UNIVERSITÄT KOBLENZ · LANDAU

Optimum Clustering Framework

Seminar: Information Retrieval

Nicolas Schönfeld

Motivation

Inhalt

- Vorwissen
- Grundlagen des OCF
- Cluster-Qualität
- Perfektes vs. optimales Clustering
- Bestandteile des OCF
- Zusammenfassung

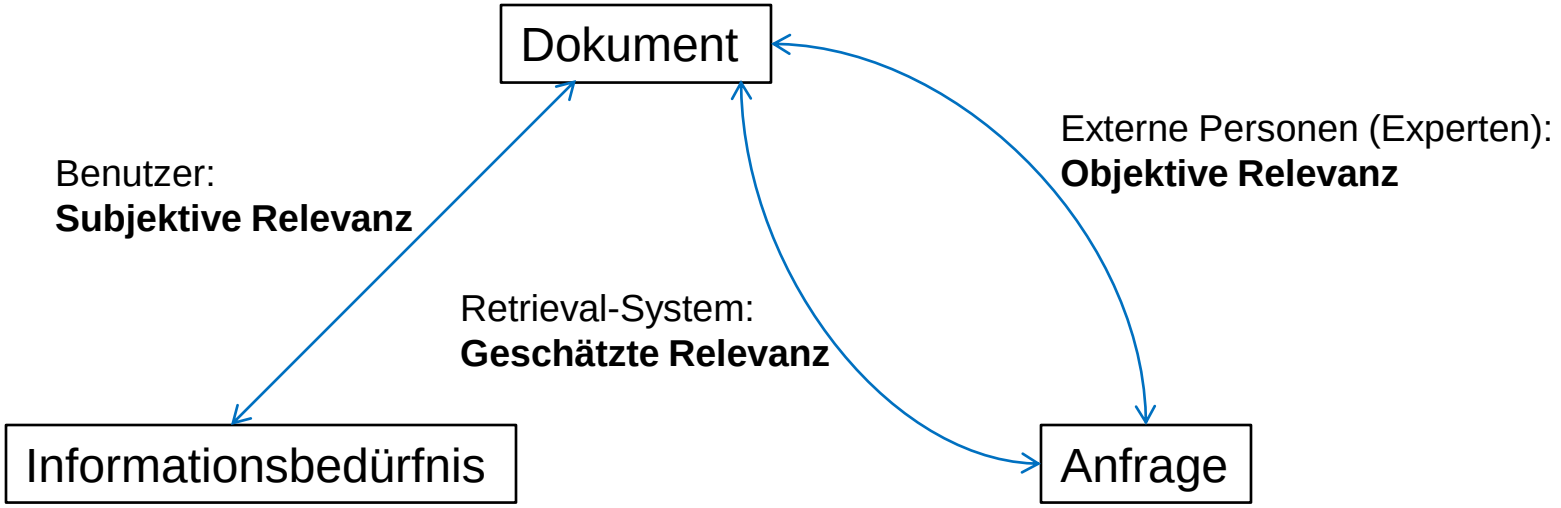
Vorwissen

- Relevanz
- Effektivität
  - Precision
  - Recall

Relevanz

- Im Information Retrieval: Beziehung zwischen einer Anfrage und einem Dokument
- Dokument gilt als relevant, wenn es das Informations-bedürfnis des Benutzer deckt
- Problem: Zur Beurteilung von Retrieval-Ergebnissen muss die richtige Antwort bekannt sein
- Praxis: Keine allgemeine Definition von Relevanz

Relevanz


Dokument
Benutzer:
Subjektive Relevanz
Externe Personen (Experten):
Objektive Relevanz
Retrieval-System:
Geschätzte Relevanz
Informationsbedürfnis
Anfrage

Effektivität

- *„Maß für die Fähigkeit eines Systems, relevante Dokumente anzuzeigen, während nicht relevante Dokumente zurückgehalten werden.“*

  [Van Rijsbergen, 1979]

- Precision: Anteil der, vom Retrieval-System gefundenen, relevanten Dokumente im Verhältnis zu allen gefundenen Dokumenten

- Recall: Anteil der relevanten Dokumente im Rechercheergebnis im Verhältnis zu allen relevanten Dokumenten der Datenbasis

Inhalt

- Vorwissen
- ➢ Grundlagen des OCF
- Cluster-Qualität
- Perfektes vs. optimales Clustering
- Bestandteile des OCF
- Zusammenfassung

Grundlagen des OCF

- Bisheriger Zustand: Die meisten Clustering-Methoden für Dokumente basieren auf Heuristiken

- Ziel des OCF: Schaffen einer theoretischen Grundlage zur Verbesserung von Clustering-Methoden

Grundlagen des OCF

Cluster-Hypothese:

„Closely associated documents tend to be relevant to the same requests.”

[Van Rijsbergen, 1979]

Grundlagen des OCF

- Ziel: Verbesserung von Dokumenten-Clustering durch Einführung einer Sammlung von Anfragen mit entsprechenden Relevanzeinschätzungen

▼

- Umkehrung der Cluster-Hypothese:

  „Documents relevant to the same queries should occur in the same cluster.”

[Fuhr et al., 2011]

Grundlagen des OCF

- Ähnlichkeit von Dokumenten neu definiert: Zwei Dokumente gelten als ähnlich, wenn sie für dieselben Anfragen relevant sind.
- Relevanz nicht bestimmbar => Berechnung der Relevanzwahrscheinlichkeit
- Optimum Clustering: Clustering, das die umgekehrte Cluster-Hypothese am besten erfüllt

Inhalt

- Vorwissen
- Grundlagen des OCF
- Cluster-Qualität
- Perfektes vs. optimales Clustering
- Bestandteile des OCF
- Zusammenfassung

Cluster-Qualität

- Es wird eine geeignete Metrik benötigt!

- Anforderungen:
  - Die Metrik muss auf einer gegebenen Anfragensammlung mit vollständigen Relevanzinformationen basieren.
  - Es sollte möglich sein, Erwartungswerte dieser Metrik durch probabilistische Retrieval-Modelle zu berechnen.

Cluster-Qualität

- Pairwise Precision:

$$P_p(D, Q, \mathcal{R}, \mathcal{C}) = \frac{1}{|D|} \sum_{\substack{C_i \in \mathcal{C} \\ c_i > 1}} c_i \sum_{q_k \in Q} \frac{r_{ik}(r_{ik}-1)}{c_i(c_i-1)}$$

c_i	Größe des jeweiligen Clusters
$r_{ik}(r_{ik}-1)$	Anzahl der Paare von relevanten Dokumenten eines Clusters für eine Anfrage q_k
$c_i(c_i-1)$	Anzahl aller Dokumentenpaare innerhalb des jeweiligen Clusters

Cluster-Qualität

- Pairwise Recall:

$$R_p(D, Q, \mathcal{R}, \mathcal{C}) = \frac{\sum_{q_k \in Q} \sum_{C_i \in \mathcal{C}} r_{ik}(r_{ik} - 1)}{\sum_{\substack{q_k \in Q \\ g_k > 1}} g_k(g_k - 1)}$$

$r_{ik}(r_{ik} - 1)$	Anzahl der Paare von relevanten Dokumenten eines Clusters für eine Anfrage q_k
$g_k(g_k - 1)$	Anzahl aller Paare von relevanten Dokumenten für eine Anfrage q_k

Cluster-Qualität

- Pairwise F-measure:

$$F_p(D, Q, \mathcal{R}, \mathcal{C}) = \frac{2}{\frac{1}{P_p(D,Q,\mathcal{R},\mathcal{C})} + \frac{1}{R_p(D,Q,\mathcal{R},\mathcal{C})}}$$

  - Harmonisches Mittel aus Pairwise Precision und Pairwise Recall

Inhalt

- Vorwissen
- Grundlagen des OCF
- Cluster-Qualität
- ➢ Perfektes vs. optimales Clustering
- Bestandteile des OCF
- Zusammenfassung

Perfektes vs. optimales Clustering

- Im klassischen Retrieval:
  - Perfektes Retrieval: Anordnung aller relevanten Dokumente vor dem ersten nicht-relevanten Dokument
    - Nur mit externen Bewertungsmaßen möglich
  - Nur optimales Retrieval im Bezug auf Dokumenten-Repräsentationen und mit limitiertem Wissen über das Informationsbedürfnis des Benutzers möglich
    - Mit internen Bewertungsmaßen möglich

Perfektes vs. optimales Clustering

- **Perfektes Clustering:**

  Es existiert kein Clustering $\mathcal{C}'$ für das gilt:

$$P_p(D, Q, \mathcal{R}, \mathcal{C}) < P_p(D, Q, \mathcal{R}, \mathcal{C}') \wedge R_p(D, Q, \mathcal{R}, \mathcal{C}) \leq R_p(D, Q, \mathcal{R}, \mathcal{C}')$$

oder

$$P_p(D, Q, \mathcal{R}, \mathcal{C}) \leq P_p(D, Q, \mathcal{R}, \mathcal{C}') \wedge R_p(D, Q, \mathcal{R}, \mathcal{C}) < R_p(D, Q, \mathcal{R}, \mathcal{C}')$$

Perfektes vs. optimales Clustering

- Voraussetzung für Definition von optimalem Clustering:
  - Schätzung der relevanten Dokumentenpaare in einem Cluster
  - Schätzung der Qualität eines Clusterings durch Berechnung von Erwartungswerten der zuvor definierten Metriken

    => Expected Precision, Expected Recall, Expected F-measure
- Definition nun analog zum perfekten Clustering
- Unterschied: Ersetzen der externen Relevanzbeurteilung (Pairwise Precision/Recall) durch Schätzungen der Relevanzwahrscheinlichkeit (Expected Precision/Recall)

Inhalt

- Vorwissen
- Grundlagen des OCF
- Cluster-Qualität
- Perfektes vs. optimales Clustering
- ➢ Bestandteile des OCF
- Zusammenfassung

Bestandteile des OCF

- Methoden zum Dokumenten-Clustering bestehen grundsätzlich aus 3 Komponenten:

  1. Anfragen-Sammlung
  2. Retrieval-Funktion
  3. Ähnlichkeitsmaß für Dokumente

- OCF: Geeignete Wahl dieser 3 Komponenten

Bestandteile des OCF

Anfragen-Sammlung:

- Herausforderung: Anfragen finden, die dem aktuellen Informationsbedürfnis des Benutzers ähnlich sind
- 3 Methoden zur Erstellung einer Anfragen-Sammlung
  - Lokal
  - Global
  - Extern

Zusammenfassung

- Jede Clustering-Methode basiert auf einer Anfragen-Sammlung, einer Retrieval-Funktion und einem Ähnlichkeitsmaß für Dokumente
- Optimale Cluster-Qualität für eine gegebene Anfragen-Sammlung und probabilistische Retrieval-Funktion dank theoretischer Grundlage
- Ersetzen der bisher vorherrschenden heuristischen Methoden durch solideren Ansatz
- Rahmenwerk ermöglicht gezieltere Forschung nach besseren Clustering-Methoden

Vielen Dank für die Aufmerksamkeit!