Prof. Dr. Maximilian Eibl
eibl@informatik.tu-chemnitz.de

PROFESSUR
MEDIENINFORMATIK

02 Evaluation of Information Retrieval Systems

Lecture Media Retrieval
Maximilian Eibl, Medieninformatik, TU Chemnitz

Why Empirical Evaluation?

- Complexity of the subject
  - Expressing the information need
  - Matching queries and documents
  - Ranking
  - Mathematical / statstical approaches to assess semantics
  - Semantic assessment by humans is not coherent

→ Quality of IRS needs to be measured empirically

Prof. Dr. Maximilian Eibl
eibl@informatik.tu-chemnitz.de

General Challenge

”In all cases, evaluation of Information Retrieval Systems will suffer from the subjective nature of information. There is no deterministic methodology for understanding what is relevant to a user’s search.”
(Kowalski, 1997: 244)

Special Challenge Multimedia

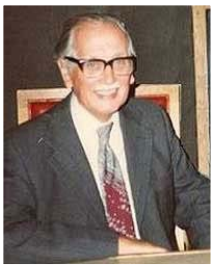
- Vagueness of images
- Semantics highly depends on context and subject
- No analogy to textual units like ‚word‘ or ‚sentence‘
- No well defined similarities



Prof. Dr. Maximilian Eibl
eibl@informatik.tu-chemnitz.de

Classic Test Setup: Cranfield-Paradigma

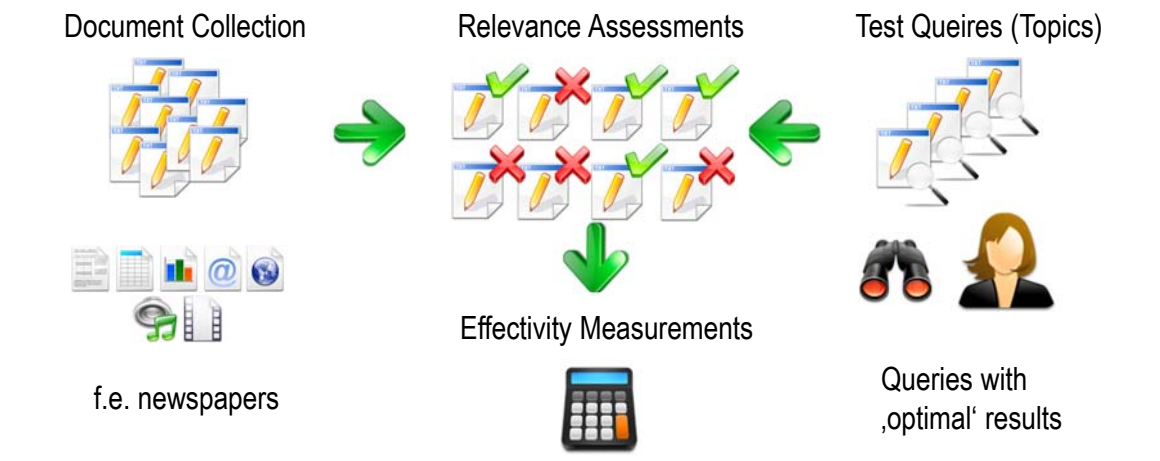
1957-1967: Cranfield Project: experiments studying the quality of index languages in an controlled environment




Cyril Cleverdon

→ Comprehensive test collections for compairing different approaches in IR

Classic Test Setup



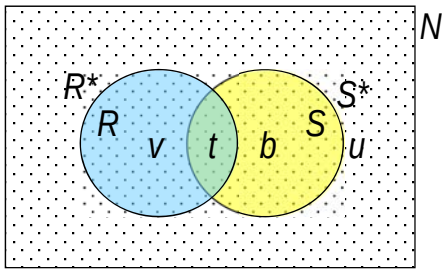
Prof. Dr. Maximilian Eibl
eibl@informatik.tu-chemnitz.de

Examples of Collection Sizes

Colllection	Cranfield	CACM	TREC2
Size (documents)	1.400	3.204	742.611
Size (MB)	1.5	2.3	2.162
Release year	1968	1983	1991
Words	8.226	5493	1.040.415
Word occurences	123.200	117.578	243.800.000
Average word count per document	88	36	328
Number of topics	225	50	100

Classic Measurements: *recall* & *precision*

Assessment… …by juror / … by system	relevant	non-relevant	sum
… found	t	b	S
…not found	v	u	S*
sum	R	R*	N



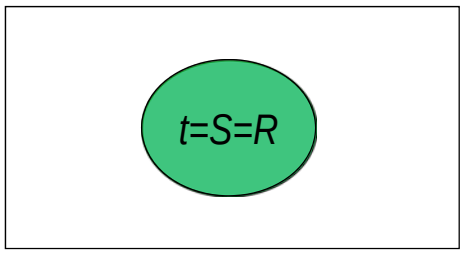
→ Which set would you want to be as big as possible?

$$recall = \frac{t}{R}$$

$$precision = \frac{t}{S}$$

Prof. Dr. Maximilian Eibl
eibl@informatik.tu-chemnitz.de

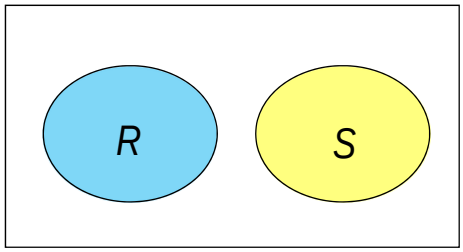
recall / *precision*: **Extremwerte**



recall = 1

precision = 1

(all documents found are relevant)



recall = 0

precision = 0

(none of the documents found is relevant)

recall / *precision*: **Graph**

recall / *precision-graph* with 10 standard measuring points:

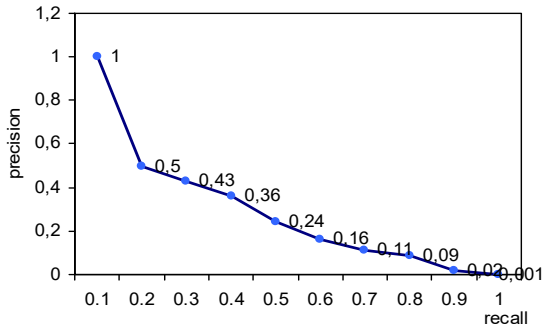
recall: 0.1 – 0.2 – 0.3 – …. – 1

example:

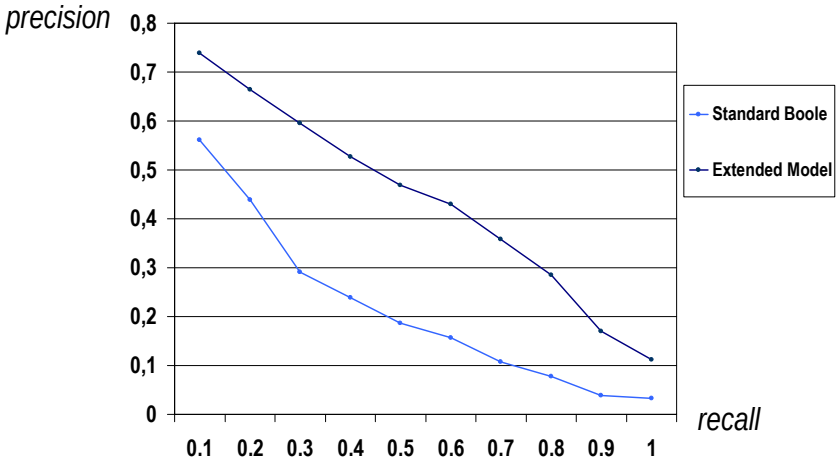
R={doc_{10}, doc_{21}, doc_{25}, doc_{50}, doc_{62}, doc_{70}, doc_{100}, doc_{105}, doc_{150}, doc_{198} }

recall / precision graph

Ranking for query q:	*recall*	*precision*	
1. doc_{10}	0.1	1	(1 / 1)
2. doc_{13}	0.2	0.5	(2 / 4)
3. doc_{14}			
4. doc_{21}	0.3	0.43	(3 / 7)
5. doc_{128}	0.4	0.36	(4 / 11)
6. doc_{250}	…	…	
7. dok_{25}			
8. doc_{158}			
9. doc_{22}			
10. doc_{270}			
11. doc_{50}			
12. doc_{526}			
13. doc_{566}			
14. doc_{62}			
15. doc_{567}			



recall / precision: example



Salton et al. 1983: Fig.5

recall & *precision*

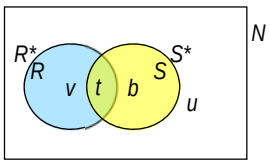
- Do not work independently
  - Recall increases with amout of retrieved documents
  - Increasing recall -> decreasing precision
- Importance depends on context
  - Expert systems, file search: Recall optimized
  - Web: Precision-optimized

Fallout

Problems with *recall* / *precision*:

- Mathematical problems: Division by 0 if no relevant documents exist (recall) or no relevant documents are found (precision)
- *recall* and *precision* behave strictly opposite

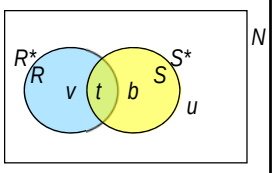
→ Effektivity $fallout = \frac{b}{R*}$



(Abfallquote)

Complementary Measurements

- miss ratio (↔ recall) - *Fehlquote* $miss\ \ ratio = \frac{v}{R}$
- noise ratio (↔ precision) - *Ballastquote* $noise\ ratio = \frac{b}{S}$
- rejection ratio (↔ fallout) - *Abweisquote* $rejection\ ratio = \frac{u}{R*}$

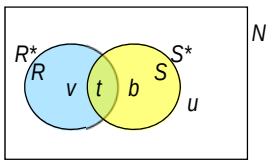


miss ratio = 1 – *recall*
noise ratio = 1 - *precision*
rejection ratio = 1 - *fallout*

Generality

- Defines proportion of relevant documents:

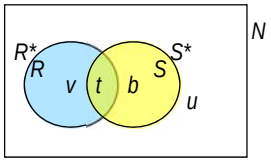
$$generality = \frac{R}{N}$$



User Centered Measures

- Coverage

$$\frac{\text{documents known by the user}}{t}$$



- Novelty

$$\frac{\text{documents not known by the user}}{t}$$

Calculating Averages

- Situation in a user test
  - Several users
  - Several topics

→many results: How kann you calculate one single score?

Macro method:

$$\mu(r_i) = \frac{1}{n}\sum_{i=1}^{n}\frac{r_i}{n_i}$$

Micro method:

$$\mu(r_i) = \frac{\sum_{i=1}^{n} r_i}{\sum_{i=1}^{n} n_i}$$

Prof. Dr. Maximilian Eibl
eibl@informatik.tu-chemnitz.de

Example Averaging for *recall*

Retrieval results	1	2	3	4	5	∑
Number of found relevant documents *t*	10	5	20	1	30	66
Number of existing relevant documents *R*	100	5	30	10	40	158
recall r	0.1	1	0.66	0.1	0.75	2.616

Micro: $r = \frac{10+5+20+1+30}{100+5+30+10+40} = \frac{66}{158} = 0.418$

Macro: $r = \frac{0.1+1+0.\overline{6}+0.1+0.75}{5} = \frac{2.61\overline{6}}{5} = 0.52\overline{3}$

Example Averaging for *recall*

Retrieval results	1	2	3	4	5	∑
Number of found relevant documents *t*	10	5	20	1	30	15
Number of existing relevant documents *R*	100	5	30	10	40	105
recall r	0.1	1	0.66	0.1	0.75	1,1

Micro: $r = \frac{10+5}{100+5} = \frac{15}{105} = 0.143$ (over all: 0.418)

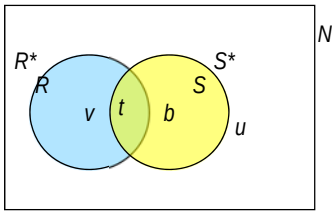
Macro: $r = \frac{0.1+1}{2} = \frac{1.1}{2} = 0.55$ (over all: 0.523)

MAP – Mean Average Precision

- Calculate by macro method
- Base:
  - Average Precision for each query
  - Standard points of the recall-precision-graph

Problem #1: How to get to R

- Initial situation:
  - N often > 6-7 digit number
  - R three digit number
- Methods of getting to R:
  - Assessment by juror(s)
  - Estimation
  - Pooling



Problem #2: Relevance

„There seems to be general agreement […] that the information retrieval process will never be fully understood without a prior understanding of that elusive notion called ‚relevance'. ‚Relevance' is one of the most fundamental, if not *the* fundamental, concept encountered in the theory of information retrieval."

Cooper 1971:19

Definition of Relevance

Relevance is the (A) *gage of relevance* of an (B) *aspect of relevance* existing between an (C) *object judged* and a (D) *frame of reference* as judged by an (E) *assessor*, where:

(A)	(B)	(C)	(D)	(E)
measure	utility	document	question	requester
degree	matching	document representation	question representation	intermediary
extent	informativeness	reference	research stage	expert
judgment	satisfaction	textual form	information need	user
estimate	appropriateness	information provided	information used	person
appraisal	usefulness	fact	point of view	judge
relation	correspondence	article	request	information specialist

Saracevic 1970b: 121 und 1975: 328 zitiert nach Schamber et al. 1990: 761

Prof. Dr. Maximilian Eibl
eibl@informatik.tu-chemnitz.de

Problems of Relevance Assessment I

- System centered vs. user centered relevance
  - Cooper 1971: logical relevance vs. utility
  - Salon&McGill 1983: objective vs. subjective relevance
- Judgements by users are not static but change over time

→ „Subjective, depending on human (user or nonuser) judgment and thus not an inherent characteristic of information or a document."
(Schamber 1994: 6)

Problems of Relevance Assessment II

- Criteria (Barry 1994):
  - Amount of information in the dokuments
  - User's background and knowledge
  - User's personal preferences
  - Source of the documents
  - Comparison to other information sources
  - Physical access to the document
  - User's personal situation

Relevance Assessment: Practice

- Relevance ~ Is the needed information in the document
- four-level scale:
  - Certainly relevant: Contains all information needed for answering the question
  - Possibly relevant: Contains essential information needed for answering the question
  - Less relevant: Contains few information needed for answering the question
  - Not relevant: Contains no information needed for answering the question

Problem #3: Collections

- Specific or general in context (f.e. scientific or general interest)
- Short descriptions or long texts
- Balanced or unbalanced
- Mediality

→ collection affects algorithms