



Processamento
e Recuperação
de Informação

Evaluation
and Relevance

Precision vs.
Recall

Other
Measures

Obtaining the
Ground Truth

Evaluation of
Classifiers

Processamento e Recuperação de Informação

Evaluation of IR and IE Systems

Departamento de Engenharia Informática
Instituto Superior Técnico

1º Semestre
2018/2019



Bibliography

Processamento
e Recuperação
de Informação

Evaluation
and Relevance

Precision vs.
Recall

Other
Measures

Obtaining the
Ground Truth

Evaluation of
Classifiers

Bing Liu, Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, 2nd edition. Chapter 6.

Ricardo Baeza-Yates, Berthier Ribeiro-Neto, Modern Information Retrieval, 2nd edition. Chapter 4.

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval. Chapter 8.



Outline

Processamento
e Recuperação
de Informação

Evaluation
and Relevance

Precision vs.
Recall

Other
Measures

Obtaining the
Ground Truth

Evaluation of
Classifiers

1 Evaluation and Relevance

2 Precision vs. Recall

3 Other Measures

4 Obtaining the Ground Truth

5 Evaluation of Classifiers



IR System Evaluation

Processamento
e Recuperação
de Informação

Evaluation
and Relevance

Precision vs.
Recall

Other
Measures

Obtaining the
Ground Truth

Evaluation of
Classifiers

Why evaluate?

- Measure the benefit of using an IR system
- Measure how well an IR system fulfills its goal
- Compare IR systems

What to evaluate?

- Collection coverage
- Processing time
- Output presentation
- User effort
- Recall and Precision



Elements of an information retrieval performance evaluation experiment

Processamento
e Recuperação
de Informação

Evaluation
and Relevance

Precision vs.
Recall

Other
Measures

Obtaining the
Ground Truth

Evaluation of
Classifiers

The Cranfield Paradigm

An IR experiment, as devised by Cyril Cleverdon (1950s), must include:

- 1 A reference collection
- 2 Relevance judgments
- 3 An evaluation metric



Relevant Documents

Processamento
e Recuperação
de Informação

Evaluation
and Relevance

Precision vs.
Recall

Other
Measures

Obtaining the
Ground Truth

Evaluation of
Classifiers

Recall and Precision

Measure the ability of a system to return **relevant** documents.

Relevance

- Subjective notion
- Usually **evaluated by a set of experts**



Outline

Processamento
e Recuperação
de Informação

Evaluation
and Relevance

Precision vs.
Recall

Other
Measures

Obtaining the
Ground Truth

Evaluation of
Classifiers

1 Evaluation and Relevance

2 Precision vs. Recall

3 Other Measures

4 Obtaining the Ground Truth

5 Evaluation of Classifiers



Evaluating Prediction

Processamento
e Recuperação
de Informação

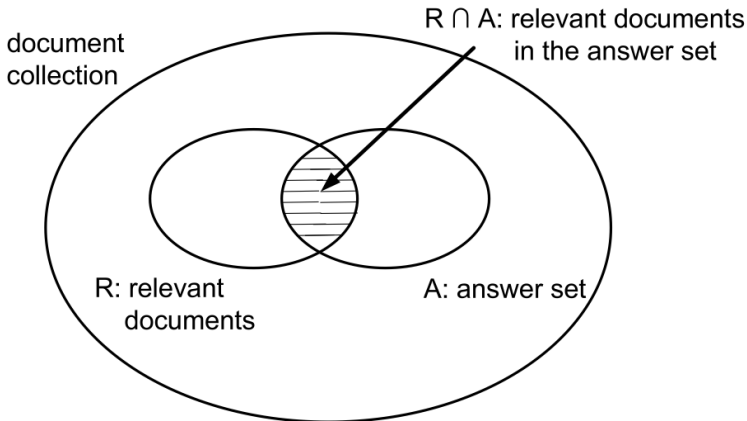
Evaluation
and Relevance

Precision vs.
Recall

Other
Measures

Obtaining the
Ground Truth

Evaluation of
Classifiers





Measuring Precision and Recall

Processamento
e Recuperação
de Informação

Evaluation
and Relevance

Precision vs.
Recall

Other
Measures

Obtaining the
Ground Truth

Evaluation of
Classifiers

Definition

Let A be the set of documents retrieved for query Q .

Let R be the set of documents that are relevant to query Q .

Precision is the proportion of retrieved documents that are relevant, i.e.:

$$Pr = \frac{|R \cap A|}{|A|}$$

Recall is the proportion of relevant documents retrieved, i.e.:

$$Re = \frac{|R \cap A|}{|R|}$$



Precision-Recall Curves

Processamento
e Recuperação
de Informação

Evaluation
and Relevance

Precision vs.
Recall

Other
Measures

Obtaining the
Ground Truth

Evaluation of
Classifiers

- Retrieved documents are ordered \Rightarrow we are interested in measuring how precision changes as recall increases

Example

Let $A = \{d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8, d_9, d_{10}\}$ be an ordered set of retrieved documents, for a query Q .

Let $R = \{d_2, d_5, d_8, d_{15}\}$ be the set of relevant documents for query Q .

Re	Pr
0.25	0.50
0.50	0.40
0.75	0.38



Interpolated Precision-Recall

Processamento
e Recuperação
de Informação

Evaluation
and Relevance

Precision vs.
Recall

Other
Measures

Obtaining the
Ground Truth

Evaluation of
Classifiers

- Precision is usually measured at 10 standard recall points: 0%, 10%, 20%, ..., 90%, 100%

- Precision at r^0 recall is defined as

$$P(r) = \max_{i \geq r} P(i)$$

- Precision is zero after no more relevant documents are found



Interpolated Precision-Recall (cont.)

Processamento
e Recuperação
de Informação

Evaluation
and Relevance

Precision vs.
Recall

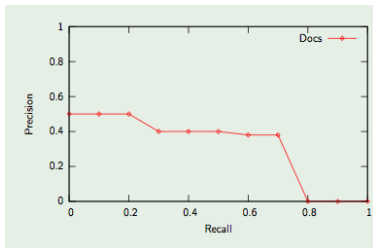
Other
Measures

Obtaining the
Ground Truth

Evaluation of
Classifiers

Let $A = \{d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8, d_9, d_{10}\}$ be an ordered set of retrieved documents, for a query Q . Let $R = \{d_2, d_5, d_8, d_{15}\}$ be the set of relevant documents for query Q .

Re	Pr
0.00	0.50
0.10	0.50
0.20	0.50
0.30	0.40
0.40	0.40
0.50	0.40
0.60	0.38
0.70	0.38
0.80	0.00
0.90	0.00
1.00	0.00





Interpolated Precision-Recall (cont.)

Processamento
e Recuperação
de Informação

Evaluation
and Relevance

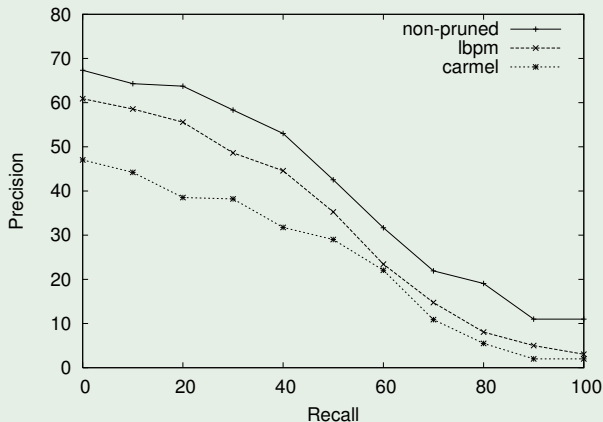
Precision vs.
Recall

Other
Measures

Obtaining the
Ground Truth

Evaluation of
Classifiers

Example





Outline

Processamento
e Recuperação
de Informação

Evaluation
and Relevance

Precision vs.
Recall

Other
Measures

Obtaining the
Ground Truth

Evaluation of
Classifiers

1 Evaluation and Relevance

2 Precision vs. Recall

3 Other Measures

4 Obtaining the Ground Truth

5 Evaluation of Classifiers



P@N, R-precision

Processamento
e Recuperação
de Informação

Evaluation
and Relevance

Precision vs.
Recall

Other
Measures

Obtaining the
Ground Truth

Evaluation of
Classifiers

P@N – Precision at the N -th retrieved document

Most commonly used

- $P@5$,
- $P@10$
- $P@20$

Usefull for Web retrieval

R-precision - Precision at the R -th document, where R is the number of relevant documents



F-measure

Processamento
e Recuperação
de Informação

Evaluation
and Relevance

Precision vs.
Recall

Other
Measures

Obtaining the
Ground Truth

Evaluation of
Classifiers

Harmonic mean of precision and recall:

$$F_1 = \frac{2 \times Re \times Pr}{Re + Pr}$$



AP, MAP

Processamento
e Recuperação
de Informação

Evaluation
and Relevance

Precision vs.
Recall

Other
Measures

Obtaining the
Ground Truth

Evaluation of
Classifiers

- **AP** - Average of the values for the precision at each recall point

$$AP = \frac{\sum_{i=1}^N Pr@i \times R_i}{|R|}$$

where $R_i = 1$ if document at rank i is relevant and $R_i = 0$ otherwise.

- **MAP** - Mean Average Precision

$$MAP = \frac{\sum_{q=1}^Q AP_q}{Q}$$

- AP can also be interpolated



Discounted Cumulative Gain

Processamento
e Recuperação
de Informação

Evaluation
and Relevance

Precision vs.
Recall

Other
Measures

Obtaining the
Ground Truth

Evaluation of
Classifiers

Cumulative gain: sum the relevance weights

- **DCG** - Discounted cumulative gain

$$\text{DCG}_p = R_1 + \sum_{i=2}^p \frac{R_i}{\log_2 i}$$

where $R_i = 1$ if document at rank i is relevant and $R_i = 0$ otherwise.

- **nDCG** - Normalized discounted cumulative gain

$$\text{nDCG}_p = \frac{\text{DCG}_p}{\text{Ideal DCG}_p}$$



MRR

Processamento
e Recuperação
de Informação

Evaluation
and Relevance

Precision vs.
Recall

Other
Measures

Obtaining the
Ground Truth

Evaluation of
Classifiers

MRR - Mean Reciprocal Rank

$$\text{MRR} = \frac{1}{N} \sum_{i=1}^N \frac{1}{\text{rank}_i}$$

where rank_1 is the rank of the first relevant document.



ERR

Processamento
e Recuperação
de Informação

Evaluation
and Relevance

Precision vs.
Recall

Other
Measures

Obtaining the
Ground Truth

Evaluation of
Classifiers

ERR - Expected Reciprocal Rank

Makes use of the cascade model as a user browsing model.

$$\begin{aligned} \text{ERR} &= \sum_{i=1}^N \frac{1}{N} \times P(\text{user stops at position } i) \\ &= \sum_{i=1}^N \frac{1}{N} \times \prod_{j=1}^{N-1} (1 - R_j) R_N \end{aligned}$$

where $R_i = 1$ if document at rank i is relevant and $R_i = 0$, or instead the result of mapping from relevance grades to probability of relevance.



Ranking Comparison

Processamento
e Recuperação
de Informação

Evaluation
and Relevance

Precision vs.
Recall

Other
Measures

Obtaining the
Ground Truth

Evaluation of
Classifiers

Spearman Coefficient

Computes the difference between the positions of a same document in two rankings

$$\rho(X, Y) = 1 - \frac{6 \sum_{i=1}^N d_i^2}{N(N^2 - 1)}$$

where $d_i = \text{rank}(X)_i - \text{rank}(Y)_i$ is the difference in rankings of document i .



Ranking Comparison (cont.)

Processamento
e Recuperação
de Informação

Evaluation
and Relevance

Precision vs.
Recall

Other
Measures

Obtaining the
Ground Truth

Evaluation of
Classifiers

Kendall's Tau

Let $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$, where each x_i is the rank of document i in ranking X , and y_i is the rank of document i in ranking Y .

$$\tau = \frac{|\text{concordant pairs}| - |\text{discordant pairs}|}{N(N-1)/2}$$

where a pair (x_i, y_i) is concordant with (x_j, y_j) if either:

$$\begin{cases} x_i > x_j \wedge y_i > y_j \\ x_i < x_j \wedge y_i < y_j \end{cases}$$

and discordant if either:

$$\begin{cases} x_i > x_j \wedge y_i < y_j \\ x_i < x_j \wedge y_i > y_j \end{cases}$$



Outline

Processamento
e Recuperação
de Informação

Evaluation
and Relevance

Precision vs.
Recall

Other
Measures

Obtaining the
Ground Truth

Evaluation of
Classifiers

- 1 Evaluation and Relevance
- 2 Precision vs. Recall
- 3 Other Measures
- 4 Obtaining the Ground Truth**
- 5 Evaluation of Classifiers



Reference Collections

Processamento
e Recuperação
de Informação

Evaluation
and Relevance

Precision vs.
Recall

Other
Measures

Obtaining the
Ground Truth

Evaluation of
Classifiers

TREC Various collections of documents (Ad hoc, Web, Blog, Clinical Decision Support, ...)

CACM Articles from Communications of the ACM

ISI Information science papers

CFC Cystic Fibrosis Collection

...

- Standards for research in IR
- Provide sets queries + evaluated documents



Human Experimentation in the Lab

Processamento
e Recuperação
de Informação

Evaluation
and Relevance

Precision vs.
Recall

Other
Measures

Obtaining the
Ground Truth

Evaluation of
Classifiers

- User preferences are affected by the characteristics of the user interface (UI)
 - For instance, the users of search engines look first at the upper left corner of the results page.
 - Changing the layout is likely to affect the assessment made by the users and their behavior.
- Proper evaluation of the user interface requires going beyond the framework of the Cranfield experiments



A/B Testing

Processamento
e Recuperação
de Informação

Evaluation
and Relevance

Precision vs.
Recall

Other
Measures

Obtaining the
Ground Truth

Evaluation of
Classifiers

- A/B testing consists of displaying to selected users a modification in the layout of a page
 - The group of selected users constitute a fraction of all users such as, for instance, 1%
 - The method works well for sites with large audiences
- By analysing how the users react to the change, it is possible to analyse if the modification proposed is positive or not

A/B testing provides a form of human experimentation, even if the setting is not that of a lab

Amazon Mechanical Turk

Make Money by working on HITs

HITs - *Human Intelligence Tasks* - are individual tasks that you work on. [Find HITs now.](#)

As a Mechanical Turk Worker you:

- Can work from home
- Choose your own work hours
- Get paid for doing good work



Get Results from Mechanical Turk Workers

Ask workers to complete HITs - *Human Intelligence Tasks* - and get results using Mechanical Turk. [Get started.](#)

As a Mechanical Turk Requester you:

- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results



<https://www.mturk.com>

- The participants execute human intelligence tasks, called HITs, in exchange for small sums of money
- The tasks are filed by requesters who have an evaluation need
- While the identity of participants is not known to requesters, the service produces evaluation results of high quality (except for **free-loaders**, etc)



Evaluation using Clickthrough Data

Processamento
e Recuperação
de Informação

Evaluation
and Relevance

Precision vs.
Recall

Other
Measures

Obtaining the
Ground Truth

Evaluation of
Classifiers

A promising alternative...

The data can be obtained by observing how frequently the users click on a given document, when it is shown in the answer set for a given query

Attractive, because...

The data can be collected at a low cost without overhead for the use



Outline

Processamento
e Recuperação
de Informação

Evaluation
and Relevance

Precision vs.
Recall

Other
Measures

Obtaining the
Ground Truth

Evaluation of
Classifiers

- 1 Evaluation and Relevance
- 2 Precision vs. Recall
- 3 Other Measures
- 4 Obtaining the Ground Truth
- 5 Evaluation of Classifiers



Classifier Evaluation

Processamento
e Recuperação
de Informação

Evaluation
and Relevance

Precision vs.
Recall

Other
Measures

Obtaining the
Ground Truth

Evaluation of
Classifiers

- Previous lectures have shown that tasks such as document classification or information extraction from text can be modeled as classification problems
 - I.e., techniques in this section also apply to IE systems
- Goal in supervised classification is the minimization of classification error on test data
- We can evaluate through measures like recall, precision, and accuracy (i.e., one minus error)
 - But classification tasks can involve more than two classes (i.e., more than distinguishing relevant from non-relevant)



Confusion Matrix

Processamento
e Recuperação
de Informação

Evaluation
and Relevance

Precision vs.
Recall

Other
Measures

Obtaining the
Ground Truth

Evaluation of
Classifiers

- $M[i, j]$ is the number of test documents belonging to class i which were assigned to class j
- Perfect classifier: diagonal elements $M[i, i]$ would be nonzero
- Example:

$$M = \left\{ \begin{array}{c|c|c} 5 & 0 & 0 \\ \hline 1 & 3 & 0 \\ \hline 1 & 2 & 4 \end{array} \right\}$$

- If M is large, we use

$$\text{accuracy} = \sum_i M[i, i] / \sum_{i,j} M[i, j]$$

- Notice that accuracy is not a good measure for *small* classes



Micro-Averaged Precision

Processamento
e Recuperação
de Informação

Evaluation
and Relevance

Precision vs.
Recall

Other
Measures

Obtaining the
Ground Truth

Evaluation of
Classifiers

In a problem with n classes, let C_i be the number of documents in class i and let C'_i be the number of documents estimated to be of class i by the classifier

- **Micro-averaged precision** is defined as

$$\frac{\sum_{i=1}^n C'_i \cap C_i}{\sum_{i=1}^n C'_i}$$

- **Micro-averaged recall** is defined as

$$\frac{\sum_{i=1}^n C'_i \cap C_i}{\sum_{i=1}^n C_i}$$

- Micro-averaged precision/recall measures correctly classified documents, thus favoring large classes



Macro-Averaged Precision

Processamento
e Recuperação
de Informação

Evaluation
and Relevance

Precision vs.
Recall

Other
Measures

Obtaining the
Ground Truth

Evaluation of
Classifiers

In a problem with n classes, let P_i and R_i be the precision and recall, respectively, achieved by a classifier for class i

- **Macro-averaged precision** is defined as

$$\frac{1}{n} \sum_{i=1}^n P_i$$

- **Macro-averaged recall** is defined as

$$\frac{1}{n} \sum_{i=1}^n R_i$$

- Macro-averaged precision/recall measures performance per class, giving all classes equal importance



F_1 measure

Processamento
e Recuperação
de Informação

Evaluation
and Relevance

Precision vs.
Recall

Other
Measures

Obtaining the
Ground Truth

Evaluation of
Classifiers

The F_1 measure is also commonly used

$$F_1 = \frac{2 \times P_i \times R_i}{P_i + R_i}$$

- Harmonic mean between precision and recall
- Discourages classifiers that trade one for the other



Processamento
e Recuperação
de Informação

Evaluation
and Relevance

Precision vs.
Recall

Other
Measures

Obtaining the
Ground Truth

Evaluation of
Classifiers

Questions?