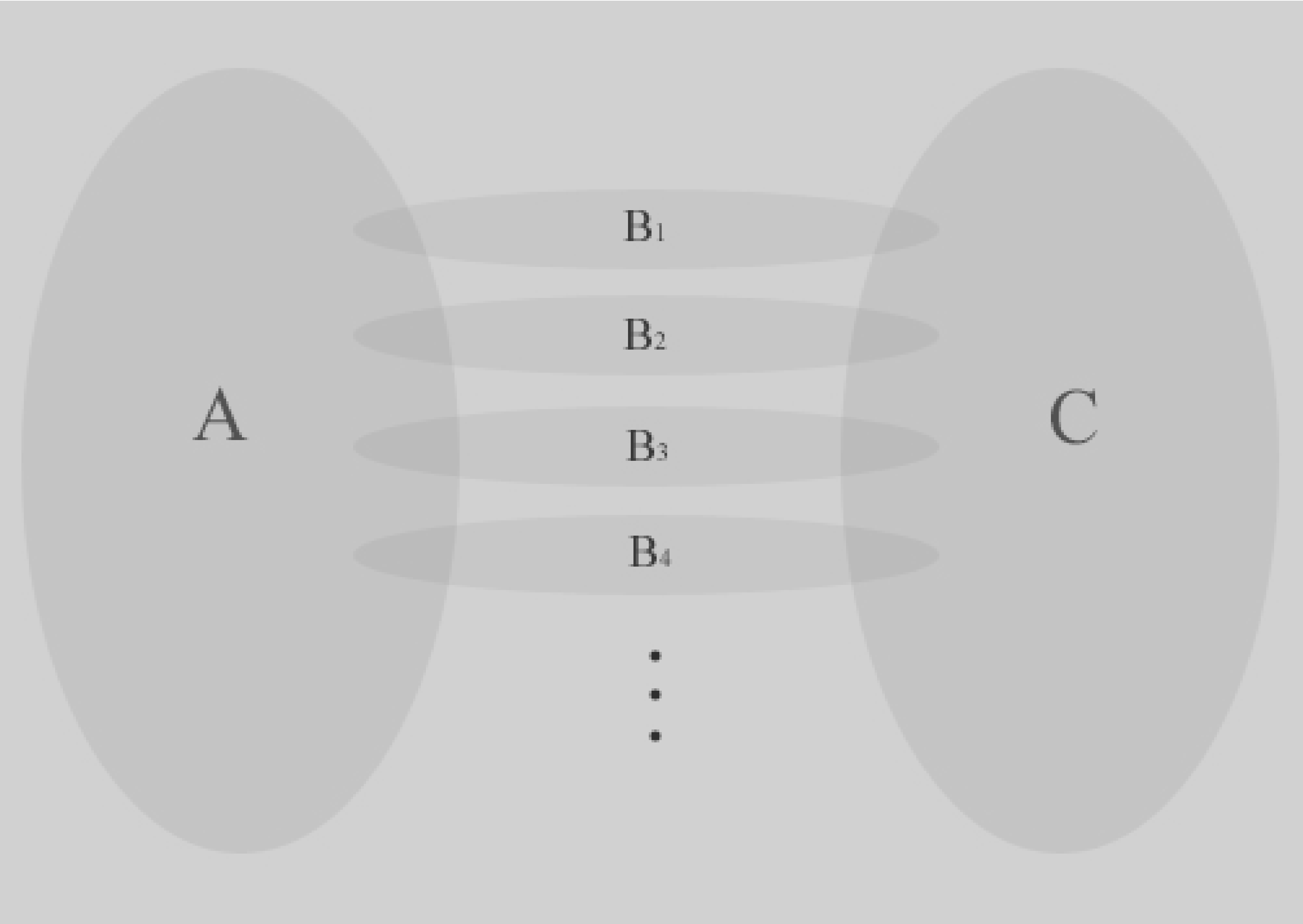
# Linking Two Disparate Sets of Articles in MEDLINE

Neil R. Smalheiser

Vetle I. Torvik

March 11, 2007


A
B1
B2
B3
B4
C

# "The Arrowsmith Project"

- UIC - Smalheiser, Torvik
- U of Chicago - Don Swanson
- 3 field neuroscience testing sites (UIUC, Stanford, UTM, UCSD, UCLA)
- http://arrowsmith.psych.uic.edu
- Focus on 2 node search problem
- Author-ity, Anne O'Tate, ADAM…

# two node searches used as gold standards

- Field testers carried out searches freely
- Chose 6 “well behaved” searches
  - 100s to a few thousand articles in each lit
  - Topically defined and coherent
  - Clear, real-life query need
  - Few or no articles shared across lits
  - Marked relevant B-terms that gave meaningful links (words AND 2- or 3-word phrases)
  - Marked non-relevant B-terms thought not to be meaningful for any query or any user

| **A-literature query** | **C-literature query** | **B-terms** | **Relevant B-terms sought** |
|---|---|---|---|
| retinal detachment[ti]<br>n = 5122 | aortic aneurysm[ti]<br>n = 5687 | n = 2294 | a) Diseases or syndromes in which both features have been described. n = 30 |
| | | | b) Surgical procedures used for diagnosis or treatment of both.<br>n = 26 |

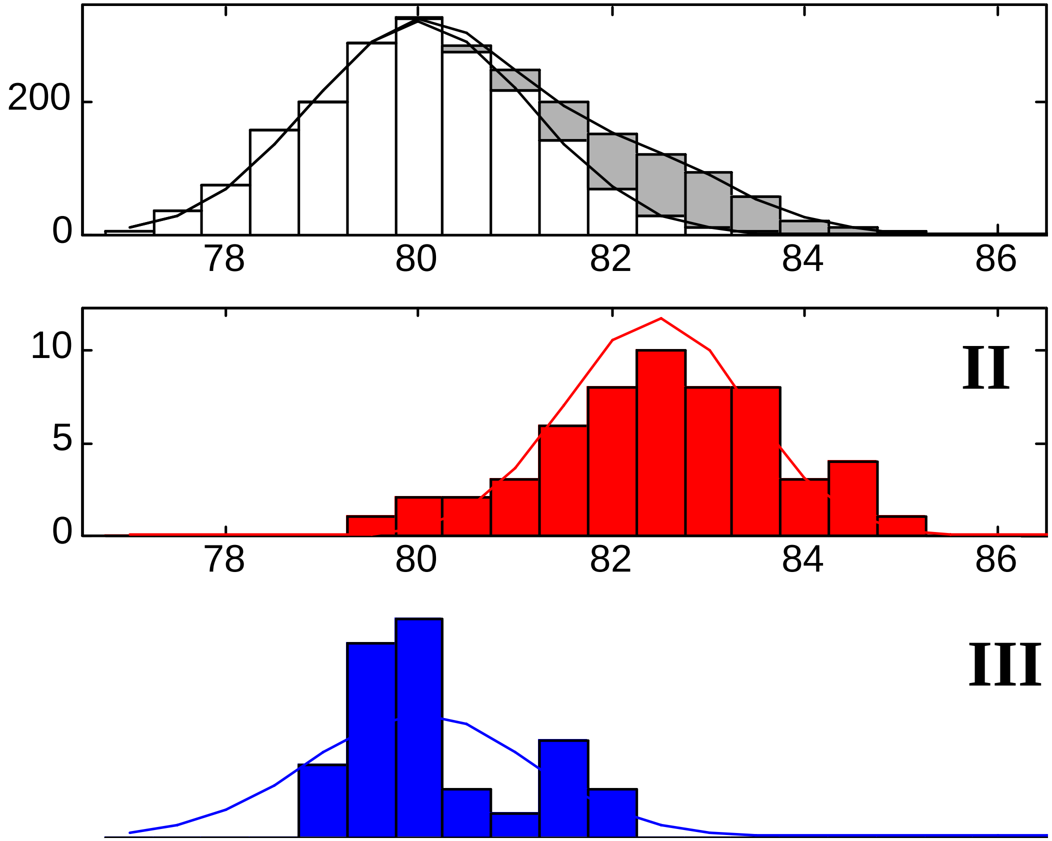
# B-term Features

1. Does the B-term occur in more than one paper within literatures A and C?
2. Do the AB and BC sub-literatures share any MeSH terms?
3. Does the B-term map to at least one UMLS semantic category?
4. Does the B-term have a high literature cohesion score?
5. Is the B-term moderately frequent within MEDLINE as a whole?
6. Did the B-term first appear recently within MEDLINE as a whole?
7. Is the B-term highly characteristic within literature A or C?
8. Do the words within the B-term all occur on the customized 1400 word stoplist?

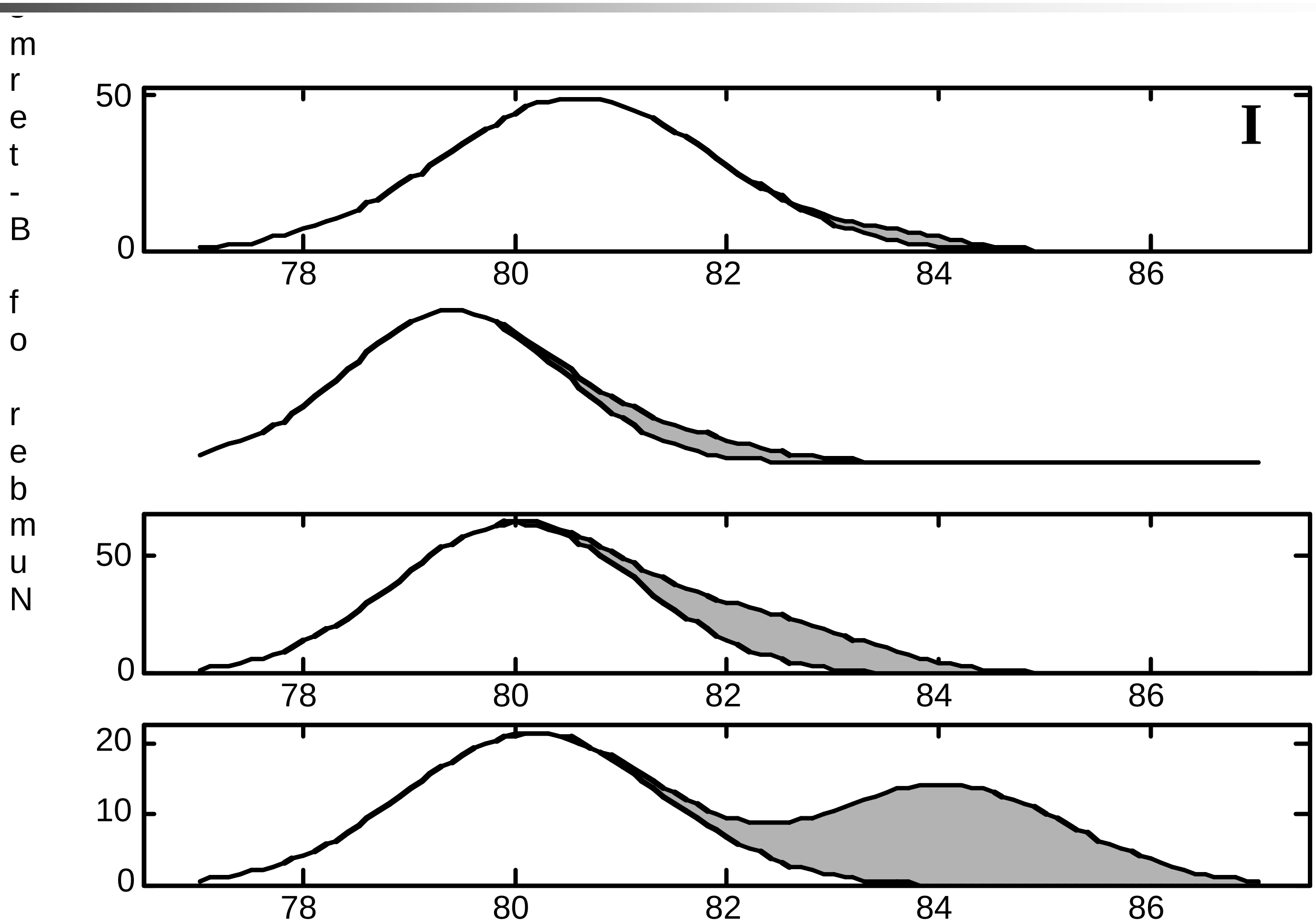
# combining the features into a single B-term score

- Our goal was to derive a mathematical formula for mapping each 8-dimensional vector to a single B-term score, so that this score optimally distinguishes the marked relevant B-terms from all other B-terms. A simple and intuitive formula is the weighted combination of the features:
- B-term score: $y(\mathbf{x}) = w_1x_1 + w_2x_2 + \ldots + w_8x_8$
- Employed logistic regression model

# Distribution of B-term scores can be decomposed into 2 bell curves



# Percentage of predicted relevant B-terms estimates amount of implicit information linking disparate two sets of articles



# How to Evaluate?

- Leave-one out, compute the weightings based on 5 gold standards and evaluate average precision of B-term ranking on the 6th
- Also create an independent, automatic set of queries for testing….

# Genomics TREC 2005 Queries

- **20 queries of the form**
  - Role of gene in disease
  - Role of gene in biological process
  - Took gold standard relevant papers
  - Extracted title terms from these
  - Filtered through Swanson stoplist
  - Treat these as marked relevant terms
- **Formed two node search automatically**
  - Computed ranked B-terms
  - Compute mean average precision

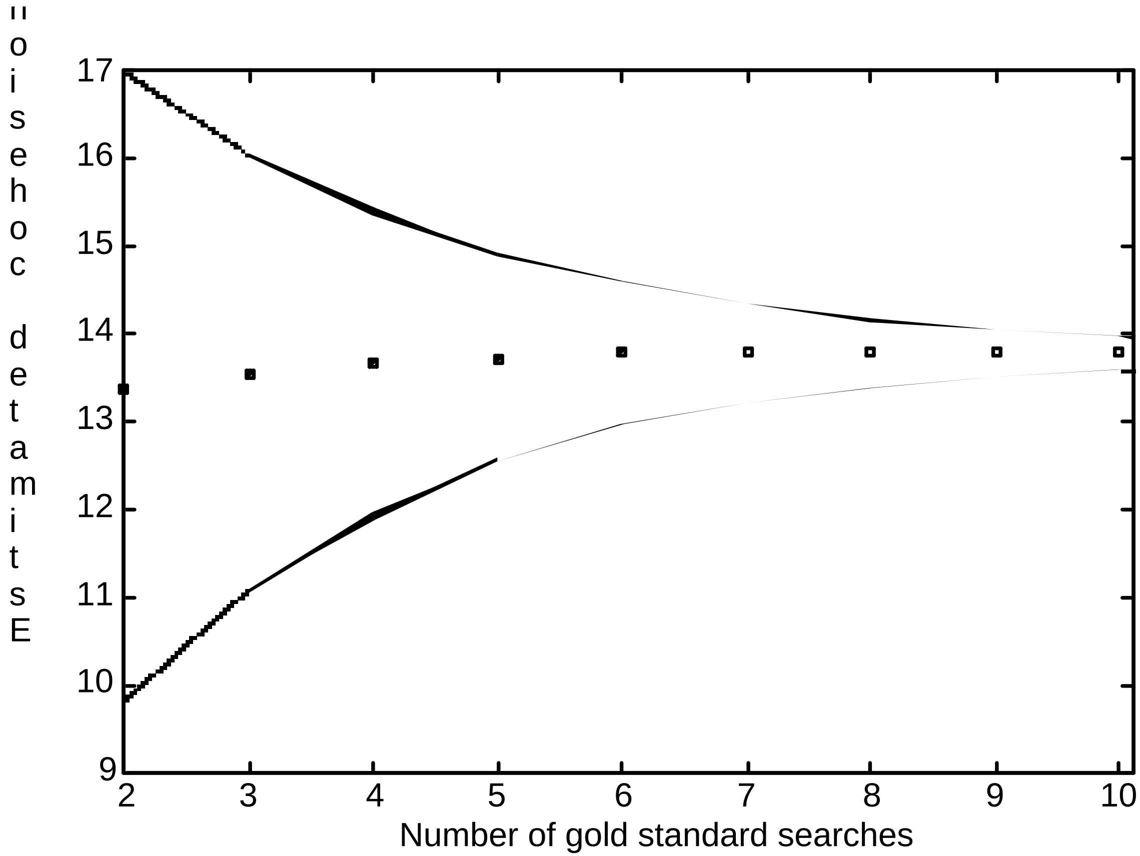
# compare our model against frequency based MIM models

| | MAP (%) two node searches | p-value | MAP (%) TREC queries | p-value ≤ |
|---|---|---|---|---|
| | | | | |
| B-term Score logistic regression model | 27.4 | | 21.0 | |
| B-term Score 6-fold cross-validation | **25.2** | | **21.0** | |
| Average MIM model | 13.3 | 0.0062 | 14.0 | 0.0001 |
| Minimum MIM model | 12.0 | 0.0056 | 12.5 | 0.0001 |

# all features are significant, none are superfluous*

| | | | | |
|---|---|---|---|---|
| B-term Score logistic regression model | 27.4 | | 21.0 | |
| B-term Score 6-fold cross-validation | **25.2** | | **21.0** | |
| Average MIM model | 13.3 | 0.0062 | 14.0 | 0.0001 |
| Minimum MIM model | 12.0 | 0.0056 | 12.5 | 0.0001 |
| | | | | |
| $x_1$ : $N_{AB} > 1$, $N_{BC} > 1$ | 5.0 | 0.0007 | 8.2 | 0.0001 |
| $x_2$ : MeSH in common | 4.7 | 0.0007 | 7.5 | 0.0001 |
| $x_3$ : mapped to UMLS | 5.4 | 0.0014 | 7.9 | 0.0001 |
| $x_4$ : cohesion | 16.3 | 0.0241 | 12.9 | 0.0001 |
| $x_5$ : frequency in MEDLINE | 11.0 | 0.0168 | 7.8 | 0.0001 |
| $x_6$ : 1st year in MEDLINE | 18.0 | 0.0730 | 10.6 | 0.0001 |
| $x_7$ : p-value of $N_{AB} + N_{BC}$ | 10.0 | 0.0047 | 16.9 | 0.0038 |
| Global features ($x_3$, $x_4$, $x_5$, $x_6$) | 20.8 | 0.0144 | 14.4 | 0.0001 |
| Local features ($x_1$, $x_2$, $x_7$) | 9.5 | 0.0043 | 17.5 | 0.0143 |

# assessing the robustness of the model



# Conclusions - 1

- Each user and each query is unique, yet we can predict relevant B-terms using features that are generic and do not involve domain knowledge
- Multiple, complementary features are needed for good performance; cf. frequency based methods.

# Conclusions - 2

- Model also provides estimate of total implicit information linking lits
- Need to generalize the model for lits that are not topical, not coherent, not equal in size, cover range of sizes
- Applicable to one node search, I.e. open ended discovery?
- Applicable to improve IR?