Evaluation: Metrics and Test collections

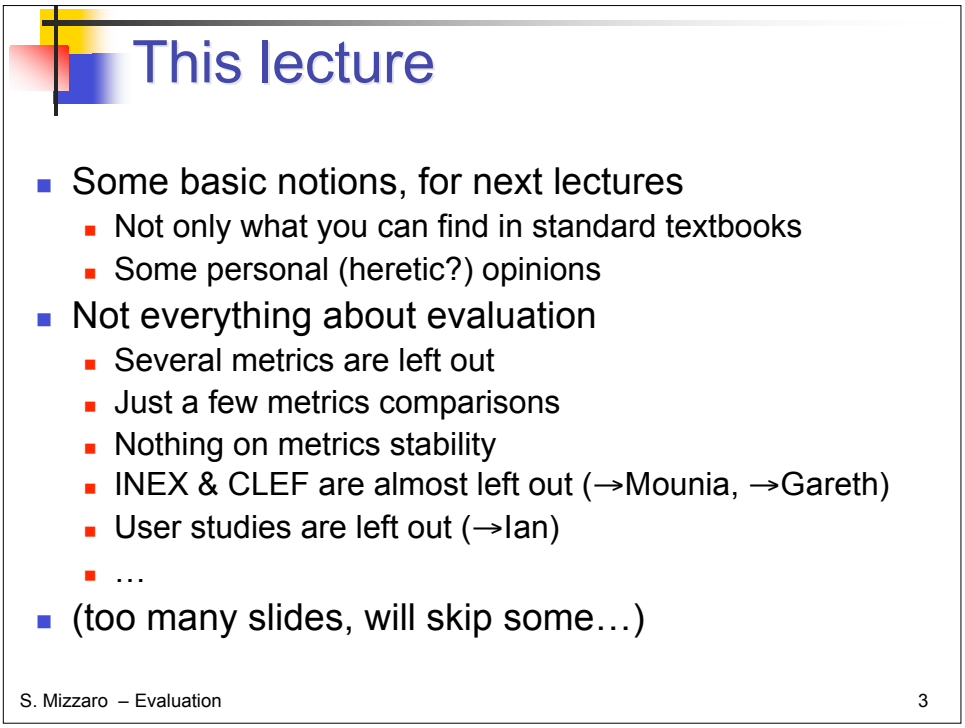
Stefano Mizzaro

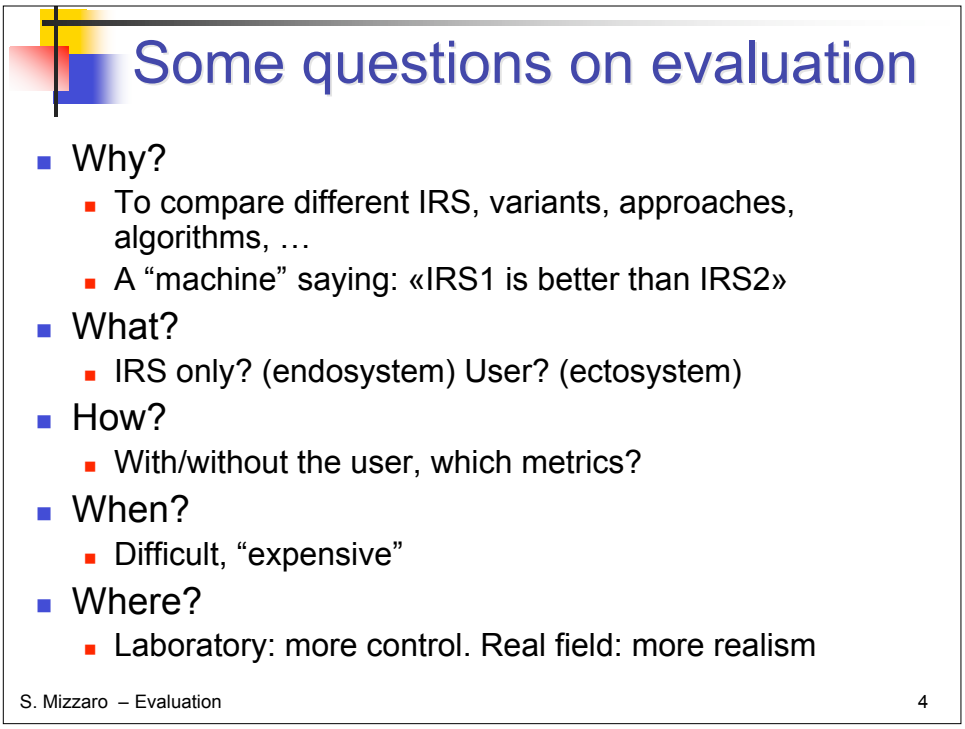
Dept. of Mathematics and Computer Science
University of Udine
http://www.dimi.uniud.it/mizzaro
mizzaro@dimi.uniud.it

ESSIR 2005 – Dublin, 6 September 2005

Outline

- Introduction
  - On evaluation (& relevance)
- Metrics (a.k.a. measures)
  - Common metrics
  - Some less common metrics
  - Classification attempt
- Test collections and Evaluation initiatives
  - Test collections concepts
  - TREC (what it is, terminology, participation, …)
  - Besides TREC (NTCIR, CLEF, INEX)


This lecture
Some basic notions, for next lectures
Not only what you can find in standard textbooks
Some personal (heretic?) opinions
Not everything about evaluation
Several metrics are left out
Just a few metrics comparisons
Nothing on metrics stability
INEX & CLEF are almost left out (→Mounia, →Gareth)
User studies are left out (→Ian)
…
(too many slides, will skip some…)
S. Mizzaro – Evaluation 3


Some questions on evaluation
Why?
To compare different IRS, variants, approaches, algorithms, …
A “machine” saying: «IRS1 is better than IRS2»
What?
IRS only? (endosystem) User? (ectosystem)
How?
With/without the user, which metrics?
When?
Difficult, “expensive”
Where?
Laboratory: more control. Real field: more realism
S. Mizzaro – Evaluation 4

Relevance

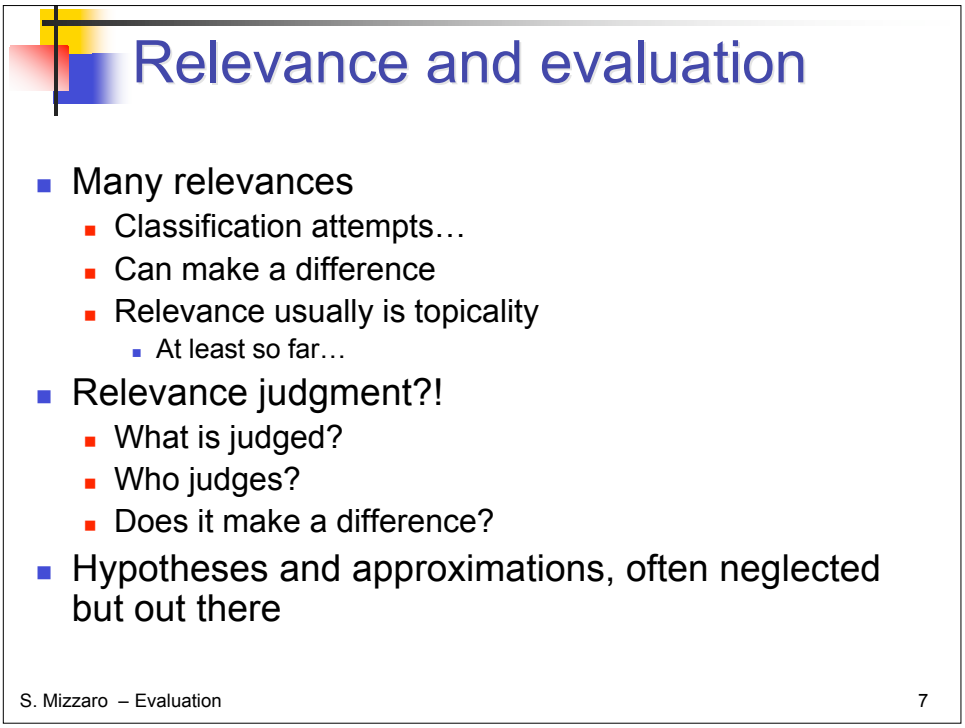
- What to evaluate? The capability of an IRS to retrieve relevant documents
- Relevance?
  - Topicality?
  - “System relevance” vs. “User relevance”?
  - User satisfaction?
  - Utility?

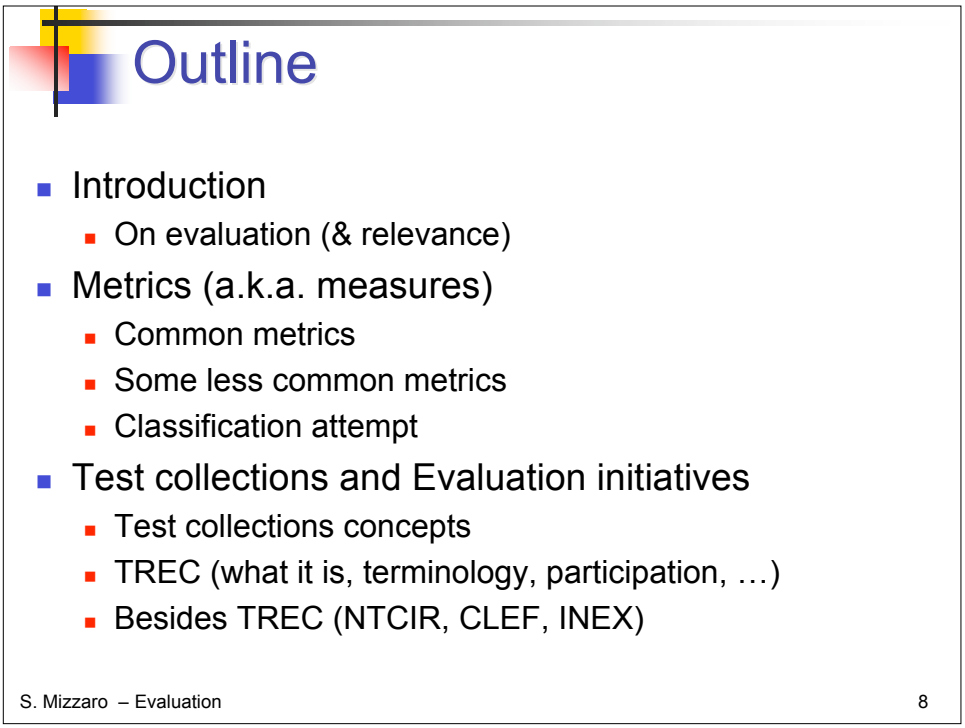
S. Mizzaro – Evaluation 5

The relevance pot

- Relevance
- Situational relevance
- Topicality
- Pertinence
- System relevance
- Utility
- User relevance
- Usefulness
- User satisfaction
- …

S. Mizzaro – Evaluation 6


Relevance and evaluation
Many relevances
Classification attempts…
Can make a difference
Relevance usually is topicality
At least so far…
Relevance judgment?!
What is judged?
Who judges?
Does it make a difference?
Hypotheses and approximations, often neglected but out there
S. Mizzaro – Evaluation
7


Outline
Introduction
On evaluation (& relevance)
Metrics (a.k.a. measures)
Common metrics
Some less common metrics
Classification attempt
Test collections and Evaluation initiatives
Test collections concepts
TREC (what it is, terminology, participation, …)
Besides TREC (NTCIR, CLEF, INEX)
S. Mizzaro – Evaluation
8

Naïve metrics

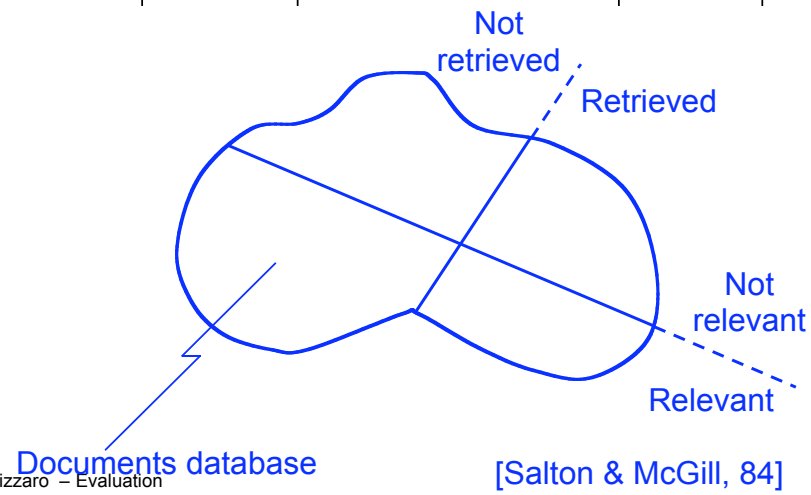
- Just count the number of relevant doc. among the retrieved ones?
  - … An IRS might retrieve the whole collection…
- Just count the number of non relevant among the retrieved?
  - … An IRS might retrieve no docs.…
- Both are needed

A non-IR example…

- Box with 10 white balls and 1000 black balls
- Task: to find the white balls
  - John: 9 white, but also 9 black
  - Mary: 5 white, but also 2 black
- Who's the best?
- Well, it depends.
- Are we more interested in:
  - Find all the white balls? John, 9/10 (5/10) ← Recall (R)
  - Find only white balls? Mary, 5/7 (9/18) ← Precision (P)
- In IR too

Precision & Recall

$$P = \frac{|\text{relevant \& retrieved}|}{|\text{retrieved}|} \qquad R = \frac{|\text{relevant \& retrieved}|}{|\text{relevant}|}$$



[Salton & McGill, 84]

Precision & Recall

- (same def., just different wording)

	Retrieved	Not Retrieved	
Relevant	a	b	$n_1=a+b$
Non relevant	c	d	
	$n_2=a+c$		$N=a+b+c+d$

$$P = \frac{a}{n_2} \qquad R = \frac{a}{n_1}$$

Beyond P&R

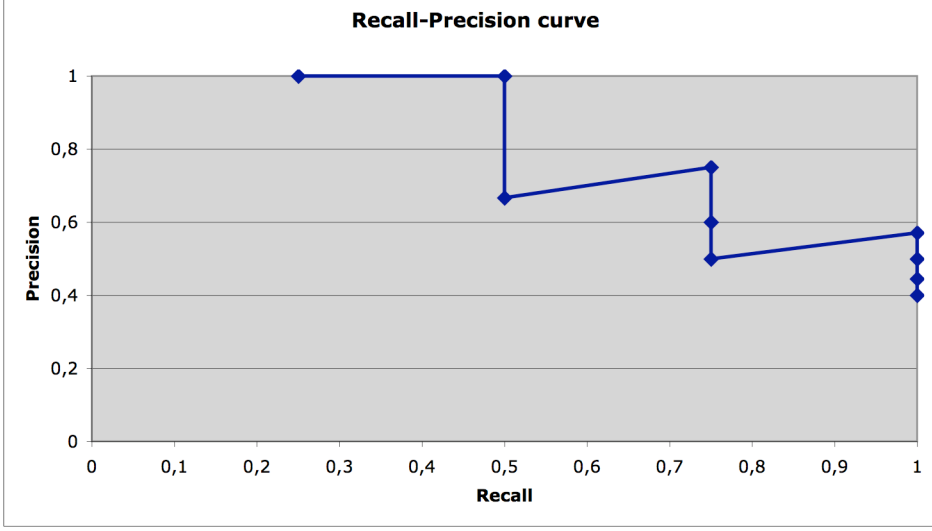
- Binary relevance
  - A document is either relevant or not relevant
- Binary retrieval
  - A document is either retrieved or not retrieved
- IRSs rank the retrieved documents
- Binary relevance, ranked retrieval
  - Classical assumptions in IR evaluation (more later…)

An example

- 1 = relevant
- 0 = not relevant
- 4 relevant docs in the collection

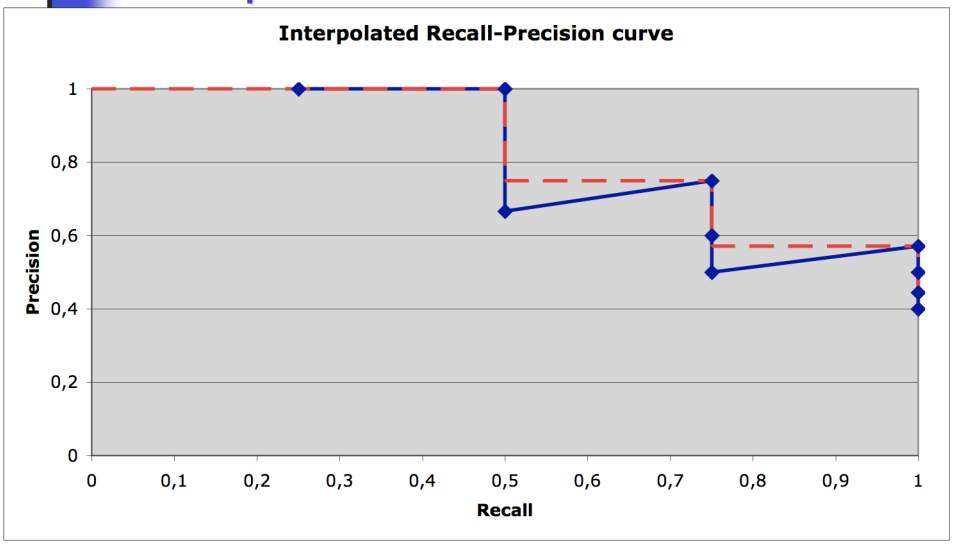
Rank	Rel?	R	P
1	1	0,25	1
2	1	0,5	1
3	0	0,5	0,67
4	1	0,75	0,75
5	0	0,75	0,6
6	0	0,75	0,5
7	1	1	0,57
8	0	1	0,5
9	0	1	0,44
10	0	1	0,4

Recall-Precision curve (graph)



S. Mizzaro – Evaluation 15

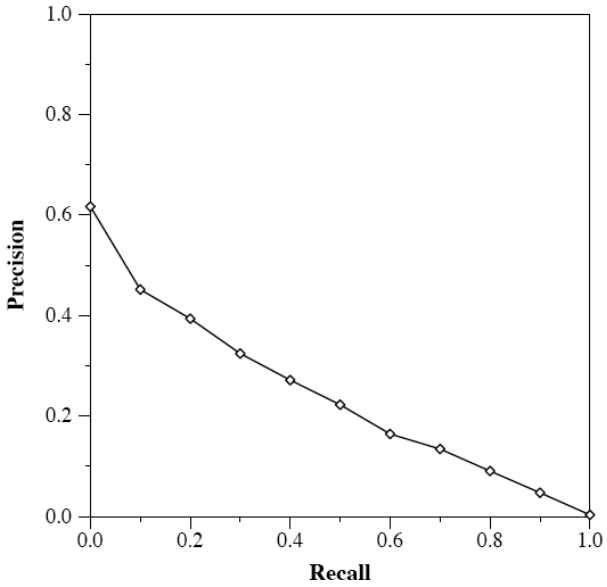
Interpolated R-P curve



S. Mizzaro – Evaluation 16

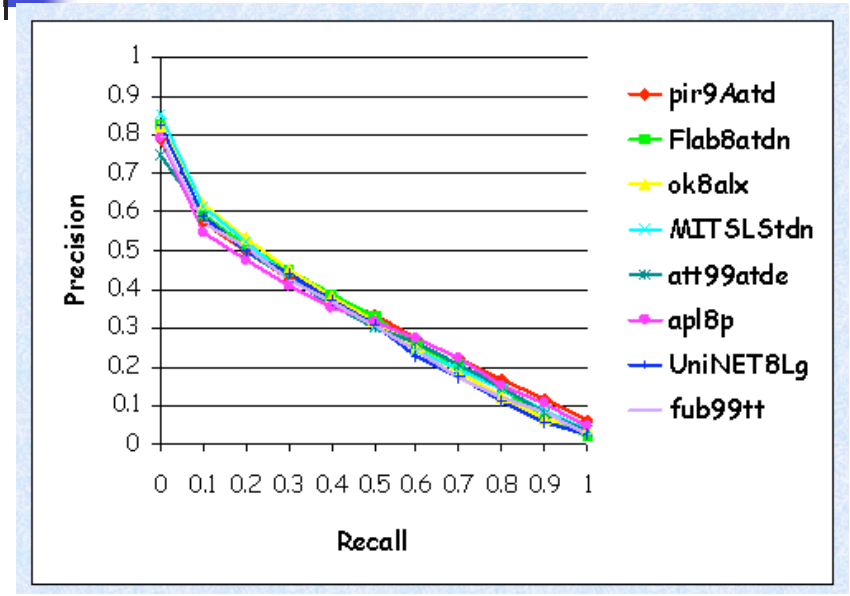
Average over several queries

- # of steps depends on # o relevant docs.
- 11 recall levels: 0, 0.1, 0.2, …, 1
- Saw-tooth → Step → Smooth



S. Mizzaro – Evaluation 17

Comparison of several curves



S. Mizzaro – Evaluation 18

R-P curve: summary

- Binary relevance, ranked retrieval
- "Golden standard"
- Often, recall can't be computed exactly
- It is not a single number
  - Comparison sometimes difficult
  - → Some single valued measures

Average precision

- Average of the precision values obtained after each relevant document is retrieved
  - If not retrieved, precision = 0
  - NOT average of P at the 11 standard recall levels!

Rank	Rel?	R	P
1	**1**	0,25	**1**
2	**1**	0,5	**1**
3	0	0,5	0,67
4	**1**	0,75	**0,75**
5	0	0,75	0,6
6	0	0,75	0,5
7	**1**	1	**0,57**
8	0	1	0,5
9	0	1	0,44
10	0	1	0,4

Mean Average Precision

- Terminology
  - **Mean Average** Precision (MAP)
    - Average Precision (AP) is for one query
    - MAP is the mean across queries of the APs
    - Often/Usually referred to as Average Precision, or Uninterpolated MAP
  - Something ≠ is Interpolated MAP:
    - Average of the average of precisions at standard recall levels (0, 0.1, 0.2, …, 1.0)
- The area below the R-P curve

Other single-valued metrics

- P@1, P@5, P@10, …, P@N
  - Precision value after N retrieved documents
  - P@10 often used for Web search
  - P@1 useful for “I’m Feeling Lucky” searches
- R-precision
  - “P@R”
  - Precision after R documents (R: # of relevant)

Some more metrics

- Beyond:
  - Binary relevance, binary retrieval
  - Binary relevance, ranked retrieval
- ESL, Expected Search Length
- DCG, Discounted Cumulative Gain
- ADM, Average Distance Measure
- …

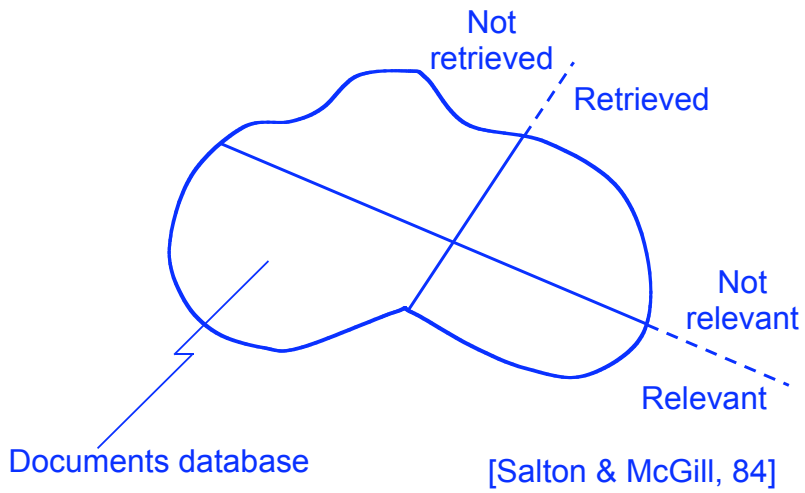
ESL, Expected Search Length

- ESL(x) = # of documents to be read (following the rank) to have x relevant documents
- Averaged over several queries
- Not a single value, a function of x
  - Average of ESL(x)/x to have a single value representing the average # of read docs. per relevant docs.
- Ok for partial ranking too

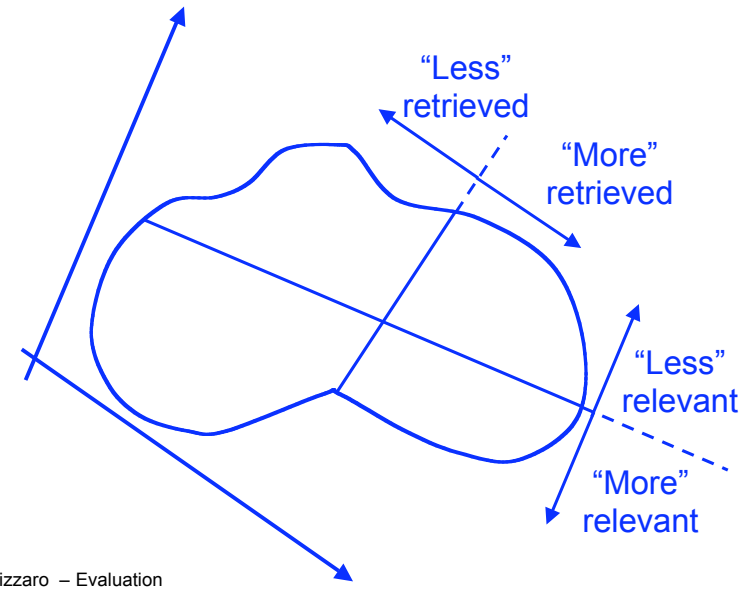
DCG, Discounted Cumulative Gain

- Category relevance, ranked retrieval
  - N relevance level: 0, 1, 2, … N-1
  - The earliest a highly relevant doc is ranked, the better
    - The highest gain the user gets
- DCG measures the gain that a doc gives to the user
- “discounting” (decreasing) with log(rank)

ADM: From binary relevance…

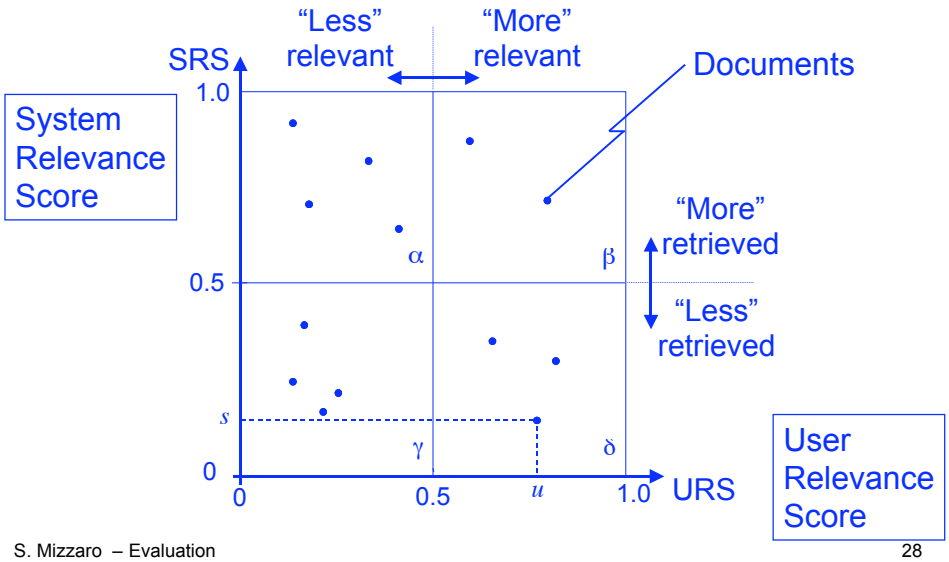


… to continuous relevance



S. Mizzaro – Evaluation 27

The URS/SRS plane



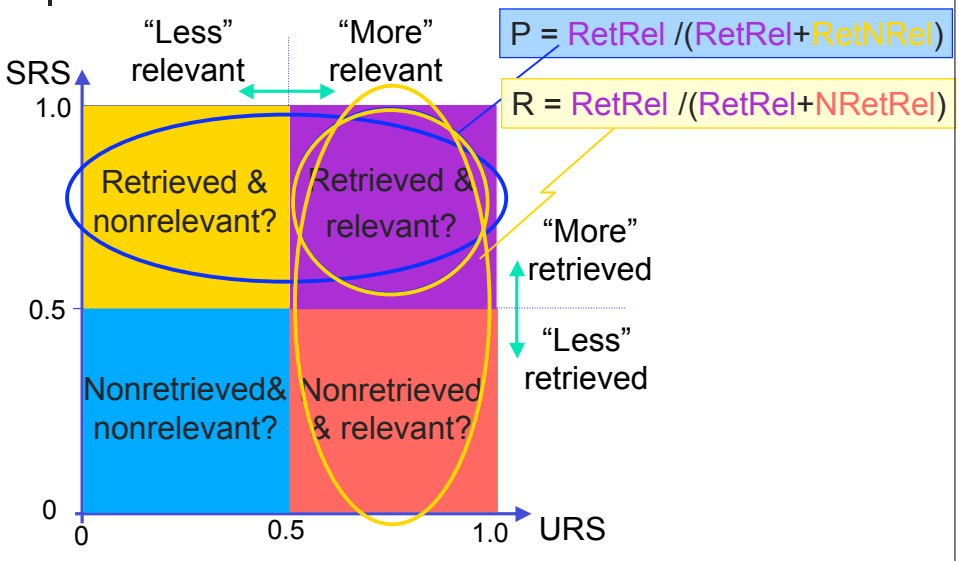
S. Mizzaro – Evaluation 28

SRS and URS

- SRS (System Relevance Score)
  - Relevance value given by the IRS
- URS (User Relevance Score)
  - Relevance value given by the user
- Real numbers, in the [0..1] range
- Different from
  - RSV (Retrieval Status Value), insensible to rank-preserving transformations
  - Estimate of the probability of relevance

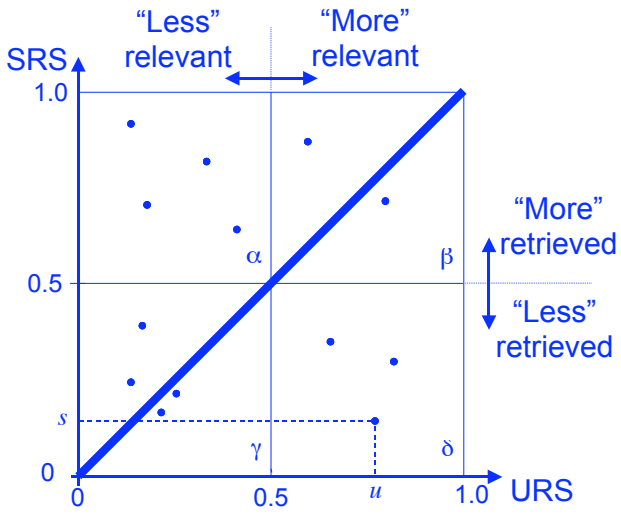
S. Mizzaro – Evaluation 29

A step backward: P & R



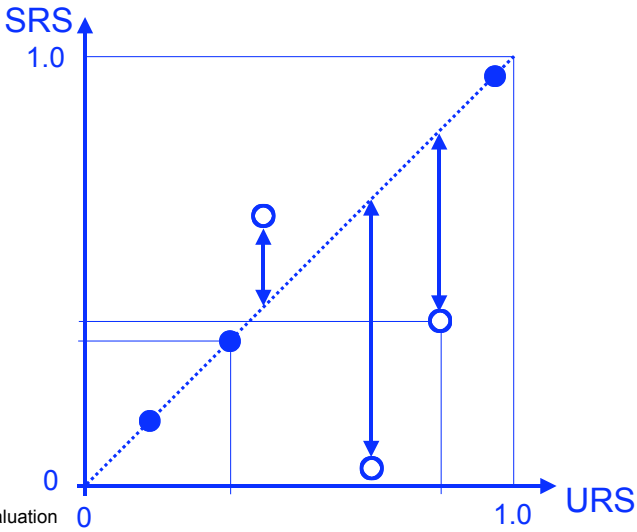
S. Mizzaro – Evaluation 30

The "right" places…



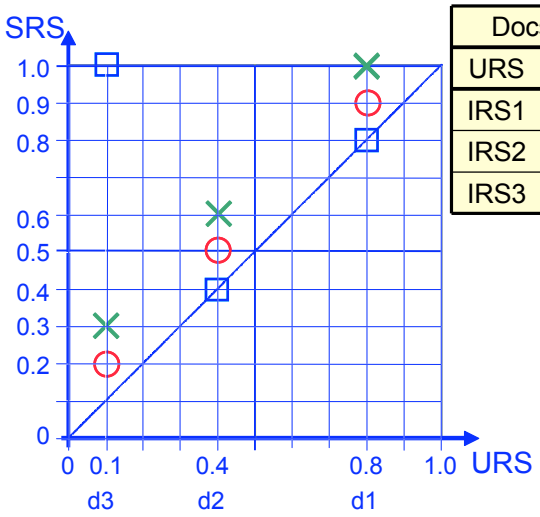
S. Mizzaro – Evaluation 31

ADM: Average Distance Measure



S. Mizzaro – Evaluation 32

An example

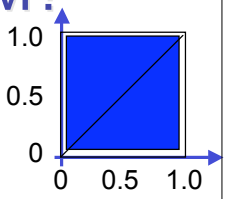


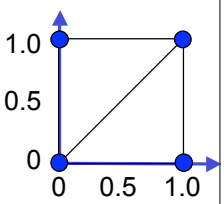
Docs.	d1	d2	d3	ADM
URS	0.8	0.4	0.1	
IRS1 ○	0.9	0.5	0.2	0.9
IRS2 ×	1.0	0.6	0.3	0.8
IRS3 □	0.8	0.4	1.0	0.7

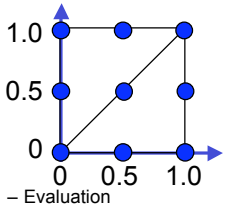
S. Mizzaro - ADM 33

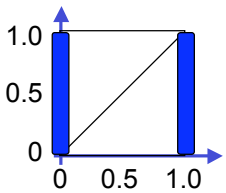
What is needed for ADM?

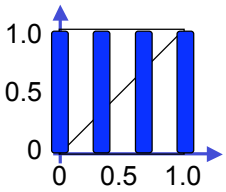
- Ideal situation: Continuous SRS & URS
- Worst situation: “binarized” ADM
  - All the documents in (0,0),(0,1),(1,0),(1,1)
  - Docs in (0,1) and (1,1) only: R
  - Docs in (1,0) and (1,1) only: P
- Intermediate situations: “discrete” ADM
  - Categories, combinations, …









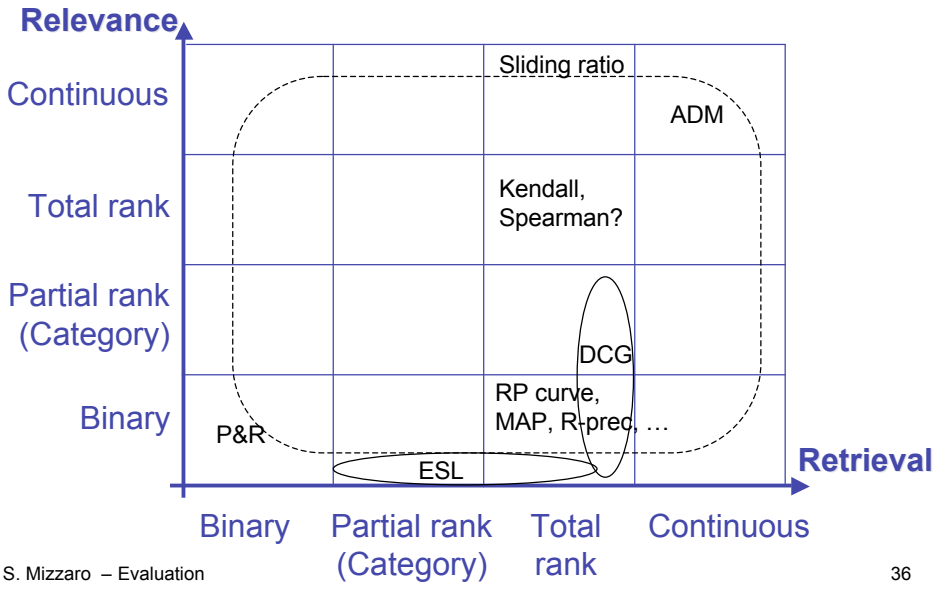


S. Mizzaro – Evaluation 34

Metrics: summary

- Binary relevance, binary retrieval
  - P & R
- Binary relevance, ranked retrieval
  - R-P curve, MAP, P@N, R-prec (← standard)
- Binary relevance, partial ranked retrieval
  - ESL
- Category relevance, ranked retrieval
  - DCG
- Continuous relevance, continuous retrieval
  - ADM

Classification (incomplete!)

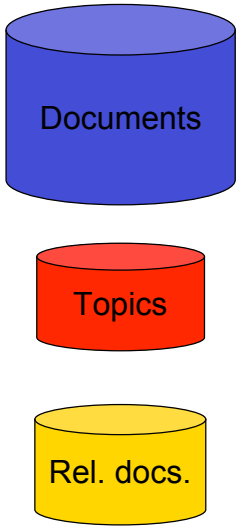


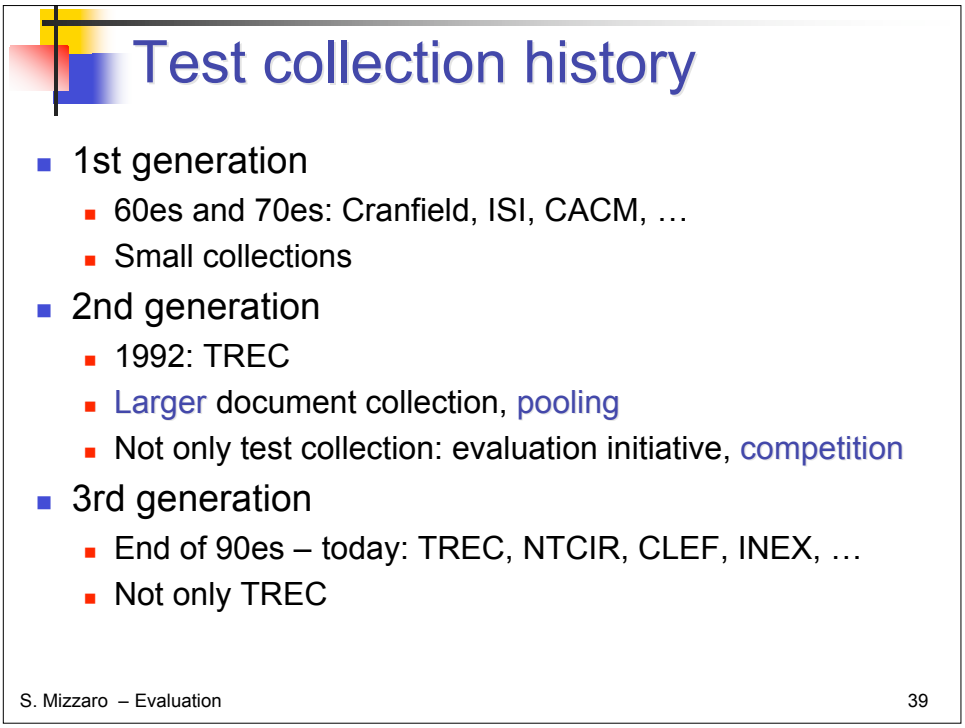
Outline

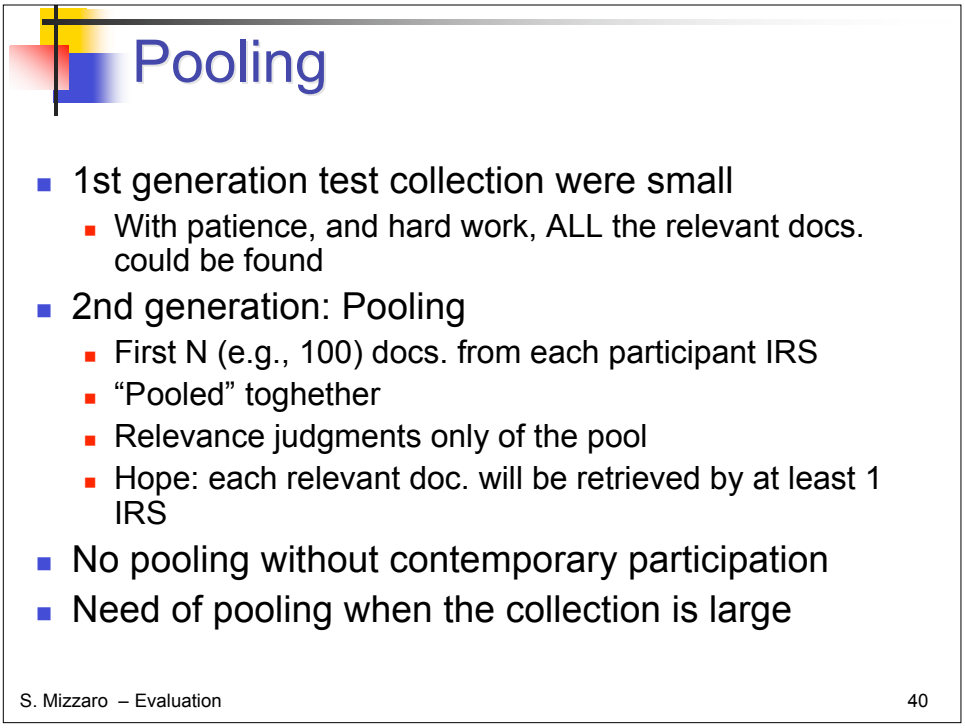
- Introduction
  - On evaluation (& relevance)
- Metrics (a.k.a. measures)
  - Common metrics
  - Some less common metrics
  - Classification attempt
- Test collections and Evaluation initiatives
  - Test collections concepts
  - TREC (what it is, terminology, participation, …)
  - Besides TREC (NTCIR, CLEF, INEX)

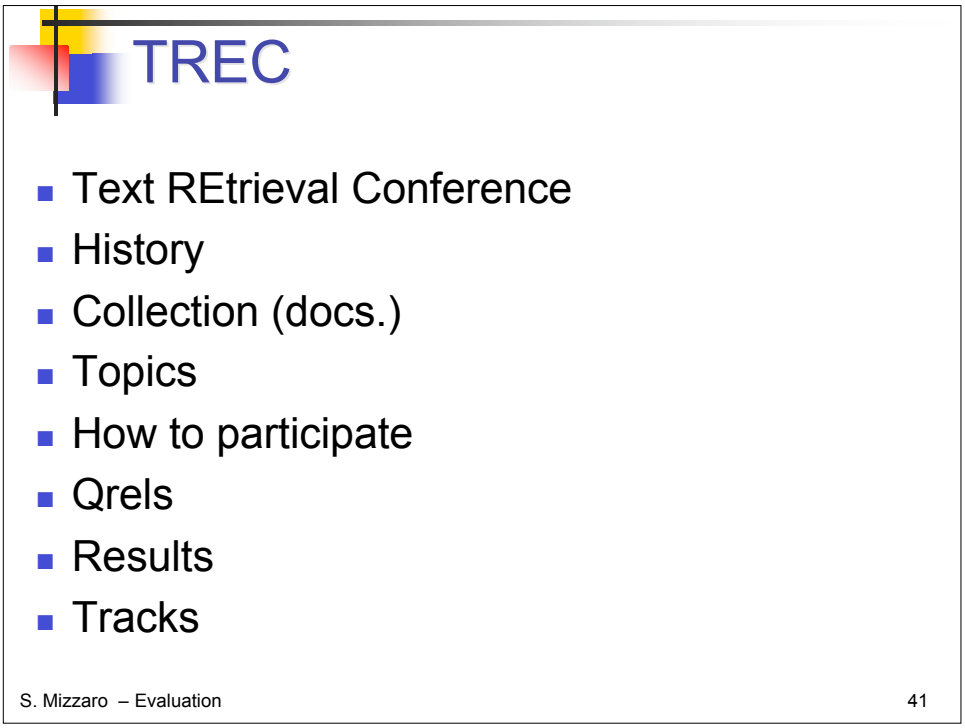
Test collection approach

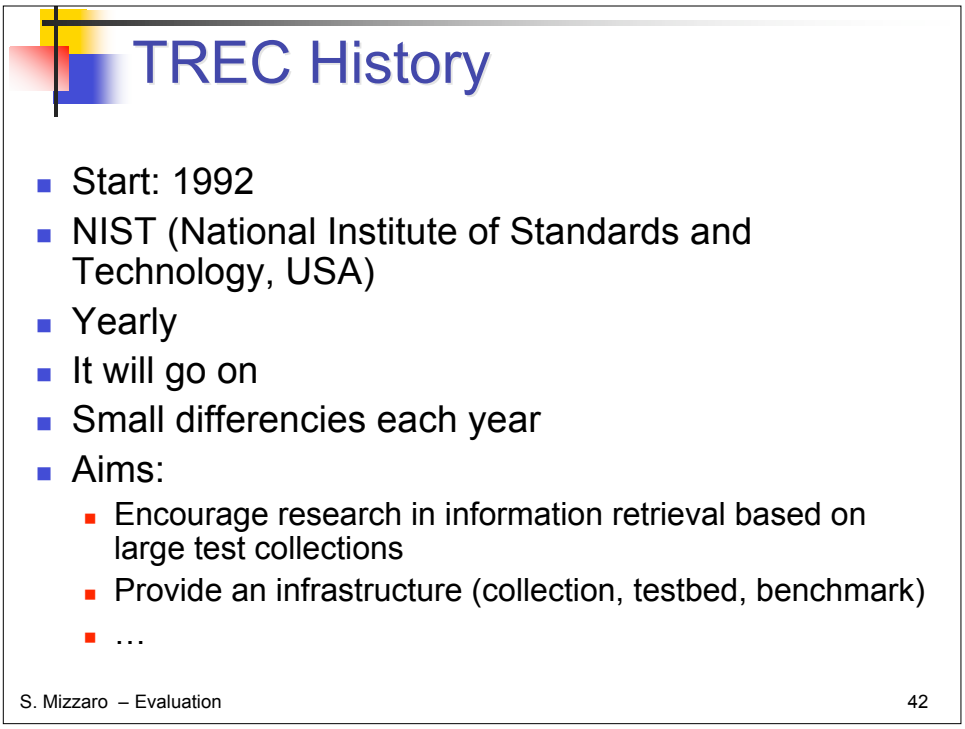
- Test collection =
  - Set of documents
  - Set of requests (“Topics”)
  - Set of relevance judgments for each request (“qrels.”)
    - Binary, categories, …



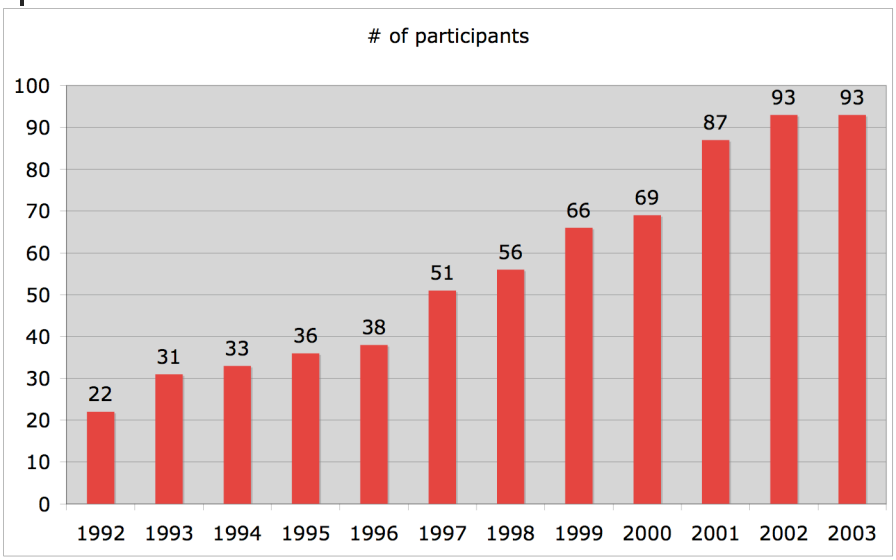

Test collection history
1st generation
60es and 70es: Cranfield, ISI, CACM, …
Small collections
2nd generation
1992: TREC
Larger document collection, pooling
Not only test collection: evaluation initiative, competition
3rd generation
End of 90es – today: TREC, NTCIR, CLEF, INEX, …
Not only TREC
S. Mizzaro – Evaluation
39


Pooling
1st generation test collection were small
With patience, and hard work, ALL the relevant docs. could be found
2nd generation: Pooling
First N (e.g., 100) docs. from each participant IRS
“Pooled” toghether
Relevance judgments only of the pool
Hope: each relevant doc. will be retrieved by at least 1 IRS
No pooling without contemporary participation
Need of pooling when the collection is large
S. Mizzaro – Evaluation
40


TREC
Text REtrieval Conference
History
Collection (docs.)
Topics
How to participate
Qrels
Results
Tracks
S. Mizzaro – Evaluation
41


TREC History
Start: 1992
NIST (National Institute of Standards and Technology, USA)
Yearly
It will go on
Small differencies each year
Aims:
Encourage research in information retrieval based on large test collections
Provide an infrastructure (collection, testbed, benchmark)
…
S. Mizzaro – Evaluation
42

TREC participation



S. Mizzaro – Evaluation 43

Collection (documents)

- Standard, “Ad hoc”
- Incrementally built year after year
- ~2GB, 500K – 1M documents, some hundreds words per document
- Newspaper articles, government docs., abstracts, …
  - Original versions, including errorsSGML formatted
- DOCID (“DOCNO”)

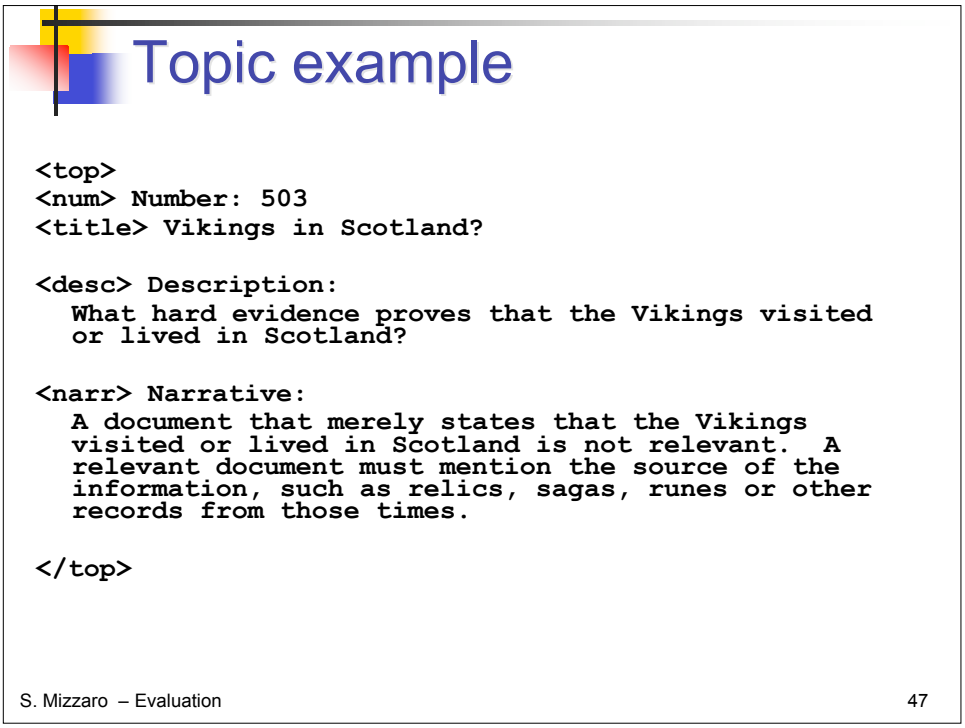
S. Mizzaro – Evaluation 44

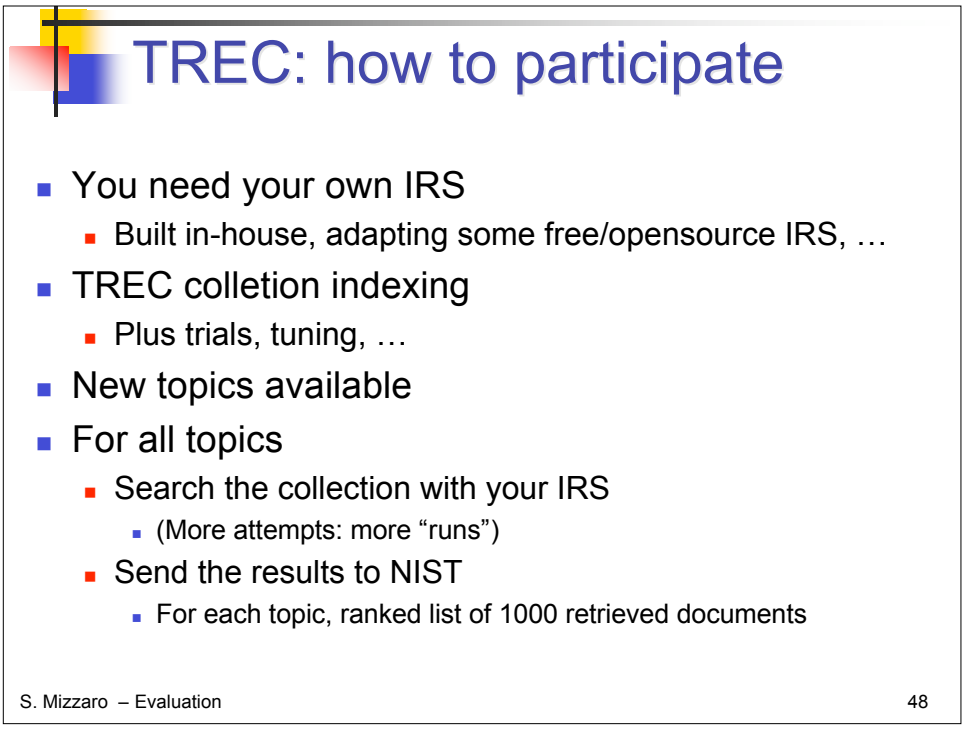
Document example

```
<DOC>
<DOCNO>FT911-3</DOCNO>
<PROFILE>AN-BEOA7AAIFT</PROFILE>
<DATE>910514
</DATE>
<HEADLINE>
FT 14 MAY 91 / International Company News: Contigas
  plans DM900m east German project
</HEADLINE>
<BYLINE>
By DAVID GOODHART
</BYLINE>
<DATELINE>
BONN
</DATELINE>
<TEXT>
CONTIGAS, the German gas group 81 per cent owned by the
  utility Bayernwerk, said yesterday that it intends
  to invest DM900m (Dollars 522m) in the next four
  years to build a new gas distribution system in the
  east German state of Thuringia. [...]
</TEXT>
</DOC>
```


"Topics" (requests)

- Information need representations
- Each year, 50 new topics
- Provide information to understand if a document is relevantg or not
- SGML, 4 fields
  - Numeric id.: 1-50, 51-100, …
  - Title
  - Brief description
  - Narrative description (longer)


Topic example
<top>
<num> Number: 503
<title> Vikings in Scotland?
<desc> Description:
What hard evidence proves that the Vikings visited or lived in Scotland?
<narr> Narrative:
A document that merely states that the Vikings visited or lived in Scotland is not relevant. A relevant document must mention the source of the information, such as relics, sagas, runes or other records from those times.
</top>
S. Mizzaro – Evaluation
47


TREC: how to participate
You need your own IRS
Built in-house, adapting some free/opensource IRS, …
TREC colletion indexing
Plus trials, tuning, …
New topics available
For all topics
Search the collection with your IRS
(More attempts: more “runs”)
Send the results to NIST
For each topic, ranked list of 1000 retrieved documents
S. Mizzaro – Evaluation
48

Results example

TopicID	?	DocID	Rank	Weight	RunID
151	Q0	G02-86-0432155	1	16.113211	VTnhpok1
151	Q0	G27-74-0229731	2	15.796911	VTnhpok1
151	Q0	G43-54-2688995	3	15.638825	VTnhpok1
151	Q0	G08-67-2638557	4	15.360800	VTnhpok1
151	Q0	G43-53-0649940	5	15.321091	VTnhpok1
151	Q0	G43-50-0606214	6	15.294382	VTnhpok1
...					

How the evaluation is done

- Results from all participants are collected (1000 docs by 50 topics by N participants – some with more runs)
- Pooling: first 100 documents from each run
- Human “assessors” judge the relevance of the documents in the pool
  - “qrels” are produced (relevance judgments, usually 0/1)
  - Not judged docs are not relevant
- A software program (“trec-eval”) computes some metrics

"qrels" format

- Columns meaning:
  - Topic id
  - Iteration (usually 0, not used)
  - Doc ID
  - Relevance (0 = not relevant; 1 = relevant)
- Example:

```
1 0 AP880212-0161 0
1 0 AP880216-0139 1
1 0 AP880216-0169 0
1 0 AP880217-0026 0
1 0 AP880217-0030 0
...
```

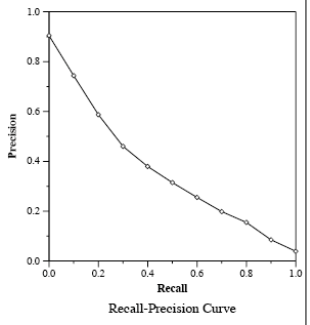

What is computed by trec-eval

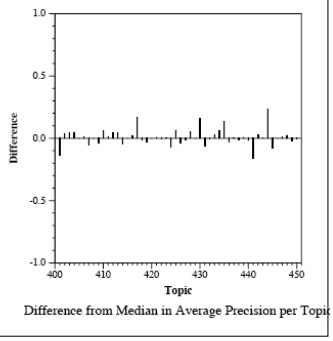
- TREC-8 (1999)
  - Precision at 11 standard recall levels
  - P/R curve
  - Average precision (single value)
  - Precision @ 5, 10, 15, 20, 30, 100, 200, 500, 1000 retrieved documents
  - R-precision
  - Average precision histogram (on each topic)

Summary Statistics	
Run Number	CL99SD
Run Description	Manual
Number of Topics	50
Total number of documents over all topics	
Retrieved:	50000
Relevant:	4728
Rel-ret:	3304

Recall Level Precision Averages	
Recall	Precision
0.00	0.9047
0.10	0.7438
0.20	0.5875
0.30	0.4601
0.40	0.3794
0.50	0.3148
0.60	0.2555
0.70	0.1984
0.80	0.1549
0.90	0.0852
1.00	0.0392
Average precision over all relevant docs	
non-interpolated	0.3537

Document Level Averages	
	Precision
At 5 docs	0.7600
At 10 docs	0.7020
At 15 docs	0.6453
At 20 docs	0.5840
At 30 docs	0.4860
At 100 docs	0.2816
At 200 docs	0.1969
At 500 docs	0.1099
At 1000 docs	0.0661
R-Precision (precision after R docs retrieved (where R is the number of relevant documents))	
Exact	0.3709



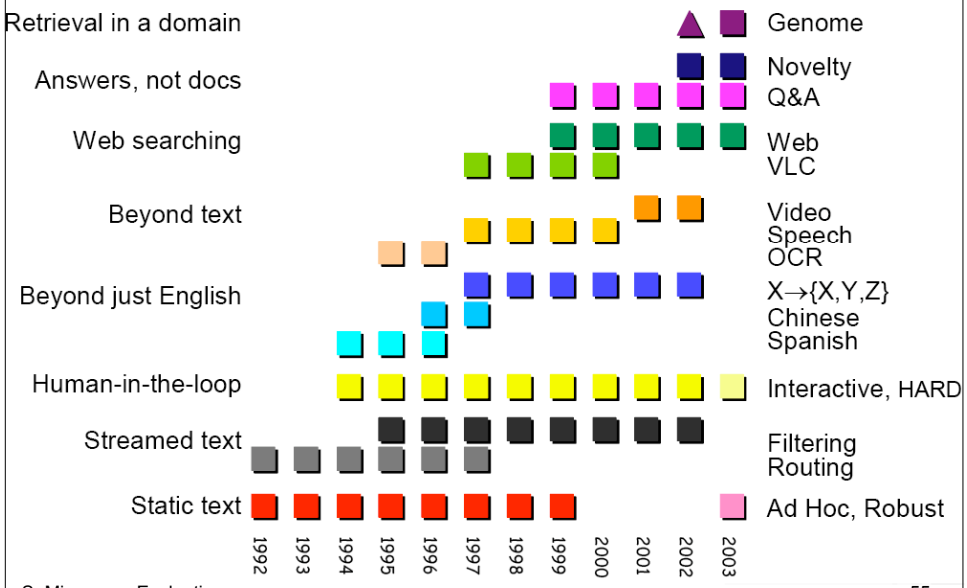


TREC, “tasks”, “tracks”

- So far: “Ad hoc” retrieval task
  - Classical: retrieve the documents that are relevant to a request and rank them in decreasing order of relevance
- Other “tasks”:
  - Information filtering/routing
  - Question answering (provide answers, not just docs.)
  - On the Web
  - …
- On the basis of these “tasks”, other “tracks”
  - Activated/deactivated year after year
  - ≠ collections, ≠ metrics, relevance assessments (non binary relevance), …

S. Mizzaro – Evaluation 54

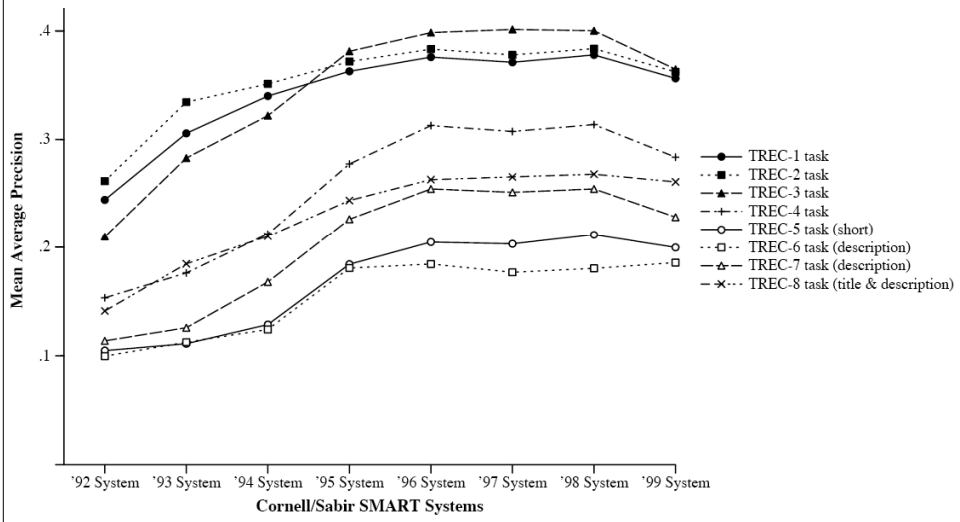
Tracks history [Harman]



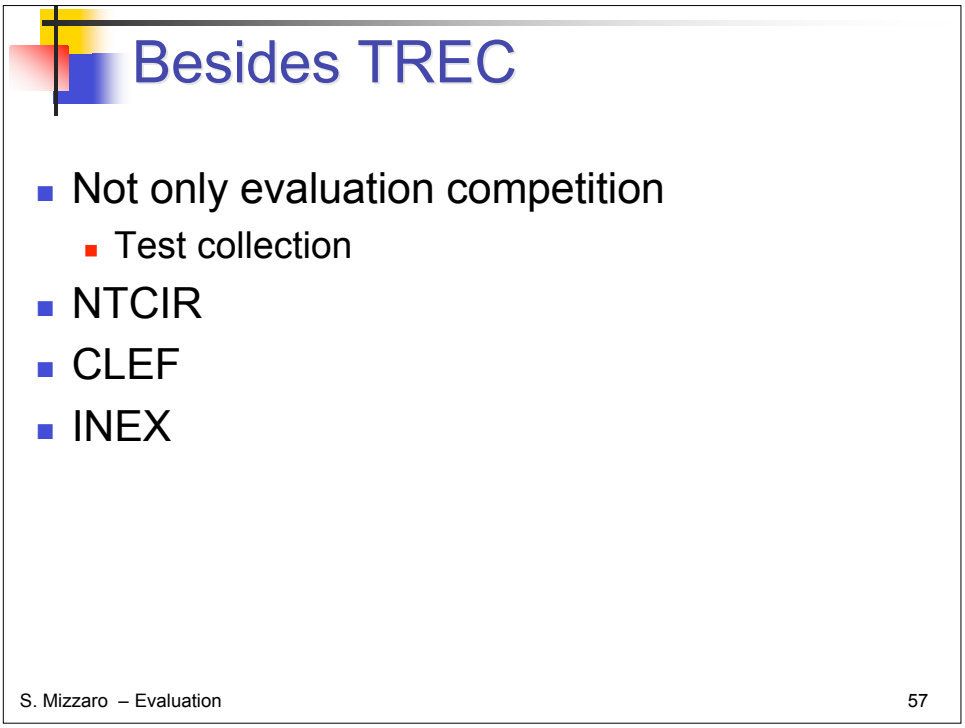
S. Mizzaro – Evaluation 55

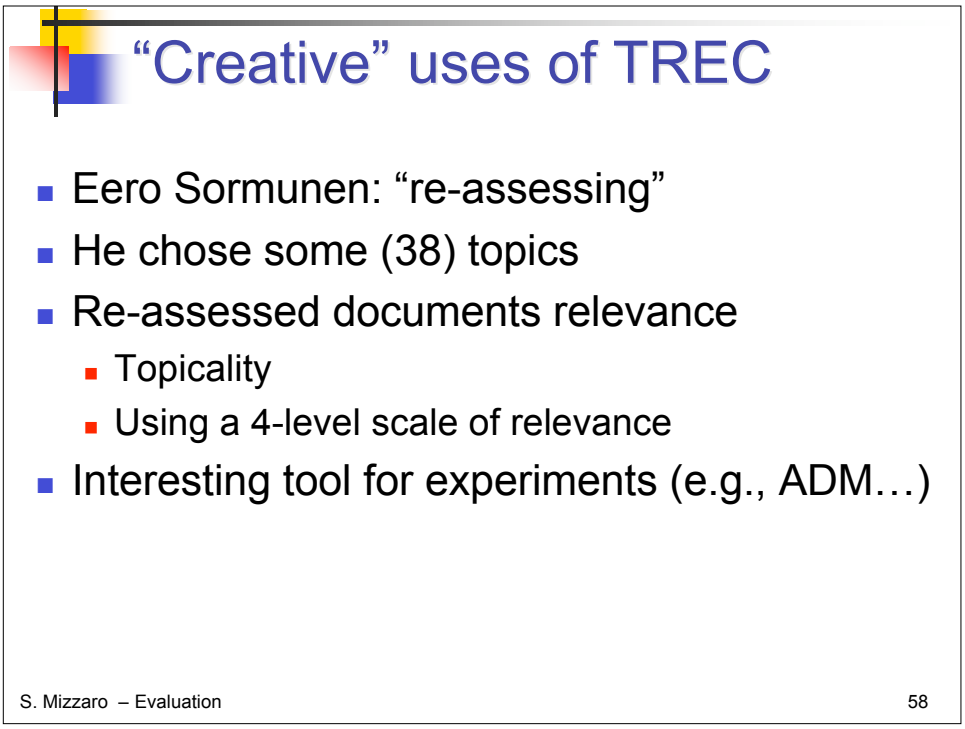
Ad hoc stopped in 1999

- Effectiveness plateau



S. Mizzaro – Evaluation 56


Besides TREC
Not only evaluation competition
Test collection
NTCIR
CLEF
INEX
S. Mizzaro – Evaluation
57


“Creative” uses of TREC
Eero Sormunen: “re-assessing”
He chose some (38) topics
Re-assessed documents relevance
Topicality
Using a 4-level scale of relevance
Interesting tool for experiments (e.g., ADM…)
S. Mizzaro – Evaluation
58

4 relevance levels

- (0) The document does not contain any information about the topic.
- (1) The document only points to the topic. It does not contain more or other information than the topic description. Typical extent: one sentence or fact.
- (2) The document contains more information than the topic description but the presentation is not exhaustive. In case of a multi-faceted topic, only some of the sub-themes or viewpoints are covered. Typical extent: one text paragraph, 2-3 sentences or facts.
- (3) The document discusses the themes of the topic exhaustively. In case of a multi-faceted topic, all or most sub-themes or viewpoints are covered. Typical extent: several text paragraphs, at least 4 sentences or facts.

NTCIR

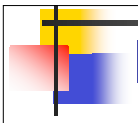
- Nii Test Collection for Information Retrieval systems
  - (NII: National Institute of Informatics, Japan)
- A TREC-like evaluation initiative
- Since 1999, every 18 months
  - (Sep99, Mar01, Oct02, Jun04)
  - # of participants: 28, 36, 65, 74
- Documents and topics on far-eastern languages (Japanese, Chinese, Korean) and English
  - X-lingual, much more complex “alphabet”, morphology, …
- Tasks (= TREC tracks): Web, Patent, QA, …

Metrics & evaluation in NTCIR

- 4 relevance levels:
  - totally relevant (“S”)
  - relevant (“A”)
  - partially relevant (“B”)
  - not relevant (“C”)
- Rigid and relaxed to compute P and R (and MAP, …)
- Study of new metrics

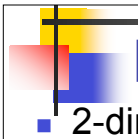
CLEF

- Cross Language Evaluation Forum
- Since 2000, yearly (’00, ’01, ’02, ’03, ’04)
- Aim: Multilingual IR for European languages
- Supported/within DELOS NoE
- # of participants: 20, 34, 37, 42, 55
- Issues:
  - Of course, bilingual, multilingual, X-lingual (→Gareth)
  - Images
  - Spoken
- Effects: significant effectiveness improvement (both multi- and mono-lingual)



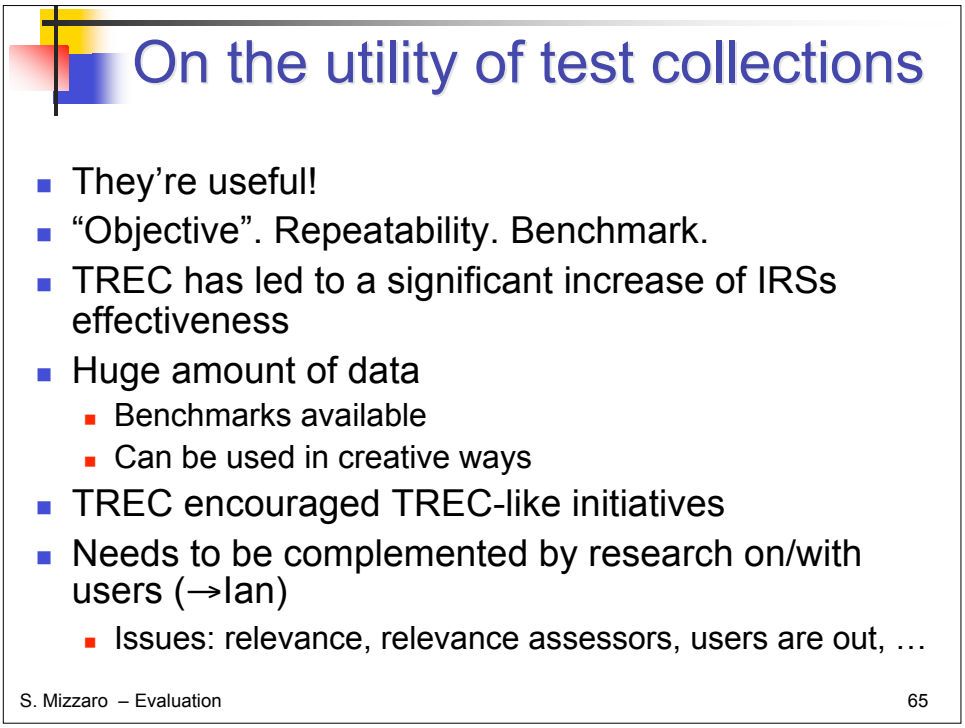
INEX

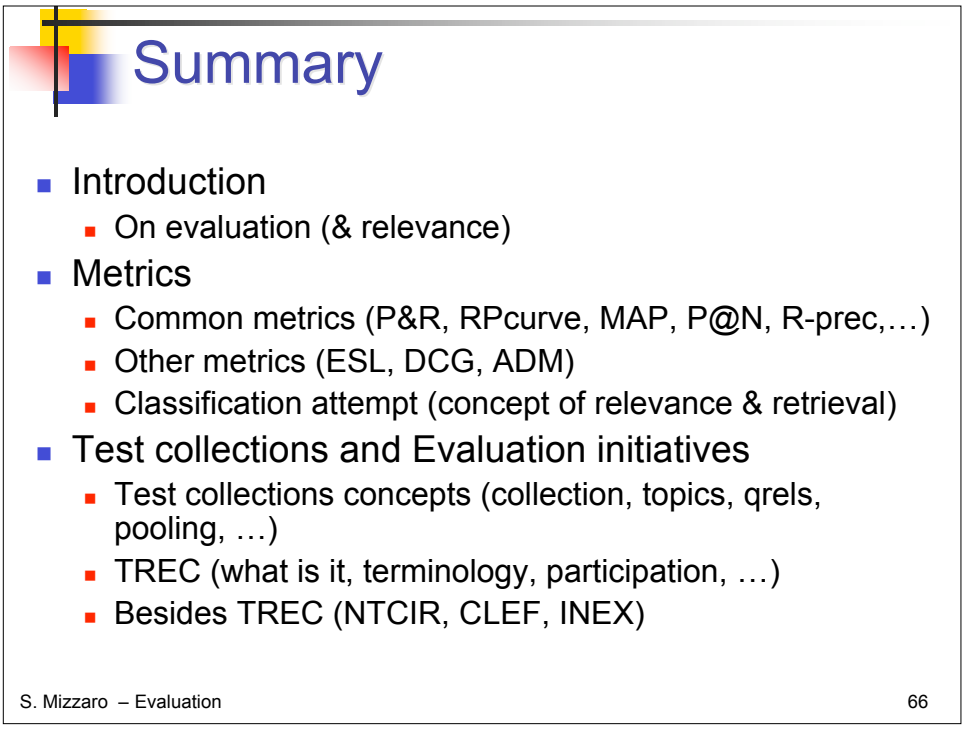
- INitiative for the Evaluation of XML retrieval
- Since 2002, yearly (’02, ’03, ’04)
- Collection:
  - ca. 12000 IEEE papers
  - 12 magazines, 6 transactions, 1995–2002
  - ca. 500MB, ca. 8M “elements”, each article on average ca. 1500 XML nodes, average depth 6.9 nodes
- Requests:
  - Topic
  - Structure (e.g., a document containing a section whose title contains certain terms)
- A lot of discussion on metrics…

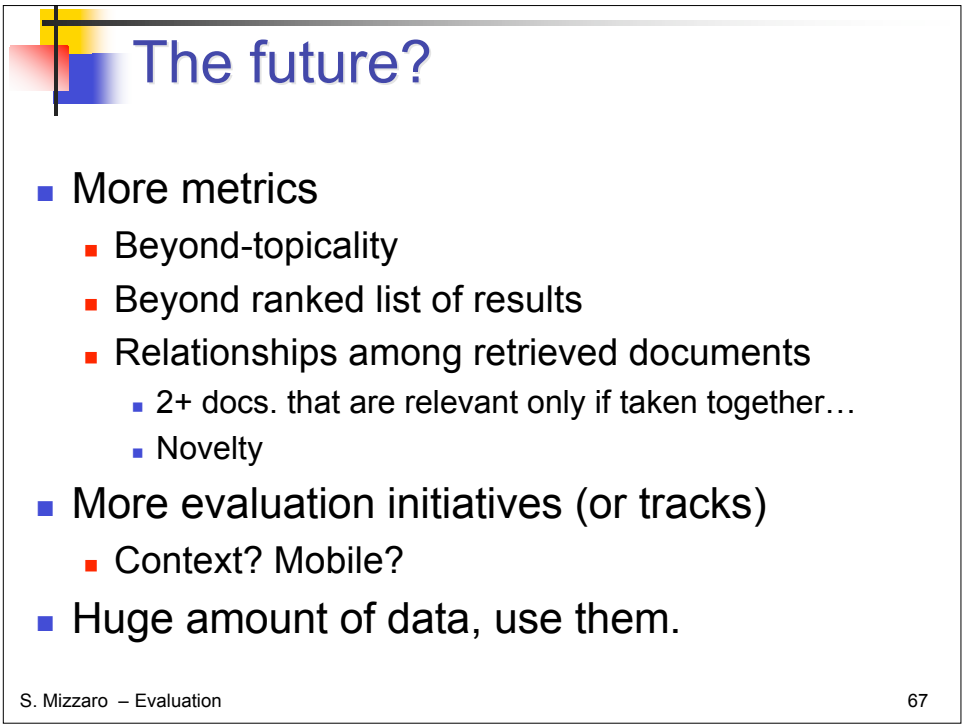


Relevance in INEX

- 2-dimensional, with 4 levels for each dimension
  - e-value: how much the document is exhaustive
    - Not exhaustive (0): the document component does not discuss the topic of request at all
    - Marginally exhaustive (1): the document component discusses only few aspects of the topic of request
    - Fairly exhaustive (2): the document component discusses many aspects of the topic of request
    - Highly exhaustive (3): the document component discusses most or all aspects of the topic of request
  - v-value: how much the document is specific
    - Not specific (0): the topic of request is not a theme of the document component (⇔ e-value=0)
    - Marginally specific (1): the topic of request is a minor theme of the document component
    - Fairly specific (2): the topic of request is a major theme of the document component
    - Highly specific (3): the topic of request is the only theme of the document component


On the utility of test collections
They're useful!
"Objective". Repeatability. Benchmark.
TREC has led to a significant increase of IRSs effectiveness
Huge amount of data
Benchmarks available
Can be used in creative ways
TREC encouraged TREC-like initiatives
Needs to be complemented by research on/with users (→Ian)
Issues: relevance, relevance assessors, users are out, …
S. Mizzaro – Evaluation 65


Summary
Introduction
On evaluation (& relevance)
Metrics
Common metrics (P&R, RPcurve, MAP, P@N, R-prec,…)
Other metrics (ESL, DCG, ADM)
Classification attempt (concept of relevance & retrieval)
Test collections and Evaluation initiatives
Test collections concepts (collection, topics, qrels, pooling, …)
TREC (what is it, terminology, participation, …)
Besides TREC (NTCIR, CLEF, INEX)
S. Mizzaro – Evaluation 66


The future?
More metrics
Beyond-topicality
Beyond ranked list of results
Relationships among retrieved documents
2+ docs. that are relevant only if taken together…
Novelty
More evaluation initiatives (or tracks)
Context? Mobile?
Huge amount of data, use them.
S. Mizzaro – Evaluation
67