Processamento e Recuperação de Informação

Evaluation and Relevance

Precision vs. Recall

Other Measures

Ranking Comparison

Obtaining the Ground Truth

Evaluation of Classifiers

Evaluation of Clustering

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

Ranking Comparison

Obtaining the Ground Truth

Evaluation of Classifiers

Evaluation of Clustering

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 edtion. Chapter 4.

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

Outline

1. Evaluation and Relevance
2. Precision vs. Recall
3. Other Measures
4. Ranking Comparison
5. Obtaining the Ground Truth
6. Evaluation of Classifiers
7. Evaluation of Clustering

IR System Evaluation

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

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

Ranking Comparison

Obtaining the Ground Truth

Evaluation of Classifiers

Evaluation of Clustering

Recall and Precision

Measure the ability of a system to return relevant documents.

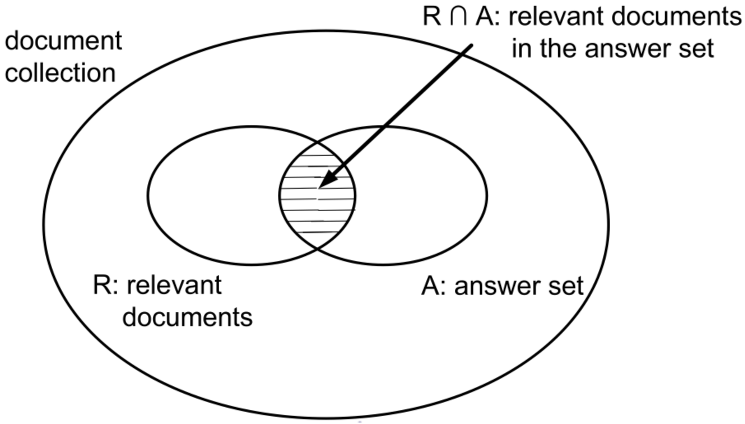
Relevance

- Subjective notion
- Usually evaluated by a set of experts

Outline

1. Evaluation and Relevance
2. Precision vs. Recall
3. Other Measures
4. Ranking Comparison
5. Obtaining the Ground Truth
6. Evaluation of Classifiers
7. Evaluation of Clustering

Evaluating Prediction



Measuring Precision and Recall

Processamento e Recuperação de Informação

Evaluation and Relevance

Precision vs. Recall

Other Measures

Ranking Comparison

Obtaining the Ground Truth

Evaluation of Classifiers

Evaluation of Clustering

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

- 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

Ranking Comparison

Obtaining the Ground Truth

Evaluation of Classifiers

Evaluation of Clustering

- Precision is usually measured at 10 standard recall points: $0\%, 10\%, 20\%, \ldots, 90\%, 100\%$
- Precision at $r\%$ recall is defined as

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

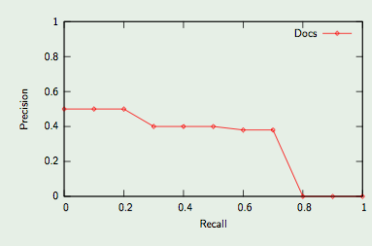
- Precision is zero after no more relevant documents are found

Interpolated Precision-Recall (cont.)

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.

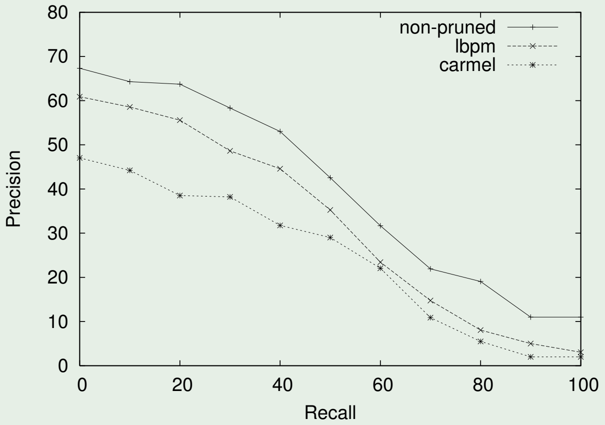
Re	Pr
0.25	0.50
0.50	0.40
0.75	0.38

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.)

Example



Outline

1. Evaluation and Relevance
2. Precision vs. Recall
3. Other Measures
4. Ranking Comparison
5. Obtaining the Ground Truth
6. Evaluation of Classifiers
7. Evaluation of Clustering

P@N, R-precision

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

Harmonic mean of precision and recall:

$$F_\beta = \frac{(1+\beta^2) \times Pr \times Re}{(\beta^2 \times Pr) + Re}$$

Ususally we adopt F_1:

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

AP, MAP

- **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

$$\mathrm{MAP} = \frac{\sum_{q=1}^{Q} \mathrm{AP}_q}{Q}$$

- AP can also be interpolated

Discounted Cumulative Gain

Cumulative gain: sum the relevance weights

- **DCG** - Discounted cumulative gain

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

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

- **nDCG** - Normalized discounted cumulative gain

$$\mathrm{nDCG}_p = \frac{\mathrm{DCG_p}}{\mathrm{Ideal\,DCG}_p}$$

Processamento e Recuperação de Informação

Evaluation and Relevance

Precision vs. Recall

Other Measures

Ranking Comparison

Obtaining the Ground Truth

Evaluation of Classifiers

Evaluation of Clustering

MRR

MRR - Mean Reciprocal Rank

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

where rank_1 is the rank of the first relevant document.

User Models

Processamento e Recuperação de Informação

Evaluation and Relevance

Precision vs. Recall

Other Measures

Ranking Comparison

Obtaining the Ground Truth

Evaluation of Classifiers

Evaluation of Clustering

- Position models: assume independence among documents in different positions and model the examination probability as a function of the position
- Cascade Models: consider the dependency among URLs on a search results page — at each position, the user has a certain probability of being satisfied depending on the relevance of the previous documents

Previous measures assumed a position model; following we show *ERR*, which assumes a cascade model.

ERR

ERR - Expected Reciprocal Rank

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

where $R_i = 1$ if document at rank i is relevant and $R_i = 0$ otherwise.
R_i can also be the result of mapping from relevance grades to probability of relevance $R_i := \mathcal{R}(g_i)$, where:

$$\mathcal{R}(g) = \frac{2^g - 1}{2^{g_{\max}}}, \; g \in \{0, \cdots, g_{\max}\}$$

Outline

1 Evaluation and Relevance

2 Precision vs. Recall

3 Other Measures

4 Ranking Comparison

5 Obtaining the Ground Truth

6 Evaluation of Classifiers

7 Evaluation of Clustering

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.

Kendall's Tau

Let $(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)$, 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

1. Evaluation and Relevance
2. Precision vs. Recall
3. Other Measures
4. Ranking Comparison
5. Obtaining the Ground Truth
6. Evaluation of Classifiers
7. Evaluation of Clustering

Reference Collections

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

Ranking Comparison

Obtaining the Ground Truth

Evaluation of Classifiers

Evaluation of Clustering

- 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

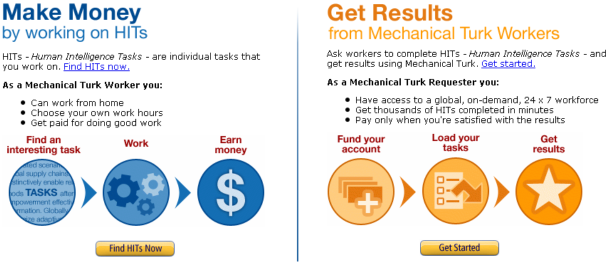
A/B Testing

- 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

Crowdsourcing

Amazon Mechanical Turk



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

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 users

Evaluation using Clickthrough Data (2)

Processamento e Recuperação de Informação

Evaluation and Relevance

Precision vs. Recall

Other Measures

Ranking Comparison

Obtaining the Ground Truth

Evaluation of Classifiers

Evaluation of Clustering

Click models

An accurate user model, which closely reflects users' interactions with the retrieval system, is essential for developing a good relevance metric from clickthrough data.

Example: Cascade model used in ERR metric, corresponding to

$$\Pi_{j=1}^{i-1}(1 - R_j)R_i$$

where the values R_i (i.e., document satisfies the user with probability R_i) can be estimated by maximum likelihood on the click logs.

Outline

1 Evaluation and Relevance

2 Precision vs. Recall

3 Other Measures

4 Ranking Comparison

5 Obtaining the Ground Truth

6 Evaluation of Classifiers

7 Evaluation of Clustering

Classifier Evaluation

- 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

- $M[i,j]$ 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

Ranking Comparison

Obtaining the Ground Truth

Evaluation of Classifiers

Evaluation of Clustering

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

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

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

- Macro-averaged recall is defined as

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

- Macro-averaged precision/recall measures performance per class, giving all classes equal importance
- The F_1 measure is also commonly used

Multi-Label Scenario

- Quality can be measured by per-instance recall and precision
  - Let C_d be the correct classes for document d and C'_d be the set of classes estimated by the classifier

$$precision = \frac{C'_d \cap C_d}{C'_d}$$

$$recall = \frac{C'_d \cap C_d}{C_d}$$

Train-Test Split

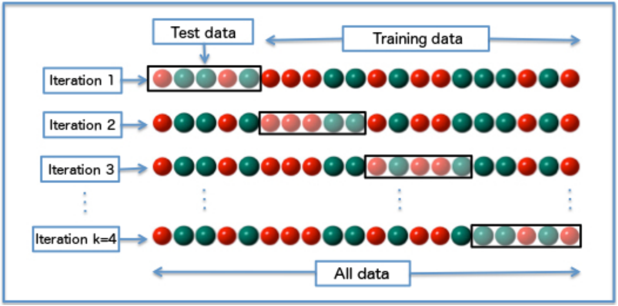
- When evaluatig a classifier, you cannot rely on the data used for training
  - You estimate is likely to be overly optimistic
  - Your model will tend to overfit
- Data must be split into a test and training sets
  - Common train/test splits: 80%/20% or 70%/30%

Cross-Validation

- Splitting the dataset and testing once may lead to a biased evaluation
- One way to avoid this is to use cross-validation
  - Leave-p-out
  - Leave-one-out
  - k-fold
  - ...

K-Fold Cross-Validation

1. Split the data into k partitions
2. For each fold $i \in [1, k]$
    1. Train your model using all partitions P_j, $j \neq i$
    2. Evaluate your model in partition P_i
3. Average your evaluation metrics overall all folds



(source of image: wikipedia)

Outline

1. Evaluation and Relevance
2. Precision vs. Recall
3. Other Measures
4. Ranking Comparison
5. Obtaining the Ground Truth
6. Evaluation of Classifiers
7. Evaluation of Clustering

What to evaluate

- Goal of clustering: attain high intra-cluster similarity (documents within a cluster are similar) and low inter-cluster similarity (documents from different clusters are dissimilar
- These are internal criteria
- Alternatively, we can evaluate the results of the application of interest
  - E.g. for search result clustering, evaluate search results
- Or we can use external criteria, comparing the clusters to a gold standard

External Criteria

- Clusters are evaluated using a gold standard (as in classification problems)
  - Each document will be assigned a class
- Unlike classification problems, we don't know to which class each cluster corresponds
  - Thus we cannot directly computed false positives, false negatives, etc.
- There are many proposals to address this issue

Purity

For a set $\Omega = \{w_1, w_2, \cdots, w_K\}$ of clusters and a set $\mathcal{C} = \{c_1, c_2, \cdots, c_K\}$ of classes, purity is defined as:

$$\text{purity}(\Omega, \mathcal{C}) = \frac{1}{N} \sum_k \max_j |w_k \cap c_j|$$

where N is the number of documents, w_k is the set of documents in cluster w_k, and c_k is the set of documents in class c_k.

Each cluster is assigned to the class most frequent in the cluster and accuracy is measured.

Normalized Mutual Information

Processamento e Recuperação de Informação

Evaluation and Relevance

Precision vs. Recall

Other Measures

Ranking Comparison

Obtaining the Ground Truth

Evaluation of Classifiers

Evaluation of Clustering

- Purity does not show the tradeoff between quality and the number f clusters
  - E.g., if each document gets a cluster, purity $= 1$
- NMI takes this tradeoff into account
  - Measures the increase in the amount of information when we are know what the clusters are,
  - But normalizes it by the entropy of the clusters

Normalized Mutual Information (cont.)

$$\text{NMI}(\Omega, \mathcal{C}) = \frac{I(\Omega, \mathcal{C})}{(H(\Omega) + H(\mathcal{C}))/2}$$

where

$$\begin{aligned} I(\Omega, \mathcal{C}) &= \sum_k \sum_j P(w_k \cap c_j) \log \frac{P(w_k \cap c_j)}{P(w_k)P(c_j)} \\ &= \sum_k \sum_j \frac{|w_k \cap c_j|}{N} \log \frac{N|w_k \cap c_j|}{|w_k||c_j|} \end{aligned}$$

is the mutual information between clusters and classes and

$$H(\Omega) = -\sum_k P(w_k) \log P(w_k) = -\sum_k \frac{|w_k|}{N} \log \frac{|w_k|}{N}$$

is the entropy (equally for $H(\mathcal{C})$)

Rand Index

- The Rand index views clustering as a series of decisions for each of the $N(N-1)/2$ pairs of documents
- Decisions can be:
  - **True Positive**: the documents are similar and in the same cluster
  - **True Negative**: the documents are *not* similar and in different clusters
  - **False Positive**: the documents are *not* similar but in the same cluster
  - **False Negative**: the documents are similar but in different clusters
- We can considere documentas as similar if they are in the same class

Rand Index (cont.)

The Rand index simply measures the percentage of correct decisions (i.e. the accuracy):

$$RI = \frac{TP + TN}{TP + FP + TN + FN}$$

The F_β Measure (Again)

- The Rand index gives equal weight to false positives and false negatives
- Separating similar documents is sometimes worse than putting pairs of dissimilar documents in the same cluster.
- We can use the F measure to penalize false negatives more strongly than false positives, giving more weight to recall
  - selecting a value $\beta > 1$

The F_β Measure (cont.)

$$P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN}$$

$$F_\beta = \frac{(1+\beta^2)\cdot P \cdot R}{(\beta^2 \cdot P) + R}$$

or, using the decision framework:

$$F_\beta = \frac{(1+\beta^2)\cdot TP}{(1+\beta^2)\cdot TP + \beta^2 \cdot FN + FP}$$

Questions?