# Evaluating Evaluation Measure Stability

Tunc Gultekin Eren Golge Havva Gulay Gurbuz Ahmet Iscen

# Outline

- Goal & Contribution
- Motivation
- Test Environment
- New Approach
- Error Rate Calculation
- Error Rates
- Conclusion

## **Goal & Contribution**

- There are many evaluation measures, which one should we trust for comparing algorithms?
- How do we **interpret** their results?
- Are they **stable**?
- A novel approach for quantifying errors of evaluation measures has been developed.

## Motivation

- To compare performances of different I.R algorithms, some experiments are performed on test collections.
- Relative performances of algorithms are expressed with evaluation measures.
  - Precision, recall...

## Motivation

- Evaluation of these measures rely on some rules of thumb;
  - Experiments must use reasonable evaluation measures.
  - Conclusions must be based on reasonable performace differences.
- The meaning of «Reasonable» can be changed among people.

An objective decision making system is required.

#### **Test Enviroment for New Approach**

- Precision(@), Recall(1000), Precision at 0.5 Recall, R-Precision, Average Precision methods were compared.
- ▶ 9 different I.R algorithms were used from TREC-8
- > 21 different query set.

## New Approach

- > 21 different query set run on 9 different I.R algorithms.
- An evaluation measure was choosen and a fuzziness\* value was defined.
- A query set was selected and the mean of evaluation measure was computed over set for each of 9 retrieval methods.

\*a threshold value that defines if the difference of measures are discriminative enough

## New Approach

- For each pair of retrieval methods, better method was found.
- Another query set was selected and comparisons were repeated multiple times.
- Results are presented in 9x9 comparison matrix.

#### New Approach

	INQa		INC	)e		INÇ	)p		Saba	a		Sab	e		Sabm	1	8	acs			pir	
APL	18 0 3	2	11	8	19	0	2	11	0	10	0	19	2	3	11	7	21	0	0	0	19	2
INQa		0	21	0	4	6	11	0	14	7	0	21	0	0	21	0	21	0	0	0	21	0
INQe					21	0	0	19	0	2	1	16	4	4	4	13	21	0	0	0	17	4
INQp	1							0	15	6	0	21	0	0	21	0	21	0	0	0	21	0
Saba								1			0 [	21	0	0	21	0	21	0	0	0	21	0
Sabe	1													21	0	0	21	0	0	2	4	15
Sabm	1																21	0	0	0	19	2
acs	1				1															0	21	0
a) Average Precision																						
I	INQa	]	INQ	e j	I	NQ	p	:	Saba			Sab	e	I	Sabr	n	1	acs		I	pir	
APL	INQa 2 12 7	0	INQ 19	e 2	I 3	<u>NQ</u> 9	р 9	2	Saba	8	0	Sab 20	e 1	-1	Sabr 14	n 6	13	acs 1	7	0	pir 19	2
APL INQa	INQa 2 12 7	0 0	INQ 19 14	e 2 7	1 3 4	NQ 9 2	р 9 15	2	Saba 11 6	8 13	0	Sab 20 21	e 1 0	1	Sabr 14 9	n 6 12	13 18	acs 1 0	73	0	pir 19 15	2 6
APL INQa INQc	INQa 2 12 7	0 0	INQ 19 14	e 2 7	1 3 4 20	NQ 9 2 0	9 15 1	2 2 16	Saba 11 6 1	8 13 4	0 0 4	Sab 20 21 6	e 1 0 11	1 0 14	Sabr 14 9 2	n 6 12 5	13 18 21	acs 1 0 0	7 3 0	006	pir 19 15 4	2 6 11
APL INQa INQc INQp	INQa 2 12 7	00	INQ 19 14	e 2 7	1 3 4 20	NQ 9 2 0	9 15 1	2 2 16 2	Saba 11 6 1 5	8 13 4 14	0 0 4 0	Sab 20 21 6 20	e 1 11 1	1 0 14 1	Sabr 14 9 2 12	n 12 5 8	13 18 21 18	acs 1 0 0	7 3 0 3	0060	pir 19 15 4 19	2 6 11 2
APL INQa INQc INQp Saba	INQa 2 12 7	00	INQ 19 14	e 2 7	1 3 4 20	NQ 9 2 0	9 15 1	2 2 16 2	Saba 11 6 1 5	8 13 4 14	0 0 4 0	Sab 20 21 6 20 19	e 1 11 1 2	1 0 14 1 0	Sabr 14 9 2 12 6	n 12 5 8 15	13 18 21 18 17	acs 1 0 0 0	7 3 0 3 4	0 0 6 0 0	pir 19 15 4 19 16	2 6 11 2 5
APL INQa INQe INQp Saba Saba	INQa 2 12 7	000	INQ 19 14	e 2 7	1 3 4 20	NQ 9 2 0	p 9 15 1	2 2 16 2	Saba 11 6 1 5	8 13 4 14	0 0 4 0 0	Sab 20 21 6 20 19	e 1 11 1 2	1 0 14 1 0 18	Sabr 14 9 2 12 6 0	n 12 5 8 15 3	13 18 21 18 17 21	acs 1 0 0 0 0	7 3 0 3 4 0	0 0 6 0 8	pir 19 15 4 19 16 1	2 6 11 2 5 12
APL INQa INQe INQp Saba Sabe Sabe	INQa 2 12 7	00	INQ 19 14	e 2 7	I 3 4 20	NQ 9 2 0	9 15 1	2 2 16 2	Saba 11 6 1 5	8 13 4 14	0 0 4 0 0	Sab 20 21 6 20 19	e 1 11 1 2	1 0 14 1 0 18	Sabr 14 9 2 12 6 0	n 12 5 8 15 3	13 18 21 18 17 21 19	acs 1 0 0 0 0 0	7 3 0 3 4 0 2	0 0 6 0 0 8 1	pir 19 15 4 19 16 1 12	2 6 11 2 5 12 8
APL INQa INQe INQp Saba Sabe Sabm acs	INQa 2 12 7	00	INQ 19 14	e 2 7	1 3 4 20	NQ 9 2 0	p 9 15 1	2 2 16 2	Saba 11 6 1 5	8 13 4 14	0 0 4 0 0	Sab 20 21 6 20 19	e 1 11 1 2	1 0 14 1 0 18	Sabr 14 9 2 12 6 0	n 12 5 8 15 3	13 18 21 18 17 21 19	acs 1 0 0 0 0 0 0	7 3 0 3 4 0 2	0 6 0 8 1 0	pir 19 15 4 19 16 1 12 21	2 6 11 2 5 12 8 0

b) Prec(10)

## **Error Rate Calculation**

- For each cell in the matrix, greater value of betterthan, worse than values were accepted as correct answer and other one is error.
- Lesser values of all cells were summed and divided by total number of decisions.
  - Error Rate for Average Precision Matrix;
    - 16 / 756 = 0.021 = 2.1%

**Error\*** = min(A > B, B > A)/ (A > B + B > A + A = = B)

\*if **Error** of the comparison matrix >~ 25% then discrimination converges to randomness

#### **Error Rates**

Measure	Error Rate (%)	Std. Dev. (%)	Ties (%)
Prec(1)	14.3	1.3	23.4
Prec(10)	3.6	0.9	24.3
Prec(30)	2.9	0.8	23.8
Prec at .5 R	2.2	0.5	11.4
Prec(100)	1.8	0.5	20.7
Ave Prec	1.5	0.4	12.8
R-Prec	1.3	0.4	19.1
Prec(1000)	1.0	0.4	22.5
Recall(1000)	0.6	0.2	20.8

## Conclusion

- Error rates of evaluation measures inversely proportional with the topic set size.
- Query sets should be carefully choosen.
  - Something may be biased.
- Recall(1000) is very stable but it appropriated for limited environments.
- Average Precision is good for general purpose.
- Precision at a cut off level is appropriate for web.