Text Categorization

By

Susanne M. Humphrey

Lexical Systems Group

National Library of Medicine

12-18-2007

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 biomedical disciplines.
- Used for indexing journals *per se*
- Assigned by human indexer to the 4100 journals used in TC
- Found in lsi2007.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

What are Journal Descriptors (JDs)?

- Examples of information from lsi2007.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)?

- lsi2007.xml produces
  List of Journals Indexed for MEDLINE (LJI)
  ftp://nlmpubs.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

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 lsi2007.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

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

$$= 0.275691$$

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

$$= \frac{\text{no. of docs in training set in which TI/AB word kidney co-occurs with JD Nephrology}}{\text{no. of docs in training set in which the word kidney occurs in TI/AB}}$$

$$= 0.140088$$

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
```

- A JD score is average of JD score for word kidney and JD 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 JD 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 …
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 JD score for words in TI/AB

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

```
PMID - 17910645
TI   - Kidney transplantation in infants and small children.
SO   - Pediatr Transplant. 2007 Nov;11(7):702-8.
```

```
1|0.092475|Transplantation
2|0.065228|Pediatrics
3|0.051550|Nephrology
4|0.023945|Hematology
5|0.021809|Urology
```

- How was score for Pediatrics calculated?

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

PMID - 17910645
TI - Kidney transplantation in infants and small children.
SO - Pediatr Transplant. 2007 Nov;11(7):702-8.

```
1|0.092475|Transplantation
2|0.065228|Pediatrics
3|0.051550|Nephrology
4|0.023945|Hematology
5|0.021809|Urology
```

- Score for Pediatrics is average of score for Pediatrics for words kidney, transplantation, infants, children (last two boost score for Pediatrics).

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

```
PMID - 15215477
TI   - Pediatric renal-replacement therapy--coming of age.
SO   - New 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):

JD Scores	Journal Descriptors
…	…
0.140088	**Nephrology**
…	…
0.000460	**Psychiatry**
…	…
0.000308	**Psychopharmacology**
…	…
0.080848	**Transplantation**
…	…

Word-JD vector

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

JD Scores	Journal Descriptors
…	…
0.223750	**Nephrology**
…	…
0.000856	**Psychiatry**
…	…
0.000429	**Psychopharmacology**
…	…
0.095716	**Transplantation**
…	…

Word-JD vector

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

JD Scores	Journal Descriptors
…	…
0.000000	Nephrology
…	…
0.314520	Psychiatry
…	…
0.067470	Psychopharmacology
…	…
0.000153	Transplantation
…	…

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, $WORD_i$ and $WORD_j$, the similarity between them may be defined as

$$COSINE(WORDi,WORDj) = \frac{\sum_{k=1}^{t}(WJD_{ik} \cdot WJD_{jk})}{\sqrt{\sum_{k=1}^{t}(WJD_{ik})^2 \cdot \sum_{k=1}^{t}(WJD_{jk})^2}}$$

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$_i$ and a document, DOC$_j$, the similarity between them may be defined as

$$COSINE(WORDi, DOCj) = \frac{\sum_{k=1}^{t}(WJD_{ik} \cdot DJD_{jk})}{\sqrt{\sum_{k=1}^{t}(WJD_{ik})^2 \cdot \sum_{k=1}^{t}(DJD_{jk})^2}}$$

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_i and DOC_j, the similarity between them may be defined as

$$COSINE(DOCi, DOCj) = \frac{\sum\limits_{k=1}^{t} (DJD_{ik} \cdot DJD_{jk})}{\sqrt{\sum\limits_{k=1}^{t} (DJD_{ik})^2 \cdot \sum\limits_{k=1}^{t} (DJD_{jk})^2}}$$

Text Categorization research based on JD vector similarity

- JD vector similarity between pairs of 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.

- JD vector similarity between word and document

  Detecting outlier (blooper) MeSH indexing terms for a document. Terms can be MTI recommendations, e.g., Stupor for “unresponsive cells” or humanly-assigned, e.g., Deception for “cheater genotypes.”

Text Categorization research based on JD vector similarity

- Automatically generate stopword list

- JD vector similarity between pairs of words in training set

- Comparing THE to:

THE	1.0
AND	0.9998
FOR	0.9977
WITH	0.9970
…	
COMLEX	0.0028

- 303,942 words in training set

Text Categorization research based on JD vector similarity

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	

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

- 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 tables representing JD indexing of each of the 304,000 words in the training set.

- System also has word-ST tables representing ST indexing of each training set word.

- Thus, STI of text can be performed exactly the same way as JDI of text. Each ST score for a text is the average of that STs score for each word in the text.

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.

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).

TC Tools

- Most of the JDI and STI in this talk can be done by using the TC Web Tools at 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 Java APIs, command line tools, and Web tools
  - First release, TC 2007
- Links to publications (click on Documentation at TC Web site)
- In coming months, we will be adding to functionality of TC Web tools as well as incorporate the ability to create new training sets.
- JAVA system developed by Chris Lu and authorized by Allen Browne.

Text Categorization research based on JDI

- Evaluating JDI. Take random sample of recent MEDLINE documents, JDI them, and use as criterion of success whether the native JD of the document is ranked highly in the JDI result.

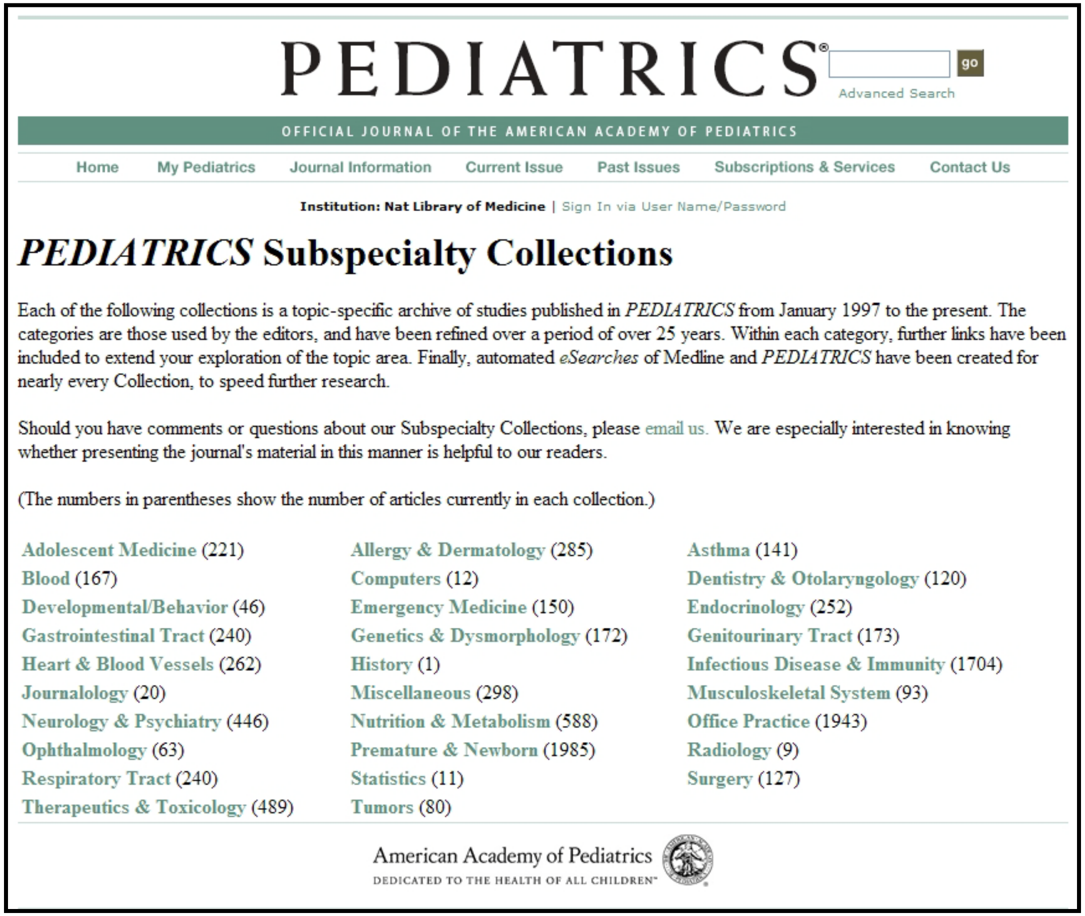
- Specialty subsets. Do JDI indexing of MEDLINE documents from general medical journal like *New England Journal of Medicine* or *JAMA* in order to partition them into specialty subsets based on JDs.

- JDI is word-based. Make it phrase-based by extracting phrases from the training set, and creating phrase-JD vectors. 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.

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

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.

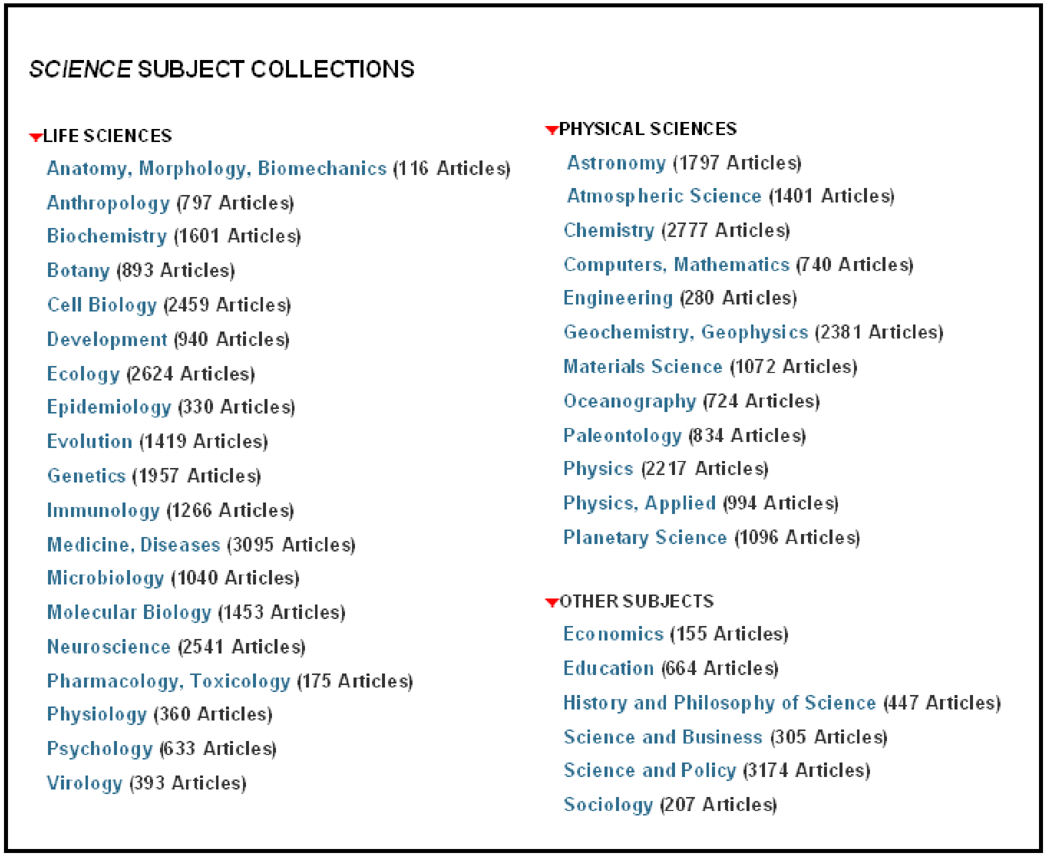
(The numbers in parentheses show the number of articles currently in each collection.)

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

- Anatomy, Morphology, Biomechanics (116 Articles)
- Anthropology (797 Articles)
- Biochemistry (1601 Articles)
- Botany (893 Articles)
- Cell Biology (2459 Articles)
- Development (940 Articles)
- Ecology (2624 Articles)
- Epidemiology (330 Articles)
- Evolution (1419 Articles)
- Genetics (1957 Articles)
- Immunology (1266 Articles)
- Medicine, Diseases (3095 Articles)
- Microbiology (1040 Articles)
- Molecular Biology (1453 Articles)
- Neuroscience (2541 Articles)
- Pharmacology, Toxicology (175 Articles)
- Physiology (360 Articles)
- Psychology (633 Articles)
- Virology (393 Articles)

▼PHYSICAL SCIENCES

- Astronomy (1797 Articles)
- Atmospheric Science (1401 Articles)
- Chemistry (2777 Articles)
- Computers, Mathematics (740 Articles)
- Engineering (280 Articles)
- Geochemistry, Geophysics (2381 Articles)
- Materials Science (1072 Articles)
- Oceanography (724 Articles)
- Paleontology (834 Articles)
- Physics (2217 Articles)
- Physics, Applied (994 Articles)
- Planetary Science (1096 Articles)

▼OTHER SUBJECTS

- Economics (155 Articles)
- Education (664 Articles)
- History and Philosophy of Science (447 Articles)
- Science and Business (305 Articles)
- Science and Policy (3174 Articles)
- Sociology (207 Articles)

Text Categorization research based on JD vector similarity

- Automatic indexing using MH-JD vectors from the training set. If you have MH-JD vectors and word-JD vectors, you can create word-MH vectors, and do MH indexing of words, and if you can do MH indexing of words, you can do MH indexing of text (phrases, MEDLINE documents, etc., consisting of words) by averaging the score of each MH across all the words in the text.

- Problem: each word-MH vector would be very long – 20,000 MH scores for each word, compared to 122 JDs for word-JD vector.