

Data-analysis and Retrieval

Introduction

Hans Philippi

April 24, 2024

- ① Part 1: Hans Philippi
- ② Part 2: Mihaela Mitici
- ③ Lab assistants: Baharak Ibrahimy (staff), Liam Thomassen (TA), Thomas vd Plas (TA)

What? part 1 (Hans)

This is not primarily a course on Information Retrieval (IR), but we are interested in:

- Processing techniques and data structures for IR queries
- Dealing with large scale unstructured/textual data: the web, libraries, scientific literature, DNA, ...
- Dealing with a NoSQL technique suited for high volume parallel computations (MapReduce)
- Ranking (classical, Google PageRank) and application of ranking to the many-answers / zero-answers problem, when querying databases
- Relation between ranking and top-k query processing
- ...

What? part 2 (Mihaela)

- Clustering: Given a set of docs, group them into clusters based on their contents
- Classification: Given a set of topics, and a new doc D , decide which topic(s) D belongs to
- Learning ranking: Can we learn how to best order a set of documents, e.g., a set of search results, based on user feedback?



application:

accounting, production

libraries, www

data type:

numbers, short strings

text

foundation:

algebra, logic

probabilistic

search paradigm:

Boolean, exact

keywords, vague, ranking

market leaders:

Oracle, IBM, ...

Google, Yahoo! ...

- DB: classical types
 - int, char, float, date, money
 - limited support for strings
- IR: text
 - granularity issues: chapters, paragraphs
- In between: XML (semi structured)

DB: theory of sets/bags

- query languages: based on logic/algebra
- queries are exact
- result is a table
- systematic query processing and generic optimization
- established paradigm; has survived several trends

IR: text, limited or no structure

- queries are vague: sets of terms
- result: basic data structure is the ordered list of document references
- quality of matching: *ranking* makes the difference
- data is vague: stemming, homonyms, synonyms, spelling variations, spelling errors, interpunction, stop words, languages, alphabet (Latin, Greek, Cyrillic, Arabic, Chinese)

Focus topic (lab exercise 1): ranking for DB queries



...

Focus topic (lab exercise 1): ranking for DB queries



- *Search for apartments in Barcelona: sleeping place for at least 4 persons, close to the city centre and close to restaurants where you can eat for 20 euros, preferably with a view at the sea; price limited to 1000 euro a week, but preferably cheaper ...*
- *... and, if possible, equipped with a dishwasher!*

Characteristics of ranked database query

- conjunction referring to many attributes
- score per attribute instead of true/false

Problems

- zero answers (or too little)
- many answers

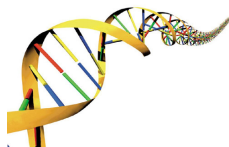
Approach

- apply concepts from IR to ranked database queries

Case study: Google Pagerank

- Web user submits a query defined by a number of keywords
- *Question:* how to determine the most relevant 10 / 20 / 30
- *Question:* how to prevent spamming

Case study: k-grams for DNA matching



- DNA data are long text strings over a limited alphabet:
- GGAGAAGACCAAGGAGGCCCTACTGGAAAAGGCCATGCT...
- biologists want to find *homologies*
- approximate string matching can be solved by dynamic programming
- often too slow: BLAST heuristic based on k-grams

Organizational issues for 2023

- Werkcollege MapReduce on Friday April 26
- All communication regarding the labs via Teams
- Deadline design lab P1 on Wednesday, May 10
- For P1, you should have studied the material of sessions 3 and 4 thoroughly
- Submission of P1 on Friday, May 24
- Exam 1: Wednesday May 22, 16:00 - 18:00, Educ beta
- Retake exam 1: Wednesday June 26 (classroom hours, in classroom)

- Literature: online books and articles
- Final grade:
 - $E = (T1 + P1 + T2 + P2)/4$
 - $P1 \geq 5.0 \wedge P2 \geq 5.0$
 - $T1 \geq 5.5 \wedge T2 \geq 5.5$