PAPER REVIEW

Presented By: Udochukwu Nweke

CS734: Introduction to Information Retrieval

Dr. Michael Nelson

November 30, 2017

Paper 1

How Reliable are the Results of Large-Scale Information Retrieval Experiments?

Justin Zobel
Department of Computer Science, RMIT, GPO Box, 2476V, Melbourne 3001, Australia
jz@cs.rmit.edu.au

Date: August 24 - 28, 1998

Paper 2

Evaluating Evaluation Measure Stability

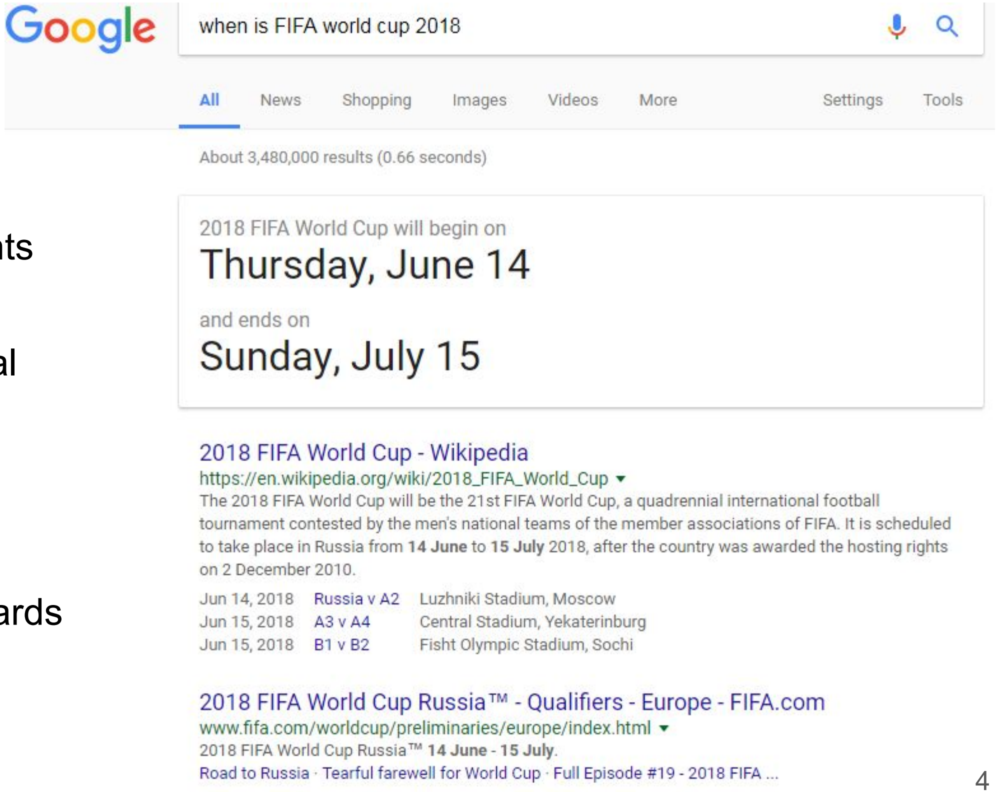
Chris Buckley Sabir
Research Inc.
Gaithersburg, MD 20878
chrisb @ sabir.com

Ellen M. Voorhees
National Institute of Standards and
Technology Gaithersburg, Maryland
20899
ellen.voorhees @nist.gov

Date: July, 2017

Measuring Information Retrieval System

- A user sends a query that represents information need
- The retrieval system identifies documents that mostly answers the query
- Based on the result, information retrieval can be measured with respect to test collection
- A Test collection is a collection of document sets, query sets, and each document’s relevant information as regards to the query.



when is FIFA world cup 2018

About 3,480,000 results (0.66 seconds)

2018 FIFA World Cup will begin on

Thursday, June 14

and ends on

Sunday, July 15

2018 FIFA World Cup - Wikipedia

https://en.wikipedia.org/wiki/2018_FIFA_World_Cup

The 2018 FIFA World Cup will be the 21st FIFA World Cup, a quadrennial international football tournament contested by the men's national teams of the member associations of FIFA. It is scheduled to take place in Russia from **14 June** to **15 July** 2018, after the country was awarded the hosting rights on 2 December 2010.

Jun 14, 2018	Russia v A2	Luzhniki Stadium, Moscow
Jun 15, 2018	A3 v A4	Central Stadium, Yekaterinburg
Jun 15, 2018	B1 v B2	Fisht Olympic Stadium, Sochi

2018 FIFA World Cup Russia™ - Qualifiers - Europe - FIFA.com

www.fifa.com/worldcup/preliminaries/europe/index.html

2018 FIFA World Cup Russia™ **14 June - 15 July**.

Road to Russia · Tearful farewell for World Cup · Full Episode #19 - 2018 FIFA ...

Reliability of System Measurement

- The reliability of the measurement of a system depends on the quality of relevance judgement

- In most cases, relevant assessors are not always in agreement. This can introduce error into information retrieval experiments

- If an assessment is done thoroughly without considering whether query terms occur in each document alone, there will be minimal error in measurement of relative performance of a system

Forming Relevant Judgement from Database of Few Records

- Before now, information retrieval experiments used databases of fewer records

- The method for choosing which document to assess and the method for measuring system performance can be unreliable

- Few database records can trivialize the retrieval problem and they are not always rich enough to distinguish between retrieval methods of different power

- The size of database allowed for complete relevant judgement to be formed

Dealing with Large Database Records

- The TREC collection used a larger experimental database which provides a more realistic test environment but prevents relevant assessment.
- The size of the collection made it difficult to determine relevant assessment
- Because of the size of the collection, a technique has to be introduced in order to determine documents to be considered for relevance assessment

Pooling Technique

- TREC used pooling technique to get some documents from a large collection
  - top *k results (for TREC, k varied between 50 and* 200) from the rankings obtained by different search engines are merged into a pool
  - duplicates are removed
  - documents are presented in some random order for assessment
- Produces a large number of relevance judgments for each query

Significance Standards

- Given two systems, A and B, and a measurement technique, how do we decide that the difference in measurement is significant?

- Significance in information retrieval has to do with the difference in the mean performance of two systems for a given set of queries

- T-test, ANOVA and Wilcoxon’s test were used to measure system performance in order to determine standard significance

Test That Determines Significance

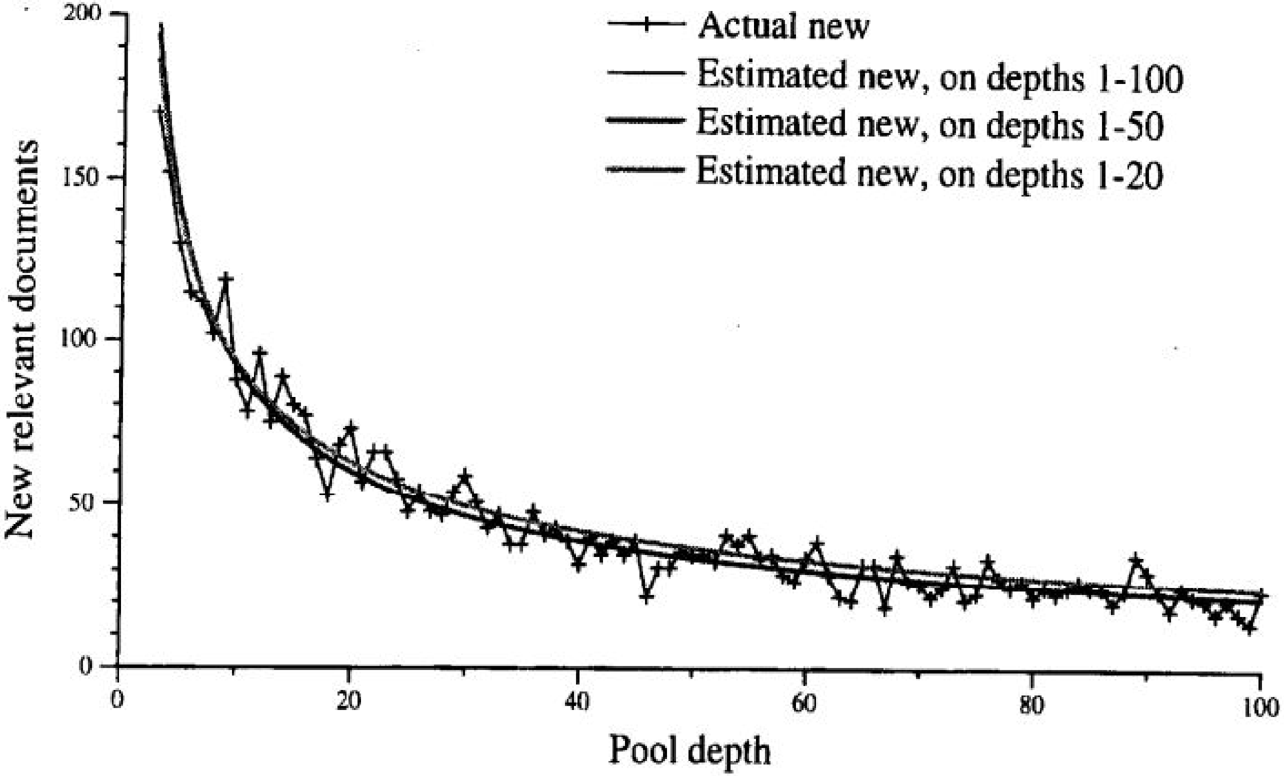
- Wilcoxon’s test confirmed 94%-98% results. Given its reliability and greater power, the Wilcoxon test should be used to determine significance.
- For Wilcoxon’s test, 94%-98% of results were confirmed. For llpt and nllpt around 97% of results were confirmed
- We conclude that, given its reliability and greater power, the Wilcoxon test should be used for determining significance.

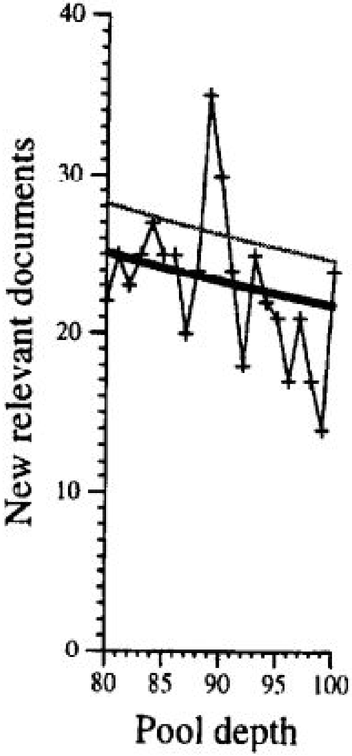
Difficulties with Pool Depth

The TREC pool size is 100 but the measurement depth is 1000; that is all 1000 documents in each run are considered during system performance measurement

- Considering a fraction of relevant document that are identified alone may have its effectiveness underestimated by system performance evaluation measurement
- System that are good at recall may continue to fetch relevant document at depth greater than 100
- If this assumption is correct,the entire measurement of system effectiveness will change a little if the pool depth is increased.

Reliability of Recall





Findings

- Results based on relevance judgement formed from limited pool depth is reliable if pool is sufficiently deep for systems that contributed to the pool and for new systems.
- TREC limit of 100 appears to be adequate
- TREC collection have been successful in identifying relevant document even if most relevant documents are not identified
- The assumption that unjudged are not relevant is not well establish

Paper 2

Evaluating Evaluation Measure Stability

Chris Buckley Sabir
Research Inc.
Gaithersburg, MD 20878
chrisb @ sabir.com

Ellen M. Voorhees
National Institute of Standards and
Technology Gaithersburg, Maryland
20899
ellen.voorhees @nist.gov

Date: July 2017

Rules of Thumb for Accepting Experimental Design

- Test collection should have a reasonable number of requests. TREC program committee used 25 requests as minimum and 50 requests as the norm
- A reasonable evaluation measure must be used for the experiment. The most common measures are: Average Precision, R-precision, and Precision at 20 (or 10 or 30) documents retrieved
- Conclusion must be based on a significant difference

Objective

The objective of this paper is as follows:

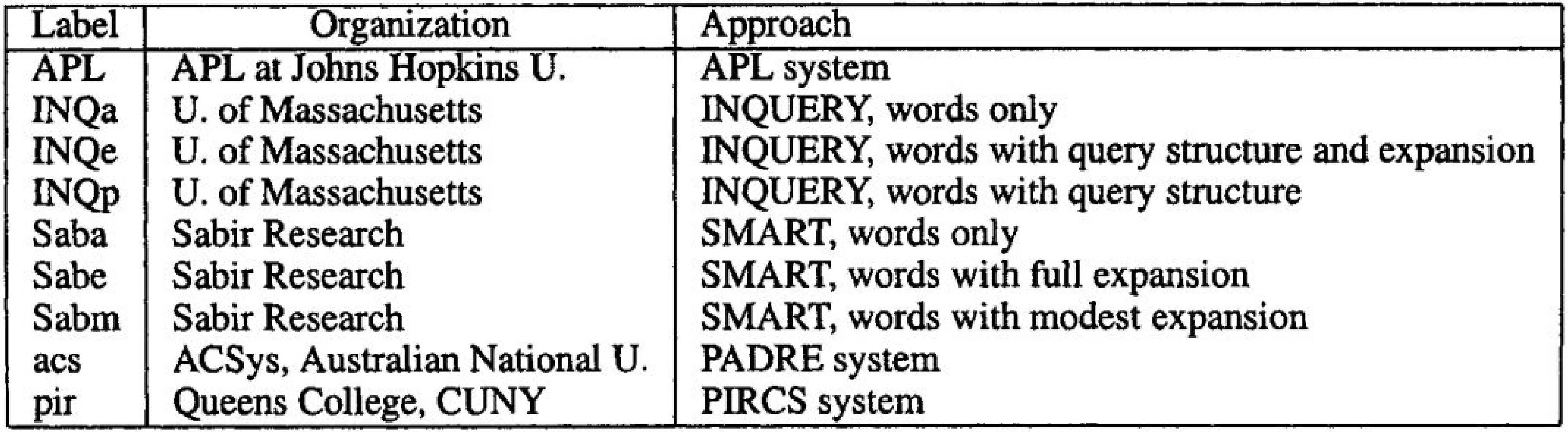
- To examine these rules of thumb and show how they interact with each other
- Present a novel approach to experimentally measure the likely error related with the given conclusion “ One method is better than the order” given a number of requests, evaluation measure, and a notion of difference
- Provide researchers with confidence intervals on how reliable the conclusions drawn from test scores are.

Test Environment

In order to help researchers design effective retrieval experiment, the following measures were investigated

- Prec(A): Precision at cut-off level A, for A = 1, 2, 5, 10, 15, 20, 30, 50, 100, 300, 1000. A cut-off level is a rank that defines the retrieved set; for example, a cut-off level of ten defines the retrieved set as the top ten documents in the ranked list.
- Recall(1000): Recall after 1000 documents have been retrieved.
- Prec at .5 Recall: Precision after half the relevant document have been retrieved.
- R-Prec: Precision after R documents have been retrieved
- Average Precision: The mean of the precision scores obtained after each relevant document is retrieved.

Retrieval Methods used in TREC-8 Query Track



Label	Organization	Approach
APL	APL at Johns Hopkins U.	APL system
INQa	U. of Massachusetts	INQUERY, words only
INQe	U. of Massachusetts	INQUERY, words with query structure and expansion
INQp	U. of Massachusetts	INQUERY, words with query structure
Saba	Sabir Research	SMART, words only
Sabe	Sabir Research	SMART, words with full expansion
Sabm	Sabir Research	SMART, words with modest expansion
acs	ACSys, Australian National U.	PADRE system
pir	Queens College, CUNY	PIRCS system

Label given to different methods and the organizations that made the run

Determining Error Rate

The following approach is used to determine if a given measure is better than the other

- Choose an evaluation measure and a “fuzziness” value
- Pick a query set and compute the mean evaluation measure over that query set for each of the retrieval methods
- For each method, compare whether the first method is better than, worse than, and equal to the second method with respect to the fuzziness value
- The fuzziness value is a percentage difference between scores

Error Rate

- Fuzziness value of 5% and 21 sets queries submitted to Query Track
- The first number in each entry gives the number of time the retrieval method of the row is better than the retrieval method of the column
- The matrix represents 756 decisions regarding the relative effectiveness of retrieval methods
- The correct answer for each pair is given by the greater of the better-than and worse than values
- The lesser of these values is the number of times a test is misleading or in error

Error Rate

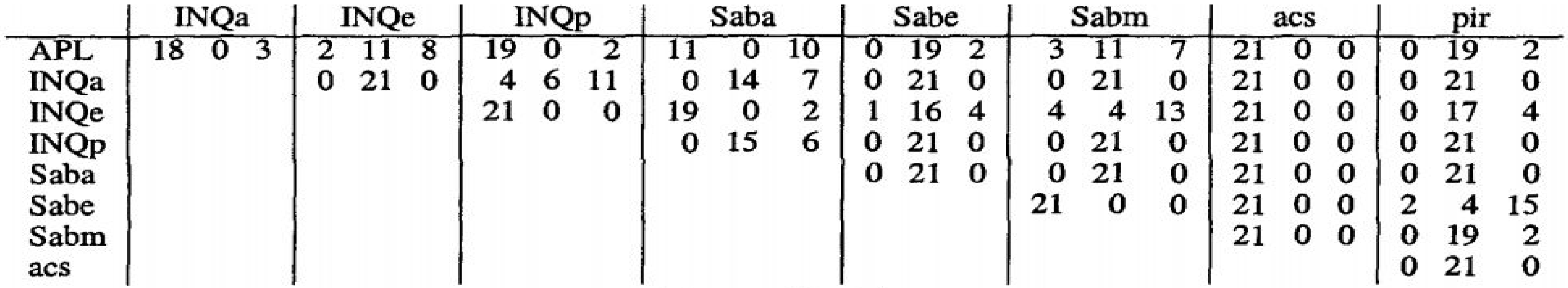
The error rate is defined by the number of errors across all method pairs divided by the total number of decisions

$$ErrorRate = \frac{\sum Min(|A > B|, |B > A|)}{\sum(|A > B| + |A < B| + |A == B|)}$$

where IA > BI is the number of times method A is better than method B in an entry

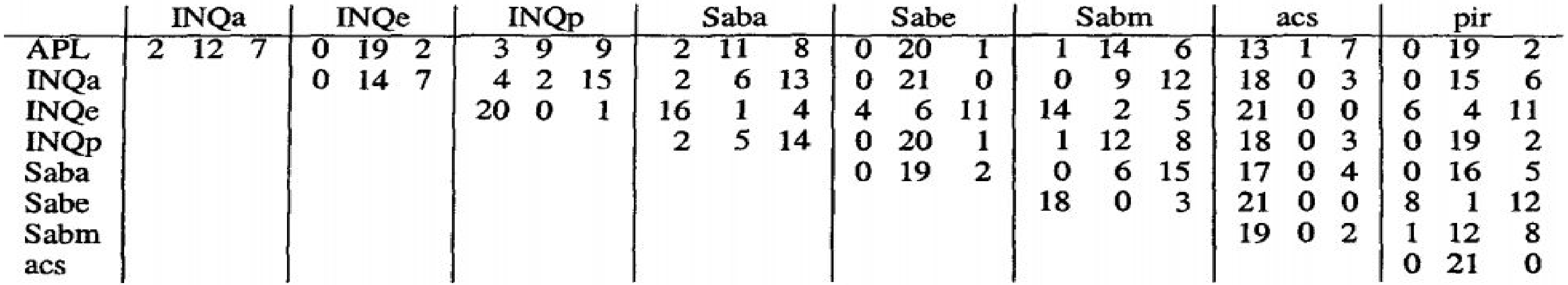
Average Precision matrix in Figure 1 is 16/756 = .021 or 2.1%. Similarly, the error rate for the Prec(10) matrix is 29/756 = .038 or 3.8%.

Testing Evaluation Measures with Fuzziness Value



	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

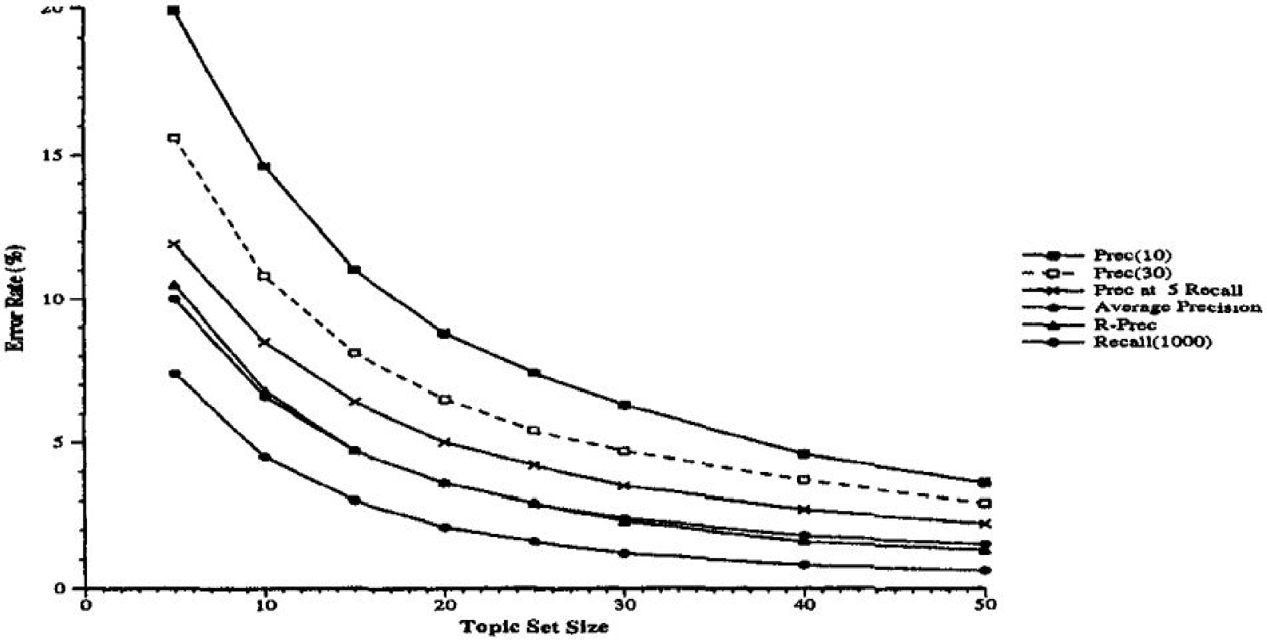
b) Prec(10)

Changing Topic Set Size

- Informational retrieval rules of thumb states that reasonable number of topics should be used for experiment
- Can Changing the number of topics used in test affect the error rate of evaluation measures?

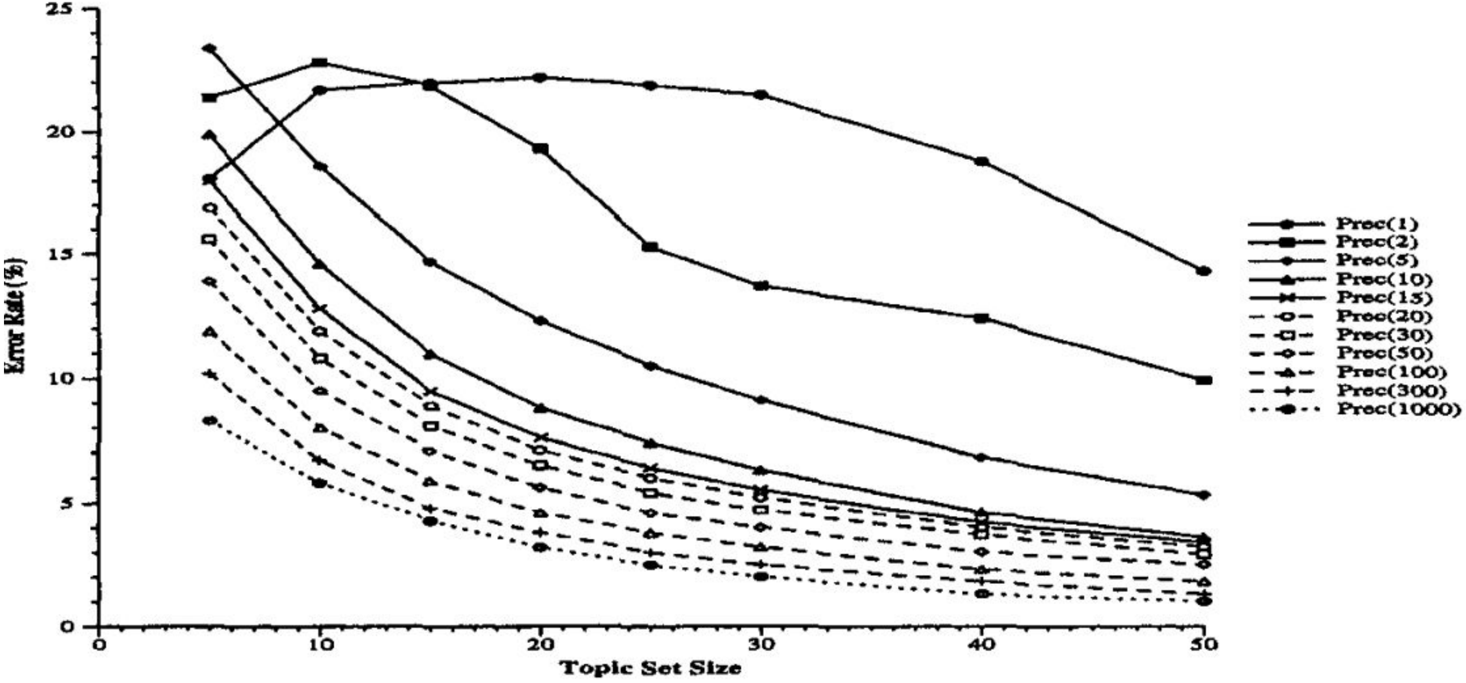
Average Error Rate of Evaluation Measures for Changing Topic Size

For all measures the average error rate decreases as the number of topics increases



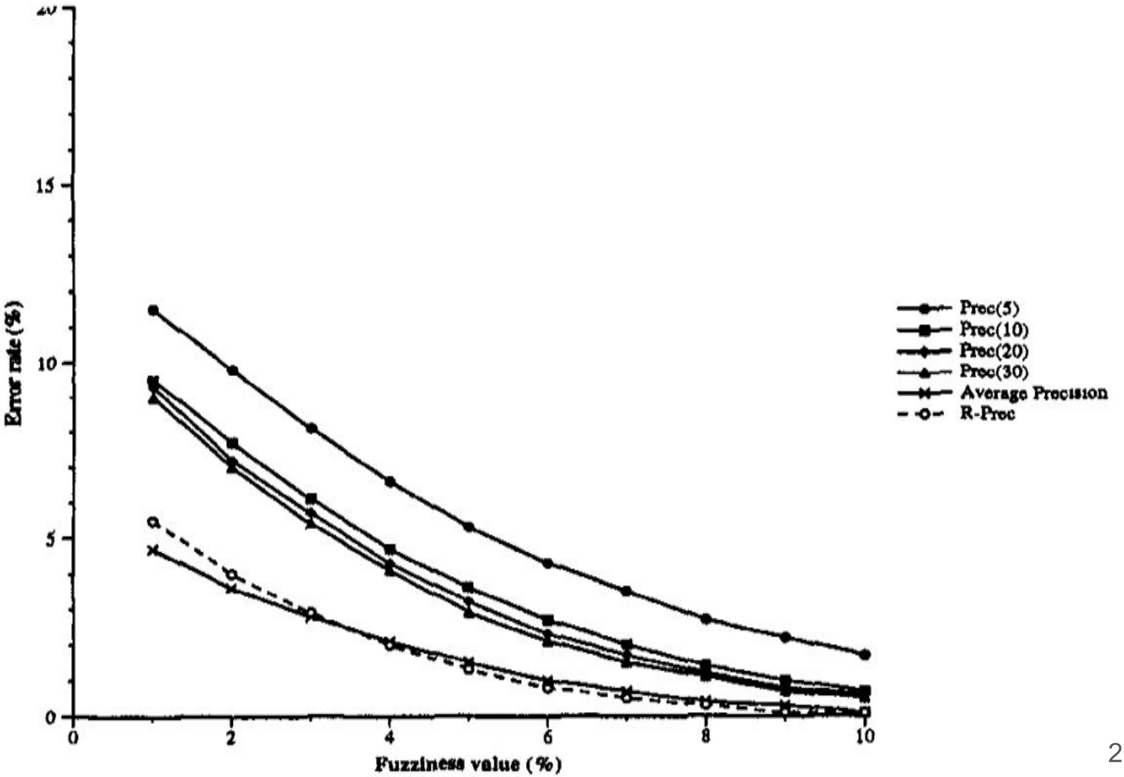
Average Error Rate of Precision at Different Cut-off Levels for Changing Topic Size Set

Precision Error rate: Precision a 5 and greater gives a smaller error rate

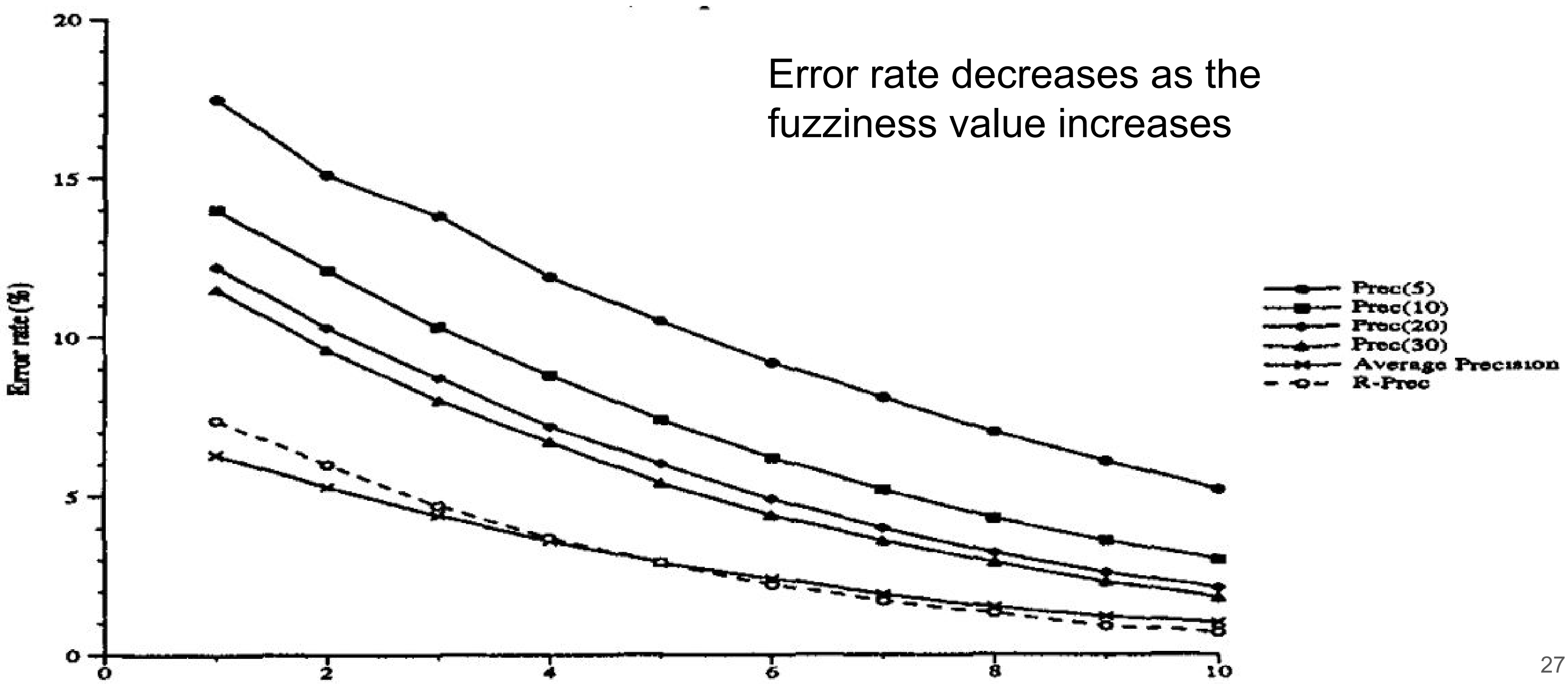


Changing Fuzziness Values (Topic Set Size= 50)

Error rate decreases as the fuzziness value increases



Changing Fuzziness Values (Topic Set Size= 25)



Error rate decreases as the fuzziness value increases

Findings

- The consistent decrease in error rate for larger topic set sizes shows that one way to increase the confidence in conclusions drawn from measures with relatively large inherent error rates is increase topics in the experiments.

- Another way of increasing the reliability of an experimental conclusion is to increase the amount of difference required between scores to conclude that the methods differ.

- The error rate of an evaluation measure is only one of the measure’s properties

- The evaluation measure to be used in retrieval experiment should be picked based on the particular aspect of retrieval behavior that is of interest

Relationship Between Paper 1 and Paper 2

- The relationship between Paper one and Paper two is that Paper one validated the test collection created through TREC workshop by demonstrating the stability of relative retrieval scores despite incomplete relevant judgement and different option as to what constitutes relevant judgement.

- Paper one also proposed a technique for building large test collections

- Paper two agrees on these results.