Maschinelle Erschließungsverfahren versus Volltextsuche

Christian Wartena

Hochschule Hannover

10. Oktober 2019

Einleitung

Christian Wartena

1989 - 1994	Computerlinguistik Katholieke Universiteit Nijmegen, Niederlande
1994 – 2000	Universität Potsdam
2000 – 2005	Lingenio GmbH, Heidelberg (Maschinelle Übersetzung)
2005 – 2011	Novay, Enschede, The Netherlands (Research and consultancy)
2011 –	Hochschule Hannover (Sprach- und Wissensverarbeitung)

Bibliothek

- Kein Bibliothekar, aber
  - Projekte mit Bibliotheken
  - Lehre im Studiengang für Bibliothekswesen
  - Glaube, das die Informatik etwas von LIS lernen kann

Outline

Einleitung

Maschinelle Erschließung vs. Volltextsuche

- Maschinelle vs. intellektuelle Erschließung wurde ausreichend diskutiert.
- Maschine vs. Maschine
  - Sachliche, emotionslose Diskussion möglich!

Äpfel und Birnen

- Wie vergleichen wir?
- Wozu macht man das eigentlich?

Maschinelle vs. Intellektuelle Erschließung

- Direkter Vergleich
- Häufige Annahme:
  Mensch ist 100% korrekt, Maschine erreicht xx %

Einleitung

Maschinelle Erschließung vs. Volltextsuche

- Maschinelle vs. intellektuelle Erschließung wurde ausreichend diskutiert.
- Maschine vs. Maschine
  - Sachliche, emotionslose Diskussion möglich!

Äpfel und Birnen

- Wie vergleichen wir?
- Wozu macht man das eigentlich?

Maschinelle vs. Intellektuelle Erschließung

- Direkter Vergleich
- Häufige Annahme:
  Mensch ist 100% korrekt, Maschine erreicht xx %

Einleitung

Maschinelle Erschließung vs. Volltextsuche

- Maschinelle vs. intellektuelle Erschließung wurde ausreichend diskutiert.
- Maschine vs. Maschine
  - Sachliche, emotionslose Diskussion möglich!

Äpfel und Birnen

- Wie vergleichen wir?
- Wozu macht man das eigentlich?

Maschinelle vs. Intellektuelle Erschließung

- Direkter Vergleich
- Häufige Annahme:
  Mensch ist 100% korrekt, Maschine erreicht xx %

Ziele

Inhaltserschließung

1. Kurze Zusammenfassung
2. Ermöglicht Schlagwortsuche
   - Vor allem für Boolsches Retrieval

Volltextindex

- Ermöglicht Volltextsuche
  - Vor allem für probabilistisches Retrieval

Retrieval

Boolsches Retrieval

- Volltextindex für boolsches Retrieval ist ziemlich daneben

Probabilistisches Retrieval

- Probabilistisches Retrieval mit Schlagwörtern/Notationen/usw. ???
  - Theoretisch sinnlos, aber....

Exkurs

Ist Relevanz binär oder gibt es eine Skala der Relevanz?

- Der Nutzer muss entscheiden, ob er ein Ergebnis „anklickt“ oder nicht, auch in neueren Systemen
- Der Kunde kauft oder er kauft nicht. Dazwischen gibt es nichts.

Aber...

- Nicht jedes relevante Dokument ist gleich *nützlich*

Relevanzurteile

von Laien

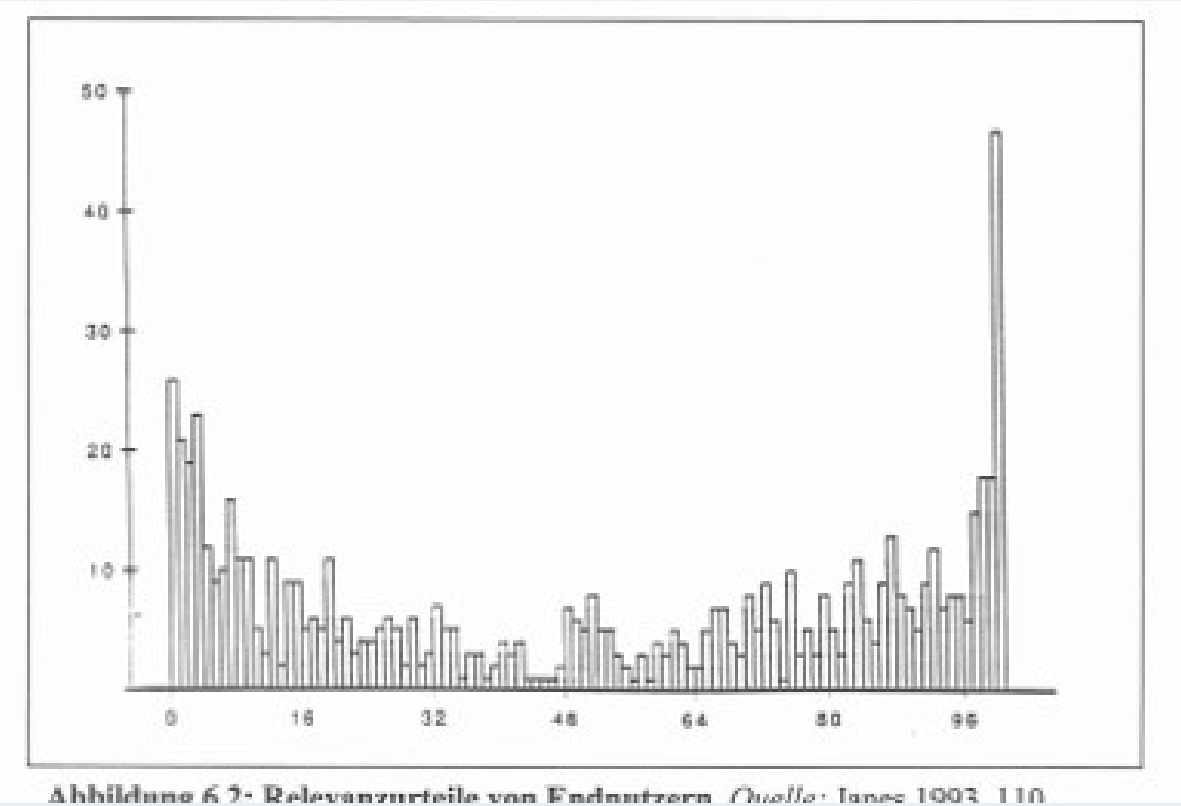


Abbildung 6.2: Relevanzurteile von Endnutzern *Quelle:* Janes 1993, 110

Relevanzurteile

von Information Professionals

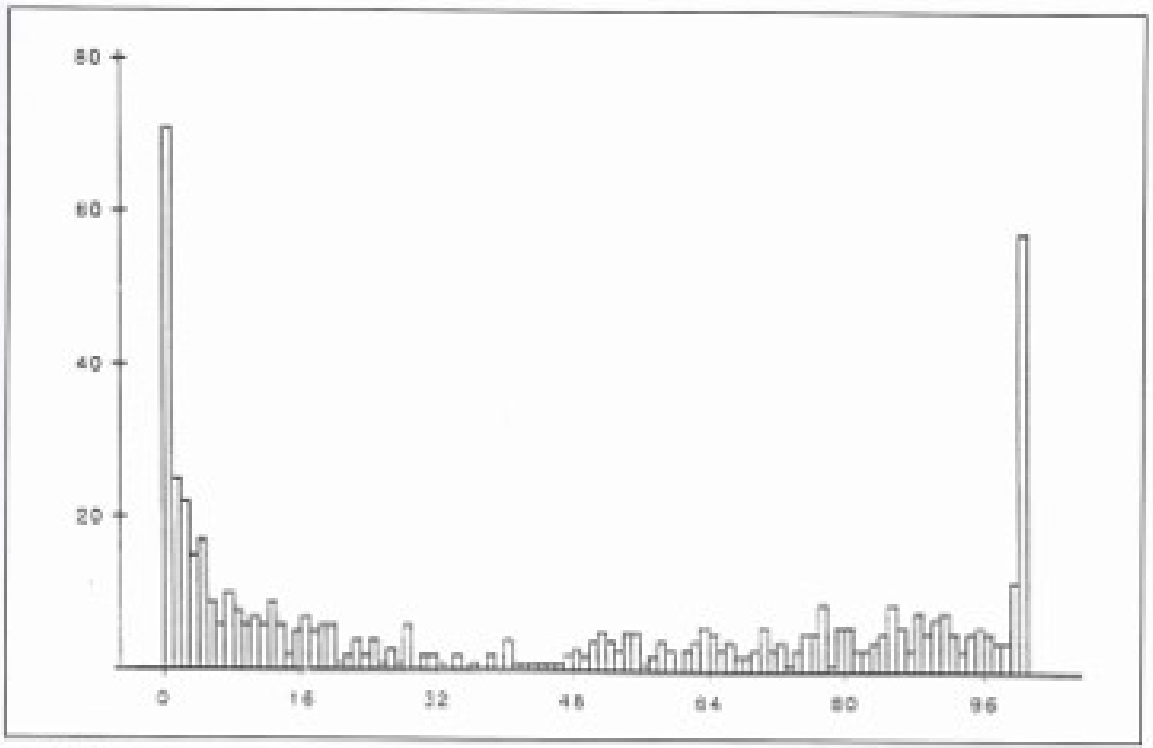


Abbildung 6.3: Relevanzurteile von Information Professionals. *Quelle:* Janes 1993, 111.

Probabilistisches Retrieval

Ranking - Theorie

- Probabilistisches Retrieval setzt binäre Relevanz voraus.
- Sortiert wird nach der Wahrscheinlichkeit auf Relevanz.
- Wenn Relevanz 100% sicher ist, kann nicht sortiert werden.
  - Über verschiedene Hintertürchen, kann man hier irgendwie auch ein Ranking erreichen und einzelne Aspekte der Volltextindexierung einschleusen.

Ranking - Praxis

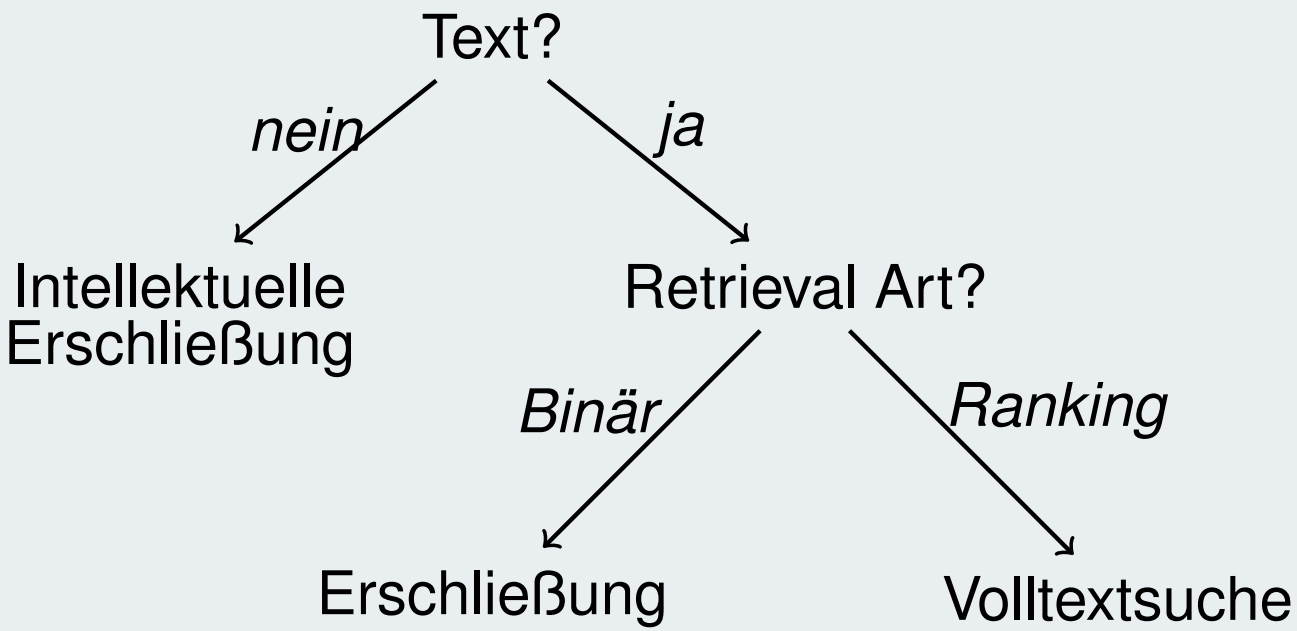
- Hohe Wahrscheinlichkeit auf Relevanz korreliert mit hoher Pertinenz/Nützlichkeit

Learning to Rank

- Jedes (konsistente) Ranking kann aus Beispielen gelernt werden
- Ranking nach Pertinenz und vielen anderen Kriterien (theoretisch) möglich

Übersicht

Volltextsuche oder maschinelle Erschließung



Übersicht

Boolsches Retrieval

- Wenn Boolsches Retrieval erwünscht ist, ist Volltextsuche nicht sinnvoll.
- Wer möchte heute wirklich noch Boolsches Retrieval

Ranking

- Welche Vorteile hat Erschließung gegenüber Volltextsuche?
- Können wir Aspekte der Erschließung in der Volltextsuche übernehmen?
- Warum kein Ranking auf Grund einer Verbalerschließung?

Vorteile der Inhaltserschließung

Vorteile

1. Trennung von Haupt- und Nebensache
2. Terminologische Kontrolle
3. Facettierung
4. Kontrolle und Korrektur
5. (Vorschau Funktion)

Vorteile der Inhaltserschließung

Trennung von Haupt- und Nebensache

- Kein Vorteil!
- Automatisch Erschließung macht letztendlich das gleiche wie Volltextindexierung
  - Schlechtes Ranking $\leftrightarrow$ Komplette Unterdrückung
  - (Relevante) Nebensachen sind nicht mehr auffindbar

Vorteile der Inhaltserschließung

Terminologische Kontrolle

- Kann man bei 210 000 GND Terme oder 414 000 LCSH überhaupt noch von term. Kontrolle sprechen?
- Mischung der Vokabulare durch Fremddatenübernahme und in Verbundkatalogen

Example

- Womit sucht man in B3Kat?
  - GND: *Deutsche Frage*
  - Friedrich Ebert Stiftung: *Deutschlandfrage*
  - RVK: *ML 1800*
  - Institut für Zeitgeschichte: *x 64* UND *x 499*

Vorteile der Inhaltserschließung

Example

- Homonyme
- Suche mit Schlagwort (erweiterte Suche!) *Leiter* in Gateway Bayern gibt u.a.
  - *Behördenleitung mit demokratisiertem Führungsstil*
  - *Leiter und pädagogischer Mitarbeiter an Volkshochschulen*
  - *Matheus Leiter, Bildhauer zu Meran (1682)*
  - *Langzeitverhalten elektrotechnischer Verbindungen unter Berücksichtigung des Kriechens der Leitermaterialien*
  - *Erhöhung der Arbeitssicherheit in der chinesischen Wirtschaft am Beispiel des Arbeitens auf Leitern*
- Vorteile der Erschließung gehen irgendwo auf dem Weg vom Bibliothekar zum Endnutzer verloren.

Studentische Arbeit zum B3Kat

Statistik über 6 Mio. Titel aus XML-Dump

- Großteil überhaupt nicht erschlossen
- Formale Metadaten als Schlagwort (Zeitschrift, Wörterbuch, . . .)
- Oft nur Titelwörter als Schlagwort

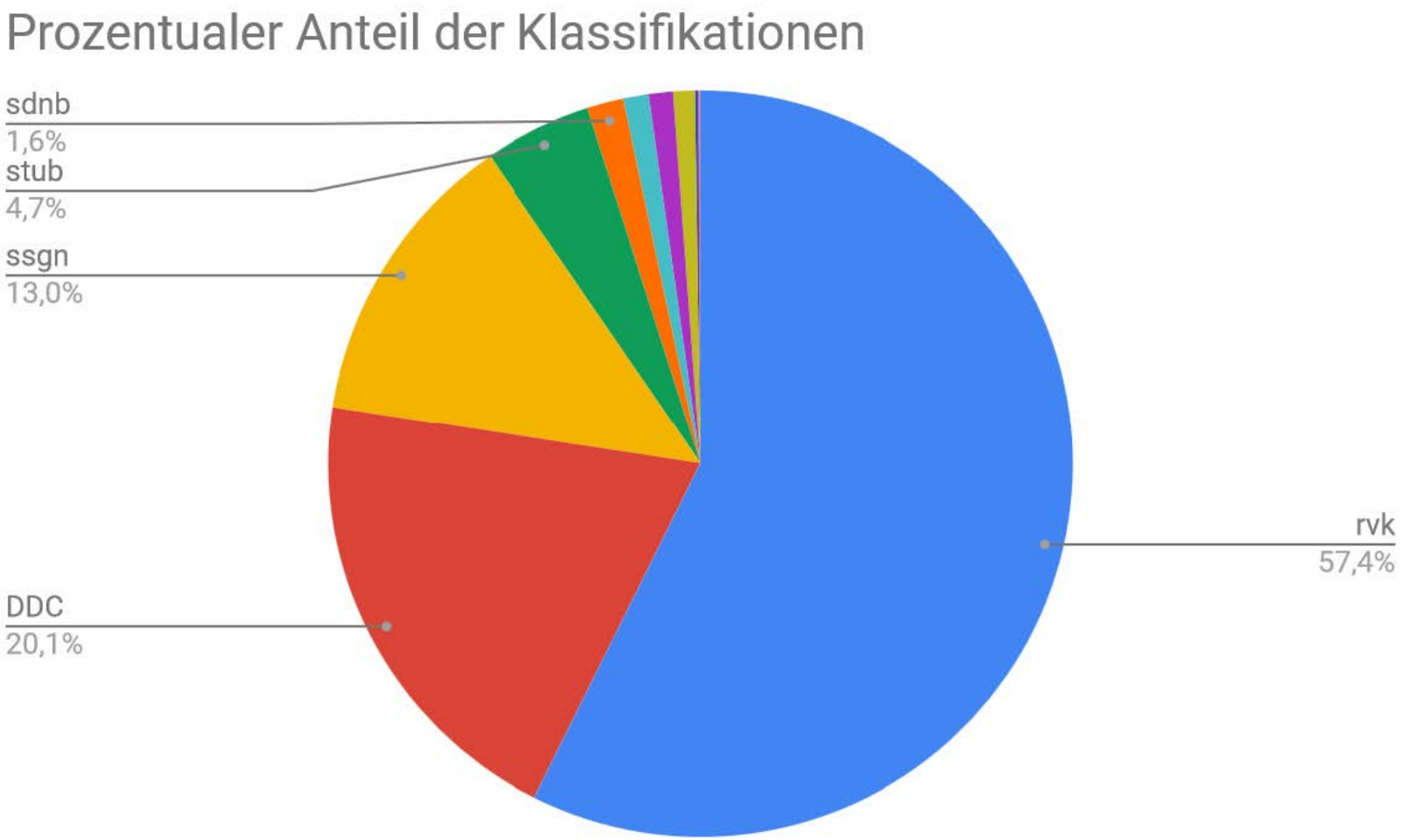
Klassen

- Etwa 50% klassiert
- Etwa 40% davon nach mehr als einer Klassifikation
- 12 Klassifikationen

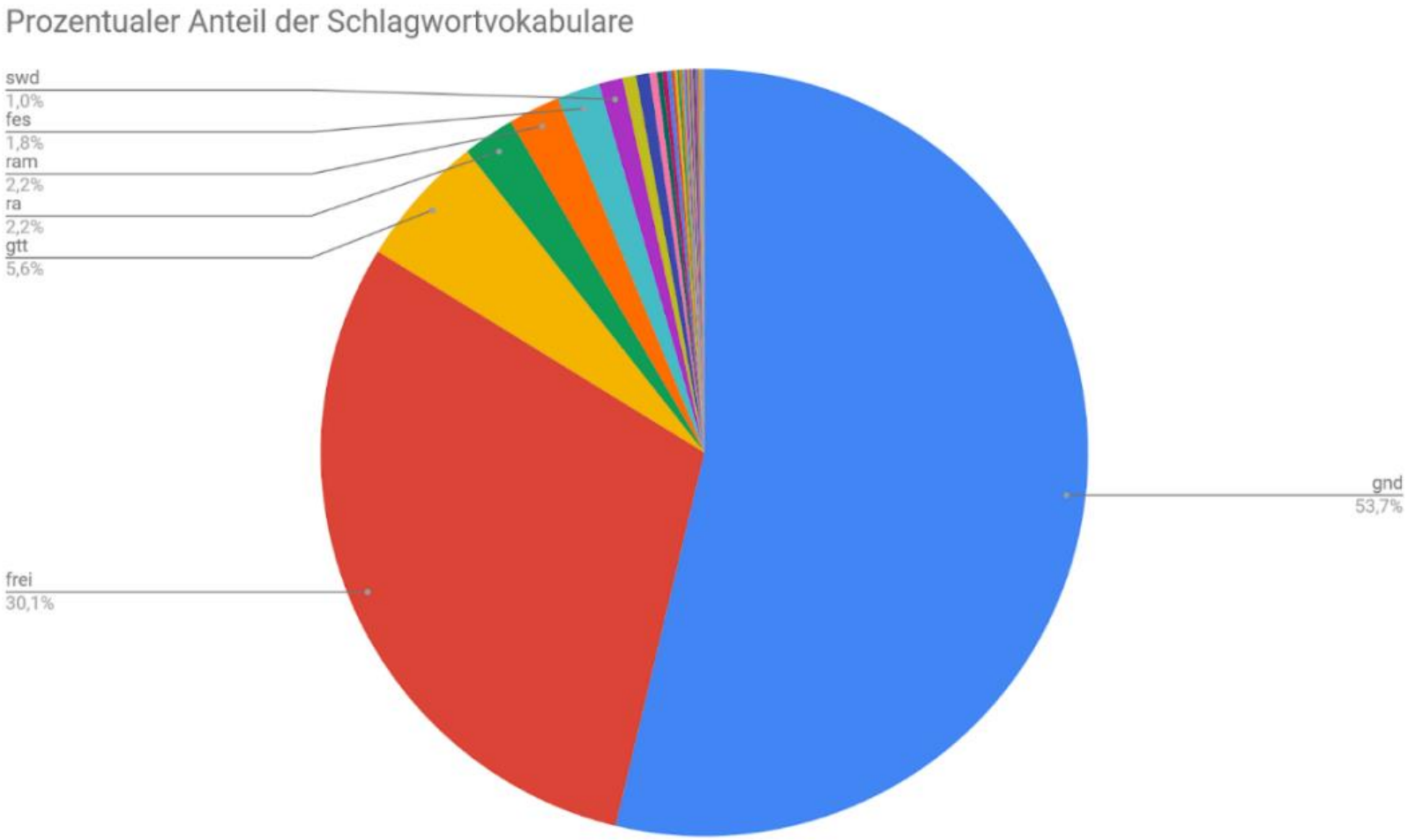
Schlagwörter

- Etwa 50% verschlagwortet
- Über 50% davon mit Schlagwörter aus mehreren Vokabularen
- 100 Vokabulare

Ranking von verschlagworteten Dokumenten



Ranking von verschlagworteten Dokumenten



Vorteile der Inhaltserschließung

Synonymkontrolle

- Sinnvoll wenn
  - Benutzer das Vokabular kennt
  - Konsistent erschlossen ist

Synonymnormalisierung in der Volltextsuche

- Ersetzung von Varianten durch standardisierten Term in Volltextindex einfach realisierbar.
  - Standard Feature von SOLR
  - Kein Wissen vom Benutzer erforderlich, da Terme auch in der Suche ersetzt werden können.

Synonymexpansion in der Volltextsuche

- Term wird durch alle Synonyme ersetzt
- Entweder im Text oder in der Suchanfrage
  - Standard Feature von SOLR

Quellen für Synonyme

Traditionell

- Wörterbücher
- Thesauri, Vokabulare, usw.

Computerlinguistik

- Distributionelle Ähnlichkeit (‘Word2vec’)
- Wörter die häufig im gleichen Kontext stehen, haben eine ähnliche Bedeutung

Synonymsuche

epigenome

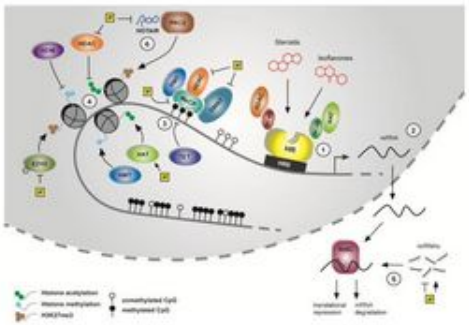
Search

Filter by image type

Filter by dis

Currently displaying results 1 to 12 of 32

Alternative search terms: methylome epigenetic epigenetics genome



Caption: Overview of the impact of soy IF on the epigenome.

Paper: Impact of Soy Isoflavones on the Epigenome in Cancer Prevention

Journal: Nutrients

Year: 2014

Author: Pudenz, Maria and Roth, Kevin and Gerhauser, Clarissa

Copyright: cc-by-4.0

Disciplines: Agriculture and forestry, Process Engineering, Biotechnology

Categories: Molecular genetics Signal transduction Molecular biology Epigenetics Cellular processes

Open in new tab

Cite

Quellen für Synonyme

Click-through data

- *Nutzer, die x (z.B. epigenome) suchen, wählen immer Dokumente mit y (z.B. epigenitics) im Snippet*

Pseudo-Relevance Feedback

- *18 der besten 20 Suchergebnisse für x (z.B. epigenetic) enthalten Term y (z.B. Methylone). Wir sollten y an der Suchanfrage hinzufügen!*
- Füge viele Terme aus der Ergebnismenge hinzu
- Finde damit relevante Dokumente, die den Suchterm nicht enthalten
- Gefahr von *query drift*

Vorteile der Inhaltserschließung

Facettierung

- Möglich wenn in einigen größeren Klassen klassiert wurde
  - Z.B. höhere Ebenen von Thesauri oder DDC
- Sehr nützlich
- Einfach kombinierbar mit Volltextsuche
- Vorherige intellektuelle oder maschinelle Klassierung erforderlich

Vorteile der Inhaltserschließung

Kontrolle und Korrektur

- Maschinelle Erschließungsergebnisse können intellektuell korrigiert werden.

Ranking von verschlagworteten Dokumenten

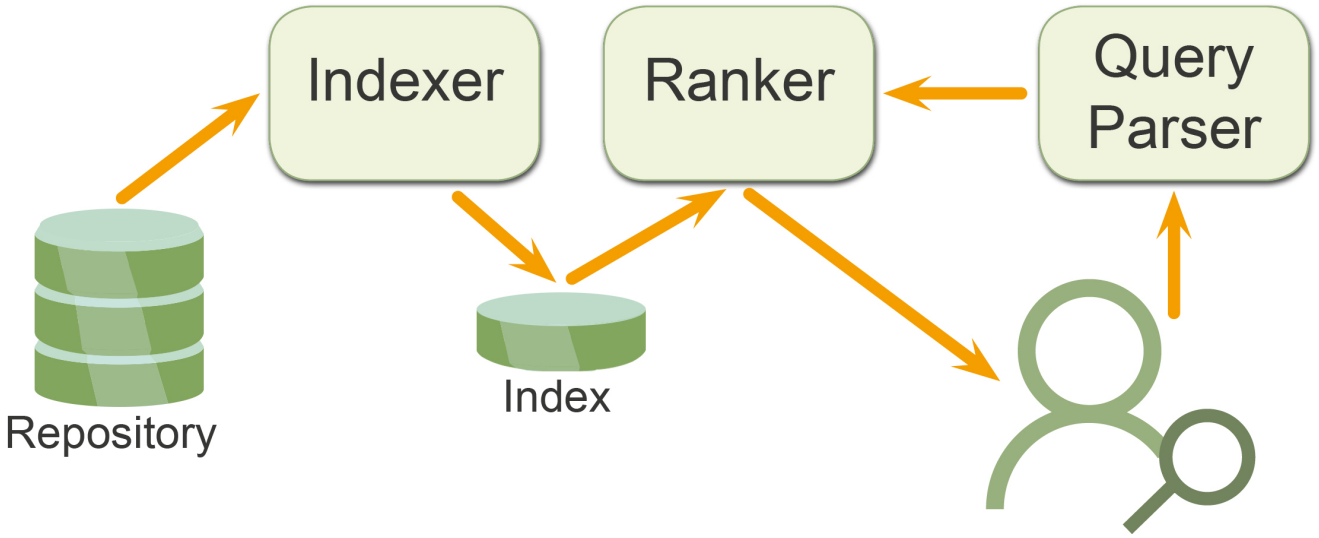
Vergleich mit Volltextindexierung

- Zwei Indexierungsschritte $\Rightarrow$ Informationsverlust
- Theoretisch kein Relevanzranking möglich
  - Klassische Relevanzmaße führen leicht zu komischen Ergebnisse
  - Fehler in der automatischen Erschließung können verheerende Folgen haben.

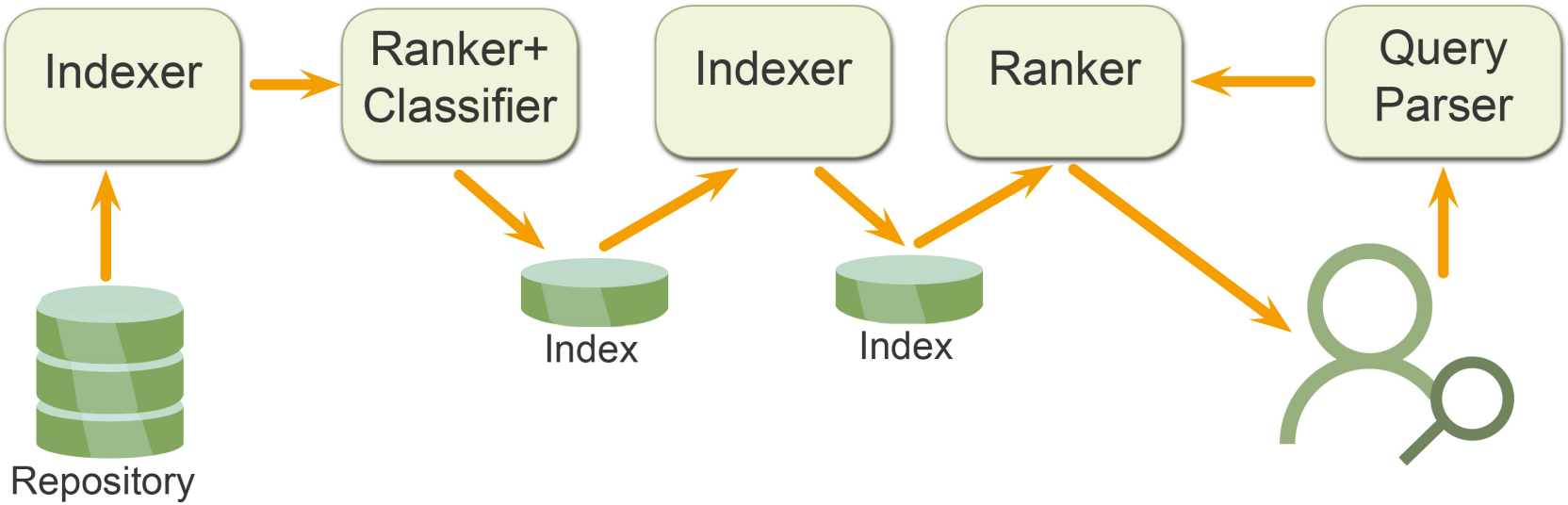
Vergleich mit Volltextindexierung

- **Direkter Vergleich in einem durchdachten Experiment notwendig.**
- Geringe Chancen für automatische Erschließung

Ranking von verschlagworteten Dokumenten



Ranking von verschlagworteten Dokumenten



Mischungen

Heterogene (teilweise erschlossene) Sammlungen

- Fehlendes Schlagwort: Nicht relevant oder nicht erschlossen
- Kombination von Volltextindexierung und verschlagworteten Dokumenten kann zur ungewollten Präferenzen bei der Suche führen.
- Maschinelle Erschließung als Ergänzung zur intellektuellen Erschließung
  - Konsistenz, Lücken Schließen.
  - Wort aus dem Volltext wird relevanter wenn es auch im Index steht und umgekehrt

Facettierung

- Automatische Klassierung in groben Klassen sehr gut möglich
- Nützlich und einfach integrierbar in Volltextsuche

Fazit: Was machen wir mit (Voll)texten?

Wann ist automatische Verschlagwortung sinnvoll?

- Als Kurzzusammenfassung / Vorschau
- Zum Erreichen einer konsistenten Erschließung?
- Komplexer Verfahren mit mehreren Quellen?

Wann ist automatische Klassierung sinnvoll?

- Vor allem wenn man die höhere Ebenen nutzt.
- (Wenn man es nutzt)

Fehler

- Erinnerung: wir reden von **Automatische** Erschließung versus Volltextindexierung
- Beide Verfahren machen (schlimmere) Fehler (als Menschen)
- Die binäre Entscheidungen in der automatischen

Ausblick

Andere Quellen für die Erschließung

- Nutze Schlagwörter und Notationen um neue zu lernen
- Ohne Konkordanzen!
- Lüschow & Wartena (2017): 81% korrekt klassifizierte MESH Terme (erste zwei Digits)

Ein Index für Beschlagwortung und Ranking

- Aut. Beschlagwortung ist Vorverarbeitung+Ranking+Selektion
- Ranking inkl. Gewichte aufheben für Suche
- Selektion nur für Anzeige
- Ermöglicht leichte Integration anderer Quellen und Verfahren

Bildnachweis

Folie 1: Hochschule Hannover, Expo Plaza 12, © Hoang
Folie 2: Librije VVV Zutphen, `www.inzutphen.nl`
Folie 8/9: Kopiert aus Stock, Wolfgang G. Information retrieval: Informationen suchen und finden;[Lehrbuch]. Vol. 1. Oldenbourg Verlag, 2007.