

INFORMATION RETRIEVAL

L13. DOMAIN-SPECIFIC IR

1

SUZAN VERBERNE 2022

TODAY'S LECTURE

- Domain-specific IR
 - Tasks and domains
 - Domain-specific test collections
 - Development of domain-specific search engines
 - Query-by-document retrieval
 - TREC Clinical Trial Track
- Your questions about the course materials

IR IN SPECIFIC DOMAINS

PROFESSIONAL SEARCH

- A large portion of information searching takes place in the workplace
- Knowledge workers:
 - lawyers, medical professionals, scientists, software developers, ...
 - process large amounts of information
- To efficiently find high quality, relevant, work-related information is a **time consuming** task that is often hampered by the already existing information overload

WEB SEARCH VS PROFESSIONAL SEARCH

- Web search engines rely on the keyword-based search paradigm
 - became popular with the rise of the world wide web in the 1990s
 - the user enters a (short) query and the search system returns a list of documents
- Focus of most IR research on ranking models for open-domain search, not for domain-specific search because
 1. open-domain search is commercially more interesting
 2. for frequent web queries, there is a lot of interaction data (clicks) available for the search engine to learn the relevance of results

WEB SEARCH VS PROFESSIONAL SEARCH

- Professional search is characterized by
 - complex, highly specific search tasks
 - long sessions (many queries) with not only query search but also browsing and link following
 - user-specific and context-specific tasks
 - users who need control over the search process
- Examples:
 - academic literature search
 - systematic reviews by clinical librarians
 - prior art retrieval by patent analysts
 - news monitoring by information specialists

TASKS AND DOMAINS

- Distinguish between
 - tasks executed by experts in the search domain (e.g. a lawyer) and
 - tasks executed by information specialists, who search on behalf of domain experts
- **Information specialists** are typically trained as librarians. They often work in a specialized domain, in particular legal or medical

SEARCH EXPERTS



“What would you like us, as Information Retrieval researchers, to focus our research on?”

“We, information retrieval experts, are the experts in information retrieval. ; Please do not (ever!) compare your Google search with my library degree. E.V.E.R.”



TASKS AND DOMAINS

Variety of search tasks within one topic domain:

- When an academic scholar is looking up a specific paper based on its authors and title, they will perform a **navigational search**
- The same academic scholar will perform an **extensive, high-recall, interactive** search task when they are collecting literature for the background section of a research proposal
 - This task will consist of multiple queries, and potentially even multiple sessions
- Most research on professional search addresses the latter type of tasks, because here the **domain-specificity and user requirements** play a larger role than in single-query, straightforward search tasks

HIGH-RECALL TASKS

- High-recall (or even: full-recall) tasks: no relevant information should be missed
 - Prior art retrieval is the search of relevant **previously published** patents for a new patent application
 - eDiscovery is the task of identifying and collecting **all relevant information** in the case of a law suit or legal investigation
 - Systematic reviewing: collect an **exhaustive** summary of current evidence relevant to a research question
- Strong demand of user **control over the search process**, which makes Boolean queries often the preferred querying method

EXAMPLE OF BOOLEAN QUERY

```
PubMed: (("CRPS" OR "Complex Regional Pain Syndromes"[Mesh] OR
"Complex Regional Pain Syndrome*" ) AND ("Dystonia"[Mesh] OR "Dystonic
Disorders"[Mesh] OR dystonia OR "Dystonic Disorder*" OR "Dystonia
Disorder*" OR "Muscle Dystonia" OR "Paroxysmal Dystonia" OR "Diurnal
Dystonia" OR "Limb Dystonia" OR "Adult-Onset Idiopathic Focal
Dystonia*" OR "Adult-Onset Idiopathic Torsion Dystonia*" OR "Autosomal
Dominant Familial Dystonia*" OR "Autosomal Recessive Familial
Dystonia*" OR "Childhood Onset Dystonia*" OR "Primary Dystonia*" OR
"Secondary Dystonia*" OR "Sporadic Dystonia*" OR "Familial Dystonia*"
OR "Hereditary Dystonia*" OR "Idiopathic Familial Dystonia*" OR "Focal
Dystonia*" OR Pseudodystonia* OR "Psychogenic Dystonia*" OR "Writer's
Cramp" OR "Writer Cramp" OR "Writers Cramp" OR "Adult-Onset
Dystonia*")) AND ("Serial casting" OR "plaster splint" OR "rigid
splint" OR "Casts, Surgical"[Mesh] OR "Surgical Cast*" OR "Plastic
Cast*" OR "Plaster Cast*" OR "Fiberglass Cast*" OR "Calcium
Sulfate"[Mesh] OR "Calcium Sulphate" OR "Plaster of Paris" OR
"Splints"[Mesh] OR Splint OR "Static Splint*" OR "Static Orthose*" OR
"Static Splinting") no filters
```

PROMINENT DOMAINS

ACADEMIC, LEGAL, MEDICAL

ACADEMIC

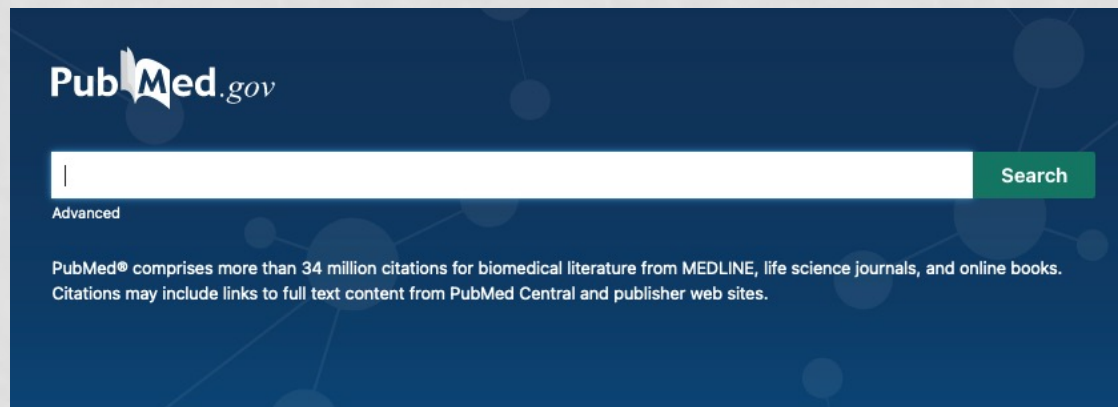
- The study of search behaviour of academic scholars has a long tradition
 - the search for academic literature was a common IR task long before the rise of the world wide web
- Traditionally, academic search heavily relied on **bibliographic references**
 - This is still important, also for document ranking. In the academic context, citations are a sign of relevance

LEGAL

- Legal IR is the search performed by legal scholars and practitioners
 - Legislation and case law retrieval by lawyers and jurists
 - Patent prior art retrieval
 - eDiscovery tasks by legal professionals (see <https://www.zylab.com/>)
- These legal search tasks are all very different from each other
- The documents all have a legal purpose, but are searched using very different tools and in very different databases

MEDICAL

- The medical domain is a very prominent professional domain, involving both scientists and practitioners



- PubMed
 - > 30 million citations and abstracts dating back to 1966 (and some even before 1900)
 - 21.5 million records have links to full-text versions
 - 1 million new records per year

DOMAIN-SPECIFIC TEST COLLECTIONS

COMMONLY USED BENCHMARK SETS

➤ Medical:

- The CLEF e-Health track [Kelly et al. 2019, Suominen et al. 2013]
- [TREC-COVID](#) [Voorhees et al. 2021]
- The PICO collection of systematic reviews [Scells et al. 2017]
- The [TREC clinical trials track](#) [2021]

➤ Legal:

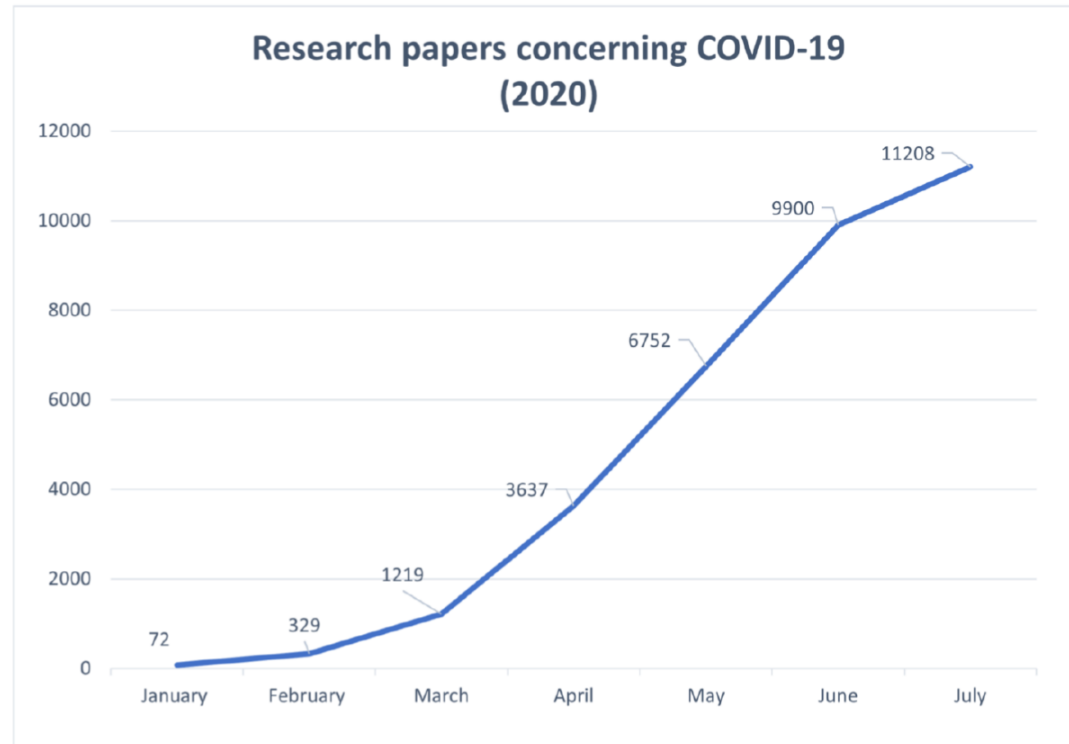
- The COLIEE shared task on legal case retrieval [Kano et al. 2018]
- FIRE AILA [Bhattacharya et al. 2019]
- The TREC total recall track [Grossman et al. 2016]
- CLEF-IP: IR in the intellectual property domain [Piroi et al. 2011]

➤ Academic

- The iSearch of academic search tasks [Lykke et al. 2010]

TREC-COVID

- In order to enhance research on IR systems to support researchers keeping track of new published trials
- Dataset:
 - Updated research corpus of Covid19 related studies
 - Updated queries about Covid19



<https://ir.nist.gov/trec-covid/>

LIMITATIONS

- For some domains there are no freely available test collections or they are very limited
- Collections often have only one type of documents instead of the complete variety that is **representative of the problem**, e.g.
 - E-discovery. Commonly used benchmark: the Enron email dataset even though it is old and limited in nature
 - Clinical benchmarks that only contain discharge summaries and not all notes and letters in an electronic health record
- Collections often have too **shallow relevance assessments**

LIMITATIONS

The problem of shallow relevance assessments

- **Pool** creation:
 - Get the top-k documents for each query from an existing ranker or multiple baseline rankers
 - Then collect human relevance labels for this pool
- The collected assessments are **biased** toward the pool, and thereby towards the initial rankers
- If we run completely different ranking models, we retrieve documents **without relevance assessment**
 - This is more severe if the new rankers are more different from the rankers in the pool (think BM25 vs BERT-based rankers)

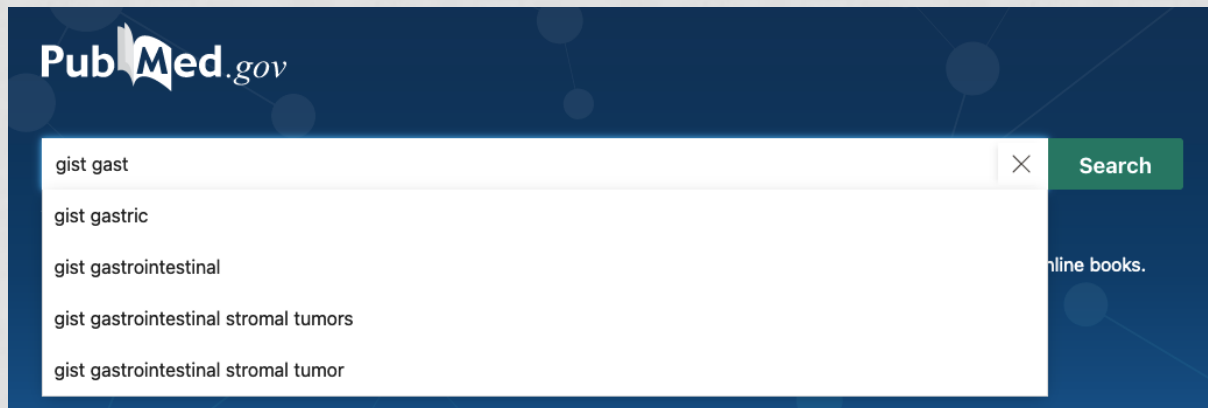
DEVELOPMENT OF DOMAIN-SPECIFIC SEARCH ENGINES

BUILD ON COMMON BACK-ENDS

- Each specific target group has – at least partly – different requirements
- Does that imply that we need to build a fully unique search engine for each professional target group to accommodate their needs?
- No, luckily not. Professional search systems are often built on top of common back-ends, such as [Apache Solr](#) or [ElasticSearch](#)

QUERY INTERFACES

1. Boolean search
2. Controlled vocabulary: queries are matched to a domain ontology and users are helped by suggesting terms from the ontology



3. Faceted search: use filters to refine the results

Save

Email

Send to

Sorted by: Best match

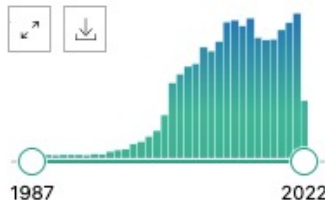
Display options ⚙

MY NCBI FILTERS

5,876 results

<< < Page 1 of 588 > >>

RESULTS BY YEAR



TEXT AVAILABILITY

- ☐ Abstract
- ☐ Free full text
- ☐ Full text

ARTICLE ATTRIBUTE

- ☐ Associated data

ARTICLE TYPE

- ☐ Books and Documents
- ☐ Clinical Trial
- ☐ Meta-Analysis
- ☐ Randomized Controlled Trial
- ☐ Review
- ☐ Systematic Review

PUBLICATION DATE

- ☐ 1 year
- ☐ 5 years
- ☐ 10 years

☐ Current clinical management of gastrointestinal stromal tumor.

1 Akahoshi K, Oya M, Koga T, Shiratsuchi Y.

Cite World J Gastroenterol. 2018 Jul 14;24(26):2806-2817. doi: 10.3748/wjg.v24.i26.2806.

PMID: 30018476 Free PMC article. Review.

Share **Gastrointestinal stromal tumors** (GISTs) are the most common malignant subepithelial lesions (SELs) of the **gastrointestinal** tract. They originate from the interstitial cells of Cajal located within the muscle layer and are characterized by over-expressi ...

☐ Gastrointestinal stromal tumours.

2 Blay JY, Kang YK, Nishida T, von Mehren M.

Cite Nat Rev Dis Primers. 2021 Mar 18;7(1):22. doi: 10.1038/s41572-021-00254-5.

PMID: 33737510 Review.

Share **Gastrointestinal stromal** tumours (**GIST**) have an incidence of ~1.2 per 10(5) individuals per year in most countries. Around 80% of **GIST** have varying molecular changes, predominantly mutually exclusive activating KIT or PDGFRA mutations, but other, rare ...

☐ Gastrointestinal stromal tumor: epidemiology, diagnosis, and treatment.

3 Mantese G.

Cite Curr Opin Gastroenterol. 2019 Nov;35(6):555-559. doi: 10.1097/MOG.0000000000000584.

PMID: 31577561 Review.

Share **PURPOSE OF REVIEW:** The purpose of this review is to review the past year's literature to provide comprehensive information to researchers, physicians, and the general public regarding the epidemiology, diagnosis, and treatment of **gastrointestinal stromal tumors** ...

☐ What is New in Gastrointestinal Stromal Tumor?

4 Schaefer IM, Mariño-Enríquez A, Fletcher JA.

Cite Adv Anat Pathol. 2017 Sep;24(5):259-267. doi: 10.1097/PAP.0000000000000158.

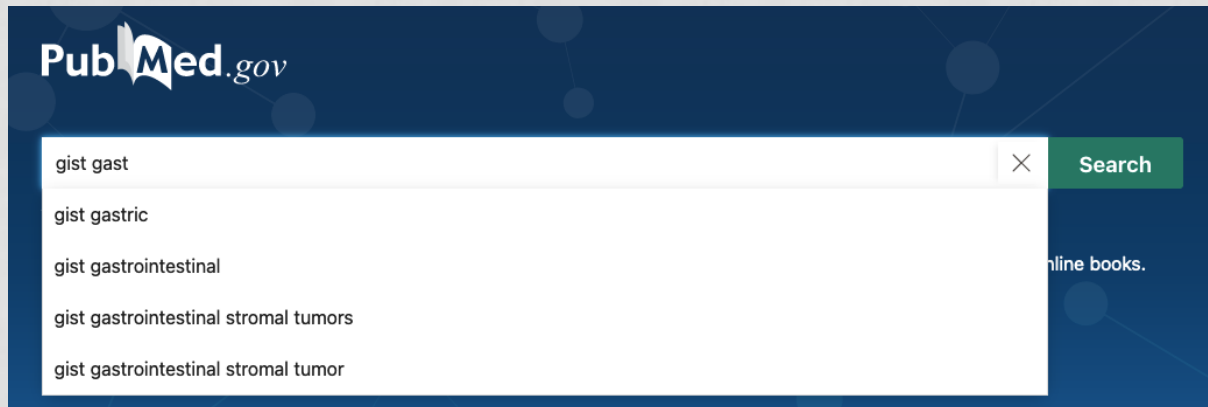
PMID: 28632504 Free PMC article. Review.

Share The classification "**gastrointestinal stromal tumor**" (**GIST**) became commonplace in the 1990s and since that time various advances have characterized the **GIST** lineage of origin, tyrosine kinase mutations, and mechanisms of response and resistance to target ...

☐ [Gastrointestinal stromal tumors (GIST)--literature review].

QUERY INTERFACES


1. Boolean search
2. Controlled vocabulary: queries are matched to a domain ontology and users are helped by suggesting terms from the ontology





3. Faceted search: use filters to refine the results
4. Fielded search

QUERY INTERFACES


- Fielded search:
specify which
document field
should contain the
query term(s)


PubMed Advanced Search Builder  [User Guide](#)

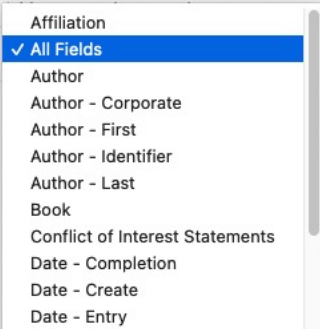

Add terms to the query box


All Fields  Enter a search term  [Show Index](#)

Query box

Enter / edit your search query here 

PubMed Advanced Search Builder  [User Guide](#)

 Enter a search term  [Show Index](#)



QUERY INTERFACES

➤ Specialized advanced query builders

AGNES version 0.2

Search through **Boolean operators** Results from **Entity types** [Help](#)

AND OR [+ Add rule](#) [+ Add group](#)

Artefact begins with afslag [X Delete](#)

AND OR [+ Add rule](#) [X Delete](#)

Time Period is neolithicum [X Delete](#)

Time Period is bronsjtd [X Delete](#)

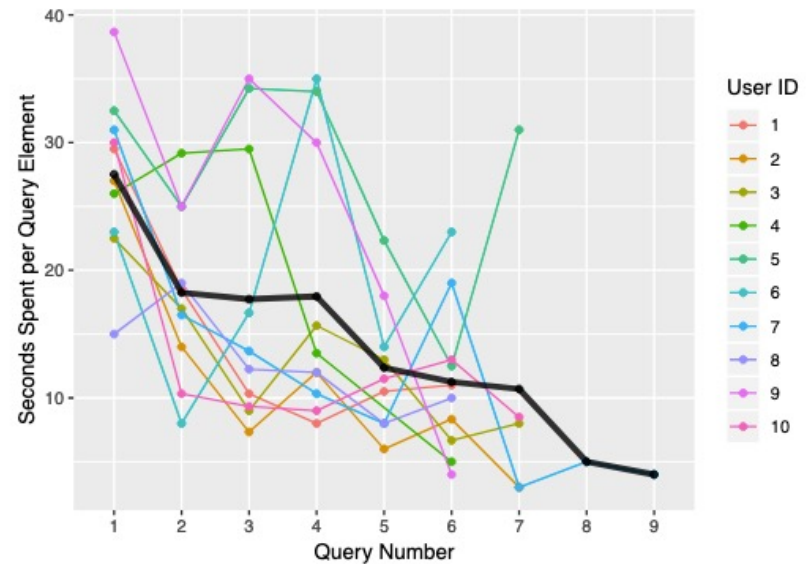
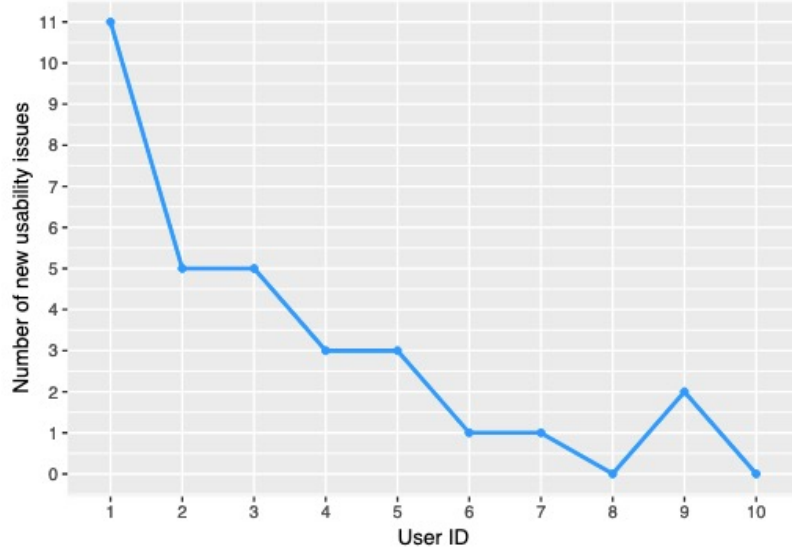
[Search](#)

USER OBSERVATION STUDIES

- Chicken-and-egg problem
 - The **requirements** for a search engine depend on the behaviour of the target group
 - But the **behaviour** of the searchers is influenced by the search system and interface
- Solution:
 1. Collect requirements
 2. Develop an initial version of the search system
 3. Observe users while they interact with the system (log data)
 4. Use a questionnaire to collect feedback for improving the system

USABILITY EVALUATION

- A small focus group (10 expert users)



QUERY-BY-DOCUMENT RETRIEVAL

QBD TASKS

Example QBD tasks:

- ‘Find similar documents’ (query-by-example)
- Prior art retrieval: given a patent application document, find similar existing patents
- Case law retrieval: given a legal case document, find similar prior cases

The screenshot displays the ColBERT search interface. At the top, a search bar contains the text 'colbert'. Below the search bar, there are filters for 'Any time', 'Sources', 'Code', 'Countries', 'Organizations', and 'Owner'. The results section shows '806 results'. The first result is from SIGIR + 1, titled 'ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT', dated 05 May 2020, by Omar Khattab & Matei Zaharia. The abstract states: 'ColBERT introduces a late interaction architecture that independently encodes the query and the document using BERT and then employs a cheap yet powerful interaction step that models ... more'. It has 119 citations and 0 PDFs. The second result is from arXiv + 1, titled 'A White Box Analysis of ColBERT', dated 17 Dec 2020, by Thibault Formal, Benjamin Piwowarski & Stéphane Clinchant. The abstract states: 'However, we propose to dissect the matching process of ColBERT, through the analysis of term importance and exact/soft matching patterns. ... Even if the traditional axioms are not formally v ... more'. It has 6 citations and 15 PDFs. Both results have a 'Find similar' button and a 'Notes' button.

CHALLENGES

- Long documents in the collection
 - One document may contain many topics
- Long queries:
 - E.g. in legal case retrieval (COLIEE), avg query length is 1269 words
 - How to process and interpret long queries?
 - Computational challenges of encoding
- Domain-specific language
- Lack of labelled data

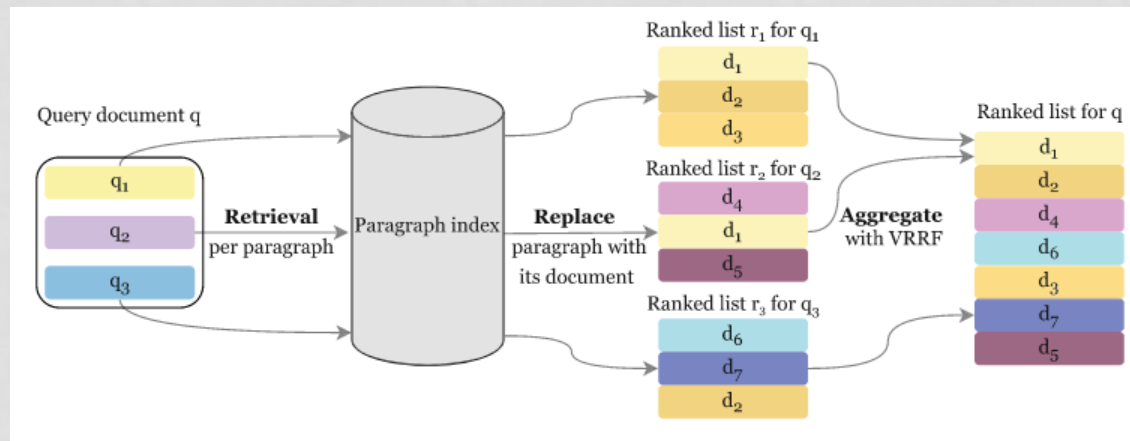
DEALING WITH LONG QUERIES

Options to deal with (long) documents as queries

- Use word-based methods (**BM25**) on the full text
- **Extract query terms** and the use regular retrieval models
 - Identify the most relevant key terms
- **Truncate** all documents or use only the abstract
 - Common: only use the first 512 tokens
- Automatic summarization
- Use paragraph-level retrieval and aggregation

DEALING WITH LONG QUERIES

- **Paragraph Aggregation** model for dense document-to-document retrieval (PARM)
- For each query paragraph, relevant documents are retrieved based on their paragraphs
- The relevant results per query paragraph are aggregated into one ranked list for the whole query document



Althammer et al (2022). PARM: A Paragraph Aggregation Retrieval Model for Dense Document-to-Document Retrieval. In: Advances in Information Retrieval.

DOMAIN-SPECIFIC LANGUAGE

Neural retrieval models require encoders ([BERT models](#))

- Medical
 - BioBERT: trained on biomedical study abstracts
 - PubMedBERT: trained on clinical abstracts from PubMed
- Legal
 - LegalBERT
 - PatentBERT
- Scientific
 - SciBERT: trained on scientific and biomedical abstracts of Semantic Scholar
- Other... ArchaeoBERTje

Models are either trained from scratch on domain data or further pre-trained from a general-domain model

LACK OF LABELLED DATA

- Test collections for domain-specific tasks are often limited (have shallow relevance assessments)
 - Alternative: learn relevance labels from click logs
 - TripClick: Large corpus of medical studies (title and abstracts), relevance annotations based on click data of users
- Additional challenge: high recall settings
 - In tasks such as systematic review or prior art search missing a relevant document can be catastrophic

TREC CLINICAL TRIALS TRACK

FINAL ASSIGNMENT

TASK MOTIVATION AND DEFINITION

- The majority of clinical trials fail to recruit the required amount of patients
- Solution: use electronic health records (EHRs)
- Task of the TREC Clinical Trials track “flips the trial-to-patients paradigm to a patient-to-trials paradigm”
 - the [query/topic is a patient description](#) (QBD retrieval)
 - the corpus is a large set of clinical trial descriptions

<http://www.trec-cds.org/2021.html>

EVALUATION

- Graded relevance
 - 2: eligible (patient meets inclusion criteria and exclusion criteria do not apply)
 - 1: excluded (patient meets inclusion criteria, but is excluded on the grounds of the trial's exclusion criteria)
 - 0: not relevant

Evaluation

The evaluation will follow standard TREC evaluation procedures for ad hoc retrieval tasks. Participants may submit a maximum of **five automatic or manual runs**, each consisting of a ranked list of up to one thousand IDs (**NCT IDs provided by ClinicalTrials.gov**). The highest ranked results for each topic will be pooled and judged by physicians trained in medical informatics. Assessors will be instructed to judge trials as either *eligible* (patient meets inclusion criteria and exclusion criteria do not apply), *excluded* (patient meets inclusion criteria, but is excluded on the grounds of the trial's exclusion criteria), or *not relevant*. Because we plan to use a graded relevance scale, the performance of the retrieval submissions will be measured using normalized discounted cumulative gain (NDCG).

As in past evaluations of medically-oriented TREC tracks, we are fortunate to have the assessment conducted by the Department of Medical Informatics of the Oregon Health and Science University (OHSU). We are extremely grateful for their participation.

<http://www.trec-cds.org/2021.html>

EVALUATION

- Official metrics for the task are **NDCG@5** and **NDCG@10**. Use the graded relevance assessment for these metrics.
- Besides the official metrics, you should report precision@10 and Reciprocal Rank
 - For these **binary metrics**, consider both labels '1' and '2' as relevant (so convert 2 to 1)
 - For RR, it is common to look at the highest ranked relevant result
 - Evaluation package: https://github.com/cvangysel/pytrec_eval

YOUR QUESTIONS

SELECTED QUESTIONS ABOUT THE COURSE MATERIALS

- I messed up some of the homework assignments. Can I still pass the course?
- Huffman coding
- Postings lists compression (week 3, exercise 4)
- The free parameters of BM25
- It would be nice if we can have a brief introduction to Elasticsearch

CAN I STILL PASS THE COURSE?

- Q: I messed up some of the homework assignments (the participation ones), been swamped. Can I still pass the course?
- A: yes (see grading on Brightspace)

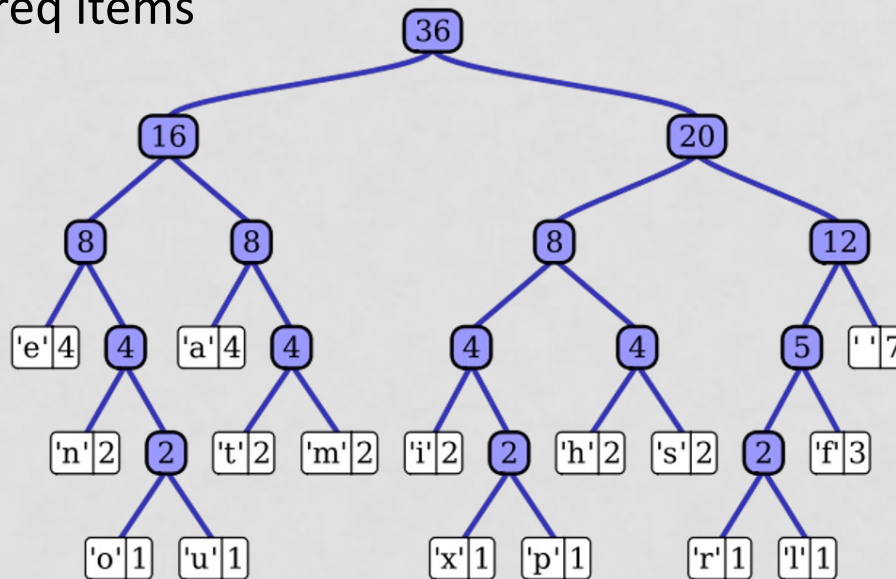
HUFFMAN CODING

- Compression is typically achieved by coding
 - We want lossless compression
 - Frequent items get short codes
- Huffman coding = Example of efficient compression scheme
 - idea: using a frequency-sorted binary tree
 - the most frequent items get the shortest codes (highest in the tree)

HUFFMAN CODING

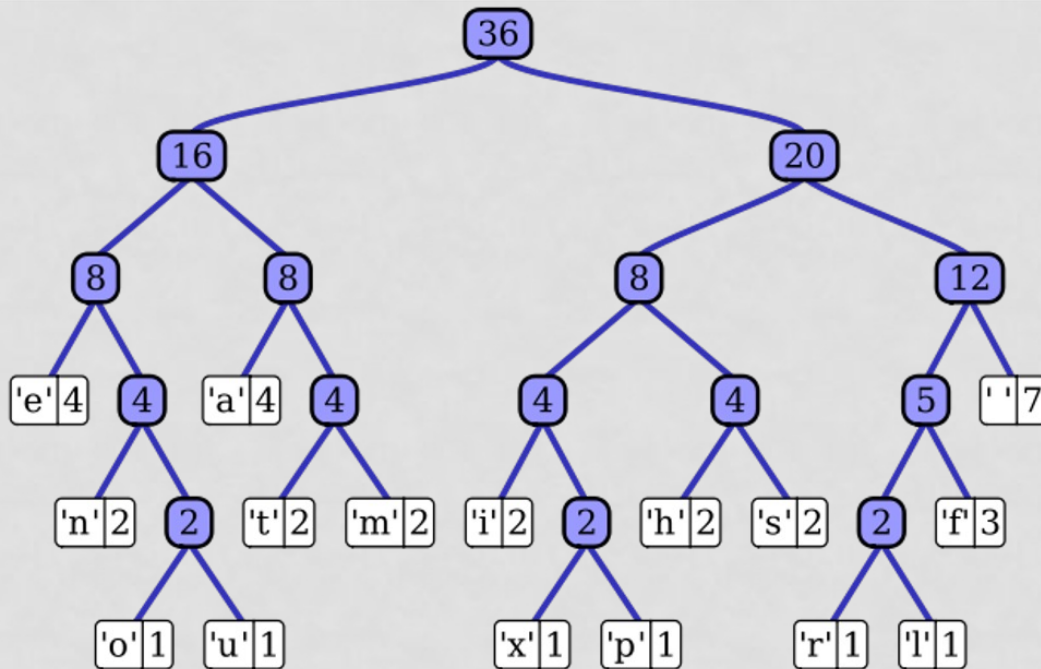
➤ Example: compress the text “this is an example of a huffman tree”

1. Sort the characters of the text by frequency
2. Build the tree bottom-up, starting from the lowest-freq items



Char	Freq
Space	7
a	4
e	4
f	3
h	2
l	2
m	2
n	2
s	2
t	2
l	1
o	1
p	1
r	1
u	1
x	1

HUFFMAN CODING



“character-based Huffman methods can typically compress English texts by about 40 percent, whereas word-based Huffman methods can reduce them by more than 75 percent, because the word distribution is more biased than the character distribution.”

Char	Freq	Code
Space	7	111
a	4	010
e	4	000
f	3	1101
h	2	1010
l	2	1000
m	2	0111
n	2	0010
s	2	1011
t	2	0110
l	1	11001
o	1	00110
p	1	10011
r	1	11000
u	1	00111
x	1	10010

POSTINGS LISTS COMPRESSION

- Coding (compression) of postings lists
 - Posting = docid (with pointer to document)
 - Lists are usually sorted by document number
 - Compression methods for storing the posting lists
- Variable byte (VB) encoding (section 5.3.1):
 - “Variable byte (VB) encoding uses an **integral number of bytes to encode a gap**. The last 7 bits of a byte are “payload” and encode part of the gap. The first bit of the byte is a **continuation bit**. It is set to 1 for the last byte of the encoded gap and to 0 otherwise. To decode a variable byte code, we read a sequence of bytes with continuation bit 0 terminated by a byte with continuation bit 1.”

WEEK 3 – EXERCISE 4

- Consider the postings list (4, 10, 11, 12, 15, 62, 63, 265, 268, 270, 400, 444) with a corresponding **list of gaps** (4, 6, 1, 1, 3, 47, 1, 202, 3, 2, 130, 44). Assume that the length of the postings list is stored separately, so the system knows when a postings list is complete. Using **variable byte encoding**:
- i. What is the largest gap you can encode in 1 byte?
 - ii. What is the largest gap you can encode in 4 bytes?
 - iii. How many bytes will the above postings list require under this encoding? (Count only space for encoding the sequence of numbers.)

WEEK 3 – EXERCISE 4

➤ Solutions Exercise 4

- i. Variable byte encoding reserves one bit for determining continuation of final byte. With **seven bits**, the maximum gap we can encode is $2^7 - 1 = 127$
- ii. Similarly, four bytes of seven bits gives $2^{28} - 1$
- iii. All numbers that are smaller than 2^7 ($=128$) can be stored in 1 byte. The numbers between 128 and 256 need two bytes:
 $10 * 1 + 2 * 2 = 14$ bytes for the gaps, one byte for the first docid. Total 15 bytes

PARAMETERS OF BM25

- “ k_1 and b are free parameters that can be optimized per collection”
- Free parameters: [hyperparameters](#) just like the parameter c in SVM, or the number of trees in RandomForest

$$RSV^{BM25} = \sum_{i \in q} \log \frac{N}{df_i} \cdot \frac{(k_1 + 1)tf_i}{k_1((1 - b) + b \frac{dl}{avdl}) + tf_i}$$

- k_1 controls term frequency scaling (saturation function)
 - $k_1 = 0$ is binary model; k_1 large is raw term frequency
- b controls document length normalization
 - $b = 0$ is no length normalization; $b = 1$ is relative frequency (fully scale by document length)

ELASTICSEARCH

“It would be nice if we can have a brief introduction to Elasticsearch”

- For groups that are struggling with ElasticSearch, we have added step-wise instructions for installation and indexing in [this document](#) under 'Assignments'.

CONCLUSIONS

HOMEWORK

- Work on the final assignment
 - Deadline: next week
- Prepare for the exam
 - Tuesday, June 14, 2022, USC

AFTER THIS LECTURE...

- You can list three characteristics of professional search compared to web search
- You can define high-recall tasks and give an example
- You can explain the problem of shallow relevance assessments
- You can define the following terms: controlled vocabulary, faceted search, fielded search
- You can describe how to set up a user observation study
- You can describe the challenges of query-by-document tasks
- You can list four solutions for dealing with documents as queries
- You know how to complete the TREC clinical track assignment