

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?*

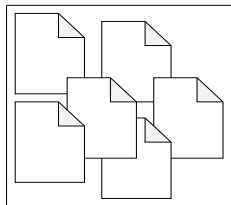
# Einleitung

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

Corpus



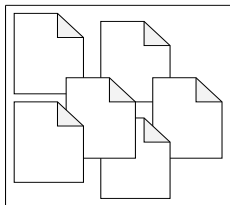
# Einleitung

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

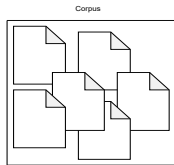
Corpus



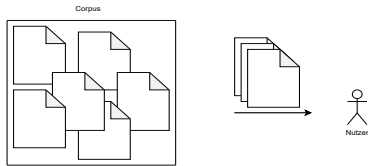
Total Recall:  
Extraktion aller  
relevanten Dokumente aus dem  
Corpus



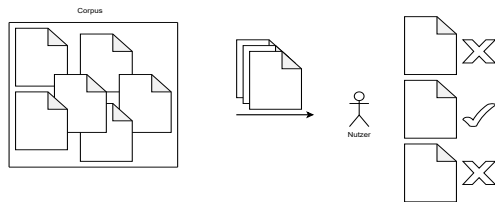
# Human-In-The-Loop zur Extraktion aller relevanter Dokumente



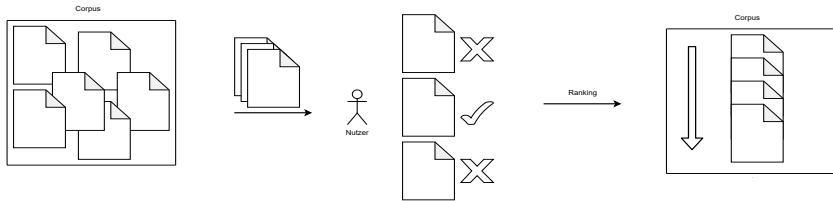
# Human-In-The-Loop zur Extraktion aller relevanter Dokumente



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

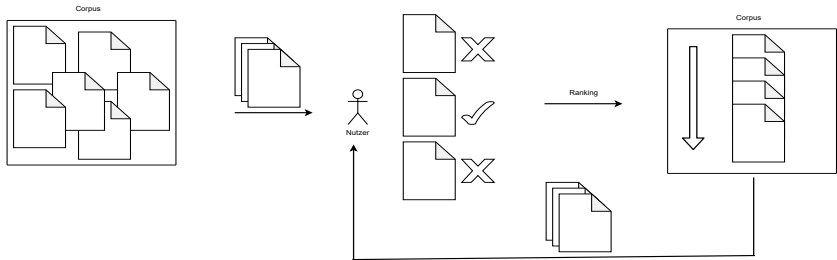


# Human-In-The-Loop zur Extraktion aller relevanter Dokumente



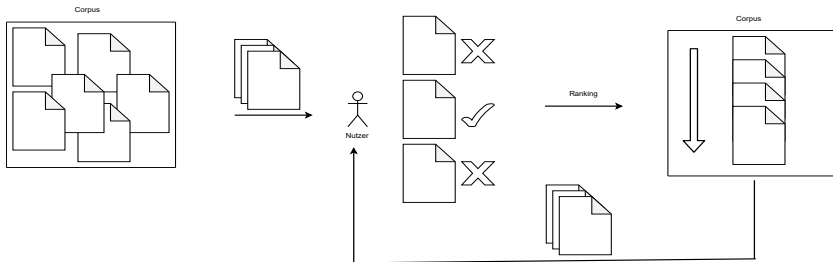


# Human-In-The-Loop zur Extraktion aller relevanter Dokumente



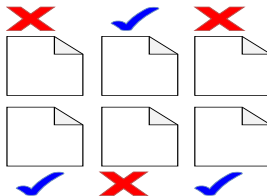
# Ranking mit Machine Learning

- 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:  $x = x_1 \dots x_n$



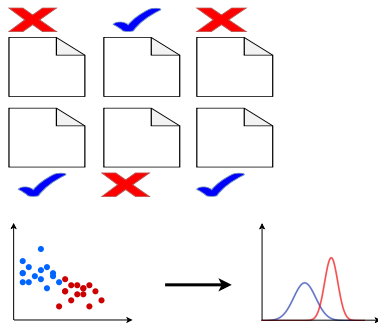
# Machine Learning - Beispiel: Naive Bayes

Trainingsdaten:  $x = x_1 \dots x_n$

Satz von Bayes:

$$P(K_i|x) = \frac{P(x|K_i) \cdot P(K_i)}{P(x)}$$

mit  $i \in \{\text{relevant}, \text{nichtrelevant}\}$



# Machine Learning - Beispiel: Naive Bayes

Trainingsdaten:  $x = x_1 \dots x_n$

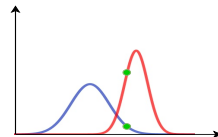
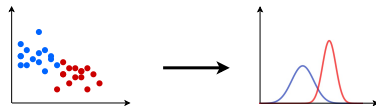
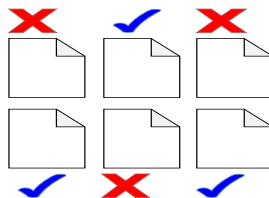
Satz von Bayes:

$$P(K_i|x) = \frac{P(x|K_i) \cdot P(K_i)}{P(x)}$$

mit  $i \in \{\text{relevant}, \text{nichtrelevant}\}$

Testdaten  $y = y_1 \dots y_k$

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 Höpfer](#) · [Anna Beyer](#) · +2 authors · [Beyers State](#) · Published in ICSE 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

[Zor-Sabaia](#) · [Matthias Beyer](#) · +1 author · [Matthias Höpfer](#) · Published in APS 2016 · Computer Science

In this paper we revisit the document clustering problem from an information retrieval perspective. The idea is to use queries as feedback 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 relevance 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

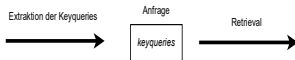
Matthias Höpfer, Anna Beyer · 12 authors · [arXiv:1606.08001](#) · Published in ICSE 2016 · Computer Science

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

Zor-Saba, Matthias Beyer · 11 authors · [arXiv:1606.08001](#) · Published in APS 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 relevance 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

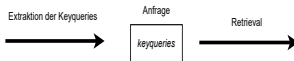
Matthias Hagen, Arno Renz, <1 author> Benno Stein, Published in ECDR 2016 • Computer Science

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

Tin Gollub, Matthias Hagen, <1 author> Matthias Hagen, Published in AIS 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 relevance search engine.



## 16 results for "keyqueries"

Fields of Study Date Range Has PDF Publication Type Author Journals & Conference

### Supporting Scholarly Search with Keyqueries

Matthias Hagen, Arno Renz, Tin Gollub, Benno Stein, Published in ECDR • 20 March 2016

TLDR We deal with a problem faced by scholars every day: identifying relevant papers on a given topic. Expand

11 PDF View via Publisher Save Alert Cite Research Feed

### Keyqueries for Clustering and Labeling

Tin Gollub, Matthias Hagen, Benno Stein, Published in AIS • 30 November 2016

TLDR In this paper we revisit the document clustering problem from an information retrieval perspective. Expand

2 PDF View via Publisher Save Alert Cite Research Feed

### Dynamic taxonomy composition via keyqueries

Tin Gollub, Michael Völkel, Matthias Hagen, Benno Stein, Published in IEEE/ACM Joint Conference on Digital Libraries • 8 September 2014

TLDR This paper presents an unsupervised framework for dynamic, subject-oriented taxonomy composition in digital libraries, which can naturally integrate existing library classification systems. Expand

11 PDF View on ACM Save Alert Cite Research Feed

### From keywords to keyqueries: content descriptors for the web

Tin Gollub, Matthias Hagen, Maximilian Michel, Benno Stein, Published in SIGR • 28 July 2013

TLDR We introduce the concept of **keyqueries** as dynamic content descriptors for documents and present an exhaustive search algorithm along with effective pruning strategies. Expand

19 PDF View on ACM Save Alert Cite Research Feed

### Webis at TREC 2020: Health Misinformation Track

Juraj Ročnik, A. Boudouk, <4 authors> Matthias Hagen • 2021

We give a brief overview of the Webis group's participation in the TREC 2020 Health Misinformation track, a baseline retrieval results of our search engine ChatNot (BM25F-based) are re-ranked in two. Expand

PDF Save Alert Cite Research Feed



# Keyquery - Eigenschaften

- mindestens  $\ell$  Ergebnisse

Relevantes Dokument:

"Supporting Scholarly Search with Keyqueries"

One result for "supporting scholarly search with keyqueries"

Fields of Study ▾ Date Range ▾ Publication Type ▾ Author ▾ Journals & Conferences

## Supporting Scholarly Search with Keyqueries

Matthias Hagen, Anna Beyer, Tim Gollub, Kristof Kormlovy, Benno Stein · Computer Science · ECIR · 20 March 2016

TLDR We deal with a problem faced by scholars every day: identifying relevant papers on a given topic.

Expand

11 PDF · View via Publisher Save Alert Cite Research Feed

# Keyquery - Eigenschaften

- mindestens  $\ell$  Ergebnisse

Relevantes Dokument:

"Supporting Scholarly Search with Keyqueries"

One result for "supporting scholarly search with keyqueries"

Fields of Study ▾ Date Range ▾ Publication Type ▾ Author ▾ Journals & Conferences ▾

## Supporting Scholarly Search with Keyqueries

Matthias Hagen, Anna Beyer, Tim Gollub, Kristof Kormiovy, Benno Stein · Computer Science · ECIR · 20 March 2016

**TLDR** We deal with a problem faced by scholars every day: identifying relevant papers on a given topic. [Expand](#)

66 11 PDF View via Publisher Save Alert Cite Research Feed

- relevantes Dokument in den top-k

Relevantes Dokument:

"Supporting Scholarly Search with Keyqueries"

About 8,630,000 results for "supporting search"

Fields of Study ▾ Date Range ▾ Has PDF ▾ Publication Type ▾ Author ▾ Journals & Conferences ▾

## Towards Supporting Search over Trending Events with Social Media

S. Kallam, M. Morris, J. Tarsen, Daniel J. Liebling, S. Dumala · Computer Science · ICWSM · 1 July 2013

**TLDR** We observe that interaction with content about trending events varies significantly with prior awareness of the event. [Expand](#)

66 44 PDF View Paper Save Alert Cite Research Feed

## Supporting Search-As-You-Type Using SQL in Databases

S. Li, Jintara Fero, Chet Li · Computer Science · IEEE Transactions on Knowledge and Data... · 1 February 2013

**TLDR** We study how to support search-as-you-type on data residing in a relational DBMS using the native query language, SQL. [Expand](#)

66 31 PDF View on IEEE Save Alert Cite Research Feed

# Keyquery - Eigenschaften

- mindestens  $\ell$  Ergebnisse

Relevantes Dokument:

"Supporting Scholarly Search with Keyqueries"

One result for "supporting scholarly search with keyqueries"

Fields of Study ▾ Date Range ▾ Publication Type ▾ Author ▾ Journals & Conferences ▾

## Supporting Scholarly Search with Keyqueries

Matthias Hagen, Anna Beyer, Tim Gollub, Kristof Kormlosy, Benno Stein · Computer Science · ECIR · 20 March 2016

**TLDR** We deal with a problem faced by scholars every day: identifying relevant papers on a given topic. [Expand](#)

66 11 · PDF · View via Publisher · Save · Alert · Cite · Research Feed

- relevantes Dokument in den top- $k$

Relevantes Dokument:

"Supporting Scholarly Search with Keyqueries"

About 8,630,000 results for "supporting search"

Fields of Study ▾ Date Range ▾ Has PDF ▾ Publication Type ▾ Author ▾ Journals & Conferences ▾

## Towards Supporting Search over Trending Events with Social Media

S. Kallam, M. Morris, J. Tarsen, Daniel J. Liebling, S. Dumala · Computer Science · ICWSM · 1 July 2013

**TLDR** We observe that interaction with content about trending events varies significantly with prior awareness of the event. [Expand](#)

66 44 · PDF · View Paper · Save · Alert · Cite · Research Feed

## Supporting Search-As-You-Type Using SQL in Databases

S. Li, Jintao Fens, Chet Liu · Computer Science · IEEE Transactions on Knowledge and Data... · 1 February 2013

**TLDR** We study how to support search-as-you-type on data residing in a relational DBMS using the native query language, SQL. [Expand](#)

66 31 · PDF · View on IEEE · Save · Alert · Cite · Research Feed

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

16 results for "keyqueries"

Fields of Study ▾ Date Range ▾ Has PDF ▾ Publication Type ▾ Author ▾ Journals & Conferences ▾

## Supporting Scholarly Search with Keyqueries

Matthias Hagen, Anna Beyer, Tim Gollub, Kristof Kormlosy, Benno Stein · Computer Science · ECIR · 20 March 2016

**TLDR** We deal with a problem faced by scholars every day: identifying relevant papers on a given topic. [Expand](#)

66 11 · PDF · View via Publisher · Save · Alert · Cite · Research Feed

## Keyqueries for Clustering and Labeling

Tim Gollub, Matthias Buse, Benno Stein, Matthias Hagen · Computer Science · AIRS · 30 November 2016

**TLDR** In this paper we revisit the document clustering problem from an information retrieval perspective. [Expand](#)

66 2 · PDF · View via Publisher · Save · Alert · Cite · Research Feed

# Keyquery - Interleaving Strategien

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

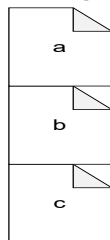
# Keyquery - Interleaving Strategien

- 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

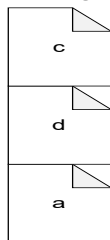
# Keyquery - Interleaving Strategien

- 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

Ranking A



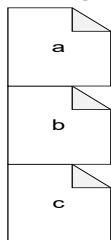
Ranking B



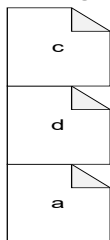
# Keyquery - Interleaving Strategien

- 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

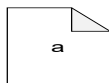
Ranking A



Ranking B



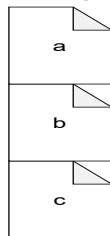
Balanced



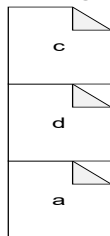
# Keyquery - Interleaving Strategien

- 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

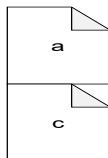
Ranking A



Ranking B



Balanced

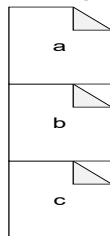




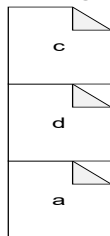
# Keyquery - Interleaving Strategien

- 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

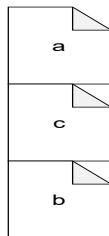
Ranking A



Ranking B



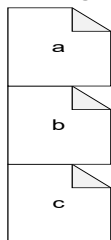
Balanced



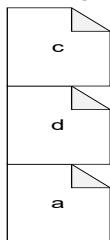
# Keyquery - Interleaving Strategien

- 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

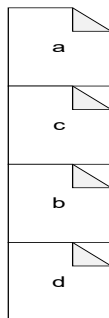
Ranking A



Ranking B



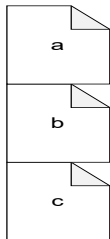
Balanced



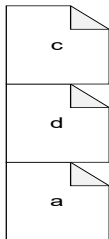
# Keyquery - Interleaving Strategien

- 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

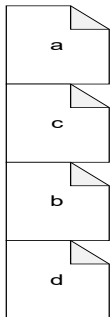
Ranking A



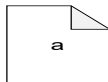
Ranking B



Balanced



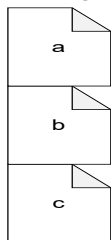
Teamdraft



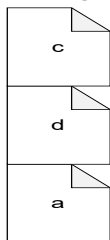
# Keyquery - Interleaving Strategien

- 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

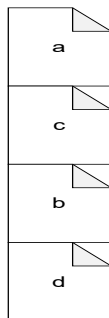
Ranking A



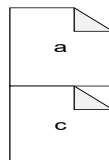
Ranking B



Balanced



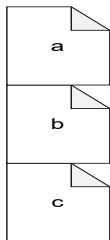
Teamdraft



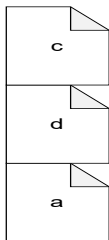
# Keyquery - Interleaving Strategien

- 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

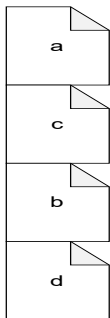
Ranking A



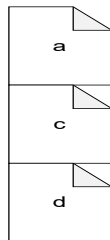
Ranking B



Balanced



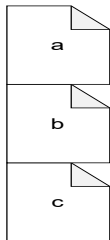
Teamdraft



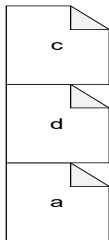
# Keyquery - Interleaving Strategien

- 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

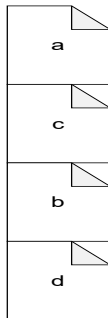
Ranking A



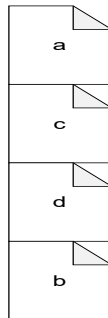
Ranking B



Balanced



Teamdraft



# 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

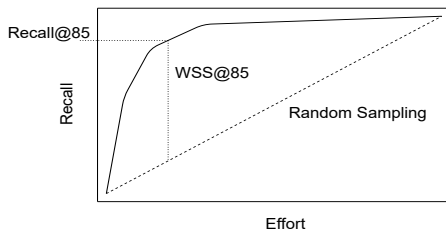
# Experimentaufbau

- 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

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



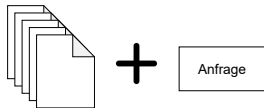
# Baseline - HiCal

- 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:  $k \leftarrow k + \frac{k+9}{10}$

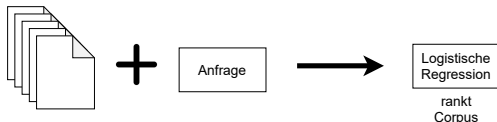
100 irrelevante Dokumente



# Baseline - HiCal

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

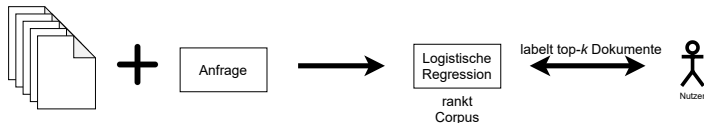
100 irrelevante Dokumente



# Baseline - HiCal

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

100 irrelevante Dokumente



# Methodologie der Machine-Learning Ansätze (1)

- Auswahl der Features

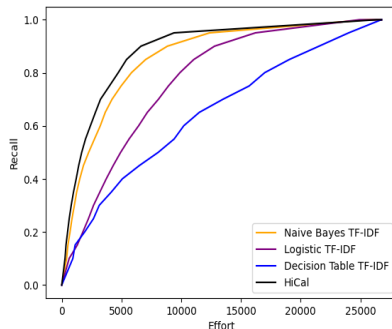
- ▶ Boolean word occurrence: Vorkommen eines bestimmten Terms im Dokument in Abhängigkeit aller Terme
- ▶ Term frequency: Anzahl eines Terms in einem Dokument
- ▶ Term frequency - inverse document frequency:  $\log \frac{N}{\sum_{D:t \in D} 1}$   
 $D$  : Dokument,  $N$  : Anzahl aller Dokumente



# Methodologie der Machine-Learning Ansätze (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



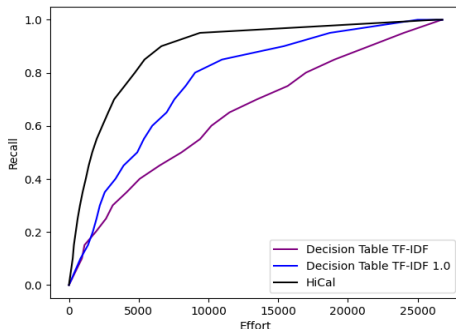
# Methodologie der Machine-Learning Ansätze (2)

- 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
- ▶  $\ell$  : Retrieval von mindestens  $\ell$  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
- ▶  $\ell$  : Retrieval von mindestens  $\ell$  Dokumenten
- ▶  $m$  : Anzahl der relevanten Dokumente in den top- $k$

Approach	$k$	$\ell$	$m$	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	<b>51.06</b>	<b>50.07</b>	37.69
Keyquery	50	50	34	20.24	11.66	9.68
Keyquery	50	100	25	23.15	20.72	11.91
Keyquery	100	100	10	50.97	50.01	<b>43.24</b>
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	<b>54.50</b>	<b>52.69</b>	<b>45.26</b>
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)

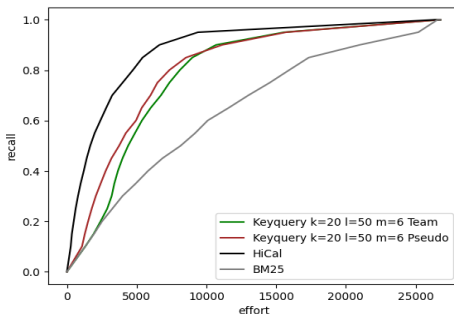
- Pseudo-Relevanz Feedback
  - ▶ Erhalten potenziell relevanter Dokumente durch Machine Learning Ansätze → Erstellen von Keyqueries
  - ▶ Vorschlag von 5 als relevant eingestuften Dokumenten durch jeden Ansatz
  - ▶ Benutzung eines vorgeschlagenen Dokuments, wenn es mit 80% Wahrscheinlichkeit relevant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



# Methodologie des Keyquery Ansatzes (3)

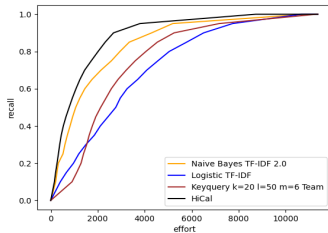
- Pseudo-Relevanz Feedback

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



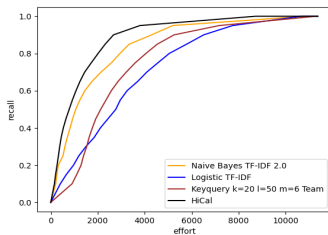
# Evaluation der Ansätze auf den neueren Datensätzen

## JKI Datensatz 2019

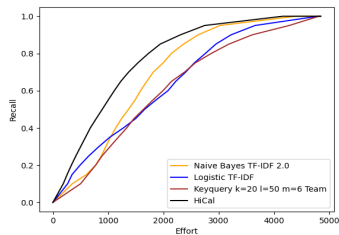


# Evaluation der Ansätze auf den neueren Datensätzen

## 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-unterstützten 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-unterstützten 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](https://www.semanticscholar.org)
- [filmofilia.com](https://www.filmofilia.com)
- Hagen et al., Supporting Scholarly Search with Keyqueries, 2016.
- Radlinski et al., how does clickthrough 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.