Information Retrieval from Electronic Health Records for Patient Cohort Discovery

William Hersh, MD
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

References


https://academic.oup.com/jamiaopen/article/1/2/188/5056817

Overview

- Re-use of clinical data
- Primer on information retrieval (IR) and challenge evaluations
- TREC Medical Records Track
- Information retrieval from electronic health record (EHR) for patient cohort discovery
- Future work
Thanks for the invitation – Corvallis is a special place for me

Re-use of clinical data

• Many re-uses (secondary use) of EHR data, including (Safran, 2007)
  – Clinical and translational research – generating hypotheses and facilitating research
  – Public health surveillance for emerging threats
  – Healthcare quality measurement and improvement
• Opportunities facilitated by widespread adoption of EHRs (Washington, 2017)
  – 96% of hospitals (Henry, 2016), 86% of physicians (Myrick, 2019)
Information retrieval – IR (Hersh, 2009)

- Aka, “search”
- Focus on indexing and retrieval of information
- Historically centered on text in documents, but increasingly associated with many types of content
- [www.irbook.info](http://www.irbook.info)

Evaluation of IR systems (Harman, 2011)

- 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 both in a single measure
    - Mean average precision (MAP)
    - Normal discounted cumulative gain (NDCG)
    - Binary preference (Bpref)
- User-oriented – how well user performs with system
  - e.g., performing task, user satisfaction, etc.
System-oriented IR evaluation

- 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
- Evaluation consists of runs using a specific IR approach with output for each topic measured and averaged across topics

Recall and precision

- Recall
  \[ R = \frac{\text{\# retrieved and relevant documents}}{\text{\# relevant documents in collection}} \]
  - Usually use relative recall when not all relevant documents known, where denominator is number of known relevant documents in collection
- Precision
  \[ P = \frac{\text{\# retrieved and relevant documents}}{\text{\# retrieved documents}} \]
Challenge evaluations

- A common approach in computer science, not limited to IR
- Develop a common task, data set, evaluation metrics, etc., ideally aiming for real-world size and representation for data, tasks, etc.
- In case of IR, this usually means
  - Test collection of content items
  - Topics to be retrieved
  - Runs from participating groups with retrieval for each topic
  - Relevance judgments of which content items are relevant to which topics – judged items derived from submitted runs

Some well-known challenge evaluations in IR

  - Mostly non-biomedical; first domain-specific track was Genomics Track (Hersh, 2009)
  - Focus on retrieval across languages, but also on image retrieval, including medical image retrieval tasks – www.imageclef.org (Hersh, 2009; Müller, 2010)
- TREC has inspired other challenge evaluations, e.g., n2c2/i2b2 NLP Shared Task, https://n2c2.dbmi.hms.harvard.edu/
TREC Medical Records Track

- Focused on use case of identifying patients for possible recruitment into clinical studies
  - Task to “retrieve” patients who fit search criteria
- Used de-identified EHR data from University of Pittsburgh Medical Center (UPMC)
- Ran for two years (Voorhees, 2011; Voorhees, 2012; Voorhees, 2013)

Test collection

VISIT LIST

RECORD-VISIT MAP

17,265 visits 101,712 reports (93,552 mapped to visits)

[Report Extract]

[Voorhees, 2013]
Some issues for test collection

• De-identified to remove protected health information (PHI), e.g., age number → range
• De-identification precludes linkage of same patient across different visits (encounters)
• UPMC only authorized use for TREC 2011 and TREC 2012 but nothing else, including any other research (unless approved by UPMC)
Topic development and relevance assessments for 2011 track

• Task – Identify patients who are possible candidates for clinical studies/trials
• Topics derived from 100 top critical medical research priorities in comparative effectiveness research (IOM, 2009)
  – Selected 35 topics assessed for appropriateness for data and with at least some relevant “visits”
• Relevance judgments by physicians who were OHSU informatics students

Participation in 2011

• Runs consisted of ranked list of up to 1000 visits per topic for each of 35 topics
  – Automatic – no human intervention from input of topic statement to output of ranked list
  – Manual – everything else
• Up to 8 runs per participating group
• Subset of retrieved visits contributed to judgment sets
  – Because resources for judging limited, could only judge relatively small sample of visits, necessitating use of BPref for primary evaluation measure
• 127 runs submitted from 29 groups
  – 109 automatic
  – 18 manual
Evaluation results for top runs...

... BUT, wide variation among topics
Easy and hard topics

• Easiest – best median bpref
  – 105: Patients with dementia
  – 132: Patients admitted for surgery of the cervical spine for fusion or discectomy
• Hardest – worst best bpref and worst median bpref
  – 108: Patients treated for vascular claudication surgically
  – 124: Patients who present to the hospital with episodes of acute loss of vision secondary to glaucoma
• Large differences between best and median bpref
  – 125: Patients co-infected with Hepatitis C and HIV
  – 103: Hospitalized patients treated for methicillin-resistant Staphylococcus aureus (MRSA) endocarditis
  – 111: Patients with chronic back pain who receive an intraspinal pain-medicine pump

Failure analysis for 2011 topics (Edinger, 2012)
Participation in 2012 track

- Task and document collection same as 2011
- Developed 50 new topics
- More judging resources than 2011 allowed for more relevance judgments
  - For each topic, pooled top 15 from all runs and 25% of all visits ranked 16-100 by any run
- 88 runs submitted from 24 groups
  - 82 automatic
  - 6 manual

Evaluation results from 2012

<table>
<thead>
<tr>
<th>Run</th>
<th>infNDCG</th>
<th>P(10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLMManual*</td>
<td>0.680</td>
<td>0.749</td>
</tr>
<tr>
<td>udelSUM</td>
<td>0.578</td>
<td>0.592</td>
</tr>
<tr>
<td>semnamed2</td>
<td>0.547</td>
<td>0.557</td>
</tr>
<tr>
<td>ohsuManBoo1*</td>
<td>0.526</td>
<td>0.611</td>
</tr>
<tr>
<td>attgeo1</td>
<td>0.524</td>
<td>0.519</td>
</tr>
<tr>
<td>UDinfoMed123</td>
<td>0.517</td>
<td>0.528</td>
</tr>
<tr>
<td>uoTrMConQRd</td>
<td>0.509</td>
<td>0.553</td>
</tr>
<tr>
<td>NICTAUTC4</td>
<td>0.487</td>
<td>0.517</td>
</tr>
</tbody>
</table>
What approaches did (and did not) work?

• Best results in 2011 and 2012 obtained from NLM group (Demner-Fushman, 2011; Demner-Fushman, 2011)
  – Top results from manually constructed queries using Essie domain-specific search engine (Ide, 2007)
  – Other automated processes fared less well, e.g., creation of PICO frames, negation, term expansion, etc.
• Best automated results in 2011 obtained by Cengage (King, 2011)
  – Filtered by age, race, gender, admission status; terms expanded by UMLS Metathesaurus
• Benefits of approaches commonly successful in IR provided small or inconsistent value for this task in 2011 and 2012
  – Document focusing, term expansion, etc.

Next step: use of bigger and better data

• Mayo Clinic-OHSU collaboration (Wu, 2017)
  – Original PI, Stephen Wu, Mayo → OHSU
• Funded by NLM R01
  – Hongfang Liu, Mayo Clinic, Co-PI
  – Steven Bedrick, OHSU, Site PI
  – William Hersh, OHSU, Co-I
• Aimed to add natural language processing (NLP) and language modeling (LM) to base IR methods on large amounts of unmodified (not de-identified) text from EHR
  – Preliminary data showed improvement over baseline IR techniques with TREC Medical Record Track collection (Zhu, 2014)
Collections of patient data

- **OHSU**
  - Extraction of patients from Oregon Clinical and Translational Research Institute (OCTRI) Research Data Warehouse (RDW) having inpatient or outpatient encounters in primary care departments (Internal Medicine, Family Medicine, or Pediatrics) with
    - 3 or more encounters
    - 5 or more text entries
    - Between 1/1/2009 and 12/31/2013
  - Stored on (highly!) secure server
- **Mayo** using comparable approach and quantities

### OHSU data model and statistics

<table>
<thead>
<tr>
<th>Type</th>
<th>Patients</th>
<th>Encounters</th>
<th>Records</th>
<th>Average</th>
<th>Median</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Administered Meds</td>
<td>47,208</td>
<td>125,831</td>
<td>6,497,157</td>
<td>51.634</td>
<td>6</td>
<td>76</td>
</tr>
<tr>
<td>Ambulatory Encounters</td>
<td>99,985</td>
<td>3,760,205</td>
<td>3,760,205</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Current Meds</td>
<td>92,783</td>
<td>-</td>
<td>31,997,402</td>
<td>344.836</td>
<td>64</td>
<td>20,102</td>
</tr>
<tr>
<td>Demographics</td>
<td>99,985</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Encounter Attributes</td>
<td>99,985</td>
<td>8,273,137</td>
<td>6,273,137</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Encounter Diagnoses</td>
<td>99,985</td>
<td>3,725,603</td>
<td>18,170,896</td>
<td>4.877</td>
<td>4</td>
<td>107</td>
</tr>
<tr>
<td>Notes</td>
<td>99,868</td>
<td>3,491,659</td>
<td>10,118,900</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Hospital Encounters</td>
<td>73,303</td>
<td>466,252</td>
<td>466,252</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Lab Results</td>
<td>83,435</td>
<td>73,461</td>
<td>20,186,748</td>
<td>27.523</td>
<td>12</td>
<td>194,888</td>
</tr>
<tr>
<td>Microbiology Results</td>
<td>27,515</td>
<td>65,373</td>
<td>29,6548</td>
<td>4.536</td>
<td>1</td>
<td>268</td>
</tr>
<tr>
<td>Medications Ordered</td>
<td>94,059</td>
<td>1,388,086</td>
<td>5,336,306</td>
<td>3.845</td>
<td>1</td>
<td>1551</td>
</tr>
<tr>
<td>Procedures Ordered</td>
<td>98,514</td>
<td>1,880,309</td>
<td>7,229,854</td>
<td>3.845</td>
<td>1</td>
<td>6681</td>
</tr>
<tr>
<td>Problem List</td>
<td>90,722</td>
<td>-</td>
<td>761,200</td>
<td>8.391</td>
<td>6</td>
<td>182</td>
</tr>
<tr>
<td>Result Comments</td>
<td>22,716</td>
<td>468,814</td>
<td>915,554</td>
<td>1.955</td>
<td>1</td>
<td>691</td>
</tr>
<tr>
<td>Surgeries</td>
<td>18,640</td>
<td>29,895</td>
<td>31,389</td>
<td>1.067</td>
<td>1</td>
<td>41</td>
</tr>
<tr>
<td>Vitals</td>
<td>99,098</td>
<td>1,362,431</td>
<td>6,647,115</td>
<td>4.879</td>
<td>2</td>
<td>6387</td>
</tr>
</tbody>
</table>
Topics

• From OHSU
  – Derived from clinical study data requests by researchers from the Oregon Clinical and Translational Research Institute (OCTRI), querying Research Data Warehouse (RDW) (29 topics)
• From Mayo
  – Phenotype KnowledgeBase (PheKB) (5 topics)
  – Healthcare quality measures from NQF (12 topics)
  – Rochester Epidemiology Project (REP) (8 topics)
  – Mayo RDW (2 topics)

Sample topics

| Adults with IBD who haven’t had GI surgery | Adults with inflammatory bowel disease who haven't had surgery involving the small intestine, colon, rectum, or anus. |
| Adults with a Vitamin D lab result | Adults with a lab result for 25-hydroxy Vitamin D collected between May 15 and October 15. |
| Postherpetic neuralgia treated with topical and systemic medication | Adults with postherpetic neuralgia ever treated by concurrent use of topical and non-opioid systemic medications. |
| Children seen in ED with oral pain | Children who were seen in the emergency department with herpetic gingivostomatitis, herpangina or hand, foot, and mouth disease, tonsillitis, gingivitis, or ulceration (aphthae, stomatitis, or mucositis) not due to chemotherapy or radiation. |
| ACE inhibitor-induced cough | Adults who have used an ACE inhibitor and experienced ACE inhibitor-induced cough. |
Evaluation across sites

- Relevance assessments being done behind firewall at each site
  - Using Web-based relevance judging system based on one used for TREC – deployed at both sites
- Early results show patient-level retrieval requires queries on both structured and unstructured data
  - Will follow with failure analysis

Future directions

- Starting collaborative project with a pharmaceutical company to detect signals of a rare disease with common symptoms, porphyria
- Developing new methods to allow access to highly private data by other researchers (Hanbury, 2015; Roegiest, 2016)
  - Used for email spam detection, corporate repositories, etc.
Evaluation as a Service (EaaS)

Beyond this research

- **Exciting times for informatics**
  - Artificial intelligence and machine learning
  - Standards and interoperability – SMART on FHIR
- **You can be part of it**
  - Estimated growth of nearly 20K jobs in clinical informatics alone (Hersh, 2018)
  - OHSU Biomedical Informatics Graduate Program
    - [http://www.ohsu.edu/informatics](http://www.ohsu.edu/informatics)