DIS17 – Search Engine Technologies

01 - Introduction

Philipp Schaer, Technische Hochschule Köln, Cologne, Germany

Version: WS 2021

Technology
Arts Sciences
TH Köln

Information Retrieval Research Group

Prof. Dr. Philipp Schaer

- Information retrieval, evaluation of IR systems, digital libraries

Timo Breuer, M.Sc.

- Living Labs Infrastructure for Information Retrieval, project STELLA

Fabian Haak, M.Sc.

- Sentiment Analysis, Query Expansion, Policital Retrieval, project ESUPOL

Björn Engelmann, M.Sc.

- Information Extraction, Machine Learning, Scientic Journalism, project JoIE

Projects, jobs, theses: **https://ir.web.th-koeln.de**

Information Retrieval Research Group

Prof. Dr. Philipp Schaer

- Information retrieval, evaluation of IR systems, digital libraries

Timo Breuer, M.Sc.

- Living Labs Infrastructure for Information Retrieval, project STELLA

Fabian Haak, M.Sc.

- Sentiment Analysis, Query Expansion, Policital Retrieval, project ESUPOL

Björn Engelmann, M.Sc.

- Information Extraction, Machine Learning, Scientic Journalism, project JoIE

Projects, jobs, theses: **https://ir.web.th-koeln.de**

Some notes on language matters...

Most of the content of this course will be in English:

- The **slides** and **worksheets** of this course will be in English.
- The **text book** we will use, is freely available in English.
- The **additional materials** and **web resources** we point to are mostly in English.
- The **assigments** will be in English.
- The **lecture itself** will be in English!

But:

- Of course, you can **speak** to me in German.
- You can still write your **assignments** in German.

Examination

There is **NO final exam** for this module!

Instead, you will work on **4 assignments** over the semester

- Assignment I: **Each of you** has to successfully complete a **tutorial** on the search engine Solr
- Assignment II: **Your team** has to submit a concept paper on your approach and plans for the Mini-TREC challenge
- Assignment III: **Your team** has build and evaluate a search engine on finding relevant documents on COVID-19 and to **submit the results (runs)**
- Assignment IV: **Your team** will prepare a **6-8-page paper** about your work. You describe, analyse and discuss your approach and results

Scoring and assignments

Individual assignment I (20 points)

- Do it on your own!
- Send a photo of your successfully finished **tutorial** and your **student id**

Group assignments II - IV (80 points)

- Concept paper (10 points)
- Run submissions (30 points)
- Term paper (40 points)
- Make sure **to give credit**! You have to document who contributed to what in you submissions!

- Your finals grade is determined using this table:

<50	50	55	60	65	70	75	80	85	90	95
n.b.	4,0	3,7	3,3	3,0	2,7	2,3	2,0	1,7	1,3	1,0

Rules for this course

No email support! Ask questions online!

- There will be a discussion forum at GitHub for this course - Use it!!
- I know – Nobody wants to be the one asking “stupid” questions.
- But: Your fellow students have the same issues – Trust me!
- Ask a lot of questions and try to **help your fellows**.

Help each other! Help yourself!

- **Tutorial** sessions are **interactive** – **get the most out of it!**
- **Active helpers** might get some **extra points**!

Code of Conduct.

- No. Plagiarism. Ever. You might get kicked out of the course!
- Whenever you pick up the ideas/works of other: Cite your sources!
- We might get **suspicious**… And we might come up with the idea to ask **embarrassing questions**. Don’t embarrass yourself!

Search Engine Technologies vs. SEO

Search Engine Optimization (SEO)

- is fancy and totally “in”, but by far not the core of search engines technologies
- focus on improving the rank of your product in a search engine developed by others

Search Engine Technologies

- developing a search engine and direct application of IR
- focus on improving the rank of products from others in your search engine to aid your users

Learning Outcomes

- We will not do any SEO, but rather

What

- understand and develop our own search engine on some data
- learn how to configure the search engine to specific requirements
- learn how to index the data and develop ranking algorithms
- learn how to evaluate our search engine

With

- use state-of-the-art technology like *Solr* with Python
- evaluate our search engine based on a test collection on COVID articles

What for

- be able to apply this knowledge on any other data
- be able to adapt to other use cases

Schedule for lectures WS 2021/22

15.10.21	Introduction	
22.10.21	IR in a Nutshell and TREC Evaluations	
29.10.21	Solr in a Nutshell	
05.11.21	Indexing	
12.11.21	Hands-on Session on Solr	
19.11.21	Queries	Assignment I due (Solr tutorial)
26.11.21	Project week (no lecture)	
03.12.21	Ranking	
10.12.21	Mini-TREC (conflict project management?)	Assignment II due (concept paper)
17.12.21	Mini-TREC	
24.12.21	Christmas (no lecture)	
31.12.21	Christmas (no lecture)	
07.01.22	Mini-TREC	Assignment III (run submission)
14.01.22	Results of Mini-TREC (World Café)	
28.02.22		Assignment IV (term paper)

Forming Teams

Form **teams of 3-4 students**

- These teams participate in the **Mini-TREC** challenge and have to submit their work
- Create a **cool team name**
  - seriously, the crazier the name the better :-)

Register your team here:

- https://forms.gle/fXXdQG8qr7YVXWXLA
- If you can't find a team, we will put you in a **random team**
- Better: Form the teams **TODAY**!!!
- Teams should be done by the end of next week

Welcome back

https://www.youtube.com/watch?v=YuOBzWF0Aws

Search engines are everywhere…

But, what exactly are search engines?

Search Engines: **The application of IR**

- performance is of high relevance (pun intended), scalability, specific problems, adaptability, …

Types of search engines

- documents and data, personalized retrieval, retail stores (e-commerce), the web and other networks, multi-modal search (“find restaurant next to my position”)

Biases and other **Limitations**

- only anticipating what the user wants, prone to trends (can be exploited), amplifies what users want (discrimination possible)

Search is all around us!

Every morning we wake up and

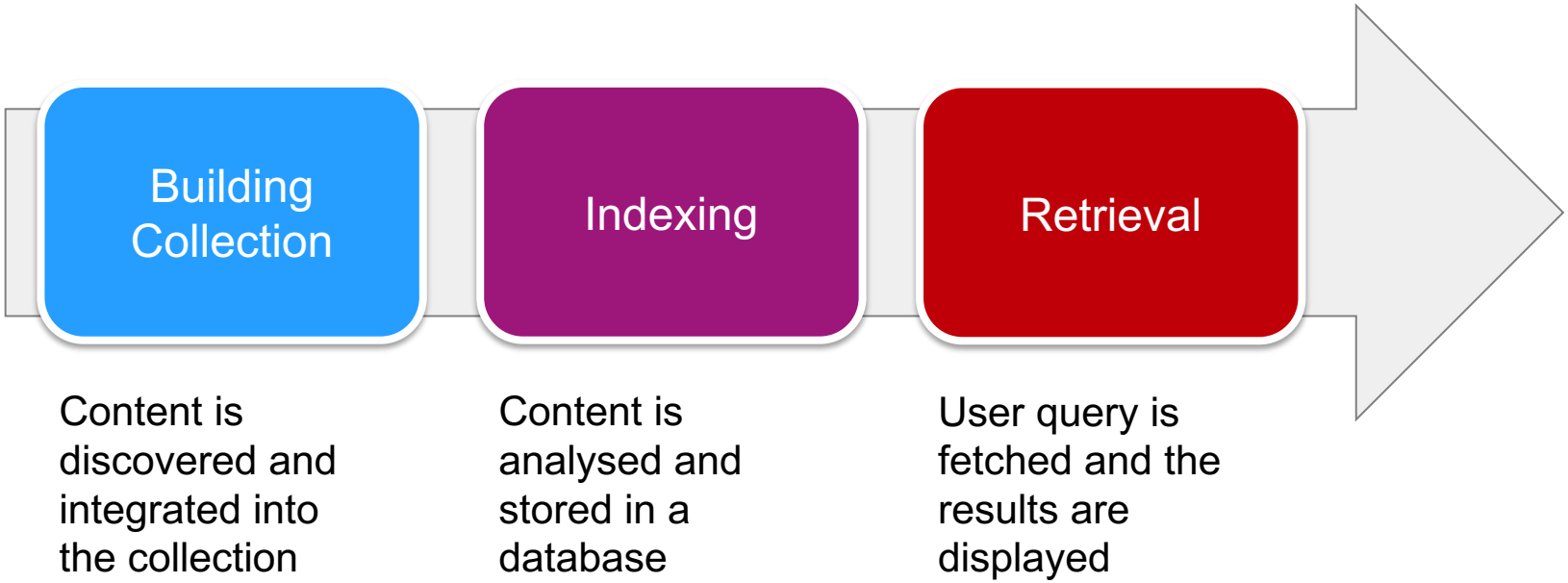
- fiddle with our smartphone to look at news, scan Facebook, read Whatsapp, and check mails
- before getting out of bed, we probably **interacted with a dozen search applications** without much thought

Ask yourself, **did you**…

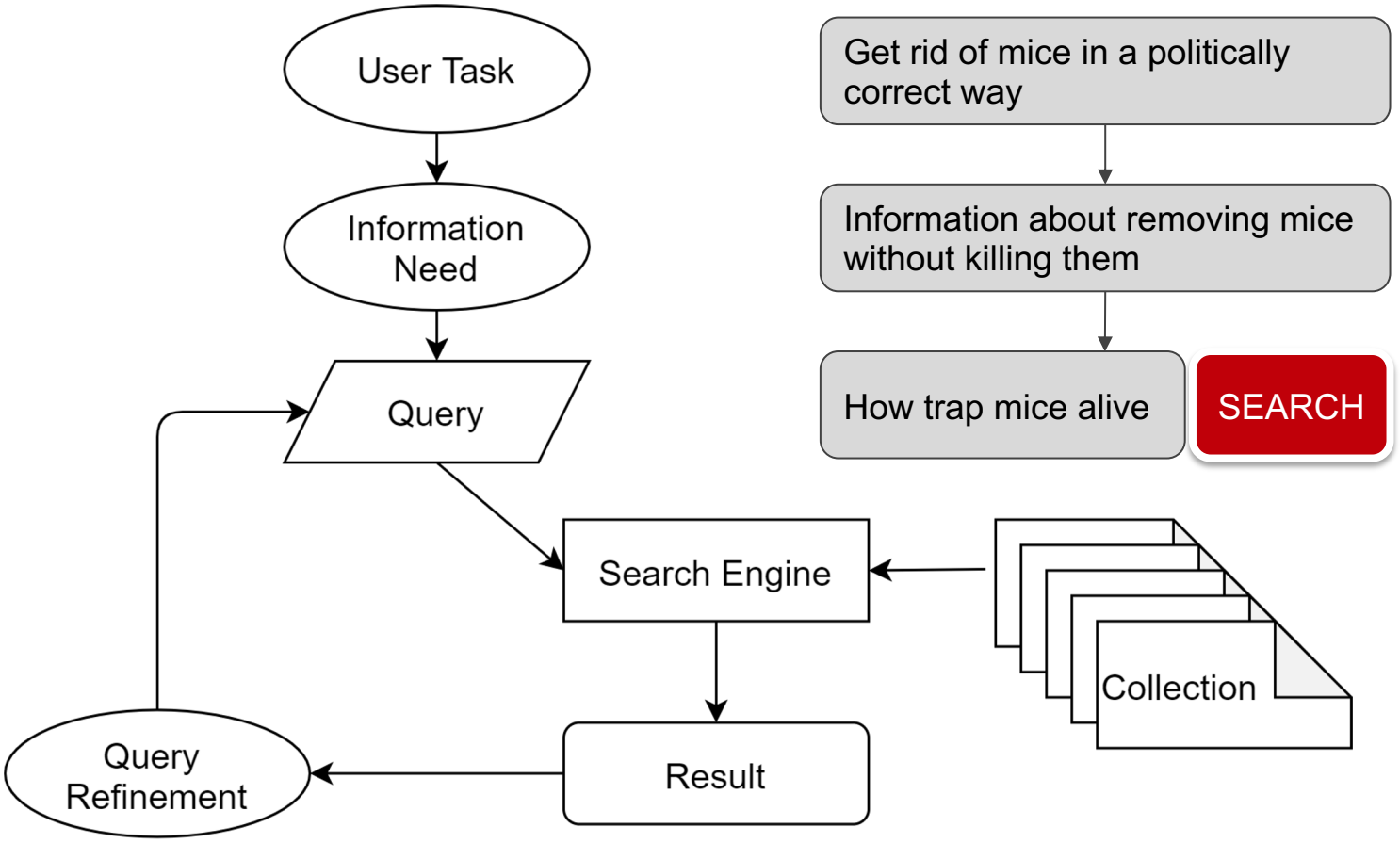
- send a message to a friend that you found in your contact list?
- search for a crucial email?
- talk to Siri?
- look for an available shared bike or when to take the next bus?

Search Engines and how they work

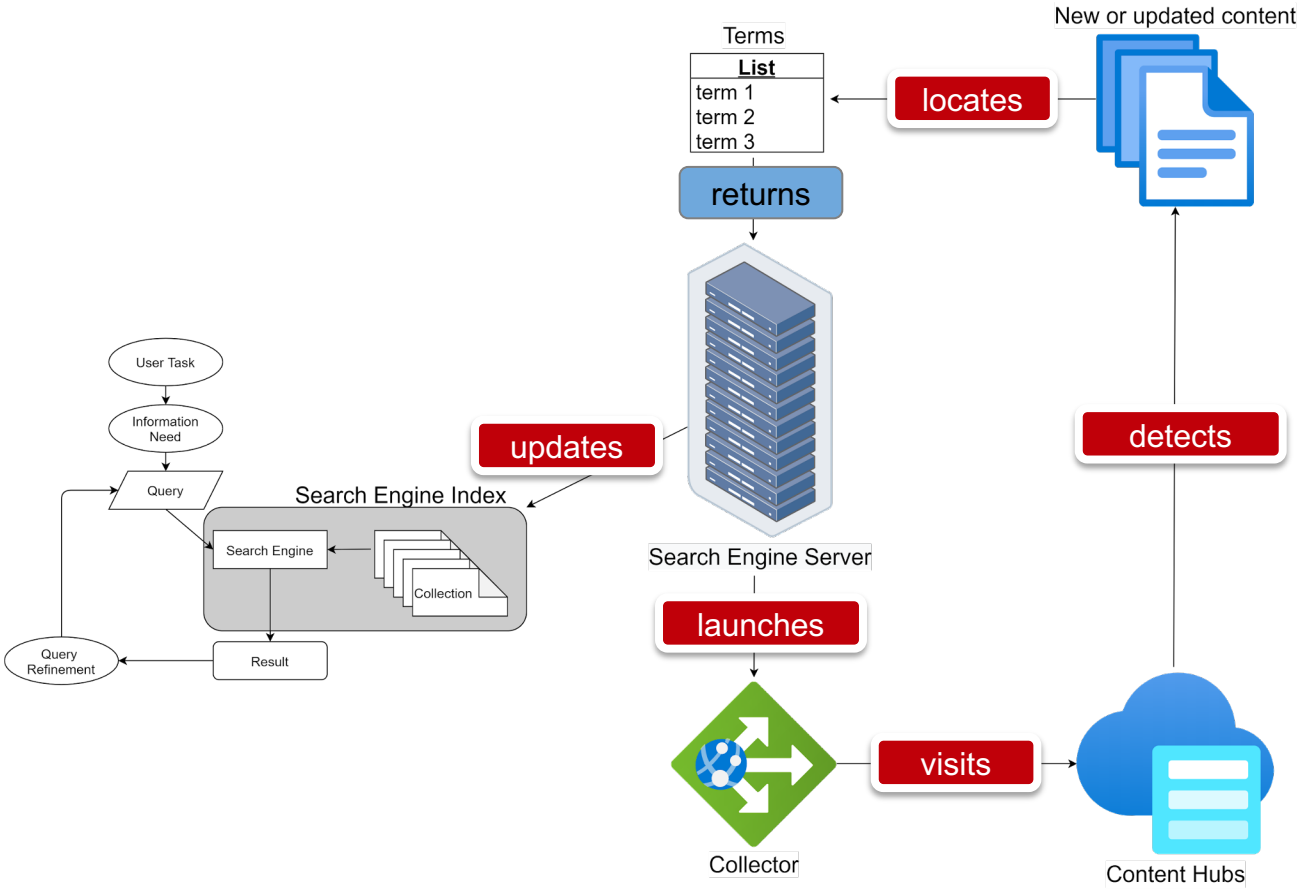
A search engine is a software program designed to identify and respond to specific questions, called keywords, and populate the result page with relevant information available in the collection!



How search engines work



How search engines work – Details



Types of search engines

Web search

- not only content but reliability in the result is needed
- no power over the content in the collection

E-commerce search

- power over the content in the collection and trustworthiness not needed
- But search is the sales person, it must sell according to specific aspects, e.g., get the old items out while making as much profit as possible

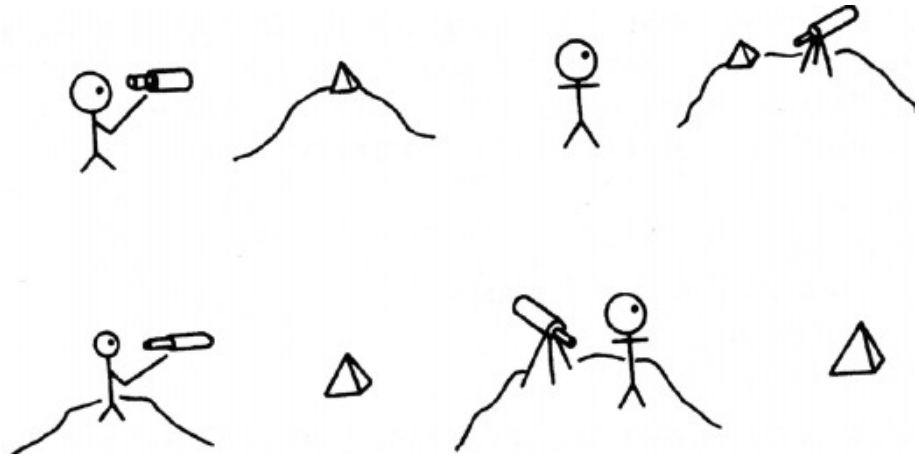
Expert search

- medicine, law, and research fields dig deeper into text for its definition of relevance
- understanding the jargon entered by specialists is key. The domain-specific relationships are key, e.g. “Heart Attack” is the same thing as “Myocardial Infarction”, or “Acute Myocardial Infarction” is a specific type of “Heart Attack”

So, which content is relevant now?

Remember Information Retrieval (IR)?

- what is relevancy?
- many interpretations possible
- which of the interpretations is the correct one?
- can there be a single correct one?



"The man saw the pyramid on the hill with a telescope"

Searching relevant content

Relevance is the art of **ranking content** for a search based on how much that content **satisfies the needs of the user** and the business.

Relevance – the devil is in the detail

Ranking search results for what content?

- Tweets? Products? People?

For what sorts of users?

- Doctors? Tech-savvy shoppers? Regular shoppers? Age groups?

For what types of searches?

- Written in Japanese? Full of brands? Filled with legal jargon?

What do those users expect?

- A shopping experience? A library card catalog?

What does your employer hope to get out of this interaction?

- Money? Page views? Goodwill?

All this needs the skills of a **relevance engineer**

What are relevance engineers?

Transforms the search engine into a **seemingly smart system** that understands the needs of users and the business.

- e.g. in e-commerce the search engine becomes the sales person

Teaches the search engine the content's **important features**:

- attributes like a restaurant's location, the words in a some text, or the color of a dress shirt

Measures what matters to users when they search:

- how far is the restaurant from me? Is this book about the topic I need help with? Will this shirt match the pants I just bought?

This job is real…!

Search Engine Evaluator

This position is restricted to current residents of Germany.

This is a Personalized Search Engine Evaluator position. As a Personalized Search Engine Evaluator, you will be given tasks that are generated from your personalized content based on your Google account linked to your Gmail address that you use to register with Leapforce. Ideal candidates will be highly active users of Google's search

Ideal Search Engine Evaluators will possess the following skills

Have in-depth, up-to-date familiarity with German social culture, media, and web culture
Excellent comprehension and written communication skills in German and English
Broad range of interests, with specific areas of expertise a plus
University degree or equivalent experience (degrees in-progress are acceptable). Advanced degrees a plus
Excellent web research skills and analytical abilities.
Ability to work independently with minimal supervision
Possess a high speed internet connection (DSL, Cable Modem, etc.)
Search Engine Evaluators are required to have currently lived in Germany for a minimum of 5 consecutive years to ensure cultural familiarity.
Use of an Android phone version 4.1 or higher, Windows phone version 8.1 or higher, or an iPhone version 4s or higher

Search Engine Evaluators provide feedback on search engine results by measuring the relevance and usefulness of web pages in correlation to predefined queries, by providing comparative analysis of sets of search engine results and various other techniques.

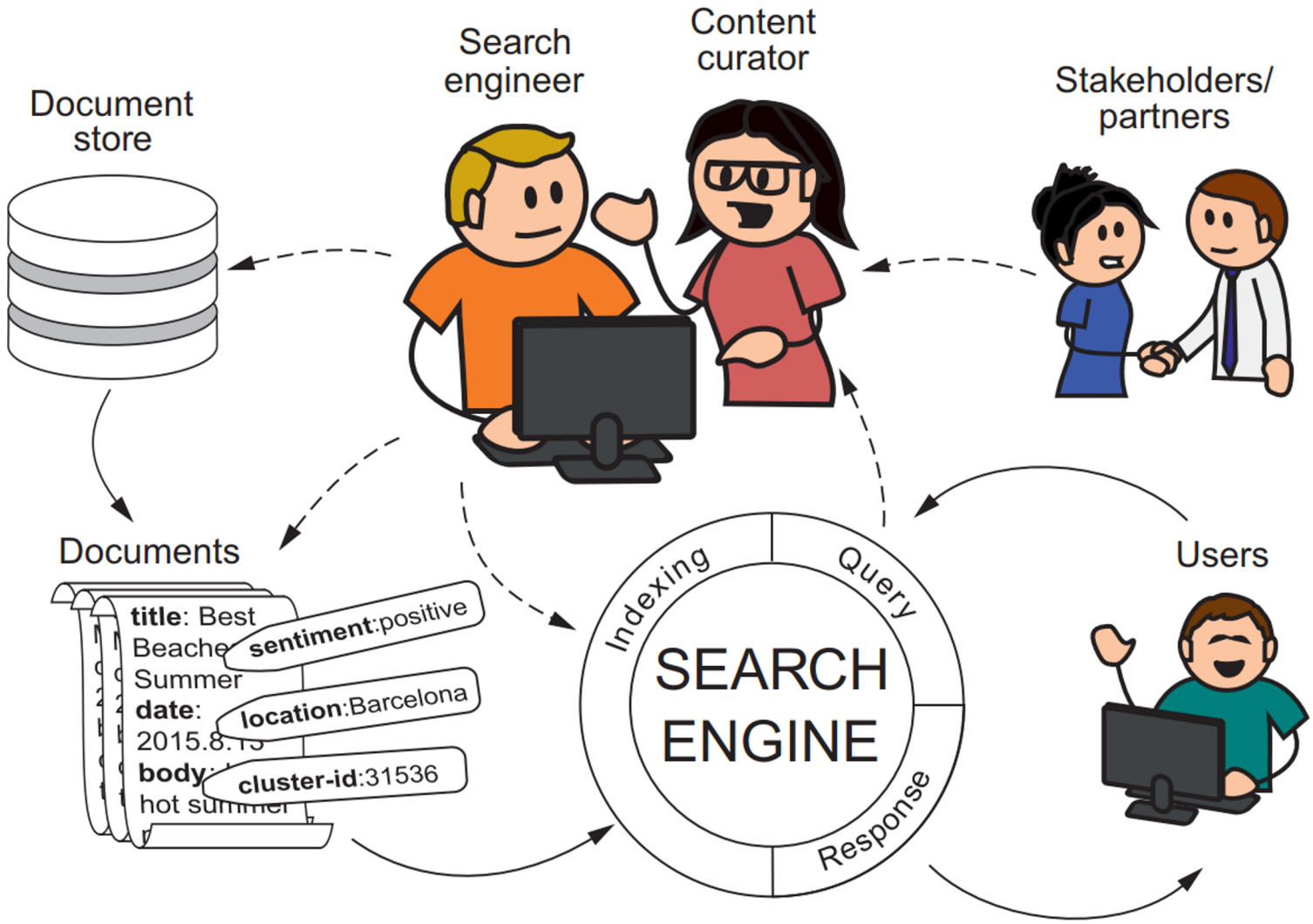
What is this job about?

What users care about, is specified in so-called **signals**

The ever-present challenge of a relevance engineer:

- **implement signals** that map to the needs of your users and business (e.g. Precision and Recall, Click-Through-Rate, Downloads, Sales, etc.)
- **select features** that you think will cause “good” signals (e.g. number of visits (trends), number of citations (importance), etc.)
- redo these steps... **over and over again**!

Implementing a search engine

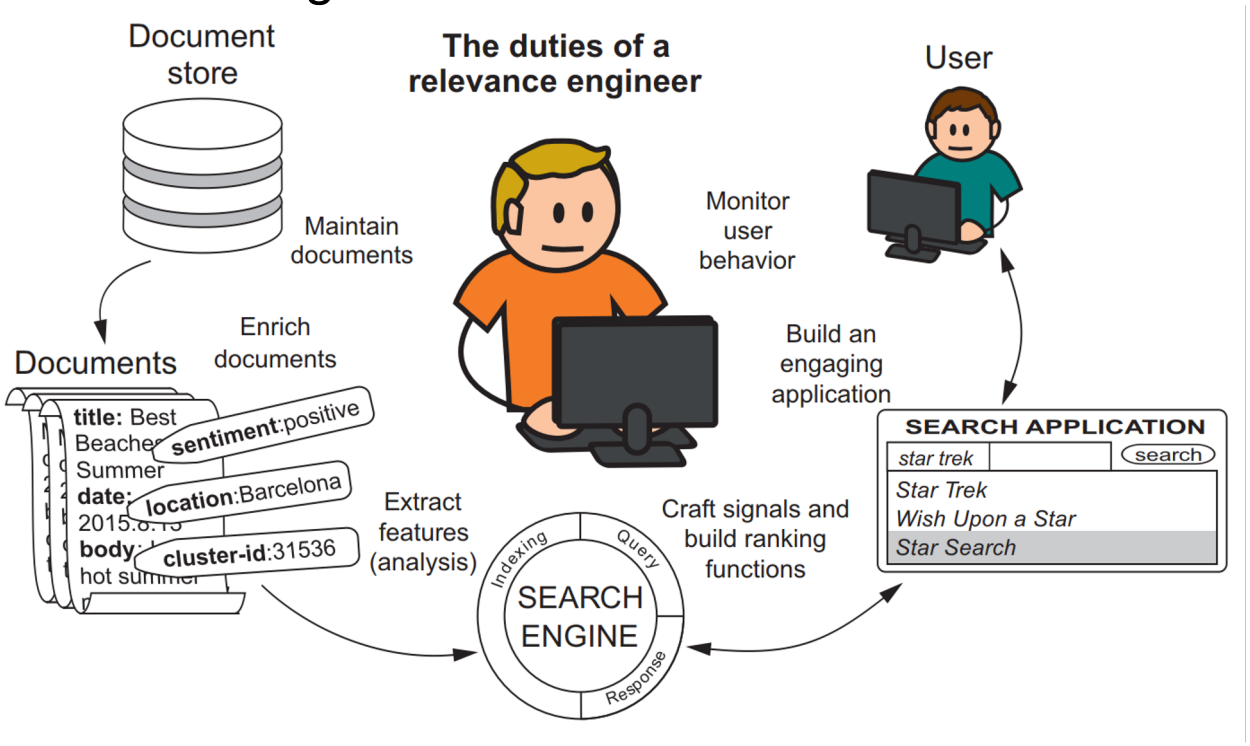


Picture from Turnball at al., Relevant Search - With applications in Solr and ElasticSearch, Manning, 2016

Tasks of a relevance engineer

To solve the relevance problem, we

- identify salient features about (content, user, queries)
- find ways to tell implement these feature into the engine
- measure what's relevant by crafting signals
- balance different signals to better re-rank the results



Picture from Turnball at al., Relevant Search - With applications in Solr and ElasticSearch, Manning, 2016

Ranking the relevant items to the top

Remember Ranking?

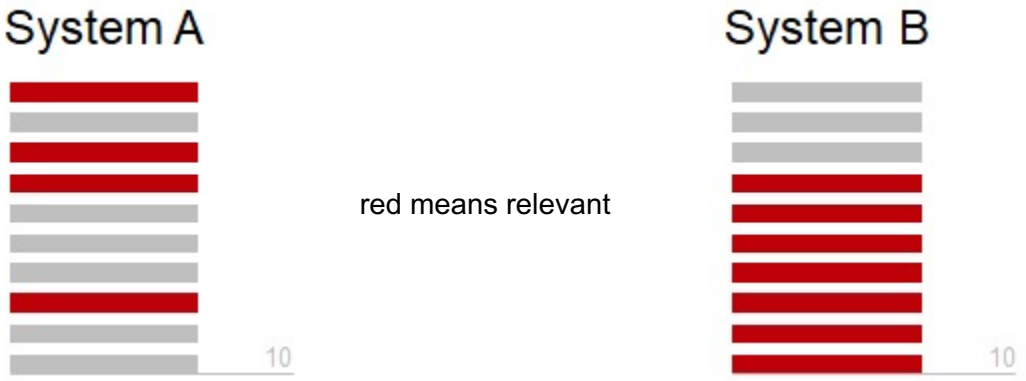
- We want to rank the most relevant items to the top

However

- Which ranking it “better” now?
- What does “better” even mean here?

For Example

- Recall vs.
- Precision@5



Recall: System B is better
Precision@5: System A is better

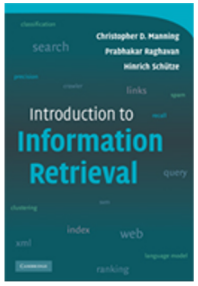
How to do search relevance

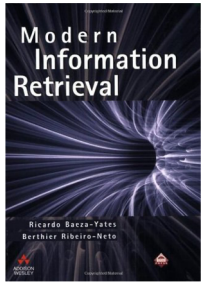
- Search relevance has not any holistic grounding or some common engineering principles
- It is instead a bag of tricks, heuristics and a lot trial and error that can’t be generally applied.
- However, there are some techniques to support this process!
- And in this lecture, we want to explore and apply these techniques, as it is...

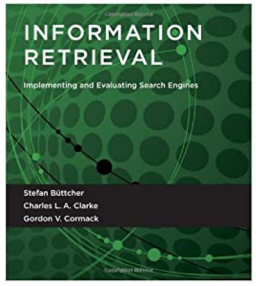
...your job is to solve relevance for your application

References

- **Introduction to Information Retrieval** by Manning et al.: http://nlp.stanford.edu/IR-book/
- **Modern Information Retrieval** by Ricardo Baeza-Yates
- **Information Retrieval - Implementing and Evaluating Search Engines** by Stefan Büttcher







- **Introduction to Solr** by Grainger and Potter
- **Relevant Search** by Turnbull and Berryman