

Enhancing Access to the Bibliome: The TREC Genomics Track

William Hersh, M.D.
Department of Medical Informatics & Clinical Epidemiology
Oregon Health & Science University
Portland, OR, USA
Email: hersh@ohsu.edu
Web: www.billhersh.info

1



There is much I could talk about (would take >>1 hour!)

- Motivations
- Past Tracks
 - 2003 – preliminary ad hoc retrieval and prediction of GeneRIFs
 - 2004 – ad hoc retrieval and two types of categorization
 - 2005 – ad hoc retrieval and more detailed exploration of one type of categorization
- Current Work
 - 2006 – retrieval of passages, aspects, and documents
- Future Directions

2



Since I only have one hour, here is what I will cover today

- Motivations
- Past Tracks
 - 2003 – preliminary ad hoc retrieval and prediction of GeneRIFs
 - 2004 – ad hoc retrieval and two types of categorization
 - 2005 – ad hoc retrieval and more detailed exploration of one type of categorization
- Current Work
 - 2006 – retrieval of passages, aspects, and documents
- Future Directions

3



Acknowledgements

- Track participants and volunteers
 - Topic collectors acknowledged in papers
- OHSU team
 - Data management – Ravi Teja Bhupatiraju, Aaron Cohen, Hair Krishna Rekapalli
 - Analysis – Aaron Cohen, Jianji Yang
 - Relevance judges – Laura Ross, Phoebe Roberts, Andrew Amata, Alita Miller, Bradley Feilmeier
- Data providers
 - National Library of Medicine
 - Mouse Genomic Informatics
 - Highwire Press
- Funder
 - National Science Foundation Grant ITR-0325160
- NIST and Ellen Voorhees
- Track steering committee

4



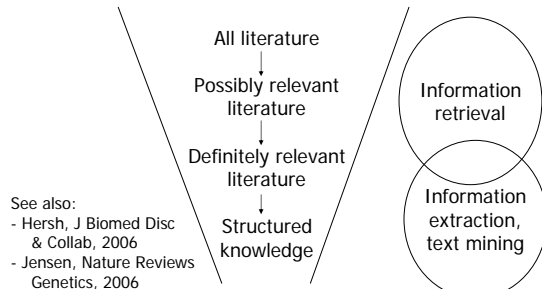
Motivation

- We are in an era of “high-throughput”, data-intensive science
- Biology and medicine provide many information challenges for information retrieval, extraction, mining, etc.
- Many reasons to structure knowledge with development of annotation, model organism databases, cross-data linkages, etc.
- Growing array of publicly accessible data resources and tools that may aid these tasks
- “A couple of months in the laboratory can frequently save a couple of hours in the library.” – Frank Westheimer

5



Emerging approach to biological knowledge management



6



Text Retrieval Conference (TREC, trec.nist.gov)

- Forum for comparative evaluation of IR systems
 - Mostly focused on newswire and government information
 - Competition minimized; collegiality maximized
- Organized by NIST
- Annual cycle consisting of
 - Distribution of test collections and queries to participants
 - Determination of relevance judgments and results
 - Annual conference for participants at NIST
- Began in 1992 and has continued annually
- Each year has ~7 tracks representing specific foci
 - e.g., Web searching, cross-language searching, question-answering, interactive searching, etc.

7



TREC (and IR) evaluation

- Historically focused on metrics of recall and precision, despite criticism of over reliance on these approaches
- Recall

$$R = \frac{\# \text{retrieved and relevant documents}}{\# \text{relevant documents in collection}}$$

- Precision

$$P = \frac{\# \text{retrieved and relevant documents}}{\# \text{retrieved documents}}$$

- Measured by use of test collections of known documents, topics, and relevance judgments
- Can be combined into a single point measure
 - Average of precision at each retrieved document (point of recall)
 - Average across all topics is the mean average precision (MAP)

8



TREC Genomics Track

- Began in 2003, continued in 2004-2006 and beyond
- Focus on real-world information problems in genomics
- 2004 and 2005 tracks focused on two tasks
 - Ad hoc retrieval task
 - Modeled after biologist with acute information needs
 - Used MEDLINE – still entry point to literature for most
 - Categorization task
 - Motivated by real-world problems faced by Mouse Genome Informatics (MGI) curators

9



Participation – growing annually; largest track in TREC

Year	Groups doing ad hoc task	Groups doing "other" task	Total groups
2003	25	14	28
2004	27	20	33
2005	32	19	41

10



TREC 2005 Genomics Track ad hoc retrieval task

- Documents
- Topics
- Relevance judgments
- Results
- Preliminary analysis

11



Ad hoc retrieval task documents

- Developed MEDLINE subset for 2004 and continued use in 2005
 - 10 years from 1994 to 2003
 - ~4.5M documents
 - About one-third of entire database, which goes back to 1966
 - ~9 GB text (MEDLINE format)
- Also promoting use of collection for other tasks beyond Genomics Track (Cohen et al., JAMIA, 2006)

12



Ad hoc retrieval task topics

- In 2004, used general information needs statements common to TREC, but in 2005 focused on more structured topics
- Still representative of common information needs but might allow other resources to be used to improve results
- Developed generic topic types (GTTs) and then interviewed real biologists to obtain real information needs that fit into template
- Transformed information needs into searchable topics

13

GTTs

Generic Topic Type (GTT)	Topics
Find articles describing standard <u>methods or protocols</u> for doing some sort of experiment or procedure	100-109
Find articles describing the role of a <u>gene</u> involved in a given <u>disease</u>	110-119
Find articles describing the role of a <u>gene</u> in a specific <u>biological process</u>	120-129
Find articles describing interactions (e.g., promote, suppress, inhibit, etc.) between two or more <u>genes</u> in the <u>function of an organ</u> or in a <u>disease</u>	130-139
Find articles describing one or more <u>mutations</u> of a given <u>gene</u> and its biological impact	140-149

14

Example topics for selected GTTs

Generic Topic Type (GTT)	Example Topic
Find articles describing standard <u>methods or protocols</u> for doing some sort of experiment or procedure	<u>Method or protocol</u> : GST fusion protein expression in Sf9 insect cells
Find articles describing the role of a <u>gene</u> in a specific <u>biological process</u>	<u>Gene</u> : Insulin receptor gene <u>Biological process</u> : Signaling tumorigenesis
Find articles describing one or more <u>mutations</u> of a given <u>gene</u> and its biological impact	<u>Gene with mutation</u> : Ret <u>Biological impact</u> : Thyroid function

15

Relevance judgments

- Pooled retrieved documents from output of runs from the 27 groups who submitted results
- Performed by five judges with varying expertise in biology
- Averages per topic
 - Documents assessed: 820 (total 41,018)
 - Definitely relevant: 50.5 (6.2%; range 0-527)
 - Possibly relevant: 41.2 (5.0%; range 0-182)
 - Definitely + possibly relevant (relevance for runs): 91.7 (11.2%; range 0-709)
- One topic (135) had no definitely or possibly relevant documents, so omitted from analysis

16

Judgment consistency – chance-corrected agreement (kappa)

Judge 2	Relevant	Not relevant	Total
Judge 1			
Relevant	1100	629	1729
Not relevant	546	8204	8750
Total	1646	8833	10479

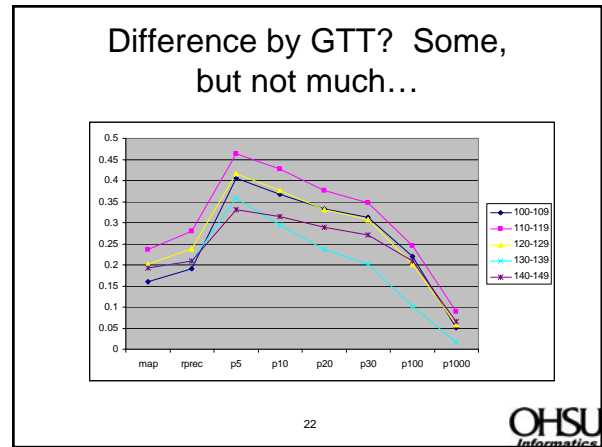
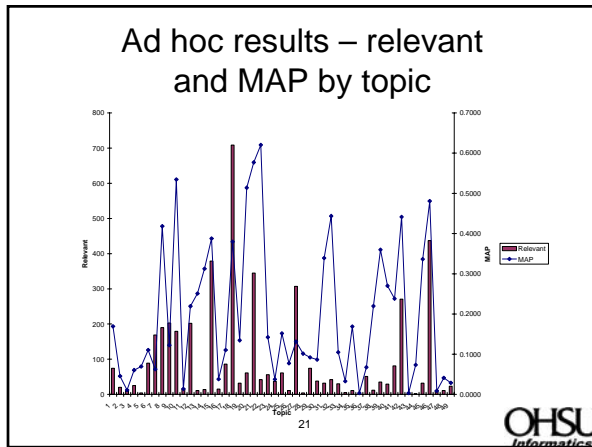
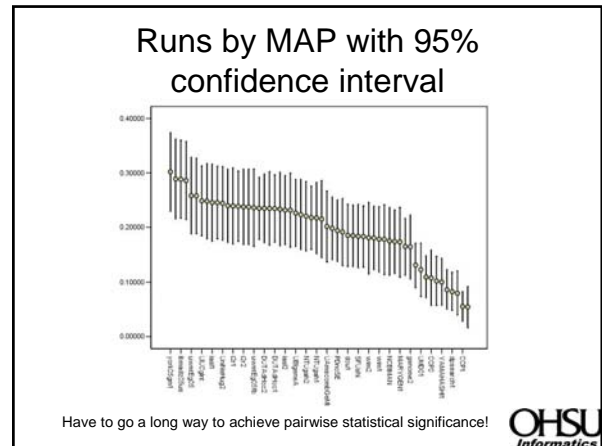
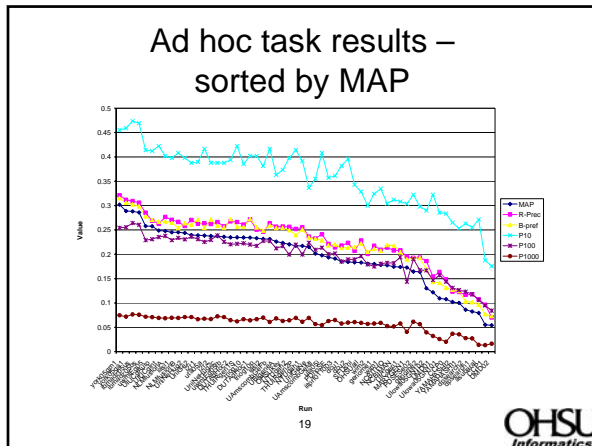
- 9 topics judged in duplicate
- One topic judged three times
- Kappa = 0.58 -> "fair" (almost "good" agreement)

17

Metrics and analysis

- Primary performance metric – mean average precision (MAP)
- Also measured B-Pref, R-Prec, and precision@N documents
- Statistical analysis – repeated measures ANOVA with posthoc pairwise comparisons
- Complete table of all official runs in paper

18



Ad hoc task analysis – preliminary observations

- Manual synonym expansion helped (York – best run with MAP of 0.3136), automated did not (IBM Watson, NLM)
- Relevance feedback without term expansion helped (UIUC)
- Basic Okapi with good parameters gives good baseline performance (several)
 - But better characterization of baseline experiments would improve our understanding

23

OHSU Informatics

Where do we want to go from 2005?

- Web-based participant survey announced to mailing list
- Posted August, 2005
- 26 respondents; results on Web site
- Clear preferences
 - Ad hoc retrieval should move to full text of journal articles
 - “Second” task should focus on information extraction, with some interest in question-answering and summarization

24

OHSU Informatics

TREC 2006 Genomics Track

- Motivations/assumptions
 - Users have questions, seek answers
 - Identifying pertinent passages would make finding relevant documents more efficient
 - Novel passages more important than repeats
- Documents will be full-text of journal articles
- Protocol not finalized; your input solicited!
- May take a second year to get it right

25

Three levels of retrieval (tasks)

- Passages – from one sentence to one paragraph
- Aspects – “normalized” passages on the same answer
 - Exploring grouping by Gene Ontology (GO) or Medical Subject Headings (MeSH)
- Documents
 - Will “roll up” passage relevance, i.e., if a document has a relevant passage, it is relevant

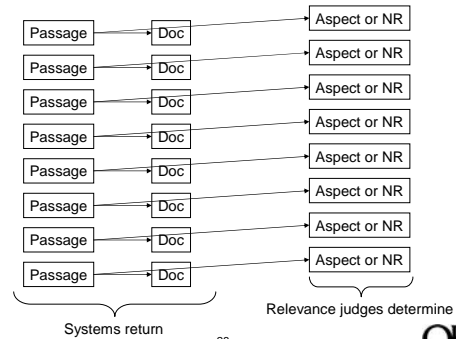
26

What participants will see and do

- Documents – agreement from 56 journals published by Highwire Press to allow full-text HTML use for research purposes
 - About 100,000 documents, 10-20 gigabytes of text
- Topics – topics from 4 of the 5 GTTs reformulated as questions
- Systems return ranked lists of passages and PMIDs they came from for each topic

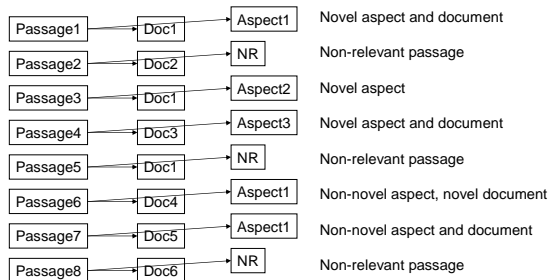
27

“Results”



28

Example results



29

Each level of retrieval presents challenges in measurement

- Passages
 - Partial vs. complete – relative value?
 - Recall difficult to measure because total number of relevant not clear
- Aspects
 - Value of repeats – novel is better but repeat may be of interest to user
 - Precision difficult to measure since systems will not nominate explicitly
- Documents
 - “Roll up” assumptions valid?

30

Current thinking to overcome challenges

- Passages
 - Need to define “gold standard” passages (TREC 2004 HARD Track approach?)
 - Need relative weight for partial overlap with gold standard
- Aspects
 - For first pass, will keep things simple and use mean of mean reciprocal rank (MMRR)
 - Will explore value of repeating aspects later
- Documents
 - Will assume roll-up correct
- If we get it totally wrong, there is always 2007!

31

OHSU
Informatics

Future directions – beyond 2006

- Reaching end of typical “life span” of TREC track
- Can continue until 2008, thanks to NSF grant
- Aim to develop enduring test collections from track data
- Additional goals (from 2003 roadmap) include
 - Interactive user experiments
 - Broader types of users, information needs, tasks

32

OHSU
Informatics

For more information

- Track Web site – <http://ir.ohsu.edu/genomics>
- TREC Web site – <http://trec.nist.gov>
- Hersh, W., Cohen, A., et al. (2005). TREC 2005 Genomics Track overview. *The Fourteenth Text Retrieval Conference - TREC 2005*, Gaithersburg, MD. National Institute for Standards & Technology. <http://trec.nist.gov/pubs/trec14/papers/GEO.OVERVIEW.pdf>.
- Hersh, W., Bhupatiraju, R., et al. (2006). Enhancing access to the biomed: the TREC 2004 Genomics Track. *Journal of Biomedical Discovery and Collaboration*, 1: 3. <http://www.j-biomed-discovery.com/content/1/1/3>.
- Cohen, A. and Hersh, W. (2006). The TREC 2004 Genomics Track categorization task: classifying full-text biomedical documents. *Journal of Biomedical Discovery and Collaboration*, 1: 4. <http://www.j-biomed-discovery.com/content/1/1/4>.

33

OHSU
Informatics