Indexing Large Text Collections Using the Vector Space Model

Kevin C. O'Kane, Ph.D.
Computer Science Professor Emeritus
University of Northern Iowa
http://www.cs.uni.edu/~okane
kc.okane@gmail.com

Additional Resources

Information Retrieval, C. J. van RIJSBERGEN
http://www.dcs.gla.ac.uk/Keith/Preface.html

Introduction to Information Retrieval, C. D. Manning, P. Raghavan and H. Schütze

http://nlp.stanford.edu/IR-book/information-retrieval-book.html

Modern Information Retrieval, R. Baeza-Yates and B. Ribeiro-Neto

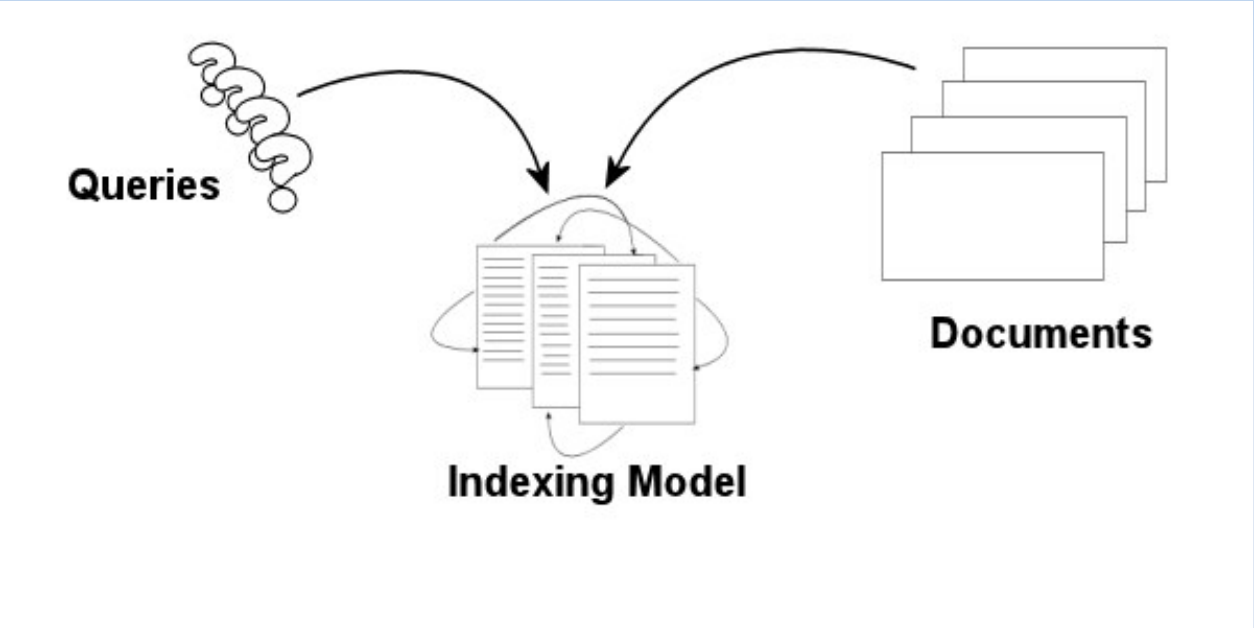
http://www.sims.berkeley.edu/~hearst/irbook

Someone, Somewhere Is Searching for It

What is Information Retrieval?

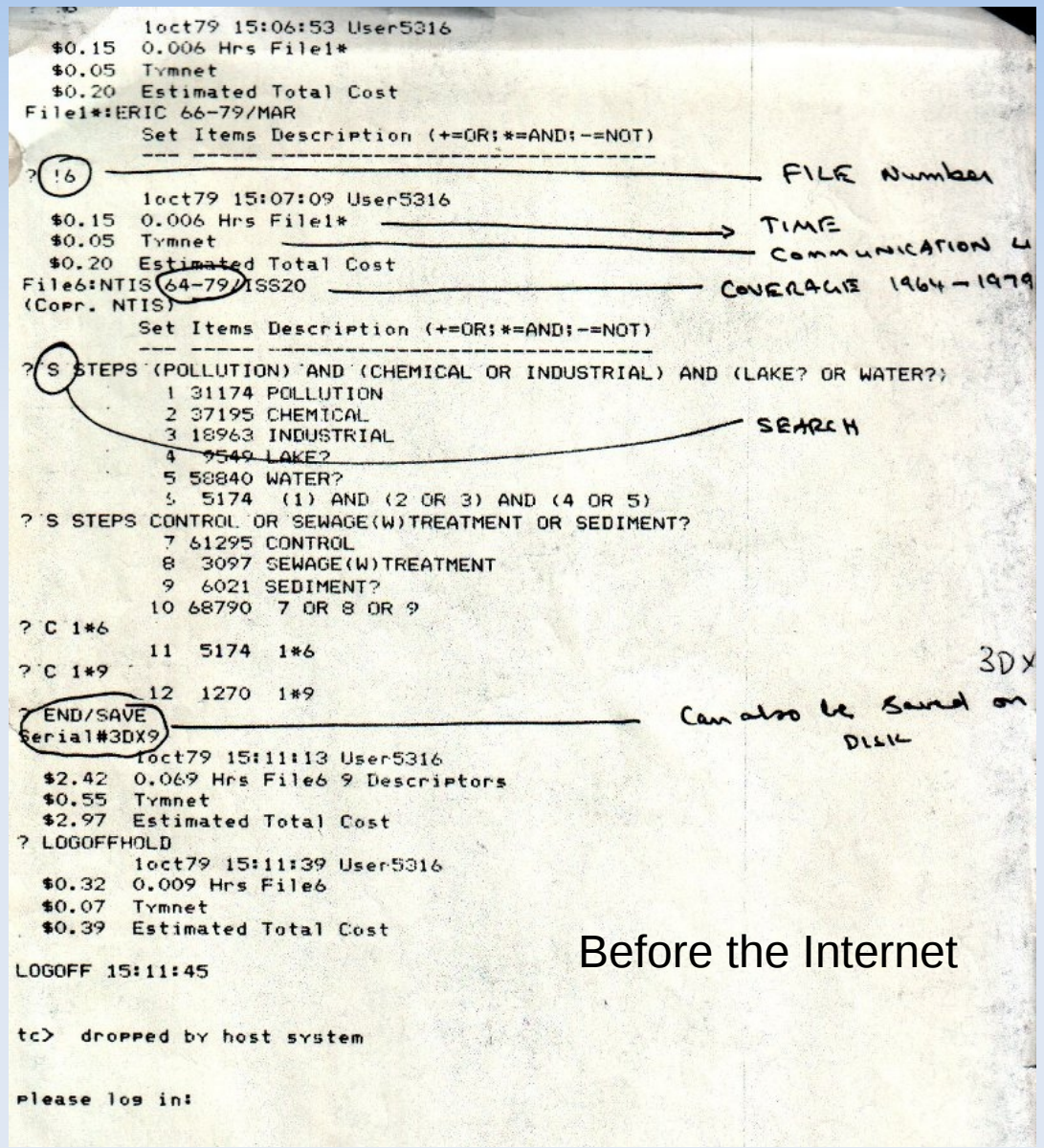
Information retrieval is the art of identifying similarities between queries and collections of objects in a database. In nearly all cases, the objects found will not be identical to the query but will only resemble it in some fashion.

Basic concept



In its simplest form, an information retrieval system maps objects and queries into a data model which makes it possible to to calculate the similarity between queries and the objects.

1979 Tymnet Search



```
        1oct79 15:06:53 User5316
  $0.15  0.006 Hrs File1*
  $0.05  Tymnet
  $0.20  Estimated Total Cost
File1*:ERIC 66-79/MAR
        Set Items Description (+=OR;*=AND;-=NOT)
        --- ----- ----------------------------
? !6
        1oct79 15:07:09 User5316
  $0.15  0.006 Hrs File1*
  $0.05  Tymnet
  $0.20  Estimated Total Cost
File6:NTIS 64-79/ISS20
(Copr. NTIS)
        Set Items Description (+=OR;*=AND;-=NOT)
        --- ----- ----------------------------
? S STEPS (POLLUTION) AND (CHEMICAL OR INDUSTRIAL) AND (LAKE? OR WATER?)
          1 31174 POLLUTION
          2 37195 CHEMICAL
          3 18963 INDUSTRIAL
          4  9549 LAKE?
          5 58840 WATER?
          6  5174  (1) AND (2 OR 3) AND (4 OR 5)
? S STEPS CONTROL OR SEWAGE(W)TREATMENT OR SEDIMENT?
          7 61295 CONTROL
          8  3097 SEWAGE(W)TREATMENT
          9  6021 SEDIMENT?
         10 68790  7 OR 8 OR 9
? C 1*6
         11  5174  1*6
? C 1*9
         12  1270  1*9
? END/SAVE
Serial#3DX9
        1oct79 15:11:13 User5316
  $2.42  0.069 Hrs File6 9 Descriptors
  $0.55  Tymnet
  $2.97  Estimated Total Cost
? LOGOFFHOLD
        1oct79 15:11:39 User5316
  $0.32  0.009 Hrs File6
  $0.07  Tymnet
  $0.39  Estimated Total Cost

LOGOFF 15:11:45

tc>  dropped by host system

please log in:
```

FILE Number

TIME

Communication

COVERAGE 1964 – 1979

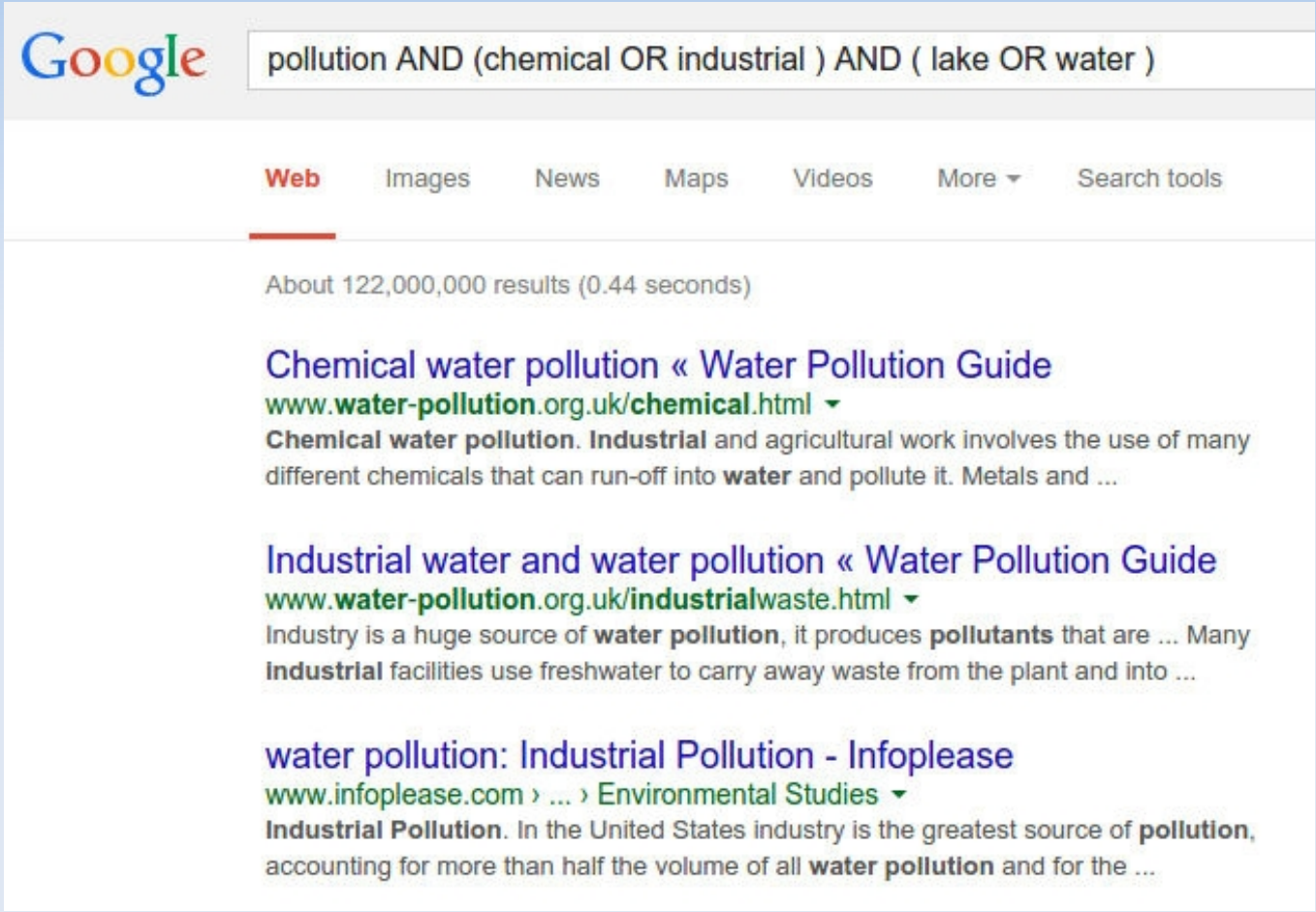
SEARCH

3DX

Can also be saved on DISK

Before the Internet

Google Search 2015



http://google.about.com/od/searchingtheweb/qt/booleantip.htm

Specialized Searching

Other forms of text searching

```
>gi|2695846|emb|Y13255.1|ABY13255 Acipenser baeri mRNA for immunoglobulin heavy chain,
TGGTTACAACACTTTCTTCTTTCAATAACCACAATACTGCAGTACAATGGGGATTTTAACAGCTCTCTGTATAATAATGA
CAGCTCTATCAAGTGTCCGGTCTGATGTAGTGTTGACTGAGTCCGGACCAGCAGTTATAAAGCCTGGAGAGTCCCATAAA
CTGTCCTGTAAAGCCTCTGGATTCACATTCAGCAGCGCCTACATGAGCTGGGTTCGACAAGCTCCTGGAAAGGGTCTGGA
ATGGGTGGCTTATATTTACTCAGGTGGTAGTAGTACATACTATGCCCAGTCTGTCCAGGGAAGATTCGCCATCTCCAGAG
ACGATTCCAACAGCATGCTGTATTTACAAATGAACAGCCTGAAGACTGAAGACACTGCCGTGTATTACTGTGCTCGGGGC
GGGCTGGGGTGGTCCCTTGACTACTGGGGGAAAGGCACAATGATCACCGTAACTTCTGCTACGCCATCACCACCGACAGT
GTTTCCGCTTATGGAGTCATGTTGTTTGAGCGATATCTCGGGTCCTGTTGCTACGGGCTGCTTAGCAACCGGATTCTGCC
TACCCCCGCGACCTTCTCGTGGACTGATCAATCTGGAAAAGCTTTT
```

If you submit the above to NCBI BLAST (National Center for Biotechnology Information), it will conduct a search of the **nr** database of nucleotide sequences. The result is a ranked list of sequences from the data base based on their similarity to the query sequence. Sequences found whose similarity score exceeds a threshold are displayed.

Example BLAST Search Result

```
>gb|U17058.1|LOU17058 Lepisosteus osseus Ig heavy chain V region mRNA, partial

 Score =  151 bits (76),  Expect = 4e-33
 Identities = 133/152 (87%), Gaps = 0/152 (0%)
 Strand=Plus/Plus

Query  242  TGGGTGGCTTATATTTACTCAGGTGGTAGTAGTACATACTATGCCCAGTCTGTCCAGGGA  301
            ||||||||| ||||||||| | | ||| || | |||||||||| ||||||||||||||||
Sbjct  4    TGGGTGGCGTATATTTACACCGATGGGAGCAATACATACTATTCCCAGTCTGTCCAGGGA  63

Query  302  AGATTCGCCATCTCCAGAGACGATTCCAACAGCATGCTGTATTTACAAATGAACAGCCTG  361
            |||||| |||||||||||||| ||||||| |    |||||| ||||| |||| |||||||
Sbjct  64   AGATTCACCATCTCCAGAGACAATTCCAAGAATCAGCTGTACTTACAGATGAGCAGCCTG  123

Query  362  AAGACTGAAGACACTGCCGTGTATTACTGTGC  393
            |||||||||||||||| ||||||||||||||
Sbjct  124  AAGACTGAAGACACTGCTGTGTATTACTGTGC  155
```

Bars indicate exact matches.

Zipf's Law

Zipf's Law states that the frequency ordered rank of a term in a document collection times its frequency of occurrence is approximately equal to a constant:

Frequency * Rank ~= constant

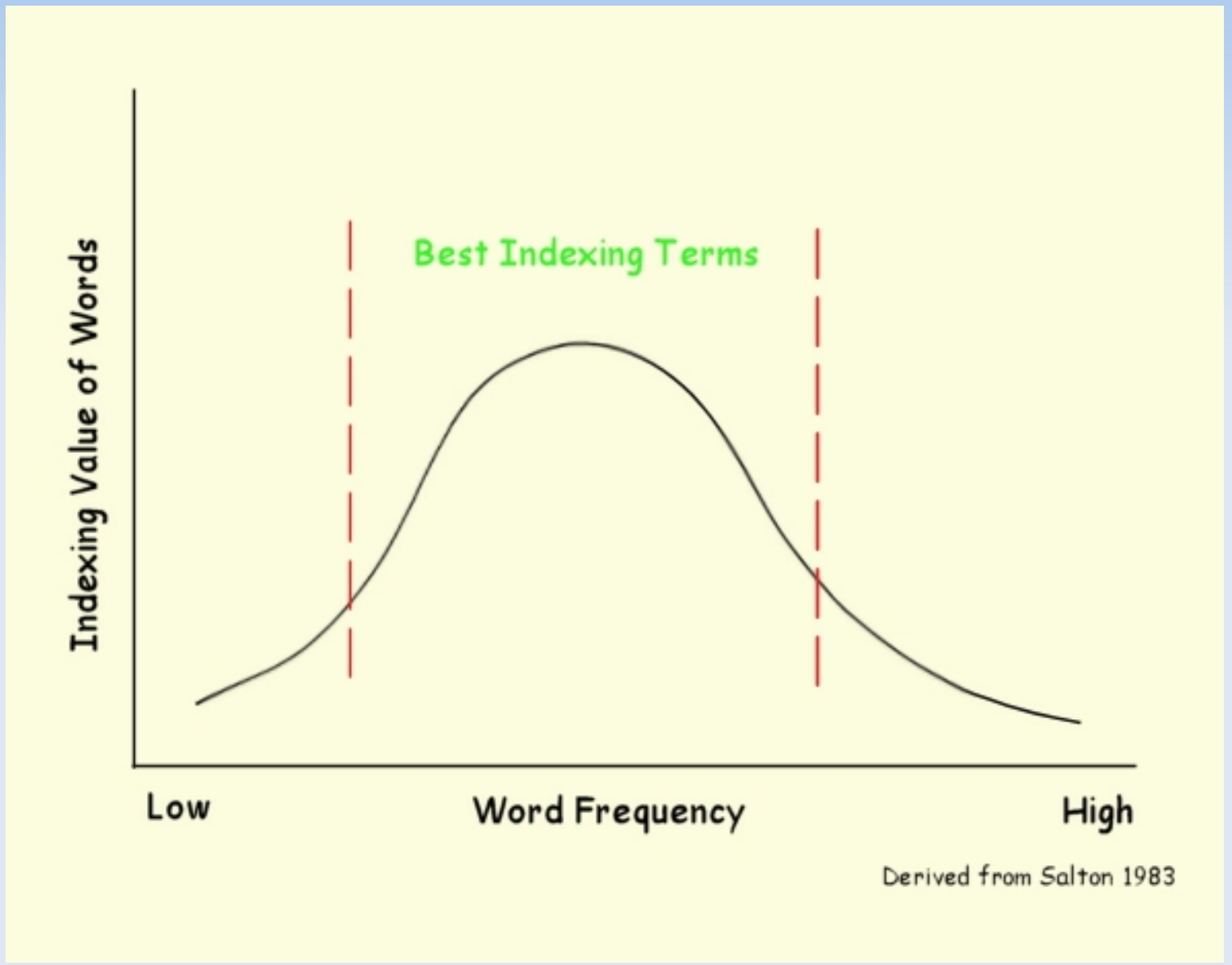
Frequency is the total number of times some term *k* occurs.

Rank is the position number of the term when the terms have been sorted by *Frequency*. The most frequently occurring term is rank 1, the second most frequently occurring term is rank 2 and so forth. Program calculates Zipf's constants.

Zipf's Constants for OHSUMED Database

369 of	381 treatment	385 who	439 normal	481 liver
425 in	393 been	391 this	435 infection	484 [news]
613 the	403 during	383 coronary	437 function	486 can
733 and	414 comment	389 used	442 which	490 effect
502 a	426 has	393 new	441 studies	493 bone
555 with	423 comments	398 protein	444 this	497 type
561 to	436 see	403 using	447 treated	497 its
514 the	428 or	403 use	448 not	494 had
529 for	437 effects	405 two	450 cases	
426 patients	428 in	411 syndrome	454 immunodeficiency	
429 on	425 disease	417 therapy	458 heart	
462 by	416 clinical	422 report	460 myocardial	
498 is	399 blood	426 virus	463 cardiac	
432 letter	408 at	430 growth	465 acid	
417 was	410 cells	434 children	469 primary	
407 from	408 effect	439 role	473 following	
425 were	417 acute	433 determine	472 diagnosis	
437 an	423 to	433 rat	472 receptor	
414 human	408 associated	422 editorial	475 analysis	
365 study	397 be	420 pulmonary	476 management	
379 we	400 case	423 cancer	476 gene	
396 that	400 studied	425 artery	479 carcinoma	
393 after	396 between	426 an	479 may	
386 as	394 renal	430 ventricular	478 we	
394 are	389 chronic	435 disease	482 care	
396 have	383 factor	437 response	484 patients	
370 cell	380 patient	439 activity		

Good Indexing Terms



IS&R Data Preliminaries

Initial processing of the data:

- Convert the data into a more easily managed format
- Remove stop words
- Remove very high and very low frequency words
- Reduce words to stems
- Construct document vectors
- Construct term vectors
- Construct document frequency vector (number of docs each term used in)
- Construct dictionary vector (number of times each word occurs)

OHSUMED TREC Text Collection

The OHSUMED test collection is a set of abstracts from MEDLINE, the on-line medical information database, consisting of titles and/or abstracts from 270 medical journals over a five-year period (1987-1991). Total number of abstracts: 293,857

Total vocabulary: ~120,000 unique words (after stemming, rejection of words less than three or longer than 25 characters in length, and words beginning with numbers)

A small number of words have very high frequencies of occurrence compared to the remainder of the file.

At the low end of the frequency spectrum, there are about 72,000 words that occur 5 or fewer times (60% of the total number of words).

If we eliminate words with total frequency of occurrence of 5 or less or greater than 40,000 (the top ranking 101 words), this results in a candidate vocabulary of about 64,000 words.

Medline Format

```
STAT- MEDLINE
MH    Acetaldehyde/*ME
MH    Buffers
MH    Catalysis
MH    HEPES/PD
MH    Nuclear Magnetic Resonance
MH    Phosphates/*PD
MH    Protein Binding
MH    Ribonuclease, Pancreatic/AI/*ME
MH    Support, U.S. Gov't, Non-P.H.S.
MH    Support, U.S. Gov't, P.H.S.
TI    The binding of acetaldehyde to the active site of ribonuclease: alterations in catalytic ...
AB    Ribonuclease A was reacted with [1-13C,1,2-14C]acetaldehyde
      and sodium cyanoborohydride in the presence or absence
      of 0.2 M phosphate. After several hours of incubation
      at 4 degrees C (pH 7.4) stable acetaldehyde-RNase adducts
      were formed, and the extent of their formation was
      similar regardless of the presence of phosphate. Although
      the total amount of covalent binding was comparable
      in the absence or presence of phosphate, this active
      site ligand prevented the inhibition of enzymatic activity
      seen in its absence. This protective action of phosphate
      diminished with progressive ethylation of RNase, indicating
      that the reversible association of phosphate with the
      active site lysyl residue was overcome by the irreversible
      ...
```

Modified Text (1 line)

148710 130 chromaticity and achromaticity. evidence for a functional differentiation in visual field defects. In the visual field defects of 10 patients who had suffered lesions in the postgeniculate part of the primary visual projection, red-green discrimination and achromatic target detection was tested. in addition, 8 of these patients were tested for detection of red and green targets. targets were presented on a low photopic achromatic background, so that the red and green targets differed from the background both in intensity and in wavelength, whereas the achromatic target differed in intensity only. six patients showed evidence of discriminating between red and green targets, 5 patients could also detect the colour targets, but none could detect the achromatic one that was presented at the same retinal position. these results imply that wavelength and intensity information are treated differentially, and suggest that these patients possess residual colour-opponent channels that subserve the defective part of the visual field.

Stop List Words

eight
either
eleven
else
elsewhere
empty
enough
etc
even
ever
every
everyone
everything
everywhere
except
few
fifteen
fifty
fill
find
fire
first
five
for

Filtered OSUMED (no stop words)

7128 6 clonidine alcohol withdrawal pilot study differential symptom responses following iv clonidine The effects iv bolus clonidine 15 mgkg withdrawal symptoms studied 7 male alcoholics compared 7 normal controls clonidine reduced subjective objective measures withdrawal reached significance scores depression muscle tension comparison controls alcoholics withdrawal showed reduced sedative effects clonidine

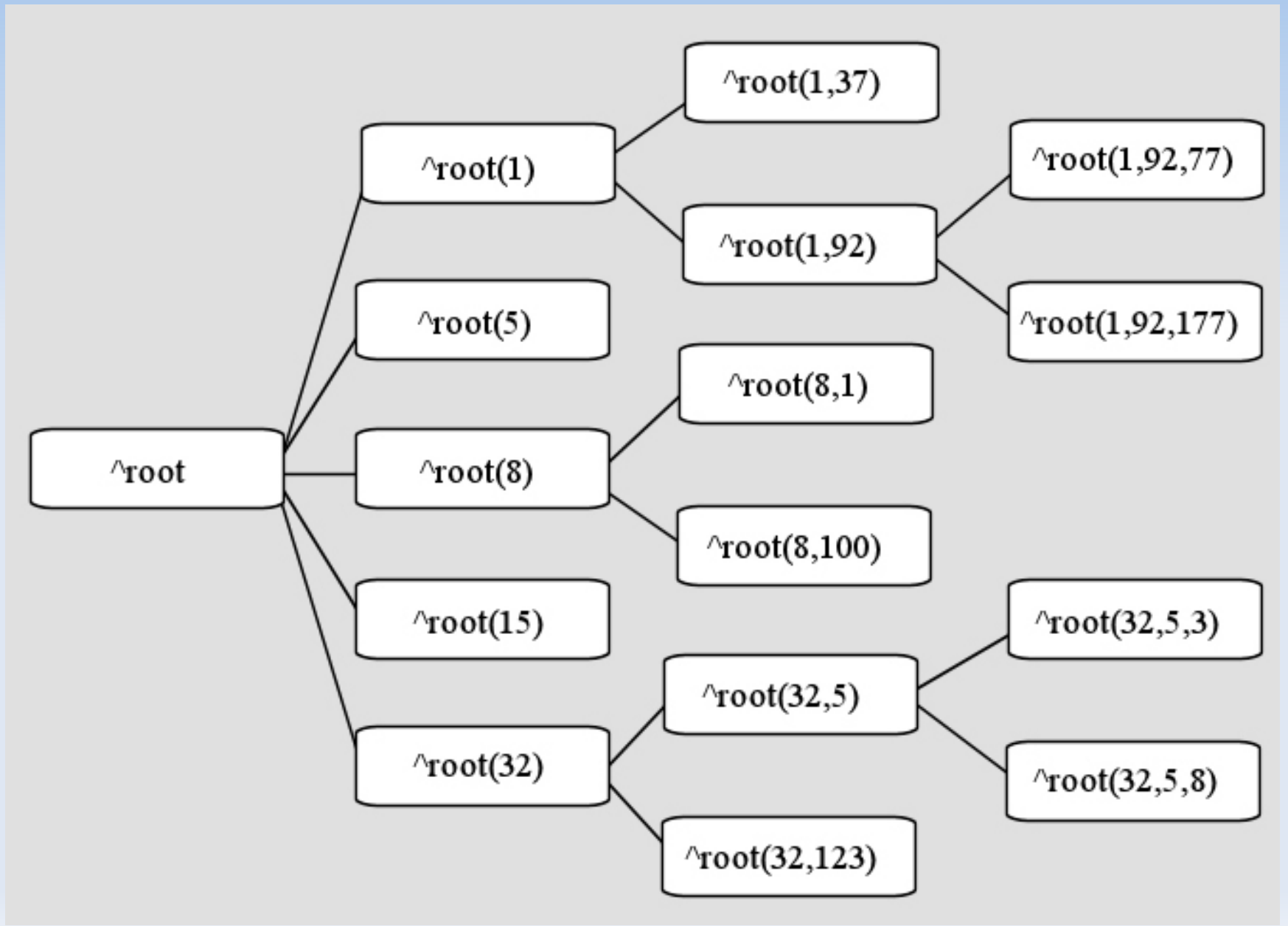
Word Stems

```
{ "nned",  4, "n",    1, fin,},
{ "rated", 5, "rate", 4, fin,},
{ "cated", 5, "cate", 4, fin,},     /* sophisticated */
{ "iated", 5, "iate", 4, fin,},     /* associated */
{ "mated", 5, "mate", 4, fin,},     /* animated automated */
{ "gated", 5, "gate", 4, fin,},     /* aggregated */
{ "lated", 5, "late", 4, fin,},     /* accumulated */
{ "bated", 5, "bate", 4, fin,},     /* debated */
{ "nated", 5, "nate", 4, fin,},     /* dominated donated */
{ "ated",  4, "te",   2, fin,},     /* automated */
{ "eted",  4, "ete",  3, fin,},     /* competed deleted */
{ "ated",  4, "ate",  3, fin,},     /* alternated */
{ "kted",  4, "kt",   2, fin,},     /* marketed */
{ "oted",  4, "ote",  3, fin,},     /* promoted */
{ "uted",  4, "ute",  3, fin,},     /* distributed */
{ "sted",  4, "ste",  3, fin,},     /* wasted */
```

Stemmed Text

7128 6 clonidine alcohol withdraw pilot study differential symptom respons follow clonidine the effect bolu clonidine mgkg withdraw symptom study male alcoholic compare norm control clonidine reduce subject objective measure withdraw reach significance score depress muscle tension comparison control alcoholic withdraw show reduce sedative effect clonidine

Mumps Global Arrays



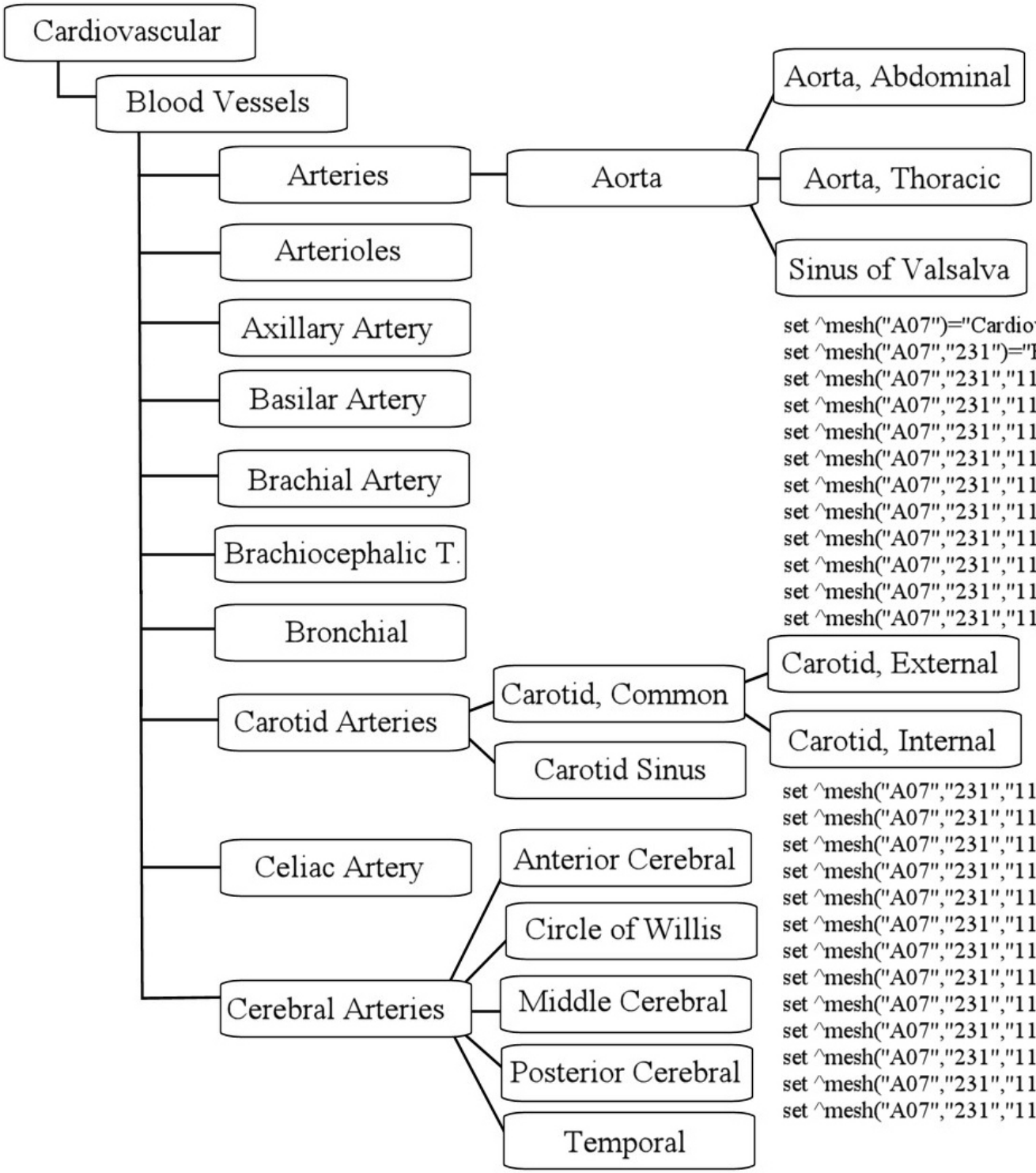
Creating Global Arrays

```
set ^root(1,37)=1
set ^root(1,92,77)=2
set ^root(1,92,177)=3
set ^root(5)=4
set ^root(8,1)=5
set ^root(8,100)=6
set ^root(15)=7
set ^root(32,5)=8
set ^root(32,5,3)=9
set ^root(32,5,8)=10
set ^root(32,123)=11
```

String Indices

```
set ^lab(1234,"hct","05/10/2008",38)=""
set ^lab(1234,"hct","05/12/2008",42)=""
set ^lab(1234,"hct","05/15/2008",35)=""
set ^lab(1234,"hct","05/19/2008",41)=""
```

Note: sometimes the indices themselves are the data and nothing ("") is actually stored at the node. That is the case here where the last index is the test result.




```
set ^mesh("A07")="Cardiovascular System"
set ^mesh("A07","231")="Blood Vessels"
set ^mesh("A07","231","114")="Arteries"
set ^mesh("A07","231","114","056")="Aorta"
set ^mesh("A07","231","114","056","205")="Aorta, Abdominal"
set ^mesh("A07","231","114","056","372")="Aorta, Thoracic"
set ^mesh("A07","231","114","056","847")="Sinus of Valsalva"
set ^mesh("A07","231","114","060")="Arterioles"
set ^mesh("A07","231","114","085")="Axillary Artery"
set ^mesh("A07","231","114","106")="Basilar Artery"
set ^mesh("A07","231","114","139")="Brachial Artery"
set ^mesh("A07","231","114","145")="Brachiocephalic Trunk"
set ^mesh("A07","231","114","158")="Bronchial Arteries"
set ^mesh("A07","231","114","186")="Carotid Arteries"
set ^mesh("A07","231","114","186","200")="Carotid Artery, Common"
set ^mesh("A07","231","114","186","200","210")="Carotid Artery, External"
set ^mesh("A07","231","114","186","200","230")="Carotid Artery, Internal"
set ^mesh("A07","231","114","186","456")="Carotid Sinus"
set ^mesh("A07","231","114","207")="Celiac Artery"
set ^mesh("A07","231","114","228")="Cerebral Arteries"
set ^mesh("A07","231","114","228","100")="Anterior Cerebral Artery"
set ^mesh("A07","231","114","228","351")="Circle of Willis"
set ^mesh("A07","231","114","228","550")="Middle Cerebral Artery"
set ^mesh("A07","231","114","228","700")="Posterior Cerebral Artery"
set ^mesh("A07","231","114","228","868")="Temporal Arteries"
```

Basic Matrices & Vectors

The *document-term matrix* *^doc(docNbr,word)* giving the number of times the term word occurs in document *docNbr*.

A *term-document matrix* *^index(word,DocNbr).*

A *dictionary vector* *^dict(word)* giving the number of times term word occurs in the total collection.

A *document frequency vector* *^df(word)* giving the number of documents term word occurs in the collection.

Document Vectors

	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10
Doc 1	5	0	3	0	1	0	2	1	0	0

Document Vectors

^doc(101,"cystic")=11.02
^doc(101,"cystic-hygroma")=2.197225
^doc(101,"cystic-sclerosing")=1.609438
^doc(101,"favouare")=6.9
^doc(101,"fever")=5.1
^doc(101,"fever-inflammatory")=1.791759
^doc(101,"hygroma")=13.8
^doc(101,"hygroma-sclerosing")=1.791759
^doc(101,"inflammatory")=8.66
^doc(101,"intracystic")=13.8
^doc(101,"last")=4.06
^doc(101,"lead")=4.26
^doc(101,"ok")=13.8
^doc(101,"origin")=4.26
^doc(101,"overy")=5.51
^doc(101,"pyogene")=6.9
^doc(101,"regress")=8.82
^doc(101,"scar")=5.51
^doc(101,"sclerose")=11.6
^doc(101,"side-effect")=6.2
^doc(101,"streptococcus")=5.8

Term Vectors

	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Doc 6	Doc 7	Doc 8	Doc 9	Doc 10
Word1	1	0	2	0	0	1	3	0	0	1

Term Vectors

^index("inflammatory",101)=8.66
^index("inflammatory",308)=12.99
^index("inflammatory",309)=4.33
^index("inflammatory",37)=4.33
^index("inflammatory",430)=4.33
^index("inflammatory",44)=8.66
^index("inflammatory",555)=4.33
^index("inflammatory",564)=4.33
^index("inflammatory",72)=4.33
^index("inflammatory",881)=4.33
^index("inflammatory",890)=4.33
^index("inflammatory",976)=8.66
^index("inflammatory",981)=4.33

Document Frequency Vector

^df("achlorhydric")=1
^df("achromatic")=1
^df("achromatic-green")=7
^df("achromatic-intensities")=7
^df("achromatic-target")=19
^df("achromatic-wavelength")=5
^df("achromaticity")=1
^df("acid")=50
^df("acid-base")=1
^df("acid-base-hypertensive")=6
^df("acid-base-spontaneously")=7
^df("acid-base-young")=7
^df("acid-cholesterol")=1
^df("acid-contain")=1
^df("acid-independent")=1
^df("acid-independent-bile")=7
^df("acid-like")=1
^df("acid-like-gamma-aminobutyric")=5
^df("acid-lod")=1
^df("acid-urea-hexadecyltrimethylammonium")=1

Dictionary Vector

^dict("achlorhydric")=1
^dict("achromatic")=4
^dict("achromatic-green")=7
^dict("achromatic-intensities")=7
^dict("achromatic-target")=19
^dict("achromatic-wavelength")=5
^dict("achromaticity")=1
^dict("acid-base")=3
^dict("acid-base-hypertensive")=6
^dict("acid-base-spontaneously")=7
^dict("acid-base-young")=7
^dict("acid-cholesterol")=1
^dict("acid-contain")=1
^dict("acid-independent")=1
^dict("acid-independent-bile")=7
^dict("acid-like")=1
^dict("acid-like-gamma-aminobutyric")=5
^dict("acid-lod")=1
^dict("acid-urea-hexadecyltrimethylammonium")=1

Vector Space Model

The vector space model views computer generated document vectors as describing a hyperspace in which the number of dimensions (axes) is equal to the number of indexing terms. This approach was originally proposed by G. Salton:

Salton, G.; McGill, M.J., *Introduction to Modern Information Retrieval*, New York: McGraw Hill; 1983.

Salton, G., The state of retrieval system evaluation, *Information Processing & Management*, 28(4): 441-449; 1992.

Videos of a conference on Salton's work and on SMART are available at:

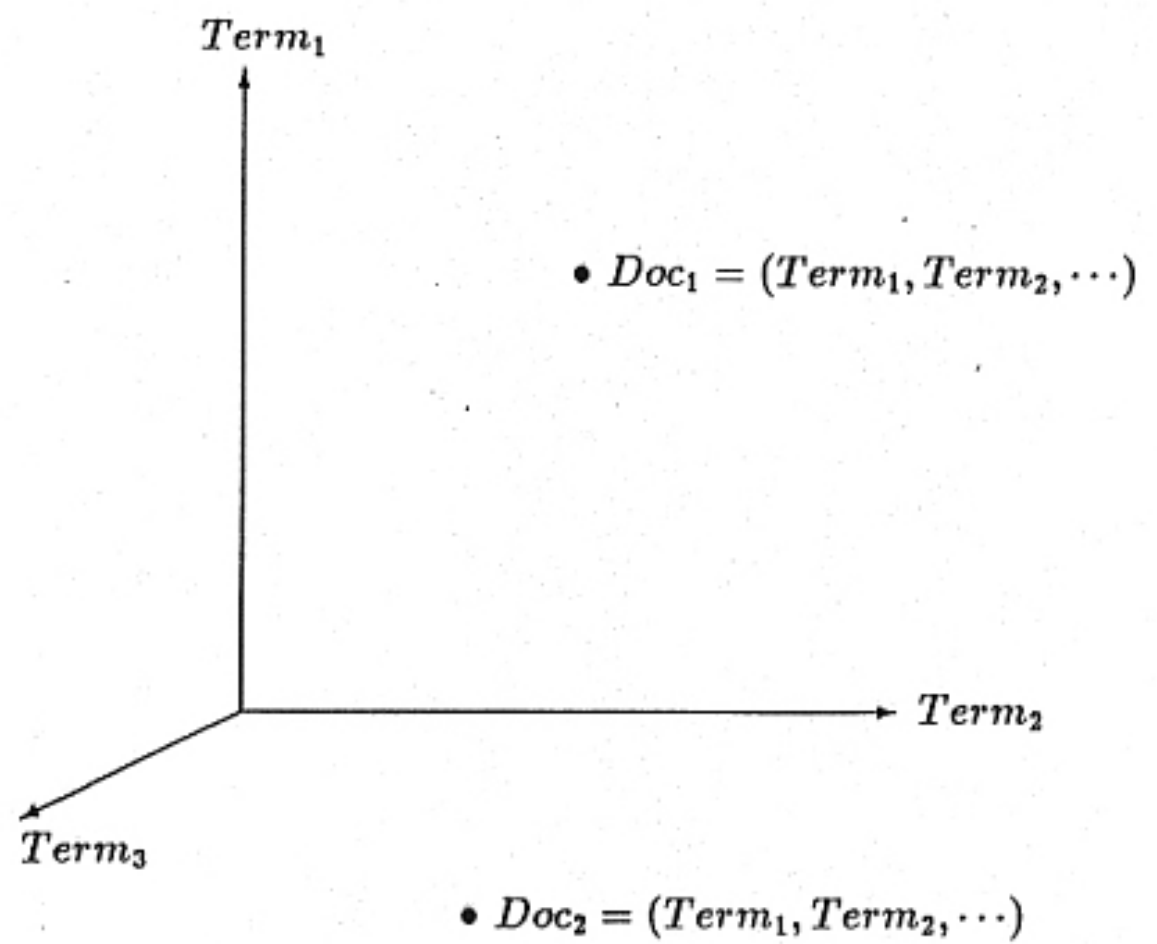
http://www.open-video.org/details.php?videoid=7053

Vector Space Model

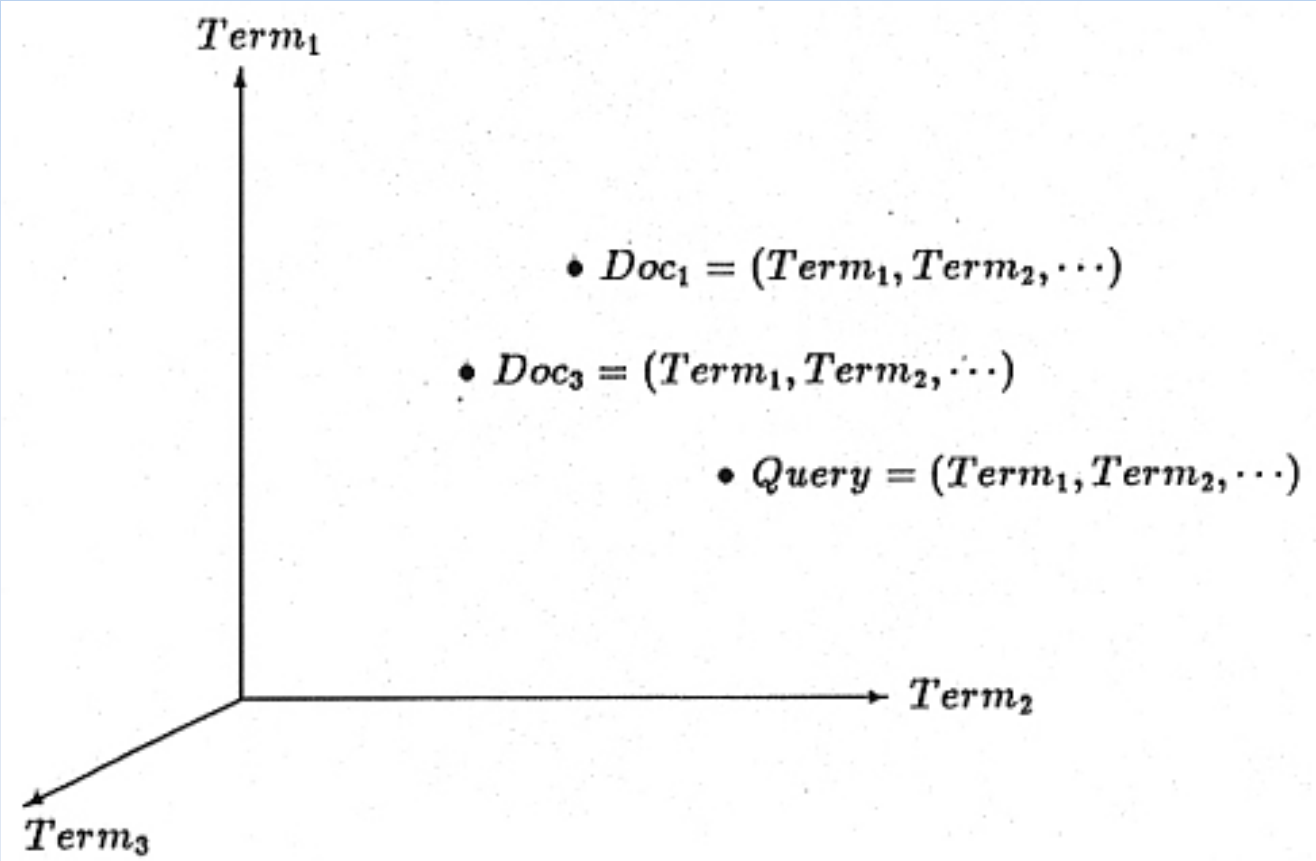
Each document vector describes a point in a space defined by the distance along the axis associated with each document term proportional to the term's importance or significance in the document being represented.

Queries are also portrayed as vectors that define points in the document hyperspace. Documents whose points in the hyperspace lie within an adjustable envelope of distance from the query vector point are retrieved.

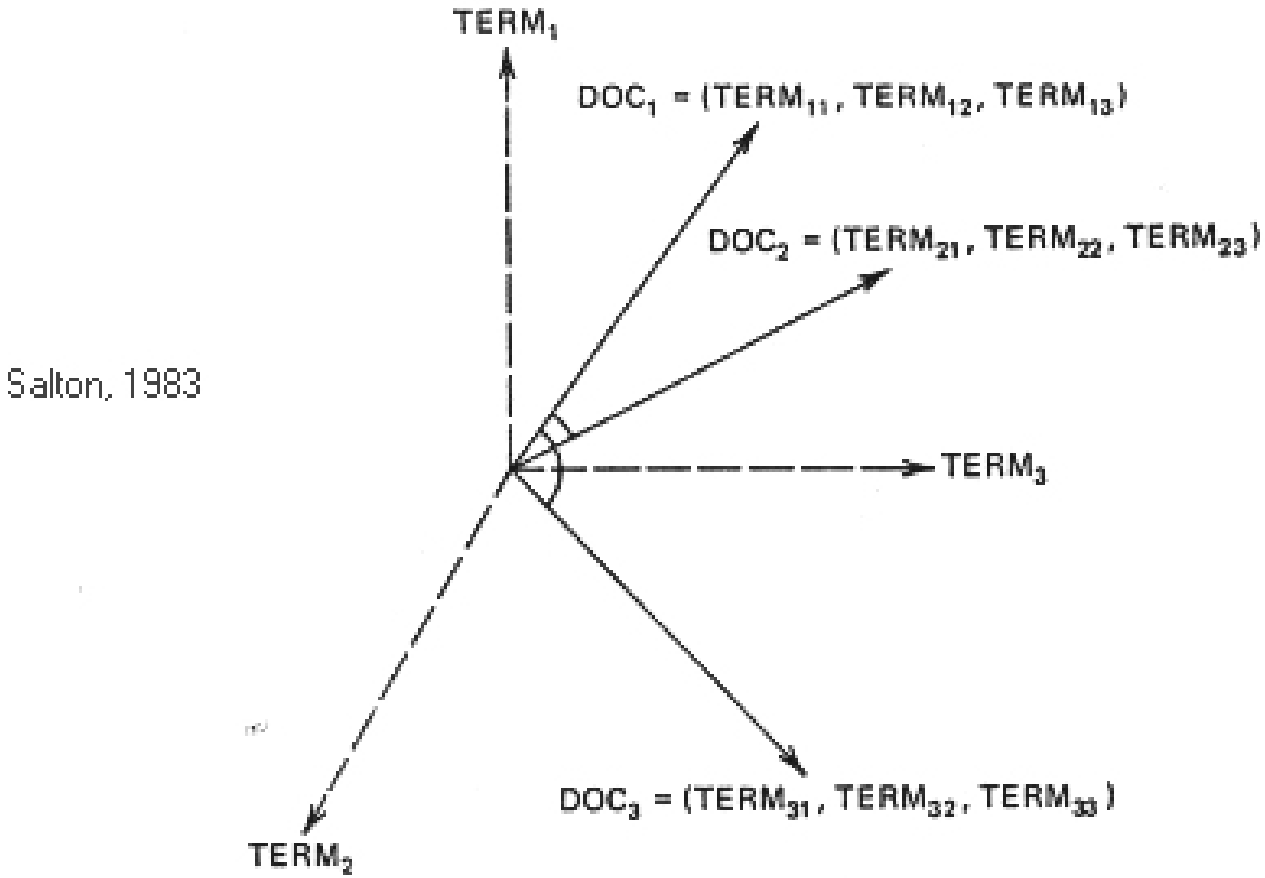
Document Hyperspace



Query and document vectors



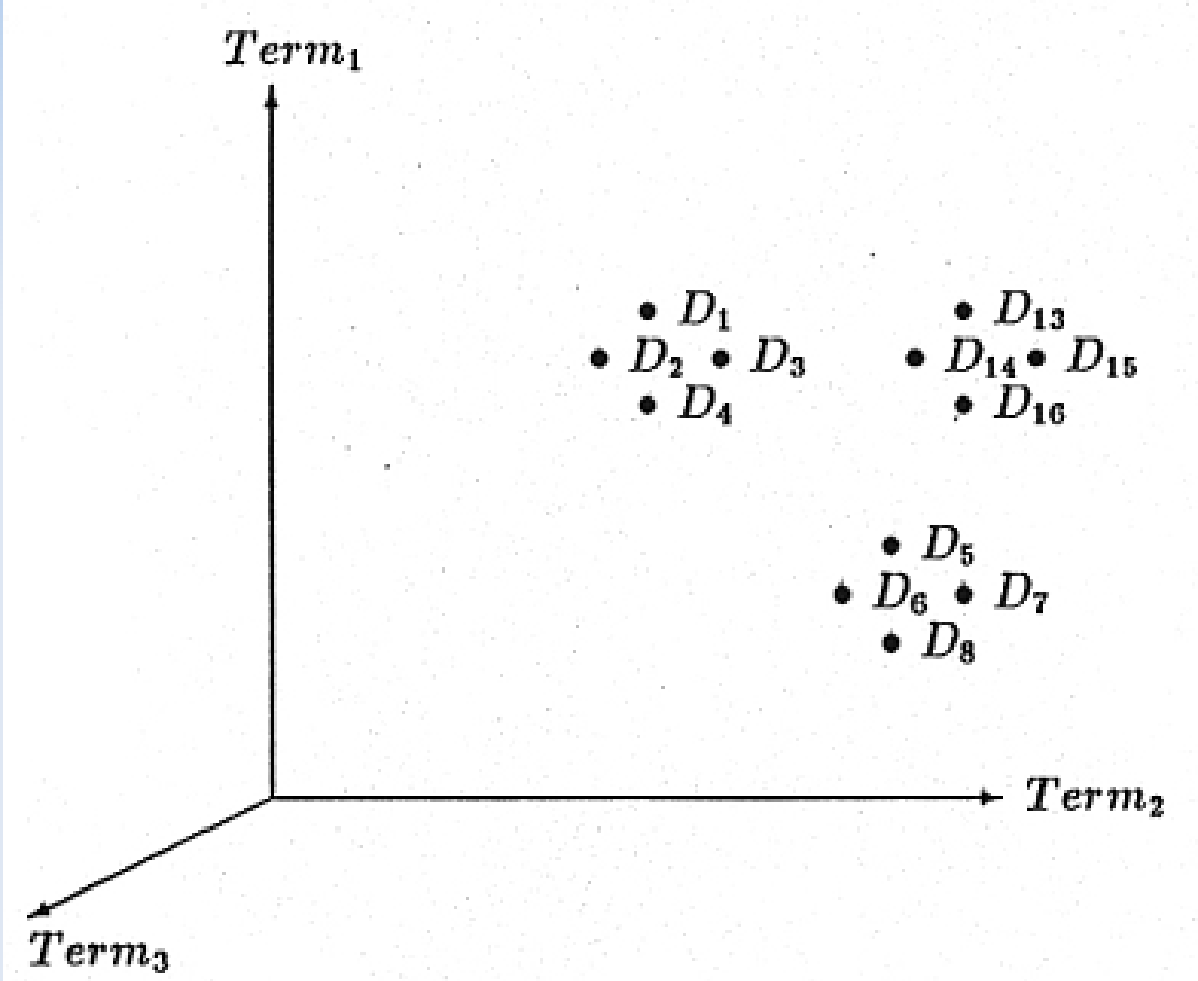
Salton's Cosine Representation



Salton, 1983

Figure 4-2 Vector representation of document space.

Document Clustering



Some Similarity functions

$$\mathrm{Sim}_1(Doc_i, Doc_j) = \frac{2\left[\sum_{k=1}^{t} (Term_{ik} \cdot Term_{jk})\right]}{\sum_{k=1}^{t} Term_{ik} + \sum_{k=1}^{t} Term_{jk}}$$

$$\mathrm{Sim}_2(Doc_i, Doc_j) = \frac{\sum_{k=1}^{t} Term_{ik} \cdot Term_{jk}}{\sum_{k=1}^{t} Term_{ik} + \sum_{k=1}^{t} Term_{jk} - \sum_{k=1}^{t} (Term_{ik} \cdot Term_{jk})}$$

$$\mathrm{Sim}_3(Doc_i, Doc_j) = \frac{\sum_{k=1}^{t} (Term_{ik} \cdot Term_{jk})}{\sqrt{\sum_{k=1}^{t} (Term_{ik})^2 \cdot \sum_{k=1}^{t} (Term_{jk})^2}}$$

Similarity functions

$$\mathrm{Sim}_4(Doc_i, Doc_j) = \frac{\sum_{k=1}^{t} (Term_{ik} \cdot Term_{jk})}{min(\sum_{k=1}^{t} Term_{ik}, \sum_{k=1}^{t} Term_{jk})}$$

$$\mathrm{Sim}_5(Doc_i, Doc_j) = \frac{\sum_{k=1}^{t} min(Term_{ik}, Term_{jk})}{\sum_{k=1}^{t} Term_{ik}}$$

Similarity Examples

Doc_i = (3,2,1,0,0,0,1,1)

Doc_j = (1,1,1,0,0,1,0,0)

$Sim_1(Doc_i,Doc_j)$= (2*6)/(8+4) -> 1

$Sim_2(Doc_i,Doc_j)$= (6)/(8+4-6) -> 1

$Sim_3(Doc_i,Doc_j)$= (6)/SQRT(16*4) -> 0.75

$Sim_4(Doc_i,Doc_j)$= 6/4 -> 1.5

$Sim_5(Doc_i,Doc_j)$= 3/8 -> 0.375

Example similarity coefficient calculations

Calculating Word Weights

Words used for indexing vary in their ability to indicate content and, thus, their importance as indexing terms.

Some words, such as *the*, *and*, *was* and so forth are worthless as content indicators and we eliminate them from consideration immediately.

Other words occur so infrequently that they are also unlikely to be useful as indexing terms.

Some words, however, with middle frequency of occurrence, are the best candidates as indexing terms.

The goal is to find a metric of the ability of a word to convey information.

Possible Word Weights

1000 documents; 29942 word instances, 563 distinct words

^doc(i,w)	Number times word w used in document i
^dict(w)	Number times word w used in total collection
^df(w)	Number of documents word w appears in
Wgt1	^doc(i,w)/(^dict(w)/^df(w))
Wgt2	^doc(i,w)*$zlog(NbrDocs/^df(w))+1
Wgt3	Wgt1*Wgt2+0.5\1

Word	^doc(i,w)	^dict(w)	^df(w)	Wgt1	Wgt2	Wgt3	MCA
[1] Death of a cult. (Apple Computer needs to alter its strategy) (column)							
apple	4	261	112	1.716	9.757	17	-1.1625
computer	4	706	358	2.028	5.109	10	-19.4405
mac	2	146	71	0.973	6.290	6	-0.0256
macintosh	4	210	107	2.038	9.940	20	-0.5855
strategy	2	79	67	1.696	6.406	11	-0.0592
[2] Next year in Xanadu. (Ted Nelson's hypertext implementations) Swaine, Michael.							
document	3	114	68	1.789	9.065	16	0.0054
operate	3	269	184	2.052	6.078	12	-2.1852
[3] WordPerfect. (WordPerfect for the Macintosh 2.0) (evaluation) Taub, Eric.							
edit	2	111	77	1.387	6.128	8	-0.0961
frame	2	9	7	1.556	10.924	17	0.0131
import	2	29	19	1.310	8.927	12	0.0998
macintosh	3	210	107	1.529	7.705	12	-0.5855
macro	3	38	24	1.895	12.189	23	0.1075
outstand	1	10	9	0.900	5.711	5	0.0168
user	4	861	435	2.021	4.330	9	-26.8094
wordperfect	8	24	8	2.667	39.627	106	0.1747

Inverse Document Frequency Weight

One of the simplest word weighting schemes to implement is the *Inverse Document Frequency* weight. The IDF weight is the measure of how widely distributed a term is in a collection.

Low IDF weights mean that the term is widely used while high weights indicate that the usage is more concentrated.

The IDF weight measures the weight of a term in the collection as a whole, rather than the weight of a term in a document.

Inverse Document Frequency Weight

The IDF weight for a term is it's weight in the *collection as a whole*.

In individual document vectors, the normalized frequency of occurrence of each term is multiplied by the IDF to give a weight for the term *in the particular document*.

Thus, a term with a high frequency but a low IDF weight could still be a highly weighted term in a particular document

On the other hand, a term with a low frequency but a high IDF weight could also be an important term in a given document.

Inverse Document Frequency Weight

$$\log_2\left(\frac{N}{DocFreq_w}\right)$$

where $DocFreq_W$ is the number of documents in which term W occurs and N is the total number of documents.

OSU MEDLINE IDF Weights

Tuning parameter thresholds other factors will produce variation in the results

```
0.189135 human
0.288966 and
0.300320 the
0.542811 with
0.737224 for
0.793466 was
0.867298 were
...
12.590849 actinomycetoma
12.590849 actinomycetomata
12.590849 actinomycoma
12.590849 actinomyosine
12.590849 actinoplane
12.590849 actinopterygii
12.590849 actinoxanthin
12.590849 actisomide
12.590849 activ
12.590849 activationin
```

Discrimination Coefficient Weights

The Discrimination coefficient measures the degree to which a term differentiates one document from another.

It is calculated based on the effect a term has on overall hyperspace density with and without a given term.

If the space density is greater when a term is removed from consideration, that means the term was making documents look less like one another (a good discriminator) while terms whose removal decreases the density are poor discriminators.

The discrimination values for a set of terms are similar to the values for the IDF weights but not exactly.

Calculating Discrimination Coefficients

The basic procedure calls for calculating the average of the pair-wise similarities between all documents in the space.

Then, for each term, the average of the pair-wise similarities of all the documents is calculated without that term.

The difference in the averages is the term discrimination value for the term.

When the average similarity increases when a word is removed, the word was a good discriminator - it made documents look less like one another.

On the other hand, if the average similarity decreased, the term was not a good discriminator since it made the documents look more like one another.

In practice, this is an expensive weight to calculate unless speed-up techniques are used.

Discrimination Coefficients

see:

http://www.cs.uni.edu/~okane/source/ISR/willett85.pdf

http://www.cs.uni.edu/~okane/source/ISR/crouch88.pdf

http://www.cs.uni.edu/~okane/source/ISR/salton83.pdf

Calculating Discrimination Coefficient with a Centroid Vector

$$\mathrm{Sim}_3(Doc_i, Doc_j) = \frac{\sum_{k=1}^{t} (Term_{ik} \cdot Term_{jk})}{\sqrt{\sum_{k=1}^{t} (Term_{ik})^2 \cdot \sum_{k=1}^{t} (Term_{jk})^2}}$$

$$\mathrm{Sim}_3(C, Doc_j) = \frac{\sum_{k=1}^{t} (C_k \cdot Term_{jk})}{\sqrt{\sum_{k=1}^{t} (C_k)^2 \cdot \sum_{k=1}^{t} (Term_{jk})^2}}$$

Average Similarity Using Centroid Vector

$$\text{AveSim} = \frac{1}{N} \sum_{j=1}^{N} \frac{\sum_{k=1}^{t} \left(C_k \cdot Term_{jk} \right)}{\sqrt{\sum_{k=1}^{t} \left(C_k \right)^2 \cdot \sum_{k=1}^{t} \left(Term_{jk} \right)^2}}$$

OSUMED Low Weights

idf	word	^dict	^df	^idf	^mca
4.01	constant	122	91	4.01	-606
4.01	operative	130	91	4.01	-698
4.01	respiratory	163	91	4.01	-1226
4.01	rest	169	91	4.01	-1244
4.02	basi	107	90	4.02	-535
4.02	evaluate	99	90	4.02	-520
4.02	fluid	192	90	4.02	-1761
4.02	undergo	132	90	4.02	-811
4.03	rang	120	89	4.03	-673
4.03	research	133	89	4.03	-1255
4.03	wall	208	89	4.03	-2018
4.03	young	128	89	4.03	-919
4.04	account	109	88	4.04	-586
4.04	disorder	148	88	4.04	-1233
4.04	head	213	88	4.04	-2246
4.04	main	110	88	4.04	-542
4.04	performance	155	88	4.04	-1041

OSUMED High Weights

idf word	^dict	^df	^idf	^mca
8.52 uninvolve	7	1	8.52	80
8.52 unna	5	1	8.52	41
8.52 ureteroureterostomy	7	1	8.52	145
8.52 urobilinogen	5	1	8.52	53
8.52 valverelate	7	1	8.52	97
8.52 vegetation	5	1	8.52	56
8.52 venomspecific	5	1	8.52	96
8.52 venoocclusive	5	1	8.52	174
8.52 ver	6	1	8.52	79
8.52 vermicular	6	1	8.52	132
8.52 vial	11	1	8.52	68
8.52 wax	5	1	8.52	41
8.52 woodchuck	8	1	8.52	202

Thesaurus & Phrase Construction

It is possible to find connections between terms based on their frequency of co-occurrence. Terms that co-occur frequently together are likely to be related and can indicate that the words may be synonyms or terms used to express a similar concept.

Thesaurus & Phrase Construction

For example, a strong relationship such as between the words artificial and intelligence in a computer science data base is due to the phrase artificial intelligence which names a branch of computing. In this case, the relationship is not that of a synonym. Similarly, in the medical data base terms such as circadian rhythm and vena cava and herpes simplex are concepts expressed as more than one term.

On the other hand words like synergism and synergistic, cyst and cystic, schizophrenia and schizophrenic, nasal and nose, and laryngeal and larynx are examples of synonym relationships.

In other cases, the relationship is not so tight so as to be a full synonym but express a categorical relationship such as anesthetic and halothane, analgesia and morphine, nitrogen and urea, and nurse and personnel

Thesaurus & Phrase Construction

Regardless of the relationship, a thesaurus table can be constructed giving a list of related terms. With this information it is then possible to:

• augment queries with related words to improve recall;

• combine multiple related infrequently occurring terms into broader, more frequently occurring categories terms; and,

• create middle frequency composite terms from otherwise unusable high frequency component terms.

Term-Term Matrix

First, we construct a square term-term correlation matrix which gives the frequency of co-occurrence of terms with one another.

Thus, if some term *A* occurs in 20 documents and if term *B* also occurs in these same documents, the term-term correlation matrix cell for row *A* and column *B* will have a value of 20.

A term-term correlation matrix's lower diagonal matrix is the same as the upper diagonal matrix since the relationship between term *A* and *B* is always the same as the relationship between term *B* and *A*.

The diagonal itself is the count of the number of documents in which the term occurs.

Term-Term Matrix

Calculating a complete term-term correlation matrix based on all documents in a large collection can be very time consuming.

In most cases, a term-term matrix potentially contains many billions of elements (the square of the number of vocabulary terms) summed over the entire collection of documents.

In practice, however, it is only necessary to sample part of the total collection.

That is, you can calculate a representative matrix by looking at every fifth, tenth or twentieth document, *etc*., depending on the total size of the collection.

Term-Term Matrix

The *term-term co-occurrence matrix,* also known as the *term connection matrix*, is an *N x N* square matrix, where *N* is the number of terms in the vocabulary, whose elements give the number of documents associated with each pair of terms. More formerly, it is:

$$TermTerm = DocTerm^{T} \cdot DocTerm$$

That is, the product of the document term matrix *DocTerm* with its transpose.

Term-Term matrix

In the *TermTerm* matrix, the rows and columns are term identifiers and the cells addressed are the number of documents which contained both terms. Here the elements of *DocTerm* are binary: a 0 indicates the term is not present and a 1 indicates the term is present.

Alternatively, in terms of the *DocTerm* matrix with elements d_{ij}, for document *i* and term *j*, the elements tt_{jk} for *terms j and k* of the *TermTerm* matrix can be calculated as:

$$tt_{jk} = d_{1j} d_{1k} + d_{2j} d_{2k} + d_{3j} d_{3k} .... d_{Nj} d_{Nk}$$

In other words, the similarity between term *i* and term *j* is the sum of the number of times they co-occur in a documents.

Term-Term Matrix

For example, is you have 3 three documents entitled:

Doc_1 = Pseudomonas-aeruginosa cystic-fibrosis.
Doc_2 = Pseudomonas-aeruginosa immune response
Doc_3 = Immune complexes in cystic-fibrosis

Then the document term matrix *DocTerm* and its transpose $DocTerm^T$ for the terms (column headings) *pseudomonas-aeruginosa, immune, cystic-fibrosis, response,* and *complexes* will be ---

Term-Term Matrix

$$DocTerm = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \end{pmatrix}$$

$$DocTerm^T = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

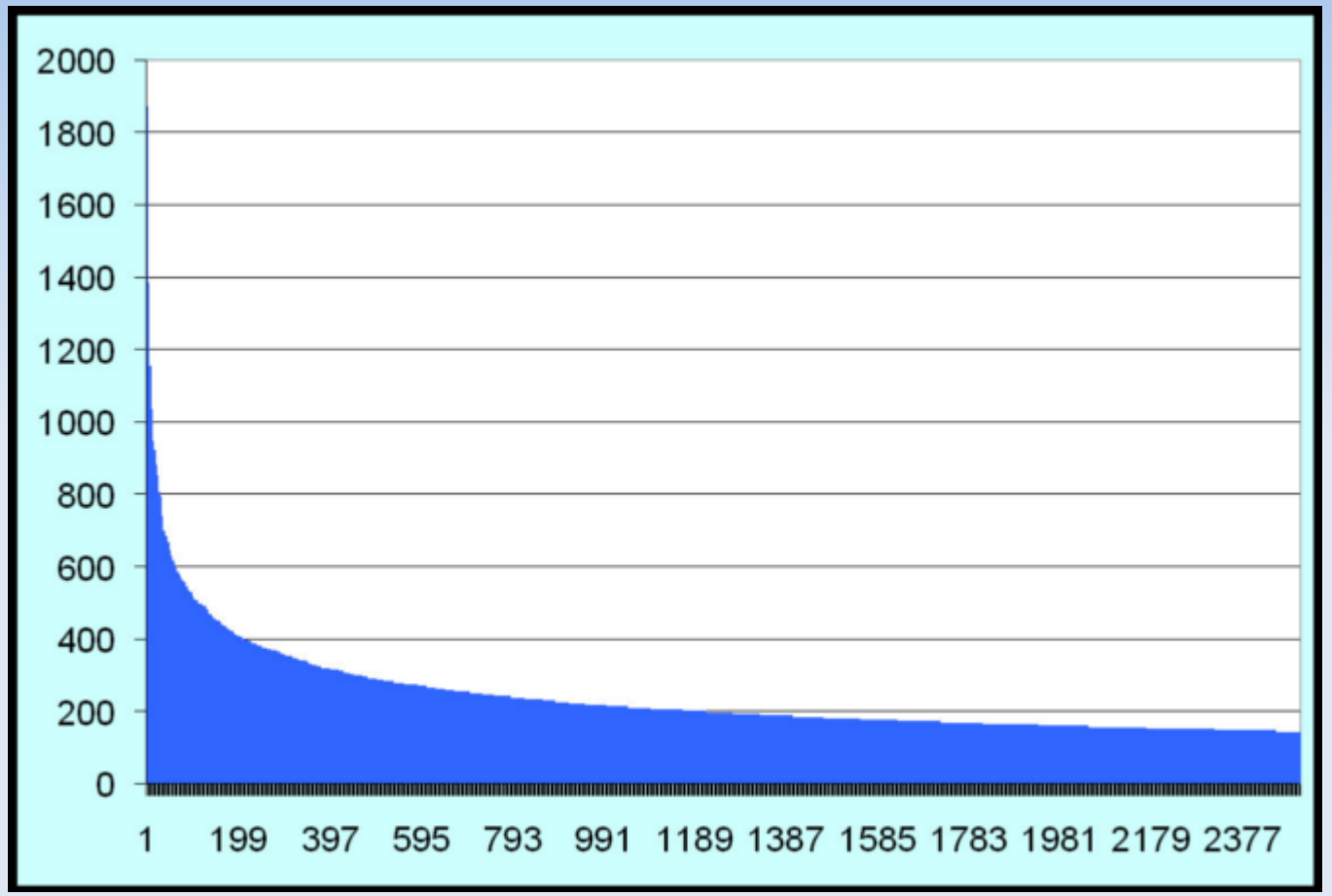
Doc_1 Pseudomonas-aeruginosa cystic-fibrosis
Doc_2 Pseudomonas-aeruginosa immune response
Doc_3 Immune complexes in cystic-fibrosis

$$TermTerm = DocTerm^T \cdot DocTerm = \begin{pmatrix} 2 & 1 & 1 & 1 & 0 \\ 1 & 2 & 1 & 1 & 1 \\ 1 & 1 & 2 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \end{pmatrix}$$

pseudomonas
immune
cystic-fibrosis
response
complexes

pseudomonas
immune
cystic-fibrosis
response
complexes

Term-Term Matrix



Frequency of co-occurrence (number of times two words co-occur together) for the entire OSU collection is plotted against rank. The vertical axis represents the number of times two terms co-occur and the horizontal axis gives the rank of the term pair when sorted by this frequency of co-occurrence. That is, the term pair that co-occur the most frequently (a bit less than 1200 times) appears first (leftmost), the term pair co-occurring next most frequently next, and so on.

Term-Term Similarity Matrix

Salton (Salton 1983) proposes the formula below where *t(i,k)* is the frequency of occurrence of term *k* in document *i* and *t(i,h)* is the frequency of occurrence of term *h* in document *I*.. The numerator is the sum of the co-occurrences of terms *k* and *h* and the denominator is the sum of the squares of the independent frequency of occurrences of the terms separately minus the frequency of co-occurrence.

Basically, if two terms never co-occur the result will be zero and if they co-occur always, the result will be one.

Unlike the raw term-term co-occurrence count, this method takes into account underlying word usage frequency. The basic term-term co-occurrence count shown above is biased in that it favors higher frequency terms. Some lower frequency terms may have usage profiles with one another that indicate a greater similarity.

$$similarity\left(Term_k, Term_h\right) = \frac{\sum_{i=1}^{n} t_{ik} t_{ih}}{\sum_{i=1}^{n} \left(t_{ik}\right)^2 + \sum_{i=1}^{n} \left(t_{ih}\right)^2 - \sum_{i=1}^{n} t_{ik} t_{ih}}$$

OSUMED Term-Term matrix

```
0.006 15 erythematosus lupu
0.006 6 endocardit infective
0.006 6 visualization visualize
0.006 7 abstract annu
0.006 7 hypothermic normothermic
0.007 6 autosomal inheritance
0.007 6 carbon monoxide
0.007 9 dismutase superoxide
0.007 9 dodecyl gel
0.008 10 erratum med
0.008 13 electrophoresis gel
0.008 17 aureus staphylococcus
0.008 9 abstract meet
0.008 9 annu meet
0.008 9 autosomal recessive
0.009 6 camera gamma
0.009 6 dysphagia swallow
0.009 6 multivariate univariate
0.010 33 erratum publish
0.010 6 corneum stratum
0.011 10 nitr oxide
0.011 11 carbon dioxide
0.012 6 pneumoniae streptococcus
0.012 8 antiinflammatory nonsteroid
0.014 6 arteriosus ductus
0.015 6 herniate herniation
0.016 6 chorionic gonadotropin
0.016 8 anorexia nervosa
0.018 7 herpe simplex
0.018 9 dodecyl electrophoresis
0.019 7 catalase dismutase
0.024 6 endoplasmic reticulum
0.033 7 aeruginosa pseudomona
0.034 6 coli escherichia
0.034 6 enzymelink immunosorbent
```

Position Specific Term-Term Matrix

Another modification to improve term-term detection involves retaining during document scanning the relative positions of the words with respect to one another in the collection. Then, when calculating the term-term matrix, proximity can be taken into account.

In DocTerm.mps a third level index was stored which is, for each distinct word in the document, a set of numbers indicating the distance of the word from the start of the document:

```
..  set ^doc(doc,w(j),k)=""
```

Where *k* is the position of the word in the document (1,2,3...). With this information, it becomes possible to attenuate the strength of co-occurrences by the proximity of the co-occurrence.

Proximity Related Terms

1721	infarct myocardial	-465
1445	dens lipoprotein	73
1184	abstract word	1184
1128	magnetic resonance	376
1095	cord spin	-309
1057	lymph node	313
964	arthrit rheumatoid	-307
936	female male	-43
878	cholesterol lipoprotein	43
872	blind double	-304
859	compute tomography	255
811	imag resonance	-189
762	carry out	262
742	determine whether	321
733	cholesterol dens	-137
686	mitr valve	114
660	state unit	-240
649	centre nerv	221
639	amino sequence	-7
638	aortic valve	194
633	guinea pig	32
629	care health	-97
625	imag magnetic	-439
610	cystic fibrosis	179
587	head neck	149
566	death sudden	-191

Term Clustering

Terms can be grouped into clusters using the results of the term-term matrix.

In a single link clustering system, a term is added to a cluster if it is related to one term already in the cluster.

Term Clustering

1. Sort the term-term correlations by strength of correlation with the most strongly correlated first.

2. Read each pair of words and their strength of correlation from step 1.

3. Search the clusters. If one of the words is in a cluster, add the other word to the cluster.

4. If neither word of a pair is in a cluster, create a new cluster containing the word pair.

Term Clusters

27 cluster
acquire aids antibiotic deficiency immune immunodeficiency infection viral viru

28 cluster
chemotherapy cisplatin metastase metastatic node

29 cluster
abdominal bowel gland hormone intestinal peptide pituitary regulate secretion testosterone thyroid

3 cluster
accurate compute electron image light microscopic microscopy scanne ultrastructure

30 cluster
diabete diabetic mellitus

31 cluster
cholesterol lipid lipoprotein triglyceride

Term Phrases

As we have seen, recall can be improved if additional, related terms are added to a query. Thus, a query for *antenna* will result in more hits in the data base if the related term *aerial* is added. An increase in recall, however, is often accompanied by a decrease in precision.

Identification of phrases, however, has the potential to increase precision.

These are composite terms of high specificity - such as *television aerial* noted above. While both *television* and *aerial* individually are broad terms, the phrase *television aerial* or *television antenna* is quite specific.

When a phrase is identified, it should become a single term in the document vector.

Term Phrases

Phrases can be identified by both syntactic and statistical methods. While techniques may take into account term proximity as well as co-occurrence such as suggested above, Salton suggests the following simpler formula for construction of term phrases:

```
^Cohesion(i,j) = SIZE_FACTOR * (^tt(i,j)/(^dict(i)*^dict(j)))
```

That is, the cohesion between term *i* and term *j* can be calculated from the frequency of co-occurrence between the terms divided by the product of their total frequencies of occurrence.

Term Phrases

Salton notes that this procedure can sometimes result in unwanted connections such as *Venetian blind* and *blind Venetian*.

For that reason, the aggregate relative order of the terms, as shown above, can help to decide when two terms are seriously linked.

That is, if the order is strongly in favor of on term preceding another, this indicates a probable phrase; on the other hand, if the relative order is in neither direction, this is probably not a phrase.

Cohesion Results

3030 erratum med
2455 carbon dioxide
2164 nitr oxide
2040 med publish
2040 erratum publish
1954 electrophoresis gel
1661 aureus staphylococcus
1372 diabete mellitus
1282 erythematosus lupu
1269 polymorphism restriction
1067 messenger rna
1025 academy america
986 expiratory forc
925 leukocyte polymorphonuclear
885 colit ulcerative
799 peptic ulcer
779 angina pector
724 doubleblind placebo

720 echocardiographic twodimension
700 adenylate cyclase
652 allergic dermatit
647 cava vena
634 boys girl
619 angioplasty transluminal
579 localization localize
560 acquire immunodeficiency
556 doppler puls
555 america society
531 magnetic resonance
520 clas york
514 echocardiographic echocardiography
508 messenger mrna
505 compute tomographic
501 lipoprotein triglyceride
491 guinea pig
489 anesthesia anesthetic

Document-Document matrix

It is also possible to construct Document-Document Matrices giving the correlation between all documents which have significant similarities with one another.

Such a matrix can be used to generate document clusters.

It can also be used for purposes of document browsing by permitting the user to navigate related documents to the one being viewed.

That is, if a user finds on of the retrieved articles of particular interest, a Document-Document matrix can be used to quickly identify other documents related to a document of interest.

The program shown next only calculates the cosines between documents that share *at least* one term rather than between all possible documents.

Document-Document Matrix

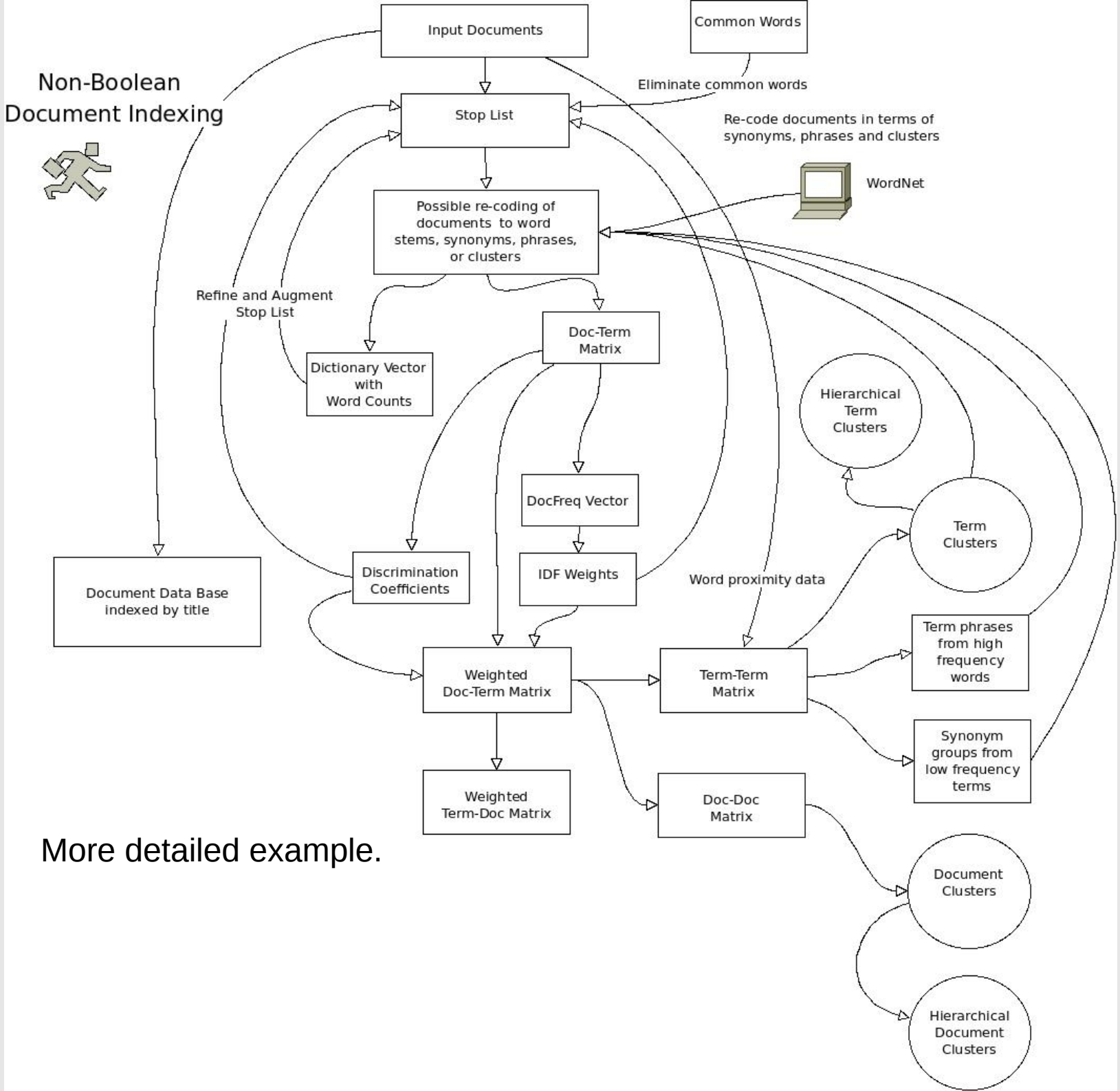
```
1:        244(0.017) 3298(0.010) 3499(0.020) 3510(0.023) 4013(0.089) 4338(0.028) 580(0.005)
100:      3774(0.170) 4942(0.052)
1000:     1897(0.064) 1946(0.056) 3919(0.301) 4014(0.079) 4015(0.173) 987(0.032) 993(0.059)
1001:     1259(0.033) 2596(0.054) 4194(0.046) 504(0.060) 852(0.081) 999(0.113)
1002:     1540(0.175) 217(0.021) 2245(0.175) 2484(0.208) 2795(0.046) 292(0.117) 3755(0.084)
1003:     1539(0.185)
1004:     1007(0.151) 1024(0.054) 1033(0.102) 1474(0.023) 208(0.013) 2940(0.039)
1005:     1007(0.092) 2866(0.031) 3291(0.102) 3298(0.042) 3418(0.103) 3971(0.072)
1006:     1007(0.166) 1897(0.091) 2444(0.011) 2661(0.058) 3206(0.029) 3285(0.039)
1007:     1737(0.142) 1946(0.138) 2666(0.078) 2951(0.072) 3596(0.055) 4008(0.095)
1009:     1540(0.228) 4794(0.297)
1018:     1021(0.050) 2519(0.062) 324(0.158) 3411(0.260) 3546(0.023) 4698(0.060)
1019:     1030(0.033) 1422(0.055) 1480(0.025) 1519(0.103) 1526(0.034) 2186(0.076)

(long lines truncated)
```

Document Clustering

We can calculate document clusters from the document-document matrix in a manner similar to that used for term clusters.

Document Clusters

```
cluster 12
1654   Energy requirements of pregnancy in Scotland.
4475   Energy requirements of pregnancy in The Netherlands.
4582   Energy requirements of pregnancy in the Philippines.

cluster 14
1407   Papers presented at the annual meeting of the Section of Urology, American Academy of
1713   Abstracts of free communications: the thirty-fourth annual meeting of the American
2373   Abstracts. American Academy of Pediatrics, Section on Cardiology, 29th annual meeting, New

cluster 15
3343   Proceedings of the British Institute of Radiology. Practical aspects of magnetic resonance
3344   Uroradiology update. Meeting held jointly by the British Institute of Radiology, the Royal

cluster 16
1180   Metaiodobenzylguanidine as an index of the adrenergic nervous system integrity and function.
1181   Metaiodobenzylguanidine to map scintigraphically the adrenergic nervous system in man.

cluster 17
1408   A nonsurgical approach to the management of primary vesicoureteral reflux.
1410   Endoscopic treatment of vesicoureteral reflux in children.
1411   The endoscopic correction of reflux by polytetrafluoroethylene injection.

cluster 18
270    Abstracts from the 60th scientific sessions. American Heart Association. Anaheim, November
3533   Cardiovascular surgery 1986. Council on Cardiovascular Surgery, American Heart Association.
608    40th annual scientific meeting of the Gerontological Society of America. November 18-

(long lines truncated)
```



More detailed example.

Retrieval

```
---> medlineRetrieve.mps staphyloccus aureus antibiotics


query term: staphylococcus

   Words in this stem group are: staphylococcus
   adding related phrase words: antibiotic-staphylococcus organism-staphylococcus
                        peritonitis-staphylococcus staphylococcus-aureus
                        staphylococcus-endocarditis staphylococcus-epidermidis
                        staphylococcus-infections staphylococcus-methicillin-
                        resistant
                        staphylococcus-strain
   words that may sound similar to staphylococcus: staphylococcal
                        staphylococci

query term: aureus

   Words in this stem group are: aureus
   adding related phrase words: aureus-endocarditis aureus-glycocalyx
                        aureus-infections aureus-osteomyelitis
                        aureus-resistant aureus-strain coagulase-aureus
                        organism-aureus peritonitis-aureus
                        s-aureus staphylococcus-aureus tibiae-aureus
                        vivo-aureus
   words that may sound similar to aureus: none
```

Retrieval

```
query term: antibiotics

   Words in this stem group are: antibiotic antibiotics
   adding related phrase words: aminoglycoside-antibiotic antibiotic-bacteremia
                          antibiotic-bacterial antibiotic-beta-lactam
                          antibiotic-burnetii antibiotic-catheter
                          antibiotic-giardiasis antibiotic-organism
                          antibiotic-parenteral antibiotic-solution
                          antibiotic-staphylococcus antibiotic-valve
                          appropriate-antibiotic assay-antibiotic
                          cartilage-antibiotic cefonicid-antibiotic
                          cephalosporin-antibiotic cost-antibiotic
                          drainage-antibiotic efficacies-antibiotic
                          endocarditis-antibiotic fever-antibiotic
                          infections-antibiotic penetration-antibiotic
                          preoperative-antibiotic prolonged-antibiotic
                          prophylactic-antibiotic prophylaxis-antibiotic
                          susceptibility-antibiotic vancomycin-antibiotic
                          wound-antibiotic
   words that may sound similar to antibiotics: antibacterial
                          antibacterials antibiotic-impregnated
                          antibody-deficiency antibody-dependent
                          antibody-guided antibody-positive

46 documents found.
```

Retrieval

```
 Doc Rel   Title
5676 0.268 Single and combination-antibiotic therapy for experimental endocarditis caused by methicillin-
           matched keys: aureus staphylococcus *staphylococcus-aureus
                         *staphylococcus-methicillin-resistant

7048 0.267 Evolution of the hyperimmunoglobulin E and recurrent infection  HIE  JOB's  syndrome in a youn
           matched keys: aureus *s-aureus staphylococcus *staphylococcus-aureus

5084 0.213 Antibiotic treatment of Staphylococcus aureus endocarditis A review of   cases
           matched keys: *aminoglycoside-antibiotic *antibiotic-beta-lactam
                         *antibiotic-staphylococcus aureus *aureus-strain
                         staphylococcus *staphylococcus-aureus
                         *staphylococcus-strain
6949 0.202 Comparison of culture media for detecting methicillin resistance in Staphylococcus aureus and
           matched keys: aureus *aureus-resistant *aureus-strain
                         *coagulase-aureus *s-aureus staphylococcus *staphylococcus-aureus
                         *staphylococcus-strain
7145 0.200 In vivo glycocalyx expression by Staphylococcus aureus phage type  A  in S aureus osteomyeliti
           matched keys: aureus *aureus-glycocalyx *organism-aureus
                         *organism-staphylococcus *s-aureus
                         staphylococcus *staphylococcus-aureus
                         *tibiae-aureus *vivo-aureus
7146 0.199 Expression of capsular polysaccharide during experimental focal infection with Staphylococcus
           matched keys: aureus *aureus-strain staphylococcus *staphylococcus-aureus
                         *staphylococcus-strain *vivo-aureus

5272 0.184 Staphylococcus aureus septicemia mimicking fulminant Rocky Mountain spotted fever
           matched keys: aureus staphylococcus *staphylococcus-aureus

 880 0.150 Protein A and coagulase expression in epidemic and non-epidemic Staphylococcus aureus
           matched keys: aureus *aureus-resistant *aureus-strain
                         *coagulase-aureus *s-aureus staphylococcus *staphylococcus-aureus
                         *staphylococcus-strain
...
*Time used: 0 seconds
```

SQL?

The workbench generates a SQL *create table* and *insert* commands that can be used to build a relational database of the documents. Given the relational model, is this efficient?

Create Database

```
create table docs (doc int, title varchar(255))
create table docVect (doc int, word varchar(100), wgt float);
create table terms (word varchar(100), TotCount int, DocCount int);
create table dd (doc1 int, doc2 int, wgt float);
create table tt (term1 varchar(100), term2 varchar(100), wgt float);

psql -d isr < SQL.commands
```

SQL Searches

```
select doc, word, round(cast(wgt as numeric),2)
    from docVect
        where word='mole'
        order by wgt desc
        limit 10;

 doc   | word | round
-------+------+-------
 5375  | mole | 34.90
 4231  | mole | 20.94
 4235  | mole | 13.96
 4233  | mole |  6.98
 4236  | mole |  6.98
  242  | mole |  6.98
  857  | mole |  6.98
 2628  | mole |  6.98
 4230  | mole |  6.98
(9 rows)
```

SQL Searches

```
select doc, substring(title,1,40), round(cast (wgt as numeric), 2), word
       from docVect natural join docs
       where word = 'interferon'
            or word = 'bacteria'
            or word = 'infection'
       order by wgt desc limit 10;
```

doc	substring	round	word
5297	Analysis of continuous ambulatory perito	42.30	infection
2178	Comparison of nosocomial infections due	38.07	infection
2461	A five-year study of the microbiologic r	33.84	infection
4467	Inhibition of theophylline metabolism by	31.45	interferon
5763	Infection prophylaxis in pulmonary surge	29.61	infection
5641	Side effects and toxicity of interferon	25.16	interferon
5261	Treatment of idiopathic mixed cryoglobul	25.16	interferon
879	Infections in British clinical laborator	21.15	infection
7134	Disseminated Kaposis sarcoma in AIDS hi	18.87	interferon
7111	Epitopes of human interferon-alpha defin	18.87	interferon

(10 rows)

SQL Searches

```
select doc, substring(title,1,40), round(cast(score as numeric),2)
    from docs natural join (
        select doc, sum(wgt) as score
                from docVect natural join docs
                where word = 'interferon'
                or word = 'bacteria'
                or word = 'infection'
                group by doc) as t1
        order by score desc limit 10;
```

doc	substring	round
5297	Analysis of continuous ambulatory perito	42.30
2178	Comparison of nosocomial infections due	38.07
4467	Inhibition of theophylline metabolism by	35.68
2461	A five-year study of the microbiologic r	33.84
5763	Infection prophylaxis in pulmonary surge	29.61
9824	Depression of murine hepatic mixed funct	26.75
5261	Treatment of idiopathic mixed cryoglobul	25.16
5641	Side effects and toxicity of interferon	25.16
879	Infections in British clinical laborator	21.15
7134	Disseminated Kaposis sarcoma in AIDS hi	18.87

(10 rows)

SQL Searches

```
select doc1, title1, rslt, doc2, title2 from (
     select doc as doc1, substring(title,1,30) as title1, round(cast(score as numeric),2) as rslt
     from docs natural join (
          select doc, sum(wgt) as score
                    from docVect natural join docs
                    where word = 'interferon' or word = 'bacteria' or word = 'infection'
                    group by doc) as t1
          order by score desc limit 10) as e1
     natural join ( select doc1, doc2, substring(title,1,30) as title2 from
               dd, docs where doc2 = doc) as e2
     order by doc1 limit 10;
```

doc1	title1	rslt	doc2	title2
2178	Comparison of nosocomial infec	38.07	6551	A clinical pharmacy-oriented d
2178	Comparison of nosocomial infec	38.07	5203	Endemic multiply resistant St
2178	Comparison of nosocomial infec	38.07	5496	Management of soft-tissue infe
2178	Comparison of nosocomial infec	38.07	5763	Infection prophylaxis in pulmo
2178	Comparison of nosocomial infec	38.07	3863	In vivo comparison of two silv
2178	Comparison of nosocomial infec	38.07	9138	Efficacy of cefazolin cefaman
2178	Comparison of nosocomial infec	38.07	292	Delayed postbacteremic prosthe
2461	A five-year study of the micro	33.84	4575	Outcome in patients on continu
2461	A five-year study of the micro	33.84	4426	Serum magnesium level and arte
2461	A five-year study of the micro	33.84	3564	Serum myoglobin in patients on

(10 rows)

SQL Searches

```
select term1, term2, wgt from tt
     where term1 = 'anesthetize'
     order by wgt desc limit 10;

    term1    |     term2      |  wgt
-------------+----------------+-------
 anesthetize | dogs           | 0.194
 anesthetize | cats           | 0.109
 anesthetize | pentobarbital  | 0.101
 anesthetize | open-chest     | 0.092
 anesthetize | ventilate      | 0.078
 anesthetize | halothane      | 0.076
 anesthetize | min            | 0.067
 anesthetize | hemodynamic    | 0.062
 anesthetize | ventile        | 0.059
 anesthetize | output         | 0.058
(10 rows)
```

SQL Searches

```
select doc, word, substring(title,1,40) as title1
     from docVect natural join docs where word in (
          select term2 from tt
               where term1 = 'anesthetize' order by wgt desc limit 10)
     limit 10;

 doc  |    word     |                  title1
------+-------------+-----------------------------------------
 1019 | min         | Partial aortic ligation  a hypoperfusion
 1020 | dogs        | Role for intrarenal adenosine in the ren
 1020 | hemodynamic | Role for intrarenal adenosine in the ren
 1020 | min         | Role for intrarenal adenosine in the ren
 1027 | min         | Flow dependence of vasopressin-stimulate
   11 | halothane   | Potentiation of the neuromuscular blocka
 1111 | min         | Changes in cerebral blood flow during an
  112 | output      | Pumping device for re-infusion excretion
 1143 | dogs        | Pharmacokinetics and toxicity of intrape
 1161 | cats        | The effect of craniectomy on the biomech
(10 rows)
```

Precision & Recall

Two important metrics of information storage and retrieval system performance are precision and recall.

Precision measures the degree to which the documents retrieved are relevant to the query.

Recall measures the degree to which the system can retrieve all relevant documents.

$$Precision = \frac{\{relevant\ documents\} \cap \{retrieved\ documents\}}{retrieved\ documents}$$

$$Recall = \frac{\{relevant\ documents\} \cap \{retrieved\ documents\}}{relavant\ documents}$$

Precision & recall

For example, if a system responds to a query by retrieving 10 documents from the collection and of these, 8 are relevant and 2 are irrelevant and if the collection actually has 16 relevant documents, we say that the recall is 50% and the precision is 80%.

That is, only 50% of the relevant documents were recalled but of those presented, 80% were correct.

Precision & recall

In general, as recall increases, precision declines.

For example, if by setting thresholds lower the system responds with 20 documents instead of 10 and if 12 of these are relevant but 8 are not, the recall has increased to 75% but the precision has fallen to 60%.

In most systems, as you lower thresholds and more documents are retrieved, the recall will rise but the precision will decline.

In an ideal system, however, as thresholds are lowered, recall increases but precision remains 100%.

Precision and Recall example

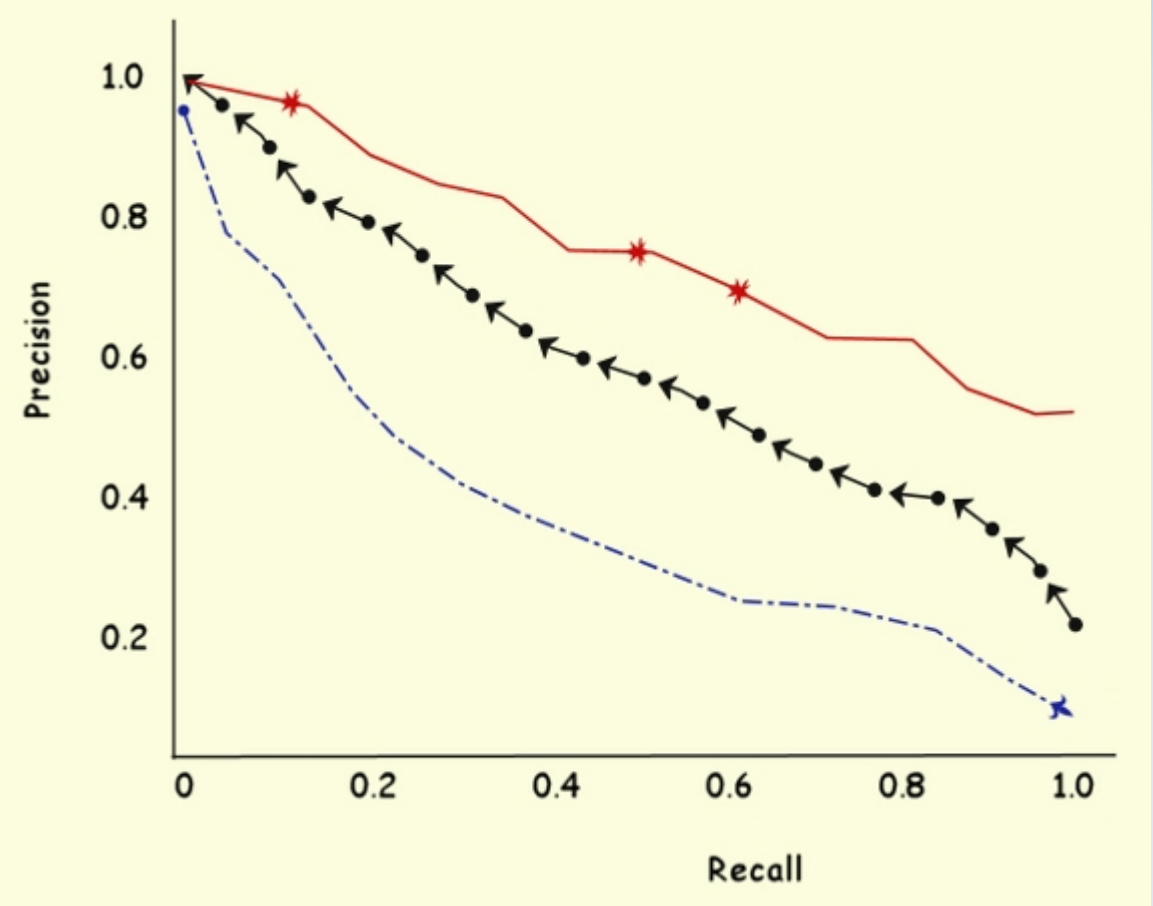
Rank	Relevant?	Recall	Precision
1	yes	0.1	1.0
2	yes	0.2	1.0
3	no	0.2	0.67
4	yes	0.3	0.75
5	yes	0.4	0.80
6	no	0.4	0.67
7	no	0.4	0.57
8	yes	0.5	0.63
9	no	0.5	0.56
10	yes	0.6	0.60

Comparing systems

Salton [1971] used precision-recall graphs similar to the one shown next in order to compare the results of different retrieval experiments.

Those experiments which resulted in a slower drop in precision as recall increased represent improvement in system response.

Precision-recall graphs



Effect of term frequency on precision and recall

Indexing vocabulary effects on precision & recall results.

Generally speaking, terms of *low frequency* tend to increase the precision of a system's responses at the expense of recall as these tend to reference more narrowly defined concepts.

On the other hand, terms of *high frequency* tend to increase recall at the expense of precision as these are often more broadly defined.

Identifying those terms which strike a balance is a major goal of any system.

Other influences on precision and recall

- High frequency terms tend to decrease precision and increase recall
- Low frequency term tend to increase precision and decrease recall
- Adding synonyms decreases precision and increases recall
- Use of term phrases during indexing can increase precision and decrease recall

Effect on recall of added terms

Recall can be improved if additional, related terms are added to a query.

Thus, a query for antenna will result in more hits in the data base if the related term aerial is added.

An increase in recall, however, is often accompanied by a decrease in precision. As is evidenced by the fact that while aerial is a commonly used synonym for antenna, as in television aerial, it can also refer to a dance move, a martial arts move, skiing, various musical groups and performances, and any activity that is done at a height (*e.g.*, aerial photography). Thus, adding it to a query with antenna has the potential to introduce many extraneous hits.