Omar

Ken

CIKM-2023 Tutorial: Some Useful Things to Know When Combining IR and NLP: the Easy, the Hard and the Ugly

Omar Alonso, Amazon, Santa Clara, CA, USA

Kenneth Church, Northeastern University, San Jose, CA, USA

Disclaimer

The views, opinions, positions, or strategies expressed in this talk are mine (Omar) and do not necessarily reflect the official policy or position of Amazon.

Tutorial Agenda

1. Introduction
2. Hard
3. Easy
4. Medium
5. Ugly
6. Conclusions

https://github.com/kwchurch/CIKM_2023_tutorial

Introduction: Outline

- High level overview of information access (search, QA, etc.)
- What are the components that we need
- IR stack (indexing, crawling, ranking, query understanding)
- NLP stack (POS, NER, specific tasks, embeddings, transformers, etc.)
- HEMU dimensions
  - Hard
  - Easy
  - Medium
  - Ugly

Information seeking

- The user has an information need
- Expressed as a query (or question)
- Examine search results or answer(s)
- Re-formulate if needed
- Systems
  - Information retrieval/Search engine,
  - Q&A systems
  - Chatbots
  - ChatGPT
  - Forums
  - Social networks

Information needs and queries

- Relevance to what?
- In Web
    - Two queries of two terms …
    - … looking each time to at most two pages and doing two clicks per page
    - Not a lot of data to guess correctly
- Relevance to the query
    - Problematic
    - Short queries
- Information need
- User intent

Overview of information needs

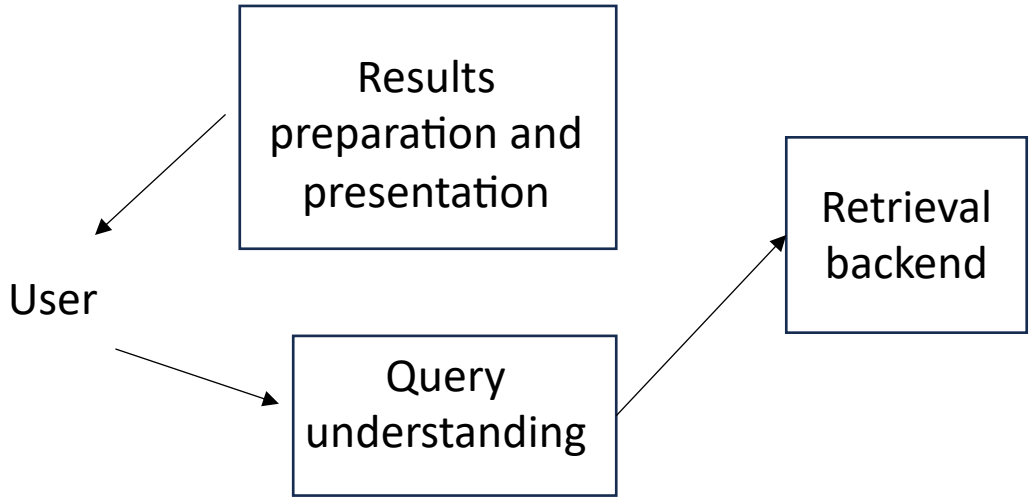
- Keyword queries
  - Free text queries
- Structured queries
  - SQL, SPARQL
- Keyword++
  - Queries with filters or facets
- Natural language
  - Natural language queries, questions
- Zero queries
  - You are the query

How do we know if users are happy?

- Search returns relevant results to users
  - How do you assess this at scale?
- Search results get clicked a lot
  - Misleading titles/summaries can cause users to click
- Users buy after using the search engine
  - Users spend a lot of $ after using the search engine
- Repeat visitors/buyers
  - Do users leave soon after searching?
  - Do they come back within a week/month/… ?

Information seeking architectures

- More or less the same
- Query understanding
  - Normalization
  - Spelling correction
  - Segmentation
  - Annotation (NER, POS, etc.)
  - Term expansion
  - Query-rewriting
- Ranking
  - Many models to chose from
- Answer generation/snippets
- SERP construction
- 10-blue links



This tutorial

- Lots of technology changes recently
- Are things really new or just another iteration?
- It is very easy to prototype new solutions
- Have we solved all problems?
- IR, NLP, ML are converging
  - How can we combine all these new tech to solve new problems?
  - Where should we focus on?

HEMU dimensions

- Computing resources
  - laptop < cloud access < cluster (1K machines is a typical industry cluster)
- Data
  - Small, medium, big
  - Public, sensitive
- Algorithmic complexity
  - Hashing/indexing, PageRank, Deep Learning
- Skills
  - Some things require the worlds' expert, and other things can be done by a software engineer
  - Things can be done by a non-programmer

HEMU dimensions - II

- Size of team
  - You
  - Pizza team
  - The number of authors per paper has been growing suggesting that you need more and more people to do certain things
- Cost
  - Includes all of the above
  - As well as externalities such as power
- Value
  - Some things are super expensive and not worth doing
  - Think about time scales... some things are good for years/decades, and other things age quickly
  - How easy is it for the competition to do a fast follow?
- Organization
  - Is your team/company ready for your idea?
  - Immediate/ medium/long-term success