Research in Biomedical Information Retrieval at OHSU

William Hersh, MD
Professor and Chair
Department of Medical Informatics & Clinical Epidemiology
School of Medicine
Oregon Health & Science University
Email: hersh@ohsu.edu
Web: www.billhersh.info
Blog: http://informaticsprofessor.blogspot.com
Twitter: @williamhersh

References

Buckley, C and Voorhees, EM (2004). Retrieval evaluation with incomplete information. *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Sheffield, England. ACM Press. 25-32.

Harman, DK (2005). The TREC Ad Hoc Experiments. TREC: Experiment and Evaluation in Information Retrieval. E. Voorhees and D. Harman. Cambridge, MA, MIT Press**:** 79-98.

Hersh, WR (2009). Information Retrieval: A Health and Biomedical Perspective (3rd Edition). New York, NY, Springer.

Jarvelin, K and Kekalainen, J (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*. 20: 422-446.

Voorhees, E and Hersh, W (2012). Overview of the TREC 2012 Medical Records Track. *The Twenty-First Text REtrieval Conference Proceedings (TREC 2012)*, Gaithersburg, MD. National Institute of Standards and Technology http://trec.nist.gov/pubs/trec21/papers/MED12OVERVIEW.pdf

Voorhees, EM and Harman, DK, Eds. (2005). TREC: Experiment and Evaluation in Information Retrieval. Cambridge, MA, MIT Press.

Wu, S, Liu, S, et al. (2017). Intra-institutional EHR collections for patient-level information retrieval. *Journal of the American Society for Information Science & Technology*: in press.

Yilmaz, E, Kanoulas, E, et al. (2008). A simple and efficient sampling method for estimating AP and NDCG. *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Singapore. 603-610.

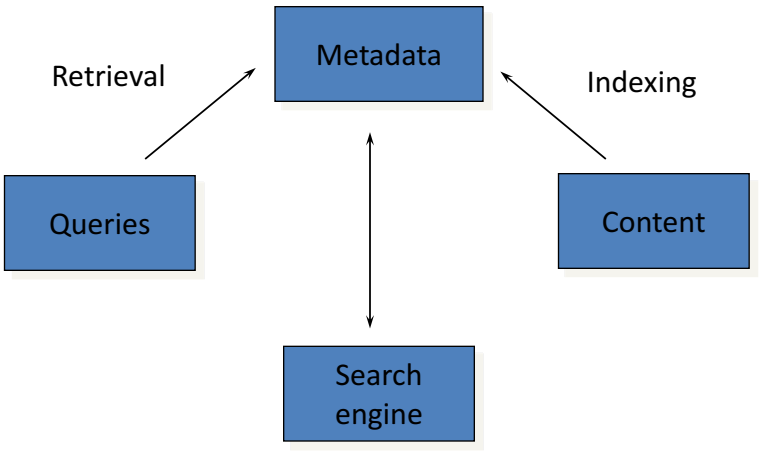
Research in Biomedical Information Retrieval at OHSU

William Hersh
Professor and Chair
Department of Medical Informatics & Clinical Epidemiology
Oregon Health & Science University
Portland, OR, USA
Email: hersh@ohsu.edu
Web: www.billhersh.info
Blog: http://informaticsprofessor.blogspot.com
Twitter: @williamhersh

Outline

- Brief primer of information retrieval (IR) systems and their evaluation
- Focus on IR from electronic health records (EHRs) based on cohort-discovery use case
- Challenges and solutions for large-scale research with private and/or proprietary data

Basics of IR (aka, search) (Hersh, 2009)



3

Use cases for IR

- Historically, retrieval of knowledge
  - Documents, especially journal articles (originally abstracts)
  - Multimedia – images, sounds, video, etc.
  - Hypermedia – Web-based content
- Newer foci
  - Clinical data – e.g., cohort discovery from electronic health records
  - Data – e.g., finding data sets

4

Evaluation of IR systems has always been important

- System-oriented – how well system performs
  - Historically focused on relevance-based measures
    - Recall and precision – proportions of relevant documents retrieved
  - When documents ranked, can combine in a single measure
- Historically assessed with *test collections*, which consist of
  - Content – fixed yet realistic collections of documents, images, etc.
  - Topics – statements of information need that can be fashioned into queries entered into retrieval systems
  - Relevance judgments – by expert humans for which content items should be retrieved for which topics
- User-oriented – how well user performs with system
  - e.g., performing task, user satisfaction, etc.

Recall, precision, and aggregate measures

- Recall $$R = \frac{\#\,retrieved\ and\ relevant\ documents}{\#\,relevant\ documents\ \ in\ collection}$$
- Precision $$P = \frac{\#\,retrieved\ and\ relevant\ documents}{\#\,retrieved\ documents}$$
- Commonly used aggregated measures
  - Mean average precision (MAP) (Harman, 2005)
  - Bpref (Buckley, 2004)
  - Normal discounted cumulative gain (NDCG) (Jarvelin, 2002)
  - MAP and NCDG can be "inferred" when there are incomplete judgments (Yilmaz, 2008)

Many challenge evaluations in IR

- Text Retrieval Conference (TREC, http://trec.nist.gov; Voorhees, 2005) – sponsored by National Institute for Standards and Technology (NIST) since 1992
  - Many "tracks" of interest, such as routing/filtering, Web search, question-answering, etc.
  - Mostly non-biomedical, but some tracks focused on
    - Genomics
    - Medical Records
    - Precision Medicine
- TREC has inspired other challenge evaluations, e.g.,
  - i2b2 NLP Shared Task, https://www.i2b2.org/NLP/
  - bioCADDIE Dataset Retrieval Challenge – https://biocaddie.org/biocaddie-2016-dataset-retrieval-challenge-registration

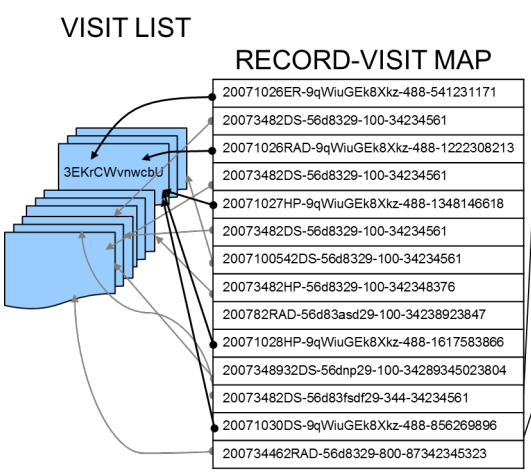
7

TREC Medical Records Track (Voorhees, 2012)

- Use case – retrieve records and data within them to identify patients who might be candidates for clinical studies
- Facilitated with development of large-scale, de-identified data set from University of Pittsburgh Medical Center (UPMC)
- Ran in 2011 and 2012, discontinued due to privacy worries

8

Test collection




```
DISCHARGE SUMMARY
...
PRINCIPAL DIAGNOSES:
1. Urinary tract infection.
2. Gastroenteritis.
3. Dehydration.
4. Hyperglycemia.
5. Diabetes mellitus.
6. Osteoarthritis.
7. History of anemia.
8. History of tobacco use.

HOSPITAL COURSE:  The patient is a **AGE[in 40s]
-year-old insulin-dependent diabetic who
presented with nausea, vomiting, and diarrhea.
She was admitted, placed on IV fluids, a sliding scale,
was found to have evidence of urinary tract infection,
and treated with oral Bactrim.  She was seen by
Endocrinology.   She was started on Lantus, and overall
is feeling better.  She is tolerating a regular diet.
Her sugars have been under better control, and
she is being  discharged to home.  Sodium was 135,
potassium was 4.5, BUN was 21, creatinine was 0.9, and
glucoses recently ranged from 80 to the highest of
219.  Her sugars were as high as 300.  Her white count
was 7.5, hemoglobin was 11, and hematocrit was 33.0.
Urinalysis was positive.
...
```

Report Extract
20071030DS-9qWiuGEk8Xkz-488-856269896

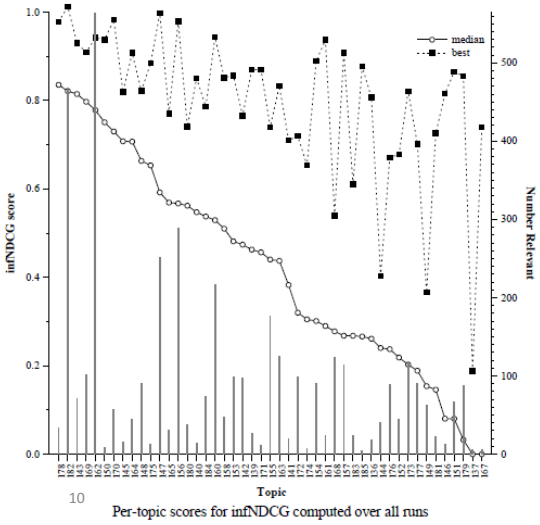
17,265 visits

101,712 reports (93,552 mapped to visits)

(Courtesy, Ellen Voorhees, NIST)

Results for 2012

Run	infNDCG	infAP	P(10)
NLMManual*	0.680	0.366	0.749
udelSUM	0.578	0.286	0.592
sennamed2	0.547	0.275	0.557
ohsuManBool*	0.526	0.250	0.611
atigeo1	0.524	0.224	0.519
UDinfoMed123	0.517	0.236	0.528
uogTrMConQRd	0.509	0.231	0.553
NICTAUBC4	0.487	0.216	0.517



Per-topic scores for infNDCG computed over all runs

Challenges for medical records data

- De-identified data – loss of realism due to
  - Decreased precision of data – e.g., age ranges, geography, etc.
  - Breaking of linkages - e.g., patient's visits, institutions visited, etc.
- Privacy concerns
  - Potential for re-identification of data, especially textual data of most interest to IR

11

Scaling to more realistic data sets (Wu, 2017; R01LM011934)

Methods

Preprocessing & Indexing

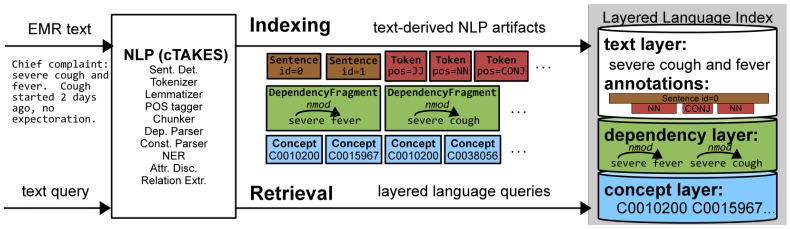
We processed a set of ~100,000 EHR text records using cTAKES (ctakes.apache.org). This clinical system augments the free text "layer" with other linguistically-motivated layers: grammatical dependencies between words, recognizable concepts from a controlled medical vocabulary, etc.

Retrieval & Layers

We started with the intuition that a user will imagine a document that he or she is looking for, and then write a query that matches that document. Defining this probabilistically following the language modeling approach to IR (Ponte and Croft, 1998; Metzler and Croft, 2005), we searched several "layers" separately with the query likelihood model and the Markov Random Field model. The "layers" we evaluated:

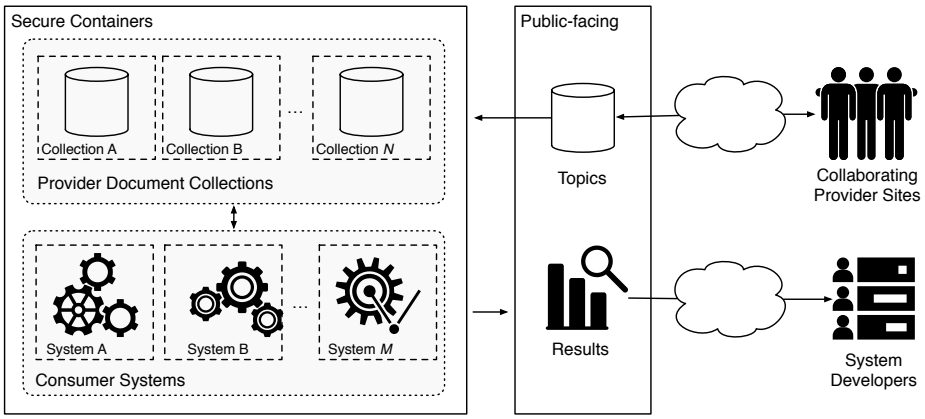
- Text (original "text" model, also Markov Random Field model "*textmrf*")
- Concepts (UMLS CUIs from cTAKES are unique identifiers for a "*concept*")
- Dependencies (dependency parse nodes "depfrag-0" and arcs "*depfrag-1*")
- Topics (inferred from an LDA model "*lda-N*")

The probabilities resulting from each layer's search were linearly combined to rank patient documents.



12

Future directions: Evaluation as a Service (EaaS) – Pending R01



Conclusions

- Importance of IR in biomedicine will not diminish as volume, variety, and velocity of science continue to expand
- Varying benefits for different use cases, but in general, medical vocabulary resources offer most value via query expansion
- While ad hoc IR for general information needs relatively solved, still challenges with
  - Novel types of data, e.g., medical records and other structured data
  - High-recall tasks, e.g., systematic reviews
- Research confounded by larger issues, e.g.,
  - Private data
  - Proprietary data