CAIM: Cerca i Anàlisi d’Informació Massiva
FIB, Grau en Enginyeria Informàtica

Slides by Marta Arias, José Balcázar, Ricard Gavaldá
Department of Computer Science, UPC

Fall 2016
http://www.cs.upc.edu/~caim
4. Evaluation and Relevance Feedback
Evaluation of Information Retrieval Usage, I

What are we exactly to do?

In the Boolean model, the specification is unambiguous:

We know what we are to do:

- Retrieve and provide to the user
- all those documents
- that satisfy the query.

But, is this what the user really wants?

Sorry, but usually... no.
Then, what exactly are we to optimize?

Notation:

\( \mathbb{D} \): set of all our documents on which the user asks one query;

\( \mathbb{A} \): answer set: documents that the system retrieves as answer;

\( \mathbb{R} \): relevant documents: those that the user actually wishes to see as answer.

(But no one knows this set, not even the user!)

Unreachable goal: \( \mathbb{A} = \mathbb{R} \), that is:

1. \( Pr(d \in \mathbb{A} | d \in \mathbb{R}) = 1 \) and
2. \( Pr(d \in \mathbb{R} | d \in \mathbb{A}) = 1. \)
The Recall and Precision measures

Let’s settle for:

- high recall, \( \frac{|R \cap A|}{|R|} \) :

  \[ Pr(d \in A|d \in R) \text{ not too much below } 1, \]

- high precision, \( \frac{|R \cap A|}{|A|} \) :

  \[ Pr(d \in R|d \in A) \text{ not too much below } 1. \]

Difficult balance. More later.
Recall and Precision, II
Example: test for tuberculosis (TB)

- 1000 people, out of which 50 have TB
- test is positive on 40 people, of which 35 **really** have TB

**Recall**
% of true TB that test positive = \(\frac{35}{50} = 70\%\)

**Precision**
% of positives that really have TB = \(\frac{35}{40} = 87.5\%\)

- **Large recall**: few sick people go away undetected
- **Large precision**: few people are scared unnecessarily (few *false alarms*)
Recall and Precision, III. Confusion matrix

Equivalent definition

Confusion matrix

<table>
<thead>
<tr>
<th>Reality</th>
<th>Answered</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>relevant</td>
<td>relevant</td>
<td>not relevant</td>
</tr>
<tr>
<td>relevant</td>
<td>$tp$</td>
<td>$fn$</td>
<td></td>
</tr>
<tr>
<td>not relevant</td>
<td>$fp$</td>
<td>$tn$</td>
<td></td>
</tr>
</tbody>
</table>

- $|\mathcal{R}| = tp + fn$
- $|\mathcal{A}| = tp + fp$
- $|\mathcal{R} \cap \mathcal{A}| = tp$
- **Recall** = $\frac{|\mathcal{R} \cap \mathcal{A}|}{|\mathcal{R}|} = \frac{tp}{tp+fn}$
- **Precision** = $\frac{|\mathcal{R} \cap \mathcal{A}|}{|\mathcal{A}|} = \frac{tp}{tp+fp}$
How many documents to show?

We rank all documents according to some measure. How many should we show?

- Users won’t read too large answers.
- Long answers are likely to exhibit low precision.
- Short answers are likely to exhibit low recall.

We analyze precision and recall as functions of the number of documents $k$ provided as answer.
Rank-recall and rank-precision plots

(Source: Prof. J. J. Paijmans, Tilburg)
A single “precision and recall” curve

\[x\text{-axis for recall, and } y\text{-axis for precision.} \]
(Similar to, and related to, the ROC curve in predictive models.)

(Source: Stanford NLP group)

Often: Plot 11 points of interpolated precision, at 0 %, 10 %, 20 %, \ldots, 100 % recall
Other measures of effectiveness

- **AUC:** Area under the curve of the plots above, relative to best possible

- **F-measure:** \[ \frac{2}{\text{recall} + \frac{1}{\text{precision}}} \]
  - Harmonic mean. Closer to min of both than arithmetic mean

- **\(\alpha\)-F-measure:** \[ \frac{2}{\alpha \text{recall} + \frac{1-\alpha}{\text{precision}}} \]
Other measures of effectiveness, II

Take into account *the documents previously known to the user.*

- **Coverage:**
  \[
  \frac{|\text{relevant & known & retrieved}|}{|\text{relevant & known}|}
  \]

- **Novelty:**
  \[
  \frac{|\text{relevant & retrieved & UNknown}|}{|\text{relevant & retrieved}|}
  \]
Relevance Feedback, I
Going beyond what the user asked for

The user relevance cycle:

1. Get a query $q$
2. Retrieve relevant documents for $q$
3. Show top $k$ to user
4. Ask user to mark them as relevant / irrelevant
5. Use answers to refine $q$
6. If desired, go to 2
Relevance Feedback, II

How to create the new query?

Vector model: queries and documents are vectors
Given a query $q$, and a set of documents, split into relevant $R$ and nonrelevant $NR$ sets, build a new query $q'$:

Rocchio's Rule:

$$q' = \alpha \cdot q + \beta \cdot \frac{1}{|R|} \cdot \sum_{d \in R} d - \gamma \cdot \frac{1}{|NR|} \cdot \sum_{d \in NR} d$$

- All vectors $q$ and $d$'s must be normalized (e.g., unit length).
- Weights $\alpha$, $\beta$, $\gamma$, scalars, with $\alpha > \beta > \gamma \geq 0$; often $\gamma = 0$.
  - $\alpha$: degree of trust on the original user’s query,
  - $\beta$: weight of positive information (terms that do not appear on the query but do appear in relevant documents),
  - $\gamma$: weight of negative information.
Relevance Feedback, III

In practice, often:

- good improvement of the recall for first round,
- marginal for second round,
- almost none beyond.

In web search, precision matters much more than recall, so the extra computation time and user patience may not be productive.
It is a form of **Query Expansion**:

The new query has non-zero weights on words that were not in the original query.
Pseudorelevance feedback

Do not ask anything from the user!

- User patience is precious resource. They’ll just walk away.
- Assume you did great in answering the query!
- That is, top-$k$ documents in the answer are all relevant
- No interaction with user
- But don’t forget that the search will feel slower.
- Stop, at the latest, when you get the same top $k$ documents.
Alternative sources of feedback / query refinement:

- Links clicked / not clicked on.
- Think time / time spent looking at item.
- User’s previous history.
- Other users’ preferences!
- Co-occurring words: Add words that often occur with words in the query - for query expansion.