Total Recall via Keyqueries: A Case Study for Systematic Reviews

Paul Alexander Cahn

28.05.2021

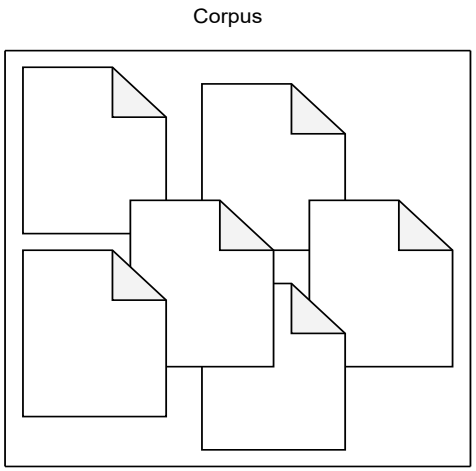
Einleitung

Anfrage:
Welche Umweltartikel gibt es, die Fragen zu Auswirkungen oder Expositionen mit einem Link zur Umweltmanagementpolitik oder -praxis beantworten?

Einleitung

Anfrage:
Welche Umweltartikel gibt es, die Fragen zu Auswirkungen oder Expositionen mit einem Link zur Umweltmanagementpolitik oder -praxis beantworten?

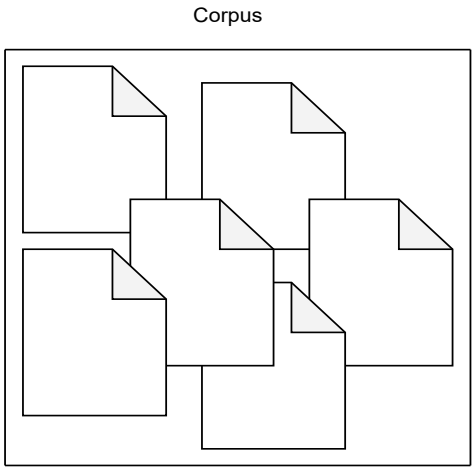
Nutzung der komplizierten Anfrage, um alle potenziell relevanten Dokumente zu sammeln



Einleitung

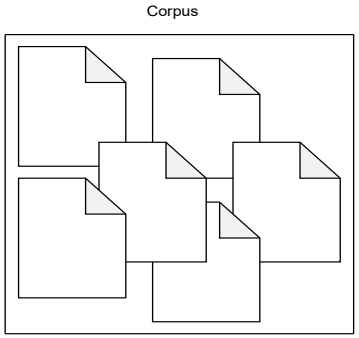
Anfrage:
Welche Umweltartikel gibt es, die Fragen zu Auswirkungen oder Expositionen mit einem Link zur Umweltmanagementpolitik oder -praxis beantworten?

Nutzung der komplizierten Anfrage, um alle potenziell relevanten Dokumente zu sammeln

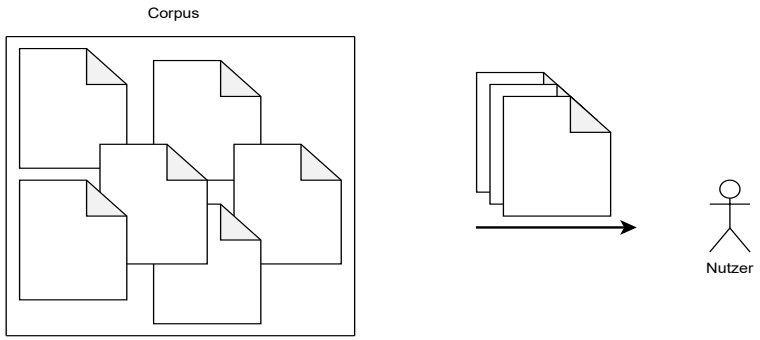


Total Recall:
Extraktion aller relevanten Dokumente aus dem Corpus

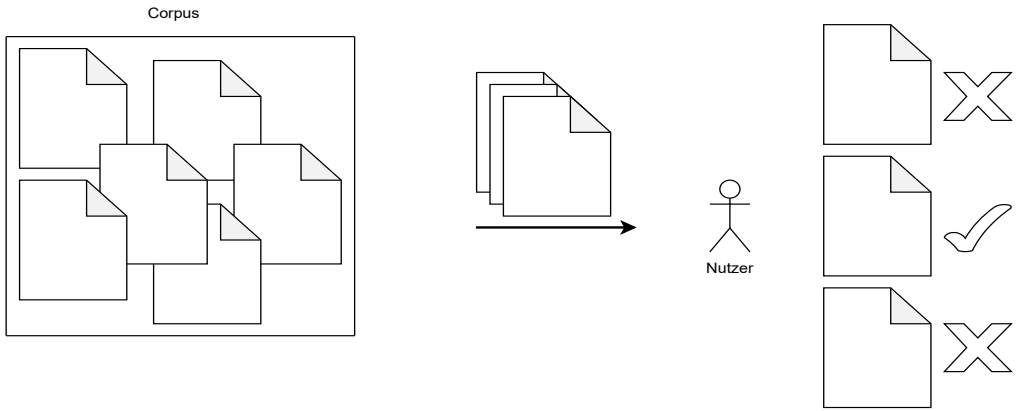
Human-In-The-Loop zur Extraktion aller relevanter Dokumente



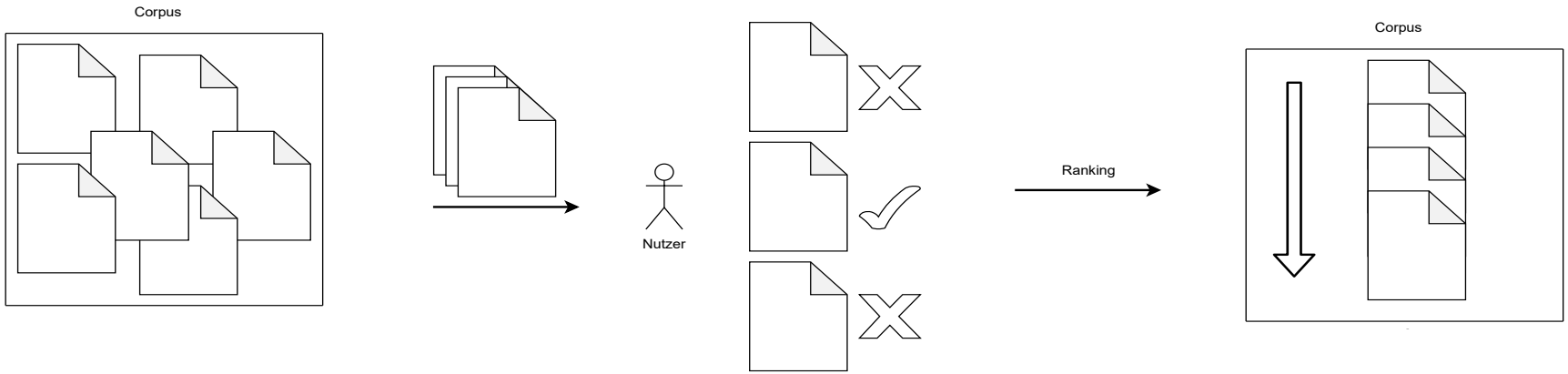
Human-In-The-Loop zur Extraktion aller relevanter Dokumente



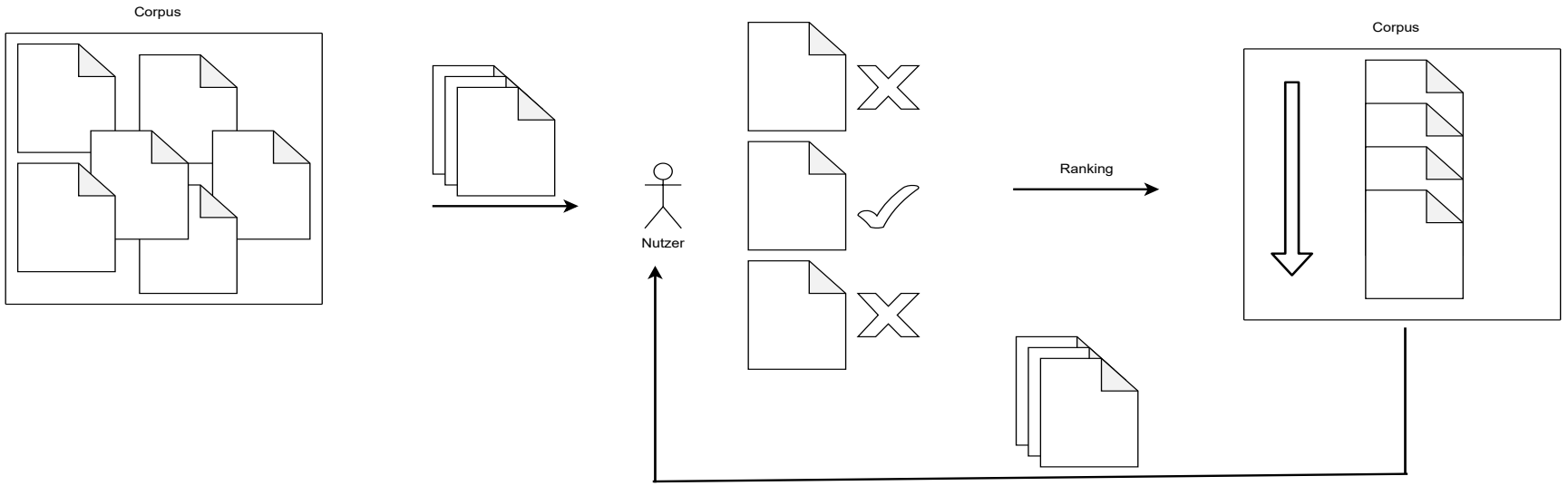
Human-In-The-Loop zur Extraktion aller relevanter Dokumente



Human-In-The-Loop zur Extraktion aller relevanter Dokumente

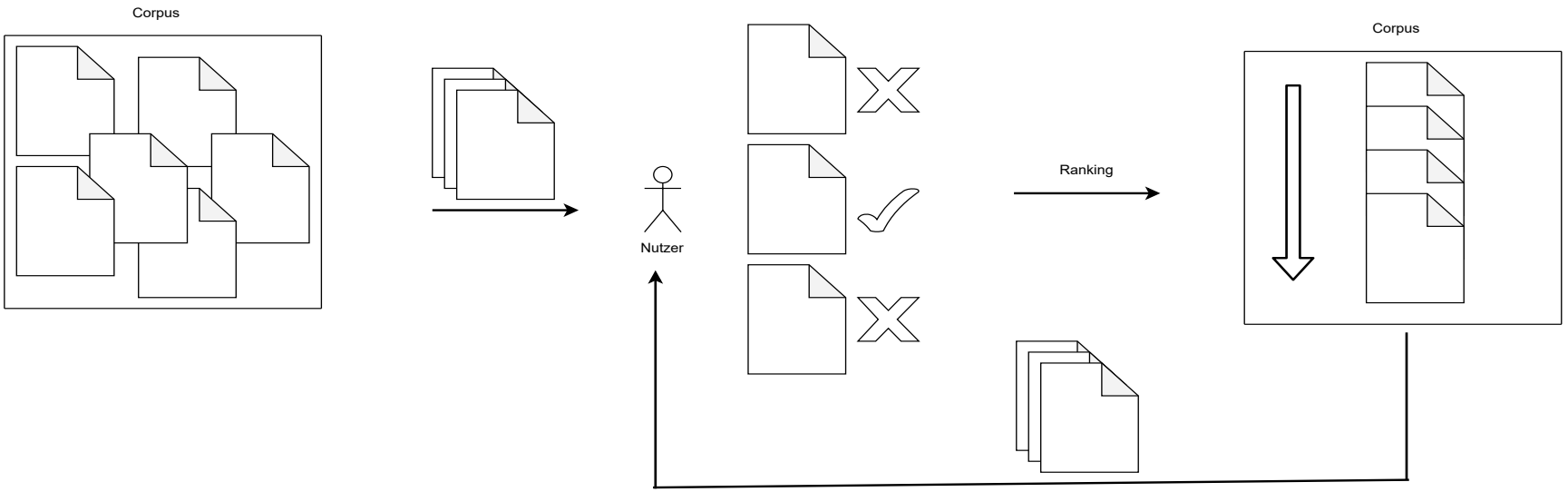


Human-In-The-Loop zur Extraktion aller relevanter Dokumente



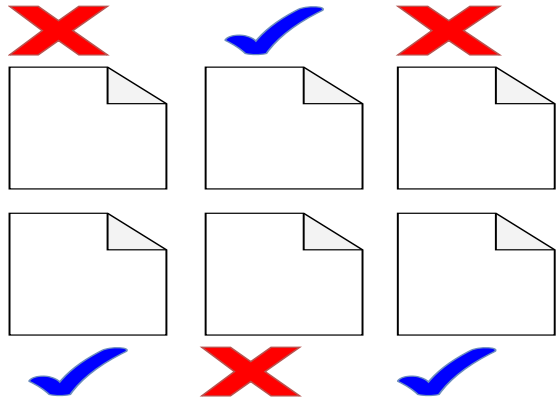
Ranking mit Machine Learning

- Labeln der Dokumente als relevant oder nicht-relevant durch Nutzer
- Training des Models mit den gelabelten Daten
- Klassifizierung nicht gesehener Dokumente durch das Model
- Sortierung der neu-klassifizierten Dokumente nach ihrer Wahrscheinlichkeit in die zwei Klassen
- Erneute Iteration; Labeln der neu-klassifizierten Dokumente durch Nutzer



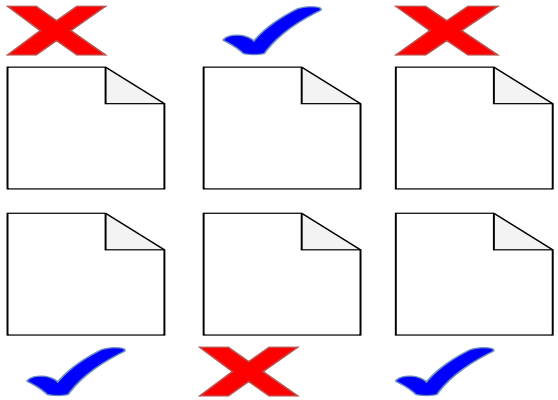
Machine Learning - Beispiel: Naive Bayes

Trainingsdaten: $x = x_1 \ldots x_n$

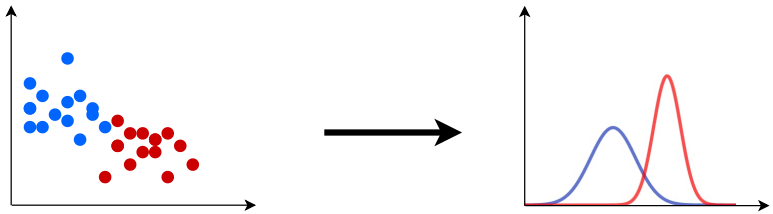


Machine Learning - Beispiel: Naive Bayes

Trainingsdaten: $x = x_1 \ldots x_n$

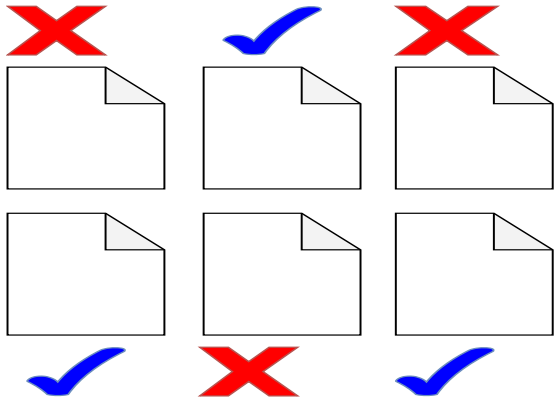


Satz von Bayes:
$P(K_i|x) = \frac{P(x|K_i) \cdot P(K_i)}{P(x)}$
mit i $\in \{relevant, nichtrelevant\}$

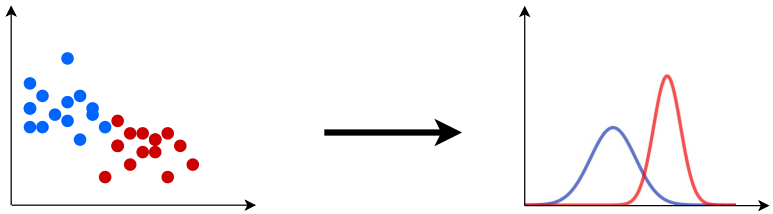


Machine Learning - Beispiel: Naive Bayes

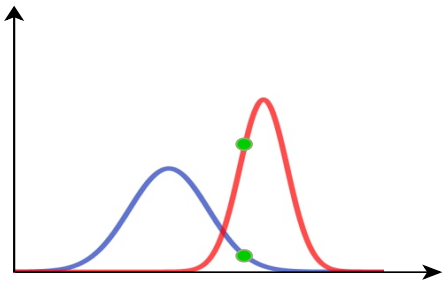
Trainingsdaten: $x = x_1 \ldots x_n$



Satz von Bayes:
$P(K_i|x) = \frac{P(x|K_i) \cdot P(K_i)}{P(x)}$
mit i $\in \{relevant, nichtrelevant\}$



Testdaten $y = y_1 \ldots y_k$
Frage: $P(K_i|y)$



Keyqueries - Motivation

- Anfragen, die relevante Dokumente in den top-k Ergebnissen zurückliefern
- Möglicher Einsatz in der related-work Suche
- Retrieval relevanter Dokumente durch Keyqueries aus relevanten Dokumenten

Supporting Scholarly Search with Keyqueries

Matthias Hagen, Anna Beyer, +2 authors Benno Stein • Published in ECIR 2016 • Computer Science

We deal with a problem faced by scholars every day: identifying relevant papers on a given topic. In particular, we focus on the scenario where a scholar can come up with a few papers (e.g., suggested by a colleague) and then wants to find "all" the other related publications. Our proposed approach to the problem is based on the concept of keyqueries: formulating keyqueries from the input papers and suggesting the top results as candidates of related work.

Keyqueries for Clustering and Labeling

Tim Gollub, Matthias Busse, +1 author Matthias Hagen • Published in AIRS 2016 • Computer Science

In this paper we revisit the document clustering problem from an information retrieval perspective. The idea is to use queries as features in the clustering process that finally also serve as descriptive cluster labels "for free." Our novel perspective includes query constraints for clustering and cluster labeling that ensure consistency with a keyword-based reference search engine.

Keyqueries - Motivation

- Anfragen, die relevante Dokumente in den top-k Ergebnissen zurückliefern
- Möglicher Einsatz in der related-work Suche
- Retrieval relevanter Dokumente durch Keyqueries aus relevanten Dokumenten

Supporting Scholarly Search with Keyqueries

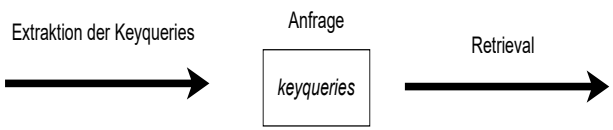
Matthias Hagen, Anna Beyer, +2 authors Benno Stein • Published in ECIR 2016 • Computer Science

We deal with a problem faced by scholars every day: identifying relevant papers on a given topic. In particular, we focus on the scenario where a scholar can come up with a few papers (e.g., suggested by a colleague) and then wants to find "all" the other related publications. Our proposed approach to the problem is based on the concept of keyqueries: formulating keyqueries from the input papers and suggesting the top results as candidates of related work.

Keyqueries for Clustering and Labeling

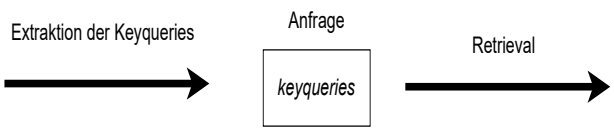
Tim Gollub, Matthias Busse, +1 author Matthias Hagen • Published in AIRS 2016 • Computer Science

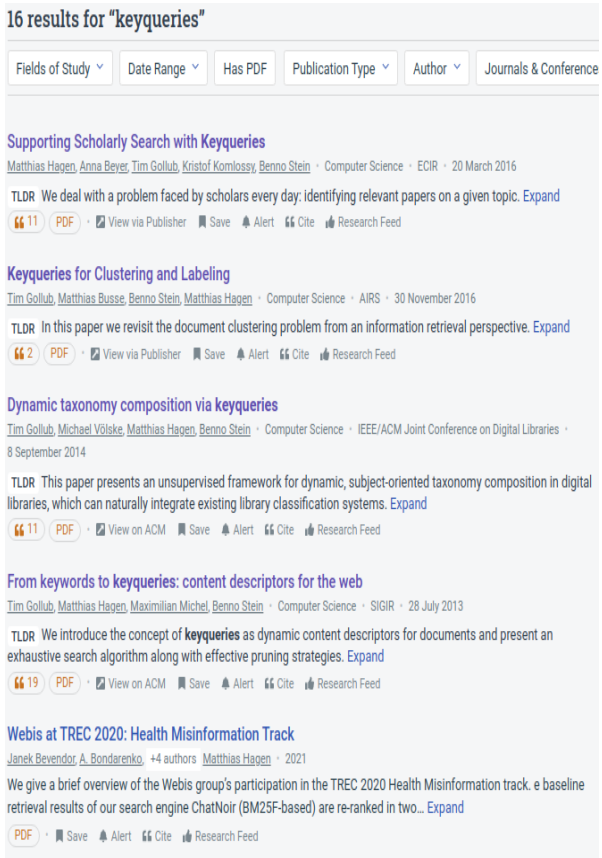
In this paper we revisit the document clustering problem from an information retrieval perspective. The idea is to use queries as features in the clustering process that finally also serve as descriptive cluster labels "for free." Our novel perspective includes query constraints for clustering and cluster labeling that ensure consistency with a keyword-based reference search engine.



Keyqueries - Motivation

- Anfragen, die relevante Dokumente in den top-k Ergebnissen zurückliefern
- Möglicher Einsatz in der related-work Suche
- Retrieval relevanter Dokumente durch Keyqueries aus relevanten Dokumenten

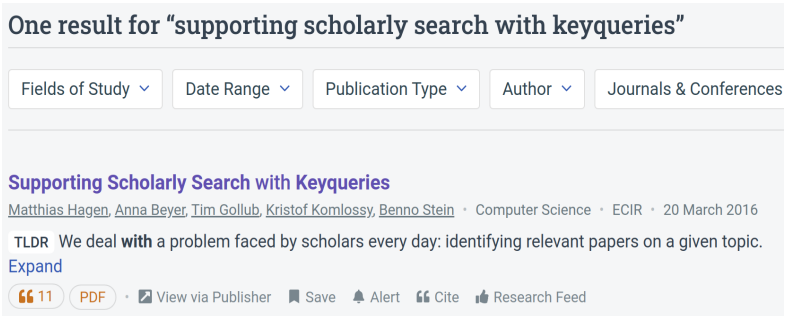




Keyquery - Eigenschaften

- mindestens ℓ Ergebnisse

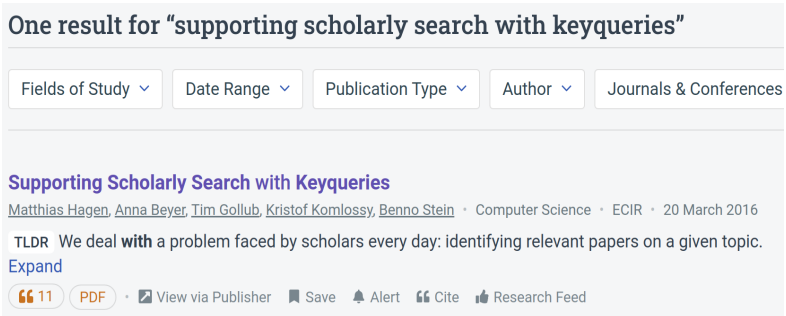
Relevantes Dokument:
"Supporting Scholarly Search with Keyqueries"



Keyquery - Eigenschaften

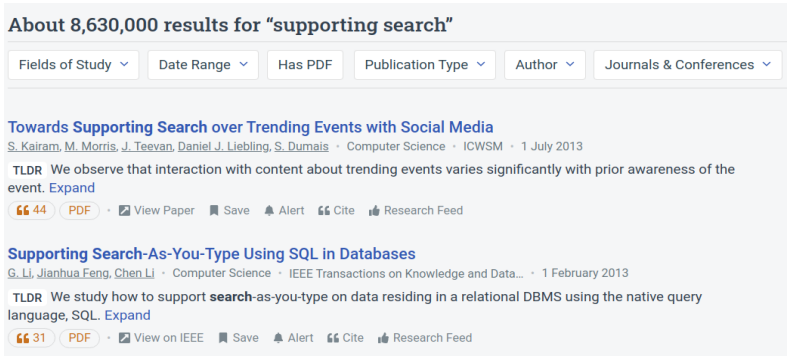
- mindestens ℓ Ergebnisse

Relevantes Dokument: "Supporting Scholarly Search with Keyqueries"



- relevantes Dokument in den top-k

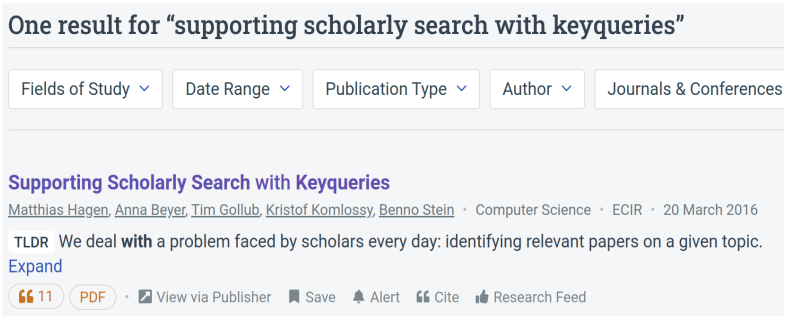
Relevantes Dokument: "Supporting Scholarly Search with Keyqueries"



Keyquery - Eigenschaften

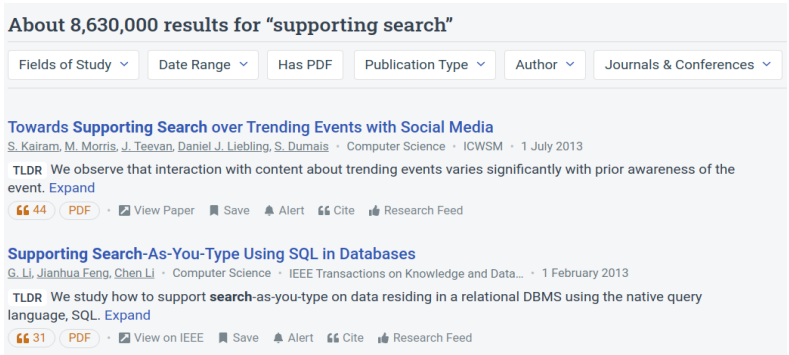
- mindestens ℓ Ergebnisse

Relevantes Dokument: "Supporting Scholarly Search with Keyqueries"

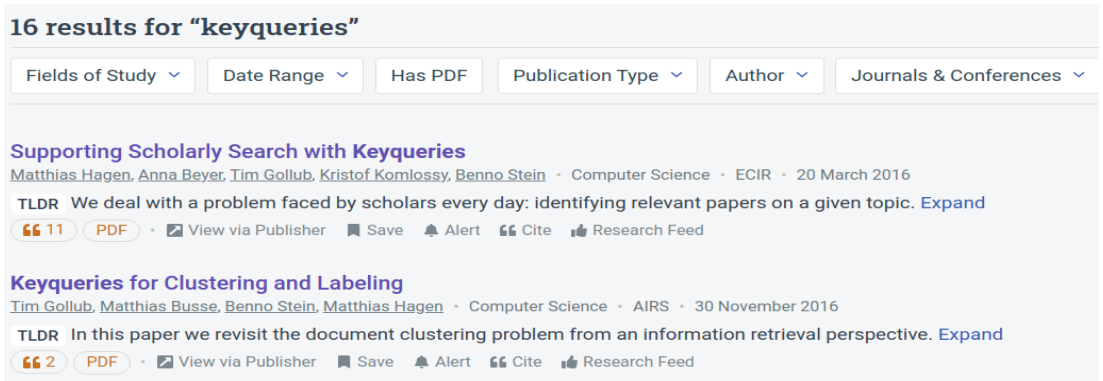


- relevantes Dokument in den top-k

Relevantes Dokument: "Supporting Scholarly Search with Keyqueries"



- Anteil m an relevanten Dokumenten in den top-k Ergebnissen
  Relevante Dokumente: "Supporting Scholarly Search with Keyqueries" und "From Keywords to Keyqueries"



Keyquery - Interleaving Strategien

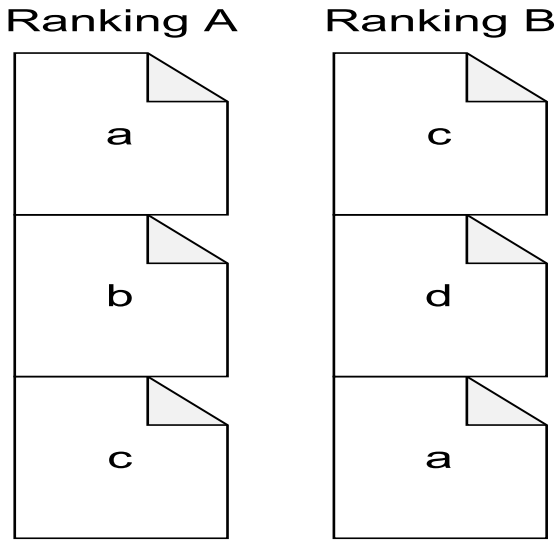
- Interleaving - Zusammenfügen verschiedener Rankings
- Motivation: Nur ein finales Ranking aller Dokumente

Keyquery - Interleaving Strategien

- Interleaving - Zusammenfügen verschiedener Rankings
- Motivation: Nur ein finales Ranking aller Dokumente
- Balanced Interleaving:
  Erstes Dokument jedes Rankings wird der Reihe nach genommen
- Teamdraft Interleaving:
  Erstes Dokument wird zufällig von einem der Rankings genommen

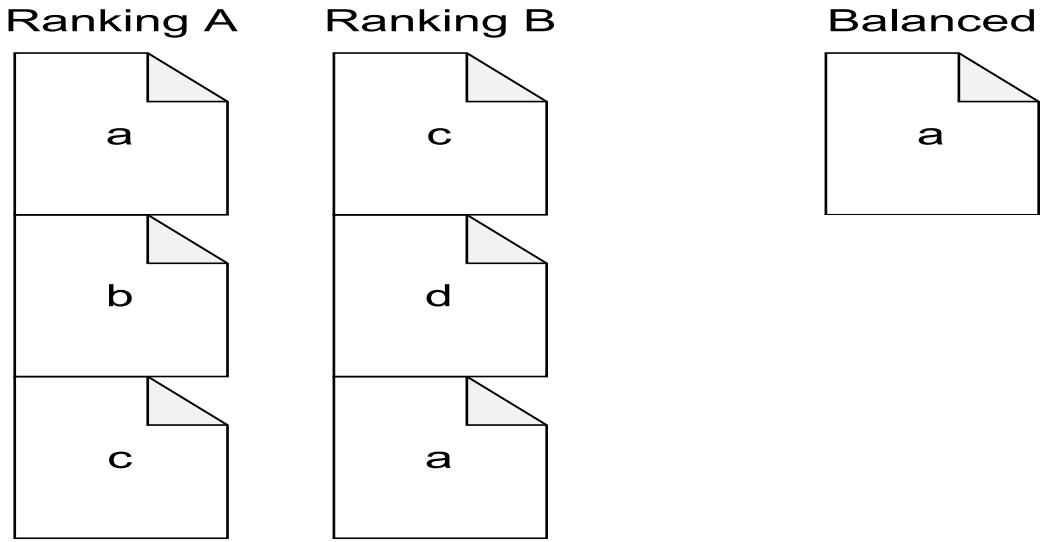
Keyquery - Interleaving Strategien

- Interleaving - Zusammenfügen verschiedener Rankings
- Motivation: Nur ein finales Ranking aller Dokumente
- Balanced Interleaving:
  Erstes Dokument jedes Rankings wird der Reihe nach genommen
- Teamdraft Interleaving:
  Erstes Dokument wird zufällig von einem der Rankings genommen



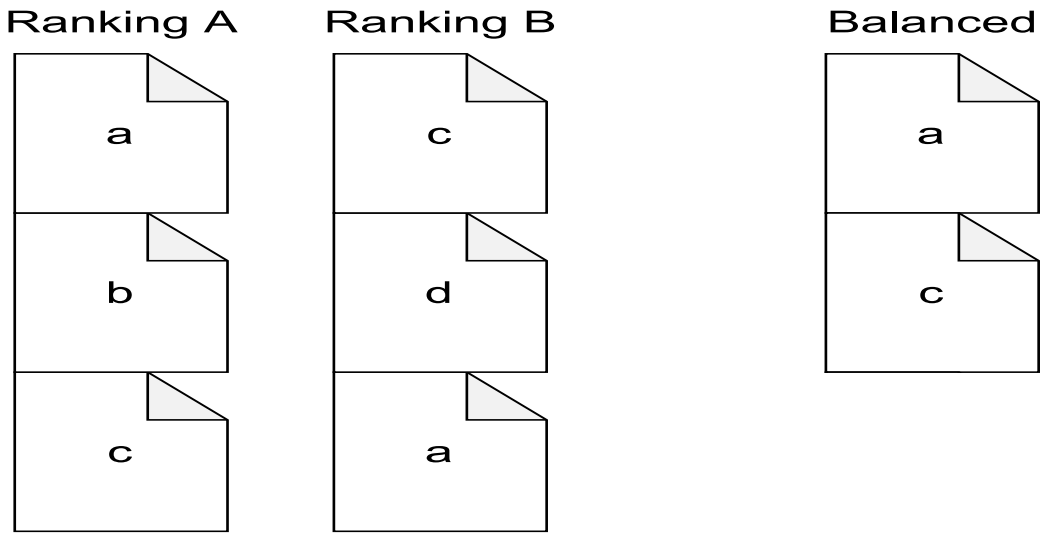
Keyquery - Interleaving Strategien

- Interleaving - Zusammenfügen verschiedener Rankings
- Motivation: Nur ein finales Ranking aller Dokumente
- Balanced Interleaving: Erstes Dokument jedes Rankings wird der Reihe nach genommen
- Teamdraft Interleaving: Erstes Dokument wird zufällig von einem der Rankings genommen



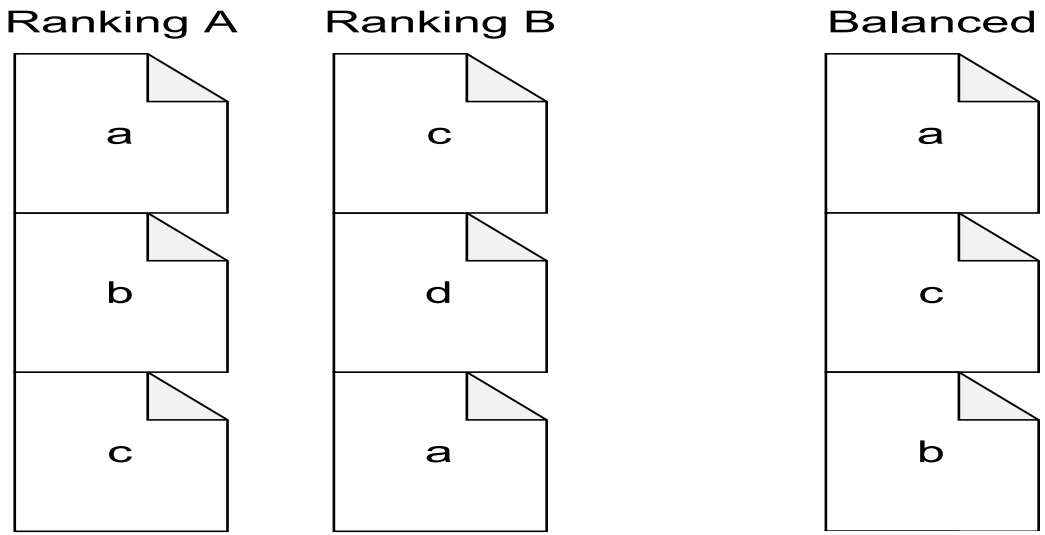
Keyquery - Interleaving Strategien

- Interleaving - Zusammenfügen verschiedener Rankings
- Motivation: Nur ein finales Ranking aller Dokumente
- Balanced Interleaving:
  Erstes Dokument jedes Rankings wird der Reihe nach genommen
- Teamdraft Interleaving:
  Erstes Dokument wird zufällig von einem der Rankings genommen



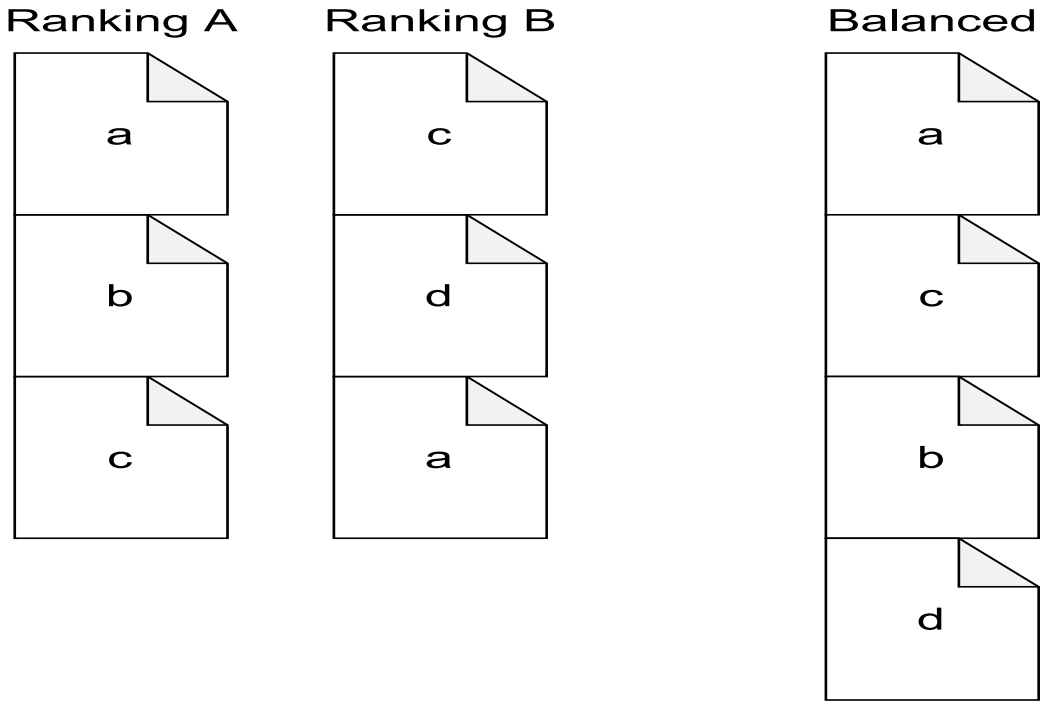
Keyquery - Interleaving Strategien

- Interleaving - Zusammenfügen verschiedener Rankings
- Motivation: Nur ein finales Ranking aller Dokumente
- Balanced Interleaving:
  Erstes Dokument jedes Rankings wird der Reihe nach genommen
- Teamdraft Interleaving:
  Erstes Dokument wird zufällig von einem der Rankings genommen



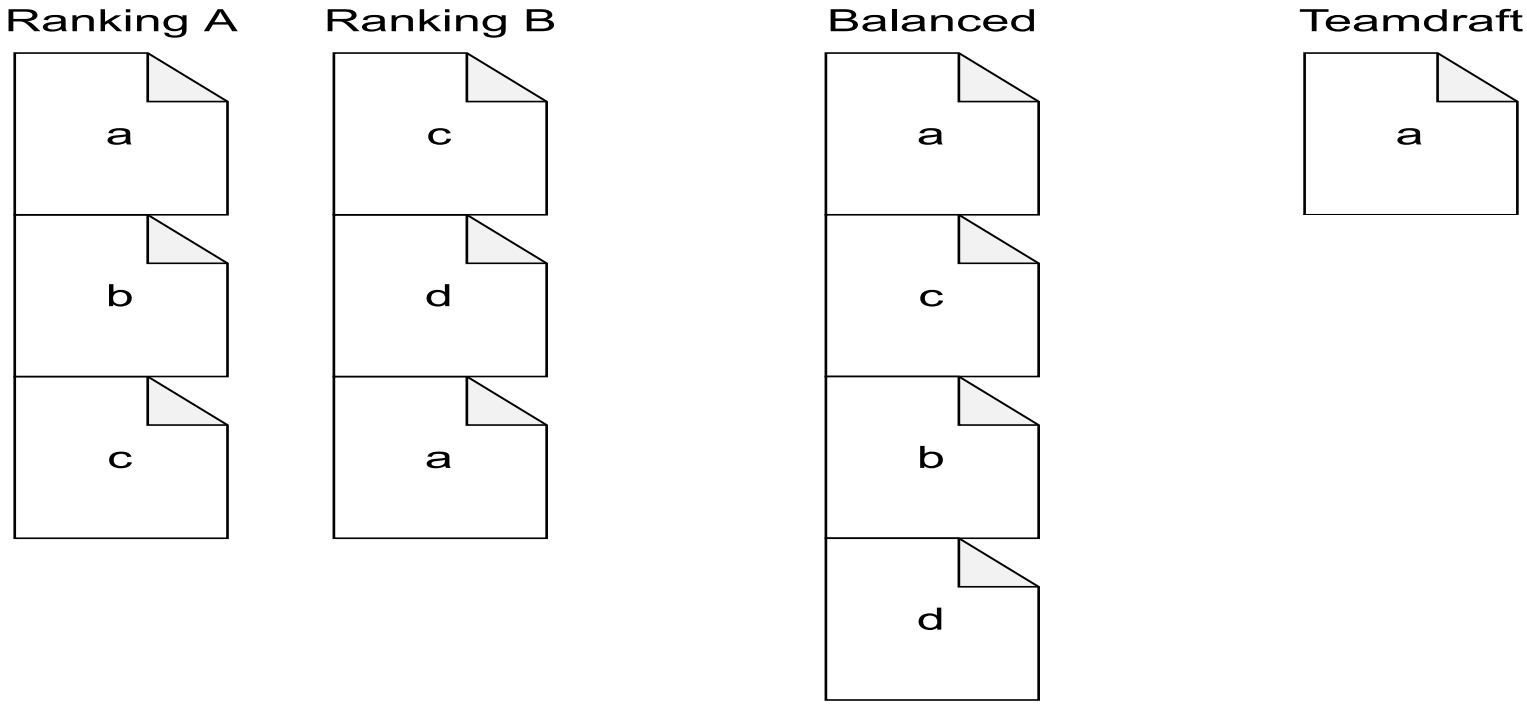
Keyquery - Interleaving Strategien

- Interleaving - Zusammenfügen verschiedener Rankings
- Motivation: Nur ein finales Ranking aller Dokumente
- Balanced Interleaving:
  Erstes Dokument jedes Rankings wird der Reihe nach genommen
- Teamdraft Interleaving:
  Erstes Dokument wird zufällig von einem der Rankings genommen



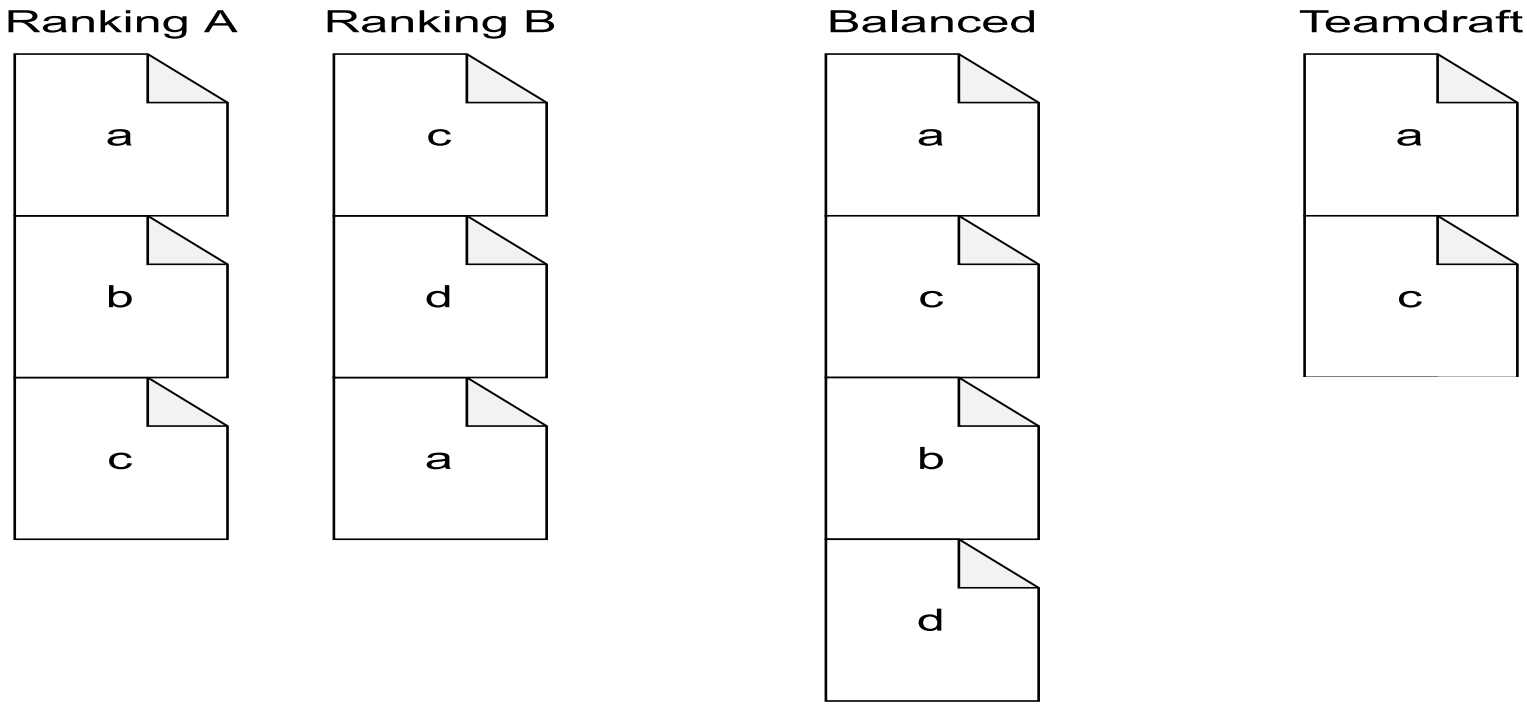
Keyquery - Interleaving Strategien

- Interleaving - Zusammenfügen verschiedener Rankings
- Motivation: Nur ein finales Ranking aller Dokumente
- Balanced Interleaving: Erstes Dokument jedes Rankings wird der Reihe nach genommen
- Teamdraft Interleaving: Erstes Dokument wird zufällig von einem der Rankings genommen



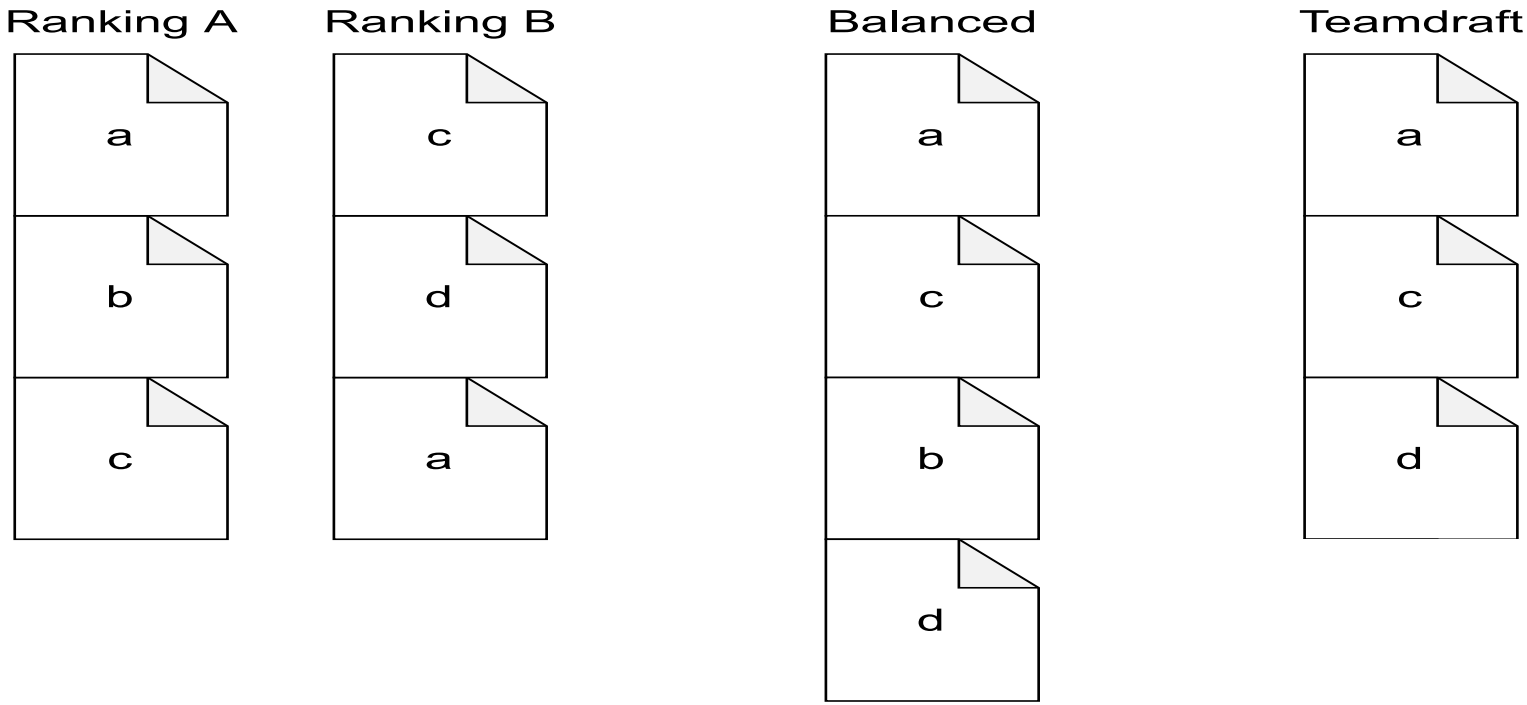
Keyquery - Interleaving Strategien

- Interleaving - Zusammenfügen verschiedener Rankings
- Motivation: Nur ein finales Ranking aller Dokumente
- Balanced Interleaving: Erstes Dokument jedes Rankings wird der Reihe nach genommen
- Teamdraft Interleaving: Erstes Dokument wird zufällig von einem der Rankings genommen



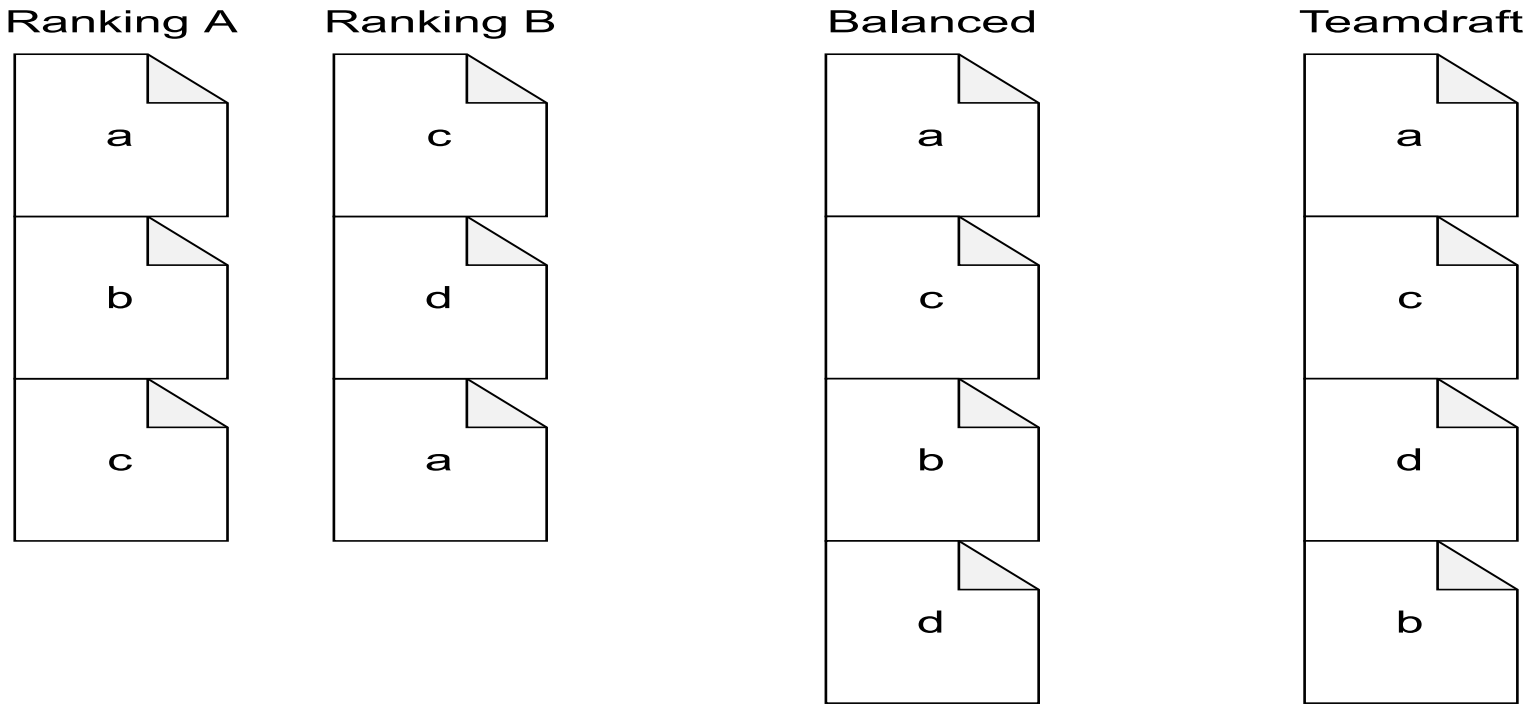
Keyquery - Interleaving Strategien

- Interleaving - Zusammenfügen verschiedener Rankings
- Motivation: Nur ein finales Ranking aller Dokumente
- Balanced Interleaving: Erstes Dokument jedes Rankings wird der Reihe nach genommen
- Teamdraft Interleaving: Erstes Dokument wird zufällig von einem der Rankings genommen



Keyquery - Interleaving Strategien

- Interleaving - Zusammenfügen verschiedener Rankings
- Motivation: Nur ein finales Ranking aller Dokumente
- Balanced Interleaving: Erstes Dokument jedes Rankings wird der Reihe nach genommen
- Teamdraft Interleaving: Erstes Dokument wird zufällig von einem der Rankings genommen



Unterschiede zwischen Keyquery-Ansatz und Machine Learning Ansatz

- Keyquery basierter Ansatz
  - Extraktion von Keyqueries aus Feedback des Nutzers
  - Nutzung der Keyqueries für Dokumentenretrieval
  - Zusammenfügen der Ergebnisse zu finalem Ergebnis
- Machine Learning basierte Ansätze
  - Training mit dem Feedback des Nutzers
  - Überprüfung für jedes Dokument, wie groß die Wahrscheinlichkeit ist, dass es relevant ist
  - Sortierung nach dieser Wahrscheinlichkeit

Experimentaufbau

- Datensätze
  - Julius-Kühn-Institut (JKI) Datensätze zur Erstellung von biologischen systematischen Reviews von 2018 und 2019
    - 5% relevante Dokumente in den JKI Datensätzen

Experimentaufbau

- Datensätze
  - Julius-Kühn-Institut (JKI) Datensätze zur Erstellung von biologischen systematischen Reviews von 2018 und 2019
    - ★ 5% relevante Dokumente in den JKI Datensätzen
  - ein Datensatz von einer TREC Konferenz zu high-recall Systemen
    - ★ 26% relevante Dokumente im TREC Datensatz

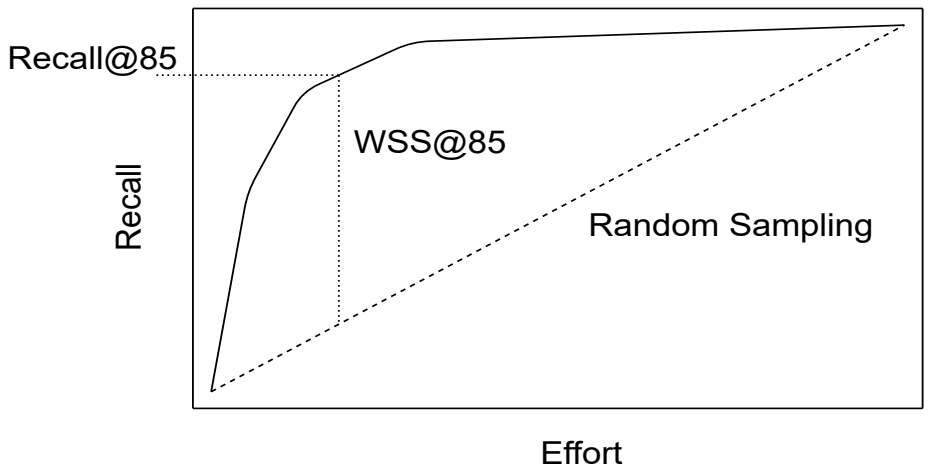
Experimentaufbau

- Datensätze
  - Julius-Kühn-Institut (JKI) Datensätze zur Erstellung von biologischen systematischen Reviews von 2018 und 2019
    - ★ 5% relevante Dokumente in den JKI Datensätzen
  - ein Datensatz von einer TREC Konferenz zu high-recall Systemen
    - ★ 26% relevante Dokumente im TREC Datensatz
  - JKI Datensatz 2018 als Validierungssatz
  - Datensätze von 2019 als Testsätze

Experimentaufbau

- Datensätze
  - Julius-Kühn-Institut (JKI) Datensätze zur Erstellung von biologischen systematischen Reviews von 2018 und 2019
    - ★ 5% relevante Dokumente in den JKI Datensätzen
  - ein Datensatz von einer TREC Konferenz zu high-recall Systemen
    - ★ 26% relevante Dokumente im TREC Datensatz
  - JKI Datensatz 2018 als Validierungssatz
  - Datensätze von 2019 als Testsätze
- Metrik zur Evaluation: Worked saved over sampling

$$WSS@Recall = \frac{True\ \ Negatives\ +\ False\ \ Negatives}{Total\ \ Number\ \ of\ \ Docs} - (1\ -\ \ Recall)$$

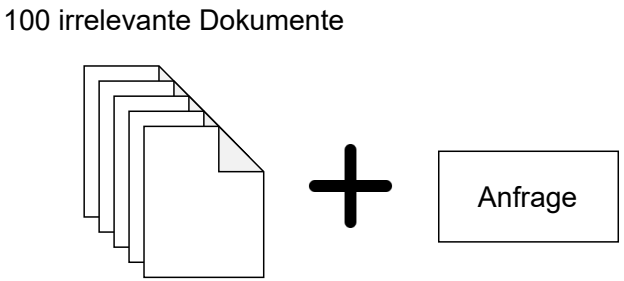


Baseline - HiCal

- HiCal - Algorithmus der Universität Waterloo
- Für High-Recall Aufgaben
- State-Of-The-Art auf diesem Gebiet
- Dynamische Batchgröße: $k \leftarrow k + \frac{k+9}{10}$

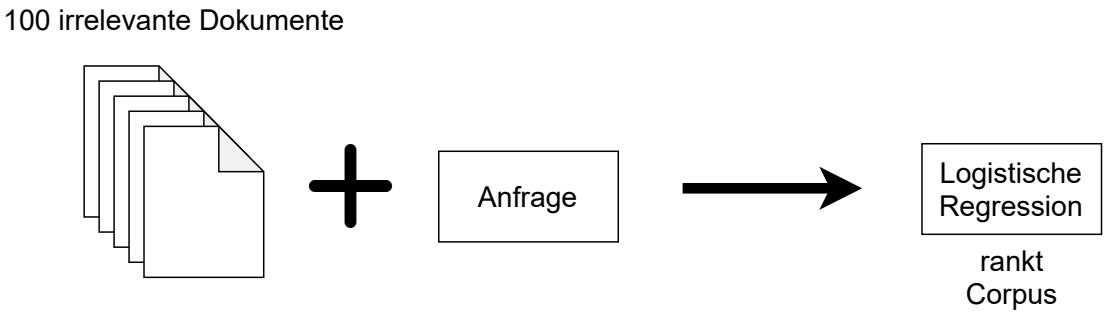
Baseline - HiCal

- HiCal - Algorithmus der Universität Waterloo
- Für High-Recall Aufgaben
- State-Of-The-Art auf diesem Gebiet
- Dynamische Batchgröße: $k \leftarrow k + \frac{k+9}{10}$



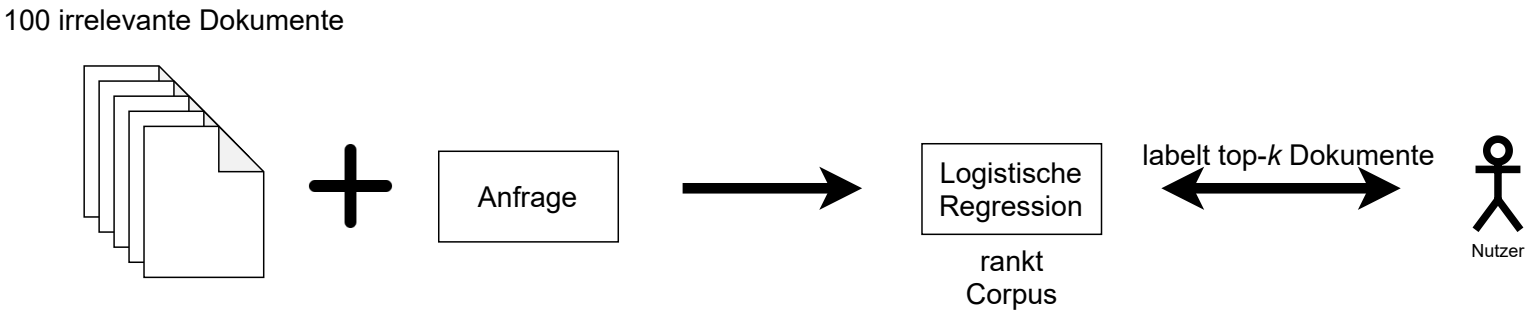
Baseline - HiCal

- HiCal - Algorithmus der Universität Waterloo
- Für High-Recall Aufgaben
- State-Of-The-Art auf diesem Gebiet
- Dynamische Batchgröße: $k \leftarrow k + \frac{k+9}{10}$



Baseline - HiCal

- HiCal - Algorithmus der Universität Waterloo
- Für High-Recall Aufgaben
- State-Of-The-Art auf diesem Gebiet
- Dynamische Batchgröße: $k \leftarrow k + \frac{k+9}{10}$



Methodologie der Machine-Learning Ansätze (1)

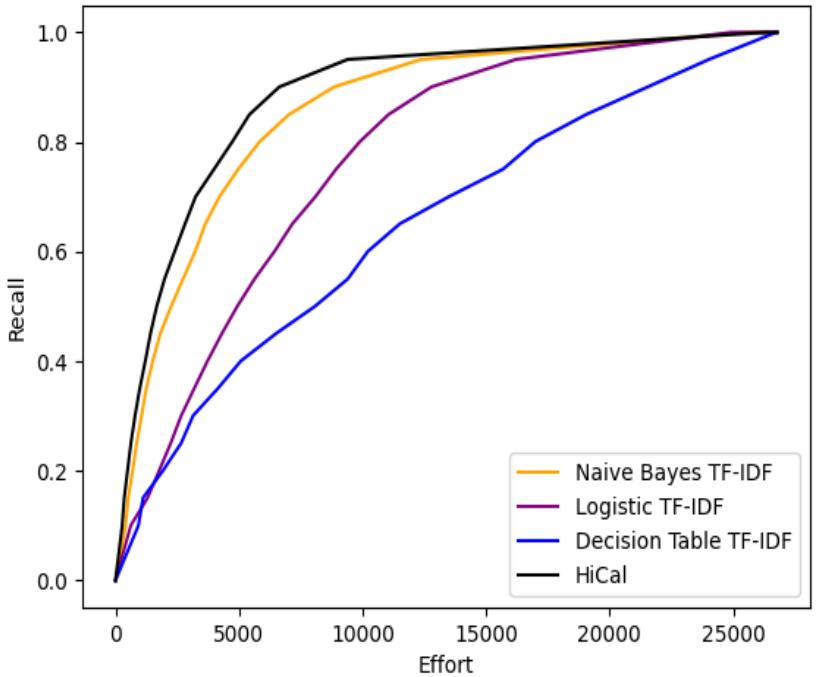
- Auswahl der Features
  - Boolean word occurrence: Vorkommen eines bestimmten Terms im Dokument in Abhängigkeit aller Terme
  - Term frequency: Anzahl eines Terms in einem Dokument
  - Term frequency - inverse document frequency: $log \frac{N}{\sum_{D:t \in D} 1}$

    D : Dokument, N : Anzahl aller Dokumente

Methodologie der Machine-Learning Ansätze (1)

- Auswahl der Features
  - Boolean word occurrence: Vorkommen eines bestimmten Terms im Dokument in Abhängigkeit aller Terme
  - Term frequency: Anzahl eines Terms in einem Dokument
  - Term frequency - inverse document frequency: $log \frac{N}{\sum_{D:t \in D} 1}$

    D : Dokument, N : Anzahl aller Dokumente

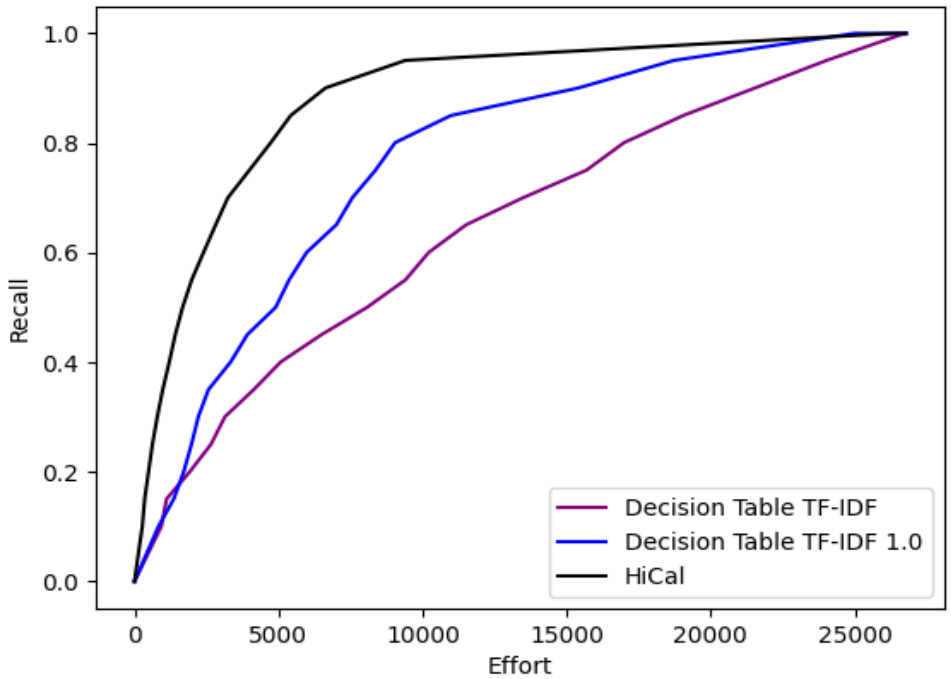


Methodologie der Machine-Learning Ansätze (2)

- Undersampling der Daten
  - Über 90% der Dokumente irrelevant
  - Testen verschiedener Verhältnisse zwischen relevanten und irrelevanten Dokumenten
    - ★ Verhältnis $\in \{0.5, 1.0, 1.5, 2.0\}$

Methodologie der Machine-Learning Ansätze (2)

- Undersampling der Daten
  - Über 90% der Dokumente irrelevant
  - Testen verschiedener Verhältnisse zwischen relevanten und irrelevanten Dokumenten
    - ★ Verhältnis $\in \{0.5, 1.0, 1.5, 2.0\}$



Methodologie des Keyquery Ansatzes (1)

- Keyquery Kriterien
  - k : Relevante Dokumente müssen in den top-k Dokumenten sein
  - ℓ : Retrieval von mindestens ℓ Dokumenten
  - m : Anzahl der relevanten Dokumente in den top-k

Methodologie des Keyquery Ansatzes (1)

- Keyquery Kriterien
  - k : Relevante Dokumente müssen in den top-k Dokumenten sein
  - ℓ : Retrieval von mindestens ℓ Dokumenten
  - m : Anzahl der relevanten Dokumente in den top-k

Approach	k	l	m	WSS @85%	WSS @90%	WSS @95%
Keyquery 10	10	1		45.19	41.12	34.71
Keyquery 10	10	7		33.85	29.10	23.98
Keyquery 20	20	2		47.70	44.61	38.50
Keyquery 20	50	6		**51.06**	**50.07**	37.69
Keyquery 50	50	34		20.24	11.66	9.68
Keyquery 50	100	25		23.15	20.72	11.91
Keyquery100	100	10		50.97	50.01	**43.24**
BM25				20.24	11.66	9.68
HiCal				64.75	65.27	59.96

Methodologie des Keyquery Ansatzes (2)

- Interleaving Strategien
  - Balanced : Beste Ergebnisse der Rankings werden abwechselnd genommen
  - Teamdraft : Zufällige Auswahl, von welchem Ranking zuerst gewählt wird

Methodologie des Keyquery Ansatzes (2)

- Interleaving Strategien
  - Balanced : Beste Ergebnisse der Rankings werden abwechselnd genommen
  - Teamdraft : Zufällige Auswahl, von welchem Ranking zuerst gewählt wird

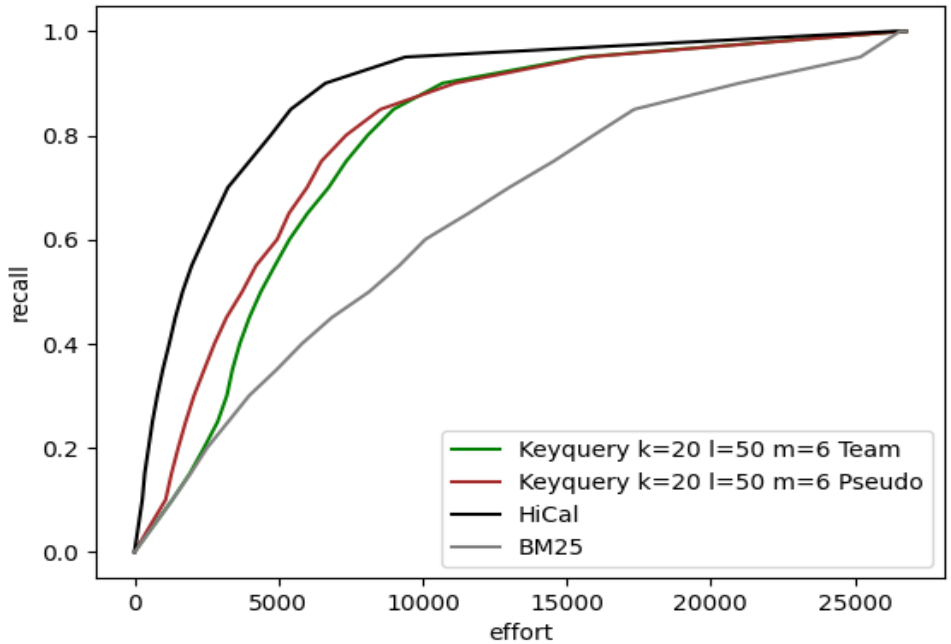
Approach	I	WSS @85%	WSS @90%	WSS @95%
Keyquery (1)	Team	51.43	50.10	36.93
Keyquery (2)	Team	**54.50**	**52.69**	**45.26**
Keyquery (1)	Bal	51.06	50.07	37.69
Keyquery (2)	Bal	50.97	50.01	43.24
BM25		20.24	11.66	9.68
HiCal		64.75	65.27	59.96

Methodologie des Keyquery Ansatzes (3)

- Pseudo-Relevanz Feedback
  - Erhalten potenziell relevanter Dokumente durch Machine Learning Ansätze $\rightarrow$ Erstellen von Keyqueries
  - Vorschlag von 5 als relevant eingestuften Dokumenten durch jeden Ansatz
  - Benutzung eines vorgeschlagenen Dokuments, wenn es mit 80% Wahrscheinlichkeit relevant ist

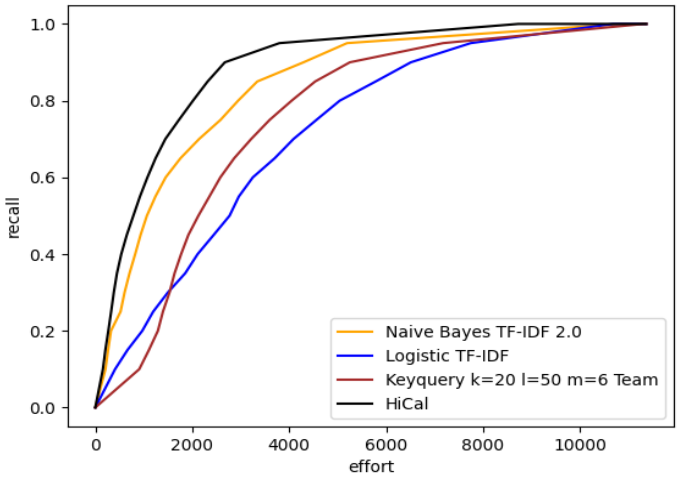
Methodologie des Keyquery Ansatzes (3)

- Pseudo-Relevanz Feedback
  - Erhalten potenziell relevanter Dokumente durch Machine Learning Ansätze $\rightarrow$ Erstellen von Keyqueries
  - Vorschlag von 5 als relevant eingestuften Dokumenten durch jeden Ansatz
  - Benutzung eines vorgeschlagenen Dokuments, wenn es mit 80% Wahrscheinlichkeit relevant ist



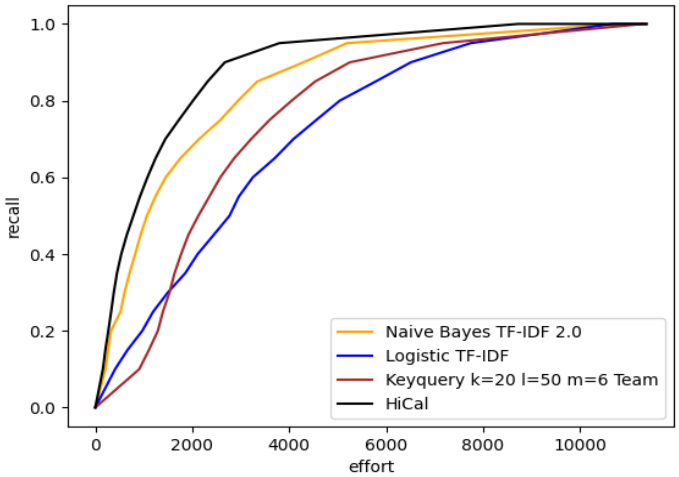
Evaluation der Ansätze auf den neueren Datensätzen

JKI Datensatz 2019

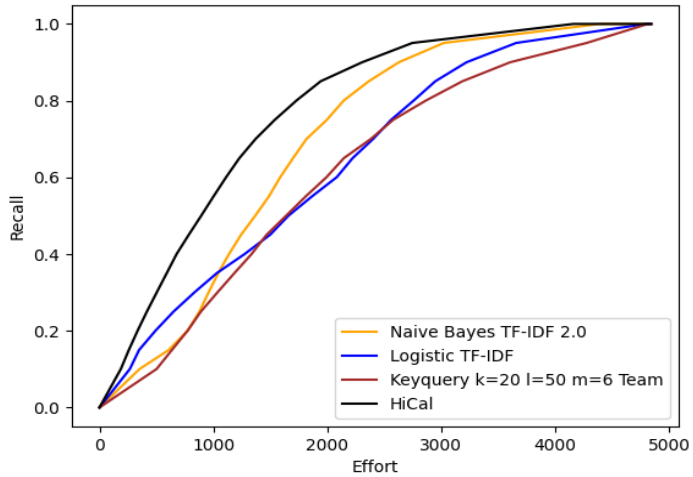


Evaluation der Ansätze auf den neueren Datensätzen

JKI Datensatz 2019



TREC Datensatz 2019



Zusammenfassung

- Keyqueries : Anfragen, die das gesuchte Dokument in den top-k Ergebnissen zurückliefern
- Human-In-The-Loop Framework gibt die Grundlage für das Extrahieren relevanter Dokumente
- Gute Ergebnisse bei System-unterstützenden systematischen Reviews durch Machine Learning Ansätze
- Keyquery-Ansatz erzielt nicht die Performance des State-of-the-Art Algorithmus

Zusammenfassung

- Keyqueries : Anfragen, die das gesuchte Dokument in den top-k Ergebnissen zurückliefern
- Human-In-The-Loop Framework gibt die Grundlage für das Extrahieren relevanter Dokumente
- Gute Ergebnisse bei System-unterstützenden systematischen Reviews durch Machine Learning Ansätze
- Keyquery-Ansatz erzielt nicht die Performance des State-of-the-Art Algorithmus

Vielen Dank für Ihre Aufmerksamkeit!

Quellenverzeichnis

- semanticscholar.org
- filmofilia.com
- Hagen et al., Supporting Scholarly Search with Keyqueries, 2016.
- Radlinski et al., how does clicktrough data reflect retrieval quality, 2008.
- Zhang et al., Increasing the Efficiency of high-recall Information Retrieval, 2019.
- Hagen, Vorlesung "Big Data Analytics", 2020.
- Cohen et al., Reducing Workload in Systematic Review Preparation Using, 2006.