



Information Retrieval

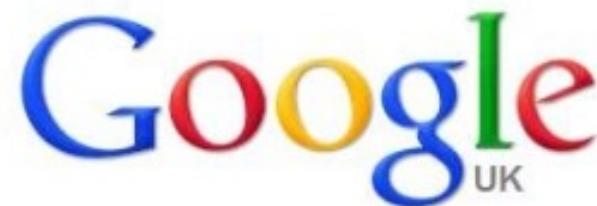
08: Query Expansion

Philipp Schaer, Technische Hochschule Köln, Cologne, Germany

Version: 2022-05-31

Leitfragen für heute

- Was ist die “Volltext-Falle”?
- Was ist das “Sprachproblem im IR”?
- Wie kann da Anfrageerweiterung helfen?
- Woher kommen die Anfrageerweiterungen?
- Why is there a dead pakistani on my couch?

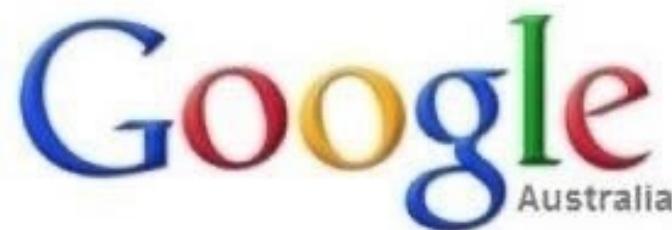


why is there

why is there **a dead pakistan** on my couch
why is there **a light in the fridge and not in the freezer**
why is there **an apple on the cover of twilight**
why is there **a worm in tequila**
why is there **blood in my poo**
why is there **a dragon on the welsh flag**
why is there **a new libby in neighbours**
why is there **salt in the sea**
why is there **a war in afghanistan**
why is there a

Google Search

I'm Feeling Lucky

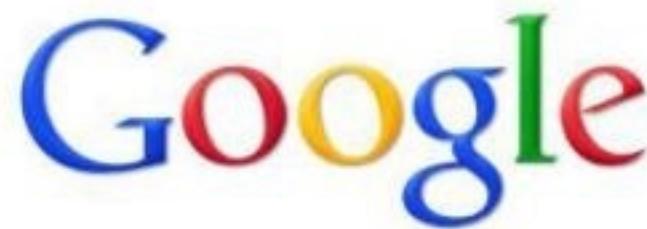


me and google are so close we finish each others sen|

sentences

Google Search

I'm Feeling Lucky



asdfasdf

[Advanced Search](#) [Language Tools](#)

asdfasdfasdfasdfasdfasdfasdfasdfasdf

asdfsdfasdfsdfasdfsdfasdfsdfasdfsdfasdfsdfasdfsdfasdfsdfasdfsdfasdfsdfasdfsdf

asdfsdfasdf

asdfasdfasdfasdft

asdfsdfasdfsdfasdfsdfasdfsdfasdfsdfasdfsdf

asdfsdfasdfsdfasdt

asdfsdfasdfsdfasdfsdf

asdfsdfasdfsdfasdfsdfasdfsdfasdfsdf

asdfsdfasdfsdfasdfsdfasdfsdfasdt

Google Search

I'm Feeling Lucky

Was war das denn?



Kann man ernsthafter ins Thema einleiten?

Ja, man kann...

Bing 

 **Angela Merkel**
Angela Dorothea Merkel ist eine deutsche Politikerin und seit dem 22. Novemb...

angela merkel

angela merkel **ehemann getrennt**

angela merkel **lebenslauf**

angela merkel **jüdischer abstammung**

angela merkel **steckbrief**

angela merkel **facebook**

angela merkel **biografie**

angela merkel **lebenslauf jüdin**

Themen der Veranstaltung

Query Expansion - Anfrageerweiterung

- Allgemeines Ziel: **Verbesserung der Anfrage**
- **Höherer Recall**, da bspw. **Synonyme** erfasst werden können (z.B. Flieger vs. Flugzeug; oder Thermodynamik vs. Hitze, etc.)

Möglichkeit zu besseren Anfragen zu gelangen...

1. Globale Methoden

- Query Expansion
 - Thesauri
 - Automatische Thesaurus-Erstellung

2. Lokale Methoden (heute nicht)

- Relevance Feedback (Nutzer bewerten Dokumente und System reagiert)
- Pseudo Relevance Feedback

Das Sprachproblem im IR



Das Sprachproblem im IR

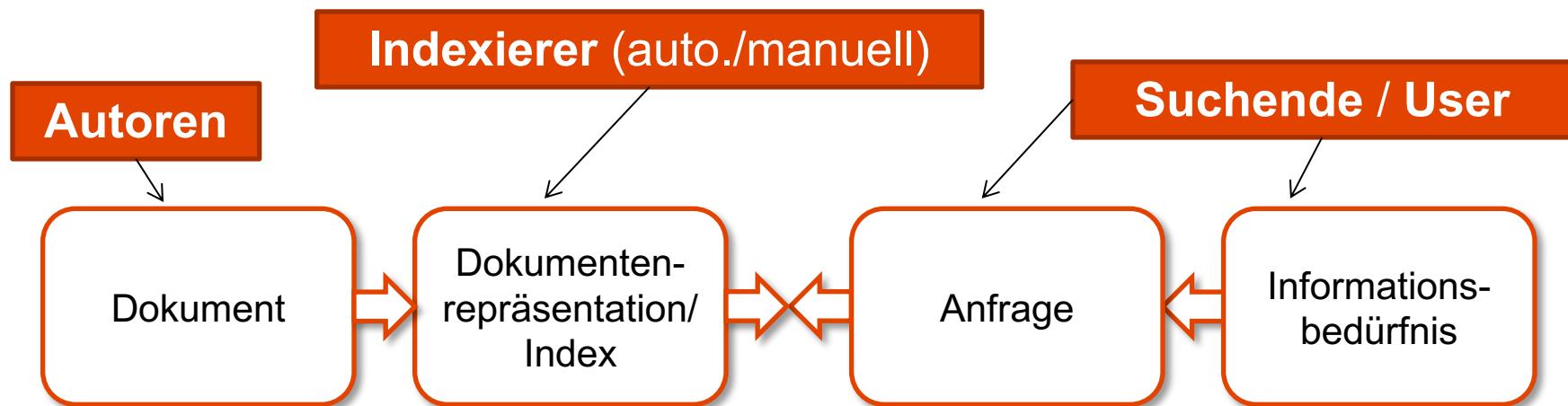
- Der Begriff des “**Sprachproblems**” wurde von Furnas et al. (1987) und später Blair (2004) geprägt.
- Ist besonders ein Problem in **Digitalen Bibliotheken**, da hier zum einen ein spezielles Vokabular verwendet wird und meist nur **Metadaten** zur Suche vorhanden sind.
- Moderne IR-Systeme versuchen den Benutzer bei seiner Suche aktiv zu unterstützen, z.B. mit Hilfe von Stopwort-Listen, Stemming, Rechtschreibkontrolle, etc.
- aber ...

„I choose search terms based not specifically on the information I want, but rather on how I could imagine someone wording [...] that information.“

from Aula et al. (2005)

Das Sprachproblem im IR

Das Sprachproblem zeigt eine Diskrepanz zwischen Vokabular des Suchenden und der automatischen oder manuellen Indexierer (z.B. kontrollierte Vokabulare, oder Volltext-Index).



- Autoren, Indexierer und Suchende haben verschiedene **Vorstellungen** und **Verständnisse** zu einem Thema.
- Information Retrieval ist oft nur ein Abgleich („match“) von Termen.

Das Sprachproblem im IR

Mindestens zwei unterschiedliche Vokabulare, meist mehr:

- Das Vokabular **des Autors** des zu suchenden Textes,
- Das Vokabular **des Suchenden**,
- Das Vokabular des **Indexierers / Katalogisierers**, das benutzt wird um eine Repräsentation des Dokumentes in der Datenbank zu hinterlegen,
- **Synonyme** für Begriffe, die die zuvorgenannten Gruppen vergeben haben (z.B. aus einem Thesaurus),
- Das Vokabular eines Suchenden, dass er tatsächlich in der **Suchanfrage** verwendet.

Wie erweitern wir die Benutzeranfrage?

Manual mit Hilfe eines Thesaurus

- z.B. STW - Standard Thesaurus Wirtschaft
- Kann schnell sehr umfangreich werden...

The screenshot shows a web page for the 'Information' descriptor in the STW thesaurus. The header includes the ZBW logo and navigation links for RDF/XML, RDF/Turtle, Concept history (RDF/Turtle), and English. The main content area has a search bar. On the left, a sidebar lists various menu items like Home, STW Relaunch, and mappings. The main content area is titled 'Information E8'. It defines 'Information' as '(engl.)' used for 'Informationsprozeß, Informationsprozess' and suggests 'Möglichst spezifischer indexieren.' Below this is a section for 'Unterbegriffe' (Subterms) listing terms like 'Produktinformation', 'Unvollkommene Information', etc., each with an 'E8' link. A section for 'Verwandte Begriffe' (Related terms) lists 'Informationsdienstleistung' and 'Informationsmarkt' with their respective 'E8' links.

Information E8

Information (engl.)

benutzt für: Informationsprozeß, Informationsprozess

Möglichst spezifischer indexieren.

Unterbegriffe

- ▶ [Volkswirtschaft](#)
- ▶ [Betriebswirtschaft](#)
- ▶ [Wirtschaftssektoren](#)
- ▶ [Produkte](#)
- ▶ [Nachbarwissenschaften](#)
- ▶ [Geographische Begriffe](#)
- ▶ [Allgemeinwörter](#)

Verwandte Begriffe

- ▶ [Informationsdienstleistung](#)
- ▶ [Informationsmarkt](#)

Wie erweitern wir die Benutzeranfrage?

Manual mit Hilfe eines Thesaurus

- z.B. STW - Standard Thesaurus Wirtschaft
- Kann schnell sehr umfangreich werden...

Globale Analyse: (statisch; alle Dokumente in der Kollektion)

- Automatisch erstellte Thesauri
 - Kookkurrenz-Analyse / z.B. Jaccard-Index
- Auswertung von Anfrage-Logdateien
 - Welche Begriffe wurden von Benutzern oft gemeinsam gesucht?
 - Gebräuchlich im WWW

Lokale Analyse: (dynamisch)

- Analyse der Dokumente in der Ergebnisliste

Kontrolliertes Vokabular: Thesaurus

- Wortherkunft: „**Wortschatz**“ (griech.)
- **Natürlichsprachig** – Anders als z.B. ein Klassifikationssystem
- **Terminologische Kontrolle**
 - Deskriptor = Vorzugsbenennung
 - Deskriptor + Nichtdeskriptoren = Äquivalenzklasse
 - Deskriptoren + Nichtdeskriptoren + Kandidatenterme (freie Deskriptoren) = Zugangsvokabular
- Relationen: Äquivalenz, Hierarchie, Assoziation
- Bevorzugte Dokumentationssprache für **fachspezifische bibliographische Datenbanken**

Kennen Sie tw. von Herrn Lepsky...

[Home](#)
[Alphabetical descriptor list](#)
[Mappings](#)
[Versions](#)
[Web Services](#)
[Downloads](#)
[About](#)

- ▶ [V Economics](#)
- ▶ [B Business economics](#)
- ▶ [W Economic sectors](#)
- ▶ [P Commodities](#)
- ▶ [N Related subject areas](#)
- ▶ [G Geographic names](#)
- ▶ [A General descriptors](#)

Consumer price index

Verbraucherpreisindex (german)

used for: Retail price index, Cost-of-living index, CPI (Consumer Price Index)

Broader Terms

- [Price index](#)

Related Terms

- [Purchasing power](#)
- [Retail price](#)

Subject Categories

- [V.05.03 Inflation](#) ▾

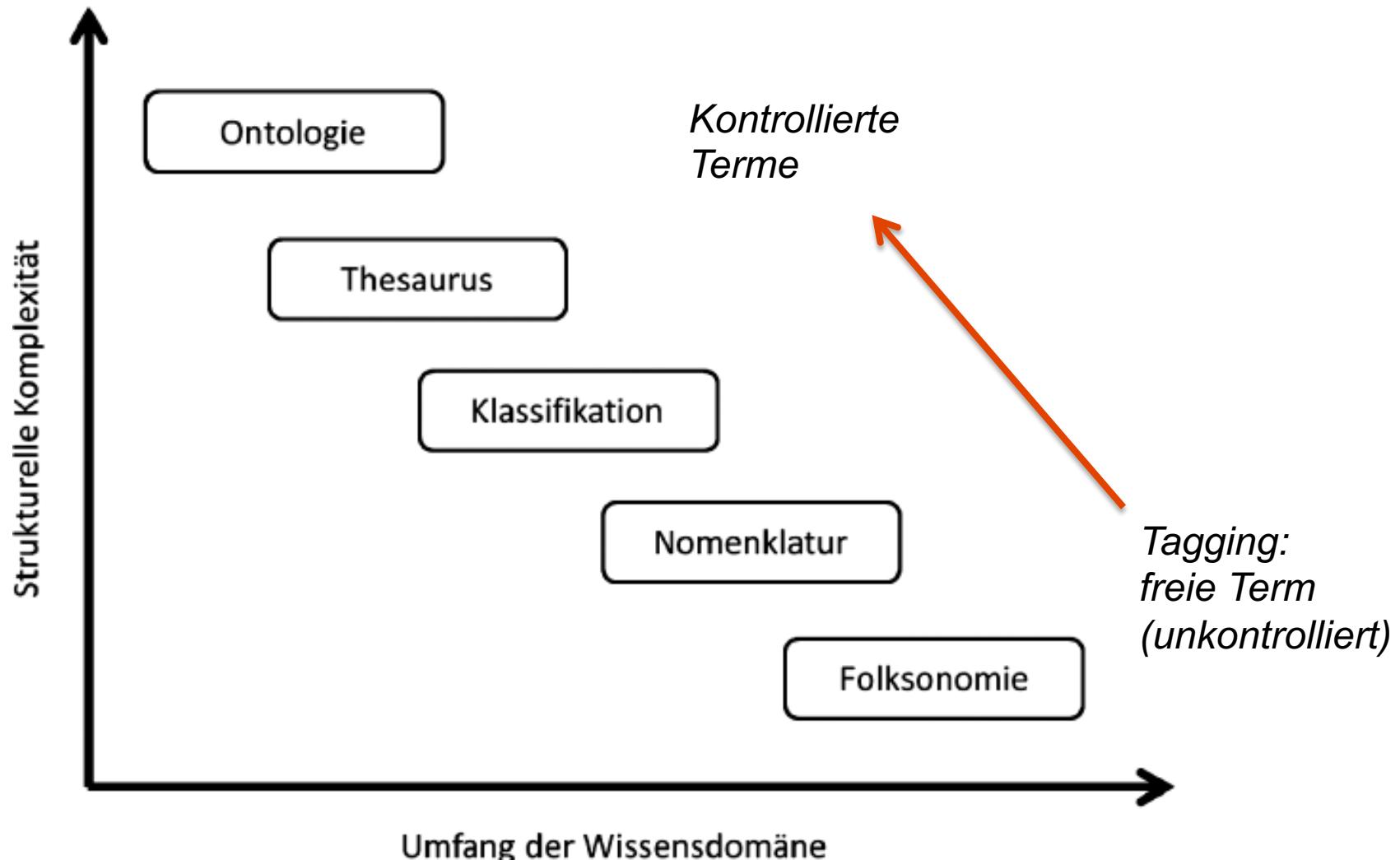
Links to other Thesauri and Vocabularies

- = [Verbraucherpreisindex](#) (from [GND](#))
- ~ [Lebenshaltungskosten](#) (from [GND](#))
- ~ [Preisindex der Lebenshaltung](#) (from [GND](#))
- ≡ [Consumer price index](#) (from [DBpedia](#))

Persistent Identifier (for bookmarking and linking)

- <http://zbw.eu/stwdescriptor/15088-6>

Unterschiedliche Klassifikationsverfahren



Kontrollierte Vokabulare

Terminologische Kontrolle

- Ambiguität der natürlichen Sprache reduzieren
- alle Bezeichnungen für einen Begriff zusammengefasst
- Erleichtert Recherche und Relevanzentscheidung

→ **Synonymkontrolle** (Bedeutungsgleichheit, z.B. Violine/Fidel)

→ **Polysemkontrolle** (Mehrdeutigkeit, z.B. Bank, Läufer, Brücke)

→ **Zerlegungskontrolle**

Suche mit Deskriptoren

Deskriptoren TheSoz	SOLIS Freitext	SOLIS Schlagwort	SW/FT
Algerien	368	275	0,75
Andorra	10	3	0,30
Ruhrgebiet	1099	868	0,79
Salzburg	1347	103	0,08
ALLBUS	441	46	0,10
Funktionsmodell	951	148	0,16
Kubakrise	72	28	0,39
Mittelalter	2145	1513	0,71
Oktoberrevolution	440	229	0,52
Perestroika	600	280	0,47
Proletarisierung	131	64	0,49
Rollenwandel	546	228	0,42
SOEP	883	384	0,43
Stalinismus	3133	2147	0,69
Überalterung	823	164	0,20

Beispiel für Deskriptorensuche

7

Die Schweiz und die EU : das Modell der bilateralen Verträge (Switzerland and the EU : the model of bilateral treaties) ▾

Autor: **Steppacher, Burkard** (Universität Köln, Albertus-Magnus-Platz, 50923 Köln, Bundesrepublik Deutschland)

Erscheinungsjahr: 2008 ; Informationstyp: Literatur; Dokumenttyp: Artikel

Datenbank: **SOLIS - Sozialwissenschaftliches Literaturinformationssystem (GESIS)**

Quelle	Das neue Europa, Martin Große Hüttmann (Herausgeber) 2008, S. 212-227 ISBN: 978-3-89974-355-5
Inhalt	'Beobachtern aus Übersee ist meist nur mit Mühe zu erklären, warum die Schweiz nach dem Ende des Ost-West-Konflikts kein EU-Mitglied ist. Sie ist nicht Teil der umgebenden Europäischen Union (EU) gelegen. Mit ihrer 'neutralen' Position in einem zusammenwachsenden Europa erscheint die Schweiz als Überbleibsel aus einer vergangenen Zeit. Doch dem aufmerksamen Betrachter kann sich auch die Frage stellen: Ist dieses Land vielleicht ein Modell für die künftige Entwicklung der Europäischen Union, die politisch weder Staatenbund noch Einheitsstaat sein kann? Ein Erfolg der Schweiz nacheifert und die in ihrer kulturellen und gesellschaftlichen Verschiedenartigkeit der Eidgenossenschaft in vielen Aspekten von Burkard Steppacher will die besondere Situation der Schweiz in Europa darstellen und das eigenständige Handeln der Schweizer gegenüber den Außenstehenden erklären. Eckpunkte sind hierbei das Nein zum EWR-Abkommen 1992 und der anschließend gewählte bilaterale Ansatz im Bereich der Wirtschaftspolitik. Es bestehen allerdings starke Bedenken, ob dieses Vorgehen 'à la carte' nicht auf Dauer zu kompliziert und zu schwerfällig ist.' (Autorenrede)
Schlagwörter	P bilaterale Beziehungen, P Bundesstaat, P EU, P EU-Beitritt, P EU-Politik, P Europa, P europäische Integration, P Leitbild, P Staatenbund

Freitext-Suche: **Andorra**
Beispiel: sog. „**Volltext-Falle**“

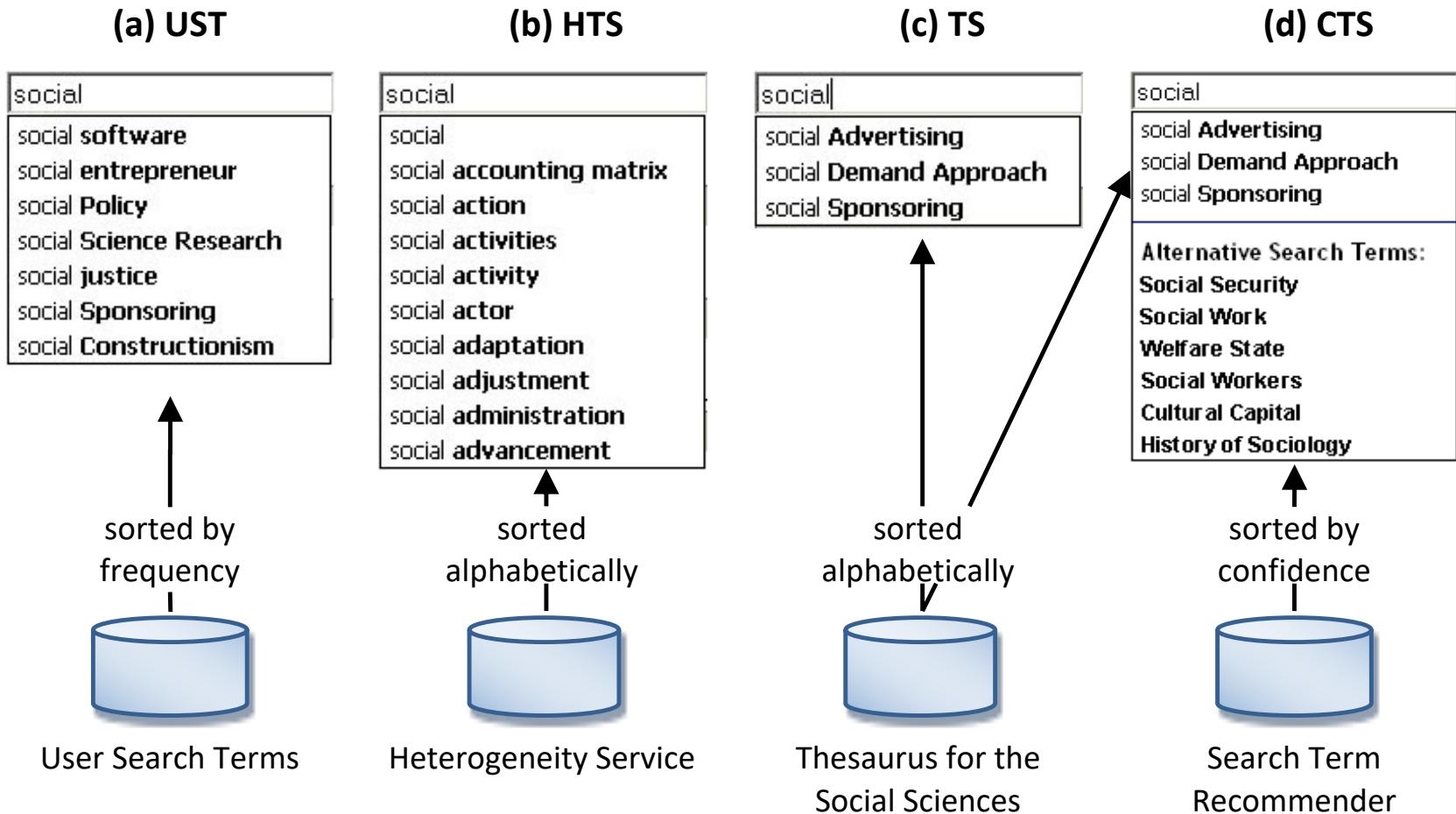
„.... Oft wird die Schweiz mit Gibraltar, **Andorra** oder den Kanalinseln verglichen,“
 → es wird die Schweiz thematisiert, **NICHT** Andorra

Thesaurus-basierte Query Expansion

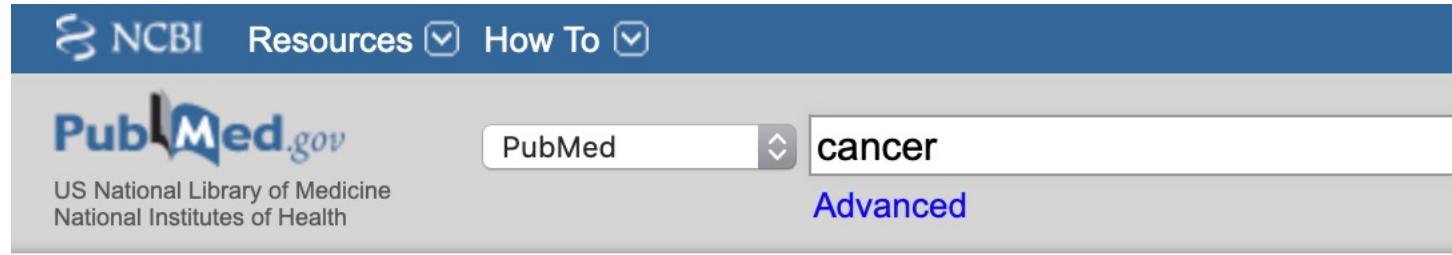
Für jeden Term t in einer Anfrage können **Synonyme** und verwandte Begriffe für t gesucht werden:

- Bestrafung → Strafe
 - Auto → Kfz
-
- Wird oft im **wissenschaftlichen Umfeld** angewendet (warum?)
 - Hierdurch wird meist der **Recall** erhöht (manchmal auch die Precision)
 - Meist sinkt aber die **Precision**, da Mehrdeutigkeiten in der Sprache nicht korrekt abgefangen werden.
 - Die **Kosten** einen Thesaurus zu betreiben und ihn zu benutzen sich vergleichsweise hoch (gemessen in Arbeitszeit).

Beispiel für interaktive Query Expansion



Beispiel bei PubMed



Search Details

Query Translation:

```
"neoplasms" [MeSH Terms] OR "neoplasms" [All Fields] OR  
"cancer" [All Fields]
```

Search

URL

Term-Term-Beziehungen lernen

GESIS-Suche: Geld und Liebe : zur symbolischen Bedeutung von Geld in Paarbeziehungen

gesis Leibniz-Institut für Sozialwissenschaften

Login Englisch Kontakt FAQ

Suche GESIS durchsuchen...

Angebot Forschung Institut

< Zurück

Freie Terme im Titel

Geld und Liebe : zur symbolischen Bedeutung von Geld in Paarbeziehungen

In: Geld und Liebe : Was die Beziehung im Innersten zusammenhält?, Konrad Paul Lieser, Christiane Wimbush, Hamburg, 121-147, 2009

Abstract: Zum Rückblick auf die Entwicklung des männlichen Ernährermodells tritt in einem die steigende Erwerbstätigkeit von Frauen bei, begünstigt durch die Bildungsexpansion der 1960er Jahre. Die Angleichung der Erwerbschancen von Männern und Frauen. Zum anderen lässt sich im Bereich des Familienzusammenlebens ein Wandel hin zu einer Legalisierung von Beziehungen zweier gleichberechtigter Partner feststellen. Mit der steigenden Erwerbstätigkeit von Frauen erhöht sich auch die Zahl der Haushalte, in denen zwei Personen erwerbstätig sind und Einkommen erzielen. Bisher wurde jedoch kaum..." [mehr](#)

Schlagworte: [Geld](#), [Liebe](#), [Partnerbeziehung](#), [Privathaushalt](#), [Symbol](#), [Partnerschaft](#), [Gleichberechtigung](#), [Einkommen](#), [Familie](#), [soziale Ungleichheit](#), [Macht](#)

Dokumenttyp: Sammelwerksbeitrag

Datenbank: GESIS-SOLIS

Freie Terme im Abstract

Aktionen

[Zitieren](#)
[in Google Scholar suchen](#)

Kontrollierte Terme

Kookkurrenzanalyse: Jaccard-Index

Term-Term-Beziehungen können z.B. durch eine **Kookkurrenzanalyse** über den gesamten Korpus **gelernt** werden.

- Analyse unterschiedlicher Dokumentattribute (Titel, Keyword)
- Einsatz des Jaccard-Koeffizient oder Jaccard-Index, der die **Ähnlichkeit von Mengen** berechnet.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Der Wert für $J(A, B)$ liegt zwischen

- 1 (absolute Gleichheit/Ähnlichkeit von A und B) und
- 0 (keine Gleichheit/Ähnlichkeit von A und B)

Der Jaccard-Index

Ein Beispiel:

- Die beiden Mengen A={1,2,3,4,7} und B={1,4,5,7,9} haben den Jaccard-Koeffizienten:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|\{1,4,7\}|}{|\{1,2,3,4,5,7,9\}|} = \frac{3}{7} = 0,429$$

- Wir teilen also die **Größe der Schnittmenge** (3 Elemente) durch die **Größe der Vereinigungsmenge** (7 Elemente)

Term-Beziehung automatisch bestimmen

- Anwendung des Jaccard-Indexes auf Textmengen, mit sowohl freie Textinhalte (Titel) und kontrollierte Inhalte (Thesaurus)
- Ziel ist es „**freie Terme**“ auf „**kontrollierte Terme**“ abzubilden

Oder anders gesagt: Wir suchen Begriffe, die sich ähnlich sind

- A ist hierbei der **freie Term**;
- B der **kontrollierte Term**.

Beispiel

Dokumenttitel 1: „Soziologie des Geldes“

- Thesaursterme: Kultur, Geld, Globalisierung, Wirtschaft

Dokumenttitel 2: „Geld und Geschlechterfragen“

- Thesaurusterme: Frau, Geld, Wirtschaft

Dokumenttitel 3: „Der Gestank des Geldes“

- Thesaurusterme: Kriminalität, Geld, Wirtschaft

Berechnen Sie den Jaccard-Index von „Geldes“ und „Geld“

- A=Geldes im Titel; B=Geld in den Thesaurustermen

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|\{1,3\}|}{|\{1,2,3\}|} = \frac{2}{3} = 0, \overline{666}$$

Was sagt uns der Wert 0,666?

Zunächst einmal wenig...

- Interessant wird es, wenn wir diesen Wert mit anderen Werten **vergleichen**, z.B. für „Soziologie“ und „Wirtschaft“:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|\{1\}|}{|\{1,2,3\}|} = \frac{1}{3} = 0, \overline{333}$$

- Die Ähnlichkeit zwischen „Soziologie“ und „Wirtschaft“ ist also geringer als die Ähnlichkeit zwischen „Geldes“ und „Geld“
- Diese Werte können wir für alle **Term-Term-Beziehungen** ausrechnen und damit eine **gerankte Liste von Term-Empfehlungen** erstellen

Angewendet auf große Menge von Texten

Wendet man das Verfahren auf **große Mengen von Texten** an, bekommt man interessante Einblicke

- **Jugendarbeit im Sport** korreliert stark mit den Thesaurustermen **Jugendlicher, Jugend, Freizeit** und **Sozialarbeit**
- **Burnout-Syndrom** steht in Verbindung zu **Burnout, Lehrer, Hauptschule** und **Krankenpflege**
- Diese Ergebnisse können für eine Anfrageerweiterung verwendet werden.

Hör auf zu sagen
was du denkst,

AI

www.animalistic.cologne



Was soll da schon schiefgehen...?

Nachteile der automatischen Verfahren

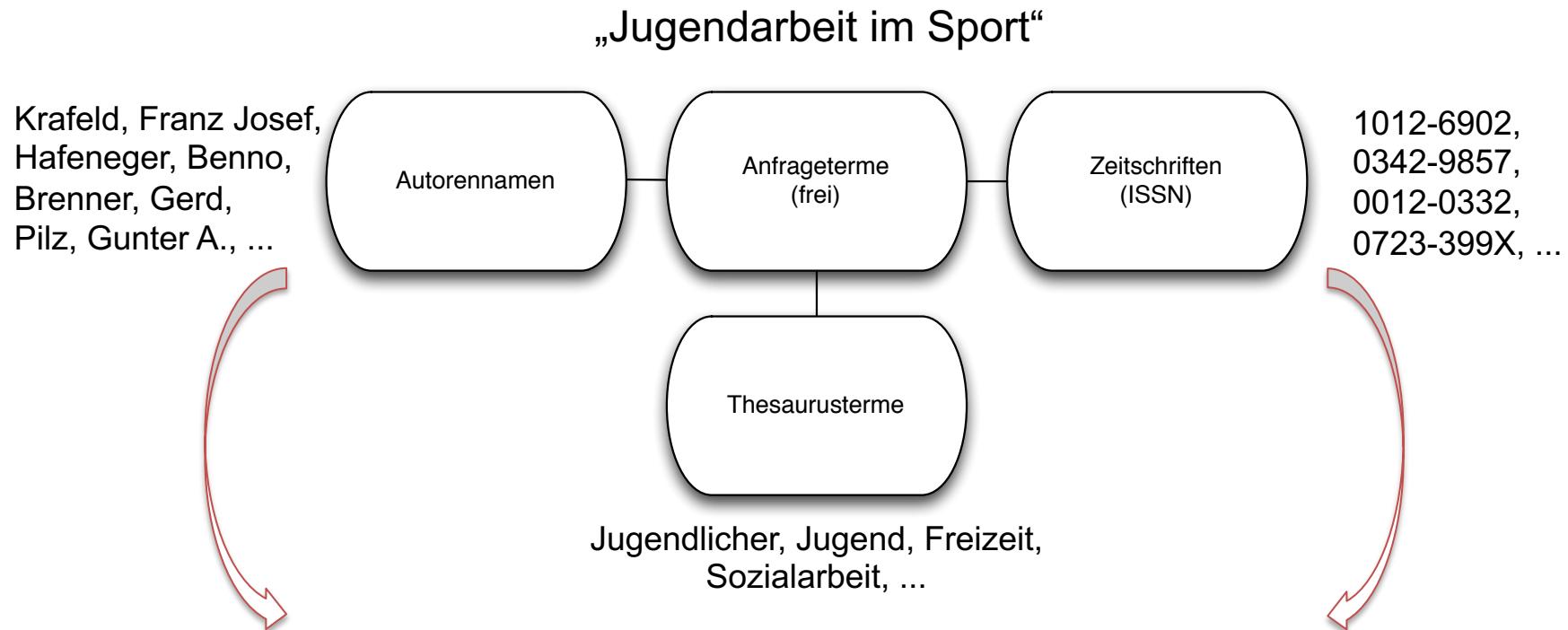
Die **Qualität der Assoziationen** ist meist ein Problem

- **Mehrdeutigkeiten** können für den Benutzer irrelevante Begriffe in die Suche einstreuen (→ **query / topic drift**)
- “Apple computer” → “Apple red fruit computer”

Typische Probleme:

- Falsch-Positive: Wörter scheinen ähnlich, sind es aber nicht
- Falsch-Negative: Wörter scheinen nicht ähnlich, sind es aber

Exkurs: Mehr als Terme...



Erweiterte Anfrage = Ursprüngliche Anfrage, ([Autoren], [Thesaurusterme], [ISSN]);

Dies ist auch durchaus sinnvoll...

Standard IR-Evaluation mit

- **GIRT-4 Korpus** (~145.000 Sozialwissenschaftliche Dokumente mit Titel, Abstract, Autoren und händisch vergebenen Thesaurustermen)
- **CLEF Topics 76-125** und ihren Relevanzurteilen
- Wir haben MAP, rPrecision, p@10, p@20 und p@100 gemessen.

3 Query Expansion-Varianten wurden mitt Baseline-System verglichen

- Baseline-System: Solr Suchmaschine mit TF*IDF Text-Ranking (B)
- Baseline und **QE mit kontrollierten Thesaurustermen** (B+TE)
- Baseline und **QE mit Autorennamen** (B+AE)
- Baseline und **QE mit kontr. Termen und Autorennamen** (B+TE+AE)

Ergebnisse der Query Expansion

- Expansion mit kontrollierten Termen (B+TE) ist besser als B (MAP +9%, aber nicht signifikant)
- B+AE ist schlechter als B (MAP -12%, signifikant) oder vergleichbar mit B (p@20 und p@100)
- Mehrwert wird erst klar, wenn alle drei zusammengenutzt werden:
 - Immer besser als B (MAP +22%, signifikant)
 - Immer besser als B+TE (MAP +11%, signifikant)

run	MAP	rPrecision	p@10	p@20	p@100
B	0,139	0,182	0,442	0,353	0,172
B+TE	0,153	0,229 *	0,430	0,399	0,217 *
B+AE	0,122 *	0,184	0,400 *	0,350	0,175
B+TE+AE	0,170 *	0,239 *	0,478	0,427 *	0,218 *

* = statistical significant for p = .05 (t-test)

Zusammenfassung Query Expansion

- Die Erweiterung der Anfrage kann helfen das „**Sprachproblem des IR**“ zu lindern.
- In bibliografischen Datenbanken eignen sich z.B. **Thesauri** für das Auffinden von **Synonymen** oder **ähnlichen Begriffen**.
- Der Einsatz von Thesauri führt meist zu einer **präziseren Suchanfrage**.
- **Term-Term-Beziehungen** können z.B. mit dem **Jaccard-Index** bestimmt werden.
- Es gibt verschiedene Verfahren zur Berechnung und Anwendung der Query Expansion: **Manuell, global** bzw. **lokal**.

Ach so... Hier noch die Antwort!

