IR intro

Mark Sanderson
University of Sheffield
m.sanderson@shef.ac.uk

Overview

- Today
  - Classic IR
  - Evaluation
  - Web IR
  - Interfaces
    - If there's time
- Tomorrow
  - Cross language IR
  - Spoken document retrieval

Introduction

- What is IR?
  - General definition
    - Retrieval of unstructured data
  - Most often it is
    - Retrieval of text documents
      - Searching newspaper articles
      - Searching on the Web
  - Other types
    - Image retrieval

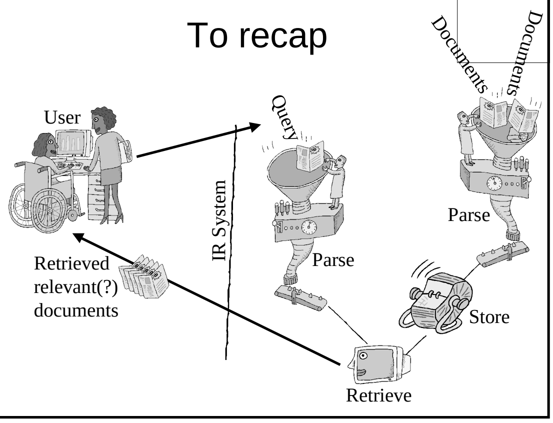
Typical interaction

- User has information need.
  - Expresses it as a query
    - in their natural language?
- IR system finds documents relevant to the query.

Text

- No computer understanding of document or query text
- Use "bag of words" approach
  - Pay little heed to inter-word dependencies:
    - syntax, semantics
  - Bag does characterise document
  - Not perfect: words are
    - ambiguous
    - used in different forms or synonymously

To recap



This section - classic IR

- Parsing
- Indexing
- Retrieving

Parsing

- Normalising format
  - Process different document formats
    - PDF, DOC
    - HTML
      - Can be very noisy, need robust parser
        - » Brin, S., Page, L. (1998) The Anatomy of a Large-Scale Hypertextual Web Search Engine
        - » http://www-db.stanford.edu/pub/papers/google.pdf
- Word segmentation
- Word normalisation

Document

```
<HTML>
<HEAD>
<META HTTP-EQUIV="Content-Type"
CONTENT="text/html; charset=windows-1252">
<TITLE>Compact memories</TITLE>
</HEAD>
<BODY TEXT="#000080" LINK="#0000ff"
VLINK="#800080" BGCOLOR="#ffffff">
<IMG SRC="pic/flex-text.jpg"
ALIGN="RIGHT" WIDTH=127 HEIGHT=166
ALT="flexible capacties">
<P>A <A HREF="http://../defs/data.htm"> digital
data storage</a>
system with capacity up to bits and random and or
sequential access is described
…
```

a	1	
b	compact	1
b	memories	9
b	have	17
b	flexible	21
b	capacities	30
b	a	41
b	digital	43
b	data	51
b	storage	60
b	system	68
b	with	75
...		

Word segmentation

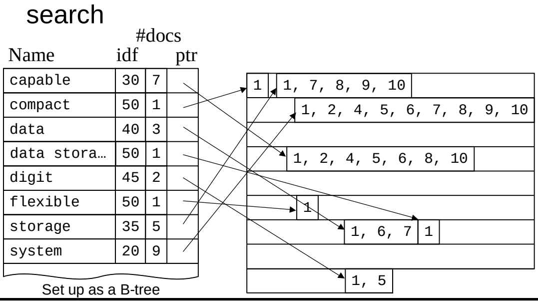
- English is easy
  - Space character? Well…
  - It is said that Google is indexing not just words, but common queries too
    - "Britney Spears"
- Other languages present problems
  - Chinese
    - no space character
    - http://www.sighan.org/bakeoff2003/
  - Japanese
    - Four alphabets
      - Romanji, Hiragana, Katakana, Kanji
  - German, Finish, URLs, etc.
    - compound words
      - "Donaudampfschiffahrtsgesellschaftsoberkapitän"
  - Arabic, Latvian, etc,
    - large number of cases to normalise European languages

Once segmented

- Case normalisation
  - If your language has case
- Stop word removal
  - Remove common words from query
- Stemming
  - Normalise word variants
  - English
    - Inflectional stemmers
      - Remove plurals (e.g. 's', 'es', etc)
      - Remove derivational suffixes (e.g. 'ed', 'ing', 'ational', etc)
        - » Porter, M.F. (1980): An algorithm for suffix stripping, in *Program - automated library and information systems*, 14(3): 130-137

Once normalised

- Create data structure to facilitate fast search



Retrieving

- Boolean
- Ranked retrieval (best match)
  - Adhoc
    - Do something that works, based on testing
  - Models
    - Vector space
    - Probabilistic
      - Classic
      - Okapi BM25
      - Language models

Boolean

- The original IR system
- User enters query
  - Often complex command language
  - Collection partitioned into set
    - Documents that match query
    - Documents that do not
- Traditionally, no sorting of match set
  - Perhaps by date

Ranked retrieval

- User enters query…
- …calculate relevance score between query and *every* document
  - Estimate what users typically want when they enter a query
- Sort documents by their score
  - Present top scoring documents to user.

Adhoc

- Popular approach
  - Create some weighting functions around notions (intuitions) of what seems sensible

$$\sum_{T \in Q} \frac{\log(t+1)}{\log(dl)} \bullet \log\left(\frac{N}{n}\right)$$

  - Term frequency (*tf*)
    - *t*: Number of times term occurs in document
    - *dl*: Length of document (number of terms)
  - Inverse document frequency (*idf*)
    - *n*: Number of documents term occurs in
    - *N*: Number of documents in collection

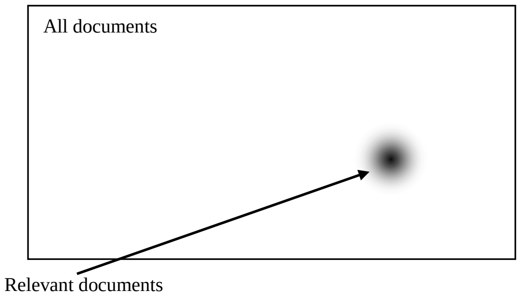
TF $\frac{\log(t+1)}{\log(dl)}$

- More often a term is used in a document
  - More likely document is about that term
  - Depends on document length?
    - Harman, D. (1992): Ranking algorithms, in Frakes, W. & Baeza-Yates, B. (eds.), *Information Retrieval: Data Structures & Algorithms*: 363-392
      - » Typo: not unique terms.
    - Singhal, A. (1996): Pivoted document length normalization, *Proceedings of the 19th ACM SIGIR conference*: 21-29

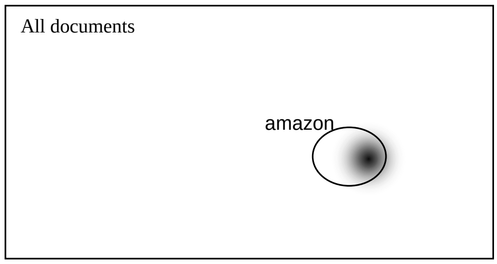
IDF $\log\left(\frac{N}{n}\right)$

- Some query terms better than others?
- Query on…
  - "destruction of amazon rain forests"
  - …fair to say that…
    - "amazon" > "forest" ≥ "destruction" > "rain"
      - Prefer documents that have amazon repeated/emphasised a lot

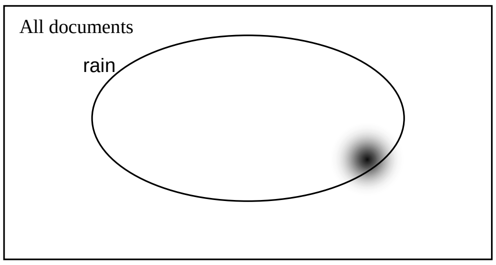
To illustrate



To illustrate



To illustrate



IDF and collection context

- IDF sensitive to the document collection content
  - General newspapers
    - “amazon” > “forest” ≥ “destruction” > “rain”
  - Amazon book store press releases
    - “forest” ≥ “destruction” > “rain” > “amazon”

Successful

- Simple, but effective
- Core of most weighting functions
  - *tf* (term frequency)
  - *idf* (inverse document frequency)
  - *dl* (document length)

Getting the balance

- Documents with all the query terms?
- Just those with high *tf•idf* terms?
  - What sorts of documents are these?
- Search for a picture of Arbour Low
  - Stone circle near Sheffield
  - Try Google and AltaVista
    - Old example

My query

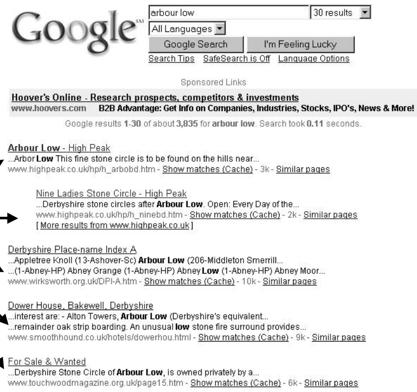
- Near Sheffield
  - "The Stonehenge of the north"

Very short "arbour" only

Longer, lots of "arbour", no "low"

"arbour low"

Arbour Low documents do exist



Lots of Arbour Low documents

Disambiguation?

Previously… every document?

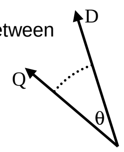
- "calculate relevance score between query and *every* document"
- In many retrieval applications
  - Not every document
  - Only those documents that have all users query words

Models

- All a little ad hoc?
  - Mathematically modelling the retrieval process
    - So as to better understand it
    - Draw on work of others
  - Overview of four
    - Vector space
    - Classic probabilistic
    - BM25
    - Language models

Vector Space

- Document/query is a vector in N space
  - N = number of unique terms in collection
- If term in doc/qry, set that element of its vector
- Angle between vectors = similarity measure
  - Cosine of angle (cos(0) = 1)
- Term per dimension
  - Model says nothing about dependencies between terms
    - Independent



Formula

$$sim(d_j, q) = \frac{\vec{d}_j \bullet \vec{q}}{\left|\vec{d}_j\right| \times \left|\vec{q}\right|} \qquad sim(d_j, q) = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{j=1}^{t} w_{i,q}^2}}$$

- $w_{x,y}$ - weight of vector element

- Vector space
  - Salton, G. & Lesk, M.E. (1968): Computer evaluation of indexing and text processing. *Journal of the ACM*, 15(1): 8-36
  - Any of the Salton SMART books

Classic probabilistic

- Like naïve Bayes classifier
  - Treat document as a binary vector
    - Probability of observing relevance given document *x* is observed?

$$P(R \mid \vec{x}) = \frac{P(R) \bullet P(\vec{x} \mid R)}{P(\vec{x})}$$

- Assume independence of terms
  - come back to this $P(\vec{x} \mid R) = \prod_{i=1}^{n} P(x_i \mid R)$
- Leads to
  - Summation of *idf* query terms

Model references

- Original papers
  - Robertson, S.E. & Spärck Jones, K. (1976): Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3): 129-146.
  - Van Rijsbergen, C.J. (1979): *Information Retrieval*
    - Chapter 6
- Surveys
  - Crestani, F., Lalmas, M., van Rijsbergen, C.J., Campbell, I. (1998): "Is This Document Relevant? ...Probably": A Survey of Probabilistic Models in Information Retrieval, in *ACM Computing Surveys*, 30(4): 528-552
  - Lavrenko, V. (2004): "A Generative Theory of Relevance" Ph.D. dissertation
    - Chapter 2
    - http://ciir.cs.umass.edu/pubfiles/ir-370.pdf

Incorporating *tf*

- Classic probabilistic model
  - Assumed binary representation of documents
- Much effort to include *tf*
  - Best example
    - BM25
      - Popular weighting scheme
        - » Robertson, S.E., Walker, S., Beaulieu, M.M., Gatford, M., Payne, A. (1995): Okapi at TREC-4, in *NIST Special Publication 500-236: The Fourth Text REtrieval Conference (TREC-4)*: 73-96

BM25

- Wanting to model notion of eliteness
  - Would indexer assign document term as a keyword?
  - Estimate with 2-Poisson model
    - Look at a term across a collection
    - Does its *tf* occur as 2 Poisson distributions?
      - One, where the term isn't important
      - One, where the term is.
  - Eventual formula not derived mathematically from derivations, but empirically found to best approximate distributions

Robertson's BM25

$$\sum_{T \in Q} w \frac{(k_1+1)tf}{K+tf} \frac{(k_3+1)qtf}{k_3+qtf} + k_2 |Q| \frac{avdl - dl}{avdl + dl}$$

- *Q* is a query containing terms *T*
- *w* is a form of *IDF*
- k_1, b, k_2, k_3 are parameters.
- *tf* is the document term frequency.
- *qtf* is the query term frequency.
- *dl* is the document length (arbitrary units).
- *avdl* is the average document length.

Continues to be developed

- Amati's divergence from random
  - How much does a term occurrence in a document differ from random?
    - Amati G. (2003): Probability Models for Information Retrieval based on Divergence from Randomness, *Thesis of the degree of Doctor of Philosophy, Department of Computing Science University of Glasgow*
      - » http://www.dcs.gla.ac.uk/~gianni/ThesisContent.pdf

Language models

- View each document as a language model
  - calculate probability of query being generated from document $P(Q \mid D)$
  - Compute for all documents in collection
  - Rank by probability
- Generated much interest
  - Ties IR into area of extensive NLP research.

Language models

- Speech recognition, machine translation
  - Work on building uni-gram, multi-gram models of language
  - Comparing language models
- Information Retrieval use work from this active field

Early language model papers

- August, 1998
  - Ponte, J., Croft, W.B. (1998): A Language Modelling Approach to Information Retrieval, in *Proceedings of the 21st ACM SIGIR conference*: 275-281
- September, 1998
  - Hiemstra D. (1998): A Linguistically Motivated Probabilistic Model of Information Retrieval, In: *Lecture Notes in Computer Science: Research and Advanced Technology for Digital Libraries* (vol. 513): 569-584
- November, 1998
  - Miller, D.R.H., Leek, T., Schwartz, R.M. (1998): BBN at TREC7: Using Hidden Markov Models for Information Retrieval. *Proceedings of TREC-7*: 80-89

Independence of terms?

- Most models assume independence of terms
  - Occurrence of one term independent of others
- Terms are dependent
  - Relevance should be calculated on term combinations as well.
- Ad hoc approximations
  - Successful
- Early attempts to explicitly model dependence
  - Probability models
    - Unsuccessful
  - Latent Semantic Indexing
    - Examining term dependencies
  - Language models
    - More success

Ad hoc approximations of dependence

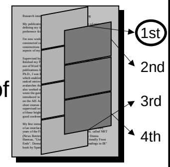
- Within query text
  - Phrase indexing
    - Documents holding query phrase
      - are more relevant
  - Passage retrieval
    - Documents holding query terms in close proximity
      - are more relevant
- Beyond query text
  - Automatic query expansion (pseudo relevance feedback)
    - Documents holding terms related to query words
      - are more relevant
  - Spell correction
    - Documents holding correctly spelled versions of query words
      - are more relevant

Phrase indexing

- Syntactic or statistical methods to locate phrases
  - Index by them too
    - Phrase in query? Up score of documents that hold phrase
- Compare statistical with syntactic, statistics won, just
  - J. Fagan (1987) Experiments in phrase indexing for document retrieval: a comparison of syntactic & nonsyntactic methods, in *TR 87-868 - Department of Computer Science, Cornell University*
- More research has been conducted.
  - T. Strzalkowski (1995) Natural language information retrieval, in *Information Processing & Management*, Vol. 31, No. 3, 397-417

Passage retrieval

- Documents holding query words close together
  - Are better
- Split document into passages
- Rank a document based on score of its highest ranking passage
- What is a passage?
  - Paragraph?
    - Bounded paragraph
  - (overlapping) Fixed window?
    - Callan, J. (1994): Passage-Level Evidence in Document Retrieval, in *Proceedings of the 17^{th} ACM-SIGIR*: 302-310



Automatic query expansion

- Collection wide, global analysis
  - Qiu, Y., Frei, H.P. (1993): Concept based query expansion, in *Proceedings of the 16^{th} ACM SIGIR*: 160-170
- Per query analysis
  - (Pseudo|Local) relevance feedback
    - Xu, J., Croft, W.B. (1996): Query Expansion Using Local and Global Document Analysis, in *Proceedings of the 19^{th} ACM SIGIR*: 4-11

Example – from LCA

- “Reporting on possibility of and search for extra-terrestrial life/intelligence”
  - extraterrestrials, planetary society, universe, civilization, planet, radio signal, seti, sagan, search, earth, extraterrestrial intelligence, alien, astronomer, star, radio receiver, nasa, earthlings, e.t., galaxy, life, intelligence, meta receiver, radio search, discovery, northern hemisphere, national aeronautics, jet propulsion laboratory, soup, space, radio frequency, radio wave, klein, receiver, comet, steven spielberg, telescope, scientist, signal, mars, moises bermudez, extra terrestrial, harvard university, water hole, space administration, message, creature, astronomer carl sagan, intelligent life, meta ii, radioastronomy, meta project, cosmos, argentina, trillions, raul colomb, ufos, meta, evidence, ames research center, california institute, history, hydrogen atom, columbus discovery, hypothesis, third kind, institute, mop, chance, film, signs

Spell correction

- Academic papers?

Modeled dependency

- Early probabilistic
  - See Van Rijsbergen's book, Ch. 6
- Vector Space
- Language models

Advances on vector space

- Latent Semantic Indexing (LSI)
  - Reduce dimensionality of N space
    - Consider dependencies between terms
      - Furnas, G.W., Deerwester, S., Dumais, S.T., Landauer, T.K., Harshman, R.A., Streeter, L.A., Lochbaum, K.E. (1988): Information retrieval using a singular value decomposition model of latent semantic structure, in *Proceeding of the 11th ACM SIGIR Conference*: 465-480
      - Manning, C.D., Schütze, H. (1999): *Foundations of Statistical Natural Language Processing*: 554-566

Language models

- Bi-gram,
- Bi-term
  - "Information retrieval", "retrieval (of) information"
    - Gao, J., Nie, J.-Y., Wu, G., Cao, G. (2004) Dependence Language Model for Information Retrieval, in the proceedings of the 27th ACM SIGIR conference: 170-177

Evaluation

- Why?
  - I've told you about IR systems and improvements
    - but how do we know they are improvements?
- Need to evaluate

What do you evaluate?

- Anyone anyone?

Precision

$$\text{Precision} = \frac{\text{Relevant and Retrieved}}{\text{Retrieved}}$$

Calculating for one query

- Precision at ?

Rank	Doc ID	Score	Rel?	
1	20	1683		
2	7	1352	Relevant	
3	18	1296		
4	10	1249	Relevant	
5	2	1241		
6	12	1184		
7	16	1074		
8	6	1045	Relevant	
9	17	1042		
10	3	1017		

Evaluate a system

- New system & collection configuration
- Go through a set of queries
- Compute precision at fixed rank for each query
  - 10, 20, 100?
- Average across the queries
- We're all happy right?

What's missing?

- Every time a new system comes along
  - Have to re-evaluate each time
    - Needs people!
- How do I compare with others?
- How many documents did we not get?

Recall

$$\text{Recall} = \frac{\text{Relevant and Retrieved}}{\text{Total relevant}}$$

Total relevant?

- How do you do that?

Test collections

- Test collections
  - Set of documents (few thousand-few million)
  - Set of queries (50-400)
  - Set of relevance judgements
    - Humans check all documents!
    - Use pooling
      - Target a subset (described in literature)
      - Manually assess these only.

      - System pooling
      - Query pooling

Test collection references

- TREC conferences
  - http://trec.nist.gov/pubs/
    - – Any of the overview papers
  - – Query pooling
    - – Cormack, G.V., Palmer, R.P., Clarke, C.L.A. (1998): Efficient Constructions of Large Test Collections, in *Proceedings of the 21st annual international ACM-SIGIR conference on Research and development in information retrieval*: 282-289
  - – Validation of pooling
    - – Voorhees, E. (1998): Variations in Relevance Judgements and the Measurement of Retrieval Effectiveness, in *Proceedings of the 21st annual international ACM-SIGIR conference on Research and development in information retrieval*: 315-323

Another ranking

Rank	Doc ID	Rel?	Recall	Precision	Rels		Total Rel
			0	0		8	3
1	8	Relevant	0.33	1.00		4	
2	17		0.33	0.50		10	
3	18		0.33	0.33			
4	1		0.33	0.25			
5	9		0.33	0.20			
6	13		0.33	0.17			
7	11		0.33	0.14			
8	16		0.33	0.13			
9	19		0.33	0.11			
10	20		0.33	0.10			
11	5		0.33	0.09			
12	4	Relevant	0.67	0.17			
13	12		0.67	0.15			
14	6		0.67	0.14			
15	2		0.67	0.13			
16	14		0.67	0.13			
17	10	Relevant	1.00	0.18			
18	7		1.00	0.17			
19	3		1.00	0.16			
20	15		1.00	0.15			

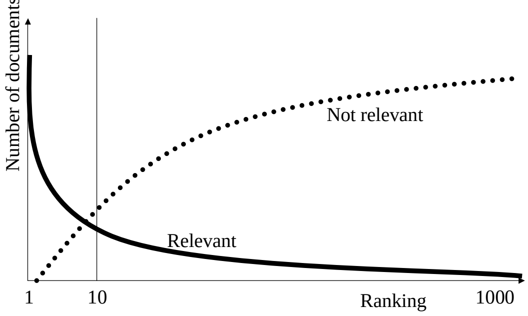
Over a ranking?

Rank	Doc ID	Rel?	Recall	Precision	Rels		Total Rel
			0.00	0.00		4	70
1	20		0.00	0.00		7	
2	7	Relevant	0.01	0.50		6	
3	18		0.01	0.33		…	
4	10		0.01	0.25			
5	2		0.01	0.20			
6	12		0.01	0.17			
7	16		0.01	0.14			
8	6	Relevant	0.03	0.25			
9	17		0.03	0.22			
10	3		0.03	0.20			
11	14		0.03	0.18			
12	4	Relevant	0.04	0.25			
13	9		0.04	0.23			
14	11		0.04	0.21			
15	19		0.04	0.20			
16	5		0.04	0.19			
17	1		0.04	0.18			
18	13		0.04	0.17			
19	8		0.04	0.16			
20	15		0.04	0.15			

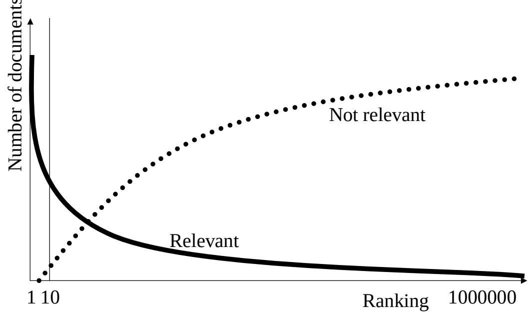
Measuring at a cut-off

- Influenced by the number of relevant documents
  - –Too few
    - Can normalise by number of relevant.
  - –Too many
    - Hawking, D., Thistlewaite, P. (1997): Overview of TREC-6 Very Large Collection Track, in *NIST Special Publication 500-240: The Sixth Text REtrieval Conference (TREC 6)*, E.M. Voorhees, D.K. Harman (eds.): 93-106

Small collection



Big collection



Measuring at recall points

- Don't measure at rank cut offs
  - Equivalent user effort
- Measure at recall values
  - 0, 0.1, 0.2, 0.3,… …, 0.9, 1.0 is popular.
  - Measure precision at each relevant document
    - Mean Average Precision (MAP)
- Good discussion
  - Hull, D. (1993) Using Statistical Testing in the Evaluation of Retrieval Experiments, in *Proceedings of the 16th annual international ACM SIGIR conference on Research and Development in Information Retrieval*: 329-338

Are test collections any good?

- "Drive by shooting"
  - One go at retrieving
    - "Never allowed to search again"
- Need to consider interaction
  - What's better
    - System that takes ages for one query
    - System that retrieve super fast
      - Allows/encourages many searches

What is relevance?

- Broder
  - Informational – almost all test collections
  - Navigational
  - Transactional
- Aspectual?
- Plagiarism?
- Readable?
- Known item?
- Authoritative
  - See further in slides

References

- Broder, A. (2002) A taxonomy of web search, *SIGIR Forum*, 36(2), 3-10.
- Excite query log analysis
  - Amanda Spink mainly in IP&M

Other forms of evaluation

- Usability of interface
- Speed
  - Appears to have dramatic impact on user ability to locate relevant documents.

Usability

- Does user understand how system works?
- Test collection says…
  - Hard to understand system retrieves more in first search
- …better than…
  - …poorer system that users understand.
- But…
  - users may be able to refine search on later system, ultimately retrieve more.

Queries answered

- MAP?
  - The density of relevant documents near the top of the ranking
    - Who cares?
- P@10?
  - Number of relevant in top 10
    - Do I really care if I get 5 or 10 relevant??
- Queries answered
  - How many queries had at least one in top N.

Web retrieval

- Brief coverage

Single Collection

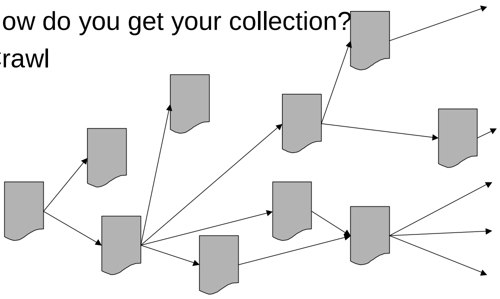
- You are Google, Alta Vista, what are your problems?
  - You are not in control of the collection you are searching
  - You have to provide a service for free!

Implication

- Must collect the collection
- Deal with
  - Changes
  - Language
- No editors (too big)
  - Undesirable content
  - Misrepresented content
  - Mistaken content
  - Boring content

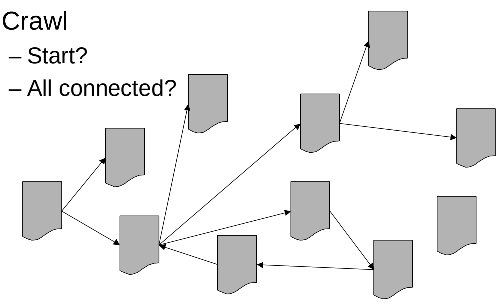
Collecting the collection

- How do you get your collection?
- Crawl



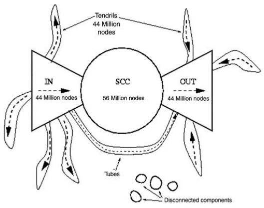
Collecting the collection

- Crawl
  - Start?
  - All connected?



The web is a bowtie

- Bowtie?
  - A.Z. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J.L. Wiener: Graph structure in the Web. *WWW9/Computer Networks* 33(1-6): 309-320 (2000)
- Six degrees of separation?
  - http://smallworld.columbia.edu/



Changes

- Need to keep checking pages
  - Pages change
    - At different frequencies
    - Who is the fastest changing?
    - Pages are removed
  - Consequence of the media?
  - Google caches its pages

Undesirable content

- You’re a family web site
  - Do you want sex pages?
  - Innocuous words can cause problems
    - “Men with hands”
  - Train recognisers
    - In general porn sites cooperate.

Spam?

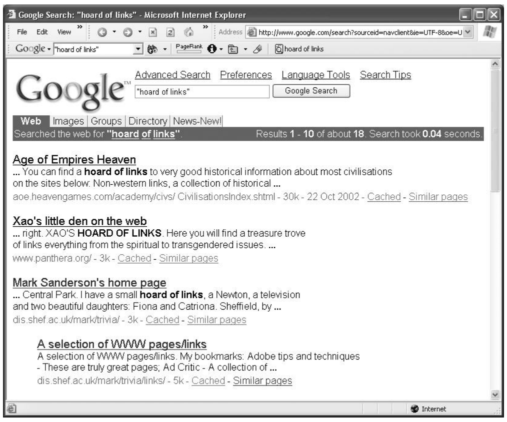
- Big problem

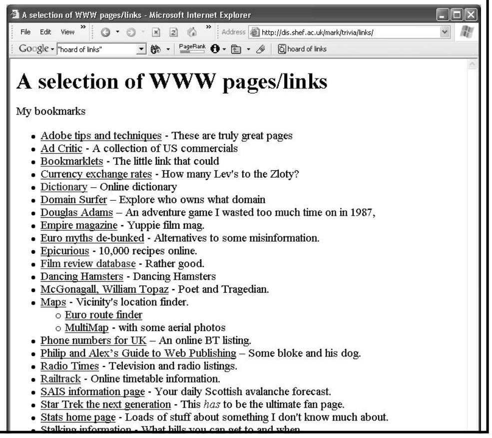
Best match (with Boolean)

- Find pages with all query words
  - Boolean AND
- Sort by a range of factors
  - Number of times words occur
  - Closeness of words
  - Word order

Other information?

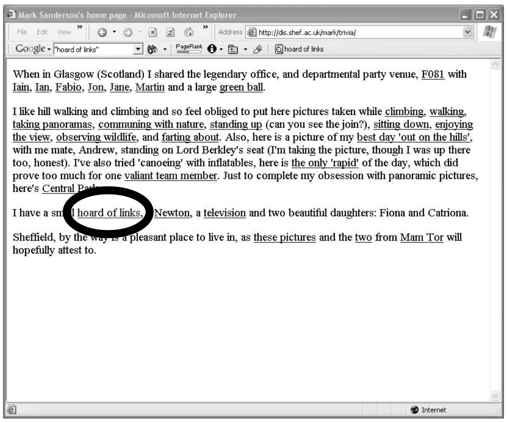
- Title
- First few lines
- Colour
- Font
- Size
  - Yahoo/Alta Vista, query help information
    - http://help.yahoo.com/help/us/ysearch/





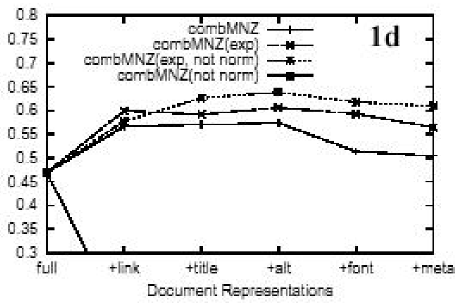
Where is the text?

- Exact phrase no-where on the page
- Google examines anchor text also
- Anchor text?
  – Blue text of link pointing to page



Another example

- “click here”
- What will happen?



- Full text, anchor text, title, image alt text, large font text, metatags
  – Ogilvie, P., Callan, J. Combining document representations for known-item search, in the *Proceedings of the 26th annual international ACM SIGIR conference on Research and Development in Information Retrieval*. Pages 143-150, 2003

Popularity?

- Query IMDB (www.imdb.org) for "Titanic"

Titanic (1915)
Titanic (1997)
Titanic (1943)
Titanic (1953)
Titanic 2000 (1999)
Titanic: Anatomy of a Disaster (1997)
Titanic: Answers from the Abyss (1999)
Titanic Chronicles, The (1999)
Titanic in a Tub: The Golden Age of Toy Boats (1981)
Titanic Too: It Missed the Iceberg (2000)
Titanic Town (1998)
Titanic vals (1964)...aka Titanic Waltz (1964) (USA)
Atlantic (1929)...aka Titanic: Disaster in the Atlantic (1999) (USA: video title)
Night to Remember, A (1958)...aka Titanic latitudine 41 Nord (1958) (Italy)
Gigantic (2000)...aka Untitled Titanic Spoof (1998) (USA: working title)
Raise the Titanic (1980)
Saved From the Titanic (1912)
Search for the Titanic (1981)
Femme de chambre du Titanic, La (1997)...aka Camarera del Titanic, La (1997) (Spain) ...aka Chambermaid on the Titanic, The (1998) (USA)
...aka Chambermaid, The (1998) (USA: promotional title)
Doomed Sisters of the Titanic (1999)

Use popularity

- Query "titanic" on IMDB
  - Titanic (1997)
- On the Web
  - Most search engines

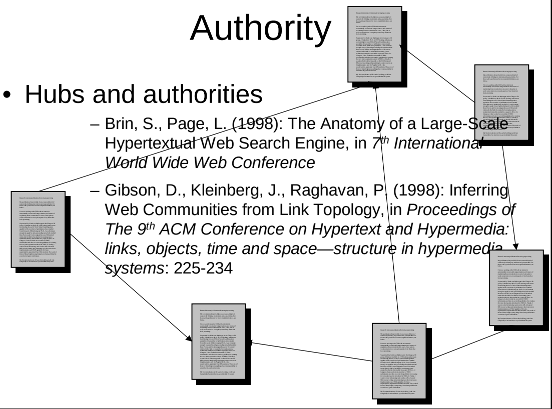
Home page finding?

- URL length
  - Good for finding home pages
- Domain name (www.sheffield.ac.uk)
  - Is query in domain name?
    - Yes good idea

Authority

- In classic IR
  - authority not so important
- On the web
  - very important (boring or misrepresented)
    - Query "Harvard"
      - Dwane's Harvard home page
      - The Harvard University home page

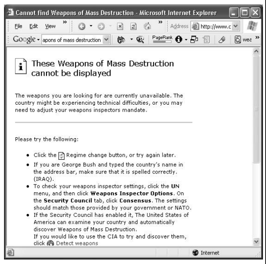
Authority



- Hubs and authorities
  - Brin, S., Page, L. (1998): The Anatomy of a Large-Scale Hypertextual Web Search Engine, in 7th *International World Wide Web Conference*
  - Gibson, D., Kleinberg, J., Raghavan, P. (1998): Inferring Web Communities from Link Topology, in *Proceedings of The 9th ACM Conference on Hypertext and Hypermedia: links, objects, time and space—structure in hypermedia systems*: 225-234

Authority?

- Search on Google for
  - "weapons of mass destruction"
    - Is this authoritative?
      - Popular?
  - "french military victories"



Spamming

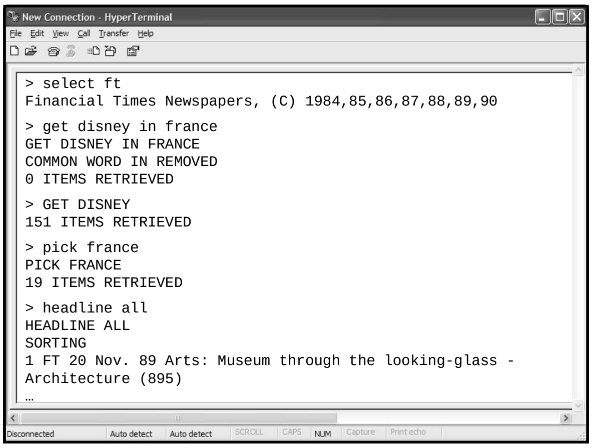
- Harder to spam a page to make it an authority?
  – Certainly not impossible
- Harder to spam a popularity system

Interface

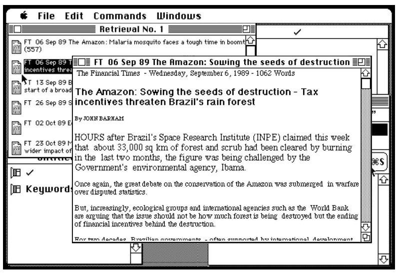
- Look
- Overview

Interface look

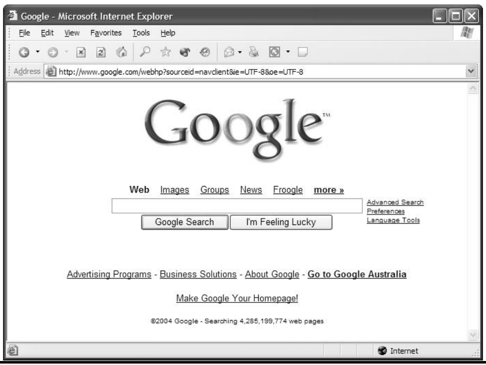
- Specialised applications
  – Classic Boolean look
  – Early WIMP interfaces
  – Web search engines
- Ubiquitous search



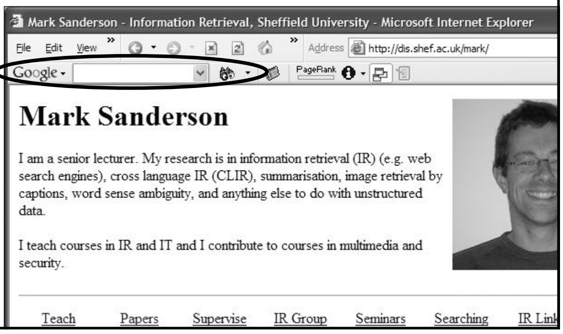
Early WIMP interfaces



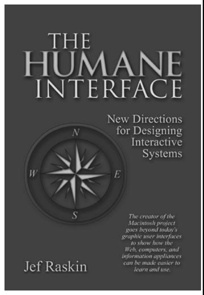
Early web search



Ubiquity



Further ubiquity?



"Leap" keys – modifiers for immediate searching

Cross Language Information Retrieval (CLIR)

Mark Sanderson

m.sanderson@shef.ac.uk

Aims

- To introduce you to issues involved in and methods used for Cross Language Information Retrieval (CLIR)

What is CLIR?

- CLIR (sometimes translingual retrieval)
  - Query written in one language (*source*)…
  - …retrieving documents written in other (*target*) language(s).
- MLIR
  - Collection holds document many languages
  - Query in many languages
  - No translation
- Monolingual IR
  - Query, collection, same language.

Where did the name come from?

- Retrieving a*cross* languages
  - Name defined at SIGIR 1996 workshop
    - Organised by Gregory Grefenstette
    - Before then, multi-lingual IR
    - www.ee.umd.edu/medlab/mlir/conferences.html

Why do it?

- Increased awareness of other languages
  - Soon only 30% Internet users native English
- User motivations
  - Retrieve documents and get them translated.
  - People can read a foreign language before they can write it.
  - Polyglots want to only enter a query in one language
  - Multimedia documents described by text
  - Minority language providers

User studies?

- Can users judge retrieved documents as relevant if they can't read document language?
  - Using machine translation?
    - Yes
    - Shown for a number of languages
  - Using word/phrase lists?
    - Yes
    - Shown for some languages

Is it possible?

- I thought machine translation was no good
  - Information Retrieval different
    - Documents and queries are bags of words.
      - No need for correct
        - » Syntax
        - » Stop words
    - IR systems tolerant of some level of error.

How to do it

- Actual working systems
  - (my own)
- Active research
  - Other approaches, are they actually used?

Working systems - how do you do it?

- What are the problems
  - Word segmentation
  - Word normalisation
  - Translation
    - How to translate
    - Picking correct translation
      - Ambiguity
      - Phrases
  - What do you translate?
    - Query

Word segmentation/normalisation

- English is easy
  - Space character? Well…
- Other languages present problems
  - Chinese
    - no space character
  - Japanese
    - Four alphabets
      - Romanji, Hiragana, Katakana, Kanji
  - German, Finish, URLs, etc.
    - compound words
      - "Donaudampfschiffahrtsgesellschaftsoberkapitän"
  - Arabic, Latvian, etc,
    - large number of cases to normalise

Picking correct translation

- Words are ambiguous, many translations
  - “grand” (in French)
    - “big”, “large”, “huge”, “massive”? (in English)
- Phrases
  - “Petit déjeuner”
    - “Little dinner”?
    - “Breakfast”!

Translation resources?

- Machine translation (MT) system
- Bilingual dictionary
- Aligned bilingual or parallel corpora
- Comparable corpora

Machine translation

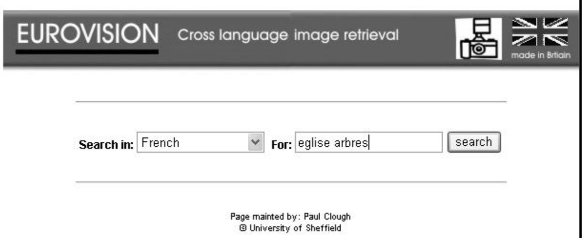
- Designed for more complex purpose
- Good with ambiguity
  - Hand built rules
  - May only have access to first guess
- Expensive to build
- Rare
  - Systran
    - Lot’s of well known languages into & out of English
    - That’s about it

Machine translation example

- Eurovision
  - Image CLIR system
    - English captions
  - Babel Fish wrapper
    - Systran professional

Eurovision

Enter query



Église Arbres

- Very simple
  - Using systran

Bilingual dictionaries

- Ballesteros's work

- Ensured phrase translation dictionary
- Sophisticated query language
- LCA
  - Query expansion
  - Pseudo relevance feedback

Sophisticated query language

- Query in French
  - "grand avion"
- Translate to English
  - "big, large, huge, massive, plane, aeroplane"
    - Translation of "grand" may dominate query
  - Solution?
    - "SYNONYM(big, large, huge, massive), SYNONYM(plane, aeroplane)"
      - Available in Inquery & Lemur

Bilingual dictionary

- Simple
- No built in support for ambiguity
- Commoner
  - Increasingly online

Good references?

- Lisa Ballesteros, a great review of past work
    - Ballesteros, L., Cross Language Retrieval via Transitive Translation, *Advances in Information Retrieval: recent research from the center for intelligent information retrieval*, Croft W.B. (ed.), 203-234

- Recent TREC/CLEF
  - http://trec.nist.gov/
  - http://www.clef-campaign.org/

No translation?

- If you have no resource?
  - Languages a bit similar?
    - French is badly spelled English
    - Query French collection with English query
      - Expand query from English collection
      - Enough will match in French
    - Works OK

No translation?

- Proper names
  - London
  - Londres
- Unlikely to be in dictionary
- Treat as spell correction problem
    - Pirkola, A. & Toivonen, J. & Keskustalo, H. & Visala, K. & Järvelin, K. (2003). Fuzzy Translation of Cross-Lingual Spelling Variants. In proceedings of the 26th ACM SIGIR Conference, pp. 345 - 352

Research - how do you do it?

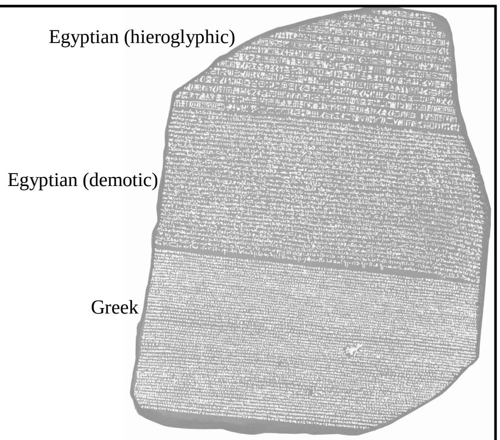
- What are the problems
  - How to translate
    - Look at some other possibilities
- *What do you translate*
  - Query or document?
    - Query less work, but less evidence
    - Document, more work, more accurate
    - Both, compare translations

Translation resources?

- Machine translation (MT) system
- Bilingual dictionary
- Aligned bilingual or parallel corpora
- Comparable corpora

Parallel corpora

- Direct translation of one text into another
  - Aligned at sentence level
  - Canadian Hansards
    - "Le chien et dans le jardin. La chat et sur la table"
    - "The dog is in the garden. The cat is on the table"
- Much rarer than dictionaries



Mining parallel texts from Web

- Get to a well funded non-English web site?
  - Often presented in English as well
- Crawl sites
  - Assume structure and layout similar

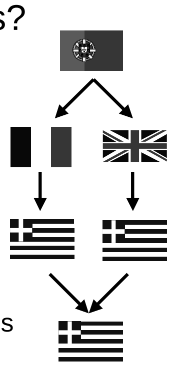
Comparable corpora

- **Not** a direct translation of one text into another
  - – Aligned at document level
  - – Swiss newspapers
- Can handle phrases and ambiguity
  - – If examples are in the corpora
- Still rare

Further work

- No translation resource.
  - – Portuguese to Greek?
    - No.
  - – Use a intermediate (pivot) language.
    - Portuguese to English
    - English to Greek
  - – Transitive retrieval common situation.

Use many pivots?



- One pivot
  - – Portuguese to English
  - – English to Greek
- Other pivot
  - – Portuguese to French
  - – French to Greek
- Intersect two Greek translations

Pivot references

- Gollins, T. & Sanderson, M. (2001) Improving Cross Language Retrieval with Triangulated Translation. *In the Proceedings of the 24th ACM SIGIR conference*, 90-95
- Ballesteros, L., Sanderson, M. (2003) Addressing the lack of direct translation resources for cross-language retrieval, in the *Proceedings of the 12th international conference on Information and Knowledge Management* (CIKM) 147-152

Query expansion

- Local Context Analysis
- Ballesteros
  - – Expand query before translating
    - From separate collection in language of the query
  - – After translating
  - – Clear improvements shown for both

Experiments

- Ballesteros's system produces very good retrieval (73% of monolingual)
  - – One of the first to make people think CLIR was being solved
    - Subsequent improvements on % of monolingual
- One question worth asking…
  - – Do users want pseudo-relevance feedback?

Spoken Document Retrieval

Mark Sanderson
m.sanderson@shef.ac.uk

Aims

- To provide an overview of the issues in the retrieval of audio recordings of speech.

Objectives

- At the end of the lecture you will be able to:
  - Provide a witty example of the problems in recognising speech
  - Explain which forms of speech are easier to recognise
  - Give a simple overview of
    - recognition methods
    - how speech retrieval is done

Why?

- Increasing interest in doing this
  - Speech recognition getting better
    - Faster
  - Speech track of TREC (SDR)
- I have/had some involvement/interest in this

How speech recognition works

- Don’t know, don’t care
  - It’s a black box with some knobs on
    - Discrete or continuous?
    - Speaker (in)dependent?
    - Vocabulary size
      - Phonemes, large vocabulary?
    - Language models
  - Output
    - Stream of words with attached probabilities
    - Other hypotheses

Problems hearing

- Say this
  - “How to wreck a nice beach”
- Now this
  - “How to recognise speech”
  - “How to wreck an ice peach”

Progress

- Unlike similar tasks, e.g. object recognition
  - Large improvements in SDR
- Follow improvements in SR
  - Improvements in computers
    - Processor speed
    - Reducing RAM and disk prices

Early work

- SR couldn't do large vocabulary
  - Consonant vowel consonant
    - Glavitsch, U., Schäuble, P. (1992): A System for Retrieving Speech Documents, in *Proceedings of the 15th ACM SIGIR conference* : 168-176
  - Small vocabulary <100 (word spotting)
    - K. Sparck Jones, G.J.F. Jones, J.T. Foote and S.J. Young, Experiments in spoken document retrieval, *Information Processing and Management*, 32(4), pp399-417, 1996, Elsevier (reprinted in Readings in Information Retrieval, Morgan Kaufman, 1997)

Passed a threshold

- Since 1996/7, had
  - Large vocabulary
    - > 60,000 words
  - Speaker independent
  - Continuous speech recognition
  - Low word error rate (WER)

SDR track of TREC

- Started in 1997 (TREC-6)
  - J. Garofolo, E. Voorhees, V. Stanford, K. Sparck Jones TREC-6 1997 Spoken Document Retrieval Track Overview and Results, *NIST Special Publication 500-240: The Sixth Text REtrieval Conference (TREC 6)*
  - Small collection
    - 100 hours of data
      - 1,500 stories
        - » 400,000 words

Techniques

- Recognise text, retrieve on it
  - Retrieval from recognised transcript almost as good as retrieval from hand transcribed.
- Combine multiple transcripts?
  - Yes, that works
  - Same as using multiple hypotheses?
    - Yes sort of similar
    - And it works

Multiple transcripts

- Hand generated transcript:
  - ...when we talk about blacks and whites we eventually get around to the tough question some of you are...
- Recogniser 1:
  - …I will talk about blacks and winds we eventually go wrong a of the tough question who he hid...
- Recogniser 2:
  - ...we talked about blanks and whites we eventually get around to the tough question his own unions say well....

Why does SDR work?

- Remember *tf*?
  - Documents with high *tf* are more likely to be what?
  - What does a document with high *tf* have?

Use of other collections

- Expand document with text from another (parallel?) source
  - Works
    - Singhal, A., Choi, J., Hindle, D., Lewis, D.D. (1998): AT&T at TREC-7, in *Proceedings of the 7th TREC conference (TREC-7)* published by NIST

New TREC areas

- TREC-8
    - Johnson, S.E., Jourlin, P., Sparck Jones, K., Woodland, P.C. (1999): Spoken Document Retrieval for TREC-8 at Cambridge University, in *proceedings of the 8th Text REtrieval Conference (TREC 8)*
  - Story segmentation
    - Remember Callan?
    - Dissimilar passages segment a story
  - Removing commercials?
    - Look for repeating sounds
      - Very effective

Other areas

- Unlimited vocabularies
  - Large vocabulary, plus phonemes
    - Wechsler, M., Munteanu, E., Schäuble, P. (1998): New Techniques for Open-Vocabulary Spoken Document Retrieval, in *Proceedings of the 21st ACM SIGIR conference*
- Retrieval of dirty/casual speech
  - Telephones
  - Conversations

SR accuracy

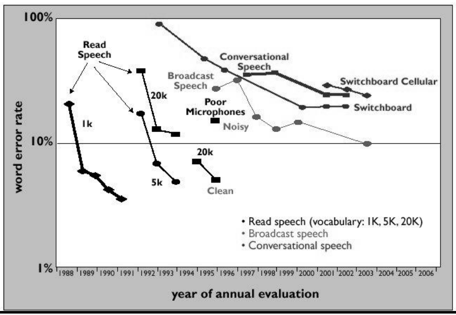


Figure reference

- Deng, L., Huang X. (2004) Challenges in adopting speech recognition, *Communications of the ACM*, 47(1), 69-75

There is more too it…

- In an SDR collection…
  - Documents badly recognised
  - Documents very well recognised.
- Retrieval ranks the well recognised
  - AAAI Spring Symposium 2003
    - "The relationship of word error rate to document ranking"
    - www.mind-project.org/papers/SS503XShou.pdf

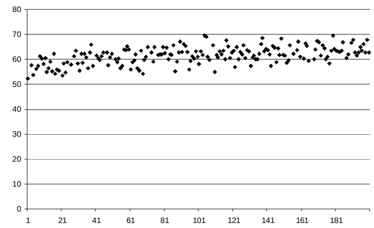
Remaining work to be done?

- Presentation of speech retrieval results
  - Snippets unreadable?
- SpeechBot
  - There's a 50% WER
    - Every other word is wrong – on average
- Looked readable to me
  - Why can users read the search result page?
- Question
  - Do top ranked documents have a lower WER than lower ranked?

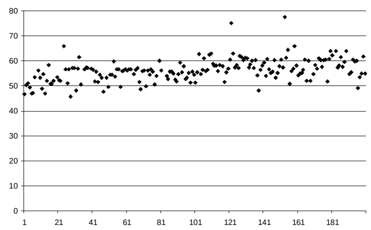
TREC-7 SDR data

- Easy to work with
  - Manual transcript of spoken documents
    - Easy to compute WER
  - Multiple transcripts of speech
  - Multiple ranks of speech documents

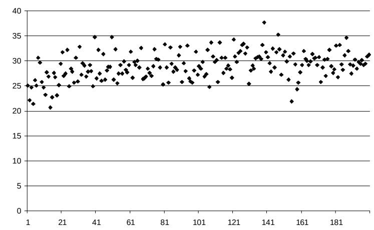
Results derasru-s1



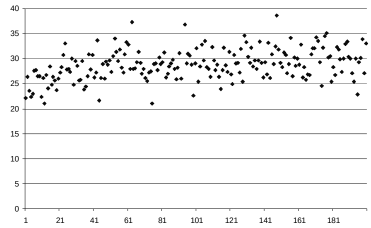
Results derasru-s2



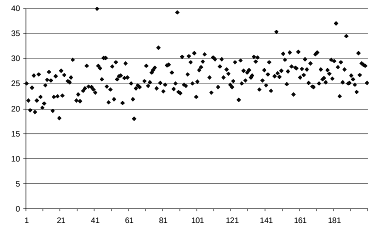
Results att-s1



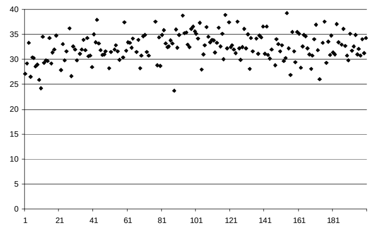
Results att-s2



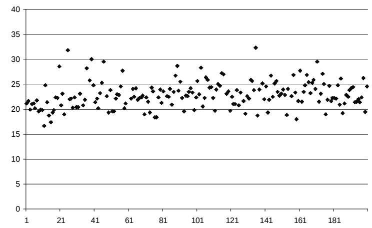
Results dragon-s1



Results shef-s1



Results cuhtk-s1



Why is this happening?

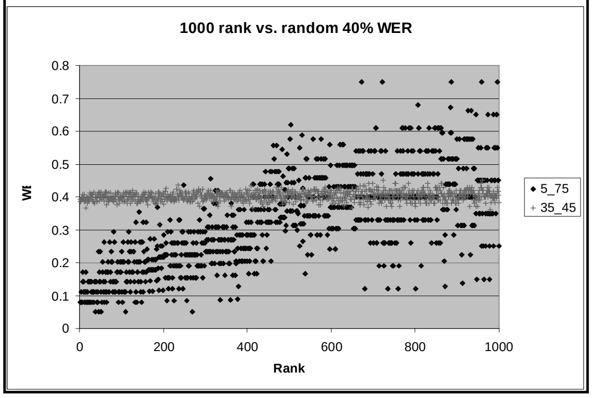
- *tf* weights
  - Documents with the highest *tf* weights are the most relevant and the best recognised?
    - Probably across audio document.
    - Good for query words probably the rest too.
- Quorum scoring
  - Documents matching on many query words again probably cleaner
    - Probability of words co-occurring in same documents very low
    - Query as a language model?

Particularly so for passages

- Match on query words in close proximity (as seen in result list), again other words in that passage likely to be recognised well.

Trend is slight

- TREC SDR more consistently clean?
- Test on SpeechBot
  - – Examined 312 retrieval result transcripts
    - Listened to audio section (not all found)
  - – Found WER of 17.6%
  - – Much lower than 50% reported across collection



Conclusion

- Speech retrieval works well and it's usable
  - – Ranking helps locate better recognized documents
- If you search in top 10, collection is large enough
  - – SDR will be very successful

Wider implications

- OCR (retrieve most readable documents)
  - – Similar problem, similar result?
- CLIR (retrieve most easily translated?)
  - – If you translate the query?
    - I think so but I can't explain why
  - – If you translate the document collection
    - Yes
    - Retrieve documents translated better?

Overview papers

- Two summary papers
  - (2001) Allan, James "Perspectives on Information Retrieval and Speech," in *Information Retrieval Techniques for Speech Applications*, Coden, Brown and Srinivasan, editors. pp. 1-10.
    - – http://ciir.cs.umass.edu/pubfiles/ir-236.pdf
  - The TREC Spoken Document Retrieval Track : A Success Story (Garofolo et al, April 2000).
    - – http://www.nist.gov/speech/tests/sdr/sdr2000/papers/01plenary1.pdf