Topics in Information Retrieval

FSNLP, chapter 15

Christopher Manning and

Hinrich Schütze

Information Retrieval

- Getting information from document repositories
- Normally text (though spoken, image, and video data are all becoming more important)
- Traditionally a rather separate field from NLP, and always very empirically based
- A field of some antiquity: the famous SMART IR system (Salton) predates the relational model in databases
- New directions: the Web, email, multimedia, ...
- There is much scope for greater profitable interaction between IR and Statistical NLP

Tasks

- “Ad hoc retrieval”: the user enters query terms which describe the desired information; the system returns a set of (sometimes ranked) documents.
- Document categorization: assign a document to one or more categories (e.g., subject codes) [chapter 16]
  - Filtering: categorization with binary choice about the relevance of a document (e.g., screen for junk email).
  - Routing: categorization for the purpose of transmitting a document to one or more users (e.g., customer service by product)

Tasks (continued)

- Document clustering: group similar documents into clusters (e.g., for making sense of ad hoc retrieval results) [chapter 14]
- Text segmentation: identify semantically coherent units within a text (e.g., for retrieval below the document level) [section 15.4]
- Text summarization: create a shorter version of a document containing just the relevant information
  - Knowledge-based: generate new text
  - Selection-based: extract the n most important summary sentences from the orginal document

[AltaVista] [Advanced Query] [Simple Query] [Private eXtension Products] [Help with Query]

Search the **Web** **Usenet**
Display results **Compact** **Detailed**

Tip: When in doubt use lower-case. Check out Help for better matches.

```
Word count: glass pyramid:  about 200; Pei:9453; Louvre:26578
```

Documents 1-10 of about 10000 matching the query, best matches first.

Paris, France
Paris, France. Practical Info.-A Brief Overview. Layout: One of the most densely populated cities in Europe, Paris is also one of the most accessible,...
http://www.catatravel.com/paris.htm - size 8K - 29 Sep 95

Culture
Culture. French culture is an integral part of France's image, as foreign tourists are the first to acknowledge by thronging to the Louvre and the Centre..
http://www.france.diplomatie.fr/france/edu/culture.gb.html - size 48K - 20 Jun 96

Travel World - Science Education Tour of Europe
Science Education Tour of Europe. B E M I D J I S T A T E U N I V E R S I T Y Science Education Tour of EUROPE July 19-August 1, 1995...
http://www.omnitravel.com/007etour.html - size 16K - 21 Jul 95
http://www.omnitravel.com/etour.html - size 16K - 15 May 95

FRANCE REAL ESTATE RENTAL
LOIRE VALLEY RENTAL. ANCIENT STONE HOME FOR RENT. Available to rent is a furnished, french country decorated, two bedroom, small stone home, built in the..
http://frost2.flemingc.on.ca/~pbell/france.htm - size 10K - 21 Jun 96

LINKS
PAUL'S LINKS. Click here to view CNN interactive and WEBNEWSor CNET. Click here to make your own web site. Click here to manage your cash. Interested in...
http://frost2.flemingc.on.ca/~pbell/links.htm - size 9K - 19 Jun 96

Digital Design Media, Chapter 9: Lines in Space
Construction planes... Glass-sheet models... Three-dimensional geometric transformations... Sweeping points... Space curves... Structuring wireframe...
http://www.gsd.harvard.edu/~malcolm/DDM/DDM09.html - size 36K - 22 Jul 95

No Title
Boston Update 94: A VISION FOR BOSTON'S FUTURE. Ian Menzies. Senior Fellow, McCormack Institute. University of Massachusetts Boston. April 1994. Prepared..
http://www.cs.umb.edu/~serl/mcCormack/Menzies.html - size 25K - 31 Jan 96

Paris - Photograph
The Arc de Triomphe du Carrousel neatly frames IM Pei's glass pyramid, Paris 1/6. © 1996 Richard Nebesky.

Results of the search '"glass pyramid" Pei Louvre' on AltaVista

IR system design

- Unlike databases, IR systems index *everything*
- Usually by an *inverted index* that contains *postings* of all word occurrences in documents
- Having position-in-file information enables *phrase matching* (where an IR “phrase” is just contiguous words)
- A *stop list* of common, meaningless words is often not indexed
- This greatly cuts the inverted index size (given Zipf’s Law)
- *Stemming* means indexing only truncated morphological roots. This sometimes helps (but not always).

Stop words: A small stop list for English

a	also	an	and	as	at	be	but
by	can	could	do	for	from	go	
have	he	her	here	his	how		
i	if	in	into	it	its		
my	of	on	or	our	say	she	
that	the	their	there	therefore	they		
this	these	those	through	to	until		
we	what	when	where	which	while	who	with
would	you	your					

The probability ranking principle (PRP)

IR fundamentally addresses this problem: Given a query W_1 and a document W_2 attempt to decide relevance of W_2 to W_1, where relevance is meant to be computed with respect to their hidden meanings M_1 and M_2.

The model underlying most IR systems (van Rijsbergen 1979: 113):

- PRP: Rank documents in order of decreasing probability of relevance is optimal.

Problems: documents that aren't independent. Any that don't give additional information (especially, duplicates!). Implies not doing word-sense disambiguation.

The Vector Space Model (Salton, TREC)

Represents terms and documents as vectors in k-dimen. space based on the bag of words they contain:

d = The man said that a space age man appeared
d' = Those men appeared to say their age

$$\vec{d} = \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_n \end{pmatrix}$$

	$\vec{d}$	$\vec{d}'$
age	1	1
appeared	1	1
man	2	0
men	0	1
said	1	0
say	0	1
space	1	0

Real-valued vector spaces

Vector dot product (how much do they have in common?):

$$\vec{x} \cdot \vec{y} = \sum_{i=1}^{n} x_i y_i$$

0 if orthogonal (no words in common)

Length of a vector:

$$|\vec{x}| = \sqrt{\sum_{i=1}^{n} x_i^2}$$

Normalized vectors

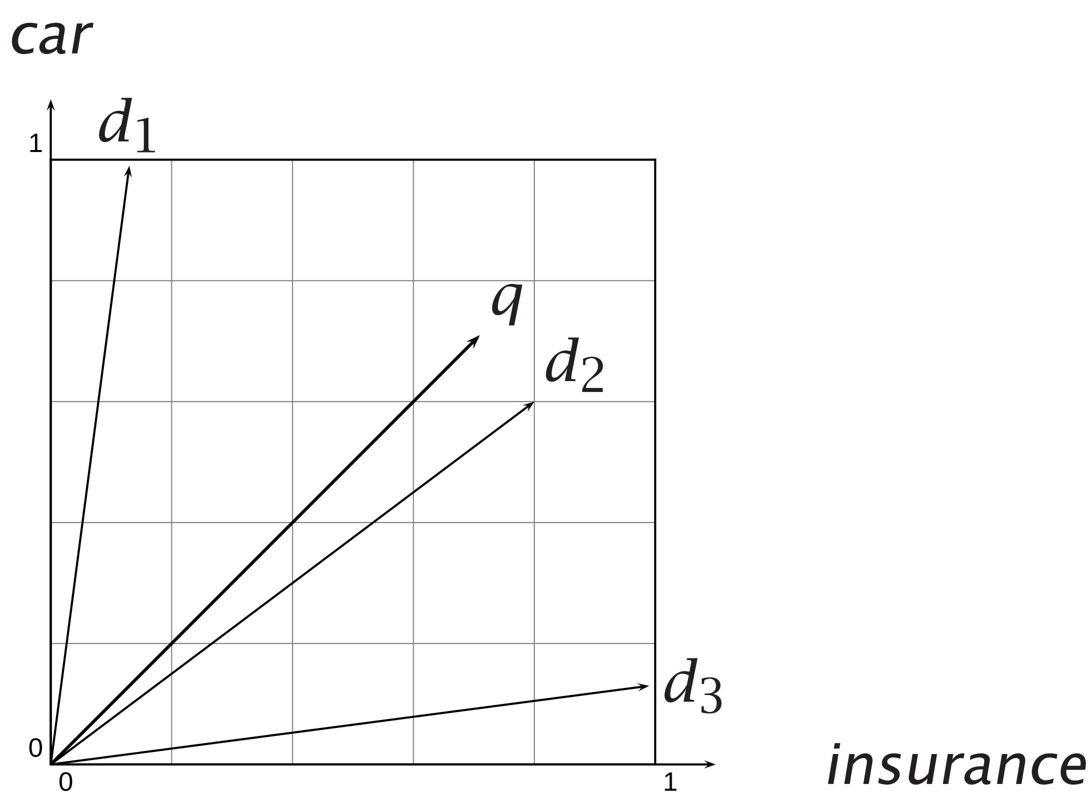
A vector can be normalized (i.e., given a length of 1) by dividing each of its components by the vector's length

This maps vectors onto the unit circle by dividing through by lengths:

Then, $|\vec{x}| = \sqrt{\sum_{i=1}^{n} x_i^2} = 1$

If we didn't normalize vectors, long documents would be more similar to each other! (By the dot product measure.)

The Vector Space Model (normalized vectors)



Cosine measure of similarity (angle between two vectors)

$$\cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

For normalized vectors, the cosine is simply the dot product: $\cos(\vec{x}, \vec{y}) = \vec{x} \cdot \vec{y}$

Developed in SMART system (Salton) and standardly used by TREC participants

Euclidean distance between vectors

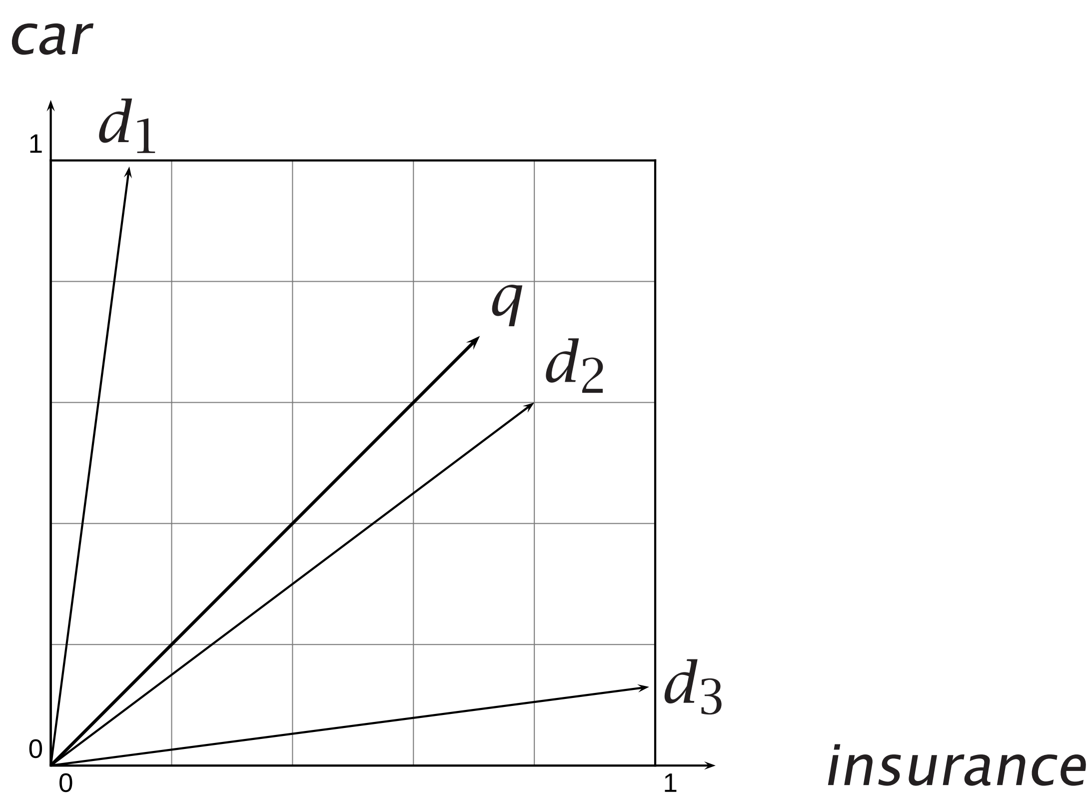
Euclidean distance:

$$|\vec{x} - \vec{y}| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

For normalized vectors, Euclidean distance gives the same closeness ordering as the cosine measure (simple exercise).

The Vector Space Model: Doing a query

We return the documents ranked by the closeness of their vectors to the query, also represented as a vector.



Measuring performance: The 2×2 contingency matrix

Black-box or "end-to-end" system performance

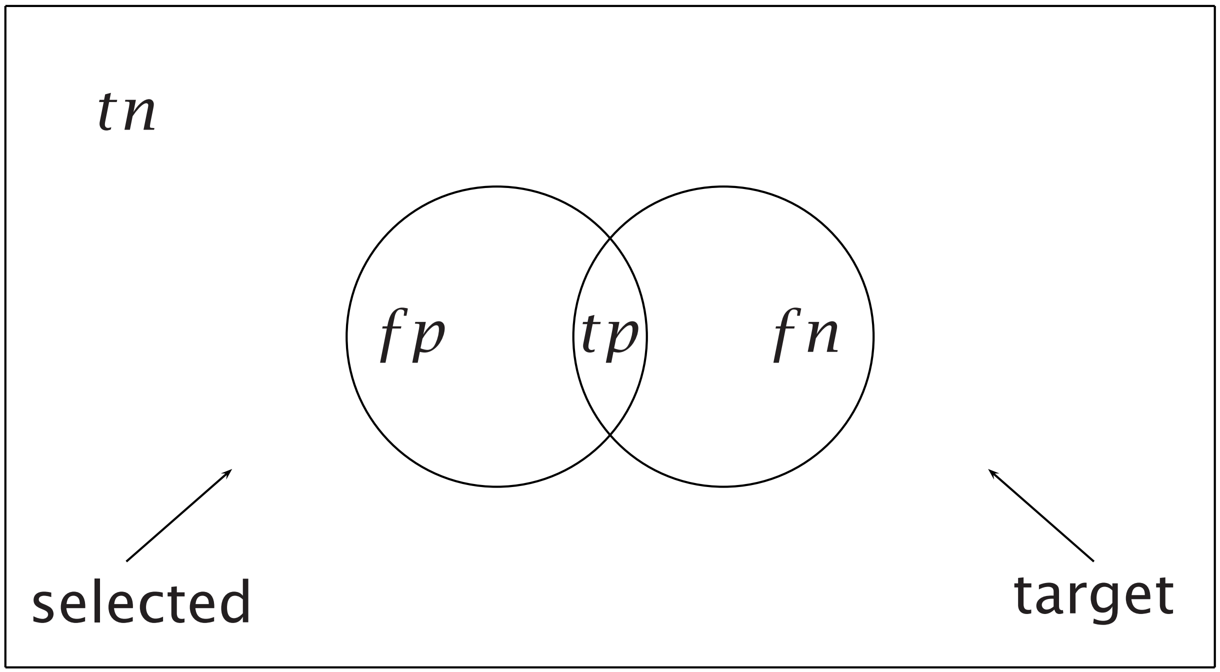
	Actual	
System	target	$\neg$ target
selected	tp	fp
$\neg$selected	fn	tn

Accuracy = $(tp + tn)/N$

Error = $(fn + fp)/N$ = $1 -$ Accuracy

Why is this measure inadequate for IR?

The motivation for precision and recall



Accuracy is not a useful measure when the target set is a tiny fraction of the total set.

Precision is defined as a measure of the proportion of selected items that the system got right:

$$\text{precision } P = \frac{tp}{tp + fp}$$

Recall is defined as the proportion of the target items that the system selected:

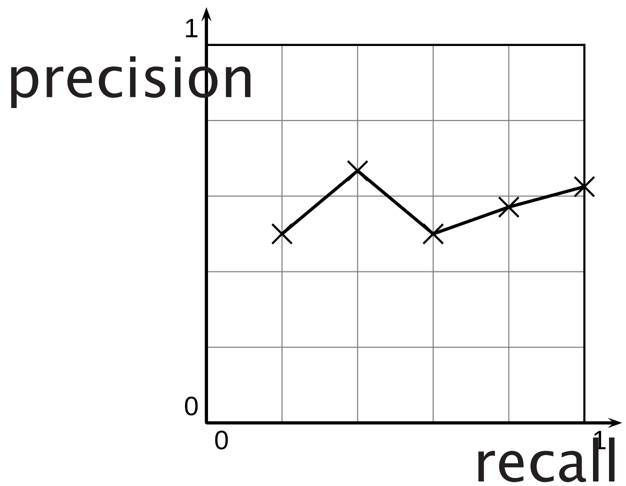
$$\text{recall } R = \frac{tp}{tp + fn}$$

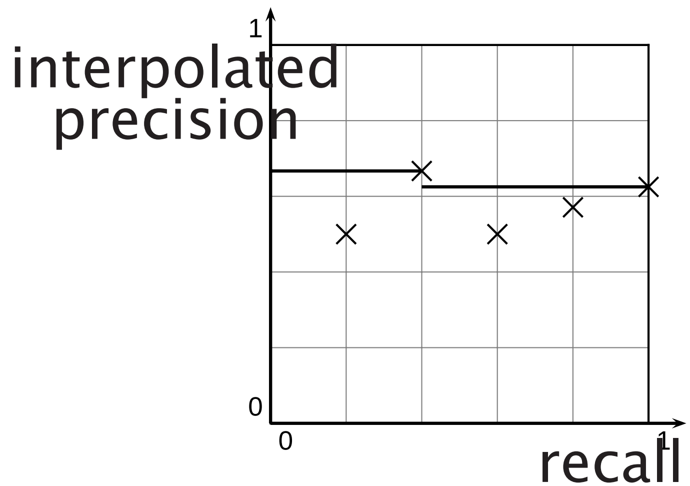
These two measures allow us to distinguish between excluding target items and returning irrelevant items.

They still require human-made “gold standard” judgements.

Evaluation of *ranked* results	Ranking 1	Ranking 2	Ranking 3
	d1: ✓	d10: ×	d6: ×
	d2: ✓	d9: ×	d1: ✓
	d3: ✓	d8: ×	d2: ✓
	d4: ✓	d7: ×	d10: ×
	d5: ✓	d6: ×	d9: ×
	d6: ×	d1: ✓	d3: ✓
	d7: ×	d2: ✓	d5: ✓
	d8: ×	d3: ✓	d4: ✓
	d9: ×	d4: ✓	d7: ×
	d10: ×	d5: ✓	d8: ×
precision at 5	1.0	0.0	0.4
precision at 10	0.5	0.5	0.5
uninterpolated av. prec.	1.0	0.3544	0.5726
interpolated av. prec. (11-point)	1.0	0.5	0.6440

Interpolated average precision





Combined measures

If we can decide on the relative importance of precision and recall, then they can be combined into a single measure.

Does one just add them? Bad, because the measures aren't independent.

What's a sensible model?

Rijsbergen (1979:174) defines and justifies the usually used alternative, the F measure
(see http://www.dcs.gla.ac.uk/Keith/Preface.html).

Assumptions:

- Interested in document proportions not absolute numbers
- Decreasing marginal effectiveness of recall and precision, e.g.:

  $$(R+1, P-1) > (R, P)$$

  but

  $$(R+1, P) > (R+2, P-1)$$

  Makes curves convex towards origin.

The F measure (where $F = 1 - E$)

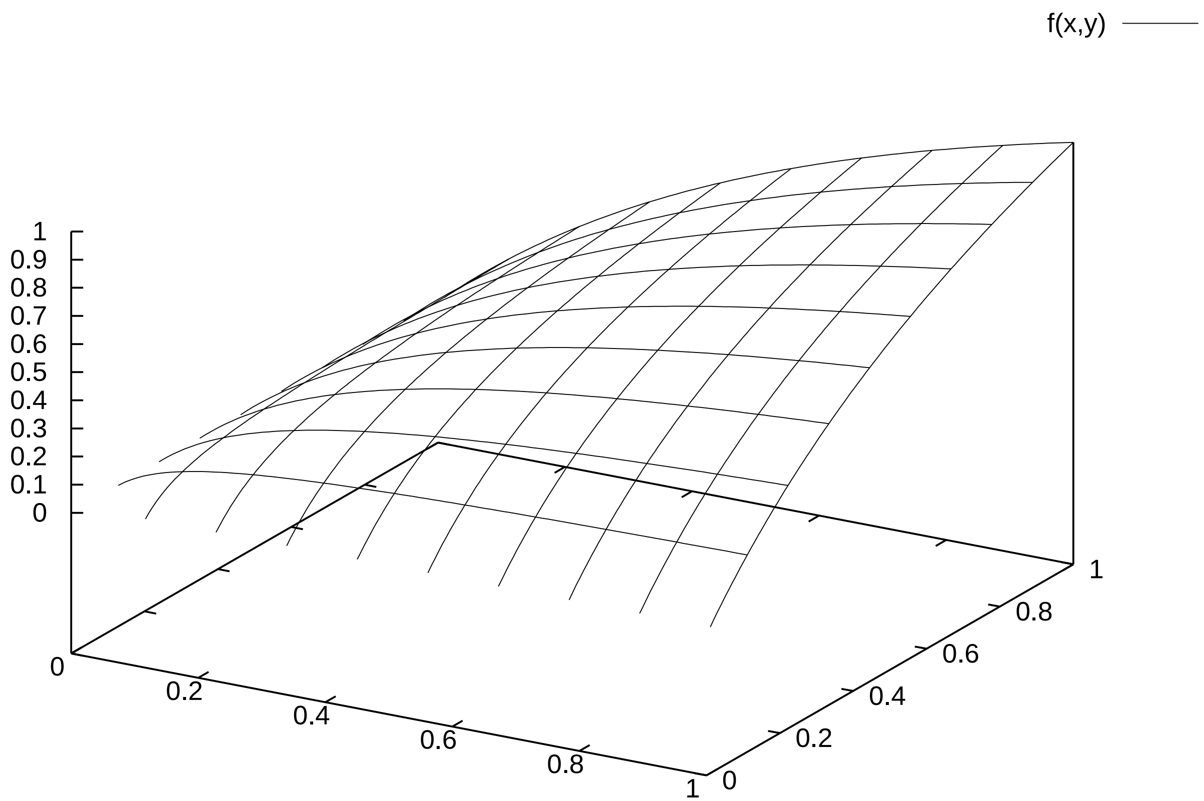
$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}}$$

where P is precision, R is recall and α weights precision and recall. (Or in terms of β, where $\alpha = 1/(\beta^2 + 1)$.)

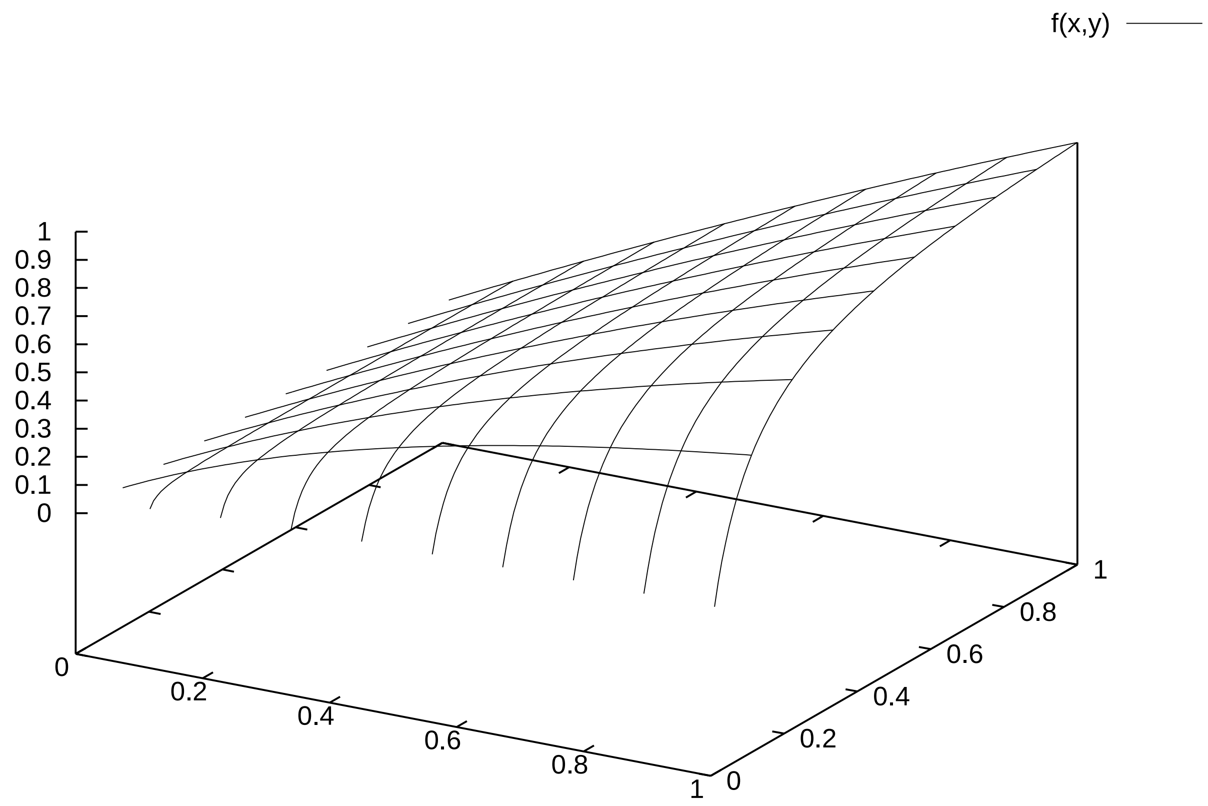
A value of $\alpha = 0.5$ is often chosen.

$$F = \frac{2PR}{R + P}$$

The F measure ($\alpha = 0.5$)



The F measure ($\alpha = 0.9$)



Term weighting

- Simplest term (vector component) weightings are:
  - count of number of times word occurs in document
  - binary: word does or doesn't occur in document
- However, general experience is that a document is a better match if a word occurs three times than once, but not a three times better match.
- This leads to a series of weighting functions that damp the term weighting, e.g., $1 + \log(x)$, $x > 0$, or $\sqrt{x}$.
- This is a good thing to do, but still imperfect: it doesn't capture that the occurrence of a term in a document is more important if that term does not occur in many other documents.

Example of term frequency (from Steven Bird)

- Documents: Austen's *Sense and Sensibility*, *Pride and Prejudice*; Bronte's *Wuthering Heights*
- Terms: affection, jealous, gossip
- SAS: (115, 10, 2); PAP: (58, 7, 0); WH: (20, 11, 6)
- SAS: (0.996, 0.087, 0.017); PAP: (0.993, 0.120, 0.0); WH: (0.847, 0.466, 0.254)

$$\begin{aligned} \cos(SAS, PAP) &= .996 \times .993 + .087 \times .120 + .017 \times 0.0 = 0.999 \\ \cos(SAS, WH) &= .996 \times .847 + .087 \times .466 + .017 \times .254 = 0.929 \end{aligned}$$

Document frequency: indicates informativeness

Word	Collection Frequency	Document Frequency
insurance	10440	3997
try	10422	8760

Adding this in (one of many ways):

$$\text{weight}(i,j) = \begin{cases} (1+\log(\mathrm{tf}_{i,j}))\log\frac{N}{\mathrm{df}_i} & \text{if } \mathrm{tf}_{i,j} \geq 1 \\ 0 & \text{if } \mathrm{tf}_{i,j} = 0 \end{cases}$$

Document frequency weighting is only possible if we have a static collection. Sometimes we don't – it's dynamically created.

Term weighting summary

term frequency $\mathrm{tf}_{i,j}$ number of occurrences of w_i in d_j

document frequency df_i number of documents in the collection that w_i occurs in

collection frequency cf_i total number of occurrences of w_i in the collection

Note that $\mathrm{df}_i \leq \mathrm{cf}_i$ and that $\sum_j \mathrm{tf}_{i,j} = \mathrm{cf}_i$.

- $tf.idf$ weighting: term frequency times inverse document frequency. This is the standard in IR (but it is really a family of methods depending on how each figure is scaled)

Language and implementation problems

- Traditional IR relies on word matching. There are two fundamental query matching problems:
  - synonymy (image, likeness, portrait, facsimile, icon)
  - polysemy (port: harbor, fortified wine, computer jack, ...)
- Effective indexing needs scale, and accuracy
- Dimensionality reduction techniques address part of the first problem, while remaining fairly efficient

Latent Semantic Indexing (LSI)

- *Approach:* Treat word-to-document association data as an unreliable estimate of a larger set of applicable words lying on 'latent' dimensions.
- *Goal:* Cluster similar documents which may share no terms in a low-dimensional subspace (improve recall).
- *Preprocessing:* Compute low-rank approximation to the original term-by-document (sparse) matrix
- *Vector Space Model:* Encode terms and documents using factors derived from SVD
- *Evaluation:* Rank similarity of terms and docs to query via Euclidean distances or cosines

Singular Value Decomposition Encoding

- Computes a truncated SVD of the document-term matrix, using the singlular vectors as axes of the lower dimensional space
- A_k is the best rank-k approximation to the term-by-document matrix A
- Want minimum number of factors (k) that discriminates most concepts
- In practice, k ranges between 100 and 300 but could be much larger.
- Choosing optimal k for different collections is challenging.

Strengths and weaknesses of LSI

- Strong formal framework. Completely automatic. No stemming required. Allows misspellings
- Can be used for multilingual search (Flournoy & Peters Stanford, Landauer Colorado, Littman Duke)
- 'Conceptual IR' recall improvement: one can retrieve relevant documents that do not contain any search terms

- Calculation of LSI is expensive
- Continuous normal-distribution-based methods not really appropriate for count data
- Often improving precision is more important: need query and word sense disambiguation