



SNS COLLEGE OF ENGINEERING

Kurumbapalayam (Po), Coimbatore – 641 107

An Autonomous Institution

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A' Grade
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

**COURSE NAME : 19CS732 INFORMATION RETRIEVAL
TECHNIQUES**

IVYEAR / VII SEMESTER

Unit 2- MODELING AND RETRIEVAL EVALUATION

Topic 8 : Precision and Recall and Reference Collection





Problem



- Makes experimental work hard
 - Especially on a large scale
- In some very specific settings, can use proxies
 - E.g.: for approximate vector space retrieval, we can compare the cosine distance closeness of the closest docs to those found by an approximate retrieval algorithm
- But once we have test collections, we can reuse them (so long as we don't overtrain too badly)



Precision and Recall



Precision:

fraction of retrieved docs that are relevant = $P(\text{relevant}|\text{retrieved})$

Recall:

fraction of relevant docs that are retrieved = $P(\text{retrieved}|\text{relevant})$

	Relevant	Nonrelevant
Retrieved	tp	fp
Not Retrieved	fn	tn



Should we instead use the accuracy measure for evaluation?



- Given a query, an engine classifies each doc as “Relevant” or “Nonrelevant”
- The **accuracy** of an engine: the fraction of these classifications that are correct
 - $(tp + tn) / (tp + fp + fn + tn)$
- **Accuracy** is a commonly used evaluation measure in machine learning classification work
- Why is this not a very useful evaluation measure in IR?



Difficulties in using Precision/Recall



- Should average over large document collection/query ensembles
- Need human relevance assessments
 - People aren't reliable assessors
- Assessments have to be binary
 - Nuanced assessments?
- Heavily skewed by collection/authorship
 - Results may not translate from one domain to another



Precision/Recall –Cont..



Combined measure that assesses precision/recall tradeoff is **F measure**
(weighted harmonic mean):

$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha) \frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

People usually use balanced F_1 measure

i.e., with $\beta = 1$ or $\alpha = \frac{1}{2}$

Harmonic mean is a conservative average

See CJ van Rijsbergen, *Information Retrieval*



Kappa measure for inter-judge (dis)agreement



Kappa measure

Agreement measure among judges

Designed for categorical judgments

Corrects for chance agreement

$$\text{Kappa} = [P(A) - P(E)] / [1 - P(E)]$$

P(A) – proportion of time judges agree

P(E) – what agreement would be by chance

Kappa = 0 for chance agreement, 1 for total agreement.



Kappa Measure: Example

Number of docs	Judge 1	Judge 2
300	Relevant	Relevant
70	Nonrelevant	Nonrelevant
20	Relevant	Nonrelevant
10	Nonrelevant	Relevant



Kappa Example



$$P(A) = 370/400 = 0.925$$

$$P(\text{nonrelevant}) = (10+20+70+70)/800 = 0.2125$$

$$P(\text{relevant}) = (10+20+300+300)/800 = 0.7878$$

$$P(E) = 0.2125^2 + 0.7878^2 = 0.665$$

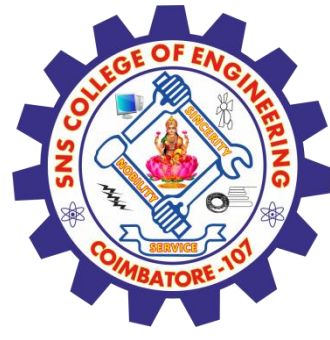
$$\text{Kappa} = (0.925 - 0.665)/(1-0.665) = 0.776$$

Kappa > 0.8 = good agreement

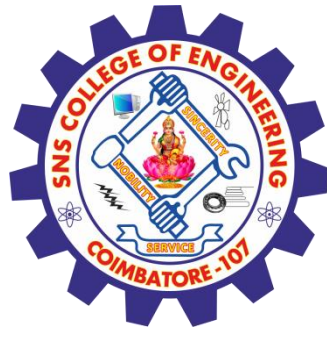
0.67 < Kappa < 0.8 -> “tentative conclusions” (Carletta '96)

Depends on purpose of study

For >2 judges: average pairwise kappas



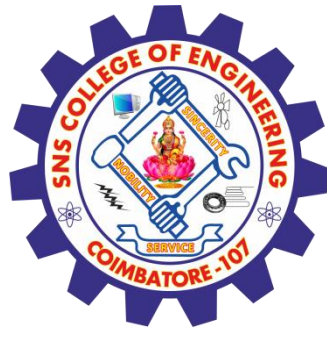
Activity



Disadvantages



- A document can be redundant even if it is highly relevant
- Duplicates
- The same information from different sources
- Marginal relevance is a better measure of utility for the user.
- Using facts/entities as evaluation units more directly measures true relevance.
- But harder to create evaluation set



Advantages



- Impact on **absolute** performance measure can be significant (0.32 vs 0.39)
- Little impact on ranking of different systems or **relative** performance
- Suppose we want to know if algorithm A is better than algorithm B
- A standard information retrieval experiment will give us a reliable answer to this question.



Assessment 1



1. List out the Advantages of Precision and Recall and Reference Collection

- a) _____
- b) _____
- c) _____
- d) _____

2. Identify the disadvantages of Precision and Recall and Collection

- a) _____
- b) _____
- c) _____
- d) _____





TEXT BOOKS:

1. Ricardo Baeza-Yates and Berthier Ribeiro-Neto, –Modern Information Retrieval: The Concepts and Technology behind Search, Second Edition, ACM Press Books, 2011.
2. Ricci, F, Rokach, L. Shapira, B.Kantor, –Recommender Systems Handbook||, First Edition, 2011.

REFERENCES:

1. C. Manning, P. Raghavan, and H. Schütze, –Introduction to Information Retrieval, Cambridge University Press, 2008.
2. Stefan Buettcher, Charles L. A. Clarke and Gordon V. Cormack, –Information Retrieval: Implementing and Evaluating Search Engines, The MIT Press, 2010.

THANK YOU