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	INQe	INQp	Saba	Sabe	Sabm	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				0 15 6	0 21 0	0 21 0	21 0 0	0 21 0
Saba					0 21 0	0 21 0	21 0 0	0 21 0
Sabe						21 0 0	21 0 0	2 4 15
Sabm							21 0 0	0 19 2
acs								0 21 0

a) Average Precision

	INQa	INQe	INQp	Saba	Sabe	Sabm	acs	pir
APL	2 12 7	0 19 2	3 9 9	2 11 8	0 20 1	1 14 6	13 1 7	0 19 2
INQa		0 14 7	4 2 15	2 6 13	0 21 0	0 9 12	18 0 3	0 15 6
INQe			20 0 1	16 1 4	4 6 11	14 2 5	21 0 0	6 4 11
INQp				2 5 14	0 20 1	1 12 8	18 0 3	0 19 2
Saba					0 19 2	0 6 15	17 0 4	0 16 5
Sabe						18 0 3	21 0 0	8 1 12
Sabm							19 0 2	1 12 8
acs								0 21 0

b) Prec(10)

Error Rate Calculation

- For each cell in the matrix, greater value of better-than, 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.