

Anticipating Annotations and Emerging Trends in Biomedical Literature

Presented by Shad Stafford

CIS 607 Data Mining and Data Integration in Bioinformatics

January 23, 2009



The paper

Anticipating Annotations and Emerging Trends in Biomedical Literature by F. Morchen *et al.*

Published in **Proceedings of KDD 2008**

BioJournal Monitor

- ◆ BioJournal Monitor is a text mining platform that performs a variety of services
 - ▶ Named-entity recognition
 - ▶ Document classification
 - ▶ Trend analysis
 - ▶ Clustering
 - ▶ Ranking

Data inputs

◆ Public data inputs

- ▶ PubMed
- ▶ MeSH ontology
- ▶ Gene Ontology
- ▶ UniProt
- ▶ FDA Clinical Trials

◆ Private data inputs

- ▶ Patent information
- ▶ news articles

Monitor Outputs

- ✦ BioJournal Monitor presents users with a web interface
 - ▶ Keyword queries return document clusters and tag clouds
 - ▶ Trends from different categories can be compared
- ✦ Goal appears to be a unified text mining tool that helps researchers find relevant documents

User Interface

BioJournalMonitor alpha Logout

Search by keywords:

Processed 10000 out of 10000 most relevant Documents out of 11,833,337

Concepts 25 Found

12-Month Window Relative Frequency

Keywords	Name	Category	Frequency	First Date	Last Date
<input type="checkbox"/> Cells	ras2	Gene	128	01-01-1998	10-23-2007
<input type="checkbox"/> Diseases	breast	Gene	125	09-24-1985	10-24-2007
<input type="checkbox"/> DNA	bc	Gene	89	03-01-1976	10-24-2007
<input checked="" type="checkbox"/> Genes	pc329444_0176	Gene	82	12-01-1994	06-15-2009
<input type="checkbox"/> Organ...	rsb1	Gene	29	03-01-1999	09-29-2007
<input type="checkbox"/> Proteins	sg11700	Gene	26	02-01-1998	09-01-2007
<input type="checkbox"/> RNA	sgn1	Gene	21	02-01-1998	09-14-2007

Topics 11 Found

Sort by:

[1918 Items]

Antineoplastic Agents, Hormonal - breast - ras2 - breast cancer
 control - developed - DNA, Neoplasm - estrogen - exon 11 - family - genetic predisposition to disease - germline - germline - group - human - pair 17 chromosomes - increase - mp11_amber - neoplasm proteins - shigella dysenteriae - tamoxifen - women

[1784 Items]

breast - breast - BBD - breast cancer cells - cancers - cell line - Contraceptives, Oral - death - estrogen - estrogen - genetic predisposition to disease - germline - human - pair 17 chromosomes - increase - mortality - mp11_amber - neoplasm proteins - solea senegalensis - study - women

Trend of Documents for selected topic [1784 Items]

Copyright (c) 2009 Esheng Computers Research Inc. All Rights Reserved

● Contributions of this paper

- ✦ This paper addresses two specific topics:
 - ▶ tagging documents with terms from the MeSH ontology (predicting the tags from curators)
 - ▶ detecting emerging trends (finding hot topics early)

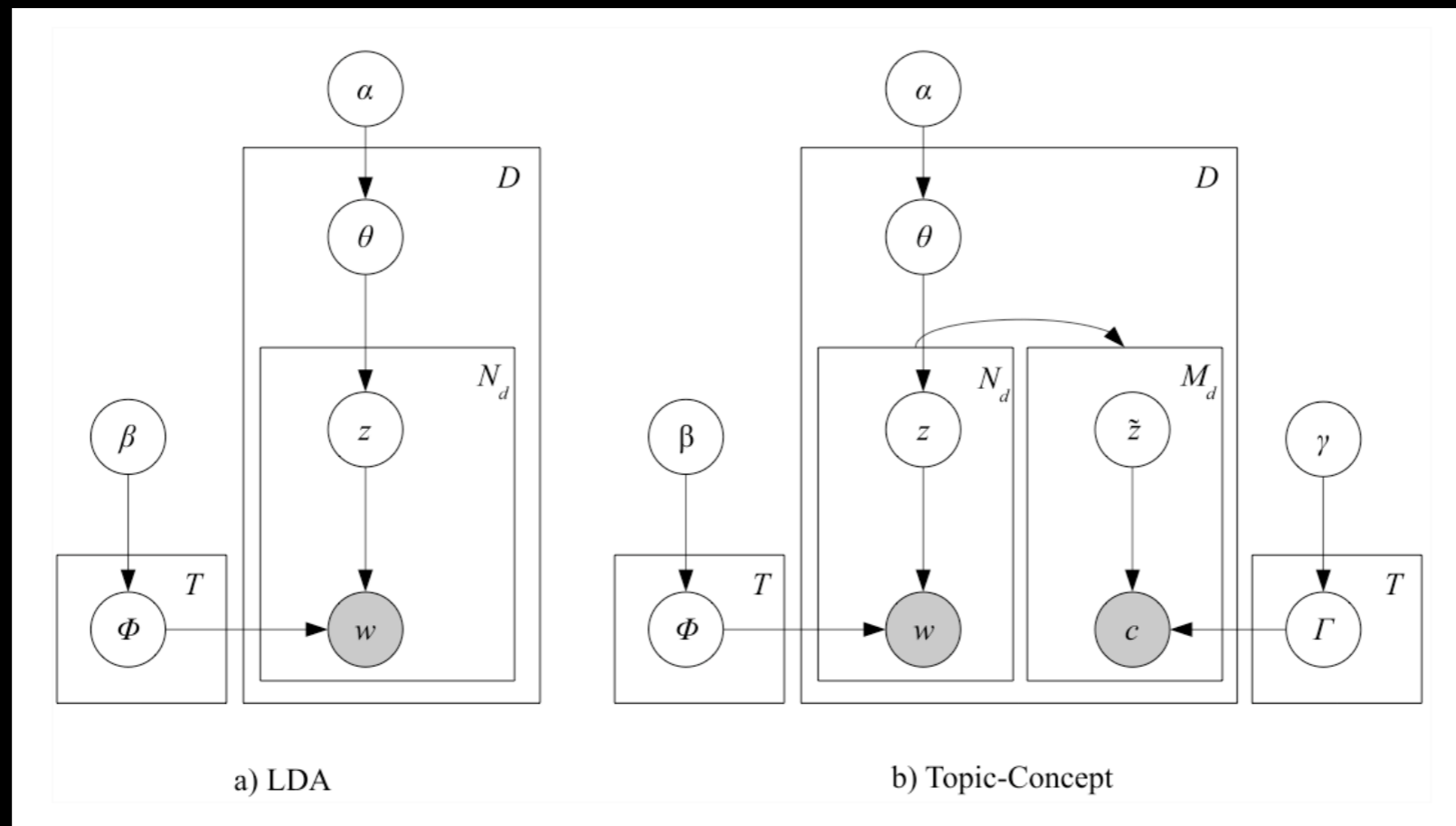
Automated Annotation

- ✦ Articles in PubMed are annotated with MeSH ontology terms manually after publication
 - ▶ this process is expensive and slow
 - ▶ the lag in annotation is a lag in document availability
- ✦ Automating tagging makes documents available faster

The Topic-Concept Model

- ✦ MeSH terms are generated probabilistically for a given document via the Topic-Concept model
- ✦ The Topic-Concept model extends LDA to include *concepts*, which are just MeSH ontology terms
 - ▶ Latent Dirichlet Allocation (LDA) posits that each document is a mixture of topics and each word in a document is attributable to one of those topics

LDA vs TC



First establish the words and topics for a given document.
 Then, for each index term you want, randomly draw a topic (\check{z})
 then choose an index term (c) from the \check{z} specific distribution Γ

Training the TC Model

- ✦ Mappings between words and topics and topics and concepts must be learned
 - ▶ ϕ = the topic-word distribution
 - ▶ Θ = the topic distribution
 - ▶ Γ = the concept-topic distribution
- ✦ ϕ and Θ are estimated using Gibbs sampling
- ✦ $p(c|\check{z})$ is approximated by this distribution:

$$\begin{aligned}
 p(\tilde{z}_i = t | c_i = m, \tilde{\mathbf{z}}_i, \mathbf{z}_{-i}, \mathbf{w}_{-i}) &\propto \\
 p(c_i = m | \tilde{z}_i = t) p(\tilde{z}_i = t | \mathbf{z}) &\propto \\
 \frac{C_{mt}^{CT} + \gamma}{\sum_{m'} C_{m't}^{CT} + M\gamma} \frac{C_{td}^{TD}}{N_d}
 \end{aligned}$$

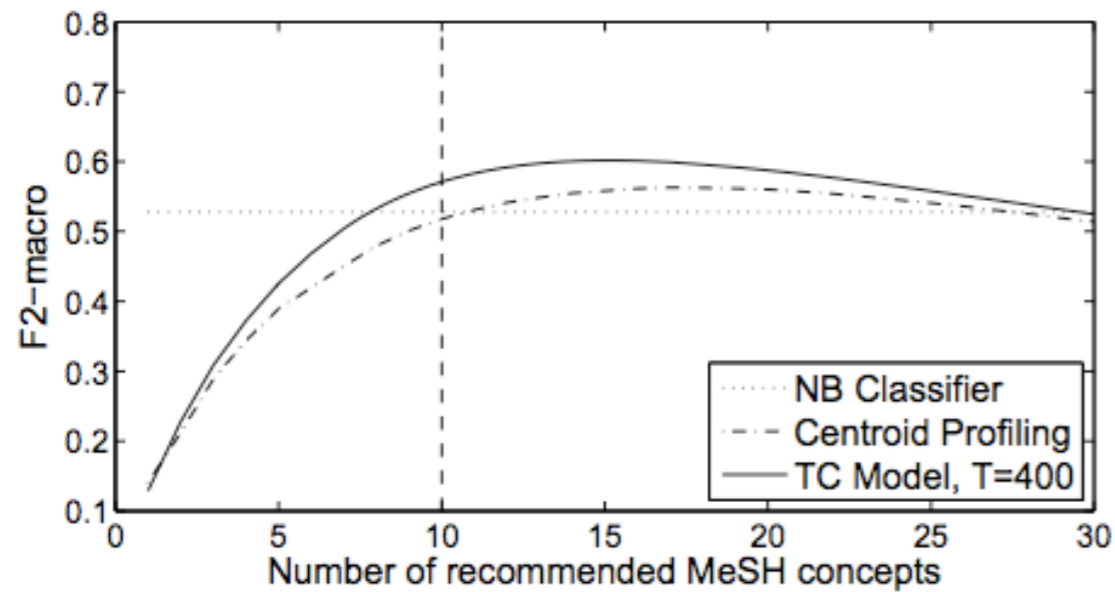
Experiments

- ◆ Two large MEDLINE corpora used to train the TC model
 - ▶ abstracts only: stemmed, stop words removed
 - ▶ MeSH descriptors pruned to first-level of sub-branch
 - results in 108 remaining labels (~10 per document)
- ◆ Compare TC model with centroid profiling and Naive Bayes classifier
- ◆ Train on 90%, test against 10%

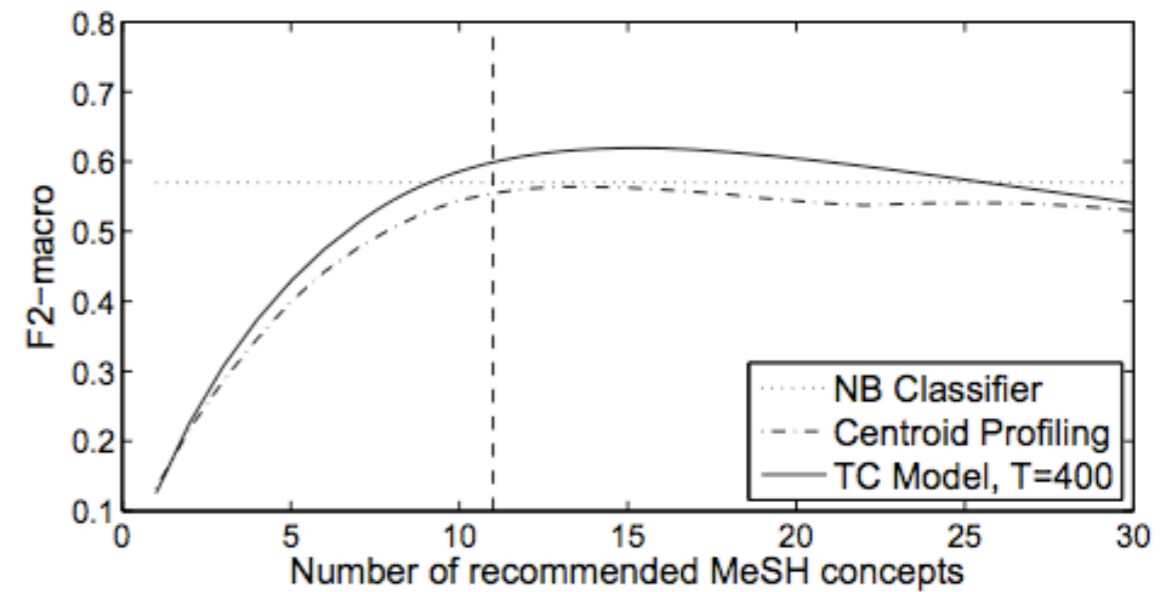
Metric

- ✦ The test is to predict MeSH labels for sample documents and compare against actual MeSH labels
- ✦ Metric is the F2-macro
 - ▶ Similar to the F-score, but weights recall twice as much as precision

Results



(a) *random 50K corpus*



(b) *genetics corpus*

TC Model marginally outperforms centroid profiling and naive bayes classifier at the 10 MeSH labels (the average number observed in the corpora)

Emerging Trend Prediction

- ✦ Emerging trend prediction seeks to predict the inclusion of new medical terms into the MeSH ontology
- ✦ This is back-tested against cancer related terms.
 - ▶ Basically, could we have predicted that cancer research was going to be such a hot topic by looking at trends in MeSH terms
 - ▶ [SS: there's an inherent selection bias, no?]

Data Collection

- ✦ Filter the PubMed DB for documents with cancer related terms
 - ▶ abstracts only again
 - ▶ date assigned is earliest appearance
- ✦ MeSH ontology filtered for cancer related terms yielding 140 terms
 - ▶ these needed to be matched to word-stems appearing in the abstract (word stem '*brca1*' maps to '*Genes, BRCA1*')
 - ▶ This narrowing yielded 81 mesh terms (true positives for cancer related biomarkers)

Trend Analysis

- ✦ To find a trend:
 - ▶ For each month: consider the frequency of a word-stem
 - ▶ Score the word-stem based on increasing frequency
 - slope of its frequency graph over time is the score
 - ▶ Words that score consistently high show a new hot topic

● Example: breast cancer gene

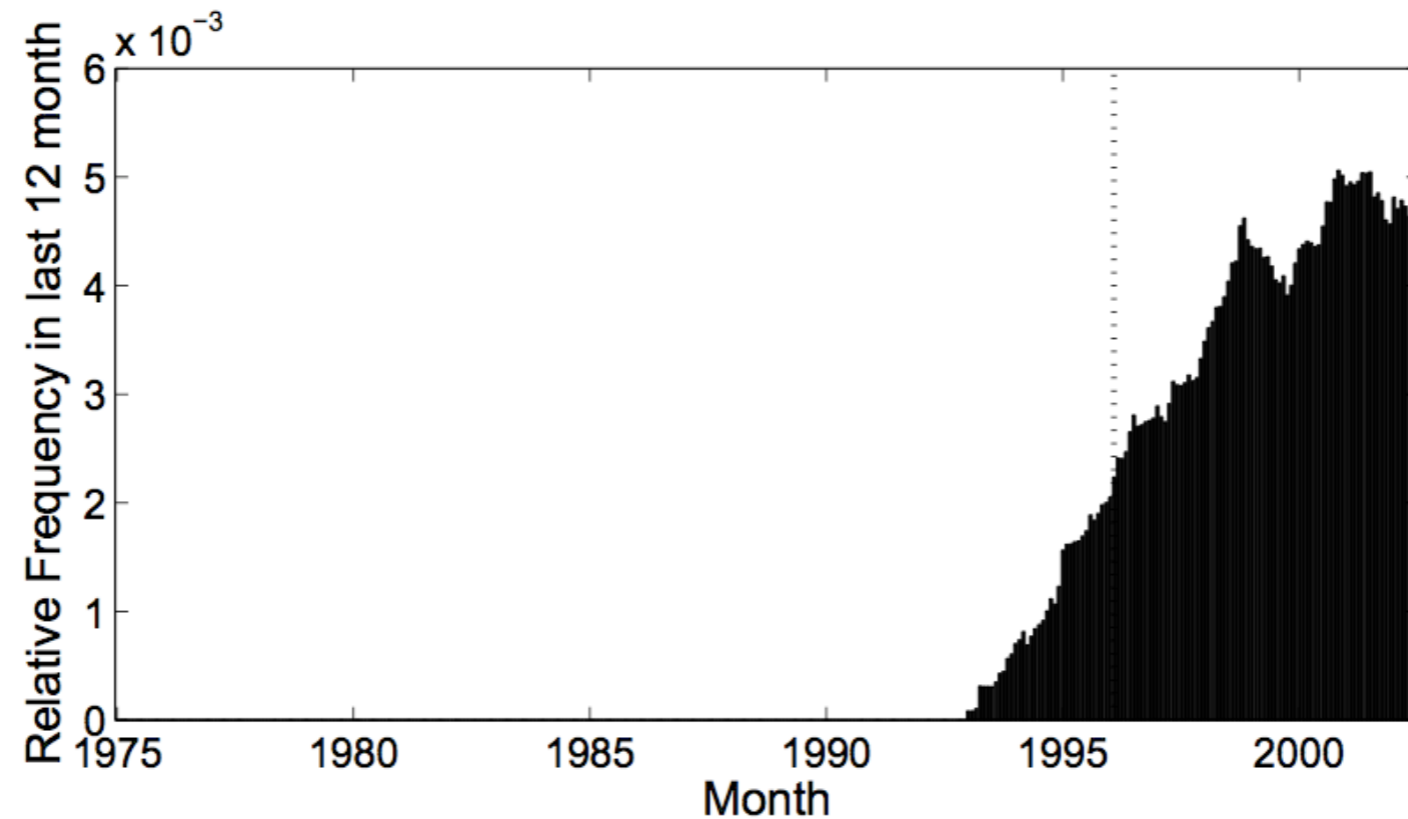


Figure 6: Trend of word stem *brca1* that corresponds to the MeSH term *Genes, BRCA1* added on 2/16/1996 (dashed line).

Evaluation

- ✦ Experiment: Measure scores of 81 known true-positives plus 10k random word stems
- ✦ Over 20 years, 5760 unique terms appear as top-200 in at least one month
 - ▶ 75 of 81 cancer terms are in top-200 words

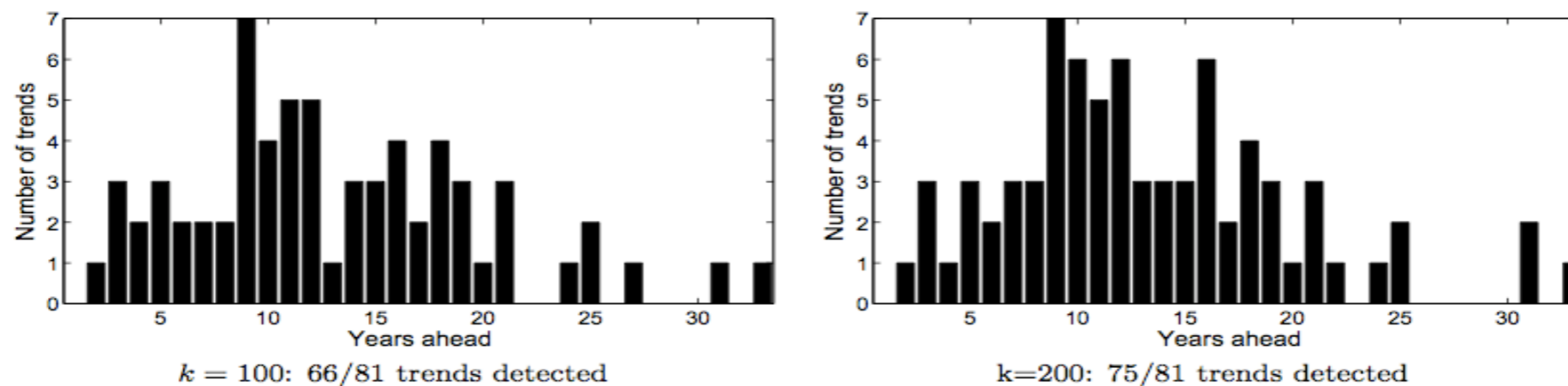


Figure 7: Time difference in years between inclusion in MeSH and earliest detection for positive trends using the top k trends every month. Larger value indicates earlier detection.

Conclusion

- ✦ It is difficult to tell from this article how useful these features are in practice
 - ▶ Is the automated MeSH annotation accurate enough that researchers can find relevant documents sooner?
 - ▶ Is the trend prediction useful? How would you measure that?