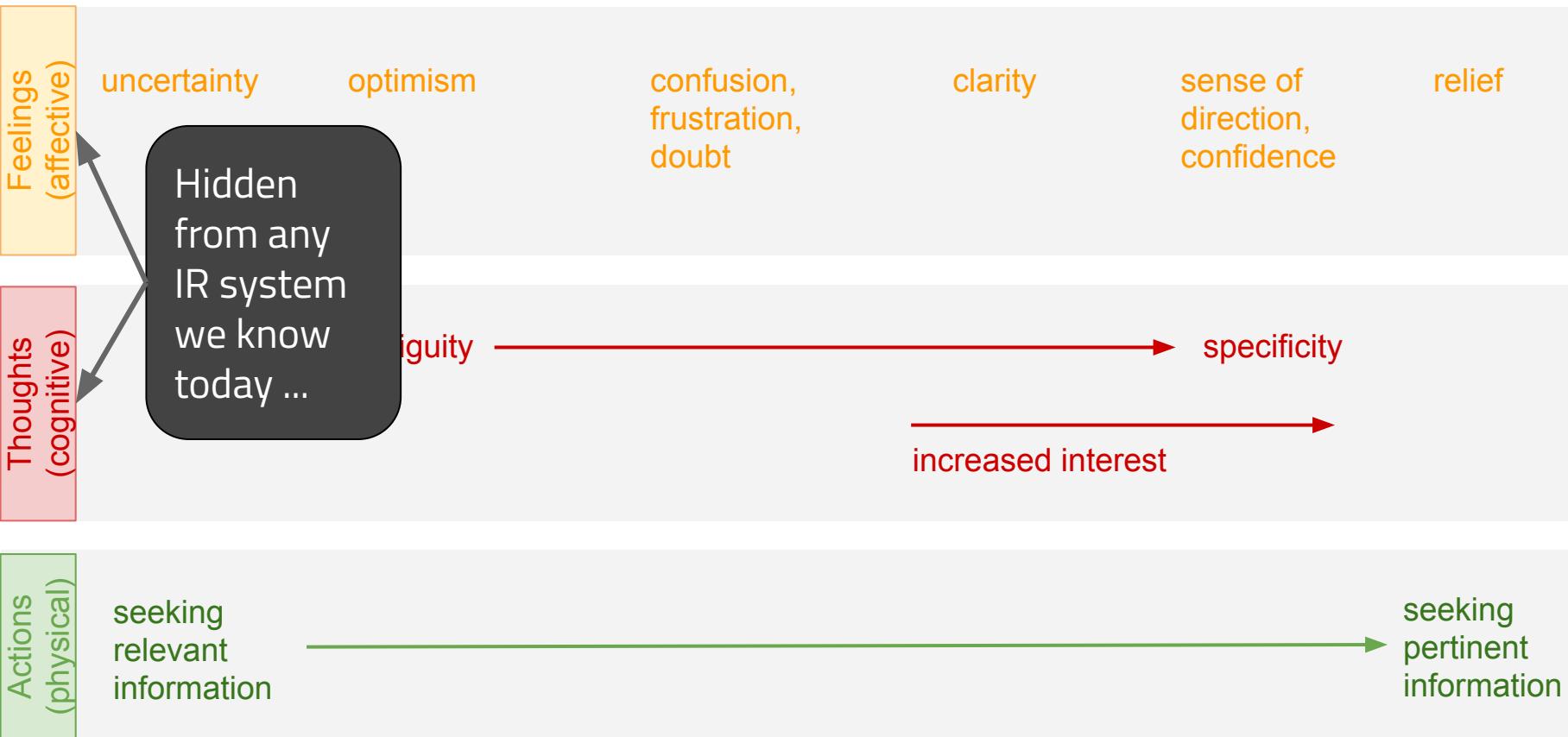
Topic selection

Prefocus exploration

Focus formulation

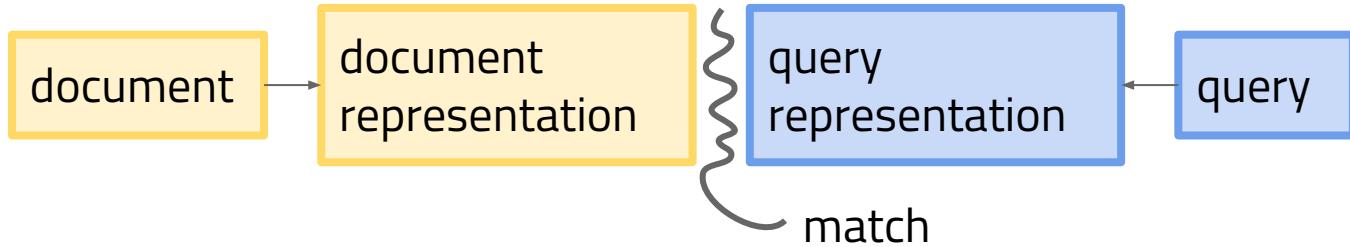
Information collection

Search closure



One of today's prevalent  
IIR modeling approaches

"classic" IR model



## Predictive models are needed

- Observational studies and descriptive models allow us to think but not to reason about interactive IR design decisions *e.g. is it better to show 20 query autocompletion items or just 3?*
- Interactive IR experiments have shown that system effectiveness and **user performance** do not necessarily correlate



space of all possible UI changes

UIs predicted to be useful by a model

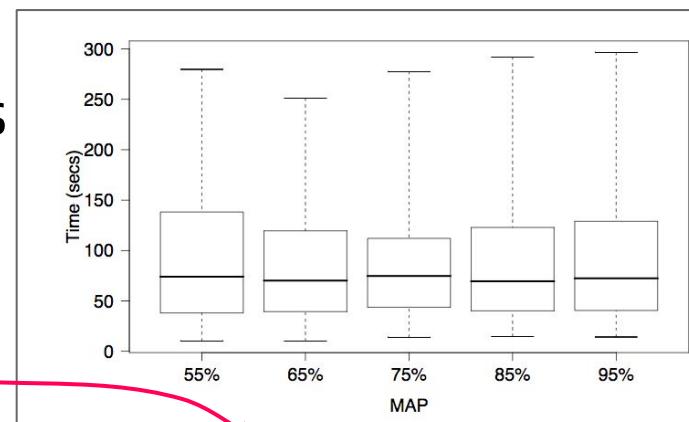


Figure 3: Time taken to find the first relevant document versus the mean average precision of the system used.

# Economic models of interaction

(Azzopardi et al, 2011-today)

Focus on understanding/predicting the behaviour of economic agents within an environment.

Economics is a field ripe with predictive models of costs and benefits; can we make use of them?

User interactions re-interpreted:

- Users take **actions** to advance towards their **goals**
- Each action has a **cost** (time, effort, cognitive load, etc.)
- An action may or may not lead to a **benefit** (saving time, finding new information, etc.)



# Economic models of interaction

(Azzopardi et al, 2011-toda

Representation of reality  
in an abstracted form;  
requires assumptions.

Having formulated a  
**mathematical model**, we can  
examine what actions:

- accrue the **most benefits**  
for a given cost
- incur the **least cost** for a  
given benefit level
- a rational user should take  
(given a task, interface,  
context, constraints) to  
achieve **optimal** results



# Economic models of interaction

(Azzopardi et al, 2011-today)

## Assumptions:

- Economic agents are **rational** and attempt to maximize their benefits
- Economic agents can **adapt** their strategies towards the optimal course of interaction



Let's look at two IR examples!

# Building economic models

1. Describe the problem context (who/what/how)
2. Specify the cost and benefit functions (keep it simple and then refine)
3. Solve the model (analytically, computationally, or graphically)
4. Use the model to generate hypotheses about behaviours (how do different variables influence interaction and behaviour)
5. Compare the predictions with observations in the literature and/or experimental data (model as a guide and evidence that [in]validates our models, leading to refinement)

iterate



# Economic model of querying

Goal: a model that describes the relationship between the length of the query and the costs/benefits of the query given its length

Longer queries tend to lead to better results; users do not use long queries.

Can we incentivize them?

How about trying this?

More to the point, does this halo around the search box:



motivate you to continue typing until the search box turns blue?



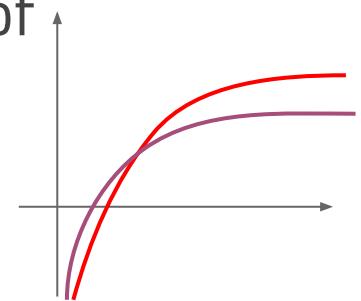
# Economic model of querying

Goal: a model that describes the relationship between the length of the query  $\mathbf{W}$  (in words) and the costs/benefits of the query given its length.

Modeling assumption: cost/benefit are a function of query length alone.

$$\underline{b(\mathbf{W}) = k \times \log_a(\mathbf{W} + 1)}$$

benefit function



$$\underline{c(\mathbf{W}) = \mathbf{W} \times c_w}$$

cost function

(i.e. the effort in querying)

Diminishing returns ( $a$  determines steepness) as the length increases with  $k$  as scaling factor (e.g. SE quality).

Effort to enter one word.

# Economic model of querying

Given the cost and benefit functions, we can compute the **profit (net benefit)** that the user receives for a query of length  $\mathbf{W}$ :



$$\pi = b(\mathbf{W}) - c(\mathbf{W}) = \mathbf{k} \times \log_a(\mathbf{W} + 1) - \mathbf{W} \times \mathbf{c_w}$$

Which query length maximizes the user's net benefit?  
Differentiate with respect to  $\mathbf{W}$  and solve:

$$\frac{\partial \pi}{\partial \mathbf{W}} = \frac{\mathbf{k}}{\log a} \times \frac{1}{\mathbf{W} + 1} - \mathbf{c_w} = 0$$

$$\mathbf{W}^* = \frac{\mathbf{k}}{\mathbf{c_w} \times \log a} - 1$$

# Economic model of querying

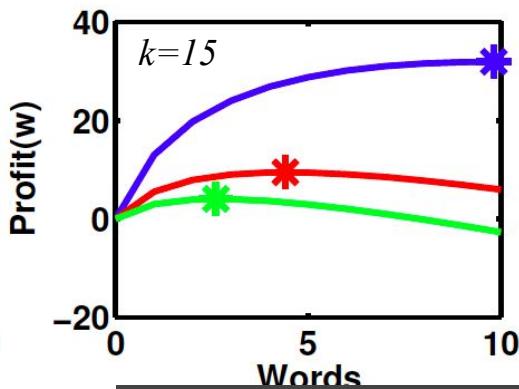
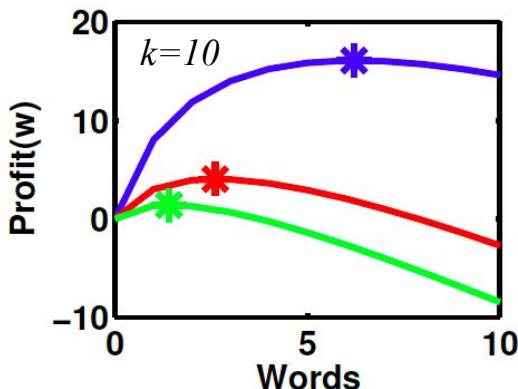
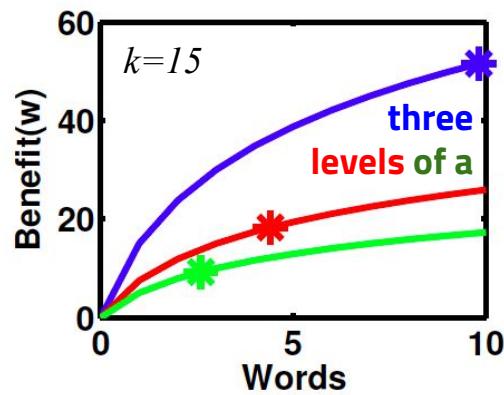
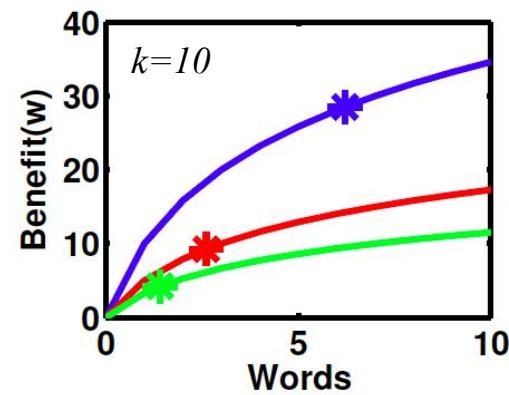
$$W^* = \frac{k}{c_w \times \log a} - 1$$

What does the model say about:

query halo effect

query autocompletion

SE with AND between query terms



Hypotheses based on this model:

- As the system performance ( $k$ ) increases, the query length increases
- If additional terms provide less and less benefit ( $a$  increases), queries decrease in length
- With decreasing cost of entering a word ( $c_w$ ), users tend to pose longer queries

# Economic model of assessing

Goal: a model that describes how users interact with a list of search results after having posed a query. Also known as "stopping behaviour".

Empirical findings: users stop when having found 'enough' or after N relevant docs or ...

Example: news retrieval

The screenshot shows a search results page from a Dutch news website. The search bar at the top contains the query 'delft verkiezingen'. Below the search bar, the results are displayed under the heading 'delft verkiezingen :'. There are six news items listed:

- Extra raadszetels in Rijswijk, Delft en Voorschoten na verkiezingen**  
Omroep West · Feb 19, 2018
- Weet jij al wat je gaat stemmen? Check hier de Delftse kieswijzer!**  
indebuurt · Feb 24, 2018
- 30 ideeën voor de stad tijdens Festival de Stem van Delft**  
Delftse Post (persbericht) (Blog) · Mar 6, 2018
- Alles over gemeenteraadsverkiezingen 2018 vind je bij Omroep West**  
Omroep West · Mar 10, 2018
- Stemwijzer gemeenteraadsverkiezingen 2018: dit zijn de stemwijzers**  
iCreate Magazine · Mar 13, 2018
- Jongeren bestormen de gemeenteraad**  
AD.nl · 14h ago

To the right of the results, there is a sidebar titled 'Related' containing a list of local names: Kieswijzer, Utrecht, Natuur & Milieu, Emmen, Noord-Brabant, de Stentor, Tilburg, Westland, Wegener, and Algemeen Dagblad.

# Economic model of assessing

Goal: a model that describes how users interact with a list of search results after having posed a query. Also known as “stopping behaviour”.

Modeling assumption: a user interacts with one list of results.

Cost function:

$$c(\mathbf{A}) = \mathbf{c_q} + \mathbf{A} \times \mathbf{c_a}$$

cost of assessing  
 $\mathbf{A}$  items

cost of the query

cost of assessing 1 doc.

Benefit function:

$$b(\mathbf{A}) = \mathbf{k} \times \mathbf{A}^{\beta}$$

Determines how quickly the benefit from information diminishes  
 $\beta < 1$  usually

# Economic model of assessing

Given the cost and benefit functions, we can compute the **profit (net benefit)** the user receives when assessing to a depth of  $\mathbf{A}$ :

$$\pi = b(\mathbf{W}) - c(\mathbf{W}) = \mathbf{k} \times \mathbf{A}^\beta - \mathbf{c_q} - \mathbf{A} \times \mathbf{c_a}$$

Differentiate with respect to  $\mathbf{A}$  and solve:

$$\frac{\partial \pi}{\partial \mathbf{A}} = \mathbf{k} \times \beta \times \mathbf{A}^{\beta-1} - \mathbf{c_a} = 0$$

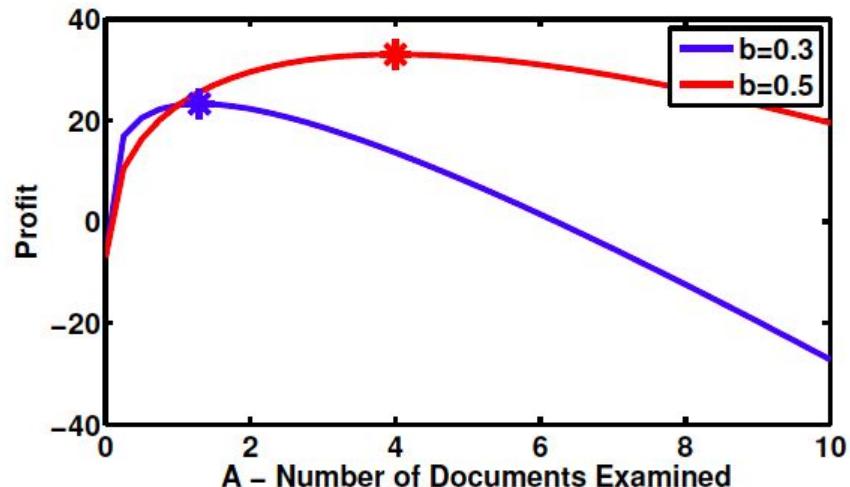
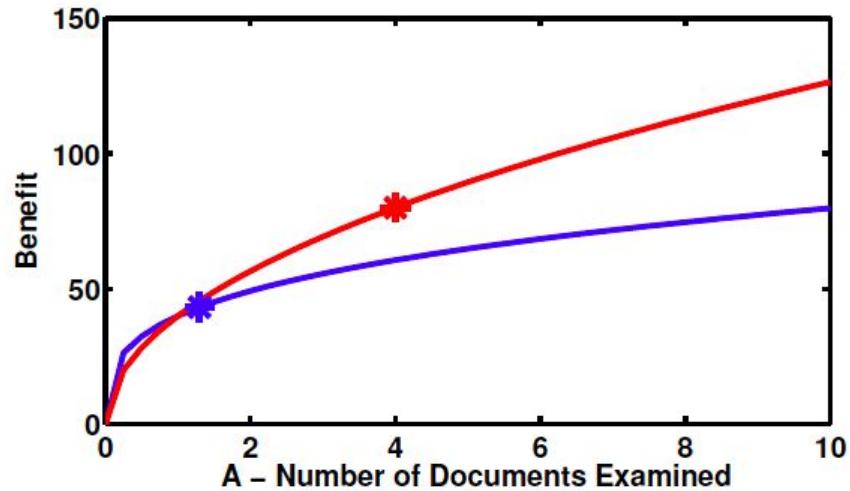
$$\mathbf{A}^* = \left( \frac{\mathbf{c_a}}{\mathbf{k} \times \beta} \right)^{\frac{1}{\beta-1}}$$

# Economic model of assessing

$$A^* = \left( \frac{c_a}{k \times \beta} \right)^{\frac{1}{\beta-1}}$$

Model interpretation:

- If the performance of the query is poor, there is little incentive to examine search results.
- If the cost of assessing documents is very high, fewer documents are examined.
- The cost of a query does not impact user behaviour (as it is a fixed cost).



# Economic model of searching

Goal: a model that describes the process of searching over a session - numerous queries can be issued, the user examines a number of items per query.

It gets more complicated quickly ...

$$c(Q, V, S, A) = c_q \cdot Q + c_v \cdot V \cdot Q + c_s \cdot S \cdot Q + c_a \cdot A \cdot Q$$

A user poses a  
number of queries

... examines a  
number of SERPs  
per query

... examines a  
number of snippets  
per query

Take-away message: models can be as simple/complex as desired.

# Economic models of interaction

(Azzopardi et al., 2011-today)

## Challenges:

- **Estimation of costs and benefits** and their respective units (temporal, fiscal, satisfaction, enjoyment, ...)
- Assumption that users seek to max. their benefit
- Is the model sufficiently realistic wrt. user and environment?
- **Design of experiments**



flickr@32193702@N07

Final note:

Make sure to know the difference  
between replication and  
reproducibility (according to ACM).

Slack: in43252019.slack.com

Email: in4325-ewi@tudelft.nl