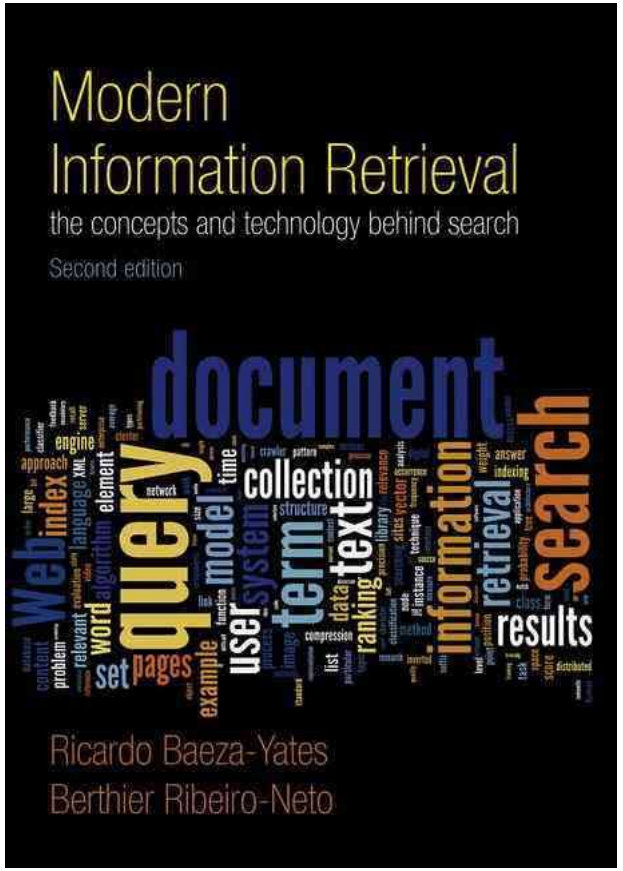
Information Retrieval

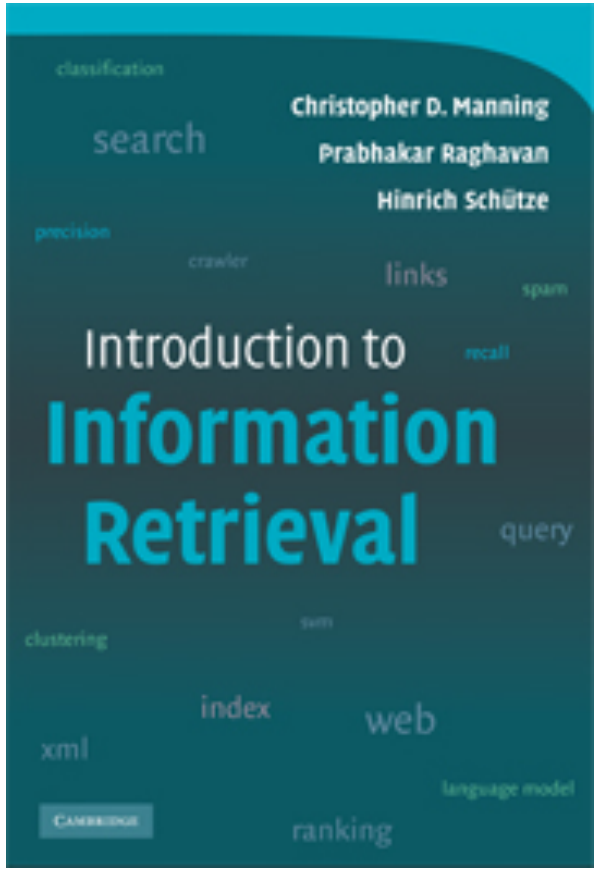
Intro and Historical Background

Dorota Glowacka

glowacka@cs.helsinki.fi

Material





Additional resources:

- Kelly, Diane. "Methods for evaluating interactive information retrieval systems with users." *Foundations and Trends® in Information Retrieval* 3.1–2 (2009).
- White, Ryen W., and Resa A. Roth. "Exploratory search: Beyond the query-response paradigm." *Synthesis lectures on information concepts, retrieval, and services* 1.1 (2009).
- Harman, Donna. "Information Retrieval: The Early Years." *Foundations and Trends® in Information Retrieval* 13.5 (2019).

Information Retrieval Forums

- ACM Special Interest Group on Information Retrieval (SIGIR)
  https://sigir.org/
- *SIGIR Forum* https://sigir.org/forum/
- **Conferences**: SIGIR, CIKM, WSDM, SAC, ECIR, JCDL, ICTIR, CHIIR, TREC
- **Journals**: TOIS, IPM, IR, JASIST

Terminology

- **General**: Information Retrieval, Interactive Information Retrieval, Exploratory Search, Information Need, Query, Retrieval Model, Retrieval Engine, Search Engine, Relevance, Relevance Feedback, Evaluation, Information Seeking, Human-Computer Interaction, Browsing, Interfaces, Filtering
- **Related**: Document Management, Knowledge Engineering
- **Expert**: term frequency, document frequency, inverse document frequency, vector-space model, probabilistic model, BM25, page rank, stemming, precision, recall, F1

Information Retrieval: Informal Definition

Representation, storage, organisation and access to information
(documents, information items, information objects)

Find relevant (useful) information

- **Goal of an IR system** – RECALL

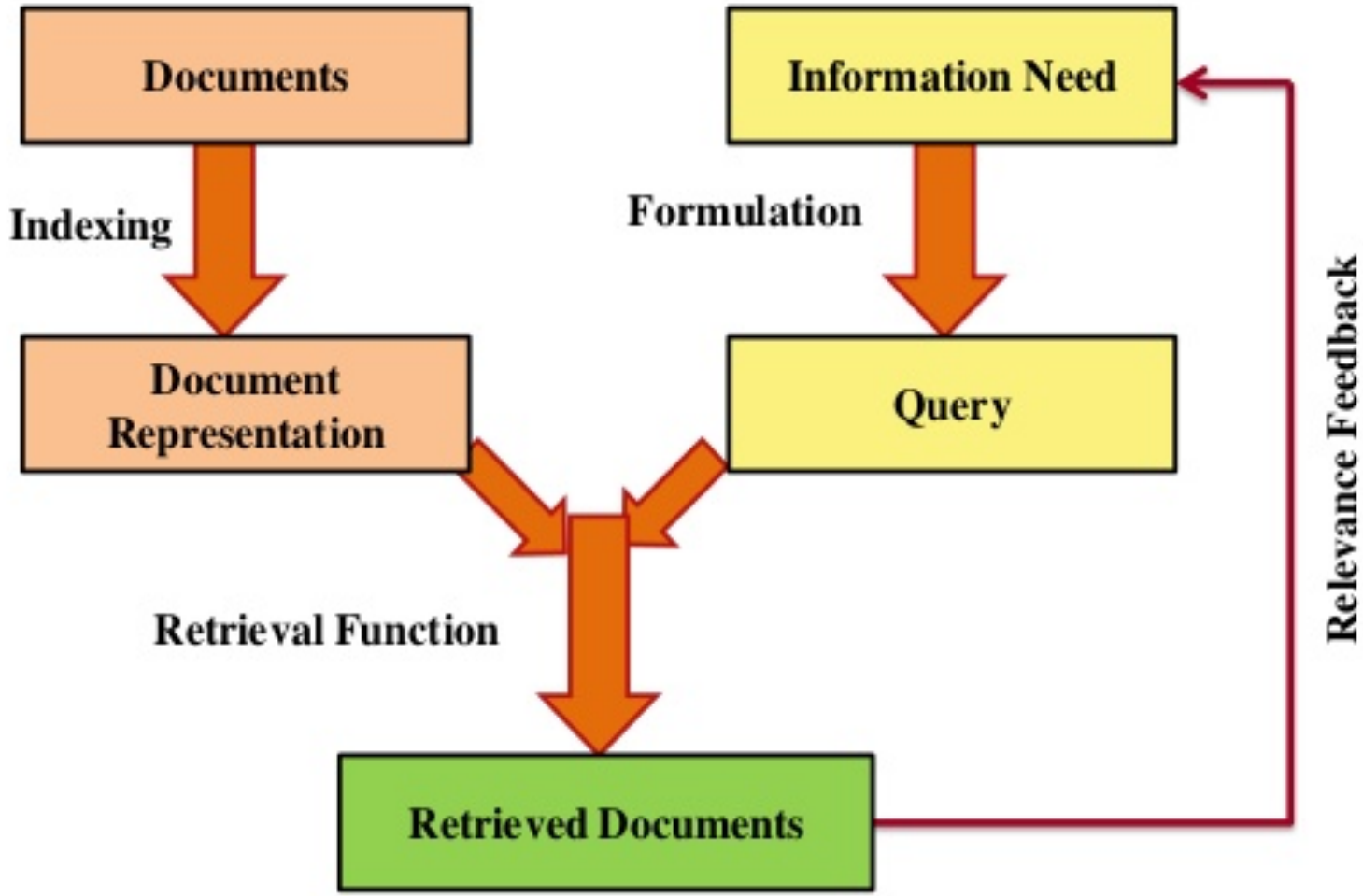
  Retrieve all relevant documents
- **Goal of an IR system** – PRECISION

  Retrieve the most relevant documents
- **Goal of an IR system**:
  - Retriev as few non-relevant documents as possible
  - Retrieve relevant documents before non-relevant documents

Some Topics in IR

- Retrieval models (ranking function, learning to rank, machine learning)
- Text processing (NLP techniques, language models)
- Interactivity and users
- Efficiency, compression, MapReduce, Scalability
- Distributed IR (data fusion, aggregated search, federated search)
- Multimedia: image, video, sound, speech
- Evaluation (crowdsourcing, user studies)
- Web retrieval and social media search
- Cross-lingual IR, Structured Data (XML)
- Digital libraries, Enterprise Search, Legal IR, Patent Search, Genomics IR

Conceptual Model of IR



Information Retrieval vs Information Extraction

- Information Retrieval
  - Given a set of terms and a set of document terms, select only the most relevant document (precision), and preferably all the relevant ones (recall)
- Information Extraction
  - Extract from the text what the document means
- IR can FIND documents but does not need to ”understand” them

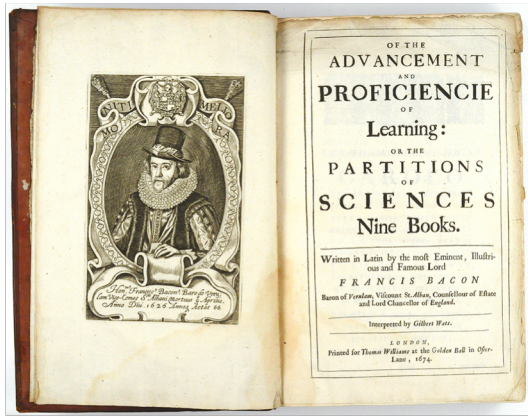
Information Retrieval vs Web Search

- Most people equate information retrieval with web search
- Information retrieval is concerned with **the finding of** relevant information

IR Beginnings (pre-1960)

There have always been libraries…

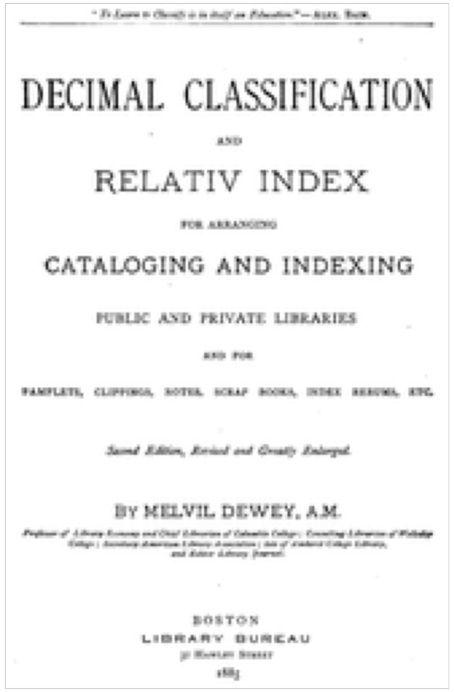
- Callimachus, a Greek poet in 3 BC, was the first known person to build a catalogue



- *Advancement of Learning* (1605) by F. Bacon: knowledge divided into three top categories (Memory, Reason, Imagination)

- Thomas Jefferson creates 42 new headings to organise his book collection

Library catalogues


DECIMAL CLASSIFICATION

AND

RELATIV INDEX

FOR ARRANGING

CATALOGING AND INDEXING

PUBLIC AND PRIVATE LIBRARIES

AND FOR

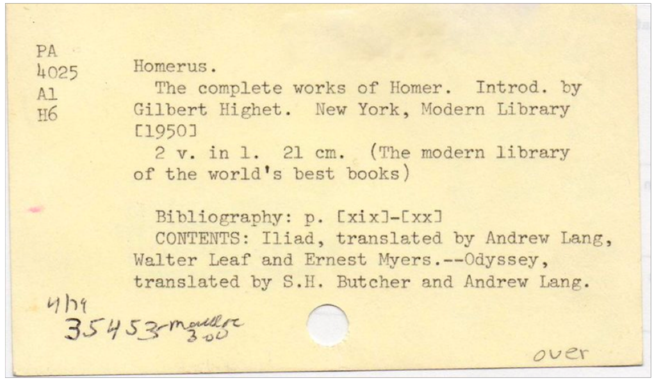
PAMFLETS, CLIPPINGS, NOTES, SCRAP BOOKS, INDEX RERUMS, ETC.

BY MELVIL DEWEY, A.M.

BOSTON

LIBRARY BUREAU

- *Dewey Decimal System* (1876) - literature divided into categories with Arabic numerals (000, 100, 200, etc.); first edition had 2000 entries
- An international version of the system (the *Universal Decimal Classification*) started in 1895 by the Belgian Paul Otlet; currently used in 130 countries
- Card catalogue – introduced by the French 1791 when confiscating library holdings of religious houses using the blank backs of playing cards.


PA
4025
A1
H6

Homerus.
The complete works of Homer. Introd. by Gilbert Highet. New York, Modern Library [1950]
2 v. in 1. 21 cm. (The modern library of the world's best books)

Bibliography: p. [xix]-[xx]
CONTENTS: Iliad, translated by Andrew Lang, Walter Leaf and Ernest Myers.--Odyssey, translated by S.H. Butcher and Andrew Lang.

4h4
35453
over

A typical title card (sorted by title)

F
Kee

The Clue of the Velvet Mask.

Keene, Carolyn.

The Clue of the Velvet Mask/ Carolyn Keene.

New York, Grosset, 1969c.

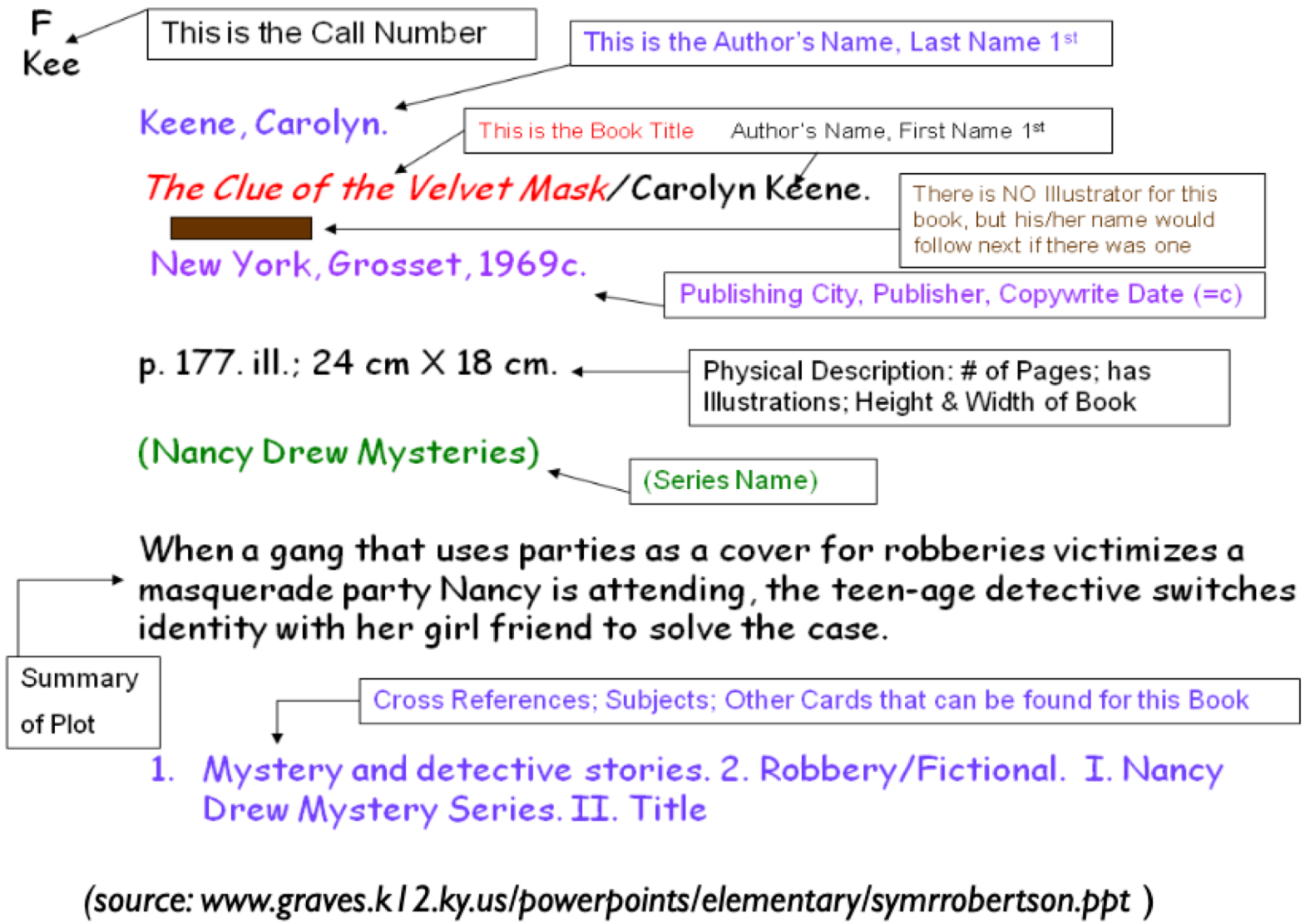
p. 177. ill.; 24 cm X 18 cm.
(Nancy Drew Mysteries)

When a gang that uses parties as a cover for robberies victimizes a masquerade party Nancy is attending, the teen-age detective switches identity with her girl friend to solve the case.

1. Mystery and detective stories. 2. Robbery/Fiction. I. Nancy Drew Mystery Series. II. Title

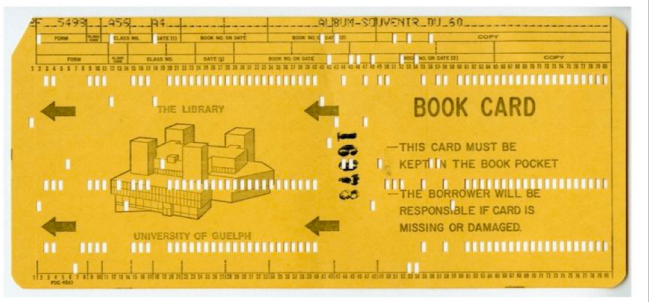
(source: www.graves.k12.ky.us/powerpoints/elementary/symrrobertson.ppt)

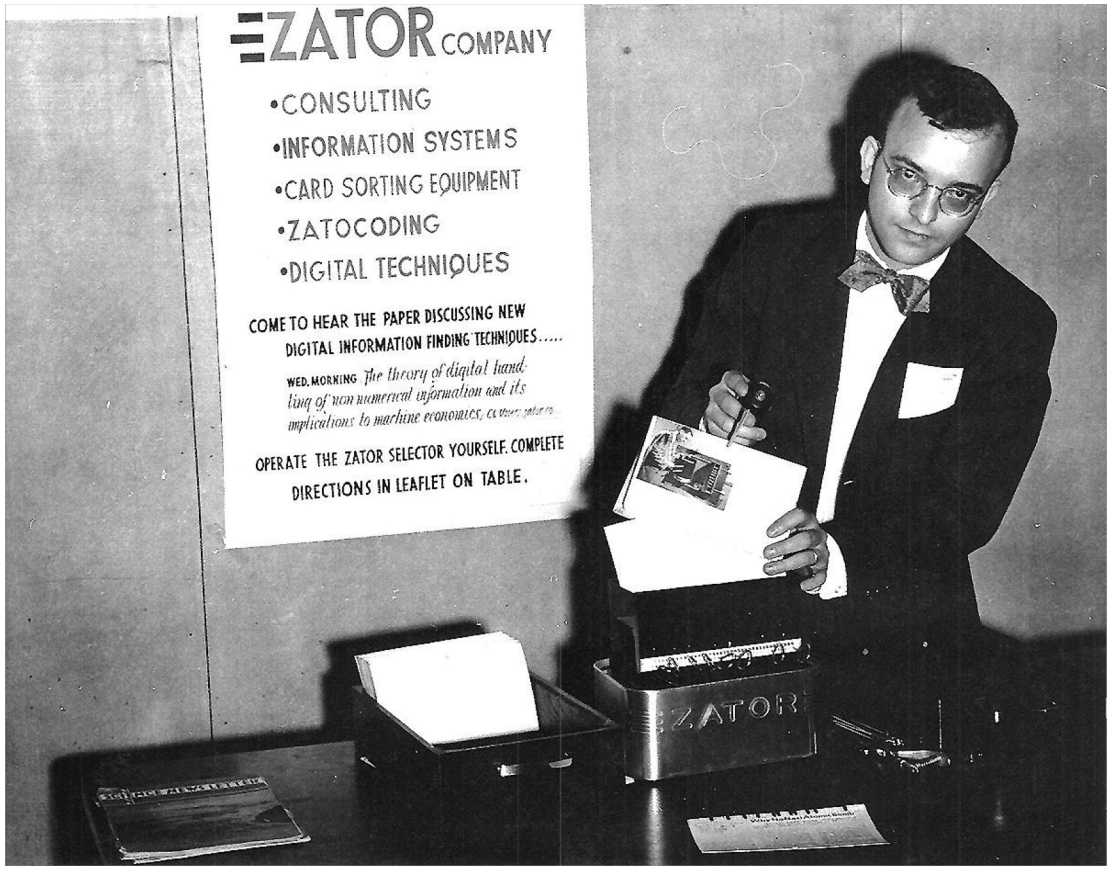
What's on a card?



(source: www.graves.k12.ky.us/powerpoints/elementary/symrrobertson.ppt)

Punchcards and Mechanical Devices

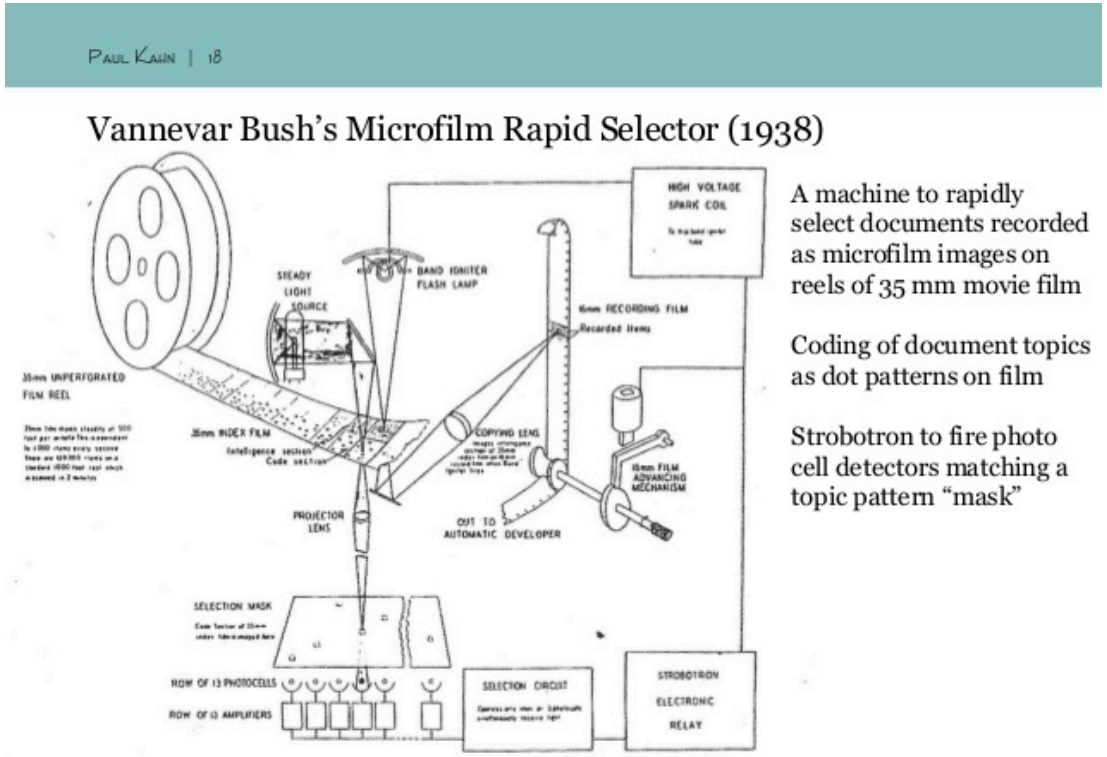




*Zato*r card company for document searching, founded by Calvin Mooers, 1947

*The problem of directing a user to stored information, some of which may be unknown to him, is the problem of "**information retrieval**"… In information retrieval, the addressee or receiver rather the sender is the active party.*

Calvin Mooers, 1950



Indexing

Coordinated and Uniterm Indexing System (Mortimer Taube 1951)

EXCURSION **43821**

90	**241**	52	63	34	25	66	**17**	58	49
130	281	92	83	**44**	75	86	57	88	119
640		122	93	104	115	146	97	158	139
							157	178	199
							207	248	269
								298	

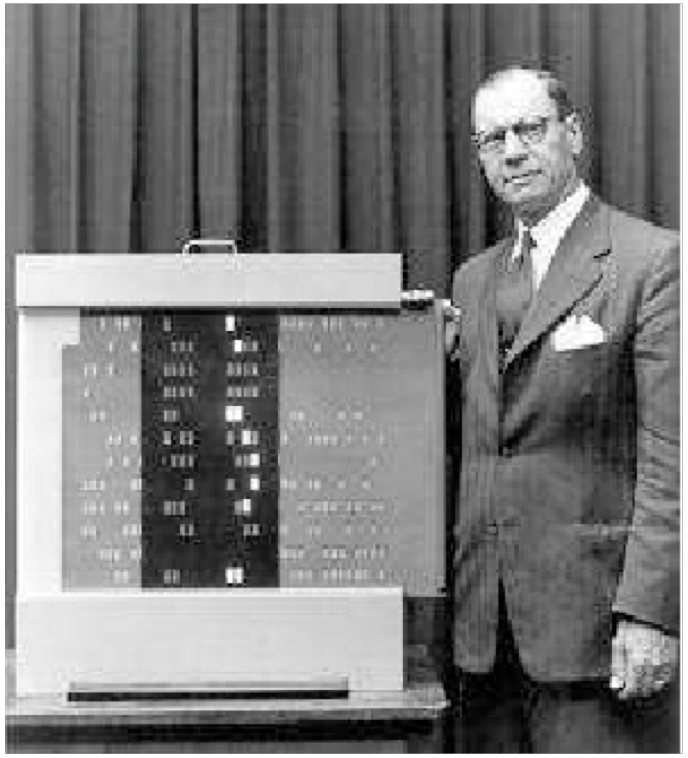
LUNAR **12457**

110	181	12	73	**44**	15	46	7	28	39
430	**241**	42	113	74	85	76	**17**	78	79
820	761	602	233	134	95	136	37	118	109
	901	982		194	165		127	198	179
							377	288	
							407		

Keyword-in-context (KWIC) system (Luhn 1960)

KEYWORD	Title of Article	Reference Code: Author	Year of Pubn.	Ident. Number
GREECE	ABANDONS PROPORTIONAL REPRESENTATION.=	POLYZO	AT29	961
BOARDS AND COMMISSIONS CREATED AND	ABOLISHED IN 1913.= STATE OFFICERS,	BATES	FG14	292
THE MONROE DOCTRINE	ABROAD IN 1823-24.=	ROBERT	WS12	225
JUDICIAL	ABROGATION OF COUNTY HOME RULE IN OHIO.=	SHOUP	EL36	1343
	ABSENT - VOTING IN NORWAY.=	SABY	RS18	474
MILITARY	ABSENT - VOTING LAWS.=	RAY	PO18	481
	ABSENT - VOTING LAWS, 1917.=	RAY	PO18	468
	ABSENT - VOTING LEGISLATION, 1924-1925.=	RAY	PO26	1910
	ABSENT VOTERS (LEGISLATION).=	RAY	PO14	294
	ABSENT VOTING (LEGISLATION).=	LAPP	JA16	374
	ABSENT VOTING LAWS.=	RAY	PO24	702
	ABSENT VOTING.=	KETTLE	C 17	426
	ABSENTEE VOTING IN THE UNITED STATES.=	STEINB	PG38	1456
D POLITICS.=	ABSOLUTISM AND RELATIVISM IN PHILOSOPHY AN	KELSEN	H 48	1941
RELATIVISM,	ABSOLUTISM, AND DEMOCRACY.=	OPPENH	F 50	2049
Y-- THE NATIONAL INTEREST VS. MORAL	ABSTRACTIONS.= THE MAINSPRINGS OF AMERICAN	MORGEN	HJ50	2039
SOCIAL SCIENCE	ABSTRACTS-- AN INSTITUTION IN THE MAKING.=	CHAPIN	FS30	1035
ON OF CASE STUDIES-- THE PROBLEM OF	ABUNDANCE (PUBLIC ADMINISTRATION).= PREP	STEIN	H 51	2071
PREME COURT DECISIONS-- THE USE AND	ABUSE OF QUANTITATIVE METHODS.= THE MATHEM	FISHER	FM58	2389

Luhn and Automatic Indexing



Hans Peter Luhn demonstrating a mock-up of an IBM card used in his scanner (1952).

Luhn's major contributions:

- Automatic indexing (using term frequency to select terms, KWIC)
- Automatic abstracting (summarization)
- Measuring similarity of documents based on their indexing terms
- Selective dissemination of information (filtering)
- Coined the term "business intelligence"

Luhn's idea: automatic indexing based on statistical analysis of text

"It is here proposed that the ***frequency of word occurrence*** in an article furnishes a useful measurement of word significance. It is further proposed that the ***relative position within a sentence*** of words having given values of significance furnish a useful measurement for determining the significance of sentences. The significance factor of a sentence will therefore be based on a combination of these two measurements. " (Luhn 1958)

LUHN, H.P., 'A statistical approach to mechanised encoding and searching of library information', *IBM Journal of Research and Development*,1, 309 - 317 (1957).

LUHN, H.P., 'The automatic creation of literature abstracts', *IBM Journal of Research and Development*, 2, 159 - 165 (1958).

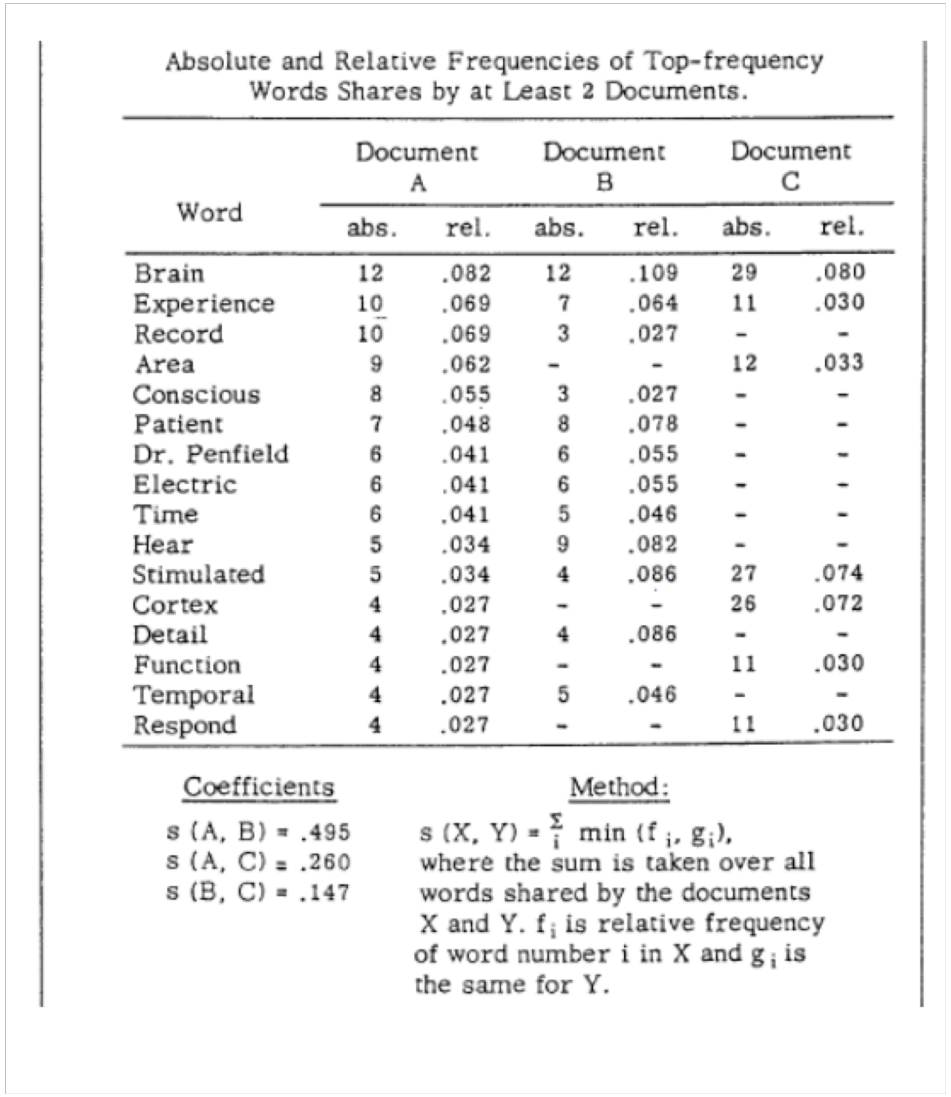
Key Word in Context (KWIC)

KWIC is an acronym for **Key Word In Context**, the most common format for concordance lines. The term KWIC was first coined by Hans Peter Luhn.

KWIC is an **acronym** for Key Word In Context, ...	page 1
... Key Word In Context, the most **common** format for concordance lines.	page 1
... the most common format for **concordance** lines.	page 1
... is an acronym for Key Word In **Context**, the most common format ...	page 1
Wikipedia, The Free **Encyclopedia**	page 0
... In Context, the most common **format** for concordance lines.	page 1
Wikipedia, The **Free** Encyclopedia	page 0
KWIC is an acronym for **Key** Word In Context, the most ...	page 1
KWIC is an acronym for Key Word ...	page 1
... common format for concordance **lines**.	page 1
... for Key Word In Context, the **most** common format for concordance ...	page 1
Wikipedia, The Free Encyclopedia	page 0
KWIC is an acronym for Key **Word** In Context, the most common ...	page 1

Sorted

Probabilistic representation and similarity computation (Luhn 1961)



Absolute and Relative Frequencies of Top-frequency Words Shares by at Least 2 Documents.

Word	Document A		Document B		Document C	
	abs.	rel.	abs.	rel.	abs.	rel.
Brain	12	.082	12	.109	29	.080
Experience	10	.069	7	.064	11	.030
Record	10	.069	3	.027	-	-
Area	9	.062	-	-	12	.033
Conscious	8	.055	3	.027	-	-
Patient	7	.048	8	.078	-	-
Dr. Penfield	6	.041	6	.055	-	-
Electric	6	.041	6	.055	-	-
Time	6	.041	5	.046	-	-
Hear	5	.034	9	.082	-	-
Stimulated	5	.034	4	.086	27	.074
Cortex	4	.027	-	-	26	.072
Detail	4	.027	4	.086	-	-
Function	4	.027	-	-	11	.030
Temporal	4	.027	5	.046	-	-
Respond	4	.027	-	-	11	.030

Coefficients

s (A, B) = .495
s (A, C) = .260
s (B, C) = .147

Method:

$s(X, Y) = \sum_i \min(f_i, g_i)$, where the sum is taken over all words shared by the documents X and Y. f_i is relative frequency of word number i in X and g_i is the same for Y.

An early idea about using unigram language model to represent text

Other early ideas related to indexing:

- [Joyce & Needham 58]: Relevance-based ranking, vector-space model, query expansion, connection between machine translation and IR
- [Doyle 62]: Automatic discovery of term relations/clusters, “semantic road map” for both search and browsing (and text mining!)
- [Maron 61]: automatic text categorization
- [Borko 62]: categories can be automatically generated from text using factor analysis
- [Edmundson & Wyllys 61]: local-global relative frequency (kind of TF-IDF)

SMART: System for Mechanical Analysis and Retrieval of Text

Gerard Salton
(Harvard, Cornell)

1961 – 1965: SMART system developed by Gerard Salton and Michael Lesk

- First automatic retrieval system
- Term weighting + vector similarity
- Experimented with many ideas for indexing
- Performed statistical significance test

Major findings:

- weighted terms are more useful than binary terms
- Cosine similarity is better than the overlap similarity measure
- automatic indexing is as good as manual indexing
- indexing based on abstracts outperforms titles
- the use of synonyms helps retrieval

About the SMART system

Developed on IBM 7094
(time-sharing system, 0.35 MIPS, 32KB memory)

Early development (1961 - 1965):
Michael Lesk

First UNIX implementation (v8, 1980):
Edward Fox

The widely used SMART toolkit (v 10/11, 1980 – 1990s):
Chris Buckley

SMART was the most popular IR toolkit (in C) widely used in 1990s by IR researchers and some machine learning researchers.

The Cranfield Evaluation Methodology

- IR is an empirically defined problem, thus experiments must be designed to test whether one system is better than another
- However, early work on IR (e.g., Luhn's) mostly proposed ideas without rigorous testing
- Catalysts for experimental IR:
  - Hot debate over different languages for manual indexing
  - Automatic indexing vs. manual indexing
- How can we experimentally test an indexing method?

Cleverdon's Cranfield Tests

1957 - 1960: Cranfield I
- Comparison of indexing methods
- Controversial results (lots of criticisms)

1960 - 1966: Cranfield II
- More rigorous evaluation methodology
- Introduced precision & recall
- Decomposed study of each component in an indexing method
- Still lots of criticisms, but laid the foundation for evaluation that has a very long-term and broad impact

Cyril Cleverdon
Librarian, Cranfield Institute of Technology, UK

Cleverdon received the ACM SIGIR Salton Award in 1991
URL : http://www.sigir.org/awards/awards.html

Cranfield II: Experimental Design

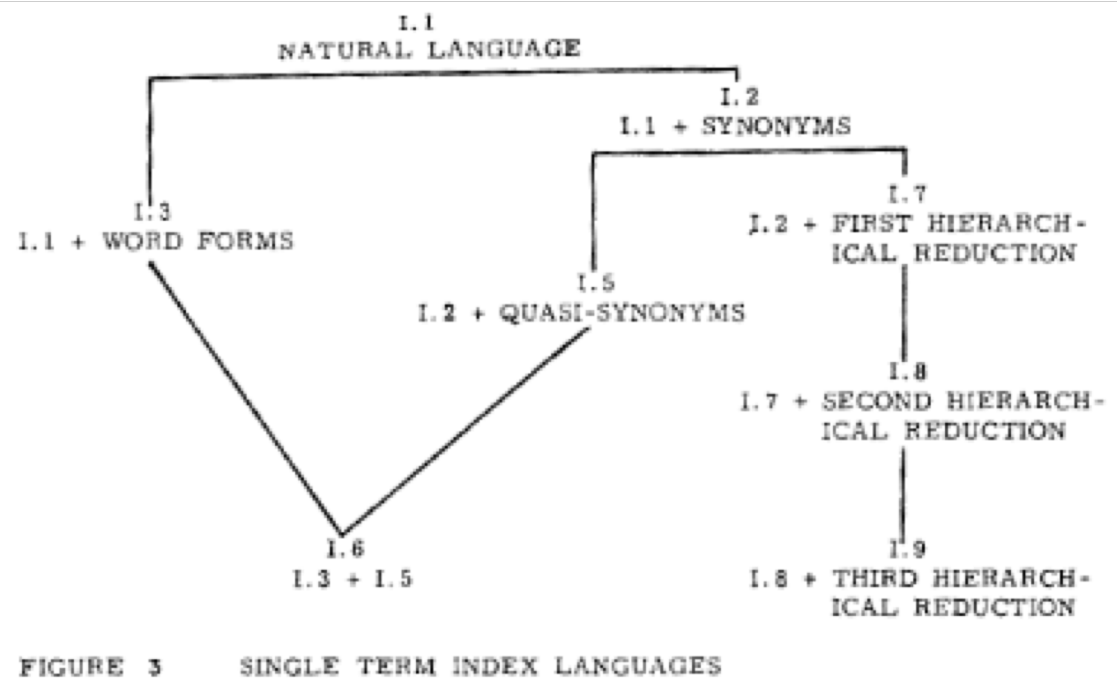


FIGURE 3 SINGLE TERM INDEX LANGUAGES

- Decomposed study of contributions of different components of an indexing language
- Rigorous control of evaluation
  - Having complete judgments is more important than having a large set of documents
  - Document collection: 1400 documents (cited papers by 200 authors, no original papers by these authors)
  - Queries: 279 questions provided by authors of original papers
  - Relevance judgments:
    - Multiple levels: 1 – 5
    - Initially done by 6 students in 3 months; final judgments by the originators
  - Measures: precision, recall, fallout, prec-recall curve
  - Ranking method: coordination level (# matched terms)

Measures: Precision, Recall, Fallout

	RELEVANT	NON-RELEVANT	
RETRIEVED	a	b	a + b
NOT RETRIEVED	c	d	c + d
	a + c	b + d	a + b + c + d = N (Total Collection)

FIGURE 1 2 x 2 CONTINGENCY TABLE

For the purpose of evaluating an information retrieval system, performance is presented by plotting the recall ratio $\left(\frac{100a}{a+c}\right)$ against either the precision ratio $\left(\frac{100a}{a+b}\right)$ or the fallout ratio $\left(\frac{100b}{b+d}\right)$. The fallout ratio is particularly useful when comparing performances of document collections of different sizes, but the precision ratio is more satisfactory for most of the results obtained in the Cranfield work.

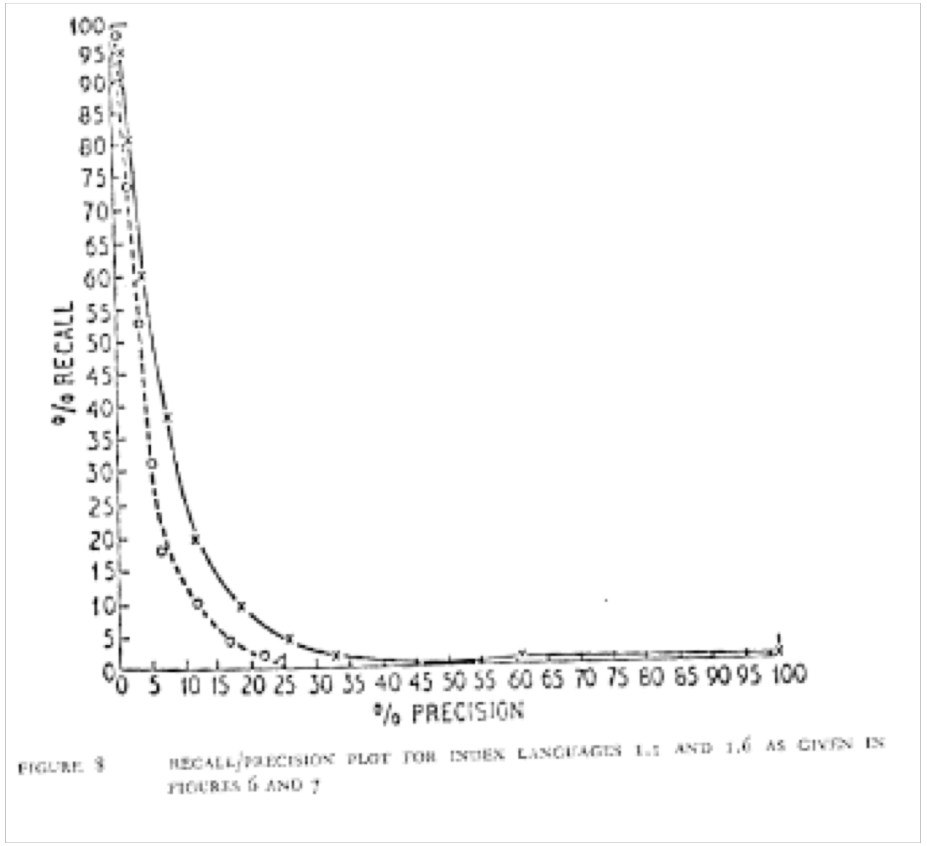


FIGURE 8 RECALL/PRECISION PLOT FOR INDEX LANGUAGES 1.5 AND 1.6 AS GIVEN IN FIGURES 6 AND 7

Cleverdon, C. W., 1967, The Cranfield tests on index language devices. Aslib Proceedings, 19, 173 - 192.

Cranfield II: Results

ORDER	NORMALISED RECALL	INDEXING LANGUAGE	
1	65.82	I-3	Single terms. Word forms
2	65.23	I-2	Single terms. Synonyms
3	65.00	I-1	Single terms. Natural Language
4	64.47	I-6	Single terms. Synonyms, word forms, quasi-synonyms
5	64.41	I-8	Single terms Hierarchy second stage
6	64.05	I-7	Single terms. Hierarchy first stage
7a	63.05	I-5	Single terms. Synonyms. Quasi-synonyms
7a	63.05	II-11	Simple concepts. Hierarchical and alphabetical selection
9	62.88	II-10	Simple concepts. Alphabetical second stage selection
10a	61.76	III-1	Controlled terms. Basic terms
10a	61.76	III-2	Controlled terms. Narrower terms
12	61.17	I-9	Single terms. Hierarchy third stage
13	60.94	IV-3	Abstracts. Natural language
14	60.82	IV-4	Abstracts. Word forms
15	60.11	III-3	Controlled terms. Broader terms

Major findings:

- Best performance obtained by the use of Single Term index language
- With these Single Term index languages, the formation of groups of terms or classes beyond the stage of true synonyms or word forms resulted in a drop of performance.
- The use of precision devices such as partitioning and interfixing was not as effective as the basic precision device of coordination

Criticism:

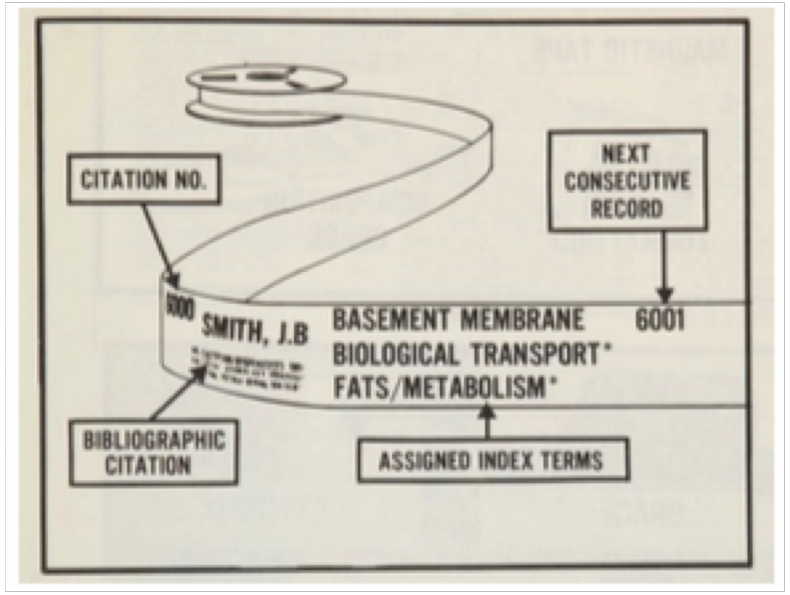
- The test did not reflect an ordinary operating system situation (inappropriate to a laboratory test)
- Unrealistic assessment procedure – queries not appropriate for the test articles
- Lack of statistical tests

Cranfield Test Methodology

- Specify a retrieval task
- Create a collection of sample documents
- Create a set of topics/queries appropriate for the retrieval task
- Create a set of relevance judgments (i.e., judgments about which document is relevant to which query)
- Define a set of measures
- Apply a method to (or run a system on) the collection to obtain performance figures

MEDLARS: Medical Literature Analysis and Retrieval System

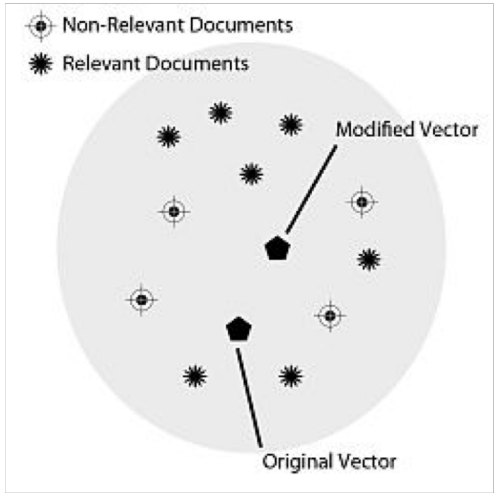
- Launched by the National Library of Medicine in 1964 with Index Medicus
- The first large-scale computer based search service available to the general public
- By April 1965, there were 265,000 citations in the database.
- Manual indexing created unit records for each citation, which were put into paper tape for input to the computer
- Users completed a search request form, which was converted by a trained medical librarian into the search format
- The request was passed against the entire file of citations, which took about 40 mins.



Into the 70's….

- Request Expansion using Relevance Feedback – the Rocchio algorithm
- Clustering experiments with the SMART system
- Inverted Document Frequency (IDF) – Sparck-Jones (1972)

  "… all terms should be allowed to match but the value of matches on frequent terms should be lower than that for non-frequent terms" (Sparck-Jones, 2004)



- Development of online retrospective search – June 1970, MEDLARS initiates an experimental service for online access to their database

Into the 70's: IR research enters a theory bulding phase

- Investigating statistical properties of term frequencies (Bookstein & Swanson 1974, Salton et al. 1974, 1975)
- Investigating term frequency properties based on relevancy (Robertson & Sparck-Jones 1976)

	Relevant	Non-relevant	
Indexed	r	$n-r$	n
Not indexed	$R-r$	$N-n-R+r$	$N-n$
	R	$N-R$	N

- The probabilistic theory of relevance weighting:
  - ***Document ordering hypothesis***: For optimum performance, the systems should order the documents and allow the searcher to search down the ordered list as far as s/he wants to go.
  - ***Probability ranking hypothesis:*** For optimum performance, the system should rank the documents according to their probability of being judged relevant or useful to the user's problem or information need. (Robertson 1977)

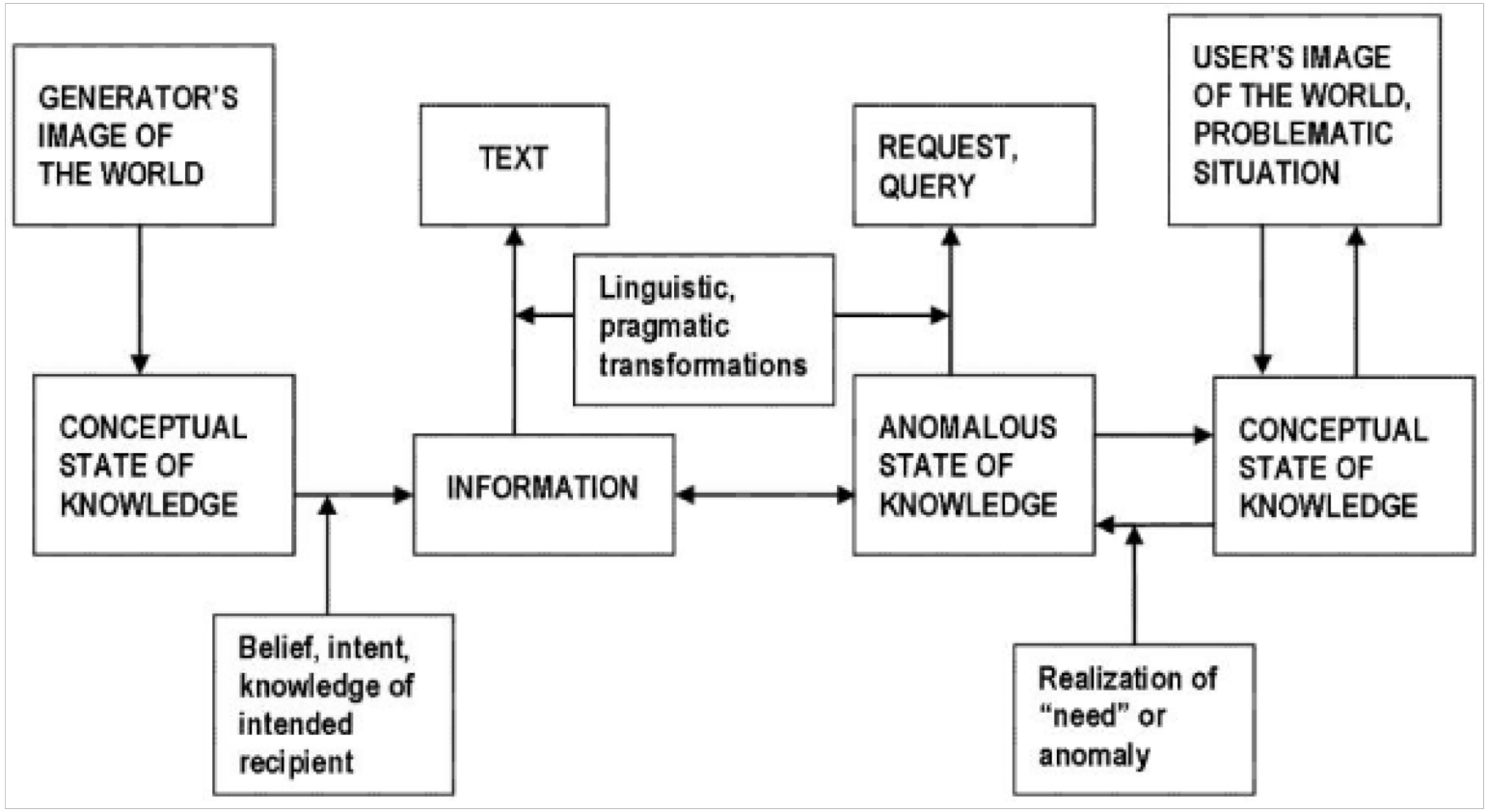
Into the 70's: the IR community expands

- The first annual ACM SIGIR (Special Interest Group for Information Retrieval) conference was held in May of 1978 in Rochester, N.Y. with 14 papers. The second SIGIR took place in Dallas, Texas, again with 14 papers and one panel. In 1980 the conference moved to Cambridge, U.K. and had expanded to 23 papers.

- Early prototypes IR systems develop:
  - SIRE (1976) – combined Boolean retrieval and SMART
  - THOMAS (1977) – built to explore MEDLARS data and based on having a dialogue with a user
  - CITE (1979) – based on MEDLARS, the system started with a user's natural language query and provided ranked output, relevance feedback and other query expansion methods.

Early 80's: research with users

ASK (Anomalous States of Knowledge, Belkin 1980) -- user is trying to fill in a gap in their knowledge, but this gap might be difficult to specify due to its complexity and the ease of expressing that need to a retrieval system. Information need has to be defined in terms of users rather than the system.



Early 80's: online services

- **In-house systems**, e.g. MASQUERADE (Brzozowski, 1983) for technical reports; CUPID (Cambridge University Probabilistic Independence Datamodel) (Porter, 1982) – could index and search 10,000 documents.

- **Online catalogues**: OCLC (Online Computer Library Centre) and RLIN (Research Libraries Information Network) – unions of catalogues of many libraries

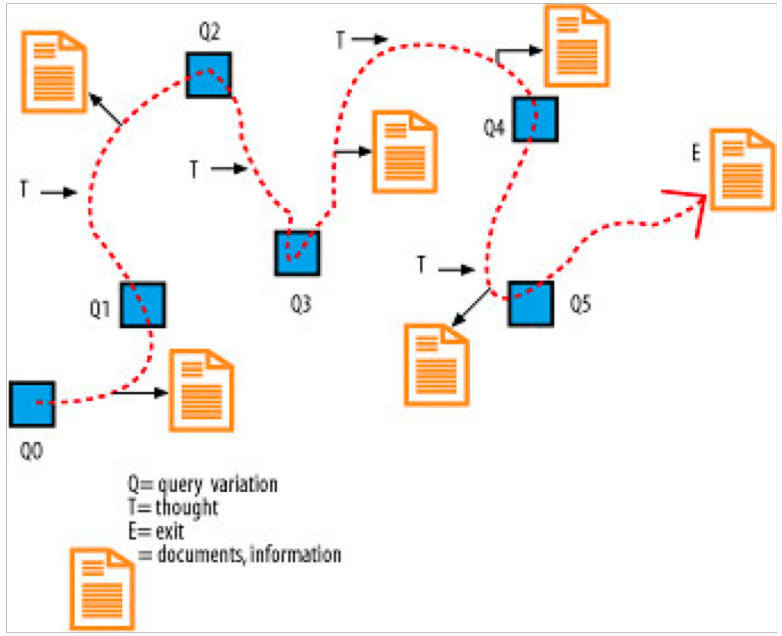
- **Science abstracting/indexing services**: BIOSIS (Biological Abstracts and BioResearch Index) with over 4M references by 1984; SCISearch – database from the Institute of Scientific Information that included references from over 4000 journals

- **Legal databases**: LEXIS and WESTLAW – full text databases

Late 80's

- Online card catalogues (OPAC) and experiments with the OPAC operational setting (the Okapi system), including variations of IDF, stemming and spelling corrections, effects of relevance feedback

- Rethinking user interfaces: increased thinking about the "end-users" as opposed to the search intermediaries and new models for end user searching were being proposed, e.g. Berrypicking (Bates, 1989)



The 1990's and arrival of the search engine

- In 1990 **Archie** was released by Peter Deutsch, Alan Emtage, and Bill Heelan at McGill University. Archie was a "search engine" that allowed users to log into a specific site (an Archie server) and using command lines, search for data that had been previously collected for that server.

- In early 1991 Tim Berners-Lee designed the **HyperText Transfer Protocol** (HTTP), the **HyperText Markup Languag**e (HTML) and the first Web browser for the NeXt environment.

- In July 1992 the **WWW** client software was made publicly available by CERN

- In January 1993 Marc Andreessen from the National Center for Supercomputing Applications (NCSA) at the University of Illinois in Urbana/Champlain released the **Mosaic** web browser, based on the Berners-Lee proposal, but built for the UNIX operating systems.

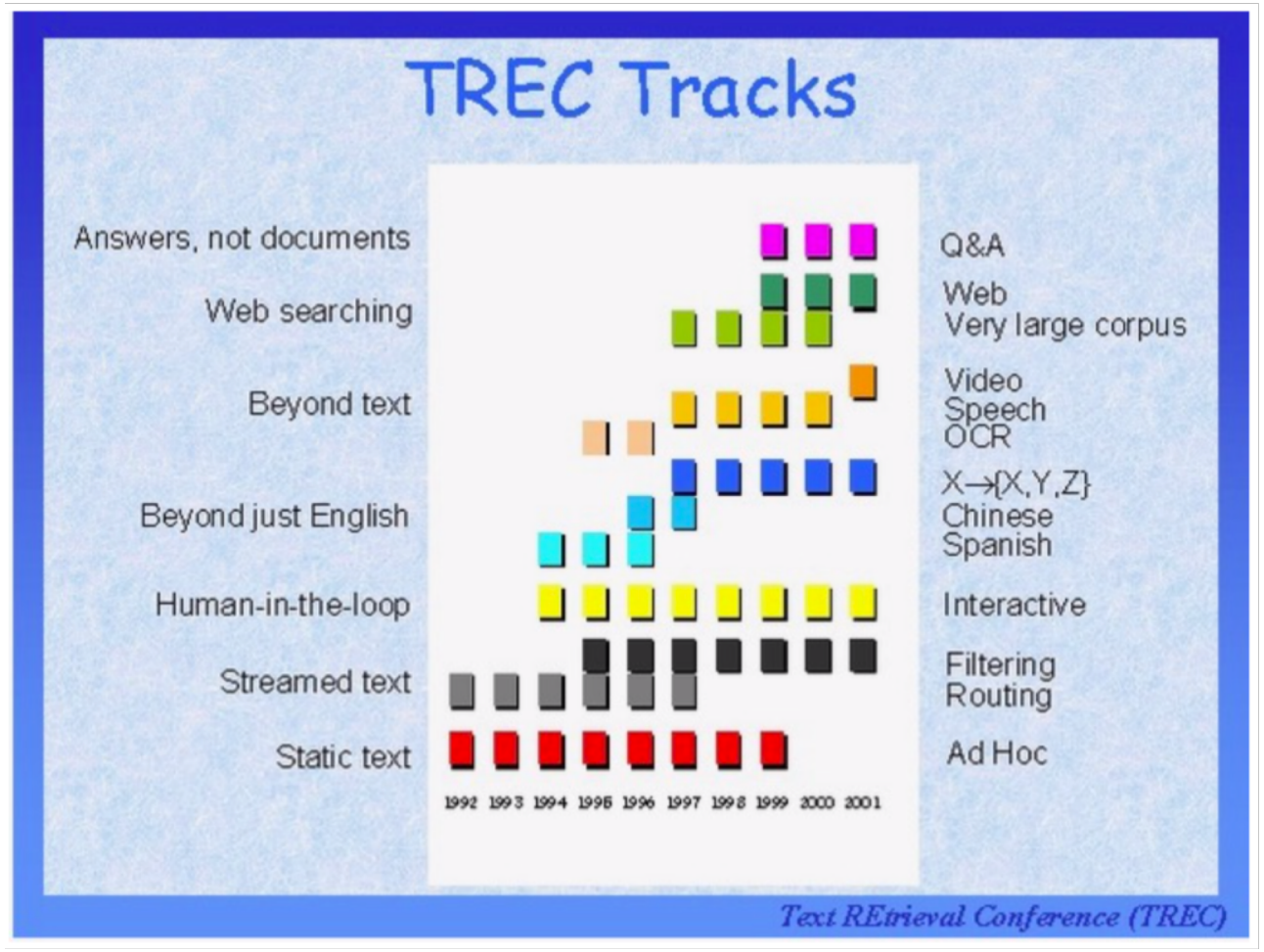
The 1990's and the arrival of the search engine

Incredible growth of the web and rapid emergence of search engines:

- 1993 – **Mosaic** released in January; 130 websites in June grew to 623 by December.
- 1994 – **Yahoo!** started in April; first WWW meeting held in May; **Lycos** went public in July; 10,222 websites by December.
- 1995 – **Infoseek** started in February; **Excite** started in October; **AltaVista** launched by DEC in December with 300,000 hits on its first day.
- 1996 – In January there were 100,000 websites, doubling by June with over half being ".com" sites.
- 1998 – **Google** Search started; **Microsoft** started a search portal called **MSN Search**, using search results from Inktomi. It did not have in-house searching until 2005 and changed its name to **Bing** in 2009.

TRIPSTER and TREC

- Late 1990: DARPA launches TRIPSTER to advance information extraction
- Test collection based on the Cranfield paradign created: 2 gig of documents from multiple domains, 50 queries, documents selected for assessment by the pooling method
- This test corpus was used in 1992 at the first TREC (Text REtrieval Conference); more test collections added over the years



And more research continues...

- **Basic retrieval algorithms**, e.g. the BM25 algorithm (Robertson & Walker 1994), neural nets (Kwok, 1995), Latent Semantic Indexing (Caid et al. 1995)
- **Extensive user studies**: pariculary at Rutgers (Belkin group), Xerox (M. Hearst), UMass (James Allan)
- **Text categorization and filtering**: Reuters and the Carnegie Group experiment with an automatic methods
- New **retrieval models** for ranking
- Research to improve **web performance**
- **Evaluation**: Kalervo Järvelin and Jaana Kekäläinen of the University of Tampere propose a new metric using graded relevance judgments and then accumulating scores while moving down the ranked list. This discounted cumulative gain metric and its successor, the normalized Discounted Cumulative Gain (**nDCG)**, have been heavily used by both the web community and the IR community.