IN4325

Query autocompletion and Interactive IR

Claudia Hauff (WIS, TU Delft)

The big picture

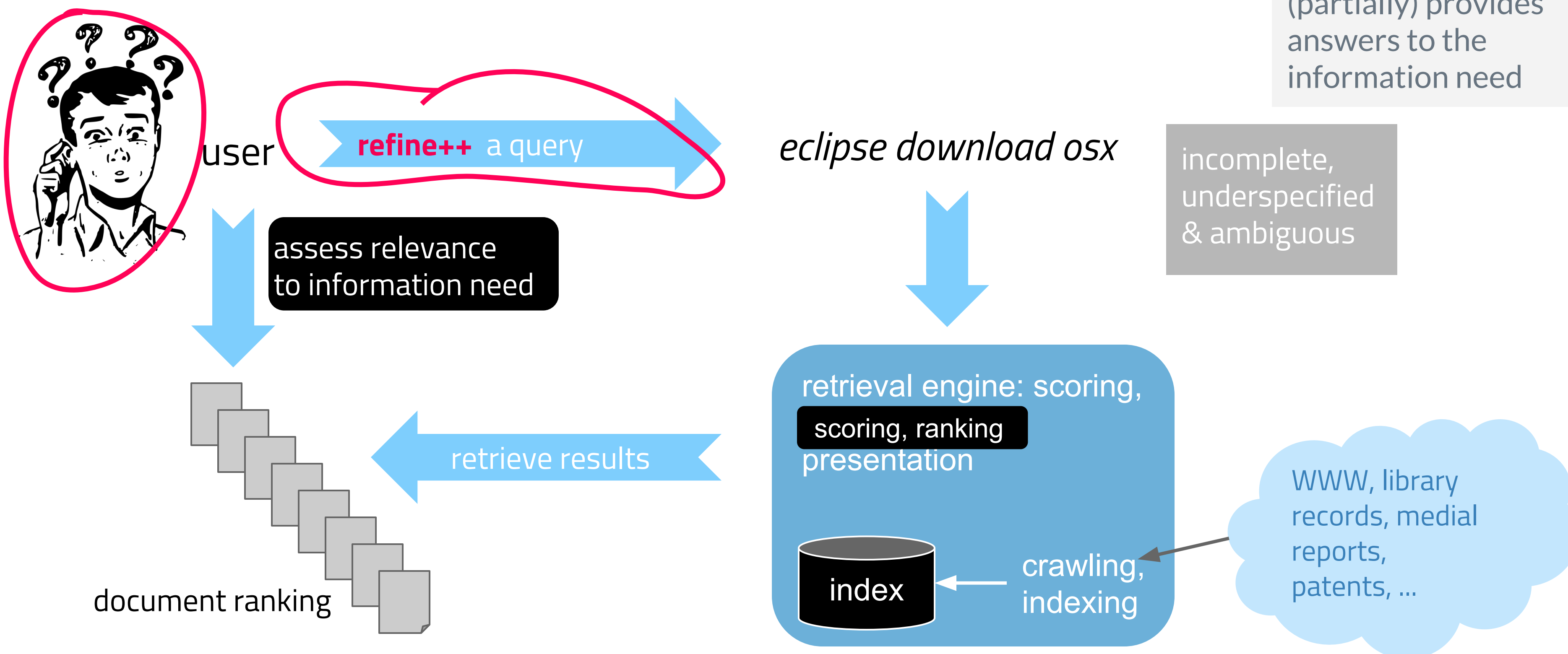
The essence of IR

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

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



Query autocompletion

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 autocompletion

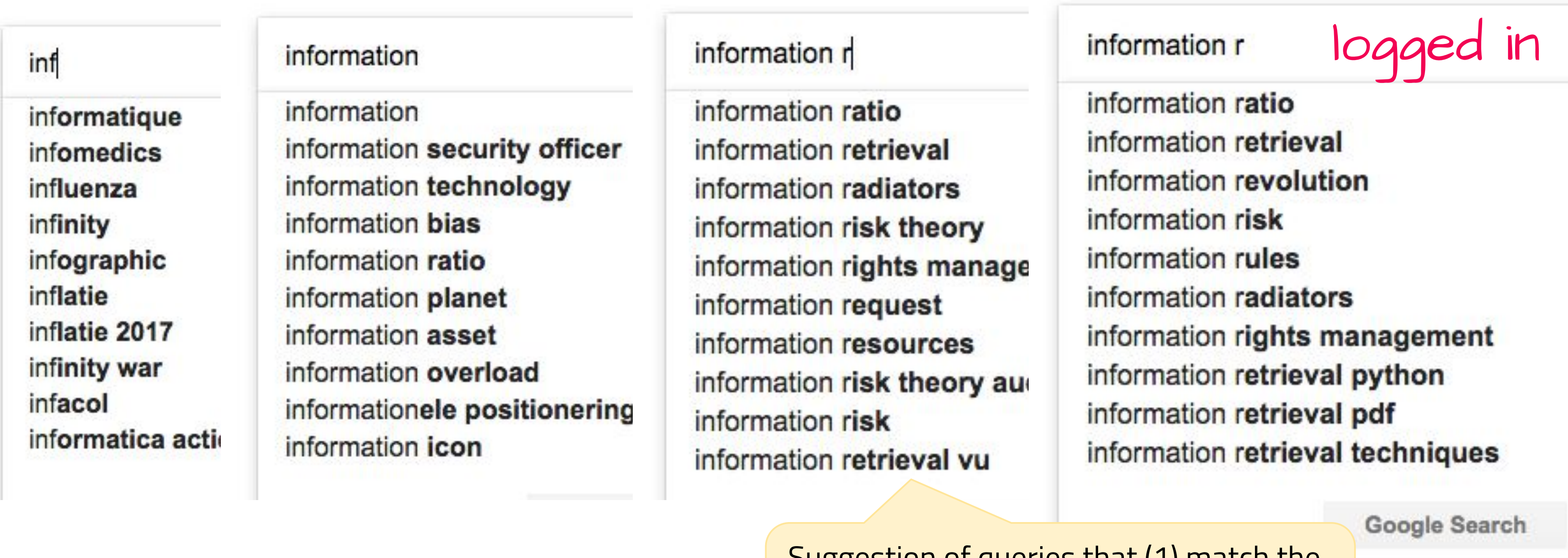
Select the complete query to replace your original query with whilst typing.

Related queries

Select the complete query to replace your original query with.

Last time

Overview



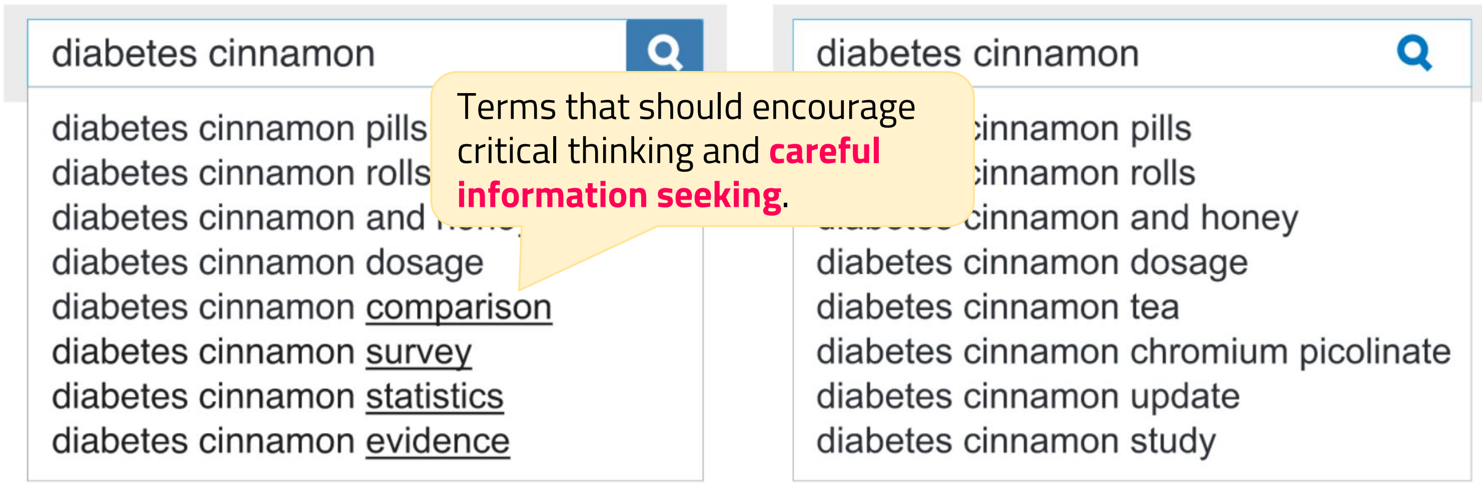
Suggestion of queries that (1) match the user's information needs and (2) yield a high-quality result ranking.

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



(1) QAC with query priming

(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 autocompletion

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

Last time

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

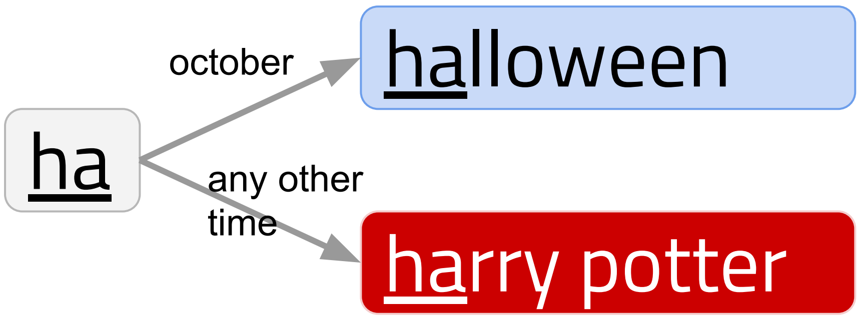
Clicked documents ranker

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)

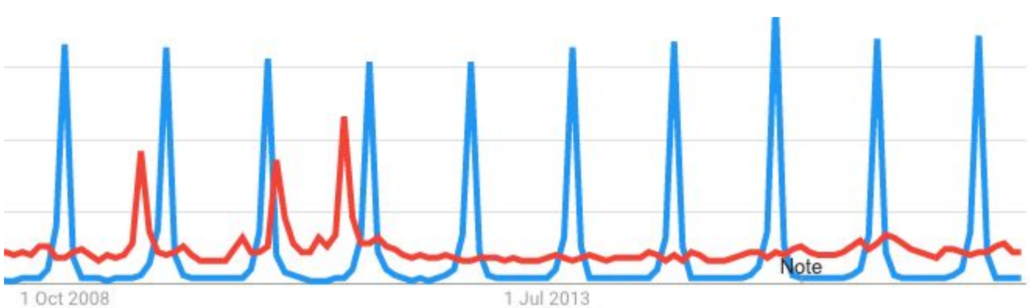
https://dl.acm.org/citation.cfm?id=2348364

Last time

Time-sensitive query autocompletion

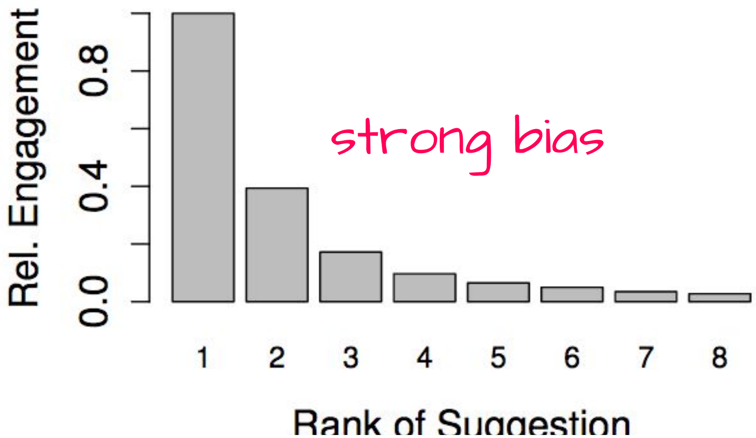


Approach: apply time-series modeling and rank candidates according to their forecasted frequencies



User engagement with query autocompletion

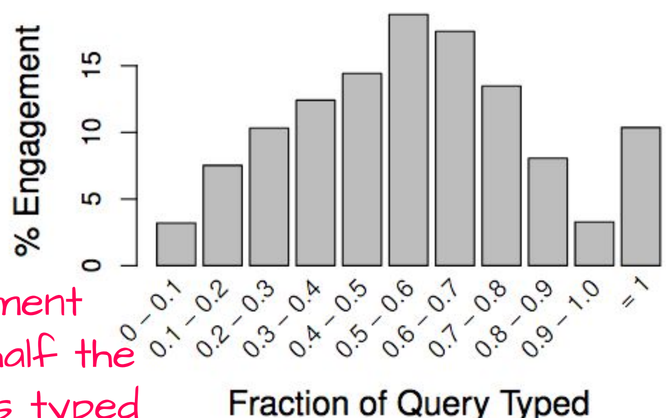
1.6M queries from Bing's search log



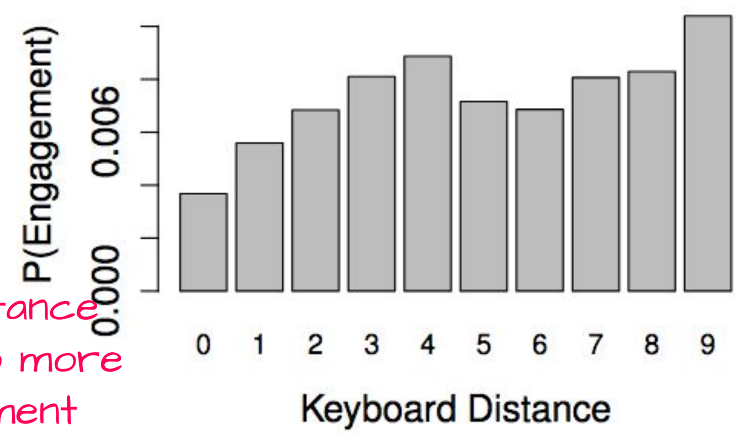
strong bias

Most engagement at word boundaries and after 3 chars of typing a word

Most engagement after half the query is typed



Larger key distance leads to more engagement



Web search engines are not everything …

Large user base

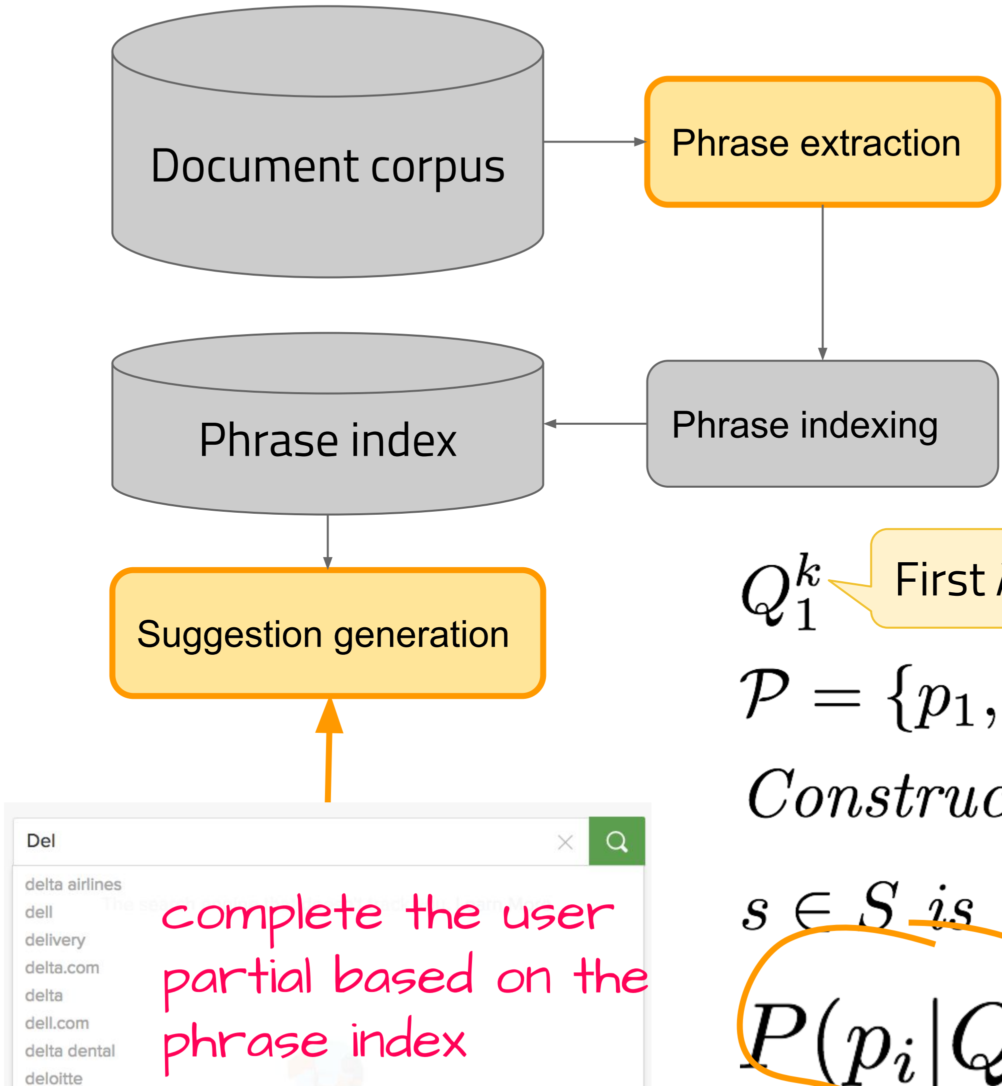
Assumptions:

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 autocompletion

Corpus-based query suggestions



- N-grams: unigrams, bigrams, trigrams
- Skip over stopwords when generating phrases

Q_1^k First k characters typed

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

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

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

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

https://dl.acm.org/citation.cfm?id=2010023

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

https://dl.acm.org/citation.cfm?id=2010023

Corpus-based query suggestions

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

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

$$Q_1^k = Q_c + Q_t$$

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

$$P(p_i|Q_1^k) = \frac{P(p_i) \times P(Q_1^k|p_i)}{P(Q_1^k)}$$

according to Bayes' theorem

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

Simplifying assumption: conditional independence

https://dl.acm.org/citation.cfm?id=2010023

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

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

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

https://dl.acm.org/citation.cfm?id=2010023

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

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

$$P(p_i)P(Q_t|p_i)$$

$$= P(p_i, Q_t)$$

$$= P(Q_t)P(p_i|Q_t)$$

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

https://dl.acm.org/citation.cfm?id=2010023

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

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

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

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

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

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.

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.

https://dl.acm.org/citation.cfm?id=2010023

Corpus-based query suggestions

Data sets

TREC: 200K news articles by the Financial Times published between 1991-1994, 40 test queries

Ubuntu: 100K discussion threads, 40 test queries

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)

https://dl.acm.org/citation.cfm?id=2010023

What are good metrics?

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

presented approach

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 government		falkland islands government
file system mount	mounter	mount the drive	1982 falklands conflict		falklands war in 1982
mount point type	mountunmount	mount the partition	falkland arms		1982 falklands war
system mount point type	mountpoints	file system mount	falklanders		invasion of the falklands

https://dl.acm.org/citation.cfm?id=2010023

12 assessors (colleagues), majority vote on top 10 suggestions

Corpus-based query suggestions

Rating	Meaning
Y	Yes, a meaningful suggestion
N	No, not a meaningful suggestion, or badly formed as a query
D	An (almost) duplicate suggestion, conveys no new information
??	Not sure

presented approach

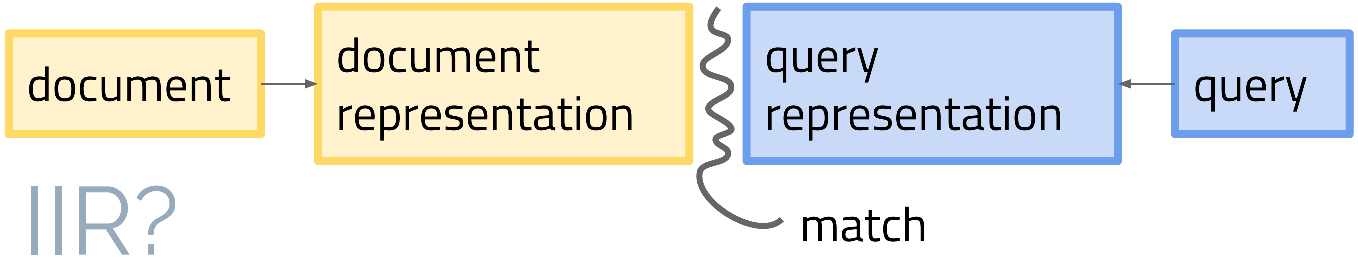
Ubuntu			
	SimSearch	**CompSearch**	**Probabilistic**
Type-A	1.00	1.00	1.00
Type-B	0.75	1.00^{s}	1.00^{s}
Overall	0.875	1.00^{s}	1.00^{s}
TREC			
	SimSearch	**CompSearch**	**Probabilistic**
Type-A	1.00	1.00	1.00
Type-B	0.15	0.95^{S}	1.00^{S}
Overall	0.575	0.975^{S}	1.00^{S}

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)

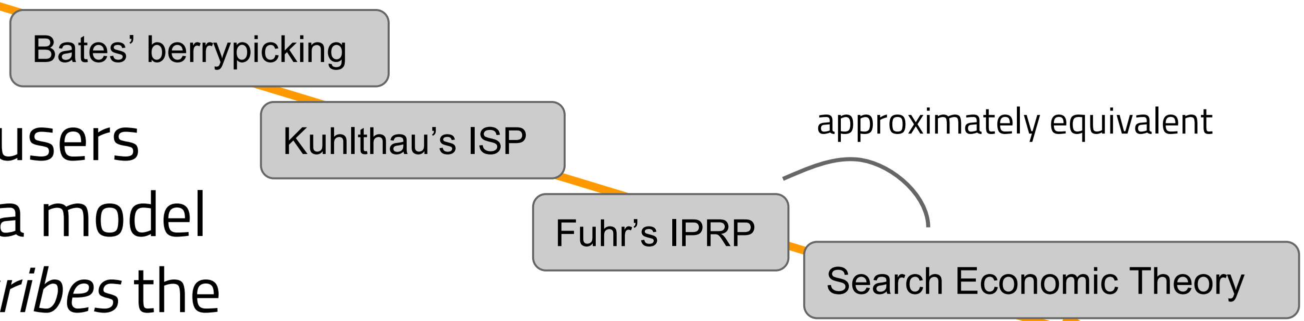
“In interactive information retrieval, ***users are typically studied*** *along with their interactions with systems and information.”*
(Diane Kelly)

*“… the interactive approach to IR has led to a f****ocus on the user-oriented activities*** *of query formulation and reformulation, and inspection and judgement of retrieved items …”* (Nick Belkin)

From past to present

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

Conceptual, observational and empirical work



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

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

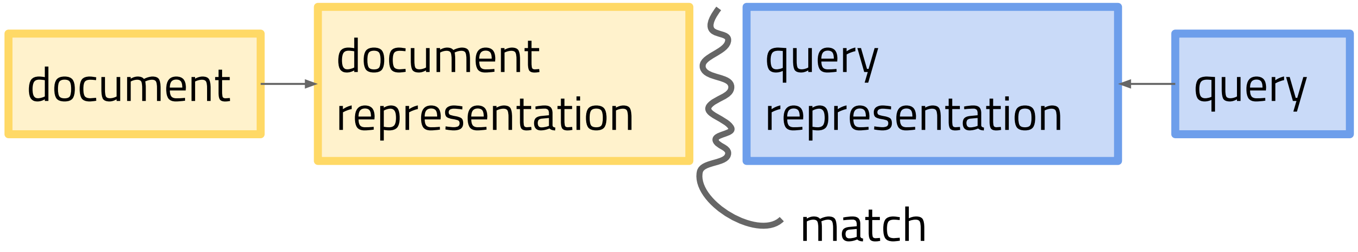
Most often in IR when we talk about models we mean retrieval models.

Not now though!

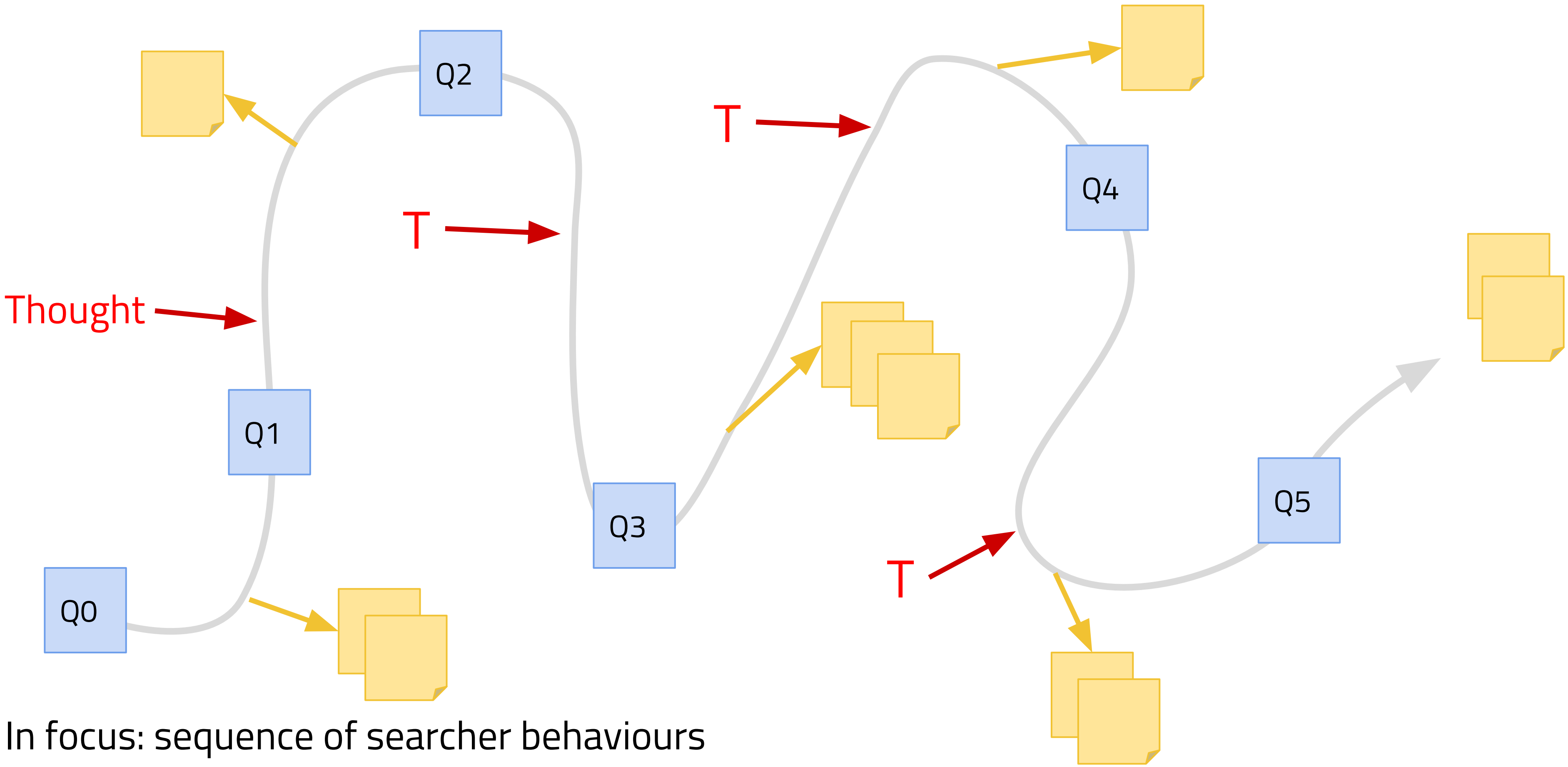
Now: models for interactive information seeking and retrieval

Two early models of IIR

"classic" IR model



Bates' berrypicking model (1989)



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

https://www.emeraldinsight.com/doi/pdfplus/10.1108/eb024320

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***

https://www.emeraldinsight.com/doi/pdfplus/10.1108/eb024320

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

flickr@juanillooo

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

Participants: 26 college-bound high school seniors

Assignment: write a paper

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?

Kuhlthau's Information Search Process model (1988)

Six stages

Task initiation	Topic selection	Prefocus exploration	Focus formulation	Information collection	Search closure
uncertainty	optimism	confusion, frustration, doubt	clarity	sense of direction, confidence	relief

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
Feelings (affective)	uncertainty	optimism	confusion, frustration, doubt	clarity	sense of direction, confidence	relief

Thoughts (cognitive)

iguity → specificity

→ increased interest

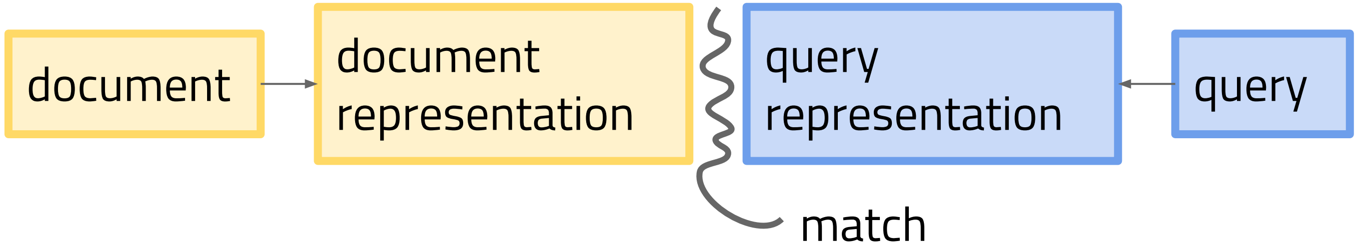
Actions (physical)

seeking relevant information → seeking pertinent information

Hidden from any IR system we know today ...

One of today's prevalent IIR modeling approaches

"classic" IR model

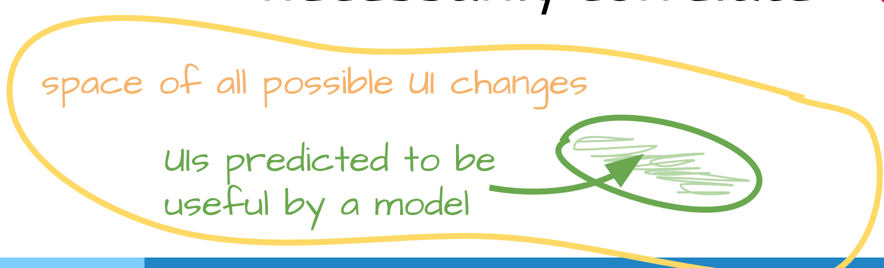


Predictive models are needed

- 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



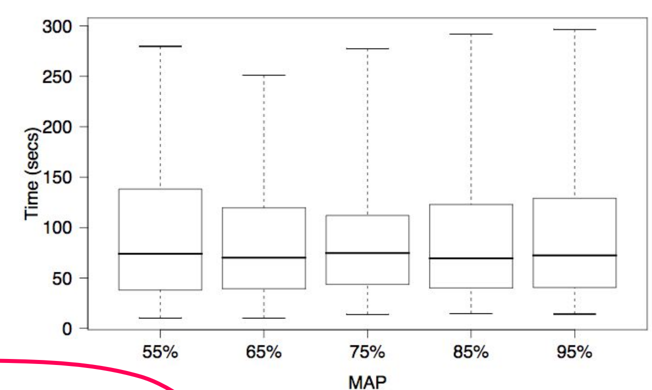


Figure 3: Time taken to find the first relevant document versus the mean average precision of the system used.

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.)

flickr@arabani

Tutorial: http://zuccon.net/publications/azzopardi-zuccon-2017-tutorial-economics.pdf

Economic models of interaction

(Azzopardi et al, 2011-today)

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

flickr@arabani

Tutorial: http://zuccon.net/publications/azzopardi-zuccon-2017-tutorial-economics.pdf

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

flickr@arabani

Let's look at two IR examples!

Building economic models

iterate

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)

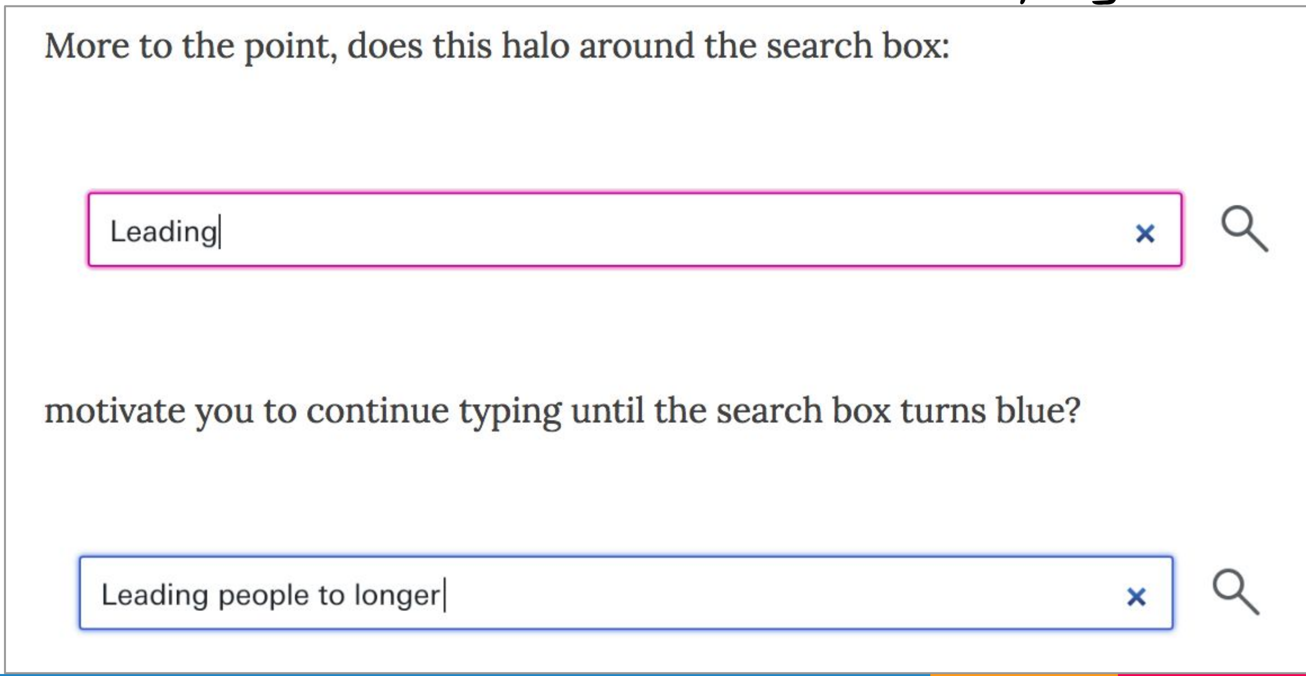
Economic model of querying

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:

Leading

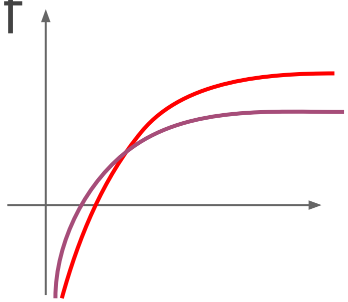
motivate you to continue typing until the search box turns blue?

Leading people to longer

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.



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

benefit function

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

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

cost function
(i.e. the effort in querying)

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 $\mathbf{W}$:

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

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

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

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

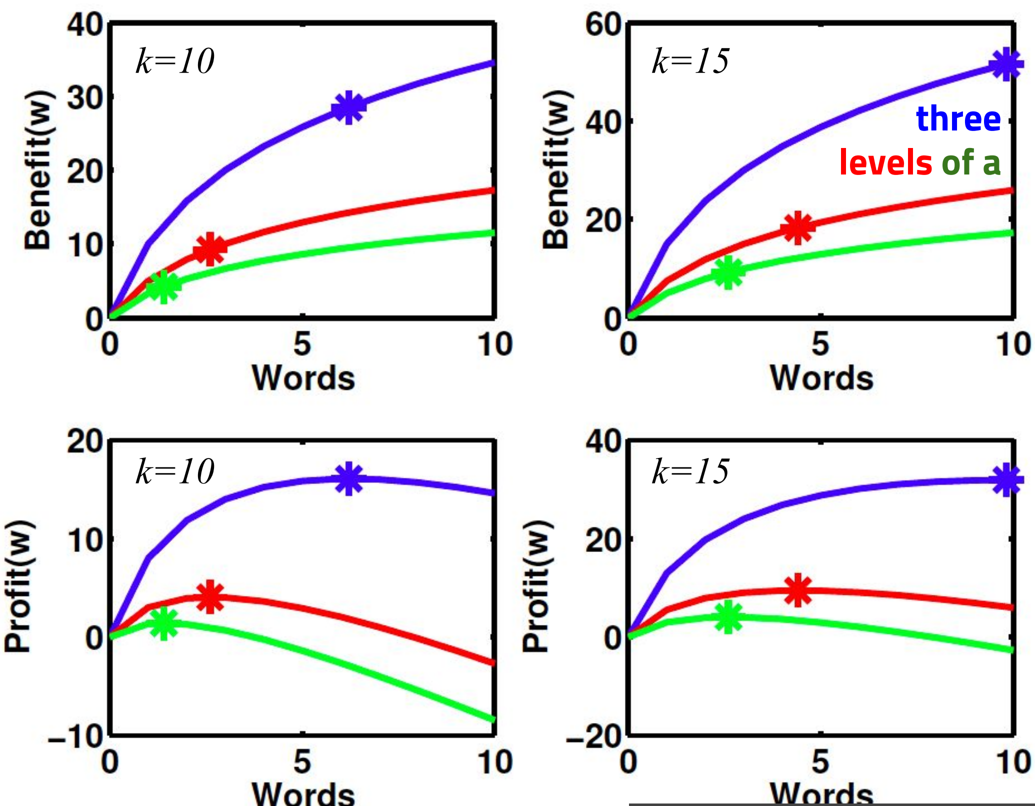
Economic model of querying

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

What does the model say about:

query halo effect

query autocompletion



Hypotheses based on this model:

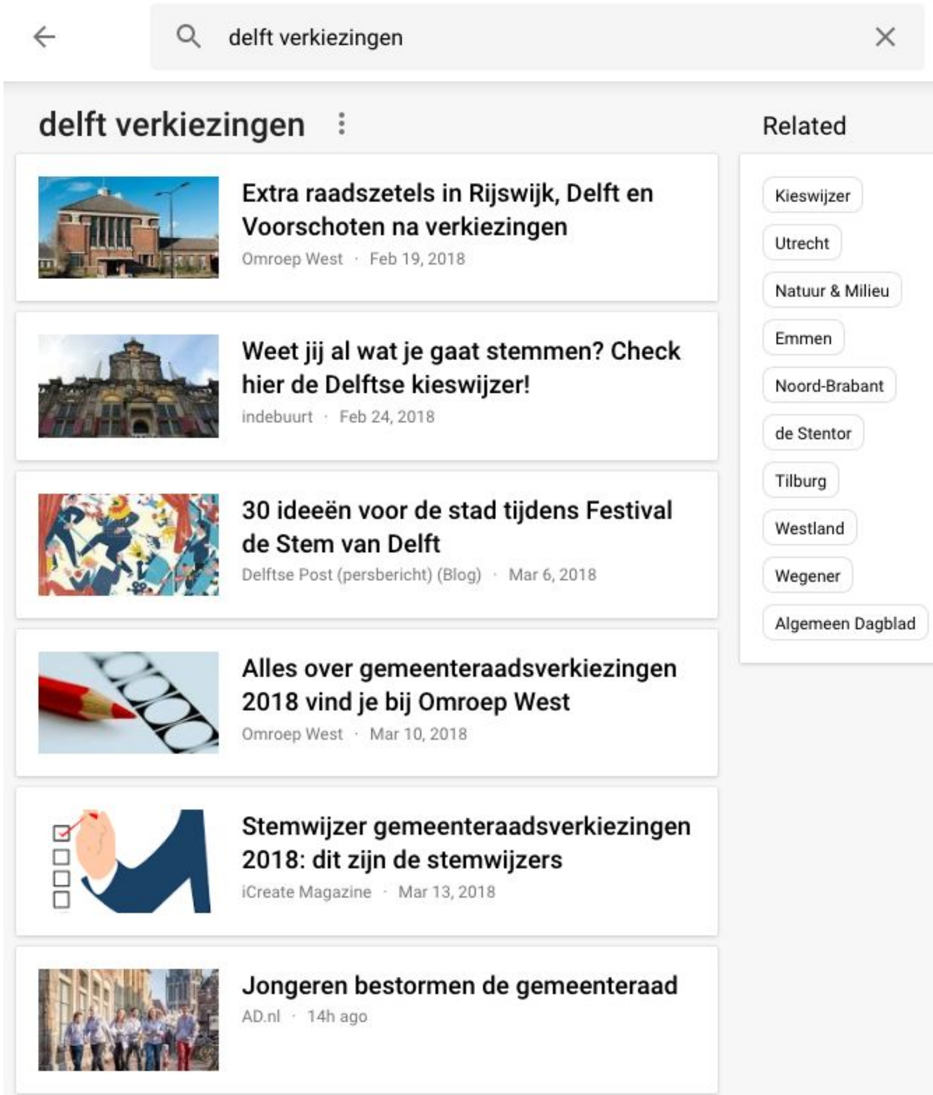
- As the system performance ($\mathbf{k}$) increases, the query length increases
- If additional terms provide less and less benefit ($\mathbf{a}$ increases), queries decrease in length
- With decreasing cost of entering a word ($\mathbf{c_w}$), users tend to pose longer queries

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



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 of assessing $\mathbf{A}$ items

cost of the query

cost of assessing 1 doc.

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

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:

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

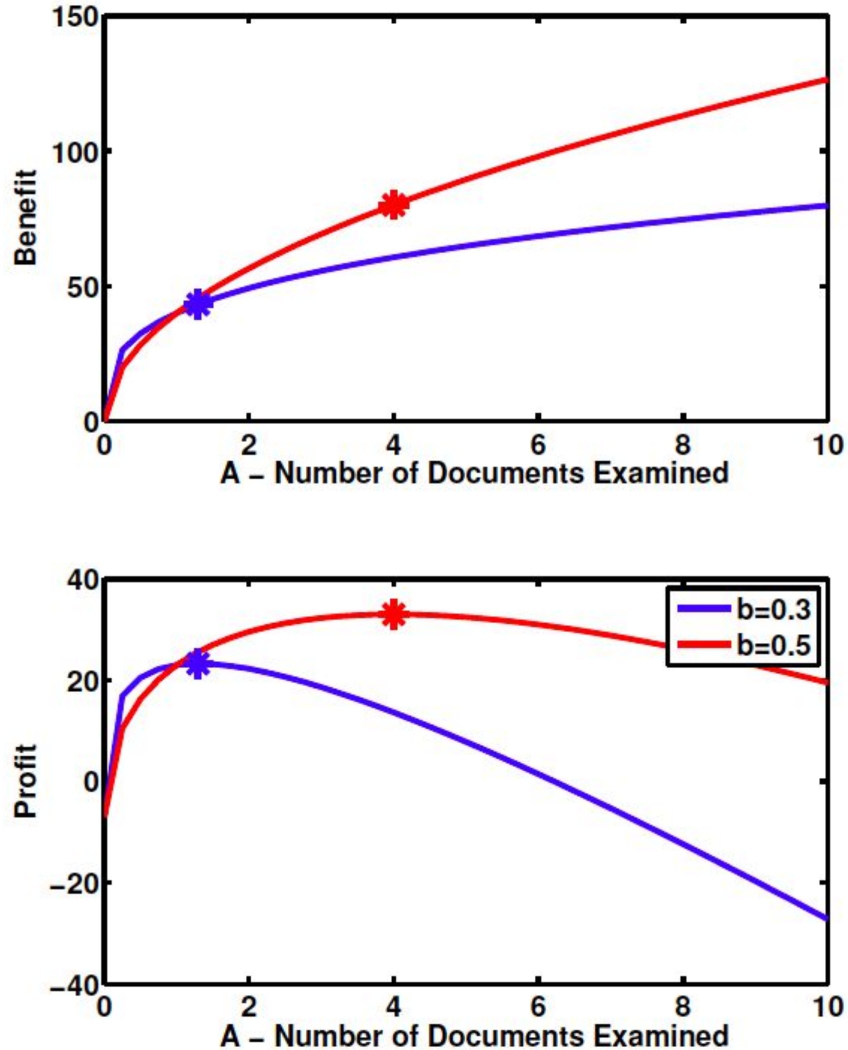
$$\mathbf{A}^{*} = \left(\frac{\mathbf{c_a}}{\mathbf{k} \times \beta} \right)^{\frac{1}{\beta - 1}}$$

Economic model of assessing

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

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 ...

$$c(\mathbf{Q}, \mathbf{V}, \mathbf{S}, \mathbf{A}) = \mathbf{c_q}.\mathbf{Q} + \mathbf{c_v}.\mathbf{V}.\mathbf{Q} + \mathbf{c_s}.\mathbf{S}.\mathbf{Q} + \mathbf{c_a}.\mathbf{A}.\mathbf{Q}$$

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**

Upcoming project deadline:

March 8 - intermediate project report (**feedback moment**).
If you are stuck, come to the support hours!

Slack: in43252019.slack.com
Email: in4325-ewi@tudelft.nl