Explicit Semantic Analysis

Seminar Information Retrieval
im SoSe2012

Christian Eiserloh

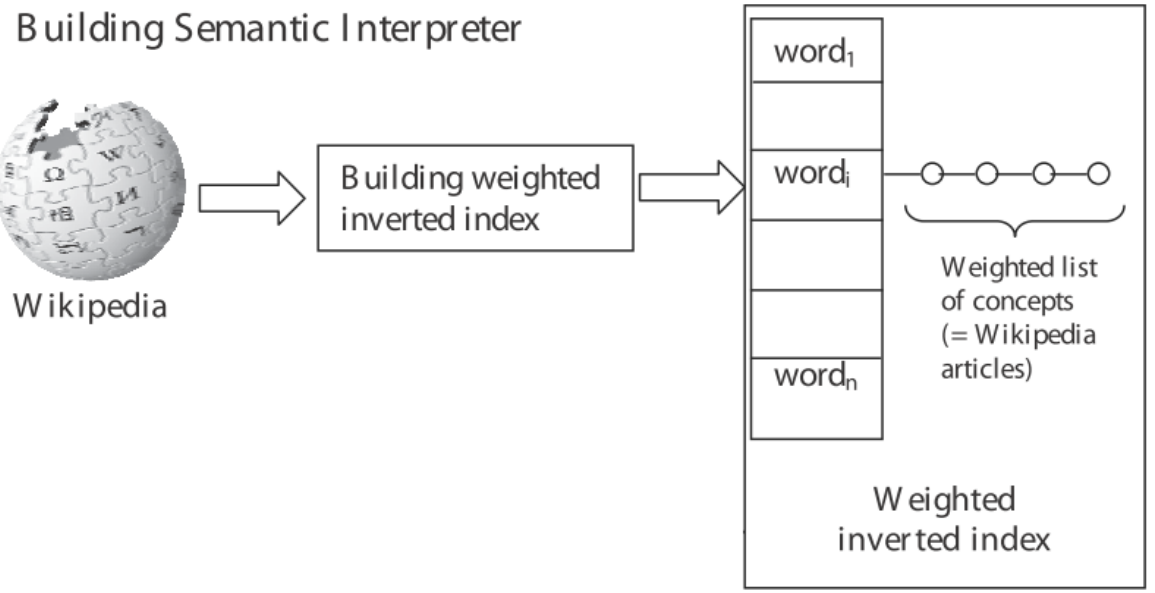
Inhalt

- Grundlegende Idee von ESA
- Konzeptbasiertes Information Retrieval mit ESA
  - ESA-basiertes Retrieval
  - Erweiterung um Selektion
  - Erweiterung um Fusion
- Weitergehende Selektionsstrategien

Grundlegende Idee von ESA

- Verfahren zum Ermitteln semantischer Verwandtschaft
  - Nicht auf Wort-Ebene, sondern auf der Ebene von Konzepten
- Repräsentation natürlichsprachlicher Texte:
  - Vektorbasiert
  - In hochdimensionalem Raum von Konzepten
- Wikipedia-basiertes ESA:
  - Konzepte aus Wikipedia abgeleitet
  - Wikipedia-Artikel = Konzepte
  - natürliche, „explizite“ Konzepte, von Menschen definiert

Inverted Index

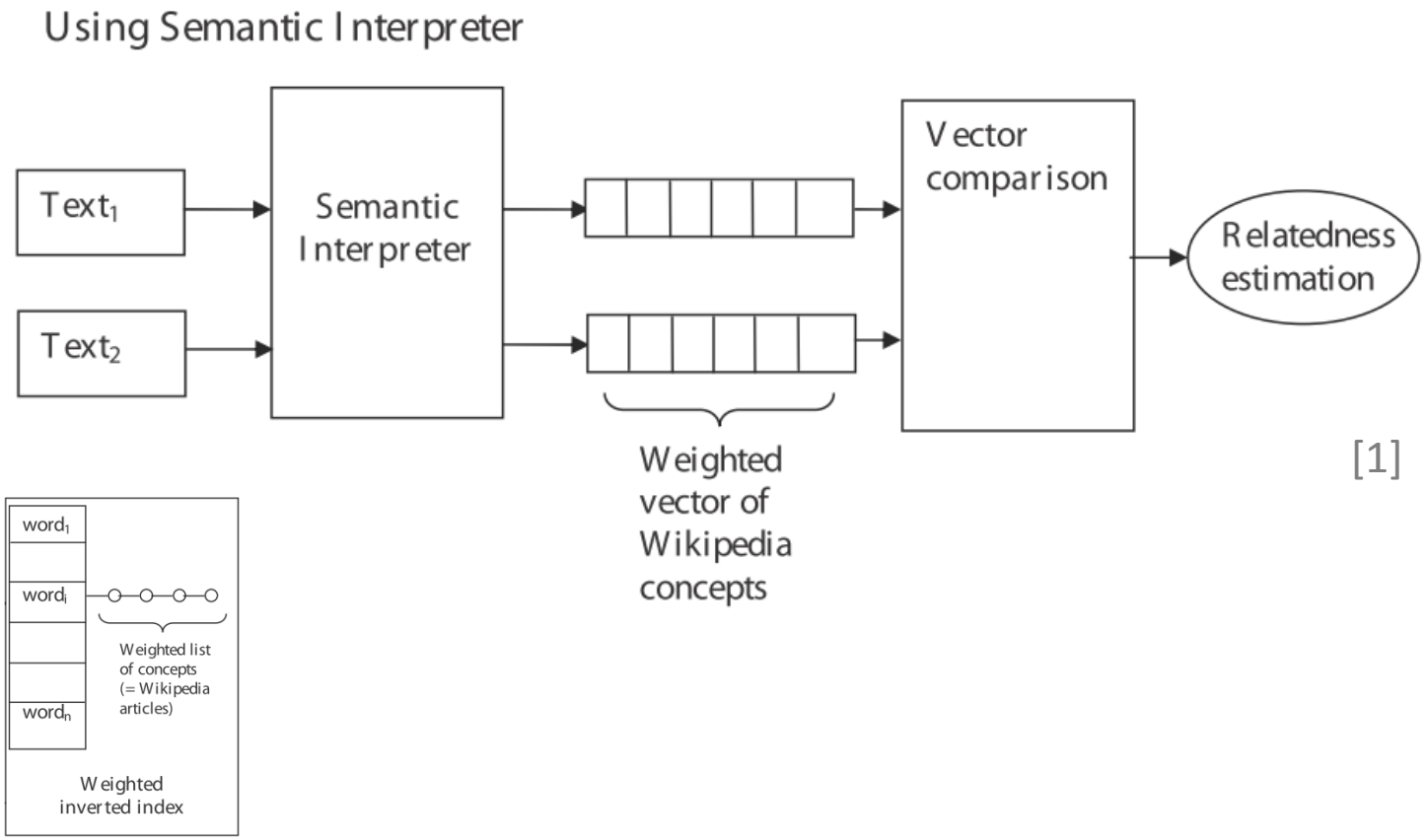


[1]

Inverted Index:

- Bildet jedes Wort auf eine (gewichtete) Liste von Konzepten (Wikipedia Artikeln) ab, in denen das Wort enthalten ist

Konzept-Vektor



[1]

Motivation für Konzept-basiertes IR

- Information Retrieval:
  - Zu einer Nutzeranfrage die relevantesten Dokumente aus einem Korpus zurückgeben
- Herkömmliche Verfahren:
  - Schlüsselwort-basiert, Bag-of-Words (BOW)-Repräsentation
  - Leiden unter sog. Vokabelproblem

Vokabelproblem

- Synonymie-Problem:
  - Nutzer verwendet andere Schlüsselwörter als Autoren
  - Synonyme werden nicht erkannt

    → Geringerer Recall

Vokabelproblem

- Synonymie-Problem:
  - Nutzer verwendet andere Schlüsselwörter als Autoren
  - Synonyme werden nicht erkannt
    → Geringerer Recall
- Polysemie-Problem:
  - Schlüsselwörter werden in verschiedenem Kontext verwendet
  - Kontext wird nicht beachtet
  - Dokumente werden fälschlicherweise als relevant zurückgegeben
    → Geringere Precision

Motivation für Konzept-basiertes IR

- Konzept-basiertes IR
  - Semantische Konzepte anstelle von (oder zusätzlich zu) Schlüsselwörtern

    → weniger abhängig von einzelnen Termen
  - Findet auch relevante Dokumente, die keine übereinstimmenden Terme mit der Anfrage haben
  - Erkennt nicht-relevante Dokumente auch dann als solche, wenn gleiche Wörter wie in der Anfrage vorkommen

ESA-basiertes Retrieval

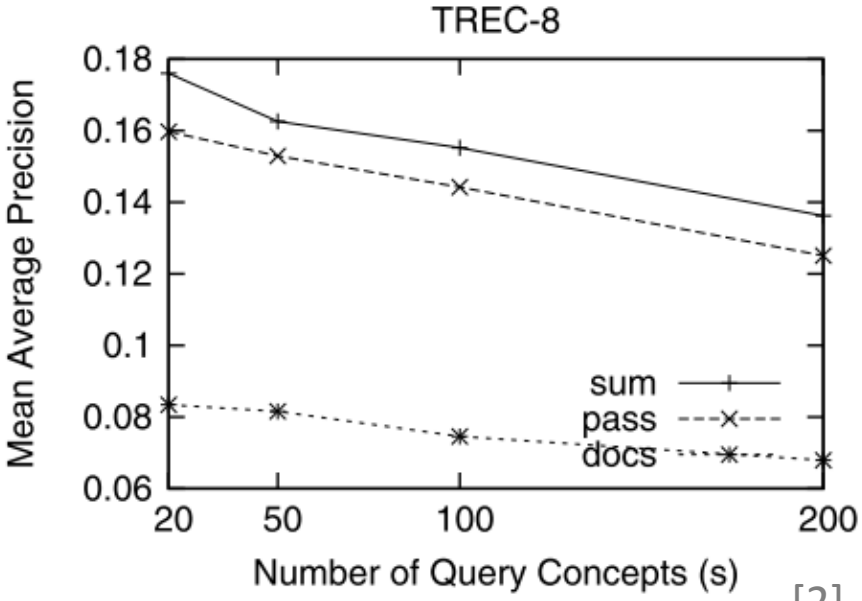
- Indexierung:
  - Inverted Index
  - Beschränkung auf Konzepte mit höchsten Gewichten (Cutoff)
  - Aufteilung langer Dokumente in überlappende Passagen fester Länge
    - Passagen erhalten jeweils eigenen Konzept-Vektor

ESA-basiertes Retrieval

- Indexierung:
  - Inverted Index
  - Beschränkung auf Konzepte mit höchsten Gewichten (Cutoff)
  - Aufteilung langer Dokumente in überlappende Passagen fester Länge
    - Passagen erhalten jeweils eigenen Konzept-Vektor
- Retrieval Algorithmus:
  - ESA Konzept-Vektor zu gegebener Anfrage
    - mit Cutoff → nur Konzepte mit höchsten Gewichten
  - Vergleich mit Konzept-Vektoren der Dokumente bzw. Passagen

ESA-basiertes Retrieval

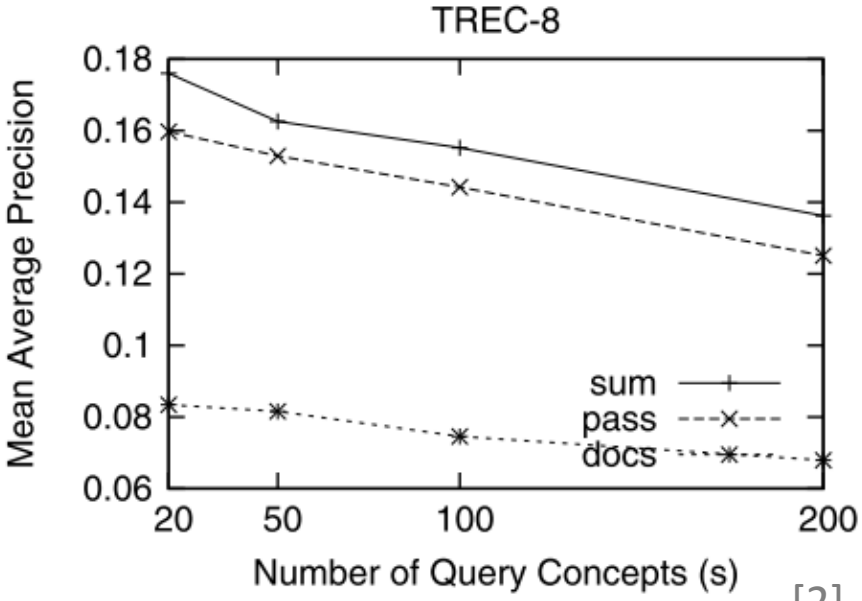
- Experiment:
  - Kurze Anfragen (1-3 Wörter)
  - Passagen von 50 Wörtern
  - Cutoff s = 20, 50, 100, 200



[2]

ESA-basiertes Retrieval

- Experiment:
  - Kurze Anfragen (1-3 Wörter)
  - Passagen von 50 Wörtern
  - Cutoff s = 20, 50, 100, 200



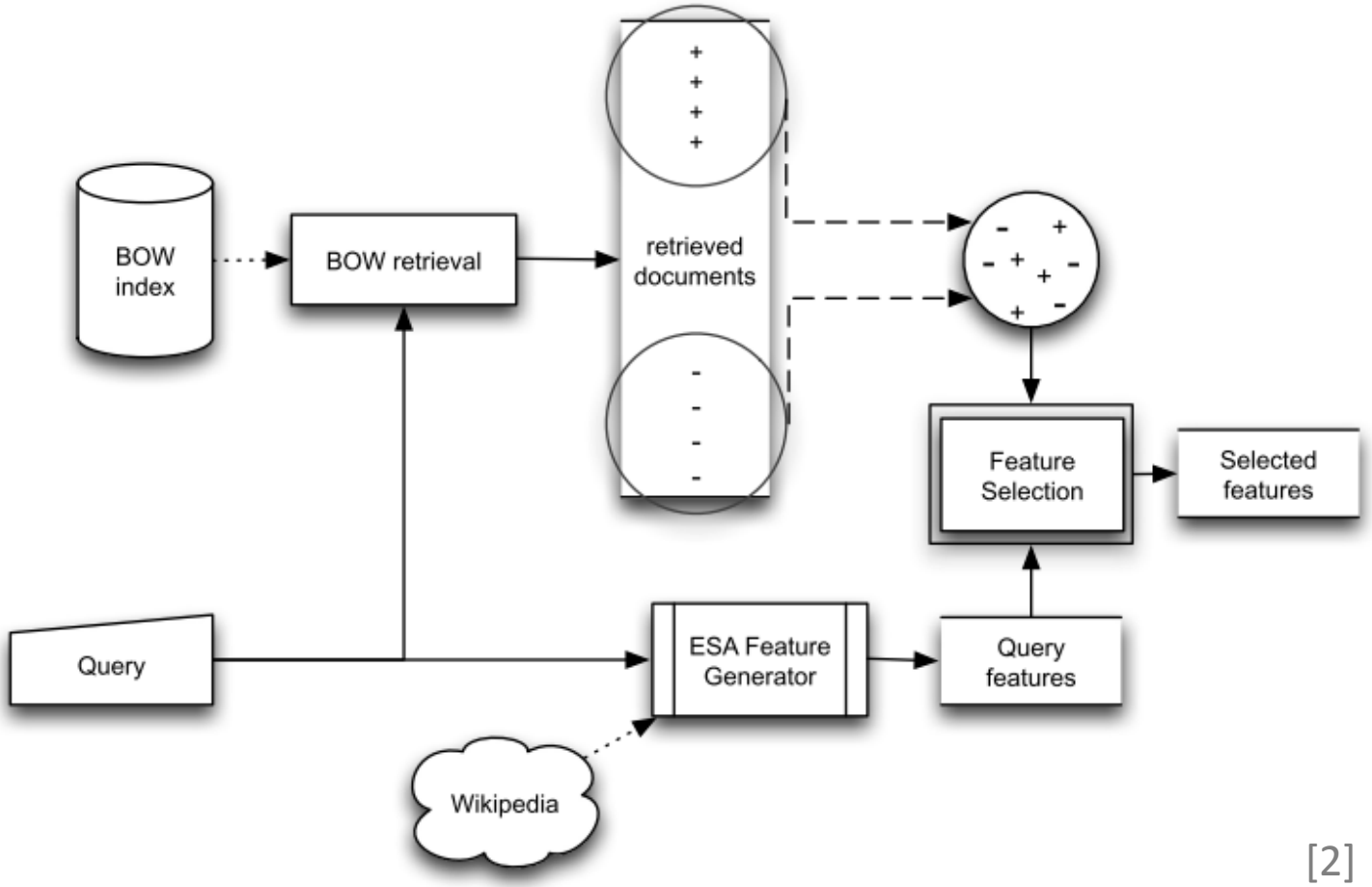
[2]

- Auswertung:
  - Höherer Cutoff bringt keine Verbesserung
    → initialer Anfrage-Vektor zu ungenau
    → weitergehende Auswahl der Anfrage-Konzepte notwendig

Selektives ESA-basiertes Retrieval

- Ziel: Optimieren der Anfrage
- Feature Selection durch Pseudo-Relevance Feedback
  - Nützlichkeit von Features (hier: Konzepten) wird bewertet
  - Umformulierung der Anfrage → modifizierter Anfrage-Vektor

Selektives ESA-basiertes Retrieval



[2]

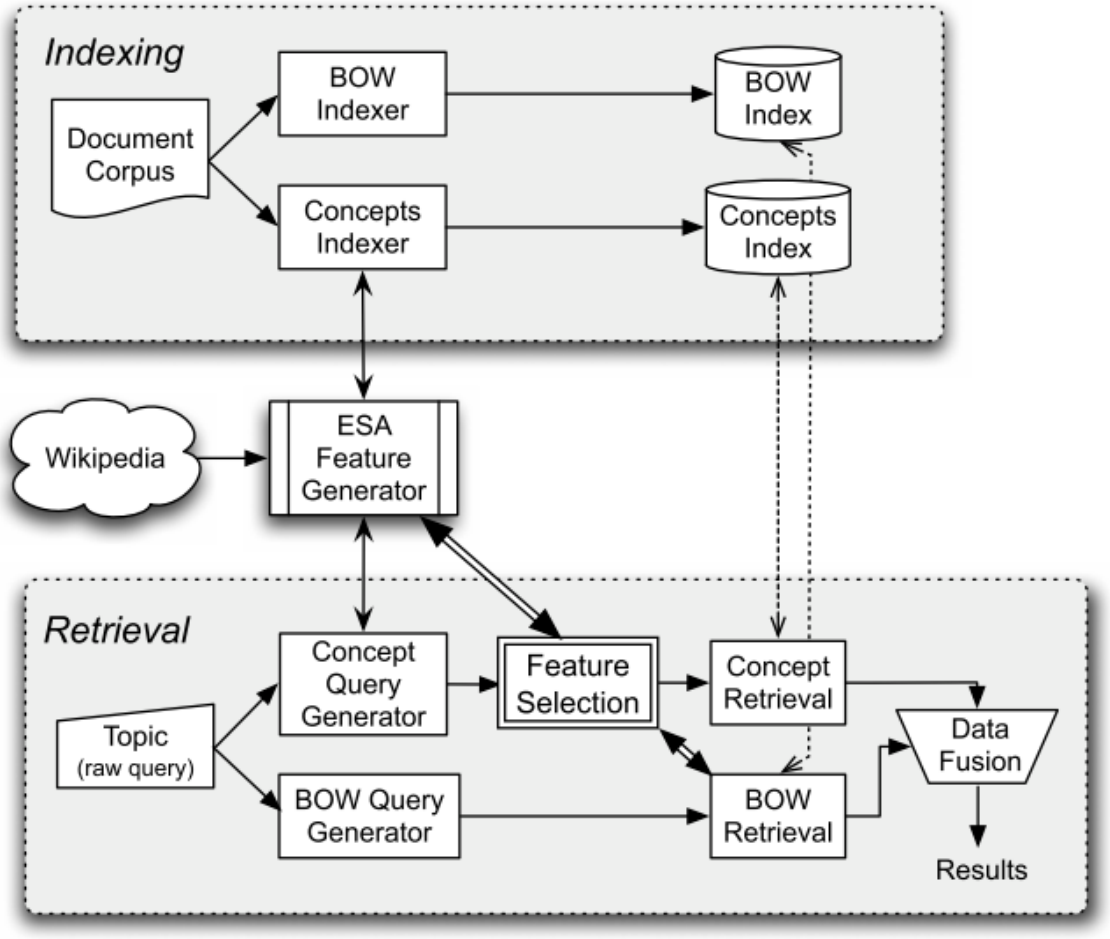
Selektives ESA-basiertes Retrieval

- Feature Selection Methoden:
  - FS mit Information Gain (IG)
  - FS mit inkrementellem IG (IIG)
  - FS mit einem Rocchio-Vektor (RV Methode)
- Auswertung:
  - Verbesserungen um bis zu 40%
  - Immer noch etwas schlechter als BOW-Retrieval
    - ESA: MAP von 0.1760
    - BOW: MAP von 0.2481

Fusioniertes selektives ESA-basiertes IR

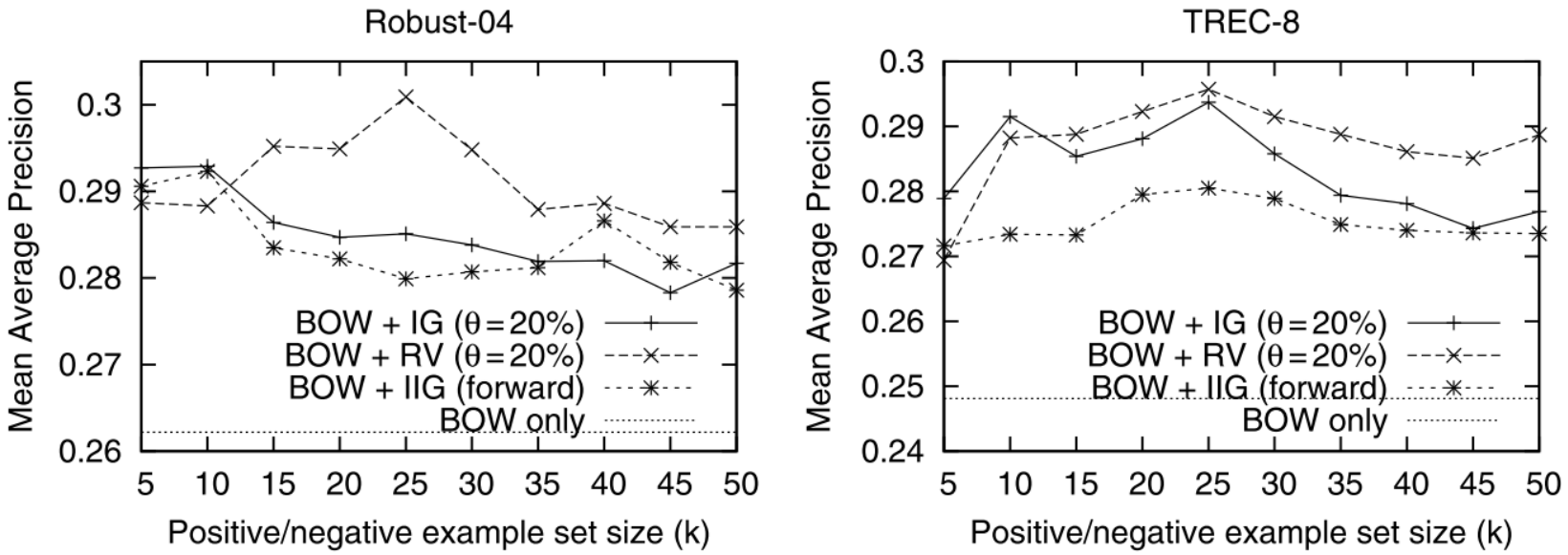
- Fusion: Kombinieren verschiedener Retrieval Methoden
  - Führt zu besseren Endergebnissen
  - Umso effektiver, je unterschiedlicher die Retrieval Methoden sind
- Das MORAG System:
  - ESA-basiertes Retrieval mit Feature Selection und Fusion
  - Fusion von BOW-Retrieval und ESA-basiertem Retrieval

MORAG


Indexing
Document Corpus
BOW Indexer
BOW Index
Concepts Indexer
Concepts Index
Wikipedia
ESA Feature Generator
Retrieval
Topic (raw query)
Concept Query Generator
Feature Selection
Concept Retrieval
Data Fusion
BOW Query Generator
BOW Retrieval
Results

[2]

MORAG



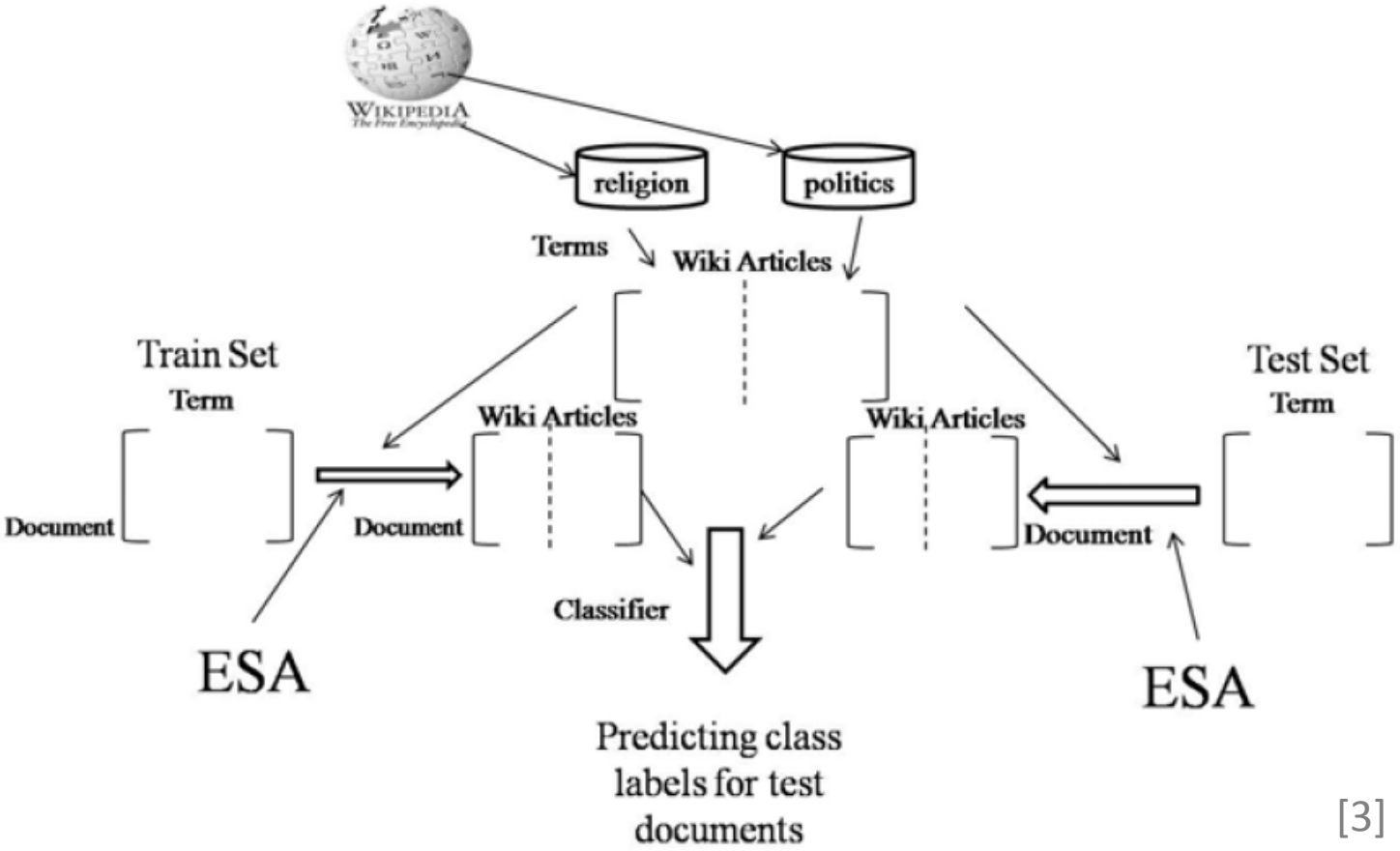
[2]

- Auswertung:
  - Beeindruckende Verbesserung zum BOW-Retrieval
  - ESA einzeln schwach, mit Fusion deutlich besser

Weitergehende Selektionsstrategien

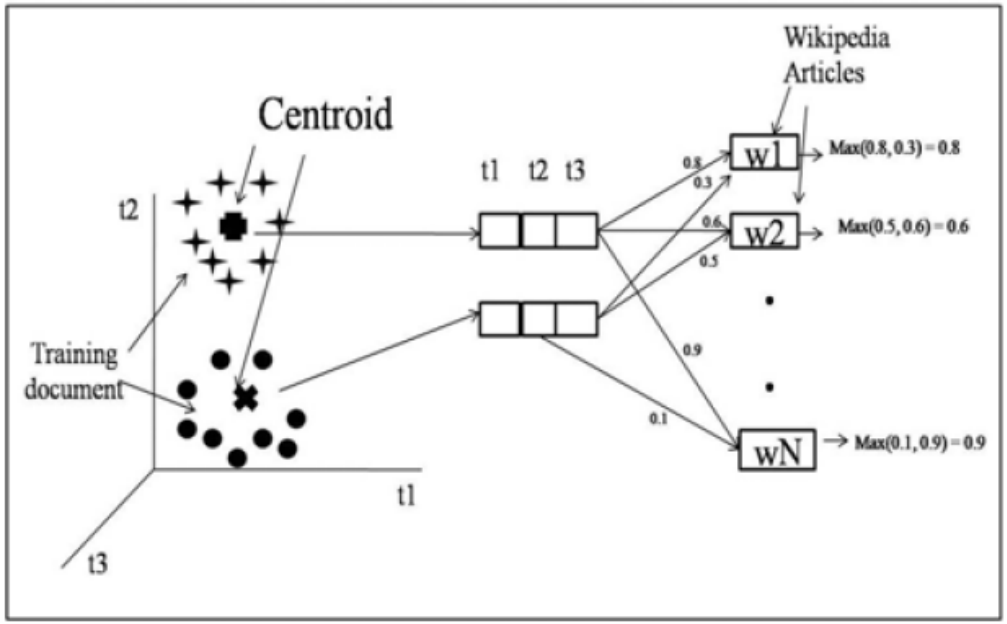
- Verbesserungspotenzial in der FS Komponente von MORAG
  - Besseres FS würde direkt zu besseren Ergebnissen führen
- Ziel: Möglichst wenige, semantisch wertvolle Features auswählen

ESA für Text-Klassifikation



[3]

Schwerpunkt Strategie



Step 1: Compute the centroids of training documents in each class.

Step 2: Compute cosine similarity between each Wikipedia article and the centroid of each class.

Step 3: Sort articles in the descending order of the maximum cosine similarity they have with any class centroid, and select the top n articles from each class.

[3]

- Idee: Artikel prototypisch für Kategorie
- Nachteile:
  - Kategorien „verhungern“
  - Wikipedia Artikel prototypisch für mehr als eine Klasse
  - Schwerpunkte nicht repräsentativ

k-nächste Nachbarn Strategie

- Für jede Klasse:
  - Diejenigen Wikipedia Artikel mit der höchsten Kosinus-Ähnlichkeit zu den 3 nächsten Trainingsdokumenten dieser Klasse
- Lokale Nachbarschaft statt Nähe zum Schwerpunkt
  - Führt eher zur korrekten Kategorie
  - Behebt den entscheidenden Nachteil der Schwerpunkt Strategie

Wahrscheinlichkeitsbasierte Strategie

- Ermittelt relative Wichtigkeit eines Artikels für die Klasse
- Wahrscheinlichkeitsberechnung auf Trainingsdaten
  - Klasse *c*, gegebener Wikipedia Artikel *wk*
  - A-posteriori Wahrscheinlichkeit P(*c*|*wk*)
  - Klasse *c* mit höchstem Wahrscheinlichkeitswert wird dem Artikel zugeordnet
  - Die besten Artikel jeder Kategorie werden ausgewählt

Erweitertes ESA

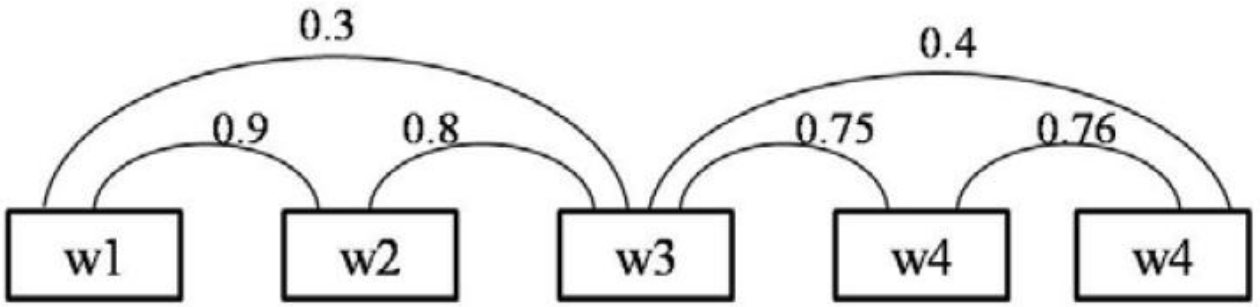
- Ziel: Unterscheidungsfähigkeit von Wikipedia Seiten ermitteln
- Mix aus Wörtern und Konzepten
- Motivation: Wörter nicht verlieren, die gut zwischen Klassen unterscheiden
- Bewertet die Nützlichkeit von Features
  - auf Wort-Ebene
  - auf Konzept-Ebene

Semantische Verwandtschaft

- ESA: Konzepte (Wikipedia Artikel) unabhängig zueinander
  - Nicht wirklich der Fall
- Kann man Retrieval verbessern, in dem man semantische Verwandtschaft modelliert?
- Wie baut man die Ähnlichkeit von Wikipedia Artikeln in die Repräsentation ein?
  - Ansatz: Case Retrieval Network (CRN)

Case Retrieval Network

- Paarweise Konzept-Ähnlichkeit → Ähnlichkeitsbögen
- Dokument als Konzept-Vektor
  - Unabhängig von Konzept: Komponente ist 0
  - Abhängig von Konzept: Komponente ist 1
- Spreading Activation: Relevante Konzepte „aktivieren“ ähnliche Konzepte
- Bsp.: Dokument D={1,1,0,0,0} → D‘={1.9,1.9.,1.1,0,0}



[3]

Auswertung

- Selektionsstrategien bewirken deutlich bessere Ergebnisse
  - Schwerpunkt Strategie insgesamt am besten
  - Lokale Modelle wie kNN in komplexen Themengebieten besonders gut
- Erweitertes ESA besser als ESA
- Modellierung von Ähnlichkeit nicht überzeugend

Zusammenfassung

- MORAG: ESA-Retrieval mit Feature Selection und Fusion
  - Verbesserungen gegenüber herkömmlichen Verfahren
  - Potenzial für weitere Verbesserungen
- ESA als Fortschritt im IR
- Verschiebung im IR Paradigma
  - Schlüsselwort-basiert → Konzept-basiert

Quellen

(1) Gabrilovich, E. & Markovitch, S. Veloso, M. M. (Ed.) Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007, 2007, 1606-1611

(2) Egozi, O.; Markovitch, S. & Gabrilovich, E. Concept-Based Information Retrieval Using Explicit Semantic Analysis ACM Trans. Inf. Syst., ACM, 2011, 29, 8:1-8:34

(3) Patelia, A.; Chakraborti, S. & Wiratunga, N. Ram, A. & Wiratunga, N. (Eds.) Selective Integration of Background Knowledge in TCBR Systems Case-Based Reasoning Research and Development, Springer Berlin / Heidelberg, 2011, 6880, 196-210

(4) Radinsky, K.; Agichtein, E.; Gabrilovich, E. & Markovitch, S. A word at a time: computing word relatedness using temporal semantic analysis Proceedings of the 20th interna-tional conference on World wide web, ACM, 2011, 337-346

(5) Gottron, T.; Anderka, M. & Stein, B. Insights into Explicit Semantic Analysis CIKM'11: Proceedings of 20th ACM Conference on Information and Knowledge Management, 2011, 1961-1964