Evaluation von Suchlösungen

::: MASTERSEMINAR SUCHMASCHINEN, CIS, SOMMERSEMESTER 2025 :::

Überblick

- Evaluation - Ziele
- Recall, Precision, F-Measure
- Mean average precision, precision at n
- DCG, NDCG

Suchmaschinen - Evaluierung

- In einer Suchlösung gibt es zahlreiche Parameter, u.a.:
  - - Welches Suchmaschine wird verwendet
  - - Vektorbasierte Suche versus Keyword-Suche
  - - Ranking-Formel
  - - Welche Embeddings werden verwendet
  - - Dokumentengewicht / Gewichtung einer Quelle
  - - Gewichtung von Dokumentabschnitten
  - - Queryvariationen (Proximity, Phrasensuche, Termgewichtung...)
  - - Stemming / Tokenisierung / Synonyme / Rechtschreibkorrektur

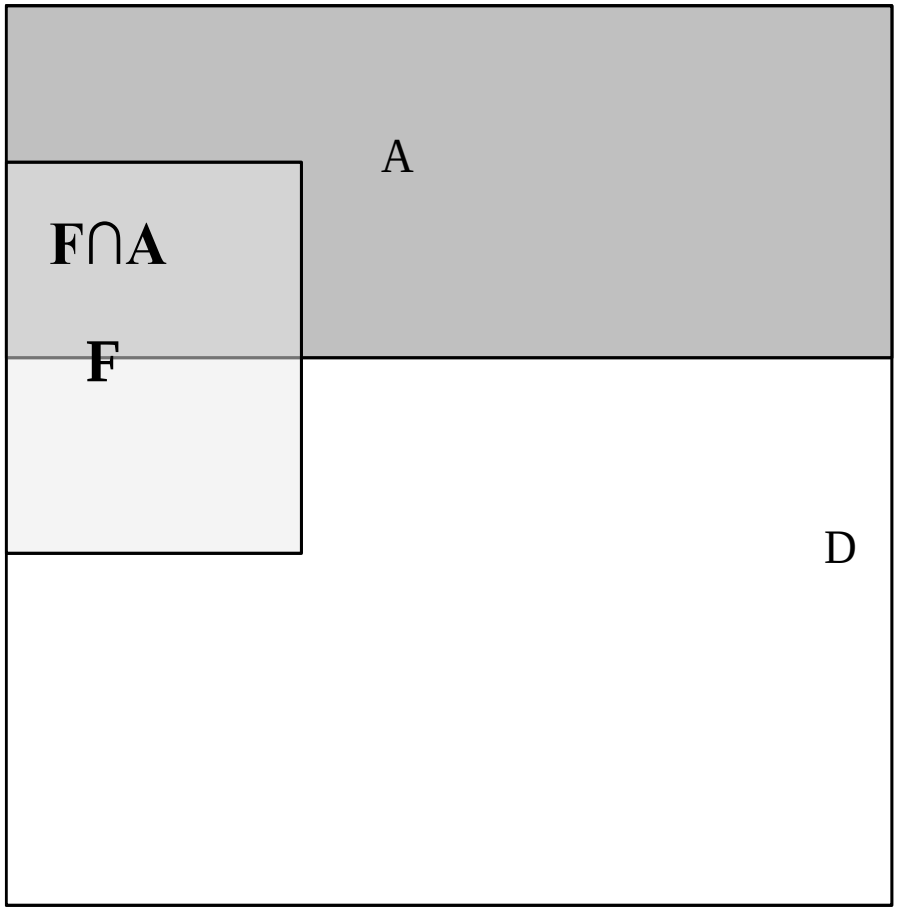
Um die verschiedenen Parameter einzustellen, müssen Evaluationskriterien definiert werden.

Suchmaschinenevaluierung - Daten

Evaluierung auf Basis von Referenzdaten

- Eine Menge von Anfragen aus der Anwendungsdomäne
  - Anfragen, die potentielle Benutzer*innen stellen könnten
  - Anfragen, die schon gestellt wurden (query logs)
    - Möglichst Abdeckung aller relevanten Bereiche
    - Einfache und komplexere Queries
- Ein Set von Dokumenten(-chunks) für jede Anfrage mit einem Relevanzmaß
  - Im einfachsten Fall: Passt/passt nicht (binär)
  - Besser: Geordnet nach Relevanz

Trefferquote (Recall) und Genauigkeit (Precision)



- Maß für die Qualität des Retrievals
- D: Alle Dokumente
- A: Relevante Dokument
- F: Gefundene Dokumente
- Recall = F ∩ A/A
- Precision F ∩ A/F

F-measure (F1 score)

F1 measure: $f = \frac{2\ p*r}{p+r}$

Hier ist p: Precision und r: Recall

(Harmonisches Mittel zwischen Precision und Recall)

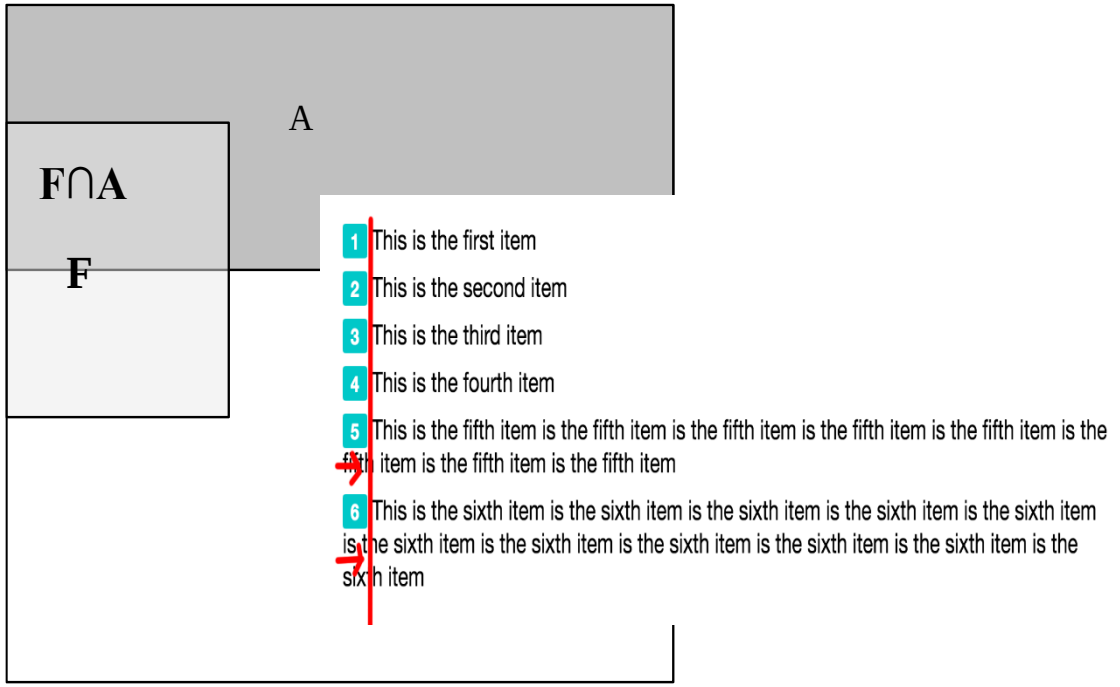
Generalisiert:

: $f_{ß} = \frac{(1+ß^2)\ p*r}{(ß^2 p)+r}$

Hier wird Recall (r) ß-mal so stark gewichtet wie Precision

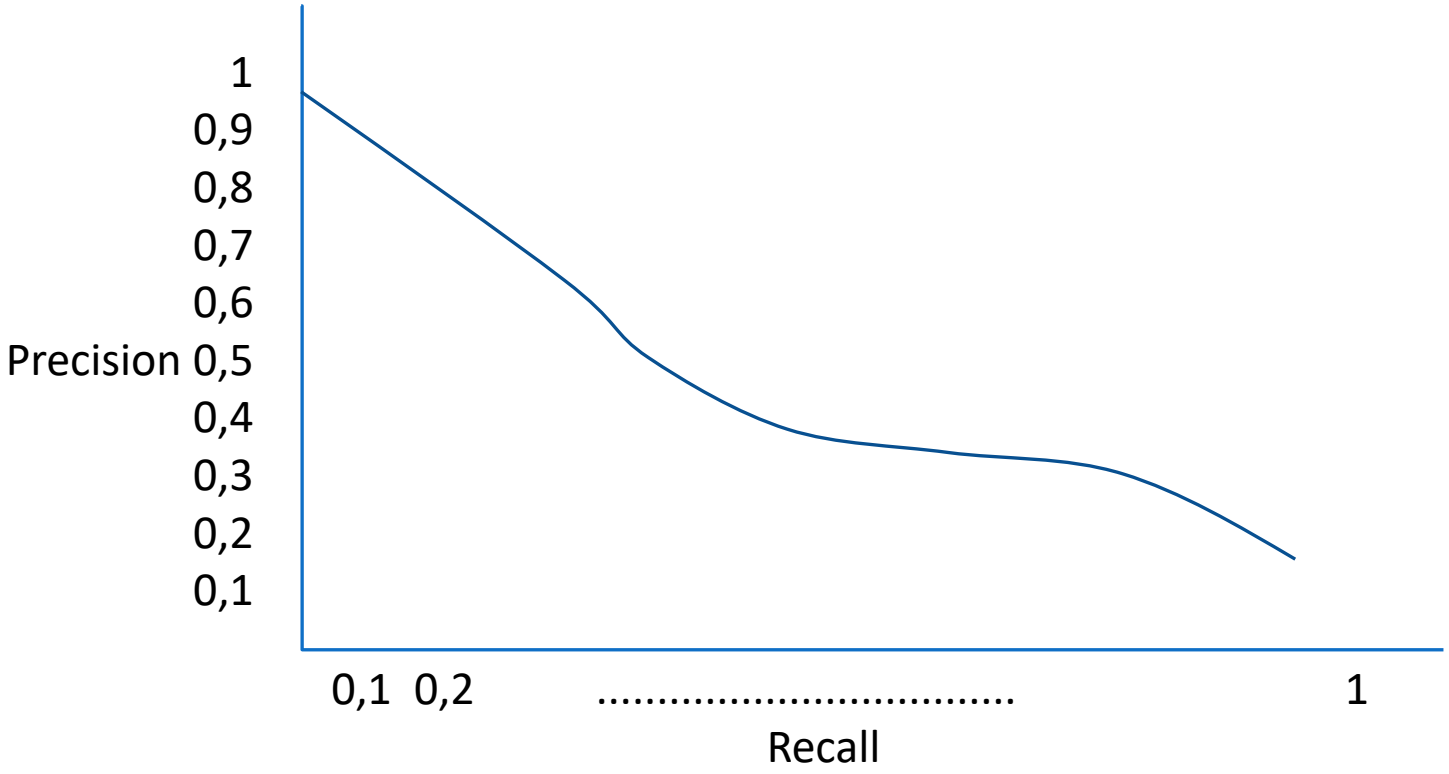
Probleme mit den bisher genannten Maßen

- Reihenfolge / Ranking spielt keine Rolle
  - In realistischen Szenarien werden nur die ersten Dokumente von den Benutzer*innen verarbeitet / angeschaut
- Schwer anwendbar auf größere Ergebnismengen
  - Relevante Dokumente müssen bekannt sein (für gesamte Ergebnismenge) um den Recall zu ermitteln
    - Dies ist in der Regel für größere Dokumentenmengen nicht der Fall



Average precision

- Average precision for all values of recall r=0 r <= 1



Mean average precision

- Das Mittel der Average Precision über mehrere Queries

Die Maße Average Precision und Mean Average Precision:

- berücksichtigen das Ranking
- können nur berechnet werden, wenn der Recall feststeht

Precision bei k

- Anteil der relevanten Dokumente unter den ersten k Dokumenten

- Kann manuell kontrolliert werden auch bei größeren Dokumentenmengen

- Unabhängig vom Recall, kann daher ohne Kenntnis der gesamten Dokumentenmenge berechnet werden

- Beurteilung der Relevanz notwendig

Aber:

- keine Abhängigkeit von der Abfolge innerhalb der ersten k Dokumente

Gesucht – Maße für weitere Anwendungsfälle

Gegeben:

- Große Dokumentenmenge

- Eine Liste von Anfragen (Queries)
  - Für jede Query sind einige sehr gute Dokumente bekannt und auch die gewünschte Reihenfolge der Dokumente, idealerweise die relevantesten Dokumente

→ Realistisches Szenario für Evaluierung von Websuche, Intranetsuche (generell bei Suche über große Dokumentenmengen)

•**Evaluate your Recommendation Engine using NDCG - Towards ...**
https://towardsdatascience.com/evaluate-your-recommendation-engine...
Evaluate your Recommendation Engine using **NDCG**. How to best
Author: Pranay Chandekar
•**NDCG - What does NDCG stand for? The Free Dictionary**
https://acronyms.thefreedictionary.com/NDCG
Looking for online definition of **NDCG** or what **NDCG** stands
Evaluation measures (information retrieval) - Wikipedia
https://en.wikipedia.org/wiki/Evaluation_measures_(information_retrieval)
The **nDCG** values for all queries can be averaged to obtain a measure of the average performance of a ranking algorithm. Note.
•**A Theoretical Analysis of NDCG Ranking Measures**
proceedings.mlr.press/v30/Wang13.pdf · PDF file
NDCG has two advantages compared to many other measures. **Page Count:** 30
•**File Size:** 612KB
•**Author:** Yining Wang, Liwei Wang, Yuanzhi Li, Di He, Tie-Yan Liu
•**National Design & Craft Gallery, Kilkenny, Ireland - NDCG**
https://www.ndcg.ie
News 19 May ZOOM Seminar - Generation: Familial Legacies. 20

Mittlerer gegenseitiger Rang

(en: mean reciprocal rank)

- Gegeben:
  - Eine Menge von Queries mit relevanten Dokumenten

- Verfahren
  - Ermittle die Position p des ersten (bekannten) relevanten Dokuments in der Ergebnisliste für jede Query
  - Der Rangwert ist 1/p

- Der mittlere gegenseitige Rang ist der Mittelwert für den Satz von Queries

[CIS]

1. Centrale Intelligenz Systeme
2. Center for Information Security
3. Centrum für Informations- und Sprachverarbeitung
4. …
5. …

Reciprocal rank: 1/3

CG, DCG

Discounted **C**umulative **G**ain

Gegeben:

- Eine Menge von Queries mit relevanten Dokumenten, jedes Dokument hat einen Relevanzwert in Bezug auf die Query (der Relevanzwert ist ein beliebiger numerischer Wert, höher ist besser)

- CG: Summe der Relevanzwerte auf den ersten Positionen (bis Position p) der Trefferliste:
  - $\mathrm{CG_p} = \sum_{i=1}^{p} rel_i$

- DCG: Summe der Relevanzwerte auf den ersten Positionen der Trefferliste, höhere Positionen zählen mehr:
  - $\mathrm{DCG_p} = \sum_{i=1}^{p} \frac{rel_i}{\log_2(i+1)}$

NDCG

Normalized **D**iscounted **C**umulative **G**ain

Gegeben:

- (= DCG): Eine Menge von Queries mit relevanten Dokumenten, jedes Dokument hat einen Relevanzwert in Bezug auf die Query

- IDCG: Ideale Summe der Relevanzwerte auf den ersten Positionen (bis Position p) der Trefferliste:
  - $\mathrm{IDCG_p} = \sum_{i=1}^{p} rel_i$ | Ideal ordering by relevancy

- NDCG: DCG in Bezug auf den idealen DCG:
  - $\mathrm{NDCG_p} = \frac{\mathrm{DCG_p}}{\mathrm{IDCG_p}}$

NDCG Beispiel

Gegeben: Query “Fratercula arctica”. Kontext: Wikipedia Suche präferiert Sprache Deutsch

Dokumente mit Relevanz:

Dok	Relevanz	Position Ergebnis	Position ideal
https://de.wikipedia.org/wiki/Papageitaucher	1	2	1
https://de.wikipedia.org/wiki/Lunde	0.8	3	2
https://en.wikipedia.org/wiki/Atlantic_puffin	0.7	1	3
https://de.wikipedia.org/wiki/Skomer	0.6	7	4
https://de.wikipedia.org/wiki/Mingulay	0.6	4	5
https://nn.wikipedia.org/wiki/Lundefugl	0.3	5	6
https://de.wikipedia.org/wiki/Eulen	0	6	-

$IDCG_6 = 2.45$

$DCG_6 = 2.1$

$NDCG_6 = 0.85$

Online metrics (search engine usage)

Monitoring the search engine

- Number of results
- Zero-result rate

Monitoring user behaviour

- Click-through rate
- Session abandonment / session success rate

Feedback collection from users