

# Text Categorization (Journal Descriptor Indexing)

By

Susanne M. Humphrey

Computer Science Branch

with Lexical Systems Group, Cognitive Science Branch

Lister Hill National Center for Biomedical Communications

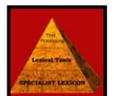
National Library of Medicine

6-27-2008



# Text Categorization (TC) Project

- Primarily concerned with developing TC Web tools; also doing research on TC using tools.
- TC Web tools do two types of categorization at this time:
  - Journal Descriptor Indexing (JDI): categorizes text according to Journal Descriptors (JDs)
  - Semantic Type Indexing (STI) categorizes text according to Semantic Types (STs)



# What are Journal Descriptors (JDs)?

- Set of 122 MeSH descriptors representing high-level categories, mostly biomedical disciplines.
- Used for indexing journals *per se*
- Assigned by human indexer to the 4100 journals used in TC
- Found in Isi2007.xml, List of Serials for Online Users file.  
Directions for ftp'ing this file at  
[http://www.nlm.nih.gov/tsd/serials/terms\\_cond.html](http://www.nlm.nih.gov/tsd/serials/terms_cond.html)



# What are Journal Descriptors (JDs)?

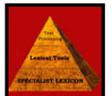
- Examples of information from Isi2007.xml used by TC
  - JID - 03132144
    - TA - Transplantation
    - JD - Transplantation
  - JID - 9802574
    - TA - Pediatr Transplant
    - JD - Pediatrics; Transplantation
  - JID - 0052631
    - TA - J Pediatr Surg
    - JD - Pediatrics; Surgery



# What are Journal Descriptors (JDs)?

- Isi2007.xml produces  
List of Journals Indexed for MEDLINE (LJI)  
<ftp://nimpubs.nlm.nih.gov/online/journals/ljiweb.pdf>
- JDs are in Subject Heading List section with  
“includes” notes and “see” and “see also” references
- JDs are headers in Subject Listing section
- online counterpart at:  
<http://www.nlm.nih.gov/bsd/journals/subjects.html>

Search Journals Database in PubMed, then select  
subject terms link





## Journal Subject Terms

 [Printer-friendly Version](#)

[Return to Journals](#)

Subject Terms are assigned by NLM® to MEDLINE® journals to describe the journal's overall scope. All of these subject terms are valid MeSH® headings. The list below is from 2008 MeSH and is the same list used for the Subject Listing in the NLM publication: [List of Journals Indexed for MEDLINE](#), 2008 edition.

Not all journals in the Journals database have subject terms. For more comprehensive subject access to journals, use the [NLM Catalog](#).

### [A](#) [B](#) [C](#) [D](#) [E](#) [F](#) [G](#) [H](#) [I](#) [J](#) [K](#) [L](#) [M](#) [N](#) [O](#) [P](#) [Q](#) [R](#) [S](#) [T](#) [U](#) [V](#) [W](#) [X](#) [Y](#) [Z](#)

#### A

[Acquired Immunodeficiency Syndrome](#)

[Aerospace Medicine](#)

[Allergy and Immunology](#) - includes Hypersensitivity, Lymphology, Serology, Serotherapy, and Interferons

See Also Transplantation

Alternative Medicine See [Complementary Therapies](#)

[Anatomy](#) - includes Morphology

See Also Cytology; Embryology; Histology; Pathology

[Anesthesiology](#) - includes Resuscitation

[Anthropology](#)

[Anti-Bacterial Agents](#)

[Antineoplastic Agents](#)

[Audiology](#)

#### B

[Bacteriology](#)

[Behavioral Sciences](#) - includes Child Behavior, Sex Behavior, and Suicide

[Biochemistry](#) - includes Biochemical Techniques, Enzymes, Lipids, Nucleic Acids, Proteins, and Vitamins



# Example of JDI

- JDI of the word “transplantation”

1|0.275691|Transplantation

2|0.070315|Hematology

3|0.044303|Nephrology

4|0.031517|Pulmonary Disease (Specialty)

5|0.029425|Gastroenterology

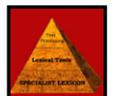
...

122|0.000000|Speech-Language Pathology



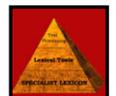
# JDI uses a training set

- Training set is about 3.4 million MEDLINE documents indexed 1999-2002
- JDI requires statistical associations between words in MEDLINE training set record TI/AB and the JD/s corresponding to the journal in the training set record
- JDs are not in a MEDLINE record
- JDs are in the NLM serial record from Isi2007.xml



# JDI uses a training set

- Example of link between MEDLINE record and serial record for *Transplantation*
  - Training set MEDLINE record:  
PMID - 10919582  
TI - Combined liver and kidney transplantation in children.  
**JID - 0132144**  
SO - *Transplantation*. 2000 Jul 15;70(1):100-5.
  - *Transplantation* serial record:  
**JID - 0132144**  
JD - Transplantation



# JDI uses a training set

- Example of Training set MEDLINE record with “imported” JD Transplantation:
  - **PMID - 10919582**
    - TI** - **Combined liver and kidney transplantation in children.**
    - SO** - *Transplantation*. 2000 Jul 15;70(1):100-5.
    - JD** - **Transplantation**



# Calculating JD score for JDI of word

- JDI of the word “transplantation”

1|0.275691|Transplantation  
2|0.070315|Hematology  
3|0.044303|Nephrology  
4|0.031517|Pulmonary Disease (Specialty)  
5|0.029425|Gastroenterology

- Transplantation score

$$\begin{aligned} & \frac{\text{no. of docs in training set in which TI/AB} \\ & \text{word transplantation co-occurs with JD Transplantation}}{\text{no. of docs in training set in which the} \\ & \text{word transplantation occurs in TI/AB}} \\ & = 0.275691 \end{aligned}$$



# Calculating JD score for JDI of word

- JDI of the word “kidney”

1|0.140088|Nephrology

2|0.080848|Transplantation

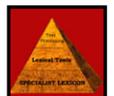
3|0.057162|Urology

4|0.032341|Toxicology

5|0.024398|Pharmacology

- Nephrology score

$$\begin{aligned} & \text{no. of docs in training set in which TI/AB} \\ & \text{word kidney co-occurs with JD Nephrology} \\ = & \frac{\hspace{10em}}{\text{no. of docs in training set in which the} \\ & \text{word kidney occurs in TI/AB}} \\ = & 0.140088 \end{aligned}$$



# Calculating JD score for JDI of phrase

- JDI of the phrase “kidney transplantation”

1|0.178269|**Transplantation**

2|0.092195|Nephrology

3|0.037875|Hematology

4|0.034381|Urology

5|0.017438|Gastroenterology

- Score for **Transplantation** is **average** of Transplantation score for **word kidney** and Transplantation score for **word transplantation**.
- A JD score is average of that JD’s score for word kidney and that JD’s score for word transplantation.



# Calculating JD score for JDI of phrase

- JDI of the phrase “kidney renal nephron glomerulus”

1|0.278721|**Nephrology**

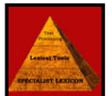
2|0.059499|Urology

3|0.054879|Transplantation

4|0.029262|Physiology

5|0.026824|Pathology

- JD score for **Nephrology** is average of that JD's score for each word in the phrase.



# Calculating JD score for JDI of MEDLINE document TI/AB outside training set

PMID - 17910645

TI - **Kidney transplantation in infants and small children.**

AB - **Transplantation is now the preferred treatment for children with end-stage kidney disease. ...**

SO - *Pediatr Transplant.* 2007 Nov;11(7):703-8.

1|0.102288|**Transplantation**

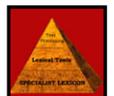
2|0.077717|**Nephrology**

3|0.051765|**Pediatrics**

4|0.023841|Hematology

5|0.021038|Urology

- Score for each JD is average of that JD's score for words in TI/AB



# Calculating JD score for JDI of MEDLINE document TI outside training set

PMID - 15215477

TI - **Pediatric renal-replacement therapy--coming of age.**

SO - *N Engl J Med* 2004 Jun 24;350(26):2637-9.

No abstract available.

1|0.123250|**Nephrology**

2|0.077300|**Pediatrics**

3|0.068716|**Transplantation**

4|0.045671|Urology

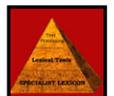
5|0.018311|Otolaryngology



# Word-JD vector

- Scores for an ordered (e.g., alphabetical) list of JDs for a word
- Word-JD vector for word “**kidney**” (showing JDs):

<b>JD Scores</b>	<b>Journal Descriptors</b>
...	...
<b>0.140088</b>	<b>Nephrology</b>
...	...
<b>0.000460</b>	<b>Psychiatry</b>
...	...
<b>0.000308</b>	<b>Psychopharmacology</b>
...	...
<b>0.080848</b>	<b>Transplantation</b>
...	...



# Word-JD vector

- Scores for an ordered (e.g., alphabetical) list of JDs for a word
- Word-JD vector for word “renal” (showing JDs):

<b>JD Scores</b>	<b>Journal Descriptors</b>
...	...
<b>0.223750</b>	<b>Nephrology</b>
...	...
<b>0.000856</b>	<b>Psychiatry</b>
...	...
<b>0.000429</b>	<b>Psychopharmacology</b>
...	...
<b>0.095716</b>	<b>Transplantation</b>
...	...



# Word-JD vector

- Scores for an ordered (e.g., alphabetical) list of JDs for a word
- Word-JD vector for word “**schizophrenia**” (showing JDs):

<b>JD Scores</b>	<b>Journal Descriptors</b>
...	...
<b>0.000000</b>	<b>Nephrology</b>
...	...
<b>0.314520</b>	<b>Psychiatry</b>
...	...
<b>0.067470</b>	<b>Psychopharmacology</b>
...	...
<b>0.000153</b>	<b>Transplantation</b>
...	...



# Vector similarity

- Similarity of **kidney**-JD vector and:
  - **kidney**-JD vector = **1.0**
  - **renal**-JD vector = **0.96**
  - **schizophrenia**-JD vector = **0.03**
- as measured by vector cosine coefficient from:  
G. Salton and M. J. McGill. Introduction to modern information retrieval. New York: McGraw-Hill.1983, p. 124.



# Vector similarity

- Vector cosine coefficient, modified for JDI, for similarity between JD vectors of two words
- Given the JD vectors for two words, **WORD<sub>i</sub>** and **WORD<sub>j</sub>**, the similarity between them may be defined as

$$\text{COSINE}(\text{WORD}_i, \text{WORD}_j) = \frac{\sum_{k=1}^t (WJD_{ik} \cdot WJD_{jk})}{\sqrt{\sum_{k=1}^t (WJD_{ik})^2 \cdot \sum_{k=1}^t (WJD_{jk})^2}}$$



# Vector similarity

- Vector cosine coefficient, modified for JDI, for similarity between JD vector of a word and JD vector of a document
- Given the JD vectors for a word, **WORD<sub>i</sub>** and a document, **DOC<sub>j</sub>**, the similarity between them may be defined as

$$\text{COSINE}(\text{WORD}_i, \text{DOC}_j) = \frac{\sum_{k=1}^t (WJD_{ik} \cdot DJD_{jk})}{\sqrt{\sum_{k=1}^t (WJD_{ik})^2 \cdot \sum_{k=1}^t (DJD_{jk})^2}}$$



# Vector similarity

- Vector cosine coefficient, modified for JDI, for similarity between JD vectors of two documents
- Given the JD vectors for a two documents, **DOC<sub>i</sub>** and **DOC<sub>j</sub>**, the similarity between them may be defined as

$$\text{COSINE}(\text{DOC}_i, \text{DOC}_j) = \frac{\sum_{k=1}^t (\text{DJD}_{ik} \cdot \text{DJD}_{jk})}{\sqrt{\sum_{k=1}^t (\text{DJD}_{ik})^2 \cdot \sum_{k=1}^t (\text{DJD}_{jk})^2}}$$

# Semantic Type Indexing (STI)

- What are Semantic Types (STs)?
- Set of 135 semantic types in the Semantic Network in NLM's Unified Medical Language System (UMLS). STs at [http://www.nlm.nih.gov/research/umls/META3\\_current\\_semantic\\_types.html](http://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html)
- For example, “aspirin” is assigned the STs Pharmacologic Substance (phsu) and Organic Chemical (orch).



# Semantic Type Indexing (STI) in the TC project

- System has word-JD vectors representing JD indexing of each of the 304,000 words in the training set.
- System also has word-ST vectors representing ST indexing of each training set word.
- Thus, STI of text can be performed exactly the same way as JDI of text. An ST score for a text is the average of that ST's score for words in the text. The scores for all the STs comprise the ST vector for the text.



# How are word-ST vectors created?

Basic principle:

When X-JD vector and Y-JD vectors, can create X-Y vector

Specifically, when word-JD vector and ST-JD vectors, can create word-ST vector

<u>Word-JD Vector</u>	<u>Semantic Type (ST)-JD Vectors</u>	
<b><i>transporting</i></b>	<b><i>Cell Function</i></b>	<b><i>Health Care Activity</i></b>
JD1 <score>	JD1 <score>	JD1 <score>
JD2 <score>	JD2 <score>	JD2 <score>
...	...	...

Cell Function (biological transport sense of transporting)

Health Care Activity (patient transport sense of transporting)



# How are word-ST vectors created?

<u>Word-JD Vector</u>	<u>Semantic Type (ST)-JD Vectors</u>	
<b><i>transporting</i></b>	<b><i>Cell Function</i></b>	<b><i>Health Care Activity</i></b>
JD1 <score>	JD1 <score>	JD1 <score>
JD2 <score>	JD2 <score>	JD2 <score>
...	...	...

Similarity between JD vector for the word ***transporting*** and JD vector for the ST ***Cell Function*** = **0.7252**

Similarity between JD vector for the word ***transporting*** and JD vector for the ST ***Health Care Activity*** = **0.3890**

Two of the STs in the ***transporting***-ST vector

***Cell Function*** **0.7252** (biological transport sense)

***Health Care Activity*** **0.3890** (patient transport sense)



# How are word-ST vectors created?

JD indexing of Semantic Types uses “semantic type documents” (ST documents) consisting of one-word Metathesaurus strings belonging to a semantic type

Cell Function document:

PMID- celf

TI - ADCC ADIPOGENESIS AFTERPOTENTIAL AMPHOPHILIA ANOIKIS  
ANTIPORT APOPTOSES APTOPOPOSIS AUTOPHAGOPHYTOSIS ...

AB - BLASTOGENESIS TRANSPORT

Health Care Activity document:

PMID- hlca

TI - ADMINISTRATIVE ADMISSION ADMIT ASSESS ASSISTING  
BLOODLETTING CHECKUP COINSURANCE ...

AB - ADJUSTMENT ADJUSTMENTS ADVOCACY AFTERCARE ...



# How are word-ST vectors created?

<u>Word-JD Vector</u>	<u>Semantic Type Document (ST)-JD Vectors</u>	
<b><i>transporting</i></b>	<b><i>celf document</i></b>	<b><i>hlca document</i></b>
JD1 <score>	JD1 <score>	JD1 <score>
JD2 <score>	JD2 <score>	JD2 <score>
...	...	...

**Similarity** between JD vector for the word ***transporting*** and JD vector for the ***celf document*** = **0.7252**

**Similarity** between JD vector for the word ***transporting*** and JD vector for the ***hlca document*** = **0.3890**

Two of the STs in the ***transporting-ST*** vector

***Cell Function*** 0.7252

***Health Care Activity*** 0.3890

When have word-ST vectors for 304,000 words in training set, can do ST indexing in same manner as JD indexing based on word-JD vectors.



# Research on STI for WSD

- Published research on STI as a tool for word sense disambiguation (WSD) in natural language processing (NLP) using UMLS Metathesaurus, disambiguating 45 ambiguous strings from NLM's WSD collection.

Humphrey SM, Rogers WJ, Kilicoglu H, Demner-Fushman D, Rindflesch TC. Word sense disambiguation by selecting the best semantic type based on Journal Descriptor Indexing: preliminary experiment. *J Am Soc Inf Sci Technol*. 2006 Jan 1;57(1):96-113. Erratum in: *J Am Soc Inf Sci Technol*. 2006 Mar;57(5):726.



# Example in research on STI for WSD

- “transport” is ambiguous:
  - Biological Transport (ST is Cell Function, celf)
  - Patient Transport (ST is Health Care Activity, hlca)
- STI of text results in ranked list of STs.
  - If **celf** ranks higher than **hlca**, then meaning is **Biological Transport**.
  - If **hlca** ranks higher than **celf**, then meaning is **Patient Transport**.



# Example in research on STI for WSD

STI of PMID 9674486 in WSD collection

Input: Preliminary results of bedside inferior vena cava filter placement: safe and cost-effective. The use of inferior vena cava filters (IVCFs) is increasing in patients at high risk for venous thromboembolism; however, there is considerable controversy related to their cost. We inserted eight percutaneous IVCFs at the bedside. The hospital charges for bedside IVCF insertion were substantially lower compared with those for IVCF insertion performed in the Radiology Department or operating room. There was one death (unrelated to the procedure) and one asymptomatic caval occlusion believed to be caused by thrombus trapping. Bedside IVCF insertion is safe and cost-effective in selected patients. This practice averts the potential complications associated with **transporting** critically ill patients.

--- ST scores and rank based on document count for word ---

**27|0.4897|hlca|Health Care Activity**

46|0.4086|celf|Cell Function



# Research on STI for WSD

- Four versions of STI for different contexts of the ambiguity:
  - ambig-sentence - sentence with ambiguity
  - doc - entire MEDLINE document
  - ambig-sentences - all sentences with ambiguity
  - doc-rule: if ambig-sentence = ambig-sentences and ambig-sentence has fewer words than some threshold, then use doc
- STI achieved an overall average precision of 0.7710 – 0.7873 (depending on STI version) compared to 0.2492 for the baseline method.
- STI continues to be investigated for WSD in NLP applications at NLM (MetaMap and SemRep).



# Can create word-SH vectors for Subheading Attachment Project

JDI method in Subheading Attachment Project uses word-SH vectors to produce a ranked list of the top-five SHs for a text to be indexed, and is combined with other methods in the project

<u>Word-JD Vector</u>	<u>MeSH subheading-JD Vectors</u>		
<b><i>surgical</i></b>	<b><i>surgery</i></b>	<b><i>blood supply</i></b>	<b><i>abnormalities</i></b>
JD1 <score>	JD1 <score>	JD1 <score>	JD1 <score>
JD2 <score>	JD2 <score>	JD2 <score>	JD2 <score>
...	...	...	...

**Similarity** between JD vector for the word ***surgical*** and:

JD vector for subheading ***surgery*** = **0.9613**

JD vector for subheading ***blood supply*** = **0.8075**

JD vector for subheading ***abnormalities*** = **0.7804**

Three of the SHs in the ***surgical-SH vector*** sorted by SH:

***abnormalities*** **0.7804**

***blood supply*** **0.8075**

***surgery*** **0.9613**



# Application of word-SH vectors

JDI method in Subheading Attachment Project uses word-SH vectors to produce a ranked list of the top-five SHs for a text to be indexed.

The text-SH vector showing the top five SHs returned by the JDI method applied to the title of MEDLINE document #15165580, **“The role of surgical decompression for diabetic neuropathy.”**

SHs	surgical	decompression	diabetic	neuropathy	avg	rank
blood supply	0.8075	0.5518	0.3348	0.5495	0.5609	5
complications	0.7400	0.4903	0.4499	0.6413	0.5804	3
etiology	0.7777	0.5140	0.4226	0.6200	0.5836	2
physiopathology	0.6009	0.4256	0.5364	0.7034	0.5666	4
surgery	0.9613	0.7455	0.1963	0.4339	0.5842	1

Research included in submission to *Journal of Biomedical Informatics*:

Névél A, Shooshan SE, Humphrey SM, Mork JG, Aronson AR.

A recent advance in the automatic indexing of the biomedical literature.



# Genetics Domain Document Classifier for Gene Symbol Disambiguation

Forthcoming AMIA 2008 paper by Andrej Kastrin and Dimitar Hristovski reports the results of their document classifier (genetics domain or not) based on MeSH indexing of genetically relevant PMIDs. Their classifier achieved predictive accuracy of 0.91 with 0.93 precision and 0.64 recall (0.76 F-score).

Authors sent us two sets of 100 PMIDs they used, annotated by human experts as to whether they were in the genetics domain or not.

JDI/STI limited to sets of genetics JDs and STs

**Genetics JDs:** Genetics; Genetics, Behavioral; Genetics, Medical; Molecular Biology

**Genetics STs:** Gene or Genome; Genetic Function; Nucleotide Sequence

**Genetics STs:** same as above + Nucleic Acid, Nucleoside, or Nucleotide; Molecular Biology Research Technique

Collaboration with Mehmet Kayaalp.



# Gene Symbol Disambiguation

PMID: 15724841

Input: **Implications of p53 in growth arrest and apoptosis on combined ...**

--- rank and score for ST based on word count ---

**1|0.6290|gngm|Gene or Genome**

18|0.4273|nusq|Nucleotide Sequence

32|0.3846|genf|Genetic Function

--- rank and score for ST based on document count for word ---

**1|0.6753|gngm|Gene or Genome**

**8|0.5241|genf|Genetic Function**

15|0.4840|nusq|Nucleotide Sequence

PMID: 15706998

Input: **Safety and feasibility of transradial coronary angioplasty in elderly ...**

--- rank and score for ST based on word count ---

**75|0.1545|gngm|Gene or Genome**

97|0.1058|genf|Genetic Function

116|0.0724|nusq|Nucleotide Sequence

--- rank and score for ST based on document count for word ---

**77|0.1773|gngm|Gene or Genome**

84|0.1586|genf|Genetic Function

117|0.0879|nusq|Nucleotide Sequence



# Genetics Domain Document Classifier

## Optimum Threshold Table (by Mehmet Kayaalp)

Threshold	Accuracy	F-Score	Precision	Recall	Specificity	TP	FP	TN	FN
1	0.87	0.58	1.00	0.41	1.00	9	0	78	13
2	0.88	0.65	0.92	0.50	0.99	11	1	77	11
3	0.89	0.69	0.92	0.55	0.99	12	1	77	10
4	0.89	0.69	0.92	0.55	0.99	12	1	77	10
5	0.89	0.69	0.92	0.55	0.99	12	1	77	10
6	0.90	0.72	0.93	0.59	0.99	13	1	77	9
...									
10	0.90	0.72	0.93	0.59	0.99	13	1	77	9
11	0.91	0.76	0.93	0.64	0.99	14	1	77	8
12	0.92	0.79	0.94	0.68	0.99	15	1	77	7
13	0.94	0.85	0.94	0.77	0.99	17	1	77	5
14	0.94	0.85	0.94	0.77	0.99	17	1	77	5
15	0.93	0.83	0.89	0.77	0.97	17	2	76	5
16	0.93	0.83	0.89	0.77	0.97	17	2	76	5
...									
55	0.72	0.60	0.44	0.95	0.65	21	27	51	1
56	0.71	0.60	0.43	1.00	0.63	22	29	49	0
57	0.69	0.59	0.42	1.00	0.60	22	31	47	0
...									
127	0.22	0.36	0.22	1.00	0.00	22	78	0	0
128	0.22	0.36	0.22	1.00	0.00	22	78	0	0



# Text Categorization research based on JD vector similarity between words

- Automatically-generated stopword list based on similarity between the JD vector for word **THE** and JD vector for each word in the training set.

- Comparing **THE** to:

<b>THE</b>	<b>1.0</b>
<b>AND</b>	<b>0.9998</b>
<b>FOR</b>	<b>0.9977</b>
<b>WITH</b>	<b>0.9970</b>
...	
<b>COMLEX</b>	<b>0.0028</b>

- 303,942 words in training set



# Text Categorization research based on JD vector similarity between indexing terms and documents

## Detecting outlier (blooper) MTI recommendations

----- PMID: 12538701 -----

-- TIAB: Human intestinal epithelial cells are broadly **unresponsive** to **Toll-like receptor 2**-dependent bacterial ligands: implications for host-microbial interactions in the gut. ...

**- Stupor 0.2352935 <= Blooper**

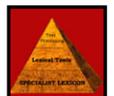
**- Toll-Like Receptor 2 0.9066665**

- Toll-Like Receptor 6 0.9066665
- Epithelial Cells 0.6258414
- Toll-Like Receptor 1 0.9066665
- Intestines 0.558997
- Ligands 0.562745
- Protein Binding 0.68266404
- Interleukin-8 0.837385
- NF-kappa B 0.6850658
- Bacteria 0.66552657
- Peptidoglycan 0.5674213
- Gene Expression Regulation 0.7048282
- Carrier Proteins 0.69688195



# Text Categorization research based on JDI

- Evaluate JDI by running JDI on MEDLINE documents from a journal, thus creating a journal-JD vector by averaging the JD scores across documents from the journal, using native JD of journal as gold standard. Evaluate STI using MeSH indexing to determine gold standard meaning (Guy Divita's idea).
- Specialty subsets. Do JDI indexing of MEDLINE documents from general journal like *New England Journal of Medicine* or *JAMA* in order to partition them into specialty subsets based on JDs. Do this for all MEDLINE to make specialty a PubMed search parameter
- JDI is word-based. Make it phrase-based by extracting phrases from the training set, and creating phrase-JD vectors in the training set itself. Also, consider variants of a word as the same word.
- Use LC call numbers (e.g., RJ1 for Pediatrics, QH431 for Genetics, NA1 for Architecture, QC851 for Meteorol. Climatol.) instead of JDs and expand to automatic indexing by LC Subclasses outside biomedicine.



# ***Bioinspiration & Biomimetics***

New journal for 2007  
20 citations in PubMed

## **Native JDs: Biology; Biomedical Engineering**

JDI of the journal (journal-JD vector) by averaging the JD scores across the 20 PubMed citations:

WC-based JDs:

**1|0.016857 Biology**

**2|0.015707 Biomedical Engineering**

...

DC-based JDs:

**1|0.033130 Biomedical Engineering**

**2|0.032420 Biology**

...

Collaboration with Mehmet Kayaalp.



# TC Tools

- TC Web site: <http://specialist.nlm.nih.gov/tc>
- The TC tools and applications are freely distributed:
  - Freely distributed with open source code
  - 100% in Java
  - Runs on different platforms
  - One complete package
  - Documentation & support
  - Provides Web tools, open source Java APIs, and command line tools
  - First release, TC 2007; new release, TC 2008
- Links to publications (click on Documentation at TC Web site)
- New release TC 2008 adds functionality, creates a new training set from MEDLINE subset, ST documents from Meta
- Intend to facilitate research (e.g., ST documents, stopwords)
- Java system developed by Chris Lu and authorized by Allen Browne; Willie Rogers, Collaborator.



# Example of Command Line

> mlt2007 -i:"<infilename>" -o:"<outfilename>" -t:TIAB

> jdi2007 -i:"<infilename>" -o:"<outfilename>" -of:on~all

> sti2007 -i:"<infilename>" -o:"<outfilename>"  
-of:can~genf -of:can~gngm -of:can~nusq



# Statistics for TC Releases

TC 2007 release training set:

4,093 journals

1,378,597 MEDLINE documents indexed 1999-2002

303,942, unique words in TI/AB

TC 2008 release training set:

5,212 journals

1,999,012 MEDLINE documents indexed 2005-2007

397,393 unique words in TI/AB

Lu, Chris J.; Humphrey, Susanne M.; Browne, Allen C. A method for verifying a vector-based text classification system. In: American Medical Informatics Association 2008 Annual Symposium Proceedings (Washington, DC AMIA 2008). Washington, DC, November 8-12, 2008. Forthcoming.



# Challenges

Normalization of counts in training set

- Word count (high frequency words)

- Document count for JDs (journals with many documents)

Thresholds in applications

ST documents for ST-JD vectors for word-ST vectors used in STI

- are fewer words for ST document better?

- ambiguity reflected in ST assignments –

- should words in an ST document belong to only one semantic group?

Stopwords

Some JD issues

- Obstetrics and Gynecology as separate JDs

- Evolution of JDs

Need for testing suite



# JAMA Topic Collections

- Published studies in *JAMA* and *Archives* journals are categorized according to Topic Collections terms at <http://pubs.ama-assn.org/collections/>



The screenshot displays the JAMA & ARCHIVES website interface. At the top, the logo "JAMA & ARCHIVES" is prominent on the left, and a search bar labeled "SEARCH ALL JOURNALS:" is on the right. Below the logo, a navigation menu includes links for HOME, SUBSCRIBE, E-MAIL ALERTS, TOPIC COLLECTIONS, CARE, PHYSICIAN JOBS, CONTACT US, and HELP. The main content area is titled "JAMA & Archives Journals Collections" and provides an overview of the collections, including a list of medical topics such as Aging/ Geriatrics, Anesthesia, Cardiovascular System, Complementary and Alternative Medicine, Critical Care/ Intensive Care Medicine, Dentistry/ Oral Medicine, and Dermatology. A "Content Access" sidebar on the right offers options like Sign in/out, Activate online subscription, and Register for E-mail Alerts.

**JAMA & ARCHIVES**

SEARCH ALL JOURNALS:  GO  
GO TO ADVANCED SEARCH >

HOME | SUBSCRIBE | E-MAIL ALERTS | TOPIC COLLECTIONS | CARE | PHYSICIAN JOBS | CONTACT US | HELP

Institution: National Institute of Health | My Account | Access Rights | Sign In

### JAMA & Archives Journals Collections

The *JAMA* & *Archives* Collections include all the *JAMA* & *Archives* Journals Topic Collection terms with articles indexed across all issues of *JAMA* & *Archives* Journals from January 1998 forward.

- [Aging/ Geriatrics](#)
- [Anesthesia](#)
- Cardiovascular System**
  - [Arrhythmias](#)
  - [Cardiac Diagnostic Tests](#)
  - [Cardiovascular Disease/ Myocardial Infarction](#)
  - [Congenital Heart Defects](#)
  - [Congestive Heart Failure/ Cardiomyopathy](#)
  - Cardiovascular Interventions**
    - [Pacemakers/ Defibrillators](#)
    - [Revascularization](#)
    - [Thrombolysis](#)
    - [Cardiovascular Interventions, Other](#)
  - [Venous Thromboembolism](#)
  - [Cardiovascular System, Other](#)
- [Complementary and Alternative Medicine](#)
- Critical Care/ Intensive Care Medicine**
  - [Adult Critical Care](#)
  - [Pediatric/ Neonatal Critical Care](#)
- [Dentistry/ Oral Medicine](#)
- Dermatology**
  - Dermatologic Disorders**
    - [Acne](#)
    - [Alopecia](#)
    - [Bites and Stings](#)
    - [Bullous Diseases](#)

**Content Access**

- Sign in/out
- Activate online subscription
- One-time access
- Individual subscriptions
- Institutional subscriptions
- Register for E-mail Alerts

NLM  
J.D.E. WSD

# Pediatric Subspecialty Collections

- Editors categorize published studies in the journal *Pediatrics* according to subspecialties similar to JDs at <http://pediatrics.aapublications.org/collections>

## PEDIATRICS

OFFICIAL JOURNAL OF THE AMERICAN ACADEMY OF PEDIATRICS

Home My Pediatrics Journal Information Current Issue Past Issues Subscriptions & Services Contact Us

Institution: Nat Library of Medicine | Sign In via User Name/Password

### *PEDIATRICS* Subspecialty Collections

Each of the following collections is a topic-specific archive of studies published in *PEDIATRICS* from January 1997 to the present. The categories are those used by the editors, and have been refined over a period of over 25 years. Within each category, further links have been included to extend your exploration of the topic area. Finally, automated *eSearches* of Medline and *PEDIATRICS* have been created for nearly every Collection, to speed further research.

Should you have comments or questions about our Subspecialty Collections, please [email us](#). We are especially interested in knowing whether presenting the journal's material in this manner is helpful to our readers.

(The numbers in parentheses show the number of articles currently in each collection.)

Adolescent Medicine (221)	Allergy & Dermatology (285)	Asthma (141)
Blood (167)	Computers (12)	Dentistry & Otolaryngology (120)
Developmental/Behavior (46)	Emergency Medicine (150)	Endocrinology (252)
Gastrointestinal Tract (240)	Genetics & Dysmorphology (172)	Genitourinary Tract (173)
Heart & Blood Vessels (262)	History (1)	Infectious Disease & Immunity (1704)
Journalology (20)	Miscellaneous (298)	Musculoskeletal System (93)
Neurology & Psychiatry (446)	Nutrition & Metabolism (588)	Office Practice (1943)
Ophthalmology (63)	Premature & Newborn (1985)	Radiology (9)
Respiratory Tract (240)	Statistics (11)	Surgery (127)
Therapeutics & Toxicology (489)	Tumors (80)	

American Academy of Pediatrics  
DEDICATED TO THE HEALTH OF ALL CHILDREN™



# Science Subject Collections

- Editors categorize articles in the journal *Science* according to fields under life sciences, physical sciences, and other subjects at <http://www.sciencemag.org/cgi/collection#clicked>

SCIENCE SUBJECT COLLECTIONS	
▼LIFE SCIENCES	▼PHYSICAL SCIENCES
<a href="#">Anatomy, Morphology, Biomechanics</a> (116 Articles)	<a href="#">Astronomy</a> (1797 Articles)
<a href="#">Anthropology</a> (797 Articles)	<a href="#">Atmospheric Science</a> (1401 Articles)
<a href="#">Biochemistry</a> (1601 Articles)	<a href="#">Chemistry</a> (2777 Articles)
<a href="#">Botany</a> (893 Articles)	<a href="#">Computers, Mathematics</a> (740 Articles)
<a href="#">Cell Biology</a> (2459 Articles)	<a href="#">Engineering</a> (280 Articles)
<a href="#">Development</a> (940 Articles)	<a href="#">Geochemistry, Geophysics</a> (2381 Articles)
<a href="#">Ecology</a> (2624 Articles)	<a href="#">Materials Science</a> (1072 Articles)
<a href="#">Epidemiology</a> (330 Articles)	<a href="#">Oceanography</a> (724 Articles)
<a href="#">Evolution</a> (1419 Articles)	<a href="#">Paleontology</a> (834 Articles)
<a href="#">Genetics</a> (1957 Articles)	<a href="#">Physics</a> (2217 Articles)
<a href="#">Immunology</a> (1266 Articles)	<a href="#">Physics, Applied</a> (994 Articles)
<a href="#">Medicine, Diseases</a> (3095 Articles)	<a href="#">Planetary Science</a> (1096 Articles)
<a href="#">Microbiology</a> (1040 Articles)	
<a href="#">Molecular Biology</a> (1453 Articles)	▼OTHER SUBJECTS
<a href="#">Neuroscience</a> (2541 Articles)	<a href="#">Economics</a> (155 Articles)
<a href="#">Pharmacology, Toxicology</a> (175 Articles)	<a href="#">Education</a> (664 Articles)
<a href="#">Physiology</a> (360 Articles)	<a href="#">History and Philosophy of Science</a> (447 Articles)
<a href="#">Psychology</a> (633 Articles)	<a href="#">Science and Business</a> (305 Articles)
<a href="#">Virology</a> (393 Articles)	<a href="#">Science and Policy</a> (3174 Articles)
	<a href="#">Sociology</a> (207 Articles)



Can do now:

**<Cardiology journals>** AND inflammation

But what if:

C-reactive protein and other circulating markers of inflammation  
in the prediction of coronary heart disease.

***N Engl J Med*** (not a Cardiology journal)

Better:

**<Cardiology specialty>** AND inflammation

Intersect specialties:

**<Cardiology specialty>** AND

**<Allergy and Immunology specialty>**

Retrieves:

The inflammation hypothesis and its potential relevance to statin  
therapy.

***Am J Cardiol***





## Can create word-MH vectors

*surgical*-MH vector sorted by score (complete vector has 19,764 MHs)

***Fibrin Tissue Adhesive*** 0.8742

***Hemostasis, Surgical*** 0.8714

***Postoperative Complications*** 0.8547

...

***Decompression, Surgical*** 0.6062 (rank 191)

...

*decompression*-MH vector sorted by score (complete vector has 19,764 MHs)

***Decompression, Surgical*** 0.9780

***Odontoid Process*** 0.9517

***Nerve Compression Syndromes*** 0.9492

...

Index “*surgical decompression*”

**1 *Decompression, Surgical*** 0.7921

**2 *Nerve Compression Syndromes*** 0.7866



# Text Categorization research based on JD vector similarity

Experiment trying 16 words. Sample results of word-MH vectors displaying top-scoring MH for each word:

<u>Word</u>	<u>Top-scoring MH</u>
abattoirs	Meat 0.8671 ...
cardiomyopathy	Cardiomyopathy, Congestive 0.9874 ...
decompression	Decompression, Surgical 0.9780 ...
congestive	Heart Failure, Congestive 0.9763 ...
diabetes	Diabetes Mellitus 0.9734 ...
diabetic	Diabetes Mellitus, Type II 0.9422 ...
failure	Peptidyl-Dipeptidase A 0.7750 ...
heart	Myocardial Diseases 0.9397 ...
intraoperative	Intraoperative Care 0.9202 ...
lymphedema	Lymphedema 0.9519 ...
mellitus	Diabetes Mellitus 0.9542 ...
neuropathy	Autonomic Nervous System Diseases 0.7512 ...
radiotherapy	Radiotherapy, Adjuvant 0.9388 ...
schizophrenia	Schizophrenia 0.9982 ...
surgical	Fibrin Tissue Adhesive 0.8742 ...
transporting	Carrier Proteins 0.8453 ...



# NLM People

## LHC

Allen Browne

Chris Lu

Willie Rogers

Mehmet Kayaalp

Dina Demner

Tom Rindflesch

Aurélie Névéol

Lan Aronson

Jim Mork

Anantha Bangalore

Guy Divita

Karen Thorn

Sonya Shooshan

## LO

Nancy Cox

Esther Baldinger

