



**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 / statistical approaches to assess semantics
 - Semantic assessment by humans is not coherent

→ Quality of IRS needs to be measured empirically



General Challenge

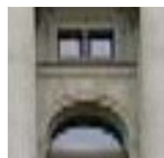
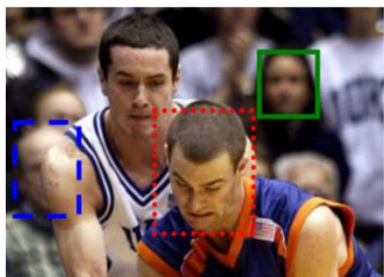
"In all cases, evaluation of Information Retrieval Systems will suffer from the subjective nature of information. There is no deterministic methodology for understanding what is relevant to a user's search."

(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





Classic Test Setup: Cranfield-Paradigm

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

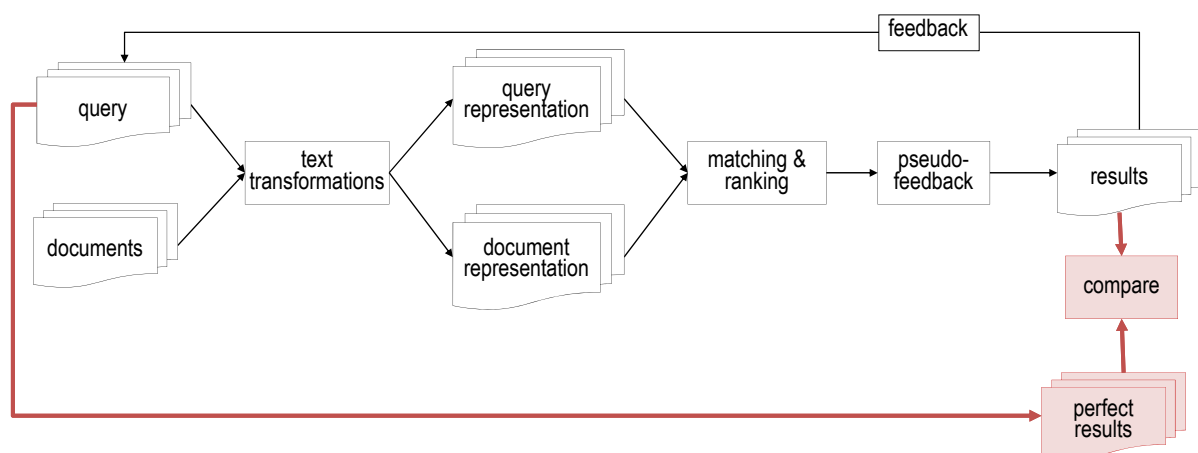


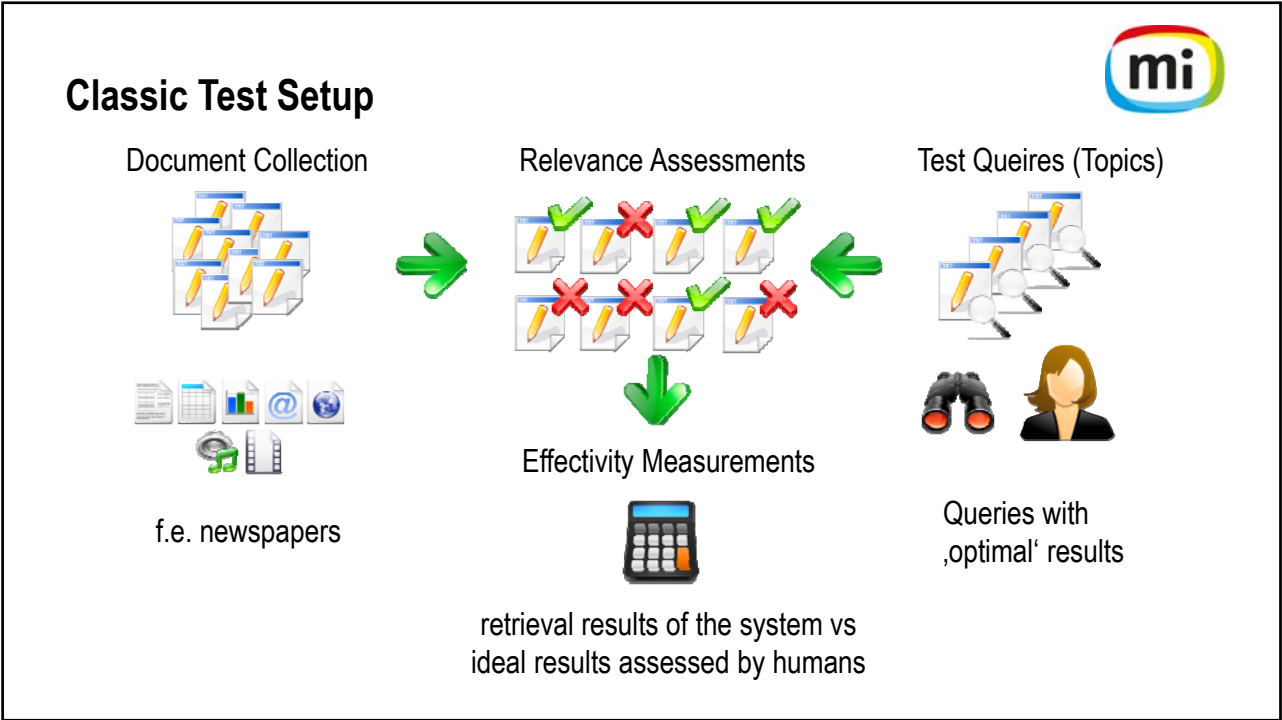
Cyril Cleverdon




→ Comprehensive test collections for comparing different approaches in IR

IRS Components: Text Retrieval







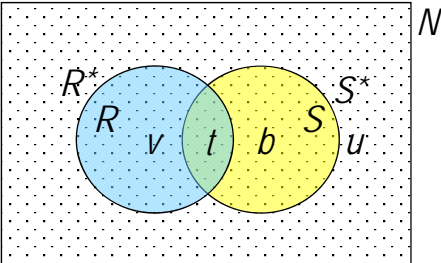
Examples of Collection Sizes

Collection	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

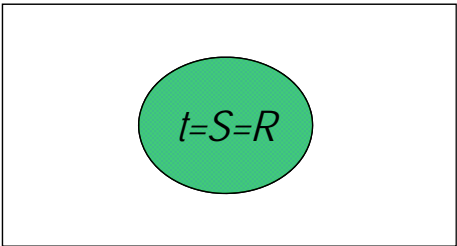


→ Which set would you want to be as big as possible?

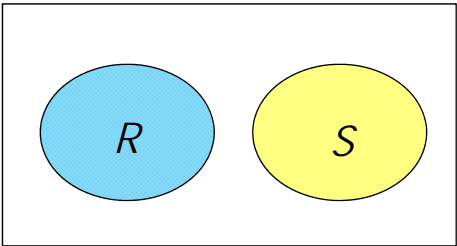
$$recall = \frac{t}{R}$$

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

recall / precision: extremes



$recall = 1$
 $precision = 1$
(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

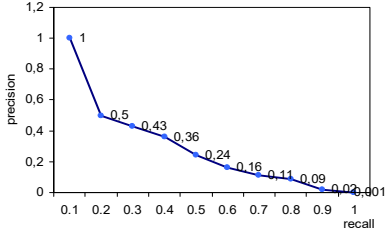
example:
 $R = \{doc_{10}, doc_{21}, doc_{25}, doc_{50}, doc_{62}, doc_{70}, doc_{100}, doc_{105}, doc_{150}, doc_{198}\}$



recall / precision graph

Ranking for query q: *recall* *precision*

1.	doc₁₀	}	0.1	1	(1 / 1)
2.	doc ₁₃		0.2	0.5	(2 / 4)
3.	doc ₁₄				
4.	doc₂₁	}	0.3	0.43	(3 / 7)
5.	doc ₁₂₈		0.4	0.36	(4 / 11)
6.	doc ₂₅₀				
7.	doc₂₅	}
8.	doc ₁₅₈				
9.	doc ₂₂				
10.	doc ₂₇₀				
11.	doc₅₀				
12.	doc ₅₂₆				
13.	doc ₅₆₆				
14.	doc₆₂				
15.	doc ₅₆₇				

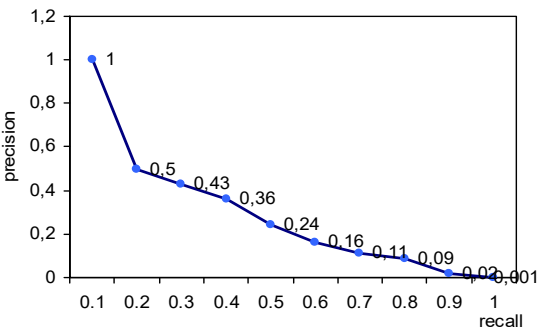




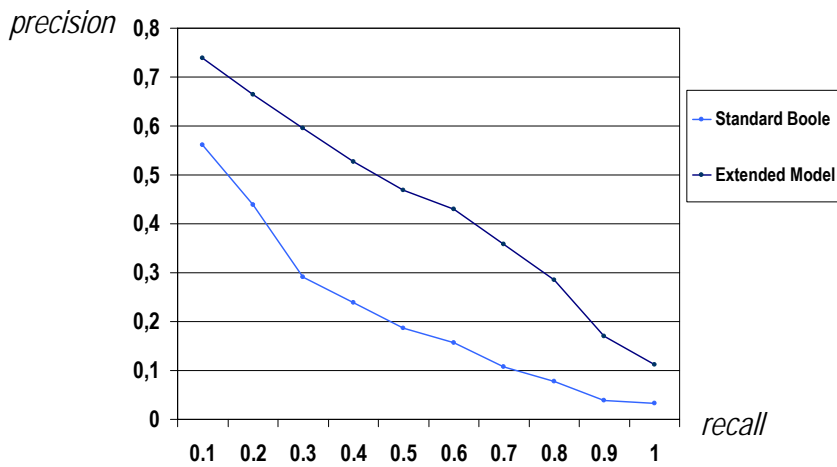
recall / precision graph

Ranking for query q:

	<i>recall</i>	<i>precision</i>	
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. doc 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 amount of retrieved documents
 - Increasing recall -> decreasing precision
- Importance depends on context
 - Expert systems, file search: Recall optimized
 - Web: Precision-optimized

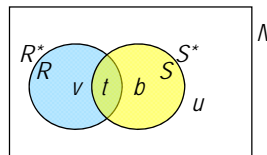


Fallout

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

→ Effektivität $fallout = \frac{b}{R^*}$



(Abfallquote)



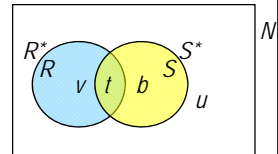
Complementary Measurements

- miss ratio (\leftrightarrow recall) - *Fehlquote* $miss\ ratio = \frac{v}{R}$

- noise ratio (\leftrightarrow precision) - *Ballastquote* $noise\ ratio = \frac{b}{S}$

- rejection ratio (\leftrightarrow fallout) - *Abweisquote* $rejection\ ratio = \frac{u}{R^*}$

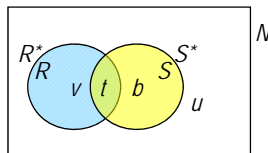
$miss\ ratio = 1 - recall$
 $noise\ ratio = 1 - precision$
 $rejection\ ratio = 1 - fallout$



Generality

- Defines proportion of relevant documents:

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

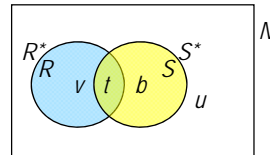




User Centered Measures

- 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:

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



Example Averaging for *recall*

Retrieval results	1	2	3	4	5	Σ
Number of found relevant documents t	10	5	20	1	30	66
Number of existing relevant documents R	100	5	30	10	40	185
<i>recall</i> r	0.1	1	0.66	0.1	0.75	2.616

$$\text{Micro: } r = \frac{10 + 5 + 20 + 1 + 30}{100 + 5 + 30 + 10 + 40} = \frac{66}{185} = 0.357$$

$$\text{Macro: } r = \frac{0.1 + 1 + 0.6 + 0.1 + 0.75}{5} = \frac{2.616}{5} = 0.523$$



Example Averaging for *recall*

Retrieval results	1	2	3	4	5	Σ
Number of found relevant documents t	10	5	20	1	30	15
Number of existing relevant documents R	100	5	30	10	40	105
<i>recall</i> r	0.1	1	0.66	0.1	0.75	1.1

$$\text{Micro: } r = \frac{10 + 5}{100 + 5} = \frac{15}{105} = 0.143 \quad (\text{over all: } 0.357)$$

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



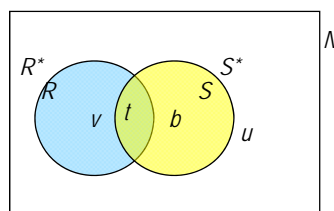
MAP – Mean Average Precision


- 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



Problems of Relevance Assessment I

- System centered vs. user centered relevance
 - Cooper 1971: logical relevance vs. utility
 - Salo&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 documents
 - 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



Evaluation Campaigns

- TREC: Text Retrieval Conference
- CLEF: Cross Language Evaluation Forum → Conference and Labs of the Evaluation Forum
- NTCIR: National Institute of Informatics Test Collection for Information Retrieval → NII Testbeds and Community for Information access Research
- INEX: Initiative for the Evaluation of XML Retrieval
- FIRE: Forum for Information Retrieval Evaluation
- RIRES / ROMIP: Russian Information Retrieval Evaluation Seminar
- MIR / MDL: Music Information Retrieval / Music Digital Library
- MediaEval: Benchmarking Initiative for Multimedia Evaluation



TREC

- since 1992: TREC-1
- National Institute of Standards and Technology (NIST)
- Goals:
 - Standardize and improve evaluation standards
 - Collections for general and specific issues
 - Comparison of system performance
 - Enable research and research exchange
 - Technology transfer

TREC: Tracks



- Contextual Suggestion Task
- Crowdsourcing Track
- Federated Web Search Track
- Knowledge Base Acceleration Track
- Microblog Track
- Session Track
- Temporal Summarization Track
- Web Track
- Ad-hoc
- Filtering Track
- Interactive Track
- Question Answering Track
- Spoken Language Track
- Cross-Language Track
- Blog Track
- Chemical IR Track
- Enterprise Track
- Genomics Track
- Legal Track
- Medical Records Track
- Video Track → TRECVID
- ...

CLEF



- since 2000
- Cross Language Evaluation Forum
- since 2010: Conference and Labs of the Evaluation Forum
- Goals:
 - multilingual and multimodal system testing, tuning and evaluation;
 - investigation of the use of unstructured, semi-structured, highly-structured, and semantically enriched data in information access;
 - creation of reusable test collections for benchmarking;
 - exploration of new evaluation methodologies and innovative ways of using experimental data;
 - discussion of results, comparison of approaches, exchange of ideas, and transfer of knowledge.



CLEF

- **Ad-Hoc:** Mono-, Bi- and Multilingual Document Retrieval on News Collections
- **Domain-Specific:** Mono- and Cross-Language Information Retrieval on Structured Scientific Data
- **iCLEF:** Interactive Cross-Language Information Retrieval
- **MLQA:** Multiple Language Question Answering
- **CL-SR:** Cross-Language Speech Retrieval
- **WebCLEF:** Multilingual Web Track
- **GeoCLEF:** Cross-Language Geographical Retrieval
- **ImageCLEF:** Cross-Language Retrieval in Image Collections



TRECVID

- 2001/2002: TREC Video Track
- since 2003: TRECVID
- Goal: „The main goal of the TREC Video Retrieval Evaluation (TRECVID) is to promote progress in content-based retrieval from digital video via open, metrics-based evaluation.“ (<http://www-nlpir.nist.gov/projects/tv2005/tv2005.html>)



Examples for Tracks

- Terminology:
 - Track
 - Collection
 - Task / Topic



Example: Topic

```
<TOP>
<Head>Tipster Topic Description</HEAD>
<NUM>066</NUM>
<DOM>Science and Technology</DOM>
<DESC>Document will identify a type of natural language
      processing technology which is being developed or marketed
      in the U.S.</DESC>
<NARR>A relevant document will identify a company or
      institution developing or marketing a natural language
      processing technology, identify the technology, and
      identify one or more features of the company's
      product.</NARR>
<CON>NLP, translation, language, dictionary, font,
      software</CON>
<NAT>U.S.</NAT>
</TOP>
```



Video Tasks

TRECVID is a laboratory-style evaluation that attempts to model real world situations or significant component tasks involved in such situations. In 2005 there are four main tasks with tests associated and participants must complete at least one of these in order to attend the workshop.

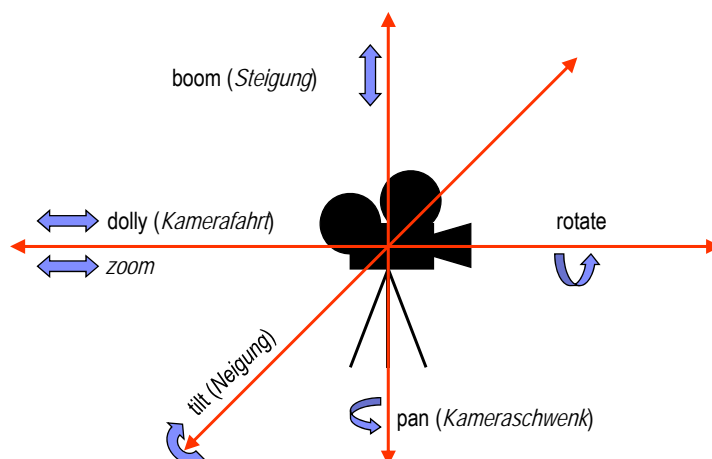
- shot boundary determination
- low-level feature extraction
- high-level feature extraction
- search (interactive, manual, and automatic)

In addition there is a pilot task which is optional. As part of putting the guidelines in final form, the details of this task and its evaluation will be worked out by those who decide to participate in it.

- Explore BBCrashes



Low-Level Feature Extraction Task



High Level Feature Extraction Task

- TRECVID Ontologie LSCOW (Large Scale Ontology Workshop)
- concepts, based on the following criteria:
 - Extractable from a single keyframe (not a full clip)
 - No context needed
 - No ambiguities (*kein Doppeldeutigkeiten*)
 - Not iconographic



Albrecht Dürer, Adam and Eve, 1504. © Sterling and Francine Clark Art Institute, Williamstown, Massachusetts, USA

High Level Feature Extraction Task

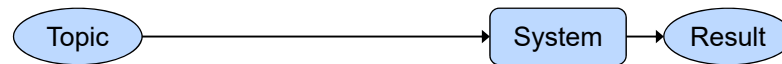


38. **People walking/running:** segment contains video of more than one person walking or running
39. **Explosion or fire:** segment contains video of an explosion or fire
40. **Map:** segment contains video of a map
41. **US flag:** segment contains video of a US flag
42. **Building exterior:** segment contains video of the exterior of a building
43. **Waterscape/waterfront:** segment contains video of a waterscape or waterfront
44. **Mountain:** segment contains video of a mountain or mountain range with slope(s) visible
45. **Prisoner:** segment contains video of a captive person, e.g., imprisoned, behind bars, in jail, in handcuffs, etc.
46. **Sports:** segment contains video of any sport in action
47. **Car:** segment contains video of an automobile

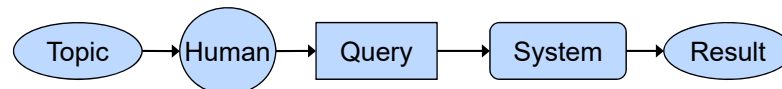
Search Task



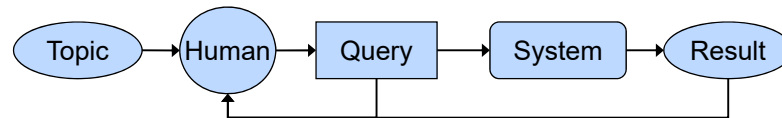
Automatic



Manual



Interactive



BBC Rushes



- Rush – special aspects:
 - Natural soundscape
 - Actors are not allways filmed
 - Repetition
 - Long shots with static camera
 - Stock videos
- Task scenario: search old raw matrial for reuse in new production
- Material: 50h BBC rushes
- Task: build a system, perform an evaluation of the material

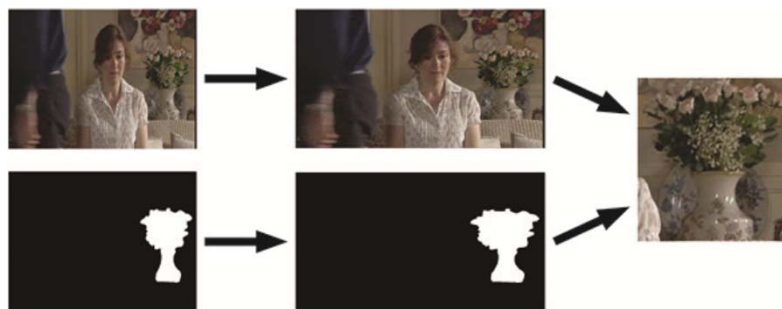


BBC Rushes: Topics Template

- Title
- Brief textual description of the information need (this text may contain references to the examples)
- Examples* of what is wanted:
 - reference to video clip
 - Optional brief textual clarification of the example's relation to the need
 - reference to image
 - Optional brief textual clarification of the example's relation to the need
 - reference to audio
 - Optional brief textual clarification of the example's relation to the need



TrecVid Instance Search



→ 41.760.000 frames

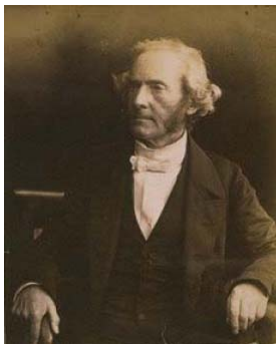


Image Retrieval Task

- St. Andrews collection of historic photographic images
 - Task 1: bilingual ad hoc task
 - Task 2: known-item interactive task:
- CasImage radiological medical database
 - Task: query-by-example



Example Image with Caption



Short title: Rev William Swan.

Long title: Rev William Swan.

Location: Fife, Scotland

Description: Seated, 3/ 4 face studio portrait of a man.

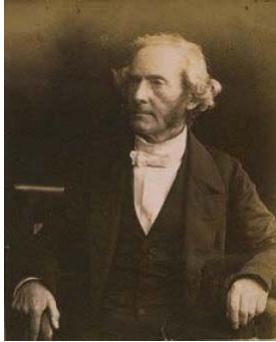
Date: ca.1850

Photographer: Thomas Rodger

Categories: [ministers][identified male][dress - clerical]

Notes: ALB6-85-2 jf/ pcBIOG: Rev William Swan () ADD:
Former owners of album: A Govan then J J? Lowson.
Individuals and other subjects indicative of St Andrews
provenance. By T. R. as identified by Karen A. Johnstone
"Thomas Rodger 1832-1883. A biography and catalogue
of selected works".

Topic



```
<top>
  <num> Number: 1 </num>
  <title> Portrait pictures of church ministers by
    Thomas Rodger </title>
  <narr> Relevant images are portrait photographs of
    ministers or church leaders by the photographer
    Thomas Rodger. Images from any era are relevant,
    but must show one person only taken within a
    studio, i.e. posing for the picture. Pictures of
    groups are not relevant. </narr>
</top>
```

Example Image Medical Case Note



<CASIMAGE_CASE>

<ID>2526</ID>

<Description>

Bassin du 28.02.1985: Status avant et après réduction. Avant réduction, luxation complète du fémur, avec fracture avec fragments du cotyle. Après réduction, interposition de l'un de ces fragments entre la tte fémorale et le toit du cotyle.

</Description>

<Diagnosis>

Luxation postérieure du fémur gauche associée? une fracture multifragmentaire ...

</Diagnosis>

...



BirdCLEF 2017: Usage scenario

The general public as well as professionals like park rangers, ecology consultants and of course ornithologists might be users of an automated bird identifying system, in the context of wider initiatives related to ecological surveillance, biodiversity conservation or taxonomy. Using audio records rather than pictures is justifiable since bird calls and songs have proven to be easier to collect and to discriminate better between species.

The 2017 bird identification task will share similar objectives and scenarios with the previous editions:

- the identification of a particular bird specimen in a recording of it,
- and the recognition of all specimens singing in raw soundscapes that can contain up to several tens of birds singing simultaneously.



BirdCLEF 2017: Data Collection

H. Glotin recorded in summer 2016 birds soundscapes [...]. These recordings have been realized into the jungle canopy at 35 meters high (the highest point of the area), and at the level of the Amazon river, in Peruvian basin [...]. Several hours were recorded few hours during mornings and evenings at the maximum of the bird acoustic activities. The recordings are sampled at 96 kHz, 24 bits PCM, stereo, dual -12 dB, [...]. These valuable data have then been annotated on the field [...] by international experts. A total of 50 species are [...] labeled in time and frequency, and by sex.

The training data, on the other side, is still built from the outstanding Xeno-canto collaborative database, which is the largest one in the world with hundreds of thousands of bird recordings associated to various metadata such as the type of sound (call, song, alarm, flight, etc.), the date, the locality, textual comments of the authors, multilingual common names and collaborative quality ratings. [...] Basically, the new training dataset will increase the total number of species of 1500 over 36,496 audio recordings in the union of Brazil, Colombia, Venezuela, Guyana, Suriname, French Guiana, Bolivia, Ecuador and Peru.



CLEF eHealth

eHealth documents are much easier to understand after expanding shorthand, correcting the misspellings, normalising all health conditions to standardised terminology, and linking the words to a patient-centric search on the Internet. This would result in “*Description of the patient's active problem: 72 year old female with dependence on [hemodialysis](#), [coronary heart disease](#), [hypertensive disease](#), and [asthma](#) who is currently presenting with the problem of significant [hyperkalemia](#) and associated [arrhythmias](#)*” with the highlighted words linked to their definitions in Consumer Health Vocabulary and other patient-friendly sources. Further, providing the required eHealth information in response to our target user groups information needs in a timely manner, where this information is reliable, accurate and available in a multilingual setting is crucial. In addition, auto converting a verbal nursing handover to text and then highlighting important information within the transcription for the next nurse would aid care documentation and release nurses time to, for example, discuss these resources and provide further information for a longer time with the patients.



CLEF

- http://clef2018.clef-initiative.eu/index.php?page=Pages/labs_info.html#centreclef



CLEF eHealth 2016

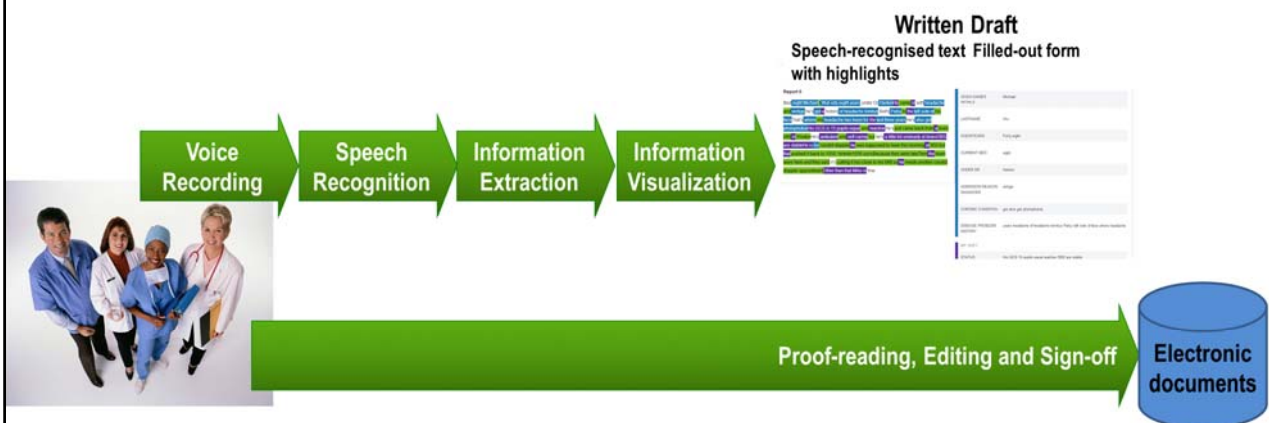
Task 1: Handover information extraction
(**New** challenge)

Task 2: Multilingual Information extraction
(**New:** clinical text dataset, causes of death extraction from French death reports)

Task 3: Patient-centred information retrieval
(**New:** web crawl, queries, evaluation criteria, modelling a whole-of-session (interactive search) scenario).


<https://sites.google.com/site/clefehealth2016/lab-overview>

CLEF eHealth: Handover information extraction



ID: 14, TYPE: cardiovascular

PROFILE:



Name: Ken Harris
Age: 71 years
Admission story: Ken is suffering from arrhythmia for the first time in his life. He is feeling pretty sick but this does not seem to be too serious.
In-patient time: He has been at the ward for three days.
Familiarity: Both you and the next nurse have looked after him earlier.

SPOKEN, FREE-FORM TEXT DOCUMENT:
WAV file (93 words, 48 seconds, 4.25 MB)
The first author typed a transcription only for this spoken document
On a bed three is Ken Harris, 71 years old under Dr Gregor. He came in with arrhythmia. He complained of chest pain this morning and ECG was done and was reviewed by the team. He was given some anginine and morphine for the pain and he is still tachycardic and new meds have been ordered in the medchart. Still for pulse checks for one full minute. Still awaiting echo this afternoon. His blood pressure is just normal though he is scoring MEWS of three for the tachycardia. Otherwise he still for monitoring.

WRITTEN, FREE-FORM TEXT DOCUMENT:
Ken harris, bed three, 71 yrs old under Dr Gregor, came in with arrhythmia. He complained of chest pain this am and ECG was done and was reviewed by the team. He was given some anginine and morphine for the pain. Still tachycardic and new meds have been ordered in the medchart. still for pulse checks for one full minute. Still awaiting echo this afternoon. His BP is just normal though he is scoring MEWS of 3 for the tachycardia. He is still for monitoring.

WRITTEN, STRUCTURED DOCUMENT:

PATIENT INTRODUCTION:

- GivenNames/Initials: Ken
- LastName: harris
- AgeInYears: 71 yrs old
- Gender: He
- CurrentBed: bed three
- UnderDr: 6.1. LastName: Dr Gregor
- AdmissionReason/Diagnosis: arrhythmia

MY SHIFT:

- Status: chest pain
- OtherObservation: tachycardic; BP is just normal; scoring MEWS of 3 for the tachycardia

APPOINTMENTS:

- Status: was done; was reviewed by the team
- Description: ECG; echo
- Time: this afternoon

MEDICATION:

- Medicine: anginine; morphine for the pain; new meds

FUTURE CARE:

- Goal/TaskToBeCompleted/ExpectedOutcome: for pulse checks for one full minute; still for monitoring

MEWS of three for the tachycardia. Otherwise he still for monitoring.

WRITTEN, FREE-FORM TEXT DOCUMENT:
Ken harris, bed three, 71 yrs old under Dr Gregor, came in with arrhythmia. He complained of chest pain this am and ECG was done and was reviewed by the team. He was given some anginine and morphine for the pain. Still tachycardic and new meds have been ordered in the medchart. still for pulse checks for one full minute. Still awaiting echo this afternoon. His BP is just normal though he is scoring MEWS of 3 for the tachycardia. He is still for monitoring.

WRITTEN, STRUCTURED DOCUMENT:

PATIENT INTRODUCTION:

- GivenNames/Initials: Ken
- LastName: harris
- AgeInYears: 71 yrs old
- Gender: He
- CurrentBed: bed three
- UnderDr: 6.1. LastName: Dr Gregor
- AdmissionReason/Diagnosis: arrhythmia

MY SHIFT:

- Status: chest pain
- OtherObservation: tachycardic; BP is just normal; scoring MEWS of 3 for the tachycardia

APPOINTMENTS:

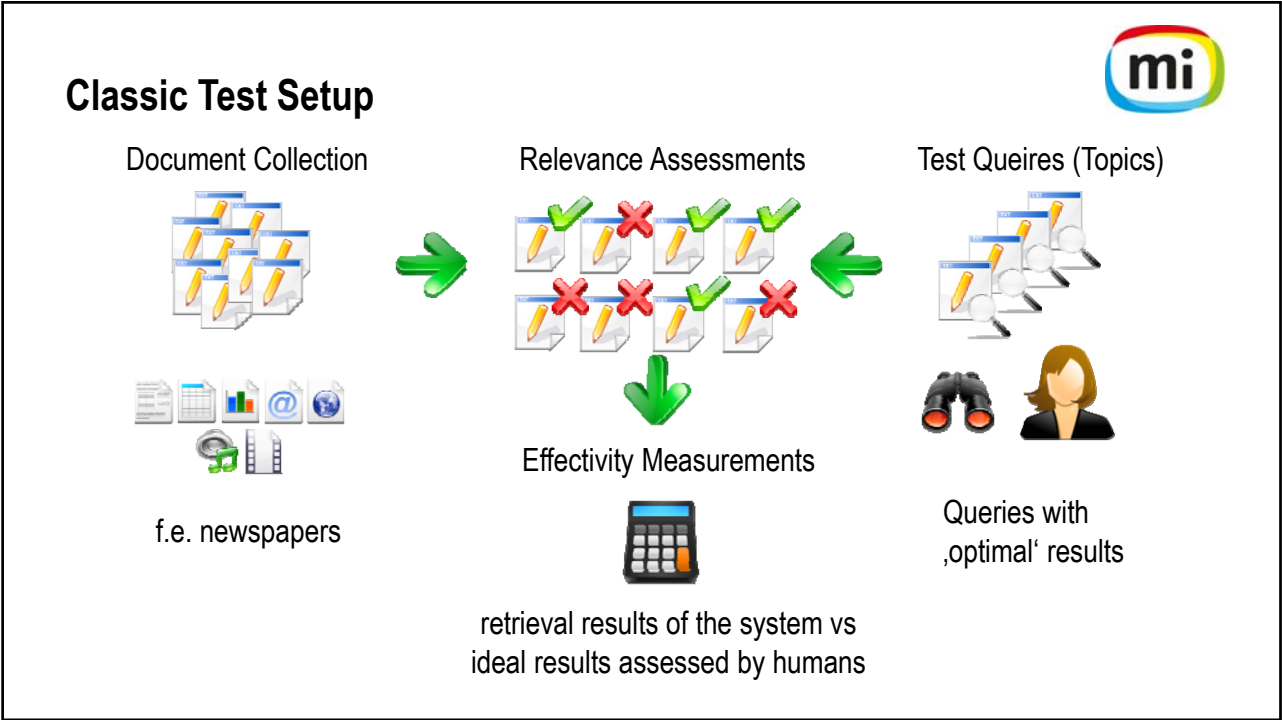
- Status: was done; was reviewed by the team
- Description: ECG; echo
- Time: this afternoon

MEDICATION:

- Medicine: anginine; morphine for the pain; new meds

FUTURE CARE:

- Goal/TaskToBeCompleted/ExpectedOutcome: for pulse checks for one full minute; still for monitoring

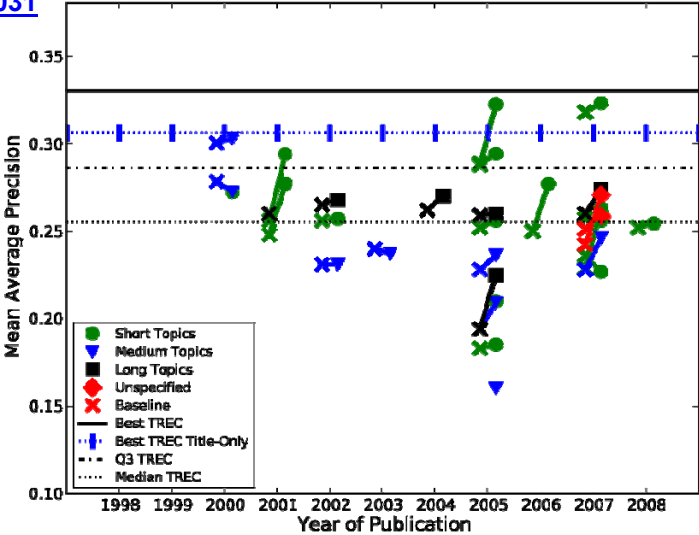


Evaluation Critique

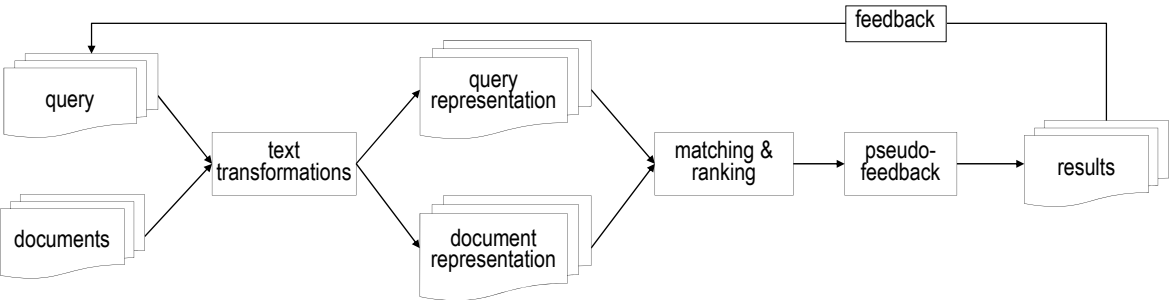
Armstrong et. al. (2008): Improvements that don't add up: ad-hoc retrieval results since 1998
<https://dl.acm.org/citation.cfm?id=1646031>

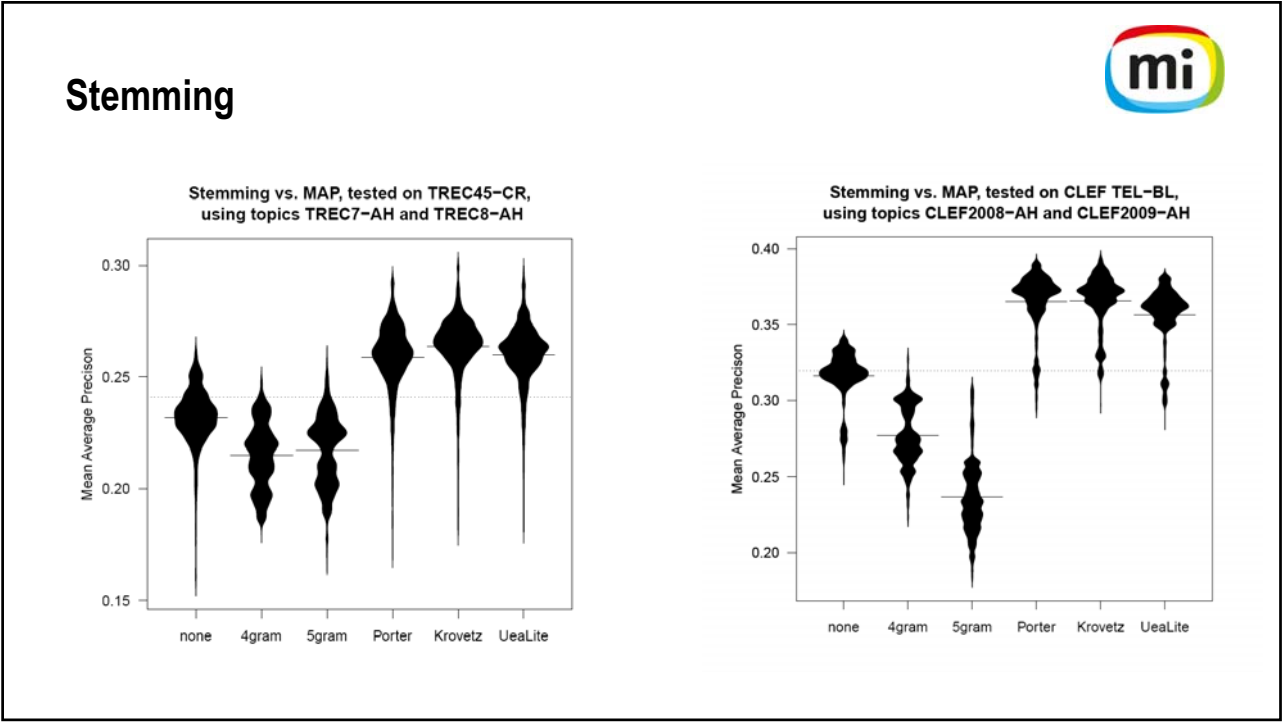
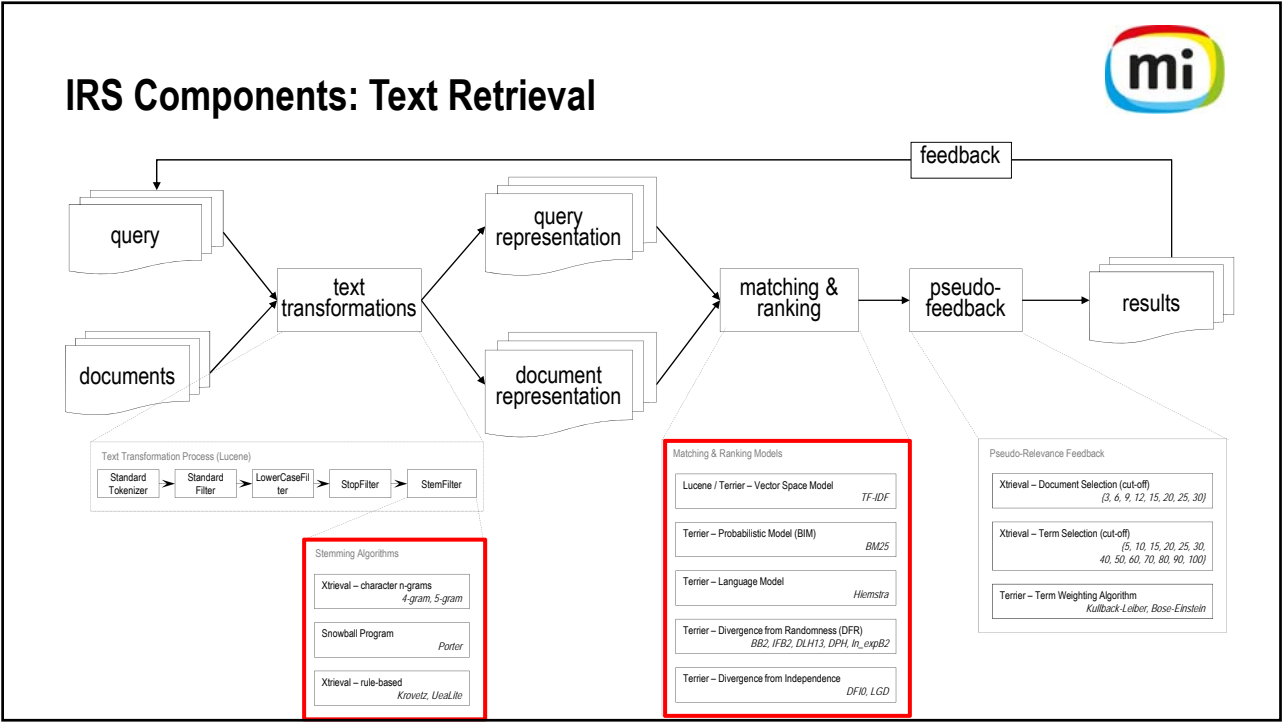
- 1 TREC collection
- 32 publications
- 10 years

→no improvement!

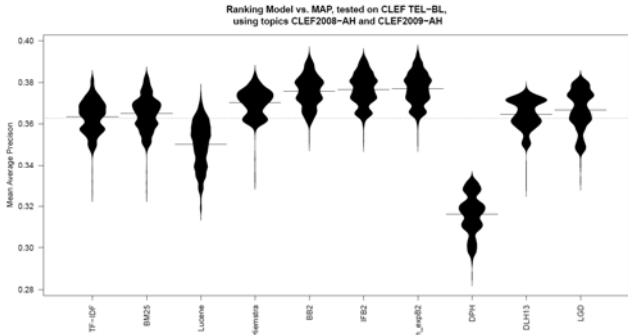
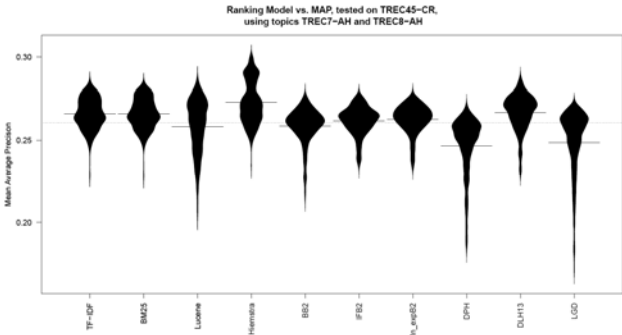


IRS Components: Text Retrieval





Ranking Models



Pseudo Relevance Feedback

