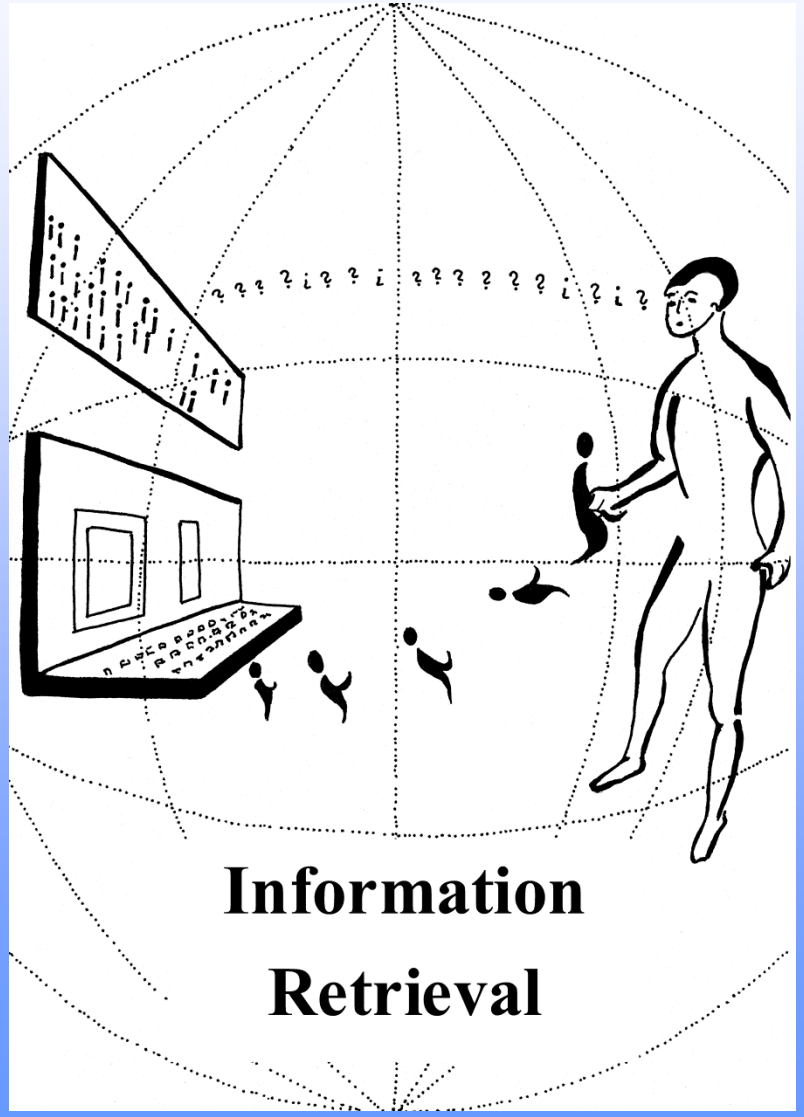
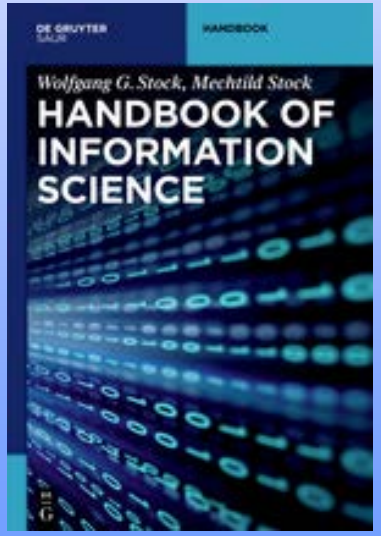

Information
Retrieval

Lehrbuch

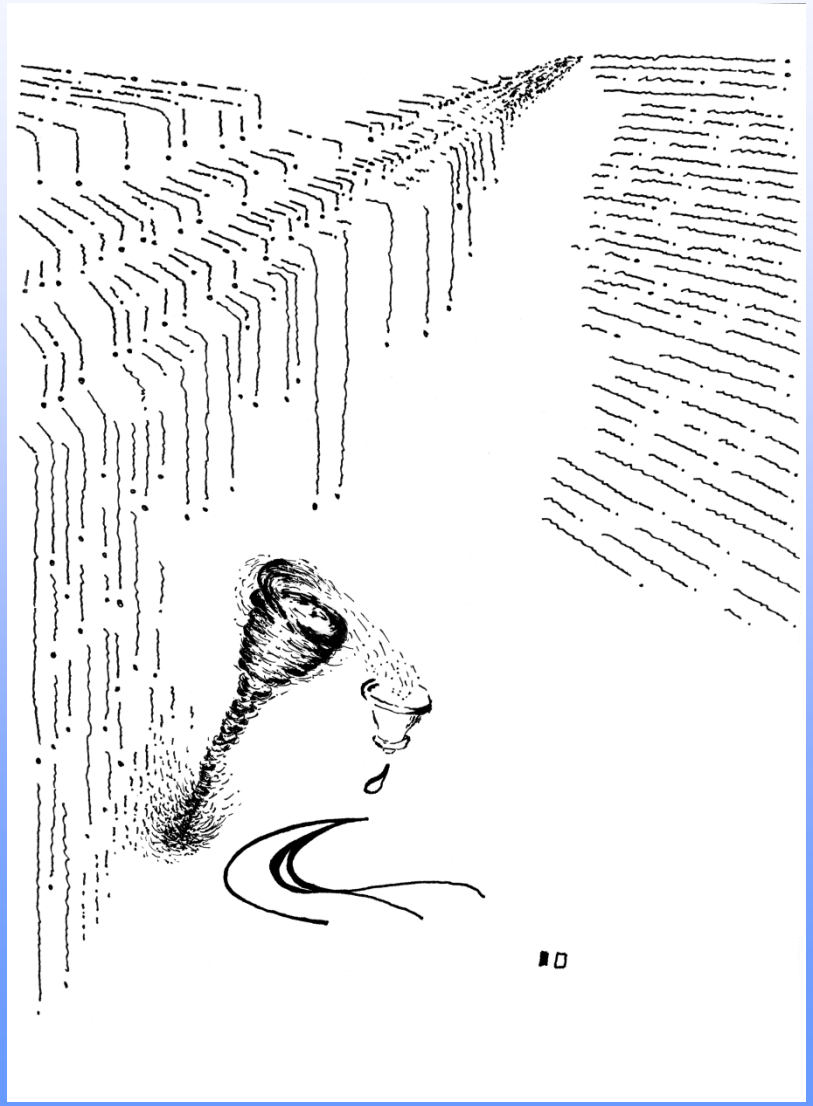
Stock, W.G., & Stock, M. (2015). Handbook of Information Science. [Paperback Edition]

Berlin, Boston, MA: De Gruyter Saur.

- **Gedrucktes Buch**
- **Online (kapitelweise)**



Teil A: Einführung in die Informationswissenschaft



Kapitel A.1

Was ist Informationswissenschaft?

A.1 Was ist Informationswissenschaft?

Informationswissenschaft

? ?

Ist Informationswissenschaft eine *Wissenschaft* (im Sinne von „*science*“)?

Hat Informationswissenschaft etwas mit Informationen zu tun?

A.1 Was ist Informationswissenschaft?

Informationswissenschaft

versus

~~**Informationswissenschaften**~~

(pluralisch)

A.1 Was ist Informationswissenschaft?

Informationswissenschaft untersucht

- **die Repräsentation und das Bereitstellen,**
- **das Suchen und Finden von relevanten (vorwiegend digitalen) Dokumenten und Wissen sowie**
- **das Umfeld von Information.**

A.1 Was ist Informationswissenschaft?

Repräsentation:

Erarbeitung von „Stellvertretern“ von Dokumenten (Surrogaten) sowie von darin enthaltenem Wissen in digitalen Systemen durch

- **Informationsverdichtung (möglichst in ganzen Sätzen),**
- **Informationsfilter (mittels Wörtern oder Begriffen)**

National Library of Medicine - Medical Subject Headings

2014 MeSH

MeSH Descriptor Data

Return to Entry Page

Standard View. Go to Concept View; Go to Expanded Concept View

MeSH Heading	Tennis Elbow
Tree Number	C05.906
Tree Number	C26.088.890
Annotation	in "housewives, artisans & violinists" as well as in tennis players; do not coord with TENNIS (NIM) unless the sport is particularly discussed; do not coord with ATHLETIC INJURIES (IM) unless tennis elbow is discussed as an athletic inj
Scope Note	A condition characterized by pain in or near the lateral humeral epicondyle or in the forearm extensor muscle mass as a result of unusual strain. It occurs in tennis players as well as housewives, artisans, and violinists.
Entry Term	Epicondylitis, Lateral Humeral
Allowable Qualifiers	BL CF CI CL CN CO DH DI DT EC EH EM EN EP ET GE HI IM ME MI MO NU PA PC PP PS PX RA RH RI RT SU TH UR US VE VI

A.1 Was ist Informationswissenschaft?

Bereitstellen:

Erarbeitung von digitalen Dokumenten, so dass sie optimal strukturiert, leicht auffindbar und gut lesbar in digitalen Speichern abgelegt und darin verwaltet werden können

(Webdesign: „Strukturieren digitaler Dokumente“)

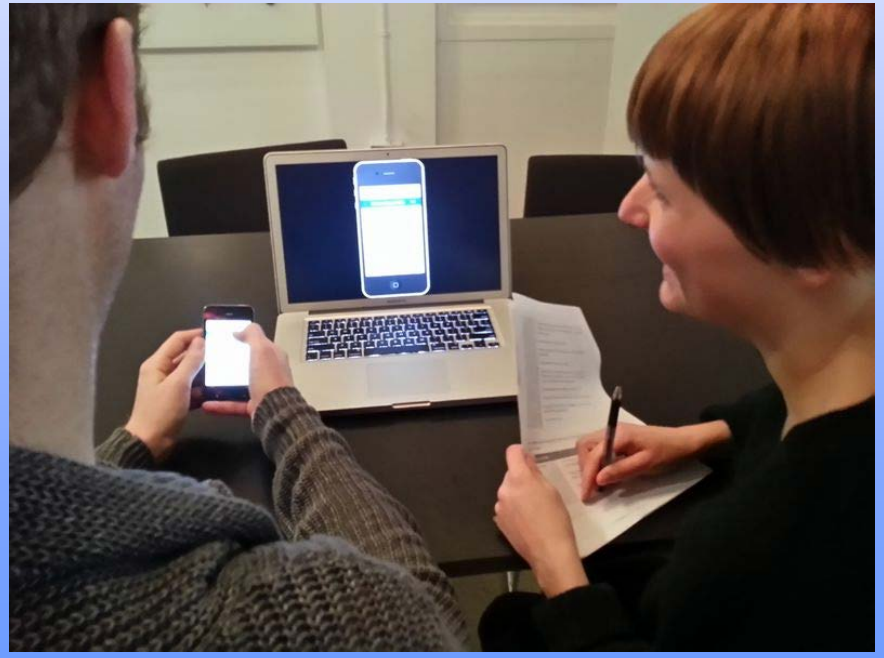
```
<!DOCTYPE html>
<html class="google" lang="en">
  <head>

    <script>
(function(H){H.className=H.className.replace(/\bgoogle\b/,'google-js')})(document.documentElement)
    </script>
    <meta charset="utf-8">
    <meta content="initial-scale=1, minimum-scale=1, width=device-width" name="viewport">
    <title>
      Information Retrieval and the Web - Research at Google
    </title>
    <script src="//www.google.com/js/google.js">
</script>
    <script>
new gweb.analytics.AutoTrack({profile:"UA-5974346-1"});
    </script>
    <link href="//fonts.googleapis.com/css?family=Open+Sans:300,400,600,700&amp;lang=en" rel=
    "stylesheet">
    <link href="/css/research.css" rel="stylesheet">
  </head>
```

A.1 Was ist Informationswissenschaft?

Suchen:

Beobachtung der Nutzer beim Abarbeiten ihrer Informationsbedürfnisse

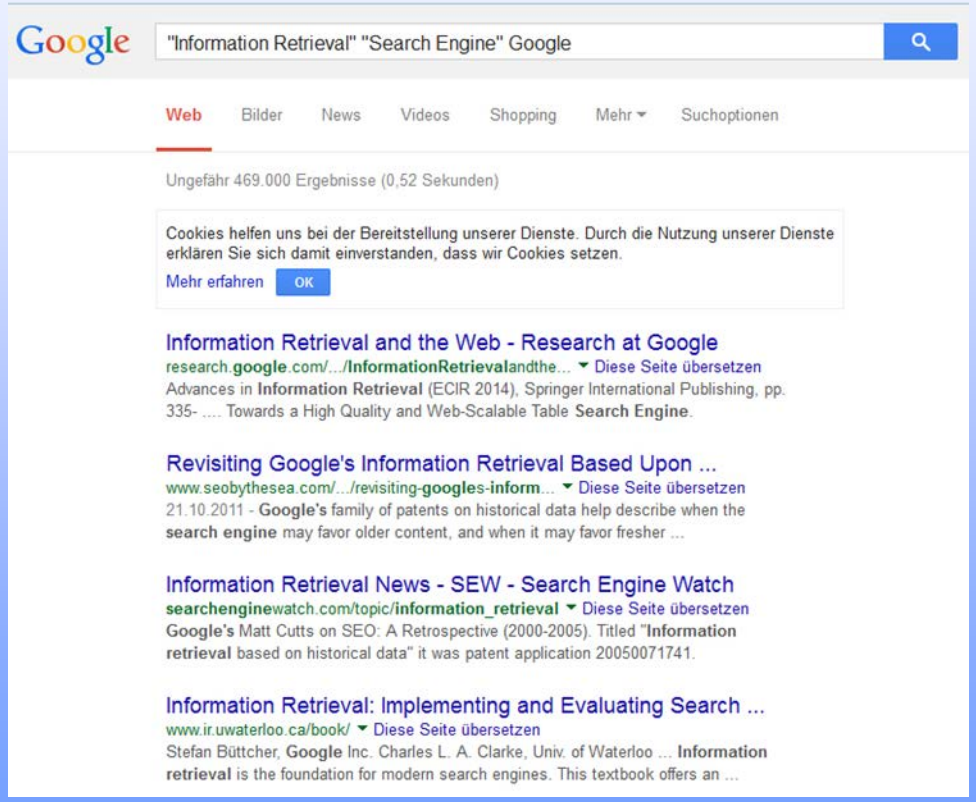
(„Nutzerforschung“ als Teil der Informetrie)



A.1 Was ist Informationswissenschaft?

Finden:

Systeme zum Recherchieren nach Wissen, u. a. Suchmaschinen in Internet und Intranets, fachliche Informationsdienste, Bibliothekskataloge („Information Retrieval“)



A.1 Was ist Informationswissenschaft?

Relevanz:

Es geht nicht um das Finden von „irgendwelchen“ Informationen, sondern nur um das Aufspüren von zutreffendem Wissen

A.1 Was ist Informationswissenschaft?

vorwiegend digital:

Digitale Informationen bilden den Schwerpunkt der Informationswissenschaft;

in Ausnahmefällen werden auch nicht-digitale Informationssammlungen (z.B. Bibliotheken, Archive) betrachtet

A.1 Was ist Informationswissenschaft?

Dokumente:

Texte und nicht-textliche Dokumente (u. a. Bilder, Musik, Videos, aber auch wissenschaftliche Fakten, Wirtschaftsobjekte, Gegenstände in Museen, Zoos usw., real-time Objekte, Personen);

nicht-digitale und digitale Dokumente

A.1 Was ist Informationswissenschaft?

Wissen:

„schlummert“ in Speichern (entweder nicht-digital: im Gehirn, in Büchern usw. oder digital: im WWW, ...).

Information bringt Wissen „in Bewegung“

in-FORM-ation

aktiv: man informiert (verbreitet Wissen)

passiv: man wird informiert (nimmt Wissen auf)

A.1 Was ist Informationswissenschaft?

Umfeld von Information:

pädagogisch: Informationskompetenz

betriebswirtschaftlich: Wissensmanagement

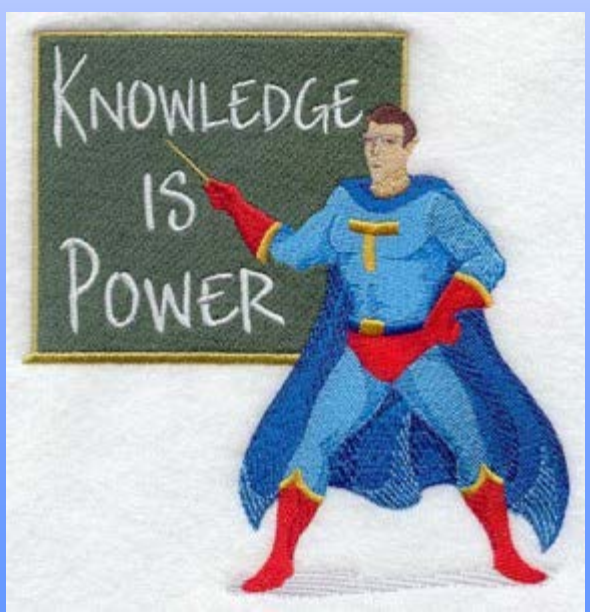
ökonomisch: Informationsmärkte

soziologisch: Wissensgesellschaft

stadt-/raumplanerisch: Informational Urbanism

juristisch: Informationsrecht

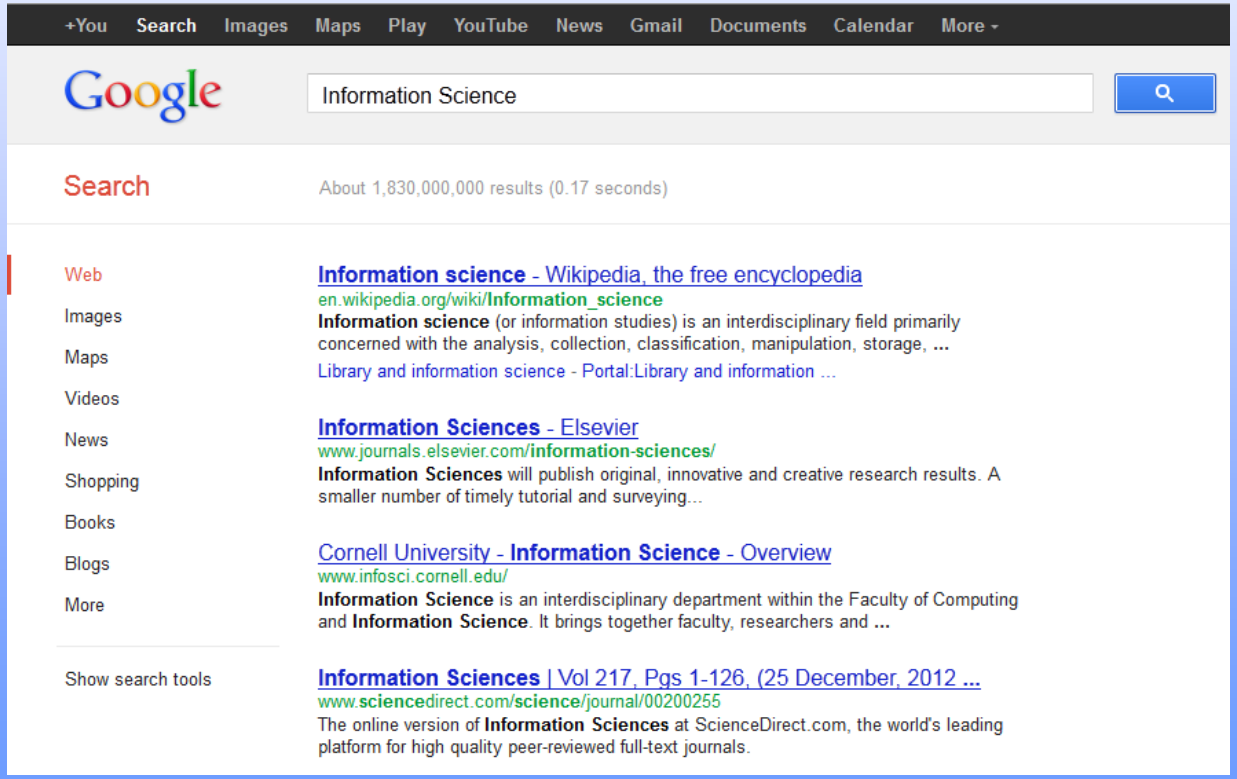
philosophisch: Informationsethik



A.1 Was ist Informationswissenschaft?

Anwendungen von Informationswissenschaft

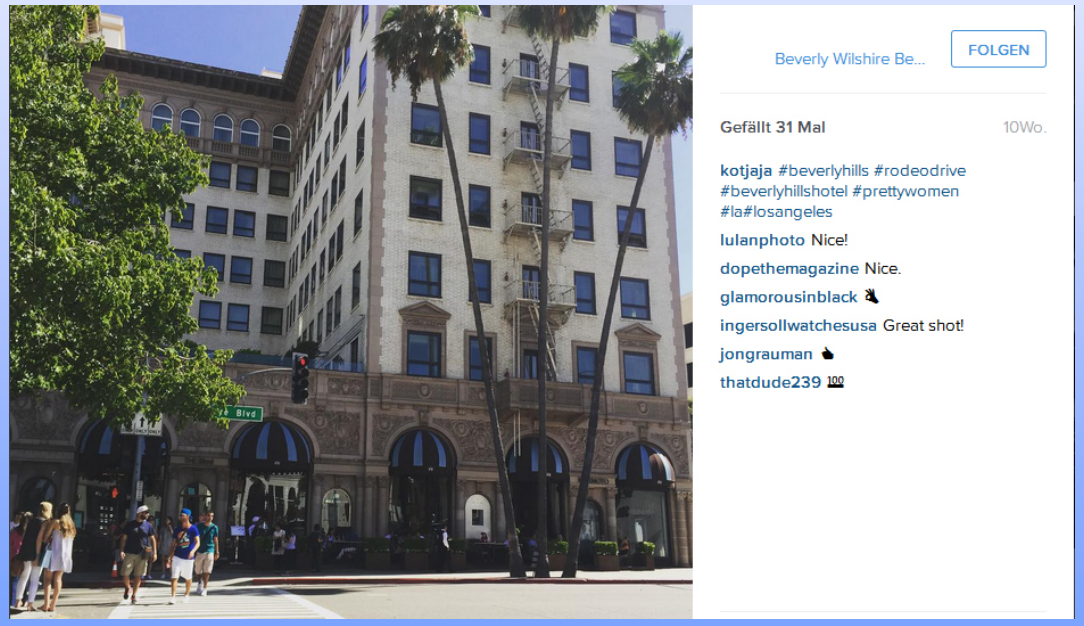
1: Suchmaschinen



A.1 Was ist Informationswissenschaft?

Anwendungen von Informationswissenschaft

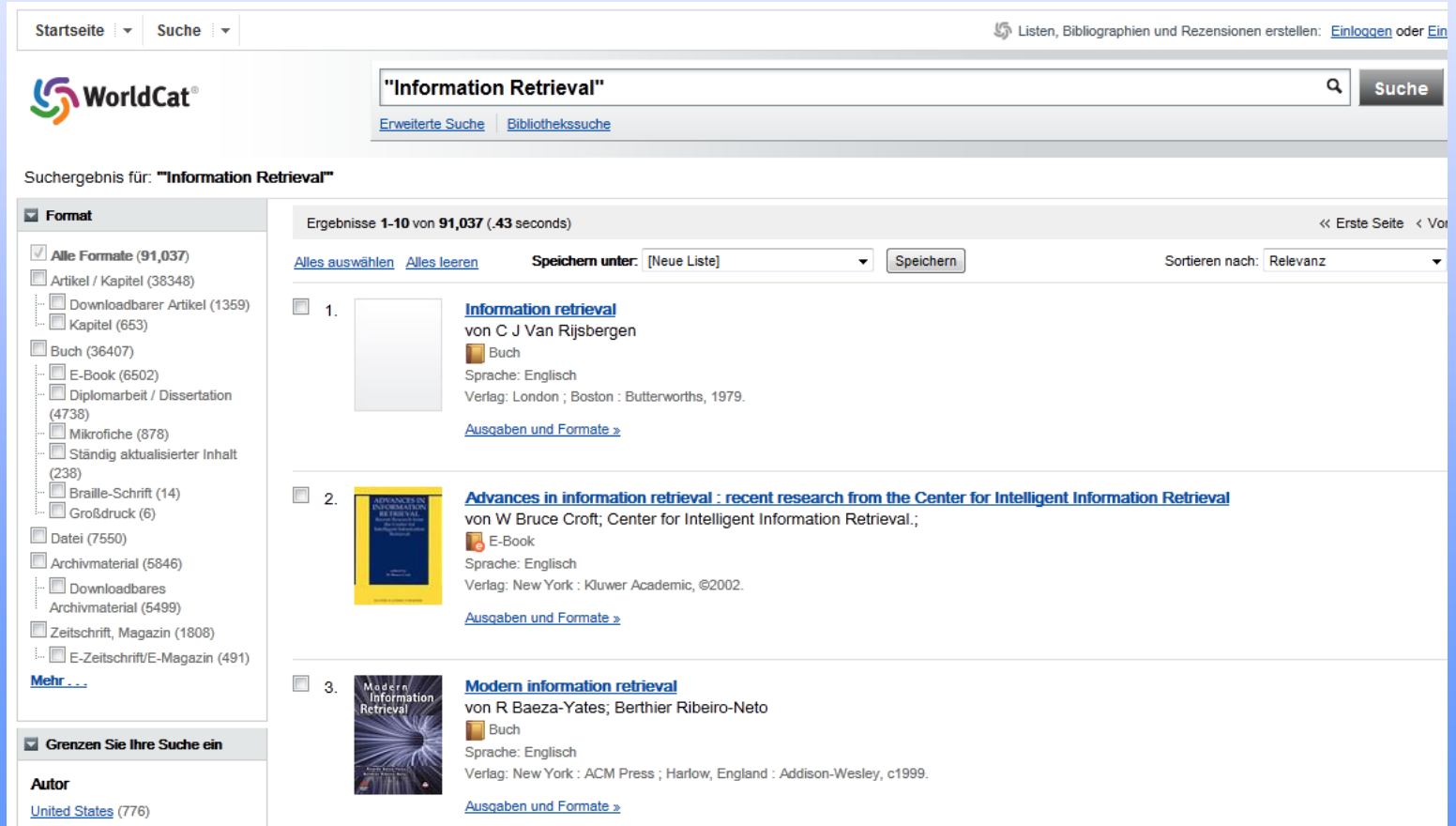
2: Social Media



A.1 Was ist Informationswissenschaft?

Anwendungen von Informationswissenschaft

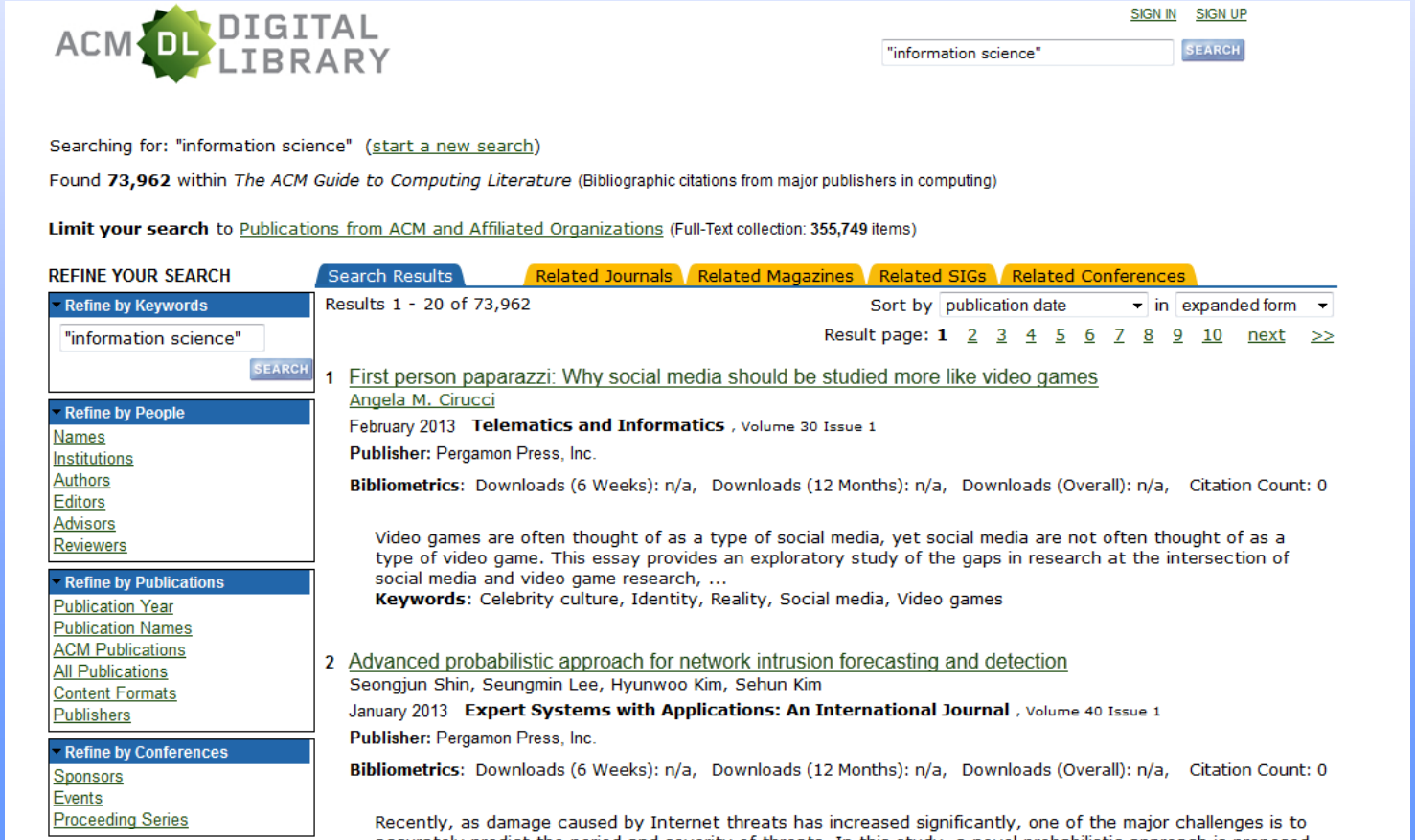
3: Online-Bibliothekskataloge



A.1 Was ist Informationswissenschaft?

Anwendungen von Informationswissenschaft

4: Digitale Bibliotheken



ACM DL DIGITAL LIBRARY

SIGN IN SIGN UP

"information science" SEARCH

Searching for: "information science" (start a new search)

Found **73,962** within *The ACM Guide to Computing Literature* (Bibliographic citations from major publishers in computing)

Limit your search to Publications from ACM and Affiliated Organizations (Full-Text collection: **355,749** items)

REFINE YOUR SEARCH

Refine by Keywords
"information science" SEARCH

Refine by People
Names
Institutions
Authors
Editors
Advisors
Reviewers

Refine by Publications
Publication Year
Publication Names
ACM Publications
All Publications
Content Formats
Publishers

Refine by Conferences
Sponsors
Events
Proceeding Series

Search Results | Related Journals | Related Magazines | Related SIGs | Related Conferences

Results 1 - 20 of 73,962 Sort by publication date in expanded form

Result page: **1** 2 3 4 5 6 7 8 9 10 next >>

1 First person paparazzi: Why social media should be studied more like video games
Angela M. Cirucci
February 2013 **Telematics and Informatics** , Volume 30 Issue 1
Publisher: Pergamon Press, Inc.
Bibliometrics: Downloads (6 Weeks): n/a, Downloads (12 Months): n/a, Downloads (Overall): n/a, Citation Count: 0

Video games are often thought of as a type of social media, yet social media are not often thought of as a type of video game. This essay provides an exploratory study of the gaps in research at the intersection of social media and video game research, ...
Keywords: Celebrity culture, Identity, Reality, Social media, Video games

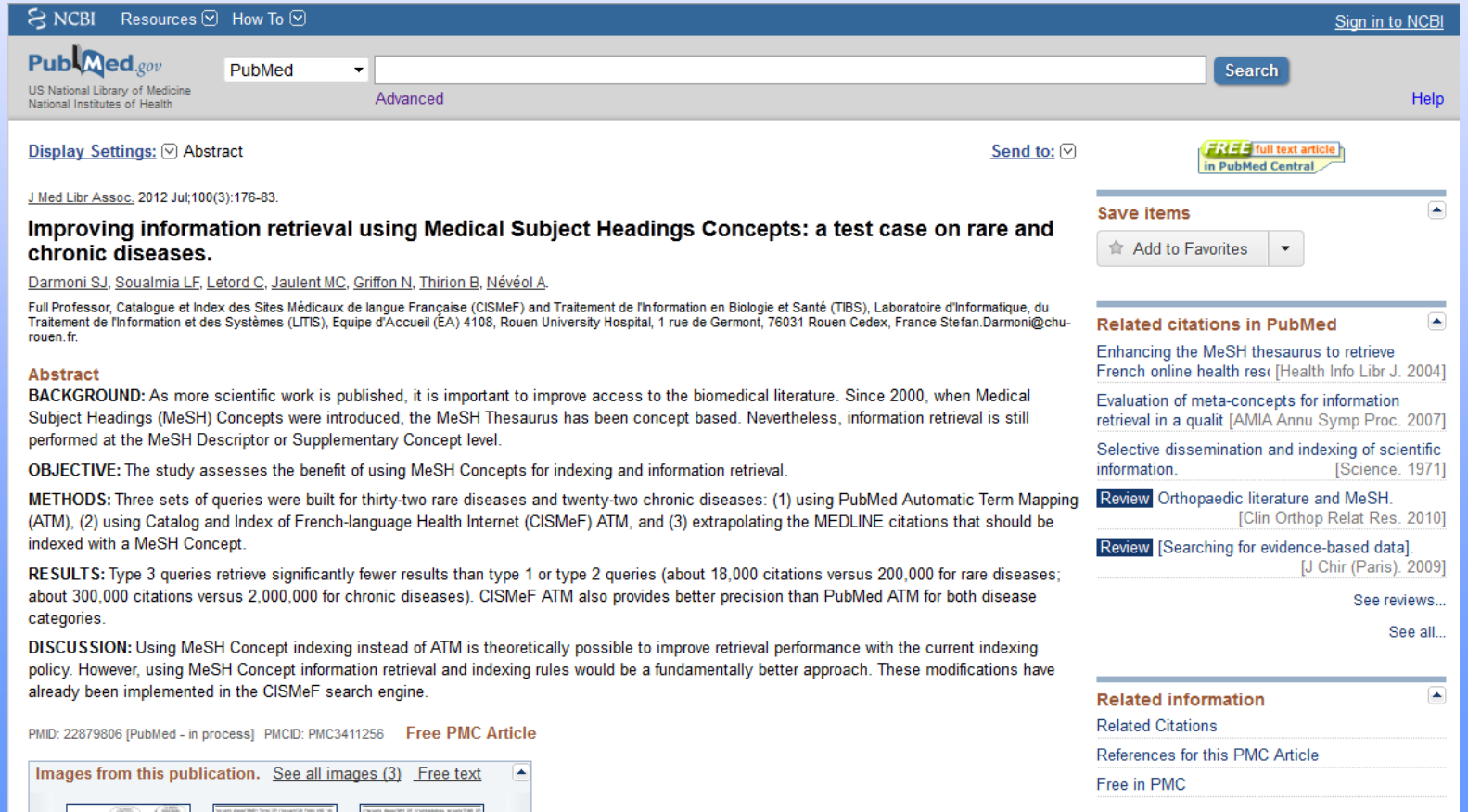
2 Advanced probabilistic approach for network intrusion forecasting and detection
Seongjun Shin, Seungmin Lee, Hyunwoo Kim, Sehun Kim
January 2013 **Expert Systems with Applications: An International Journal** , Volume 40 Issue 1
Publisher: Pergamon Press, Inc.
Bibliometrics: Downloads (6 Weeks): n/a, Downloads (12 Months): n/a, Downloads (Overall): n/a, Citation Count: 0

Recently, as damage caused by Internet threats has increased significantly, one of the major challenges is to

A.1 Was ist Informationswissenschaft?

Anwendungen von Informationswissenschaft

5: Deep-Web-Informationsdienste



NCBI Resources How To Sign in to NCBI

PubMed.gov US National Library of Medicine National Institutes of Health

PubMed Search Advanced Help

Display Settings: Abstract Send to:

FREE full text article in PubMed Central

J Med Libr Assoc. 2012 Jul;100(3):176-83.

Improving information retrieval using Medical Subject Headings Concepts: a test case on rare and chronic diseases.

Darmoni SJ, Soualmia LF, Letord C, Jaulent MC, Griffon N, Thirion B, Névéol A.

Full Professor, Catalogue et Index des Sites Médicaux de langue Française (CISMeF) and Traitement de l'Information en Biologie et Santé (TIBS), Laboratoire d'Informatique, du Traitement de l'Information et des Systèmes (LITIS), Equipe d'Accueil (EA) 4108, Rouen University Hospital, 1 rue de Germont, 76031 Rouen Cedex, France Stefan.Darmoni@chu-rouen.fr.

Abstract

BACKGROUND: As more scientific work is published, it is important to improve access to the biomedical literature. Since 2000, when Medical Subject Headings (MeSH) Concepts were introduced, the MeSH Thesaurus has been concept based. Nevertheless, information retrieval is still performed at the MeSH Descriptor or Supplementary Concept level.

OBJECTIVE: The study assesses the benefit of using MeSH Concepts for indexing and information retrieval.

METHODS: Three sets of queries were built for thirty-two rare diseases and twenty-two chronic diseases: (1) using PubMed Automatic Term Mapping (ATM), (2) using Catalog and Index of French-language Health Internet (CISMeF) ATM, and (3) extrapolating the MEDLINE citations that should be indexed with a MeSH Concept.

RESULTS: Type 3 queries retrieve significantly fewer results than type 1 or type 2 queries (about 18,000 citations versus 200,000 for rare diseases; about 300,000 citations versus 2,000,000 for chronic diseases). CISMeF ATM also provides better precision than PubMed ATM for both disease categories.

DISCUSSION: Using MeSH Concept indexing instead of ATM is theoretically possible to improve retrieval performance with the current indexing policy. However, using MeSH Concept information retrieval and indexing rules would be a fundamentally better approach. These modifications have already been implemented in the CISMeF search engine.

PMID: 22879806 [PubMed - in process] PMCID: PMC3411256 Free PMC Article

Images from this publication. See all images (3) Free text

Save items

Add to Favorites

Related citations in PubMed

Enhancing the MeSH thesaurus to retrieve French online health resc [Health Info Libr J. 2004]

Evaluation of meta-concepts for information retrieval in a qualit [AMIA Annu Symp Proc. 2007]

Selective dissemination and indexing of scientific information. [Science. 1971]

Review Orthopaedic literature and MeSH. [Clin Orthop Relat Res. 2010]

Review [Searching for evidence-based data]. [J Chir (Paris). 2009]

See reviews...

See all...

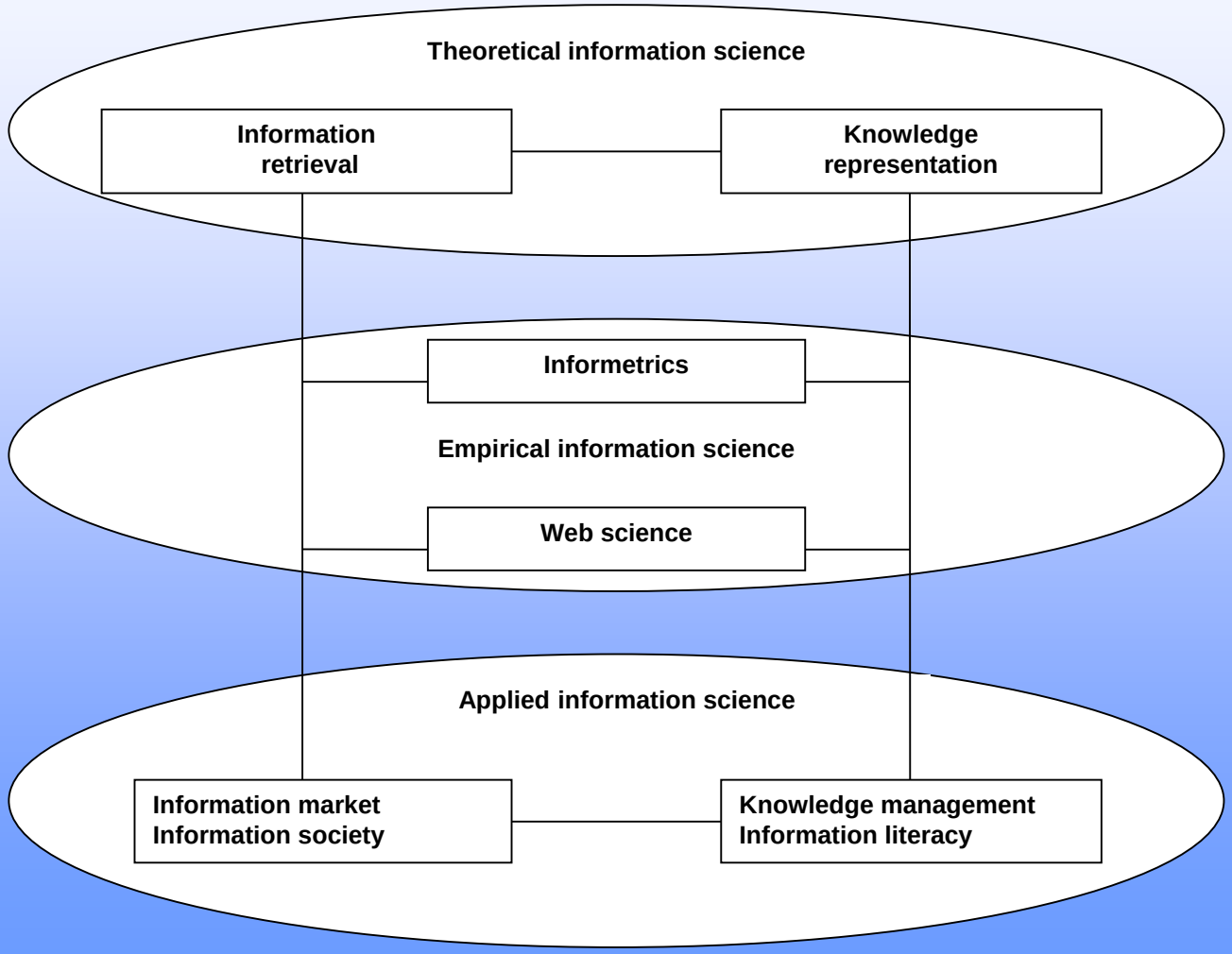
Related information

Related Citations

References for this PMC Article

Free in PMC

A.1 Was ist Informationswissenschaft?



A.1 Was ist Informationswissenschaft?

Klassikerzitate

1: Wolf Rauch

Für die *Informationswissenschaft* ... ist Information *Wissen*. Genauer: Wissen, das zur Bewältigung von problematischen Situationen gebraucht wird. Wissen ist also gleichsam *mögliche* Information. Information ist wirksam gewordenes, handlungsrelevantes Wissen.

Rauch, W. (1988). Was ist Informationswissenschaft? Graz: Kienreich. (Grazer Universitätsreden; 32), S. 26.

A.1 Was ist Informationswissenschaft?

Klassikerzitate

2: Rainer Kuhlen

Information existiert nicht für sich. Information referenziert auf Wissen. Information wird in der Regel als Surrogat bzw. Repräsentation oder Manifestation von Wissen verstanden.

Kuhlen, R. (2004). Information. In R. Kuhlen, T. Seeger, & D. Strauch (Hrsg.), Grundlagen der praktischen Information und Dokumentation (S. 3-20), 5. Aufl. München: Saur, hier: S. 6.

A.1 Was ist Informationswissenschaft?

Klassikerzitate

3: A.I. Michajlov, A.I. Cernyi und R.S. Giljarevskij

(Für die Informationswissenschaft) ist es gleichgültig, ob beispielsweise von einem neuen Insekt oder von einer fortschrittlichen Methode zur Metallbearbeitung die Rede ist.

Michajlov, A.I., Cernyi, A.I., & Giljarevskij, R.S. (1979). Informatik. Informatik, 26(4), 42-45, hier: S. 45.

A.1 Was ist Informationswissenschaft?

Klassikerzitate

4: Harold Borko

Information science is that discipline that investigates the properties and behavior of information, the forces governing the flow of information, and the means of processing information for optimum accessibility and usability. It is concerned with that body of knowledge relating to the origination, collection, organization, storage, retrieval, interpretation, transmission, transformation, and utilization of information. This includes the investigation of information representation in both natural and artificial systems, the use of codes for efficient message transmission, and the study of information processing devices and techniques such as computers and their programming systems.

Borko, H. (1968). Information science: What is it?
American Documentation, 19, 3-5, hier: S. 3.

A.1 Was ist Informationswissenschaft?

Klassikerzitate

5: Norbert Henrichs

Die Praxis der modernen Fachinformation kann gegenüber ihrer Theorie ... unzweifelhaft auf einen erheblichen zeitlichen Vorlauf verweisen. Im Anfang war also, darüber kann es wenigstens hierzulande keinen Zweifel geben, die Praxis.

Henrichs, N. (1997). Informationswissenschaft. In W. Rehfeld, T. Seeger, & D. Strauch (Hrsg.), Grundlagen der praktischen Information und Dokumentation (S. 945-957). 4.Aufl. München: Saur, hier: S. 945.

A.1 Was ist Informationswissenschaft?

Klassikerzitate

6: Tefko Saracevic

First, information science is interdisciplinary in nature; however, the relations with various disciplines are changing. The interdisciplinary evolution is far from over.

Second, information science is inexorable connected to information technology. A technological imperative is compelling and constraining the evolution of information science (…).

Third, information science is, with many other fields, an active participant in the evolution of the information society. Information science has a strong social and human dimension, above and beyond technology.

Saracevic, T. (1999). Information Science.
Journal of the American Society for Information Science, 50, 1051-1063, hier: S. 1052.

A.1 Was ist Informationswissenschaft?

Klassikerzitate. 7: Matthäus

10 Die Jünger kamen zu Jesus und fragten: "Warum sprichst du in Gleichnissen, wenn du zu den Leuten redest?“

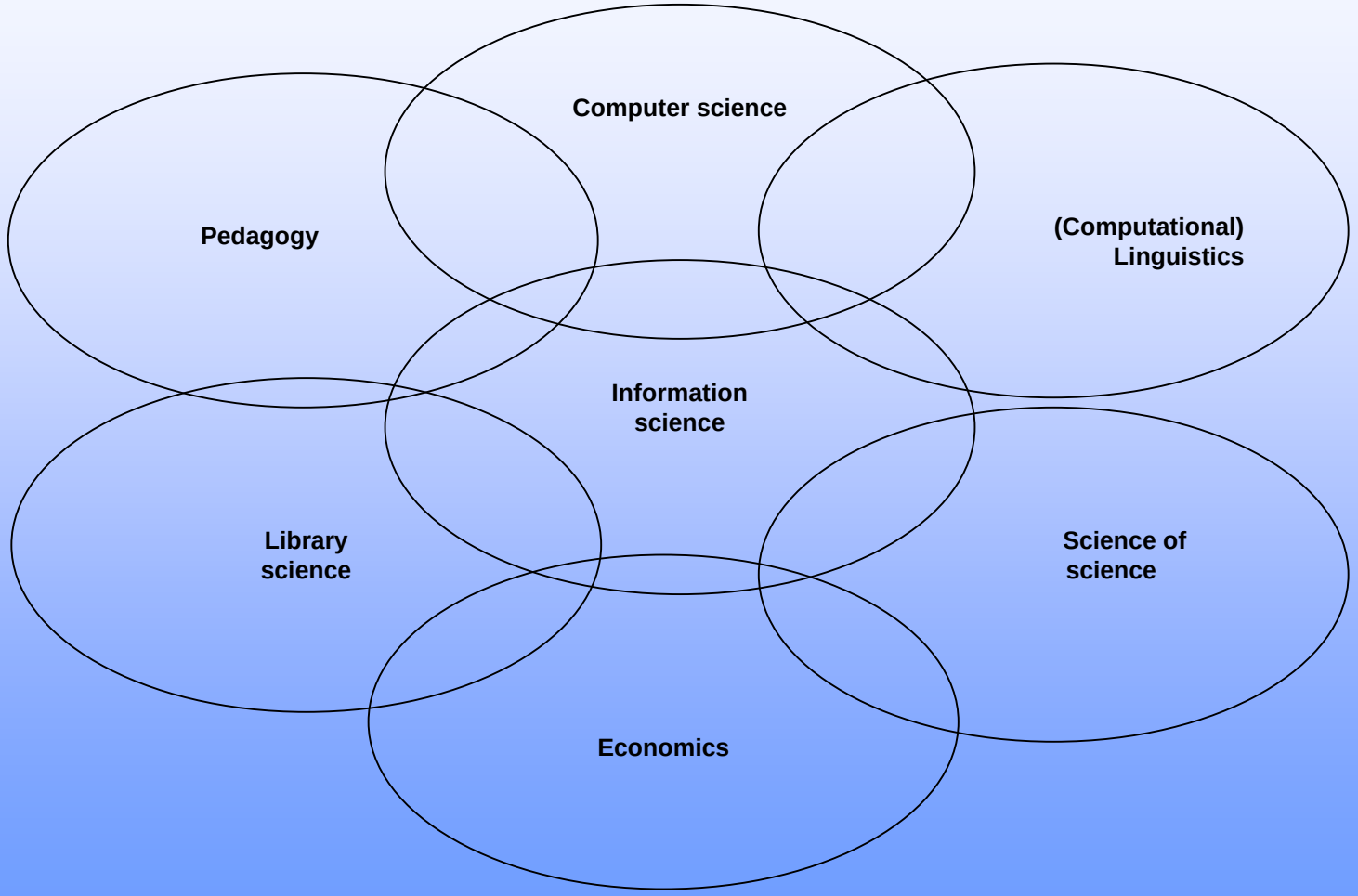
11 Jesus antwortete: "Euch hat Gott die Geheimnisse seines Planes erkennen lassen, nach dem er schon begonnen hat, seine Herrschaft in der Welt aufzurichten; den anderen hat er diese Erkenntnis nicht gegeben.

12 Denn wer viel hat, dem wird noch mehr gegeben werden, so dass er übergenug haben wird. Wer aber wenig hat, dem wird auch noch das wenige genommen werden, das er hat.

13 Mit diesem Grund rede ich in Gleichnissen, wenn ich zu ihnen spreche. Denn sie sehen zwar, aber erkennen nichts; sie hören zwar, aber verstehen nichts. ...“

Der Bezug in Matthäus 13:12 ist die Erkenntnis, also Wissen. Wer viel Wissen hat, dem wird noch mehr gegeben; wer wenig Wissen hat, der verliert im Laufe der Zeit auch noch das wenige.

A.1 Was ist Informationswissenschaft?



A.1 Was ist Informationswissenschaft?

Historische Stränge der Informationswissenschaft

Wissensrepräsentation	**ab Ende 19. Jhd.**
Informetrie	**ab Anfang 20. Jhd.**
Information Retrieval	**ab ca. 1950 Boom ab 1995**
Informationsmarkt	**ab 60er Jahre**
Wissensmanagement	**ab 80er Jahre**

Kapitel A.2

Wissen und Information

A.2 Wissen und Information

Grundbegriffe

Signale

Zeichen

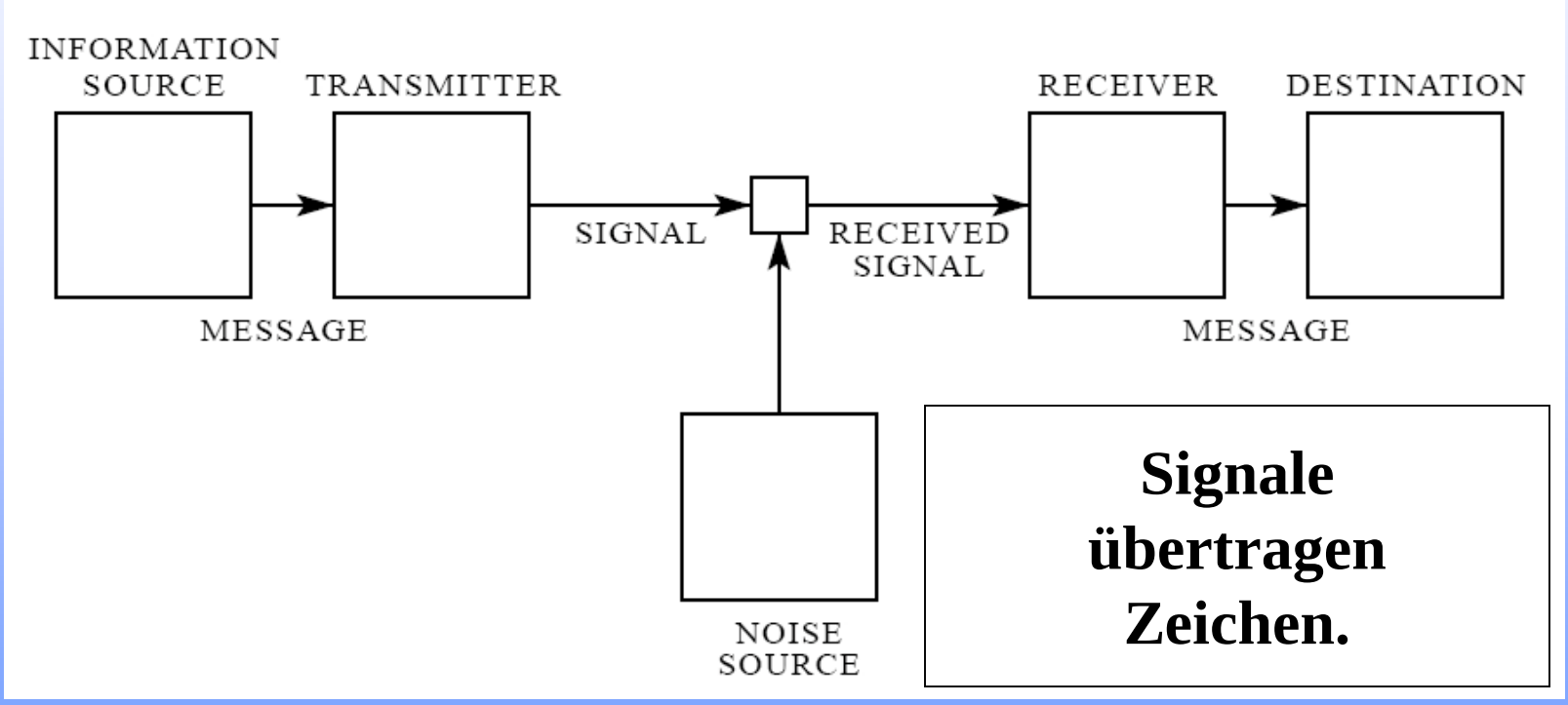
Daten

Information

Wissen

A.2 Wissen und Information

Signalübertragung nach Shannon



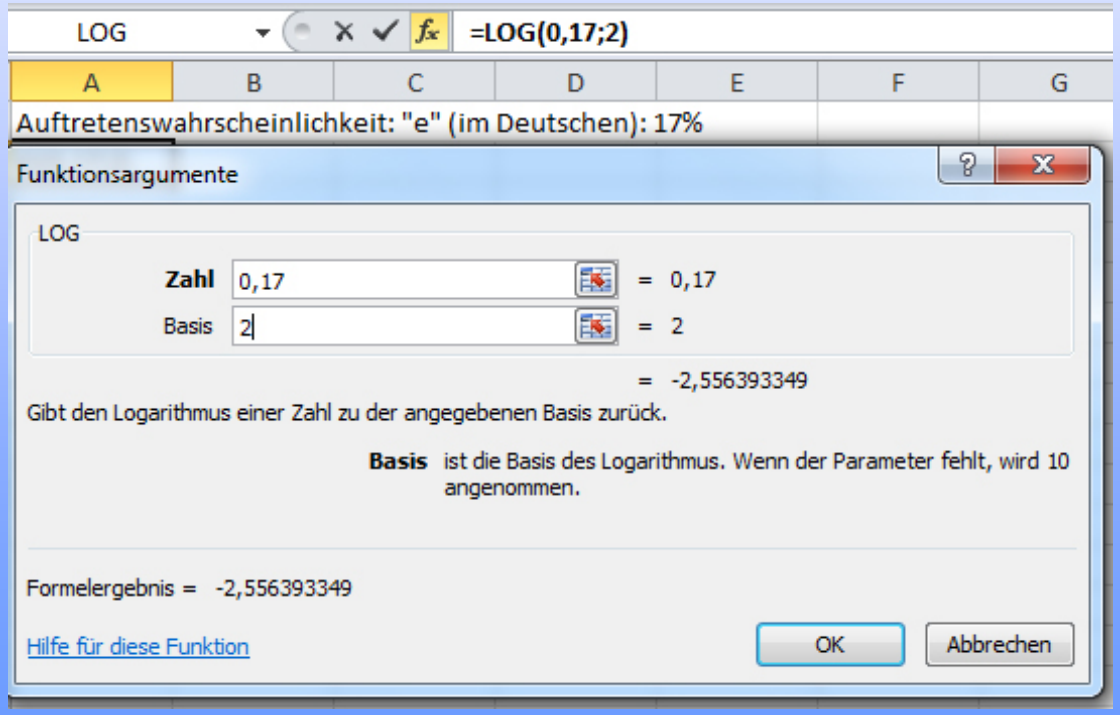
Signale übertragen Zeichen.

Shannon, C. (2001 [1948]). A mathematical theory of communication.
ACM SIGMOBILE Mobile Computing and Communications Review, 5(1), 3-55. (Original: 1948).

A.2 Wissen und Information

Informationsgehalt eines Zeichens

$$I(z_i) = \text{ld } 1/p_i \text{ bit} =$$
$$I(z_i) = -\text{ld } p_i \text{ bit}$$



A.2 Wissen und Information

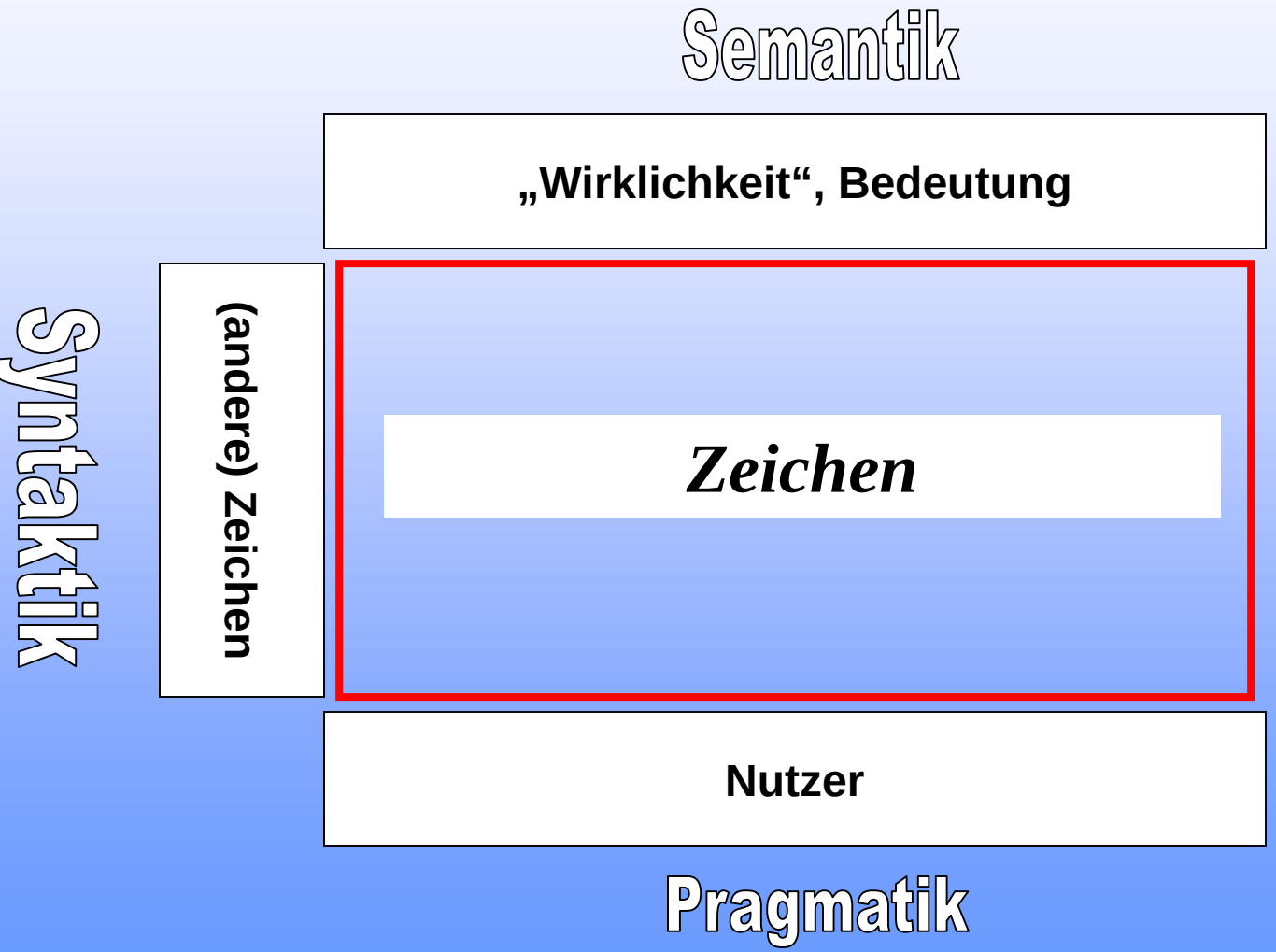
Informationsgehalt eines Zeichens

- **Der Informationsgehalt eines Zeichens ist abhängig von der Auftretenswahrscheinlichkeit des Zeichens (je seltener desto größer)**

- **Shannon: "The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem".**

- **Also: Shannons Theorie hat wenig Relevanz für die Informationswissenschaft**

A.2 Wissen und Information



A.2 Wissen und Information

Daten und Information

Daten: Übertragene Zeichenfolgen (möglicherweise mittels kryptographischer Methoden verschlüsselt)

Information: Daten mit Bedeutung (natürlich unverschlüsselt)

A.2 Wissen und Information

Wissen

(1) Wissen als Fähigkeit,

(1a) einen Gegenstand korrekt zu erfassen (englisch: "to know *that*")

(1b) mit einem Gegenstand korrekt umzugehen (englisch: "to know *how*");

(2) Wissen als Zustand

(2a) einer Person, die (etwas) weiß,

(2b) das Gewusste selbst, der Inhalt (englisch: "content")

(2c) der sprachliche Ausdruck davon

rot: an ein Subjekt gebundenes Wissen

schwarz: objektives Wissen

A.2 Wissen und Information

Wissen

im Sinne von Karl R. Popper

- **physikalische Welt (Poppers „Welt 1“)**
- **subjektives Wissen (an menschliches Denken gebunden; Poppers „Welt 2“)**
- **objektives Wissen (unabhängig von menschlichem Denken gespeichert; Poppers „Welt 3“)**

Popper, K.R. (1973 [1972]). Objektive Erkenntnis. Ein evolutionärer Entwurf. Hamburg: Hoffmann und Campe. (Original: 1972).

A.2 Wissen und Information

Information

- **Wie geschieht der Übergang aus Welt 3 in Welt 2? D.h. wie nimmt ein Subjekt objektives Wissen auf?**
- **Wie geschieht der Übergang innerhalb von Welt 2? D.h. wie nimmt ein Subjekt Wissen anderer Subjekte auf?**
- **Der formlose Content muss jeweils in eine Form gegossen werden. Also: in-FORM-ation**

Capurro, R. (1978). Information. Ein Beitrag zur etymologischen und ideengeschichtlichen Begründung des Informationsbegriffs. München: Saur.

Capurro, R. (2000). Einführung in den Informationsbegriff. Online: http://www.Capurro.de/infovorl-index.htm.

A.2 Wissen und Information

Information

- **„Wissen in Aktion“ (gemäß Kuhlen)**
- **aber auch: andere epistemische Gegenstände (z.B. Annahmen oder Lügen) „in Aktion“**
- **Wissen: statisch – Information: in Bewegung**
- **Information: gebunden an Signale (die sorgen für die Bewegung)**
- **Wissen: mit Wahrheitsanspruch – Information: ohne Wahrheitsanspruch**
- **Information: „Neuheit“ als Mittelweg zwischen Erstmaligkeit und Bestätigung**

Kuhlen, R. (1995). Informationsmarkt. Chancen und Risiken der Kommerzialisierung von Wissen. Konstanz: UVK. (Schriften zur Informationswissenschaft; 15).

A.2 Wissen und Information

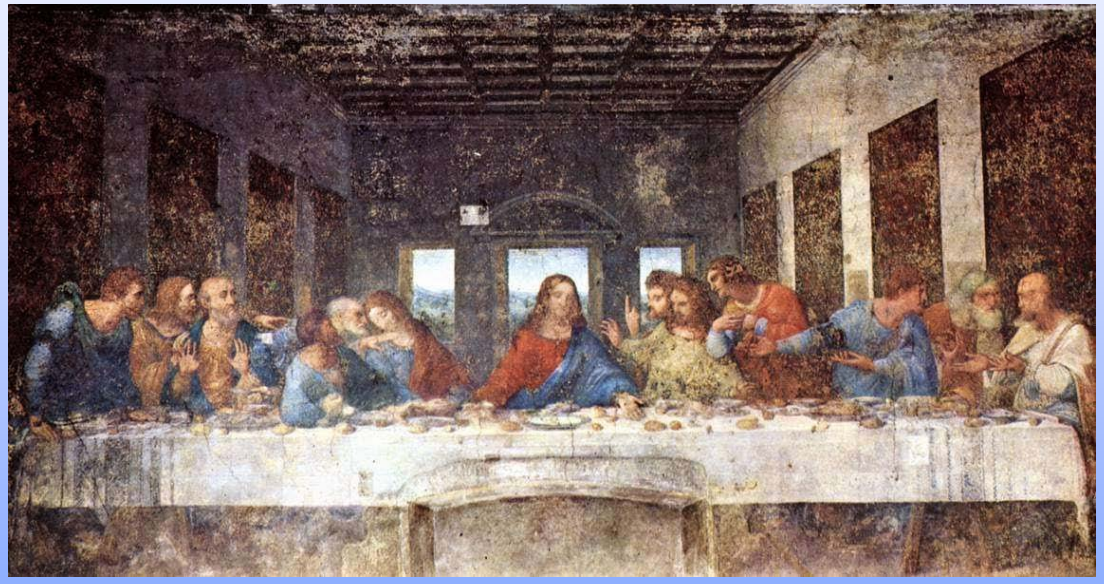
Information		
	intangibel	**tangibel**
Zustand	Wissen	Information als Ding (Dokument)
Prozess	Informations-prozess	Informationsverarbeitung (EDV)

Buckland, M.K. (1991). Information as thing.
Journal of the American Society for Information Science, 42(5), 351-360.

A.2 Wissen und Information

Wissen in Bildern (Panofsky) - analog auch bei Videos, Musik, Belletristik

- **prä-ikonographische semantische Ebene**
- **ikonographische Ebene**
- **ikonologische Ebene**



Panofsky, E. (1975). Sinn und Deutung in der bildenden Kunst. Köln: DuMont.
Panofsky, E. (2006). Ikonographie und Ikonologie. Köln: DuMont.
(darin: Zum Problem der Beschreibung und Inhaltsdeutung von Werken der bildenden Kunst, 5-32 [Original: 1932], Ikonographie und Ikonologie, 33-60 [Original: 1955]).

A.2 Wissen und Information

Knowing that – Knowing how

Knowing that: Wissen als wahre Aussage (Chisholm):

h wird von *S gewusst* =df *h* wird von *S* akzeptiert; *h* ist wahr; und *h* ist für *S* fehlerlos evident (*h* ist eine Aussage und *S* ein Subjekt)

Knowing how (Ryle)

a) Eine Person weiß implizit, gewisse Dinge zu tun (z.B.: gute Witze machen; sich auf Beerdigungen richtig zu benehmen)

b) Eine Person weiß aufgrund von gewissen Prinzipien, Dinge zu tun (z.B.: ein Omelett kochen; Kleider schneidern)

Chisholm, R.M. (1979[1977]). Erkenntnistheorie. München: dtv Wissenschaft. (Original: 1977).
Ryle, G. (1946). Knowing how and knowing that. Proceedings of the Aristotelian Society, 46, 1-16.

A.2 Wissen und Information

Subjektives implizites Wissen (Polanyi)

- *"we can know more than we can tell*"
- proximales implizites Wissen (nahe liegendes Wissen), z.B. Benutzung eines Werkzeuges
  - Weitergabe: 1. körperlich durch Vor- und Nachmachen (z.B. Handwerk)
  - 2. intellektuell durch Nach-Denken (z.B. Schachspielen)
- distales implizites Wissen (entfernt liegendes Wissen), z.B. Erinnerung an eine (nur einmal gesehene) Person; Beschreibung: kaum erschöpfend möglich - aber Erinnerung (beim Vorlegen eines Bildes)
  - Weitergabe: problematisch
- optimal: Externalisierung (Fassung des subjektiven Wissens in objektiver Form; z.B. schriftlich fixiert) - allerdings nicht immer machbar

Polanyi, M. (1967). The Tacit Dimension. Garden City, NY: Doubleday (Anchors Books).

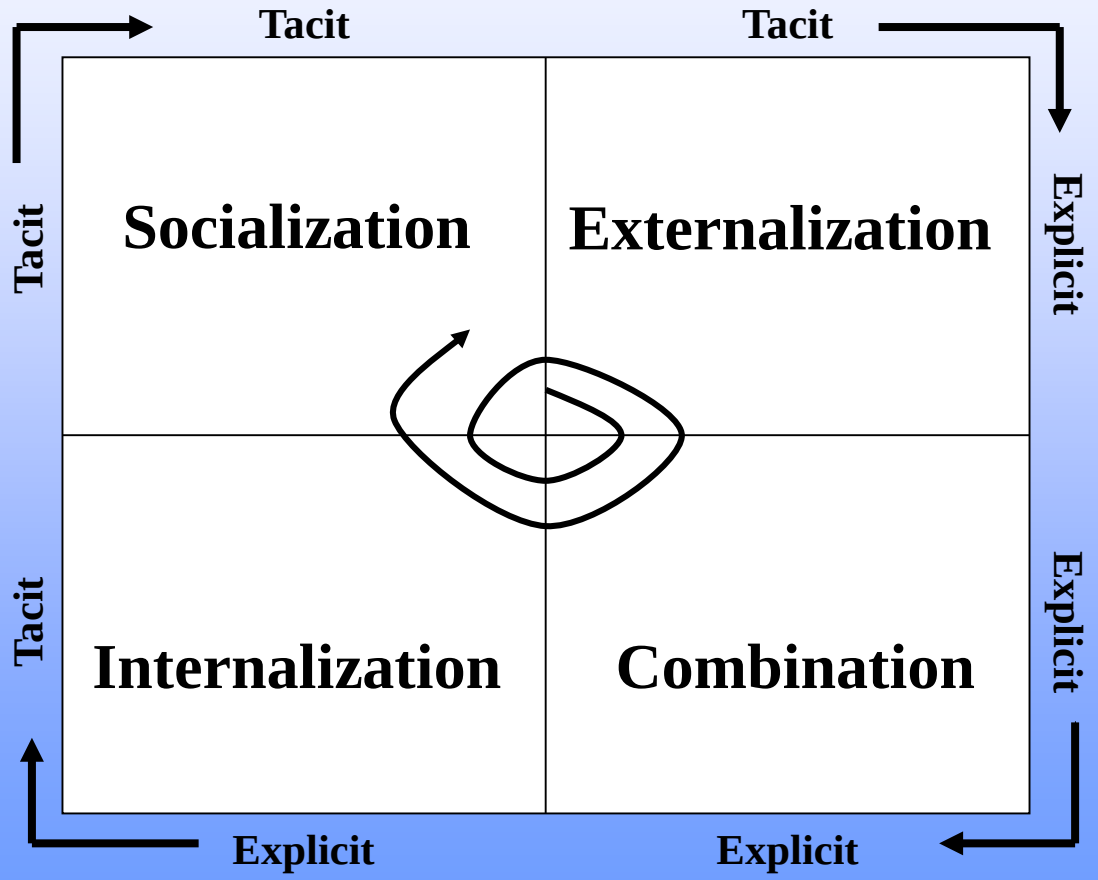
A.2 Wissen und Information

Wenn Externalisierung nicht möglich ist:

- Wissensträger beschreiben (z.B. Expertendatenbank, Yellow Pages) nutzlos, wenn Person das Unternehmen verlassen hat
- Artefakte des Wissensträgers beschreiben (z.B. Video, Foto); nutzlos, wenn kein anderer mit dem Video etwas anfangen kann
- "Skill-Management"; nutzlos, wenn keine andere Person über dieselben Fähigkeiten wie der Wissensträger verfügt
- beim Wissensträger "in die Lehre" gehen

A.2 Wissen und Information

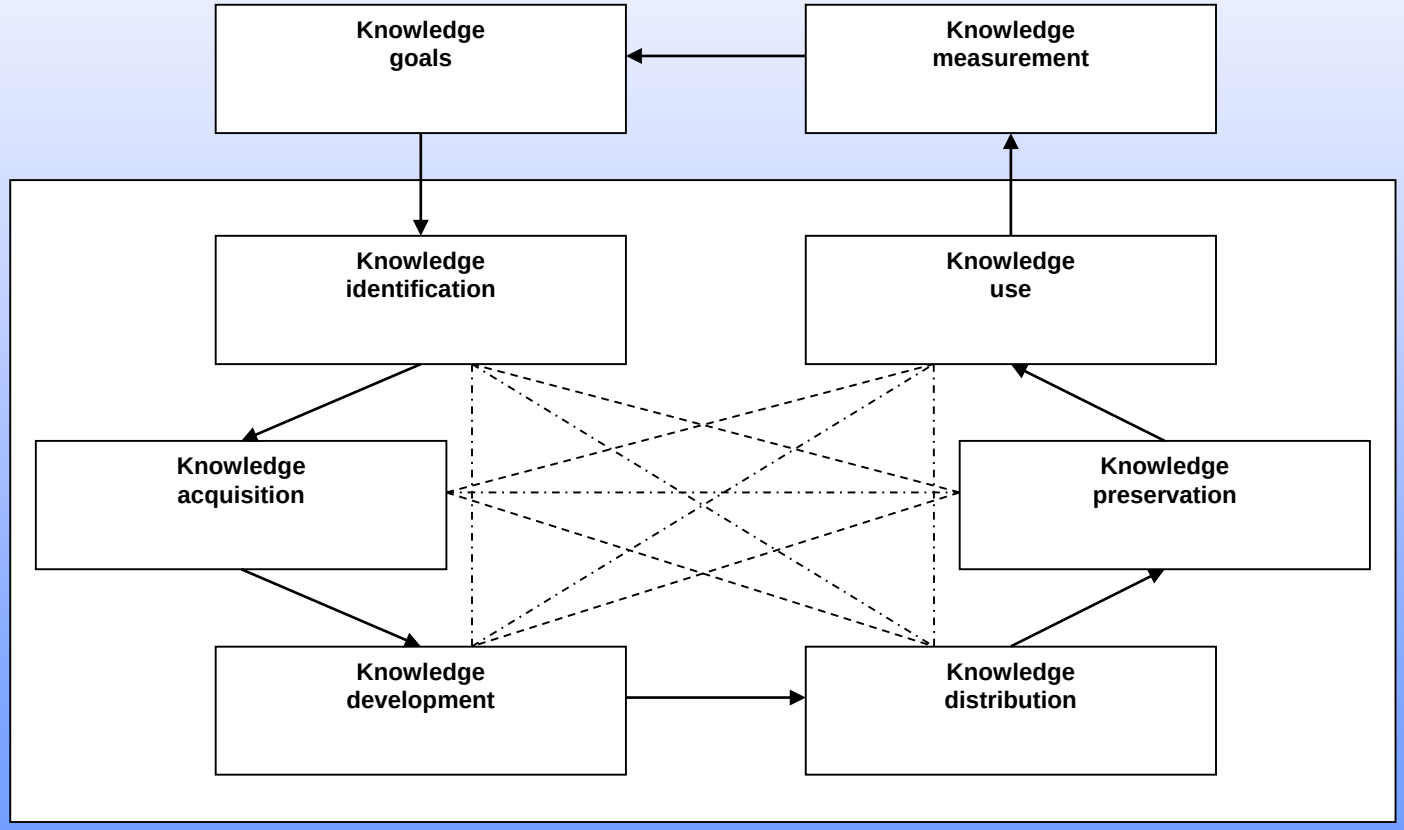
Wissensmanagement



Nonaka, I., & Takeuchi, H. (1995). The Knowledge-Creating Company.
How Japanese Companies Create the Dynamics of Innovation. Oxford: Oxford University Press.

A.2 Wissen und Information

Wissensmanagement



Probst, G.J.B., Raub, S., & Romhardt, K. (2000). Managing Knowledge. Buildings Blocks for Success. Chichester: Wiley.

A.2 Wissen und Information

Wissensarten nach Spinner

- E-Wissen versus U-Wissen
- E-Wissen: wissenschaftliche Dokumente, Patente, Rechtsakte, Nachrichten (aus seriösen Quellen), ...
- U-Wissen: Bilder, Videos, (Micro-)Blogs, persönliche Webseiten, Facebook-Posts, ...

Spinner, H.F. (1994). Die Wissensordnung. Ein Leitkonzept für die dritte Grundordnung des Informationszeitalters. Opladen: Leske + Budrich.

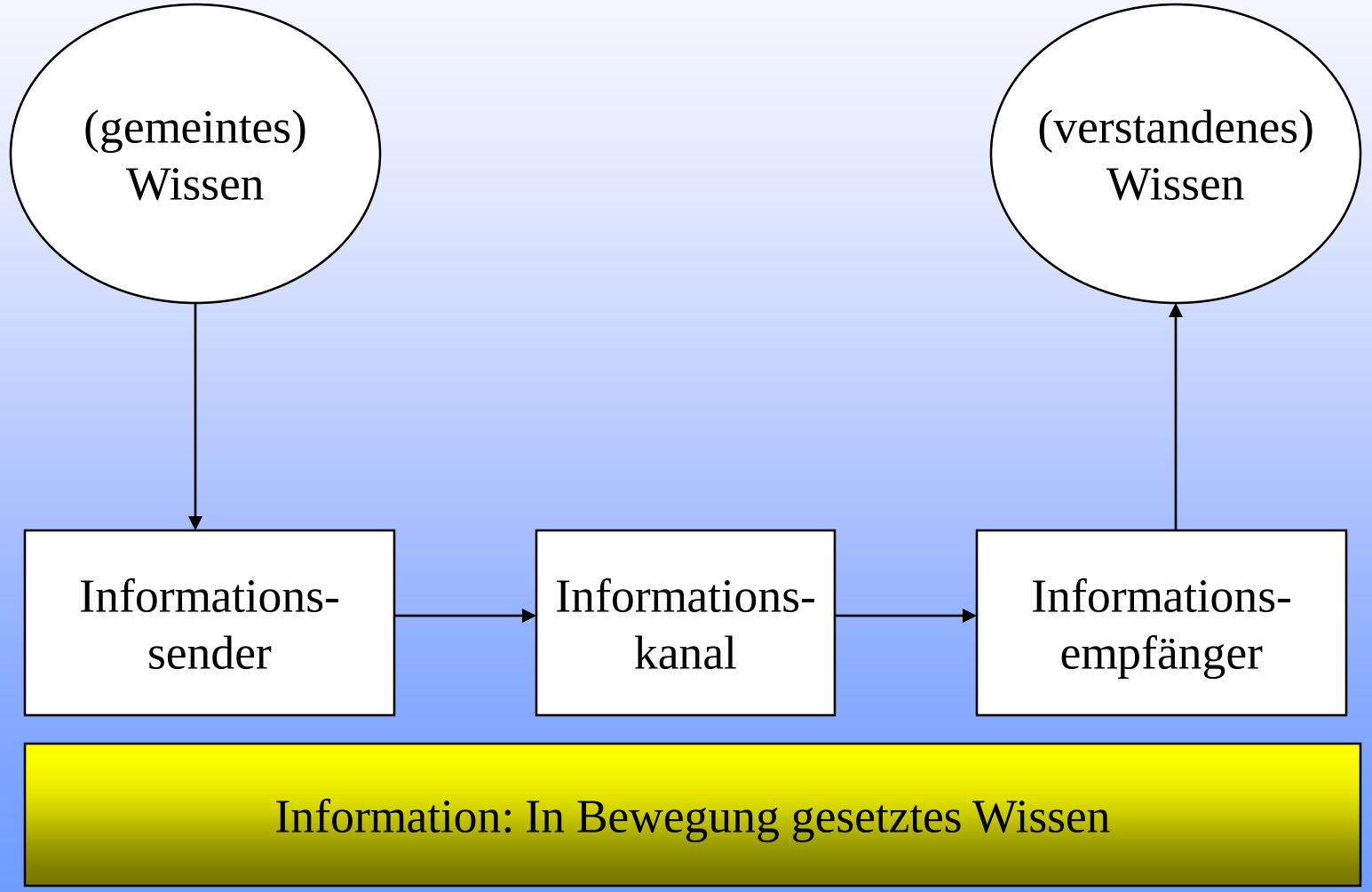
A.2 Wissen und Information

Normalwissenschaftliches Wissen nach Kuhn

- prä-paradigmatische Phase (Wissenschaftler als "Einzelkämpfer"; keine einheitliche Terminologie)
- paradigmatische Phase, "Normalwissenschaft"; einheitliche Terminologie
- Paradigmenwechsel, "wissenschaftliche Revolution"; Auswechslung der Terminologie

Kuhn, T.S. (1979[1962]). Die Struktur wissenschaftlicher Revolutionen. 4. Aufl. Frankfurt: Suhrkamp. (Original: 1962).

A.2 Wissen und Information



A.2 Wissen und Information

Information und Wissen

$$K[S] + \Delta I = K[S + \Delta S].$$

K	**Wissen (knowledge) eines Subjektes**
S	**Struktur (von Begriffen und Aussagen)**
Δ I	**übertragene Information**
Δ S	**Strukturänderung (auf der Basis von S)**

Δ I kann bei verschiedenen Subjekten unterschiedliche Strukturänderungen Δ S bewirken; Δ S kann 0 sein.

Brookes, B.C. (1980). The foundations of information science. Part I. Philosophical aspects. Journal of Information Science, 2, 125-133.

A.2 Wissen und Information

Information, Wissen und Intermediation

Wissen wird durch Information in Bewegung gesetzt

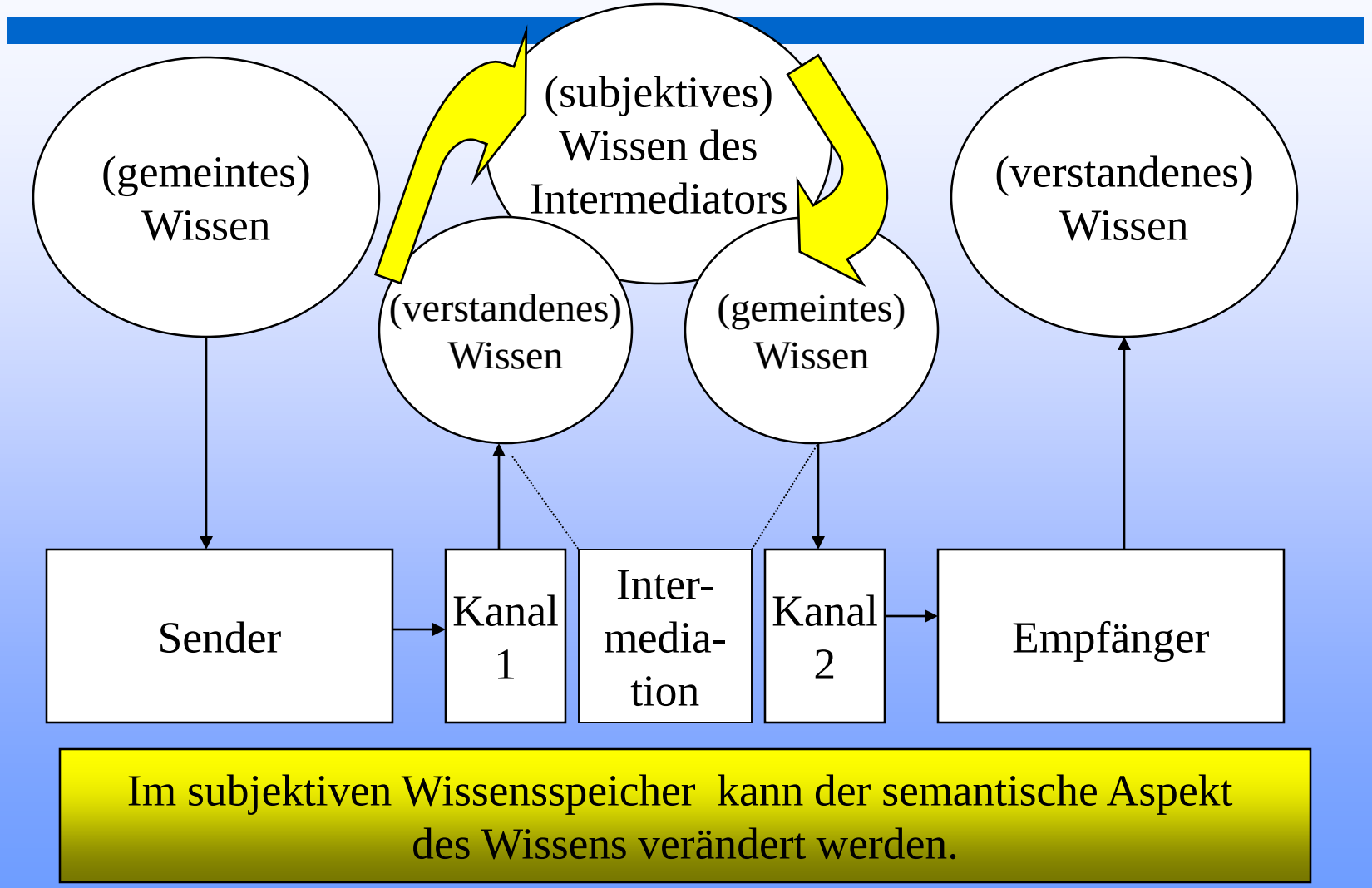
Information wird (außer im direkten Gespräch) vermittelt:

durch einen (menschlichen) Informationsvermittler (mit dessen subjektivem Wissen)

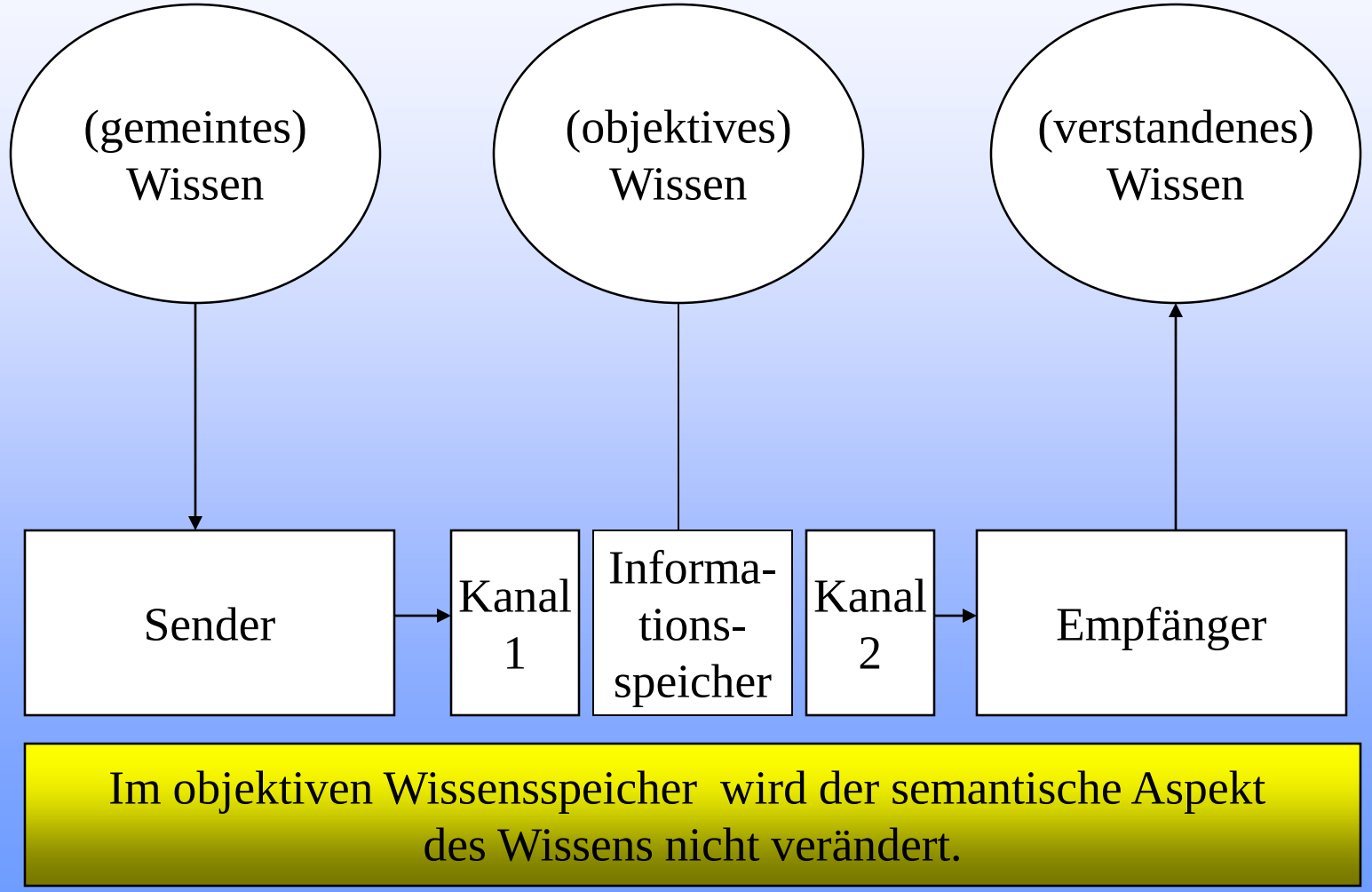
oder

durch einen objektiven Wissensspeicher (z.B. eine Suchmaschine)

A.2 Wissen und Information



A.2 Wissen und Information



A.2 Wissen und Information

„Wissen ist Macht“

- **Francis Bacon: „(D)ie Natur wird nur besiegt, indem man ihr gehorcht. Daher fallen jene Zwillingsziele, die menschliche Wissenschaft und Macht, zusammen ...“**
- **Wissen als Macht über die Natur**
- **Wissen auch als Macht über andere Menschen (oder andere Unternehmen)?**

Bacon, F. (1990 [1620]). Novum Organum = Neues Organon, hrsg. v. W. Krohn. Hamburg: Meiner. (Original: 1620), S. 65.

A.2 Wissen und Information

Welche Art Macht bekommt man durch Informationswissenschaft?

Ansatz von Waldemar Wittmann (1959):

- **Ausgang: Unsicherheit (Diskrepanz zwischen als notwendig erachteter und tatsächlich vorhandener Information)**
- **Lösung: Man sucht (und findet) zweckorientiertes Wissen**
- **auch bei Kuhlthau (2004): Uncertainty Principle**

Wittmann, W. (1959). Unternehmung und unvollkommene Information. Köln, Opladen: Westdeutscher Verlag.
Kuhlthau, C.C. (2004). Seeking Meaning. A Process Approach to Library and Information Services. 2nd Ed. Westport, CT: Libraries Unlimited.

A.2 Wissen und Information

Welche Art Macht bekommt man durch Informationswissenschaft?

Ansatz von Gernot Wersig (1974):

- **Ausgang: problematische Situation (analog zu Wittmann)**
- **Lösung: Reduktion von Unsicherheit durch Kommunikationsprozesse**

Wersig, G. (1974). Information – Kommunikation – Dokumentation. Darmstadt: Wissenschaftliche Buchgesellschaft.

A.2 Wissen und Information

Welche Art Macht bekommt man durch Informationswissenschaft?

Ansatz von N.J.Belkin et al. (1980/1982):

- **Ausgang: anomalous state of knowledge (ASK)**
- **Lösung: Information Retrieval**

Belkin, N.J. (1980). Anomalous states of knowledge as a basis for information retrieval. Canadian Journal of Information Science, 5, 133-143.
Belkin, N.J., Oddy, R.N., & Brooks, H.M. (1982). ASK for information retrieval. Journal of Documentation, 38, 61-71 (part 1); 145-164 (part 2).

A.2 Wissen und Information

ASK

- **bei Entscheidungsvorbereitungen**
- **bei Wissenslücken**
- **bei Frühwarnfällen**

- **Informationen reduzieren hierbei die Unsicherheiten**
- **Informationen führen zu Handlungen (oder zu Unterlassungen)**

- **praktisches Ziel von Nutzern:**
- **Wissen über Informationen in Handlungen umsetzen**

A.2 Wissen und Information

Handlungsrelevantes Wissen:
Wissensautonomie - Informationsautonomie

- **Hat ein Nutzer das handlungsrelevante Wissen, so ist er *wissensautonom*.**
- **Hat er die Fähigkeiten, sich das handlungsrelevante Wissen zu beschaffen, so ist er *informationsautonom*.**
- **Ziel: Schaffen von *Informationskompetenz* („information literacy“)**

Kuhlen, R. (2004). Information. In R. Kuhlen, T. Seeger, & D. Strauch (Eds.), Grundlagen der praktischen Information und Dokumentation (pp. 3-20). 5th Ed. München: Saur.

A.2 Wissen und Information

Unvollständige Information

- **Wissen ist nie ideal komplett (Reduktion der Unsicherheit auf Null ist nicht möglich)**
- **Die Vergrößerung der Informationsbasis birgt jedoch große Vorteile**
- **Allerdings nur solange, bis die Wettbewerber über die selbe Information verfügen**
- **Folge: kreative Unsicherheit**
- **Information führt so zu innovativem Wettbewerb**

A.2 Wissen und Information

Information als Wirtschaftsgut

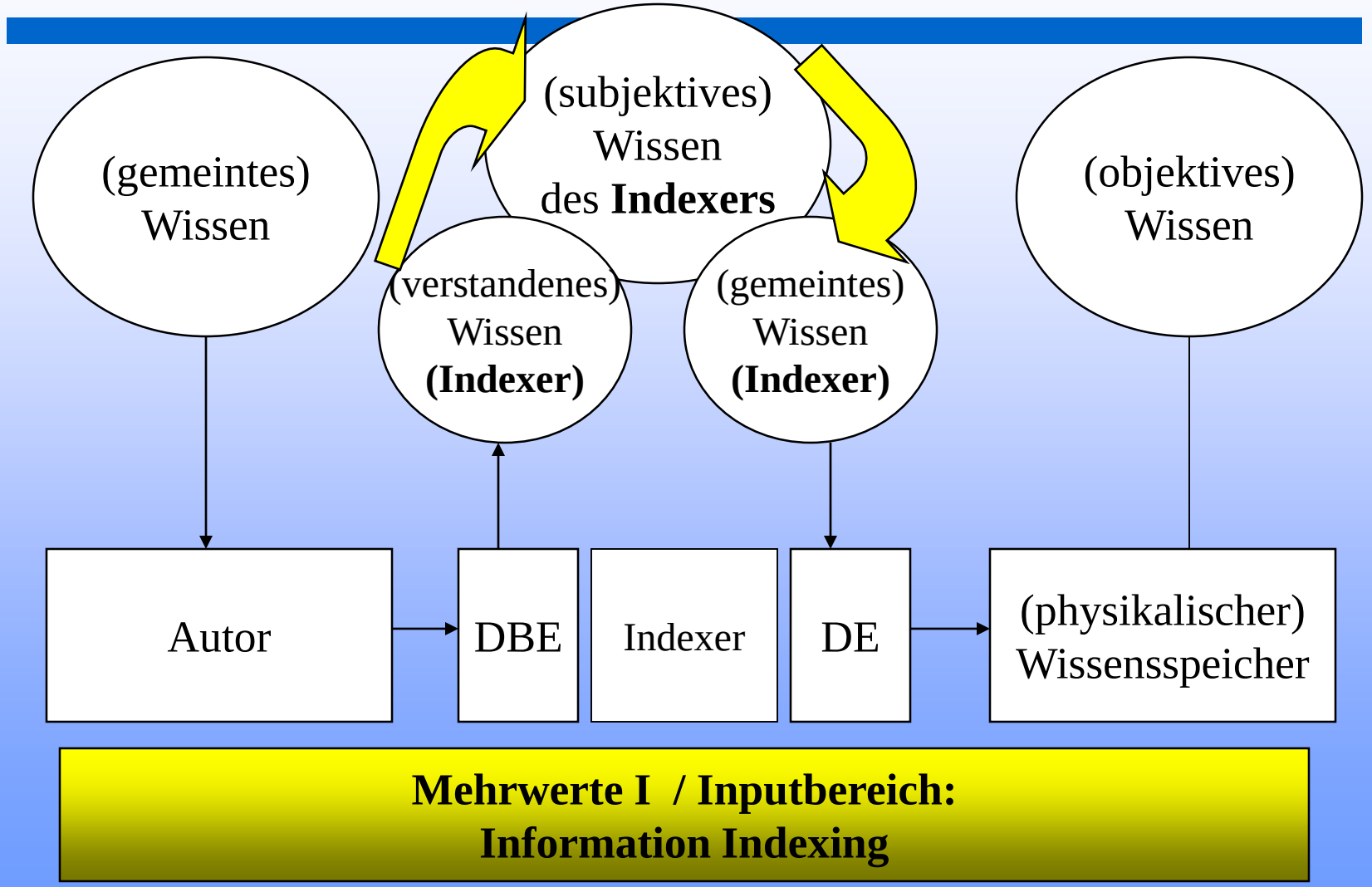
- **Information ist eine Ware**
- **allerdings eine Ware besonderer Art: auch nach einem Verkauf verfügt der Verkäufer noch über die Ware; es werden nur Kopien vertrieben**
- **Zahlungsbereitschaft je nach Nutzer stark unterschiedlich**
- **Preise für Content ebenfalls stark unterschiedlich**
  - **kostenlos (z.B. bei Web-Suchmaschinen); Finanzierung durch Werbung (z.B. Google AdWords) oder durch Universaldienst (z.B. DPMA oder Medline)**
  - **kostenpflichtig: Fachinformationen (teuer bis prohibitiv überteuert)**

A.2 Wissen und Information

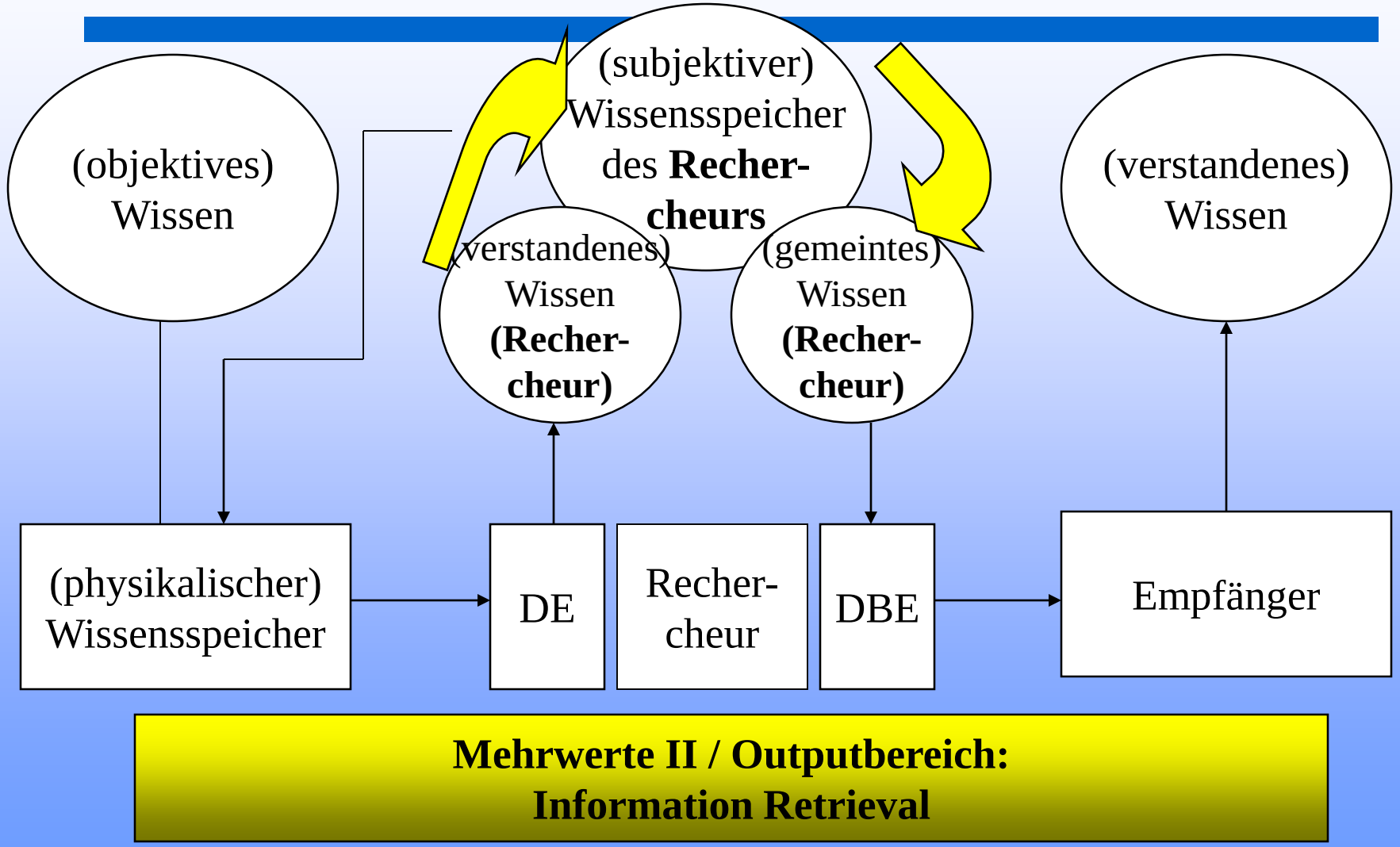
Informationelle Mehrwerte

- **Dokumente (z.B. Patentschrift, wissenschaftlicher Artikel, Zeitungsartikel) – „dokumentarische Bezugseinheiten“ (DBE)**
- **bearbeitete Dokumente: Dokumente sowie das darin enthaltene Wissen werden such- und findbar als „Dokumentationseinheiten“ (DE) aufbereitet und in speziellen System gespeichert**
  → Erarbeitung informationeller Mehrwerte beim Input (Information Indexing)
- **Bereitstellen von Recherchesystemen**
  → Erarbeitung informationeller Mehrwerte beim Output (Information Retrieval)
- **informationelle Mehrwerte: *Knowing about***

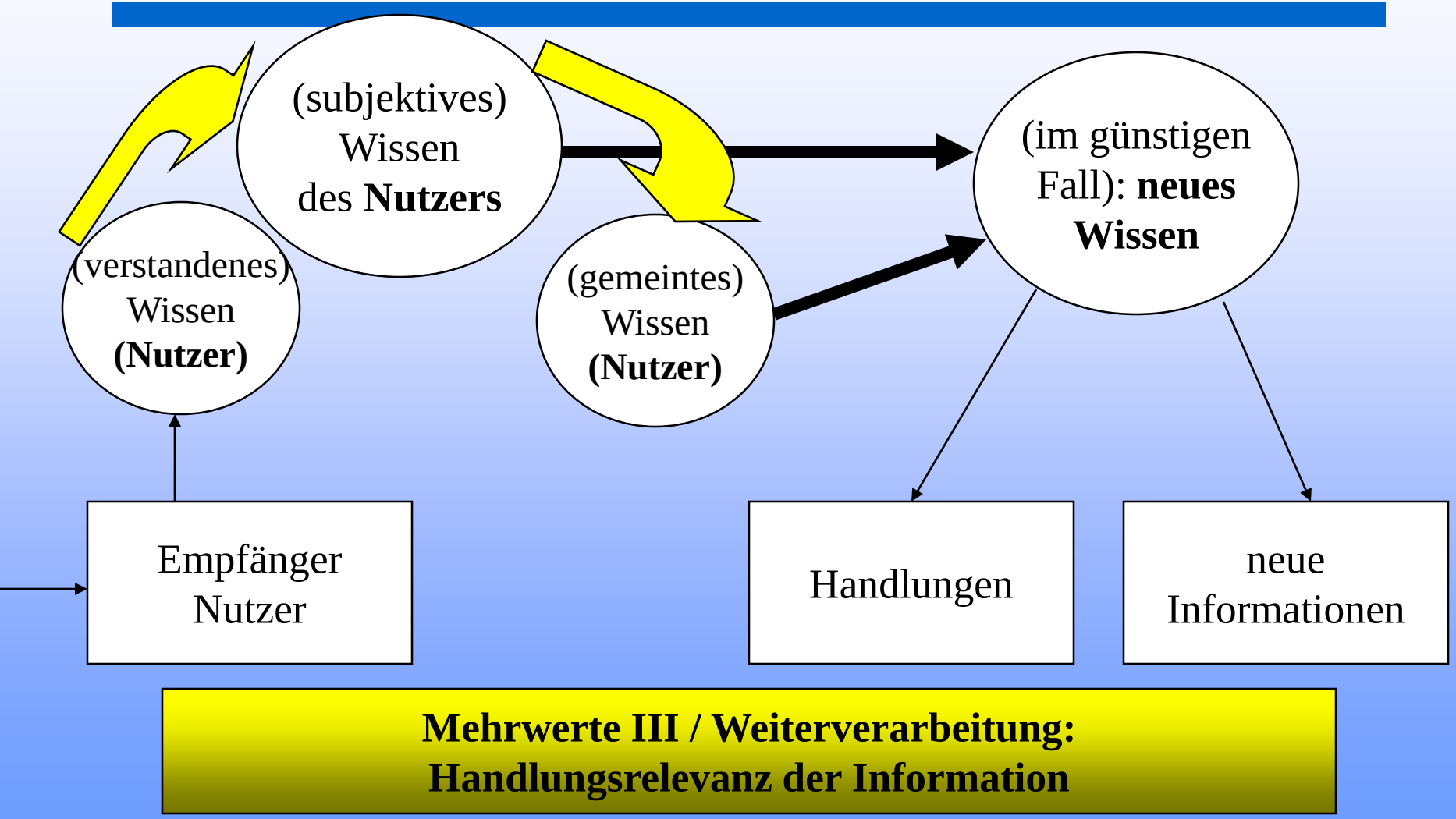
A.2 Wissen und Information



A.2 Wissen und Information



A.2 Wissen und Information



Kapitel A.3

Information und Verstehen

A.3 Information und Verstehen

- **Hermeneutik: Lehre vom (richtigen) Verstehen**

- **Informationshermeneutik: Lehre von den Verstehensprozessen bei der Wissens-repräsentation und im Information Retrieval**
  - **beim Aufbau von Begriffsordnungen**
  - **beim Indexieren und Referieren**
  - **beim Recherchieren**
  - **beim Design eines Informationssystems**

Gadamer, H.G. (1974). Hermeneutik. In J. Ritter (Ed.), Historisches Wörterbuch der Philosophie. Band 3 (pp. 1061-1074). Darmstadt: Wissenschaftliche Buchgesellschaft.

Gadamer, H.G. (1975). Wahrheit und Methode. 4. Aufl. Tübingen: J.C.B. Mohr (Paul Siebeck).

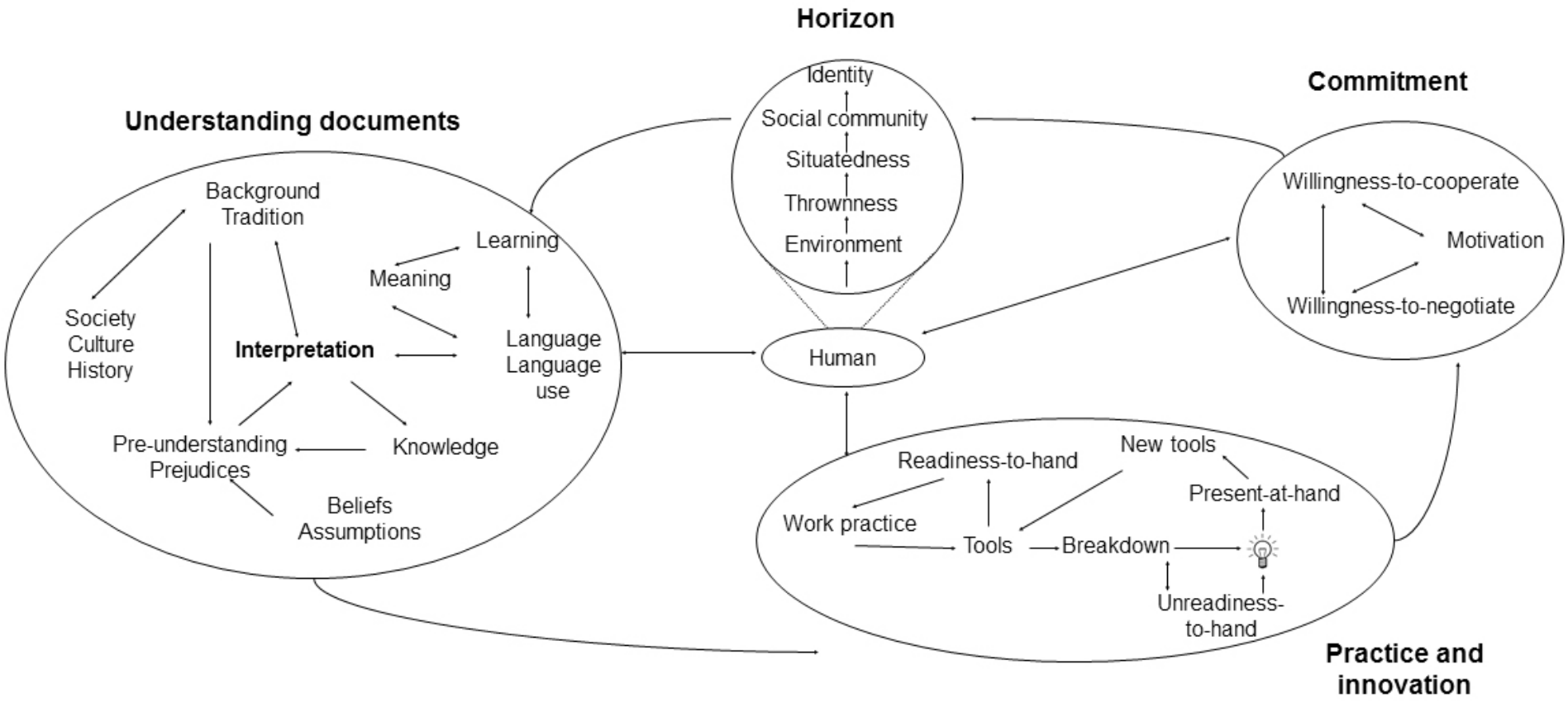
Heidegger, M. (1967[1927]). Sein und Zeit. 11. Aufl. Tübingen: Niemeyer. [Original: 1927].

A.3 Information und Verstehen

Verstehen

- **etwas *als etwas* verstehen**
- **Finden und Anwenden des "passenden" Auslegungsschlüssels**
- **Konzentration auf den überlieferten Text (d.h. den Gegenstand) und nicht etwa darauf, was der Autor möglicherweise "meinte"**
- **dabei: Rekonstruktion der Frage, auf die der Text eine Antwort ist**
- **das Verstehen der Teile bedingt das Verstehen des Ganzen, *und* das Verstehen des Ganzen bedingt das Verstehen der Teile (hermeneutischer Zirkel)**
- **Text und Leser stehen in unterschiedlichen Horizonten, die im Verstehen verschmolzen werden**
- **dabei ist das eigene Vorverständnis des Lesers positiv – als Vorurteil – einzuschätzen, sofern es dynamisch (im hermeneutischen Zirkel) dazu beiträgt, die Horizontverschmelzung voranzutreiben**

A.3 Information und Verstehen



A.3 Information und Verstehen

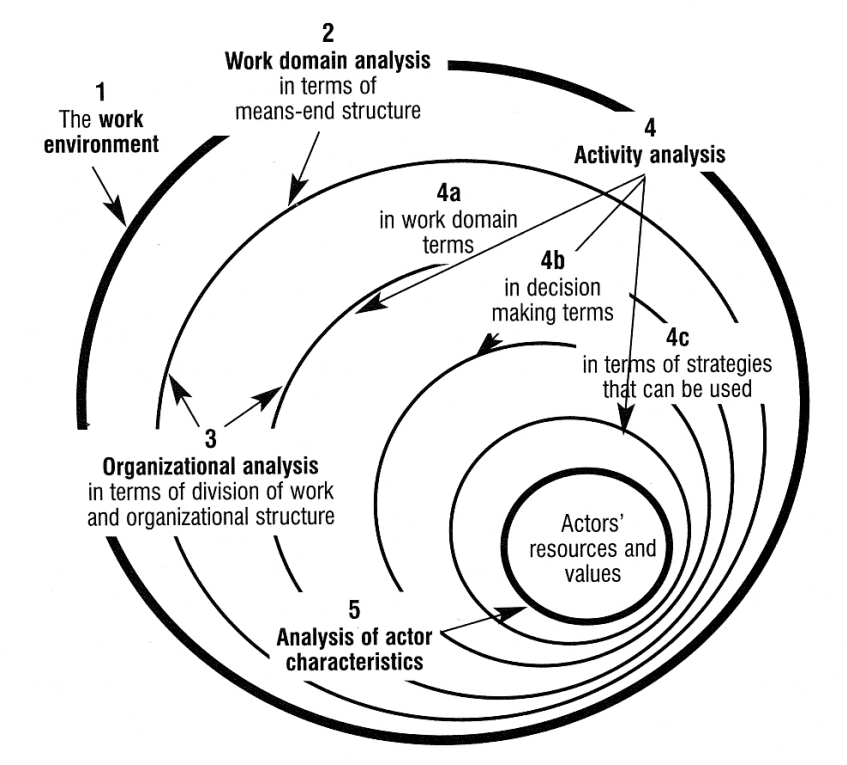
Informationsarchitektur

- **bei der Konstruktion und beim Design von Informationssystemen:**
  - **Basis: sozialer Hintergrund / Sprache der Beteiligten**
  - **Hilfsmittel: Computer**
  - **stets zu berücksichtigen: Tradition und Sprache (unterschiedliche "Traditionen" und "Sprachen" innerhalb einer Institution)**
  - **methodisches Konstrukt: Systemzusammenbruch ("breakdown")**
  - **solange Prozesse und Systeme reibungslos laufen, werden sie nicht infrage gestellt - erst beim Breakdown beginnt das Nachfragen**

Winograd, T., & Flores, F. (1986). Understanding Computers and Cognition: A New Foundation for Design. Norwood, NJ: Ablex.

A.3 Information und Verstehen

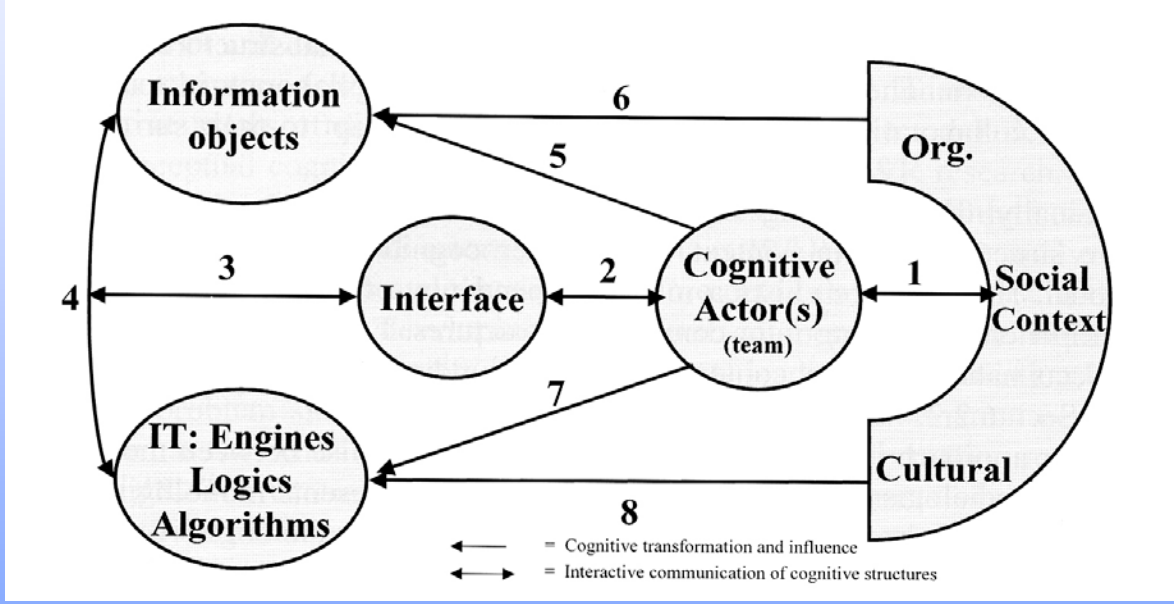
Cognitive Work Analysis (CWA)



Rasmussen, J., Pejtersen, A.M., & Goodstein, L.P. (1994). Cognitive Systems Engineering. New York, NY: Wiley.
Vicente, K. (1999). Cognitive Work Analysis. Mahwah, NJ: Lawrence Erlbaum Associates.

A.3 Information und Verstehen

Kognitive Arbeit im Information Retrieval



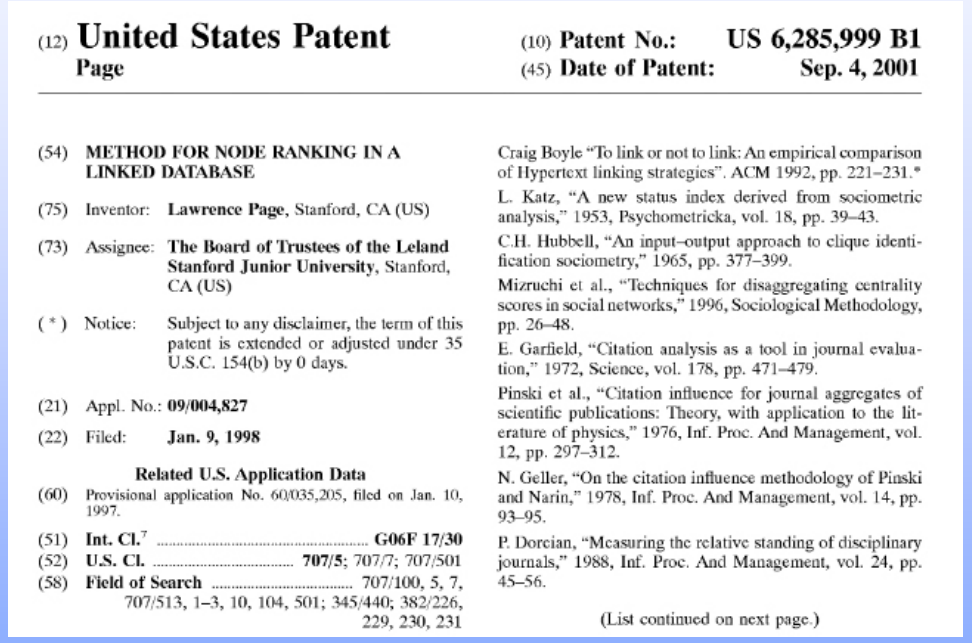
Ingwersen, P., & Järvelin, K. (2005). The Turn. Integration of Information Seeking and Retrieval in Context. Dordrecht: Springer.

Kapitel A.4

Dokumente

A.4 Dokumente

Dokumente?



(12) **United States Patent**
Page

(10) **Patent No.: US 6,285,999 B1**
(45) **Date of Patent: Sep. 4, 2001**

(54) **METHOD FOR NODE RANKING IN A LINKED DATABASE**

(75) Inventor: **Lawrence Page**, Stanford, CA (US)

(73) Assignee: **The Board of Trustees of the Leland Stanford Junior University**, Stanford, CA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **09/004,827**

(22) Filed: **Jan. 9, 1998**

Related U.S. Application Data

(60) Provisional application No. 60/035,205, filed on Jan. 10, 1997.

(51) **Int. Cl.**[7] **G06F 17/30**
(52) **U.S. Cl.** **707/5**; 707/7; 707/501
(58) **Field of Search** 707/100, 5, 7, 707/513, 1–3, 10, 104, 501; 345/440; 382/226, 229, 230, 231

Craig Boyle "To link or not to link: An empirical comparison of Hypertext linking strategies". ACM 1992, pp. 221–231.*

L. Katz, "A new status index derived from sociometric analysis," 1953, Psychometricka, vol. 18, pp. 39–43.

C.H. Hubbell, "An input–output approach to clique identification sociometry," 1965, pp. 377–399.

Mizruchi et al., "Techniques for disaggregating centrality scores in social networks," 1996, Sociological Methodology, pp. 26–48.

E. Garfield, "Citation analysis as a tool in journal evaluation," 1972, Science, vol. 178, pp. 471–479.

Pinski et al., "Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics," 1976, Inf. Proc. And Management, vol. 12, pp. 297–312.

N. Geller, "On the citation influence methodology of Pinski and Narin," 1978, Inf. Proc. And Management, vol. 14, pp. 93–95.

P. Dorcian, "Measuring the relative standing of disciplinary journals," 1988, Inf. Proc. And Management, vol. 24, pp. 45–56.

(List continued on next page.)

A.4 Dokumente

Dokumente?

A.4 Dokumente

- **klar: (gedruckter, digitaler) Text ist „Dokument“**
- **aber: Sind Objekte auch Dokumente?**
- **Suzanne Briet (1951):**

Gegenstand	Dokument?
Stern am Himmel	nein
Foto des Sterns	ja
Stein in einem Fluss	nein
Stein in einem Museum	ja
Antilope in der Wildnis	nein
Antilope im Zoo	ja

Briet, S. (1951). Qu'est-ce que la documentation?
Paris: Editions Documentaires Industrielles et Techniques.

A.4 Dokumente

- **Charakteristika von Dokumenten**
  - **1. Materialität (sie sind physikalisch – einschließlich digital – vorhanden),**
  - **2. Intentionalität (sie tragen Sinn bzw. Bedeutung),**
  - **3. Erarbeitung (sie werden geschaffen),**
  - **4. Wahrnehmung (sie werden als Dokument bezeichnet)**

Buckland, M.K. (1997). What is a "document"?
Journal of the American Society for Information Science, 48, 804-809.

A.4 Dokumente

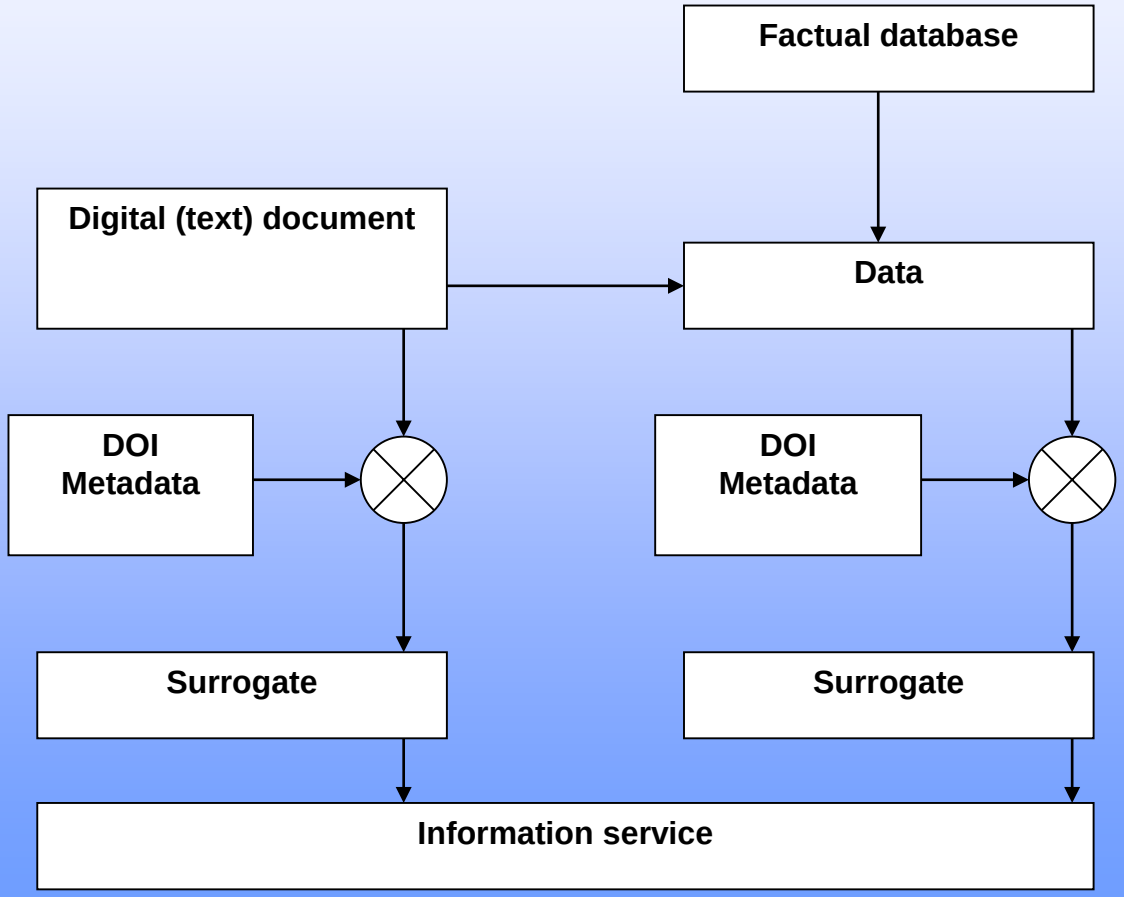
- **Intellektueller Hintergrund von Sender und Empfänger**
  - **Symmetrisches Wissen: gleicher/ähnlicher Hintergrund**
  - **Asymmetrisches Wissen: unterschiedlicher Hintergrund**
  - **„Boundary Objects“ (Boundary Documents)**
    - **sprechen Empfänger mit unterschiedlichem Hintergrund an**
    - **erläutern ihren Gebrauch**

A.4 Dokumente

- **Digitale Dokumente**
  - **können verändert werden („fluidity“)**
    - **entweder: haben aber auch „beständige“ Phasen (z.B. PDF)**
    - **oder: sind auf stetige Veränderung ausgelegt (z.B. Einträge bei Wikipedia)**
  - **RTP-DOC (Réseau Thématique Pluridisciplinaire-Documents et contenu) – Pseudonym: R.T.Pédauque)**
    - **Syntax digitaler Dokumente: „Form“ (Struktur / Daten)**
    - **Semantik digitaler Dokumente: „Zeichen“ (Text / Wissen)**
    - **Pragmatik digitaler Dokumente: „Medium“ (Inskription / Legitimität: formal publiziert, informell publiziert, nicht publiziert)**

A.4 Dokumente

- **Digitale Dokumente und Fakten**

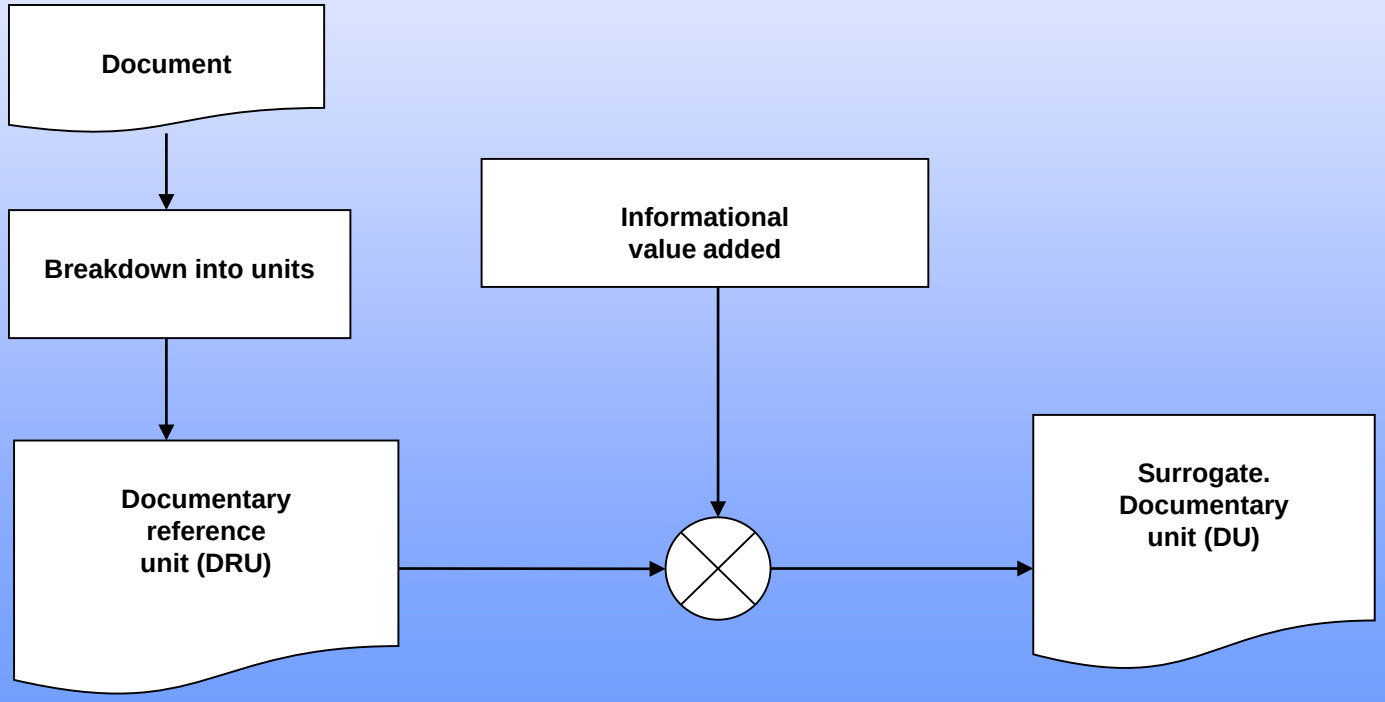


A.4 Dokumente

- **Linked (Open) Data**
  - **Kombination von Daten (soweit mit Metadaten adressiert) aus unterschiedlichen Quellen**
  - **Daten sind formal beschrieben (mittels maschinenlesbarer Sprache)**
  - **Daten sind inhaltlich beschrieben (durch Begriffe eines Knowledge Organization Systems KOS)**
  - **Linked Open Data: wenn die Daten frei zugänglich sind**

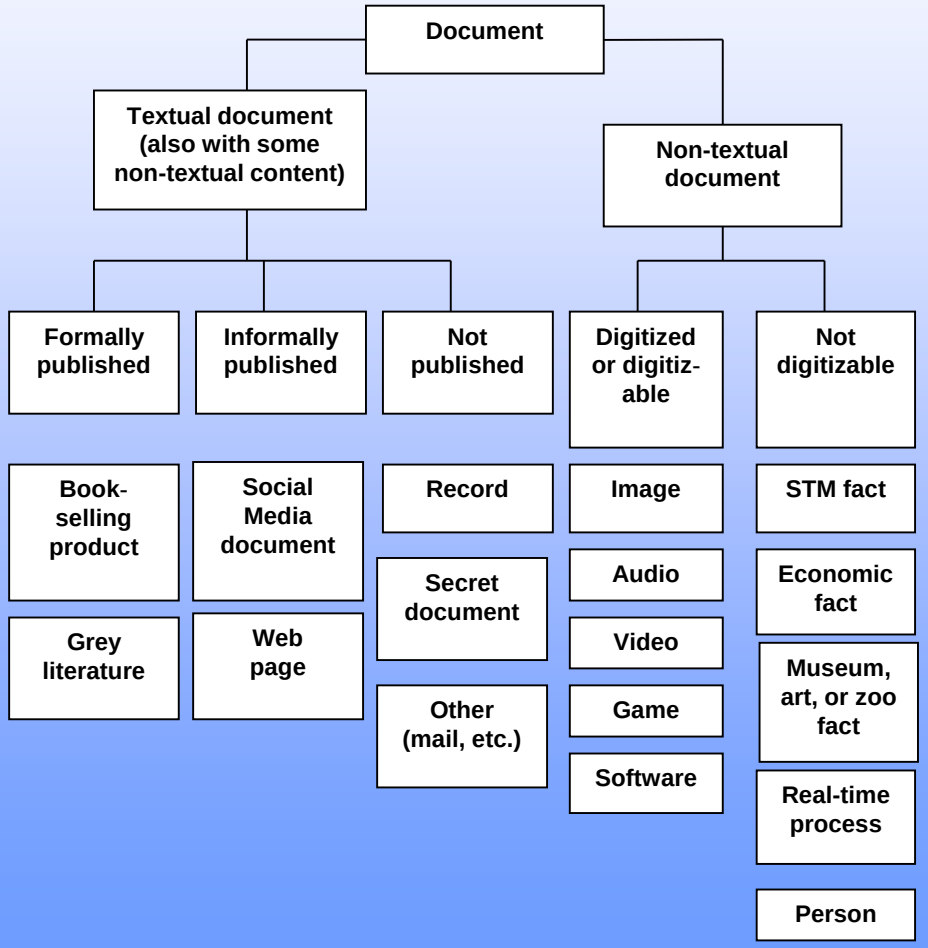
A.4 Dokumente

- **Dokumente und Surrogate**



A.4 Dokumente

- **Dokumenttypen**



A.4 Dokumente

- **formal publizierte Textdokumente (formaler Veröffentlichungsprozess mit Prüfung)**
  - **Buchhandelsdokumente (Bücher, Zeitschriften, Noten, Karten usw.)**
  - **Zeitungen, Dokumente der Nachrichtenagenturen**
  - **Dokumente des gewerblichen Rechtsschutzes (Patente, Gebrauchsmuster, Marken, Geschmacksmuster)**
  - **Rechtsnormen, Entscheidungen**
  - **graue Literatur (als Grenzfall)**

A.4 Dokumente

- **formal publizierte Textdokumente (nach Inhalt)**
  - **Wirtschafts-, Markt- und Pressedokumente**
  - **Rechtsdokumente**
  - **WTM-Dokumente (Wissenschaft, Technik, Medizin)**
  - **Dokumente aus Fiktion und Kunst**

A.4 Dokumente

- **informell publizierte (Text)Dokumente**
  - **Websites / Webseiten**
  - **Beiträge in Newsgroups und Foren**
  - **Posts in Weblogs und Microblogs (Twitter/Weibo)**
  - **Posts in Social Networking Services / SNSs (teilweise nicht universell publiziert)**
  - **Beiträge in Wikis (kooperativ „geprüft“)**
  - **Bilder/Videos/Musik in Sharing Services**

A.4 Dokumente

- **nicht publizierte Textdokumente**
  - **E-Mails**
  - **anderer Schriftverkehr**
  - **Chats, „Telegramme“**
  - **Akten**
  - **Memos, weitere Geschäftsdokumente**
  - **private Dokumente**

A.4 Dokumente

- **Akten („Records“)**
  - **persistente Repräsentationen von Handlungen oder Ereignissen**
  - **erstellt von Beteiligten**
  - **werden nicht verändert**
  - **Eine Akte kann aus mehreren Dokumenten bestehen.**

A.4 Dokumente

- **Nicht-textuelle Dokumente**
  - **digital vorliegend (oder digitalisierbar)**
    - **Bilder**
    - **gesprochene Sprache**
    - **Musik, Geräusche**
    - **Bewegtbilder (Video)**
    - **(digitale) Spiele**
    - **Software**

A.4 Dokumente

- **Nicht-textuelle Dokumente**
  - **nicht digitalisierbar**
    - **WTM (chemische Stoffe und Reaktionen, Werkstoffe, Krankheiten, ...)**
    - **Wirtschaft (Branchen, Märkte, Unternehmen, Produkte)**
    - **Museumsobjekte / Kunstwerke / Tiere im Zoo**
    - **Personen**
    - **Real-time Fakten (Wetter, Verspätungen, „Smart Home“, Flight Tracking, …)**

Kapitel A.5

Informationskompetenz

A.5 Informationskompetenz

- **Informationskompetenz**
  - **Anwendung informationswissenschaftlichen Wissens in Beruf und Alltag**
  - **Kompetenz von Laien**
  - **eng verwandt mit Medienkompetenz**
  - **„Medien- und Informationskompetenz“ (MIK)**
  - **UNESCO: „Media and Information Literacy“ (MIL)**

A.5 Informationskompetenz

United Nations Educational, Scientific and Cultural Organization

Communication and Information

UNESCO » Communication and Information » Themes » Media development » Media literacy » MIL as Composite Concept

Media and Information Literacy

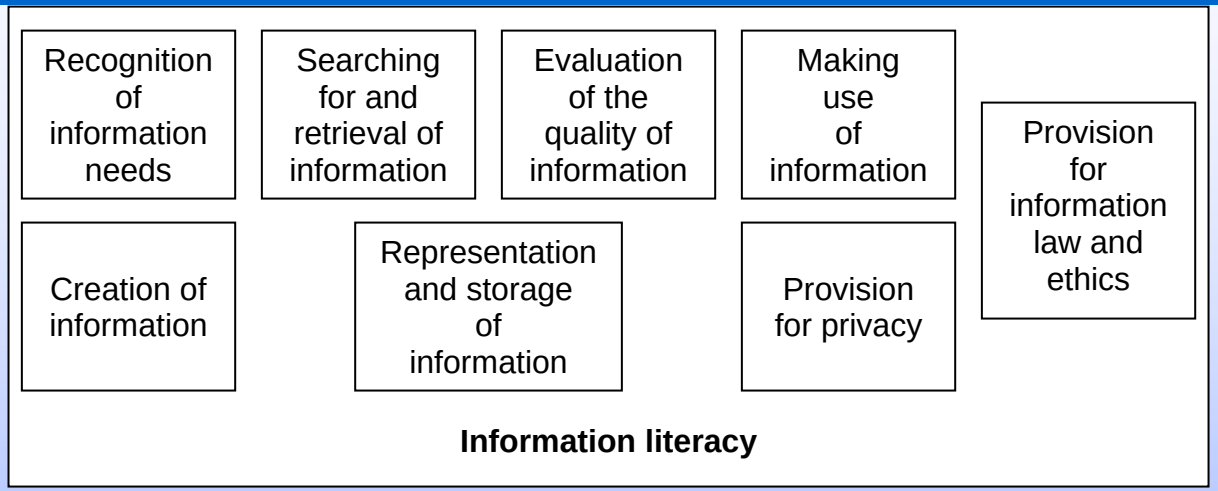
© UNESCO - Screen shot from the video on UNESCO/UNAOC-MILID

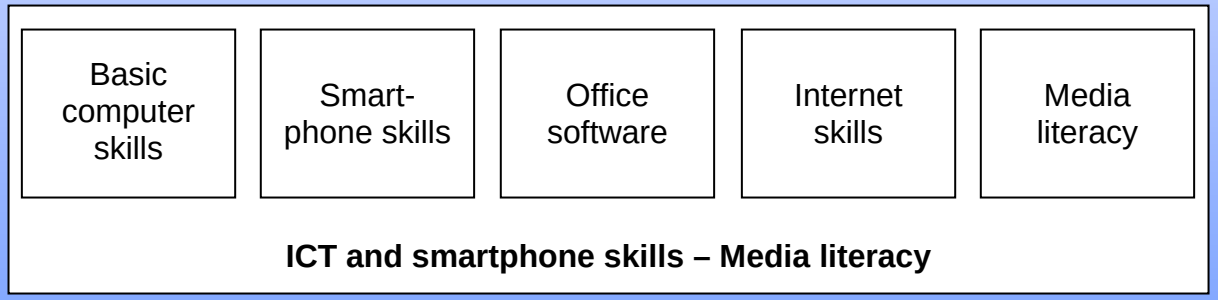
Empowerment of people through Media and Information Literacy (MIL) is an important prerequisite for fostering equitable access to information and knowledge and promoting free, independent and pluralistic media and information systems.

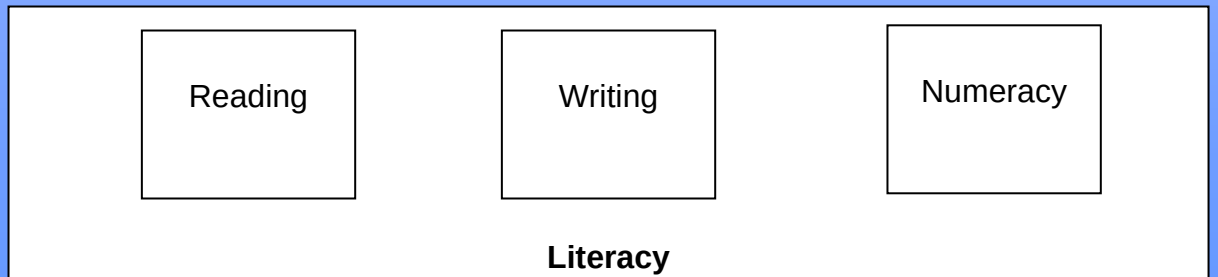
Media and Information Literacy recognizes the primary role of information and media in our everyday lives. It lies at the core of freedom of expression and information - since it empowers citizens to understand the functions of media and other information providers, to critically evaluate their content, and to make informed decisions as users and producer of information and media content.

Information Literacy and Media Literacy are traditionally seen as separate and distinct fields. UNESCO's strategy brings together these two fields as a combined set of competencies (knowledge, skills and attitude) necessary for life and work today. MIL considers all forms of media and other information providers such as libraries, archive, museums and Internet irrespective of technologies used.

A.5 Informationskompetenz







A.5 Informationskompetenz

„Information Literacy“ (1): Retrieval Literacy

- **ALA: „To be information literate, a person must be able to recognize when information is needed and have the ability to locate, evaluate, and use effectively the needed information.”**

und

- **“The information literate student understands many of the economic, legal, and social issues surrounding the use of information and accesses and uses information ethically and legally.”**

Presidential Committee on Information Literacy (1989). Final Report.
Washington, DC: American Library Association / Association for College & Research Libraries.

ACRL (2000). Information Literacy Competency Standards for Higher Education.
Chicago, IL: American Library Association / Association for College & Research Libraries.

A.5 Informationskompetenz

„Information Literacy“ (2): Knowledge Representation Literacy

- **Wichtig seit dem Aufkommen der Social Media**
- **„Prosumer“**

- **Kreation von Informationen**
- **Soweit nötig: Upload der Informationen**
- **Wiederauffindbar beschreiben (Titel, Tags, Hash-Tags)**
- **Ggf. Reaktion: Kommentieren / „Liken“ / Teilen**

A.5 Informationskompetenz

Informationskompetenz am Arbeitsplatz

- using information technology
- finding information from appropriate sources
- executing a process (information processing)
- controlling information (document management, records management, archiving, etc.)
- building up a personal knowledge base
- working with knowledge and personal perspectives adopted in such a way that novel insights are gained
- using information wisely for the benefit of others

Bruce, C.S. (1997). The Seven Faces of Information Literacy. Adelaide: Auslib.

A.5 Informationskompetenz

Informationskompetenz in der Schule

International Standard Classification of Education (ISCED)

- **ISCED 0: Vorschulische Erziehung. Stoff: Medien kennenlernen, grundlegende Smartphone-Skills, Suchen von Informationen**
- **ISCED 1: Primarstufe. Stoff: Grundlagen von Information Retrieval (IR) und Wissensrepräsentation (WR)**
- **ISCED 2: Sekundarstufe I. Stoff: IT-Skills, Medienkompetenz, IR, WR, Privacy**
- **ISCED 3: Sekundarstufe II. Stoff: IR und WR fortgeschritten (Facharbeit!), Websitegestaltung, Informationsrecht, Informationsethik**

A.5 Informationskompetenz

Informationskompetenz in der Schule
Probleme

- **eigenes Fach oder eingebettet in anderen Fachunterricht**
- **Lehrerausbildung, Lehrerweiterbildung („teacher librarian“)**
- **Schulbibliotheken**
- **Lehrpläne sind bereits jetzt voll**

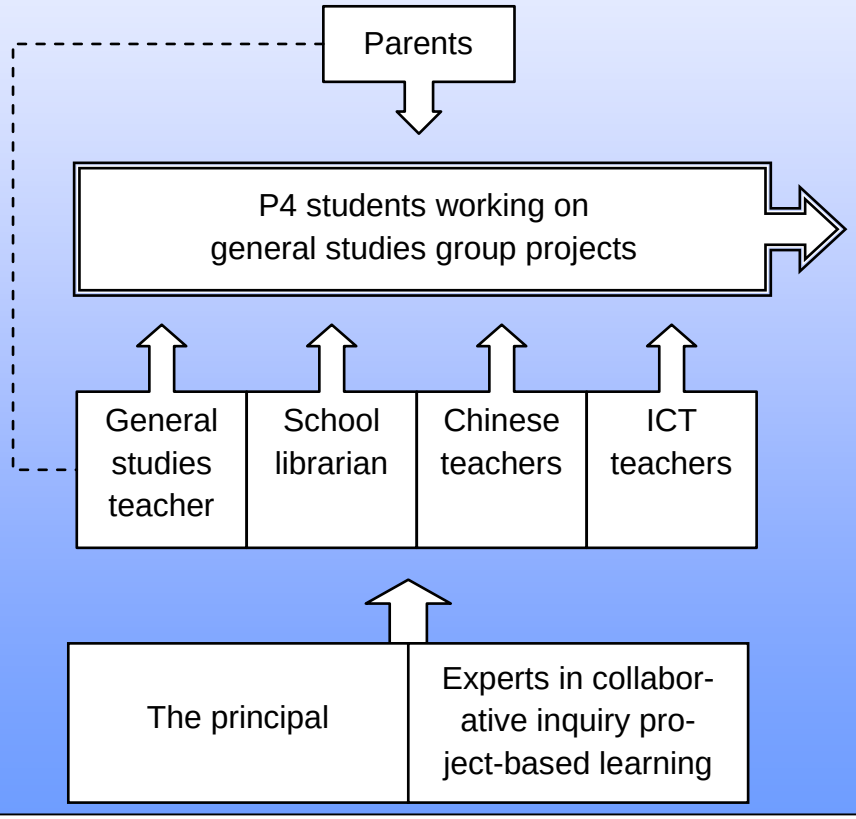
A.5 Informationskompetenz

Fachdidaktik der Informationskompetenz

- **Resource-based learning (document-based learning)**
- **Inquiry-based learning**
- **Teacher-centered learning**
- **Team-based learning**
- **Game-based learning**

A.5 Informationskompetenz

Informationskompetenzunterricht in der Primarschule (Beispiel: Hongkong)



Dimensions in students' improvement

1. Reading comprehension
2. Writing abilities
3. Information literacy
4. IT skills
5. Subject knowledge
6. Social and communication skills
7. Presentation skills
8. Research skills

Chu, S.K.W. (2009). Inquiry project-based learning with a partnership of three types of teachers and the school librarian. Journal of the American Society for Information Science and Technology, 60(8), 1671-1686.

A.5 Informationskompetenz

Wie kann man Informationskompetenz messen?

- **„Rubrics“**
- **Fragebögen (Selbsteinschätzung)**
- **Tests**

Voraussetzung: Existenz von Standards

A.5 Informationskompetenz

Rubrics (Beispiel von van Helvoort)

	Criterion	*Professional behaviour*	*Insufficient behaviour*	
2	Reference list	☐ The student product has a reference list that is complete and the citation style is used correctly. With the reference list it is easy to identify the documents that the student used. Remark: the last point is more important than a correct bibliographic description in accordance with a standard citation style. However, for the score 'very good' the citation style must also be used correctly.	☐ There is no reference list in the student product and / or ☐ The reference list is not complete (documents that are cited in the text are not listed in the reference list) or ☐ Important bibliographic data (title, author, year of publication) are missing. An example that often recurs in educational practice: for internet resources only the URL is mentioned.	Grade 1-10=

Score:	0 very good	0 good	0 sufficient	0 poor	0 bad	0 very bad

	Criterion	*Professional behaviour*	*Insufficient behaviour*	
3	Quality of the primary sources (books, journal articles, websites etc.)	☐ The reference list of the student product makes clear that the student has used relevant, reliable (preferably authentic) and up-to- date information sources that discuss the topic or the question from different points of view.	☐ The information sources the student has used are insignificant, outdated or not relevant enough. An example of 'insignificance' is that the student only used Internet-sites as an information source. And / or … ☐ The information sources the student used are one-sided (too much from one point of view). The student has, for instance, only used government information(.gov-sites) or publications from one particular author.	Grade 1-20=

A.5 Informationskompetenz

Tests (BILT / Beutelspacher Information Literacy Test)

10) Which query will retrieve more documents? [ONE ANSWER]

- Dog AND cat
- Dog OR cat
- Both queries above will yield the same amount of results
- I don't know

11) Using a search engine, you wish to research the following recipe: Cookies, either with nuts or with almonds, but definitely without cinnamon. Which of the queries below (including operators) would you use to retrieve the recipe? [ONE ANSWER]

- Cookies AND (nuts OR almonds) NOT cinnamon
- (Nuts OR almonds) (AND cookies NOT cinnamon)
- NOT cinnamon AND cookies (nuts OR almonds)
- Cookies AND almonds AND nuts NOT cinnamon
- I don't know

A.5 Informationskompetenz

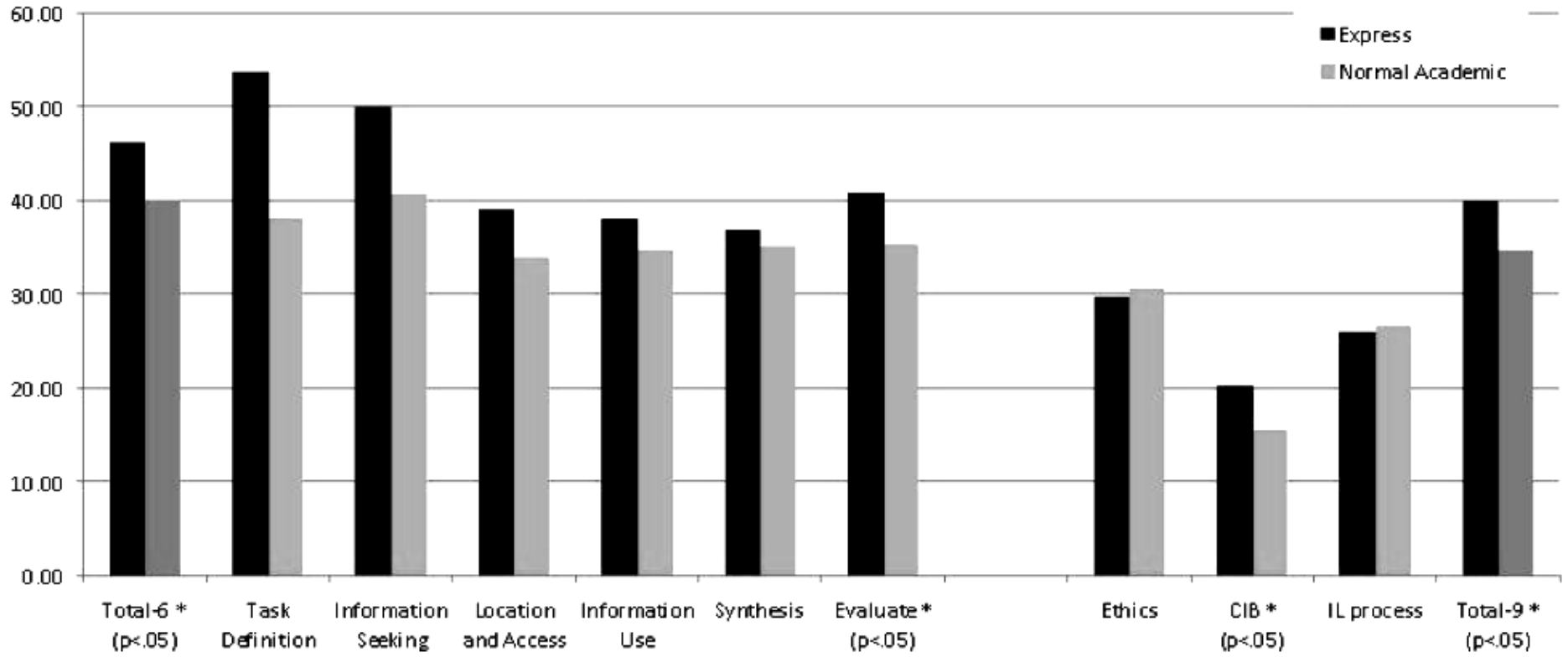
Tests (BILT / Beutelspacher Information Literacy Test)

27) Which tags (keywords) would you use for the following image of the Brooklyn Bridge if you wanted to upload it to a photosharing service for other users to find? [ANY NUMBER OF ANSWERS]

- Bridge
- Brooklyn
- Water
- Brooklyn Bridge
- My_city
- East River
- House
- New York
- Photo
- Suspension bridge
- Day
- World

A.5 Informationskompetenz

Study by Chang et al. (2012): Information Literacy of Singaporean secondary school students



Teil B: Propädeutik des Information Retrieval

Kapitel B.1

Geschichte des Information Retrieval

B.1 Geschichte des Information Retrieval

Memex

- **Vision von Vannevar Bush (1945): maschinelle Bereitstellung des Wissens**
- **nicht mittels eindimensionaler Klassifikationssysteme (wie seinerzeit in Bibliotheken üblich), sondern über assoziative Verbindungen („trails“)**

Bush, V. (1945). As we may think. The Atlantic Monthly, 176(1), 101-108.

B.1 Geschichte des Information Retrieval

1950

Erstes Auftreten des Wortes „Information Retrieval“ bei Calvin N. Mooers

Mooers, C.N. (1952). Information retrieval viewed as temporal signaling. In Proceedings of the International Congress of Mathematicians 1950 (Vol. 1, pp. 572-573.) Providence, RI: American Mathematical Society.

B.1 Geschichte des Information Retrieval

Frühe Forschungen

Hans-Peter Luhn:

„machine talents“ entdecken!

Textstatistik

Automatisches Abstracting

SDI

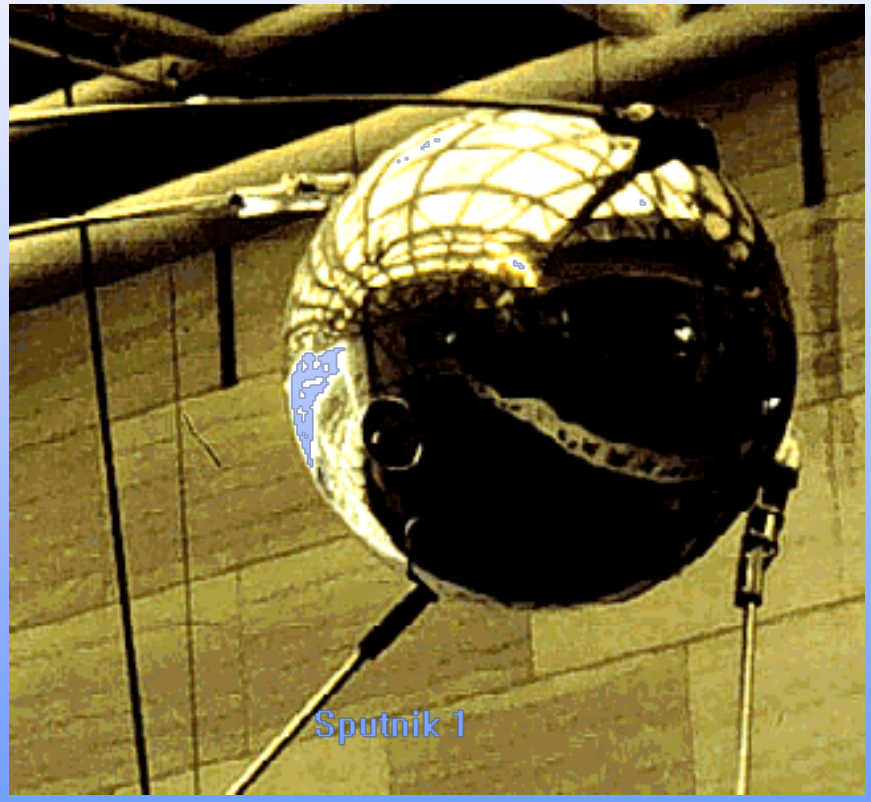
KWIC

Luhn, H.P. (1957). A statistical approach to mechanized encoding and searching of literary information. IBM Journal, 1(4), 309-317.
Luhn, H.P. (1958). The automatic creation of literature abstracts. IBM Journal, 2(2), 159-165.
Luhn, H.P. (1961). The automatic derivation of information retrieval encodements from machine-readable texts. In A. Kent (Ed.), Information Retrieval and Machine Translation, Vol. 3, Part 2 (pp. 1021-1028). New York, NY: Interscience.

B.1 Geschichte des Information Retrieval

Der Sputnik-Schock

- **4.10.1957: Start von Sputnik1**
- **Schock 1: der Westen kann nichts Vergleichbares; Folge: Apollo-Programm**
- **Schock 2: die Signale können nicht entschlüsselt werden, obwohl die entsprechende Publikation (in englisch!) vorliegt; Folge: staatliches Interesse am Informations- und Dokumentationswesen**



B.1 Geschichte des Information Retrieval

Weinberg-Bericht

„Informationsexplosion“ erkannt

Aus Informationsüberfluss kann Informationsmangel entstehen

Gegenmittel: Informations-wissenschaft (besonders: Information Retrieval)

Der Weinberg-Bericht hatte große Wirkung (Vorwort von John F. Kennedy)

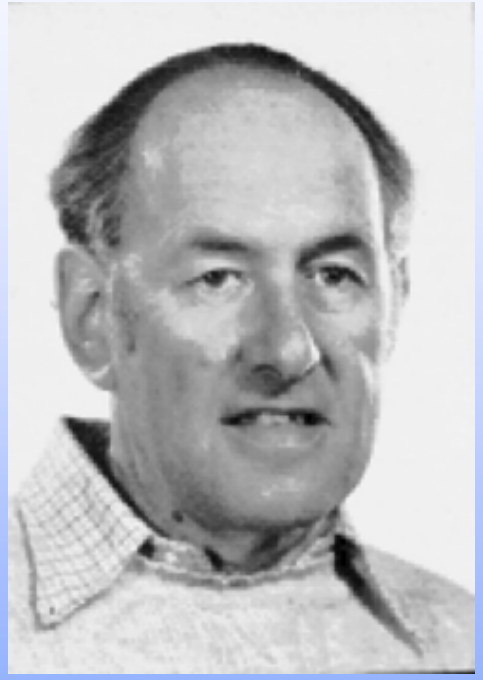
Alvin M. Weinberg

Weinberg, A.M. (1964). Wissenschaft, Regierung und Information. Genehmigte deutsche Übersetzung des Weinberg-Berichtes vom 10. Januar 1963. Frankfurt/M.: Deutsche Gesellschaft für Dokumentation. (Beiheft zu den Nachrichten für Dokumentation; 12). (Original: 1963).

B.1 Geschichte des Information Retrieval

Vektorraummodell

- **Experimente mit natürlichsprachigen Systemen: Gerard Salton**
- **Dokumente und Suchanfragen sind Vektoren in einem n-dimensionalen Raum**
- **SMART (System for the Mechanical Analysis and Retrieval of Text)**

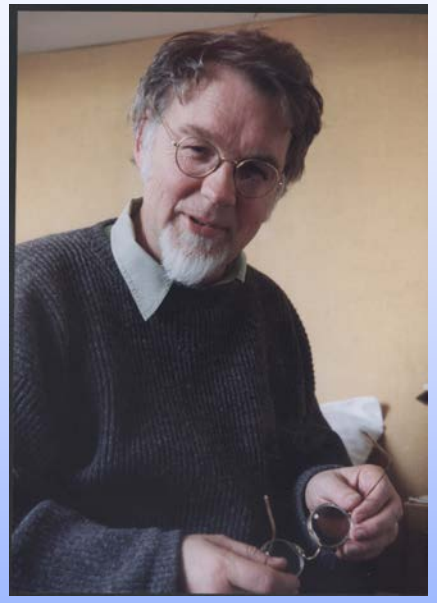


Salton, G. (1968). Automatic Information Organization and Retrieval. New York, NY: McGraw-Hill.
Salton, G., Ed. (1971). The SMART Retrieval System – Experiments in Automatic Document Processing. Englewood Cliffs, NJ: Prentice Hall.
Salton, G., & McGill, M.J. (1983). Information Retrieval – Grundlegendes für Informationswissenschaftler. Hamburg: McGraw-Hill.

B.1 Geschichte des Information Retrieval

Probabilistisches Modell

- **Ein Dokument ist mehr oder weniger relevant in bezug auf eine Suchanfrage**
- **bedingte Wahrscheinlichkeit (Relevanz unter der Bedingung der Query)**
- **Relevance Ranking**
- **1960: Maron & Kuhns**
- **ausgearbeitet vor allem von Cornelis Joost van Rijsbergen**



C.J van Rijsbergen

Maron, M.E., & Kuhns, J.L. (1960). On relevance, probabilistic indexing and information retrieval. Journal of the ACM, 7, 216-244.
van Rijsbergen, C.J. (1979). Information Retrieval. London: Butterworths, 2. Ed. (Chapter 6: Probabilistic retrieval).

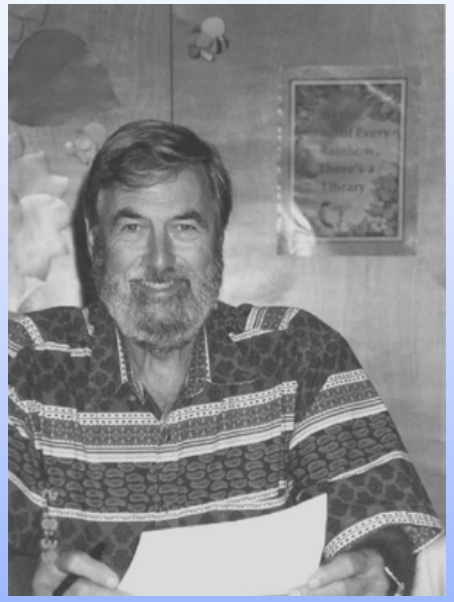
B.1 Geschichte des Information Retrieval

Zitationsdatenbanken

- **Eugene Garfield gründet 1960 das „Institute for Scientific Information“**
- **Vermarktung von Fußnoten akademischer Zeitschriften („Science Citation Index“)**
- **Vertrieb von Inhaltsverzeichnissen der Zeitschriften als „Current Contents“**
- **Erarbeitung von Kennwerten für den Einfluss der Zeitschriften („impact factor“)**

Garfield, E. (1955). Citation Indexes for Science. Science, 122(3159), 108-111.
Garfield, E. (1979). Citation Indexing. New York, NY: Wiley.
Cawkell, T., & Garfield, E. (2001). Institute for Scientific Information. Information Services & Use, 21, 79-86.

B.1 Geschichte des Information Retrieval



Summit

Online-Hosts

Vorarbeiten kommerzieller Online-Systeme:

- **Roger Kent Summit: DIALOG (FuE ab ca. 1960, online seit 1972)**
- **Carlos A. Cuadra: SDC, später ORBIT (Start: 1962; online: 1972)**
- **Richard H. Giering: Data Central, später: Lexis (Start: Ende der 60er Jahre; online: 1973)**

Bourne, C.P., & Hahn, T.B. (2003). A History of Online Information Services, 1963-1976. Cambridge, MA: MIT Press.

B.1 Geschichte des Information Retrieval

Ausarbeitung des probabilistischen Modells (1970er Jahre)

- **Robertson – Sparck-Jones - Formel**

Karen Sparck-Jones

$$w_i = \log \frac{r_i/(R - r_i)}{(n_i - r_i)/(N - n_i - R + r_i)}$$

Stephen E. Robertson

Robertson, S.E., & Sparck-Jones, K. (1976). Relevance weighting of search terms. Journal of the American Society für Information Science, 27, 129-146.

B.1 Geschichte des Information Retrieval

Erfolg der kommerziellen elektronischen Informationsdienste (ab 1980er Jahre)

Internationale Online-Hosts

Westlaw

Dialog DataStar

factiva

LexisNexis

Questel • Orbit
Intellectual Property Group

STN

Deutsche Online-Hosts

fiz technik
Fachinformation Technik

GBI
the contentmachine

GENIOS
WIRTSCHAFTSDATENBANKEN

DIMDI

juris

B.1 Geschichte des Information Retrieval

Boom durch Suchwerkzeuge im Word Wide Web (ab 1990er Jahre)

- **Suchmaschinen der ersten Generation: textorientiert (z.B. AltaVista)**
- **Suchmaschinen der zweiten Generation: zusätzlich an der Webstruktur orientiert (z.B. Google)**

Louis Monier

Sergej Brin (li.)

Larry Page

B.1 Geschichte des Information Retrieval

Ausarbeitung linktopologischer Modelle

- **PageRank (von Brin und Page)**
- **Kleinberg-Algorithmus (von Jon M. Kleinberg)**

Jon M. Kleinberg

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30, 107-117.
Page, L. (1998). Method for node ranking in a linked database. Patent Nr. US 6,285,999. Priorität: 9.1.1998. Patentinhaber: The Board of Trustees of the Leland Stanford Junior University.

Kleinberg, J.M. (1999). Authoritative sources in a hyperlinked environment. Journal of the ACM, 46, 604-632.

B.1 Geschichte des Information Retrieval

Social Media / Web 2.0

- **Kollaborative Dienste**
- **„Web 2.0“ Begriff geprägt von Tim O‘Reilly**
- **Wissensrepräsentation durch „tagging“ und „folksonomies“**

Instagram facebook flickr YouTube Broadcast Yourself tumblr. YOU NOW

common tags cloud | list

2.0 ajax article articles blog business definition design development internet o'reilly oreilly programming reference social socialsoftware software tech technology toread web **web2.0** web20 webdesign webdev

O‘Reilly, T. (2005). What is the Web 2.0?
www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

Kapitel B.2

Grundbegriffe des Information Retrieval

B.2 Grundbegriffe des Information Retrieval

Frage- und Antworttypen

- **Konkreter Informationsbedarf (Faktenfrage)**
  - **Welchen Umsatz hatte Unternehmen X im Dezember 1998 in der Region Z?**
  - **Wo liegt der Schmelzpunkt von Kupfer?**
  - **Wie schloß der Dollarkurs letzten Freitag an der Frankfurter Börse?**
  - **Wann hat mein Geschäftspartner X Geburtstag?**
- **Problemorientierter Informationsbedarf (Literatur)**
  - **Welche Methoden der fuzzy logic lassen sich beim Data Mining einsetzen?**
  - **Wie hängen Marketing und Qualitätsmanagement zusammen?**
  - **Wie bewerten Analysten das Unternehmen X?**
  - **Wie beschreiben Marktforscher das Konsumklima für ausländischen Wein in Ungarn?**

I. Frants, V.I., Shapiro, J., & Voiskunskii, V.G. (1997). Automated Information Retrieval. San Diego; CA: Academic Press, 1997.

B.2 Grundbegriffe des Information Retrieval

Konkreter Informationsbedarf

- 1. Thematische Grenzen sind klar angesteckt.
- 2. Die Suchfrageformulierung ist durch exakte Terme ausdrückbar.
- 3. *Eine* Faktenfrage reicht aus, um den Bedarf zu decken.
- 4. Mit der Übermittlung der Fakteninformation ist das Informationsproblem erledigt.

Problemorientierter Informationsbedarf

- 1. Thematische Grenzen sind nicht exakt.
- 2. Die Suchfrageformulierung lässt terminologische Varianten zu.
- 3. Es müssen diverse Dokumente aus unterschiedlichen Quellen beschafft werden.
- 4. Mit der Übermittlung der Literaturinformation wird ggf. das Informationsproblem modifiziert oder ein neuer Bedarf entdeckt.

B.2 Grundbegriffe des Information Retrieval

Informationsbedarf / Informationsbedürfnis

- **Bedarf: objektiv betrachtet (unter Abstraktion eines konkreten Subjektes)**
- **Bedürfnis: subjektiv betrachtet (die konkret empfundene Mangelsituation eines Menschen)**

B.2 Grundbegriffe des Information Retrieval

Umformulierung des Informationsbedarfs in eine konkrete Anfrage an ein Retrievalsystem

- **Problem: etwas klar und deutlich ausdrücken, was man *nicht* weiß**
- **Grundkenntnisse im thematischen Gebiet des Informations-bedarfs müssen gegeben sein**
- **Formulierung des Informationsbedarfs 1. umgangssprachlich und 2. in der Syntax des Retrievalsystems**

B.2 Grundbegriffe des Information Retrieval

Anfrageformulierung

- **Informationsbedarf: „Ich suche Informationen über Julia Roberts in Notting Hill.“**
  - **LexisNexis:**
  - **HEADLINE(Julia Roberts w/5 Notting Hill)**
  - **DIALOG:**
  - **(Julia ADJ Roberts AND Notting ADJ Hill)/ti**
  - **Google:**
  - **"Notting Hill" "Julia Roberts"**

B.2 Grundbegriffe des Information Retrieval

Wortorientierte Anfragebearbeitung *versus* begriffsorientierte Anfragebearbeitung

- **wortorientiert: Suche nach Zeichenfolgen (z.B. Java, der Informationsbedarf richte sich auf die Insel).**

- **begriffsorientiert: semantische Suche**
  - **O Java (indonesische Insel)**
  - **O Java (Kaffee)**
  - **O Java (Programmiersprache)**

  bitte ankreuzen!

B.2 Grundbegriffe des Information Retrieval

Dokumentarische Bezugseinheit (DBE)

stets gleichbleibende Einheit der Vorlagen, die in einen Informationsspeicher aufgenommen werden, hierbei ggf. analytische „Zerlegung“ der Vorlagen

Beispiele:

Buch (als Ganzes)	**-**	**Buchkapitel**
dto.	**-**	**Abbildung; Tabelle**
Zeitschrift (als Ganzes)	**-**	**Artikel**
Korrespondenz	**-**	**einzelner Brief**
Film	**-**	**Filmsequenz**

B.2 Grundbegriffe des Information Retrieval

Dokumentarische Bezugseinheit (DBE): Beispiel

Pathology in the Era of Web 2.0

William E. Schreiber, MD,[1] and Dean M. Giustini, MLS, MEd[2]

Key Words: Web site; Web 2.0; Internet; Social software; Continuing education; Communities of practice

DOI: 10.1309/AJCPEC9FZSB4DEDH

Abstract

In the past few years, the term Web 2.0 *has become a descriptor for the increased functionality of Web sites, including those with medical content. Most physicians do not know what Web 2.0 means or how it can impact their work lives. This review provides some background on the evolution of Web 2.0 and describes how its features are being incorporated into medical Web sites. Some potential applications of Web 2.0 in pathology and laboratory medicine are discussed, as are the issues that must be considered when adopting this new technology.*

None of us is as smart as all of us.
—Japanese proverb

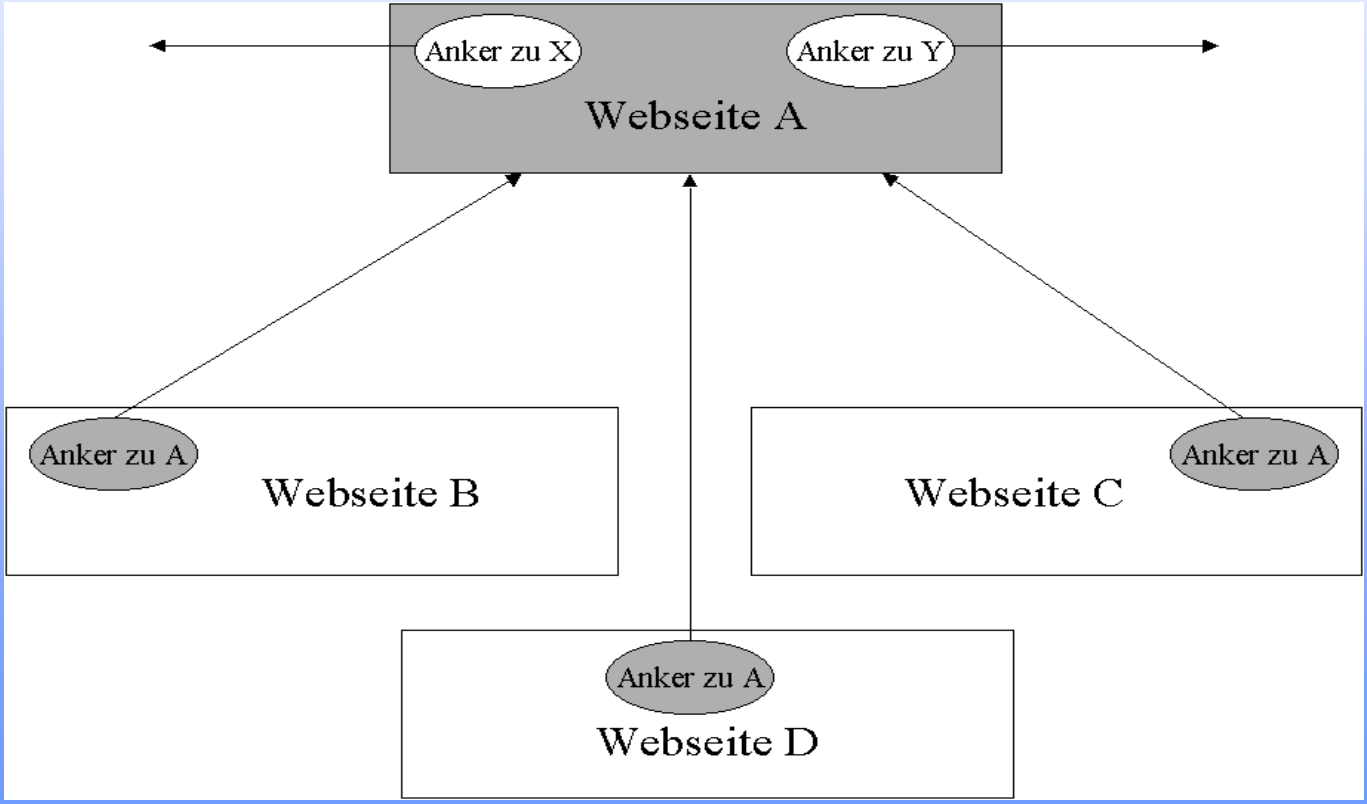
You may have heard about the Web sites Facebook and Twitter. Perhaps you have read entries in Wikipedia or watched videos on YouTube. Each of these Web sites has a different purpose, but all of them enhance communication and information sharing among a network of users through the exchange of text, images, sound, and video. These features, and the environment of collaboration that they create, are referred to as *Web 2.0*.

The term Web 2.0 was coined at the beginning of the

B.2 Grundbegriffe des Information Retrieval

Dokumentarische Bezugseinheit (DBE)

Beispiel: Google



B.2 Grundbegriffe des Information Retrieval

Dokumentationswürdigkeit

- **Kriterienkatalog, der die Entscheidung fundiert, ob eine bestimmte DBE in den Informationsspeicher aufgenommen wird oder nicht**

  Aspekte:
  - **Informationsbedarf der Nutzer**
  - **thematische Kriterien**
  - **formale Kriterien (Bsp.: nur wissenschaftliche Artikel; nur HTML-Dateien)**
  - **Finanzrahmen - Personalressourcen - Zeit**
  - **ggf.: Neuigkeit**
  - **ggf.: kritische Prüfung des Inhalts**
  - **ggf.: juristische Aspekte / Zensur**

B.2 Grundbegriffe des Information Retrieval

Dokumentationseinheit

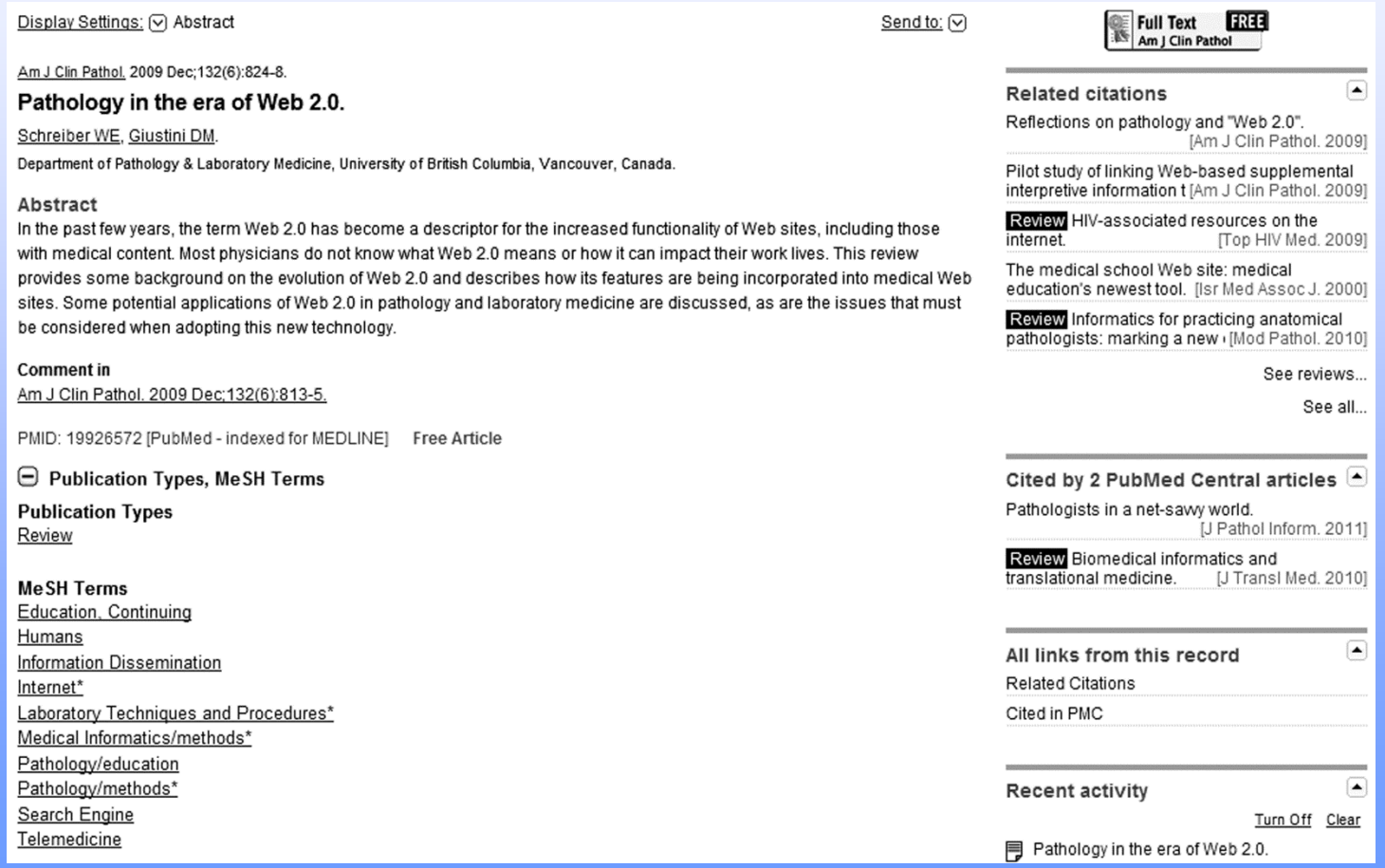
Repräsentant (Surrogat) der DBE in einem Informations-speicher

Bestandteile:

- **formale Beschreibung**
- **inhaltliche Beschreibung**
- **(nicht immer) dokumentarische Bezugseinheit in Vollform**

B.2 Grundbegriffe des Information Retrieval

Dokumentationseinheit (Beispiel). *Quelle:* Medline



Display Settings: Abstract

Send to:

Am J Clin Pathol. 2009 Dec;132(6):824-8.

Pathology in the era of Web 2.0.

Schreiber WE, Giustini DM.

Department of Pathology & Laboratory Medicine, University of British Columbia, Vancouver, Canada.

Abstract

In the past few years, the term Web 2.0 has become a descriptor for the increased functionality of Web sites, including those with medical content. Most physicians do not know what Web 2.0 means or how it can impact their work lives. This review provides some background on the evolution of Web 2.0 and describes how its features are being incorporated into medical Web sites. Some potential applications of Web 2.0 in pathology and laboratory medicine are discussed, as are the issues that must be considered when adopting this new technology.

Comment in

Am J Clin Pathol. 2009 Dec;132(6):813-5.

PMID: 19926572 [PubMed - indexed for MEDLINE] Free Article

Publication Types, MeSH Terms

Publication Types

Review

MeSH Terms

Education, Continuing
Humans
Information Dissemination
Internet*
Laboratory Techniques and Procedures*
Medical Informatics/methods*
Pathology/education
Pathology/methods*
Search Engine
Telemedicine

Full Text FREE
Am J Clin Pathol

Related citations

Reflections on pathology and "Web 2.0". [Am J Clin Pathol. 2009]

Pilot study of linking Web-based supplemental interpretive information t [Am J Clin Pathol. 2009]

Review HIV-associated resources on the internet. [Top HIV Med. 2009]

The medical school Web site: medical education's newest tool. [Isr Med Assoc J. 2000]

Review Informatics for practicing anatomical pathologists: marking a new [Mod Pathol. 2010]

See reviews...

See all...

Cited by 2 PubMed Central articles

Pathologists in a net-savvy world. [J Pathol Inform. 2011]

Review Biomedical informatics and translational medicine. [J Transl Med. 2010]

All links from this record

Related Citations

Cited in PMC

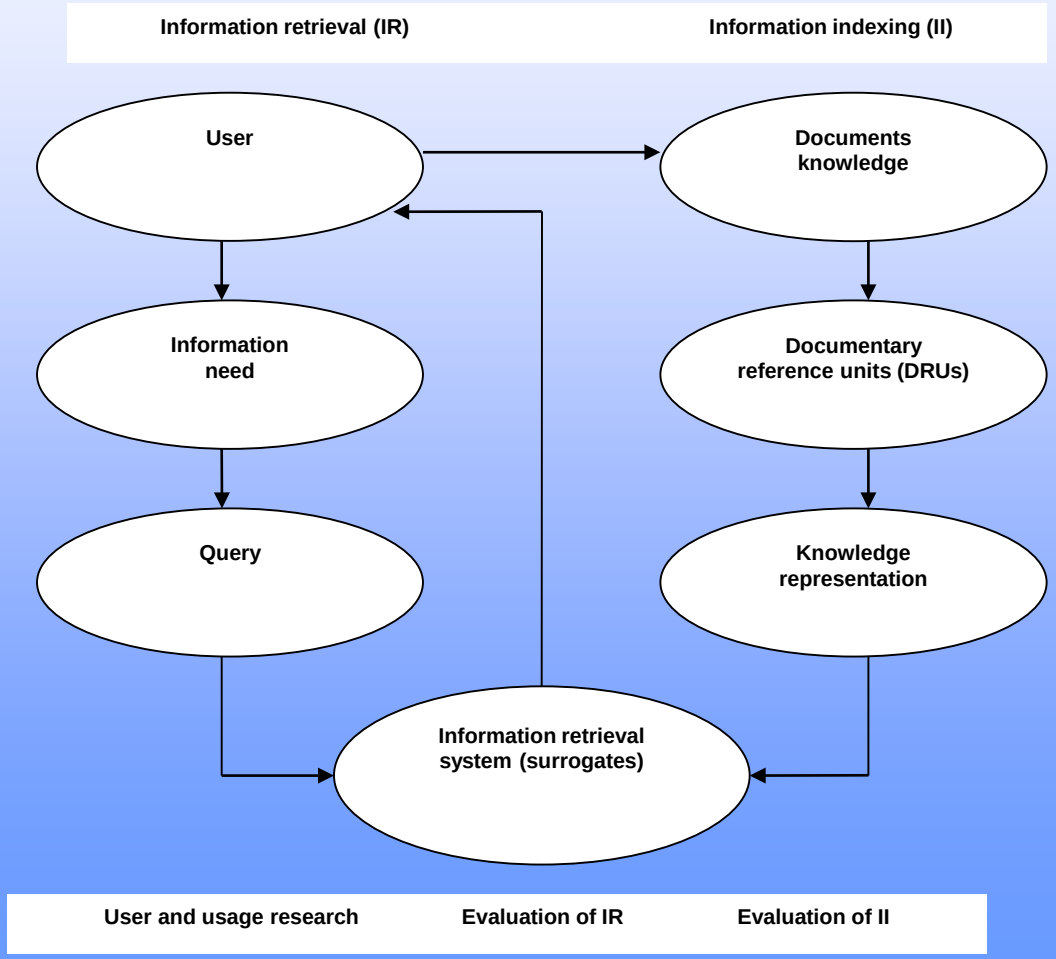
Recent activity

Turn Off Clear

Pathology in the era of Web 2.0.

B.2 Grundbegriffe des Information Retrieval

Information Indexing und Information Retrieval

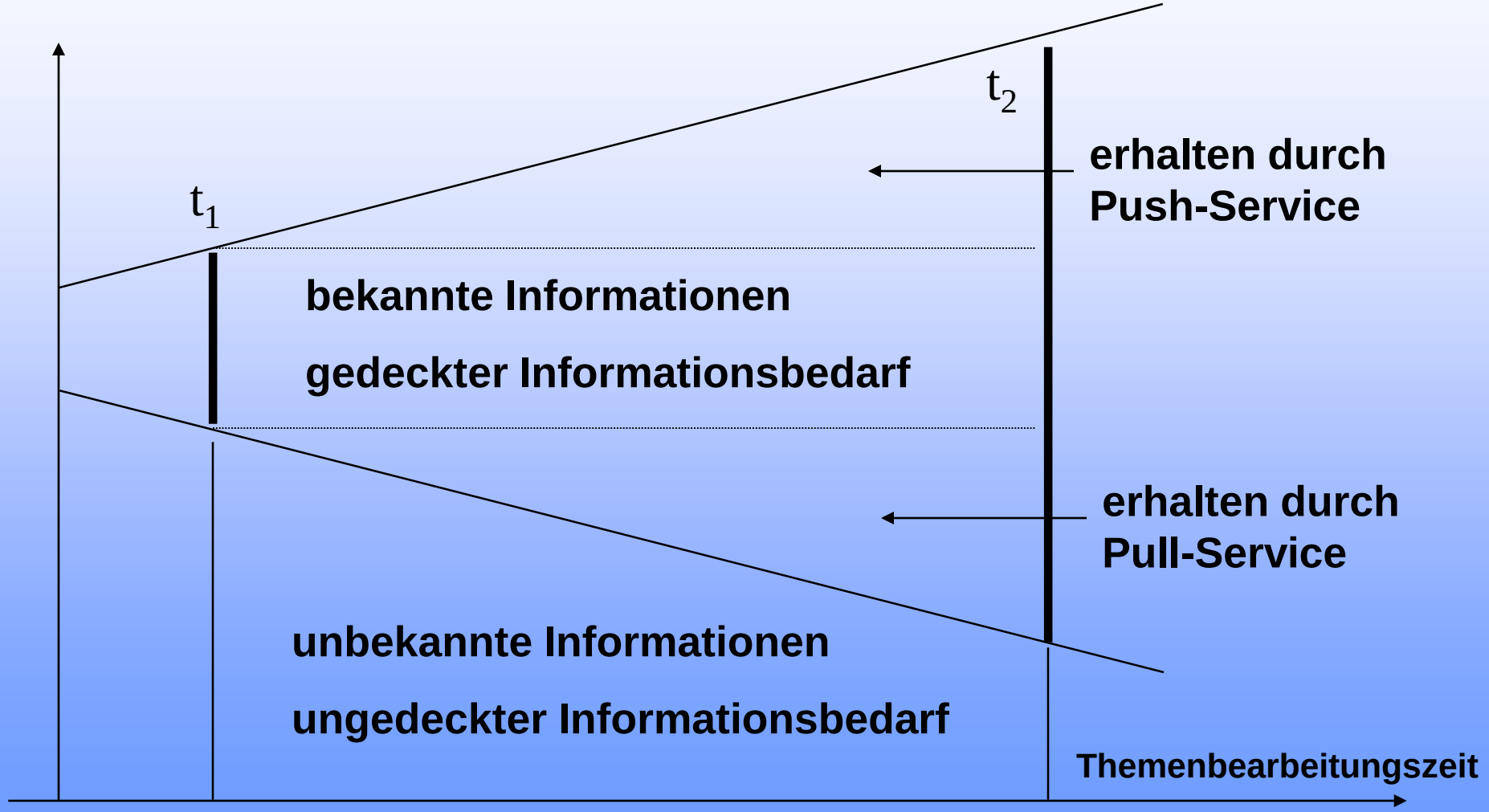


B.2 Grundbegriffe des Information Retrieval

Pull und Push

- **Pull-Service: zur Befriedigung von ad-hoc auftretendem Informationsbedarf sucht ein Nutzer aktiv in Informationssystemen nach Wissen**
- **Push-Service: zur Befriedigung eines (über einen gewissen Zeitraum) andauernden Informationsbedarfs wird ein Nutzer vom Informationssystem mit jeweils aktuellem, neuem Wissen versorgt. Push-Services sorgen für current awareness**
  - **Arbeitsschritt 1: Festlegen eines Informationsprofils (führt Nutzer oder Information Professional durch)**
  - **Arbeitsschritt 2: Periodische Lieferung von Wissen (führt Informationssystem automatisch durch) – „SDI“ (selective dissemination of information) oder „Alert“**

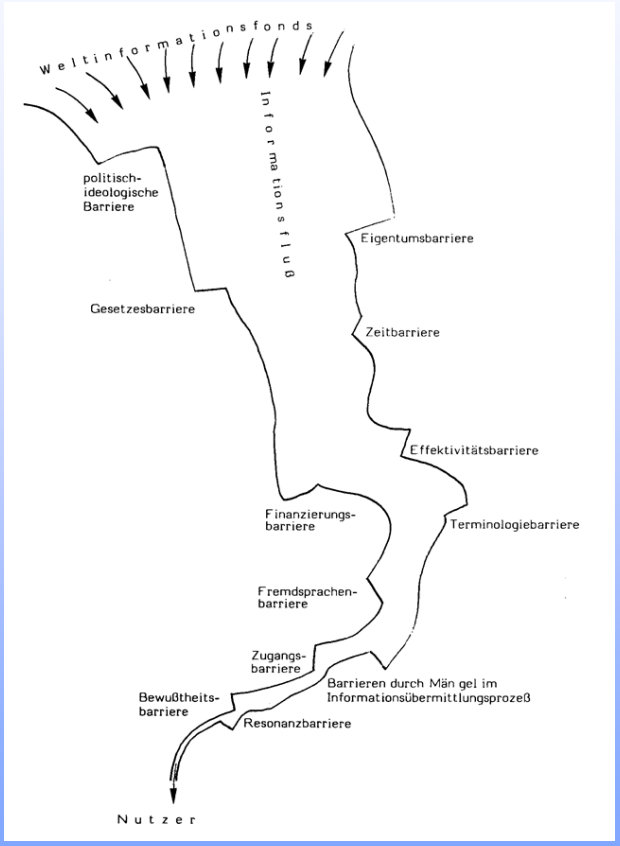
B.2 Grundbegriffe des Information Retrieval



B.2 Grundbegriffe des Information Retrieval

Informationsbarrieren

- politisch-ideologische Barriere
- Eigentumsbarriere
- Gesetzesbarriere
- Zeitbarriere
- Effektivitätsbarriere
- Finanzierungsbarriere
- Terminologiebarriere
- Fremdsprachenbarriere
- Zugangsbarriere
- Barrieren durch Mängel beim Information Retrieval
- Bewusstheitsbarriere
- Resonanzbarriere



Engelbert, H. (1976). Der Informationsbedarf in der Wissenschaft. Leipzig: Bibliographisches Institut.

B.2 Grundbegriffe des Information Retrieval

Recall und Precision

- **Haben wir alle Datensätze gefunden, die handlungsrelevantes Wissen beinhalten? (Vollständigkeit; Recall)**

  Recall = a / (a + c)

- **Haben wir nur solche Datensätze gefunden? (Genauigkeit, Precision)**

  Precision = a / (a + b)

a =: gefundene relevante Treffer

b =: nichtrelevante Datensätze, die in der Treffermenge enthalten sind (Ballast)

c =: relevante Datensätze in der Datenbank, die nicht gefunden wurden

B.2 Grundbegriffe des Information Retrieval

- **Recall und Precision beim *konkreten* Informationsbedarf**

  Recall = 1

  Precision = 1

- **Recall und Precision beim *problemorientierten* Informationsbedarf**

  inverse Relation zwischen Recall und Precision

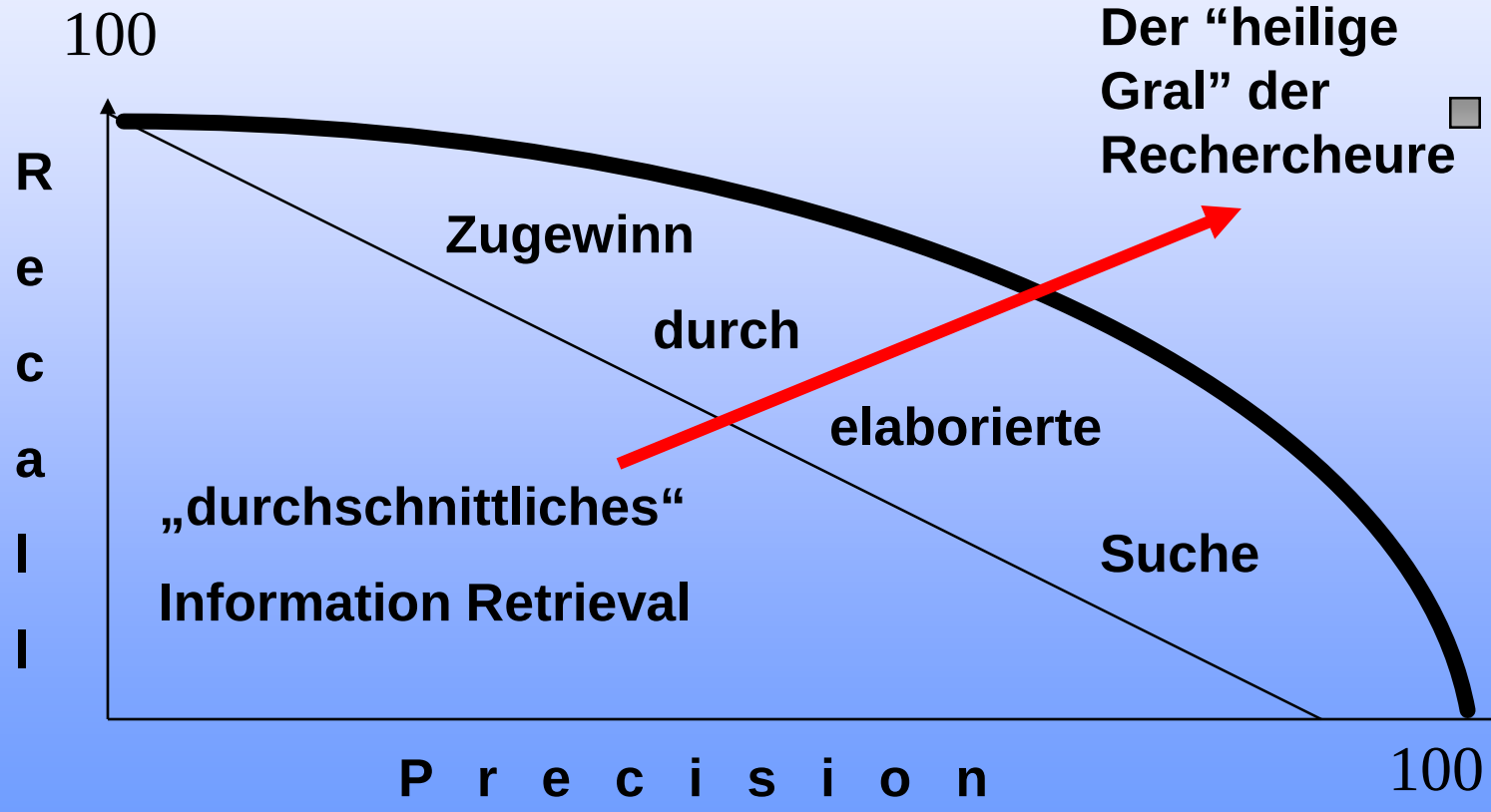
  – **bei Erhöhung des Recall: Absinken der Precision**

  – **bei Erhöhung der Precision: Absinken des Recall**

  empirischer Schätzwert: Recall + Precision = 1

- **Achtung Problem: dies ist ein theoretisches Modell; genaue Messergebnisse sind unmöglich, da der Wert c in großen Datenbanken prinzipiell unbekannt ist.**

B.2 Grundbegriffe des Information Retrieval



B.2 Grundbegriffe des Information Retrieval

Ähnlichkeit

- **Jaccard-Sneath**

  $SIM(D1\text{-}D2) = g / (a + b - g)$

- **Dice**

  $SIM(D1\text{-}D2) = 2g / (a + b)$

- **Cosinus**

  $SIM(D1\text{-}D2) = g / \sqrt{a * b}$

a: Anzahl der Wörter in D1; b: Anzahl der Wörter in D2

g: Anzahl der gemeinsamen Wörter in D1 und D2

Kapitel B.3

Relevanz und Pertinenz

B.3 Relevanz und Pertinenz

Relevanz – Pertinenz – Nützlichkeit

Relevanz: Wann ist ein Dokument für ein Suchargument relevant?

- **(1) wenn es objektiv zur Vorbereitung einer Entscheidung dient**
- **(2) wenn es objektiv eine Wissenslücke schließt**
- **(3) wenn es objektiv eine Frühwarnfunktion erfüllt**

Pertinenz: Wann ist ein Dokument für einen Nutzer pertinent?

- **(1) wenn es subjektiv zur Vorbereitung der Entscheidung eines Nutzers dient**
- **(2) wenn es subjektiv eine Wissenslücke des Nutzers schließt**
- **(3) wenn es subjektiv eine Frühwarnfunktion für den Nutzer erfüllt**

B.3 Relevanz und Pertinenz

Relevanz – Pertinenz – Nützlichkeit

Ziele des Information Retrieval:

- **Gewinnung relevanter / pertinenter Dokumente, die objektives Wissen enthalten**
- **Umwandlung des gefundenen objektiven Wissens in subjektives Wissen beim Nutzer (was auch heißt: der Nutzer muss die Fähigkeit haben, das entsprechende Wissen zu verstehen): Nützlichkeit**
- **Ableitung von Handlungen – aus dem gefundenen Wissen auf der Basis der eigenen Vorkenntnisse neues, handlungsrelevantes Wissen zu kreieren**

B.3 Relevanz und Pertinenz

Relevanz – Pertinenz

Voraussetzungen für erfolgreiches Retrieval:

- das richtige Wissen
- zum richtigen Zeitpunkt („just in time“)
- am richtigen Ort
- im richtigen Umfang
- in der richtigen Form
- mit der richtigen Qualität,

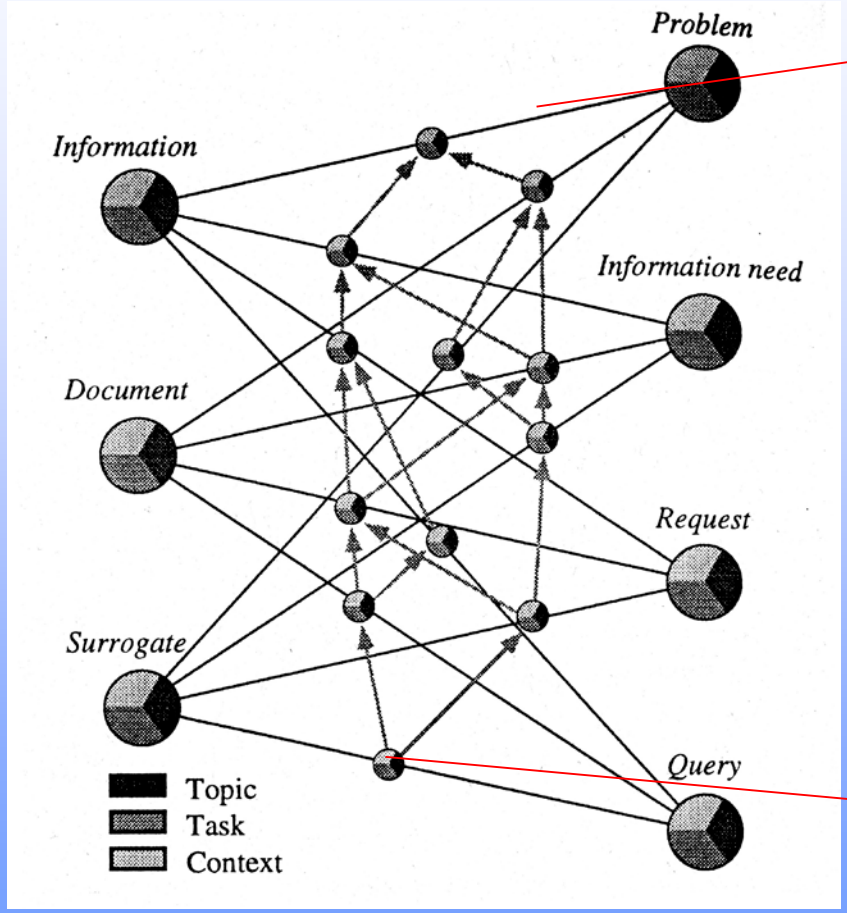
wobei „richtig“ heißt:

- (1) Wissen, Zeitpunkt usw. haben (objektiv betrachtet) Relevanz
- (2) Wissen, Zeitpunkt usw. werden vom Nutzer (subjektiv betrachtet) als passend eingeschätzt: haben Pertinenz

B.3 Relevanz und Pertinenz

- **Aspekte der Relevanz**
- **1. systemseitig:**
  - **Wissen**
  - **Dokument**
  - **Surrogat**
- **2. nutzerseitig:**
  - **Problem**
  - **Informationsbedarf**
  - **Anfrage**
  - **Konkrete Formulierung der Anfrage**
- **3. nach Thema:**
  - **Topic**
  - **Aufgabe**
  - **Kontext**
- **4. Zeit**

B.3 Relevanz und Pertinenz



Nützlichkeit

Pertinenz

„Topical Relevance“

Mizzaro, S. (1997). Relevance: The whole story.
Journal of the American Society for Information Science, 48, 810-832.

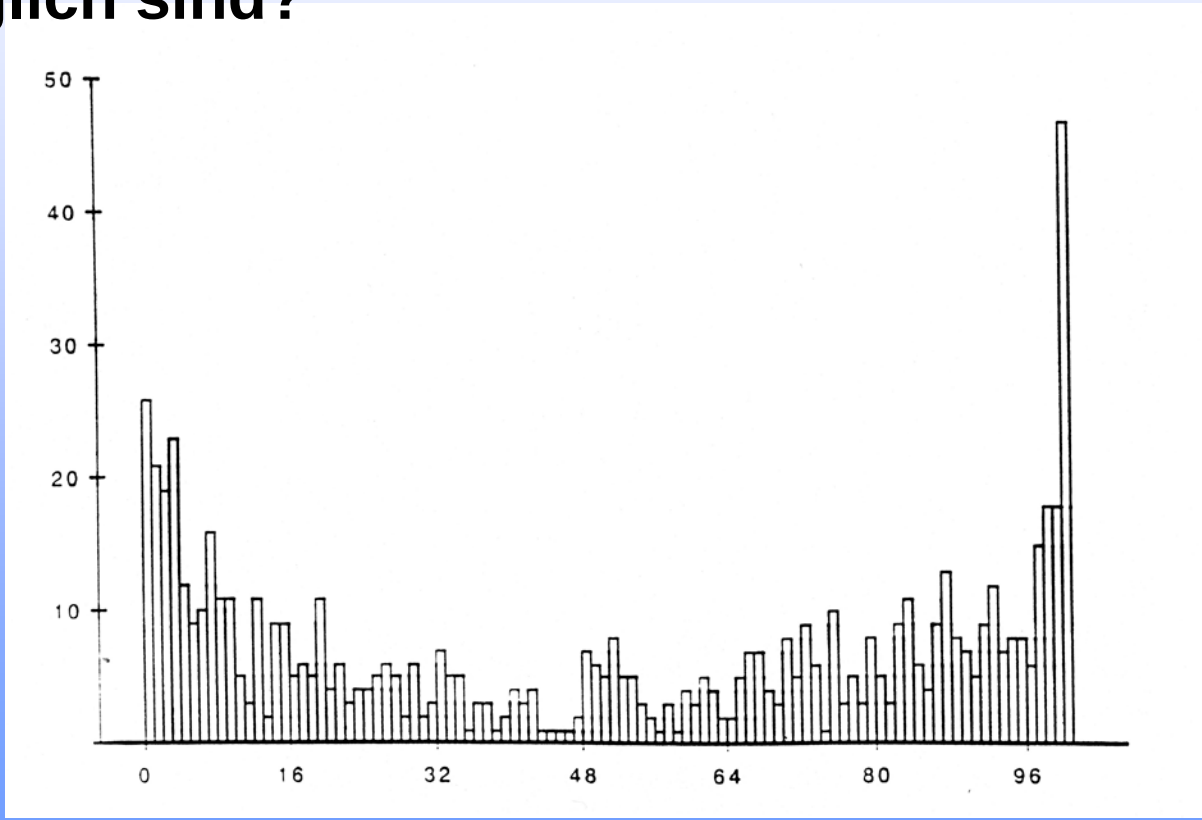
B.3 Relevanz und Pertinenz

- **Relevant oder nicht relevant: Der binäre Ansatz**
  - **0/1 Sicht**
  - **die Maße Recall und Precision bauen darauf auf**
  - **bei professionellen Retrievaltests (z.B. TReC):**
    - **Experten beurteilen, ob das Dokument Stellen enthält, die bei einer vorgegebenen Aufgabe Verwendung finden können**
    - **dies unabhängig davon, ob schon Parallelstellen vorliegen (das Problem eigentlich schon gelöst ist)**

Voorhees, E.M. (2005). The philosophy of information retrieval evaluation. Lecture Notes in Computer Science, 2406, 355-370.

B.3 Relevanz und Pertinenz

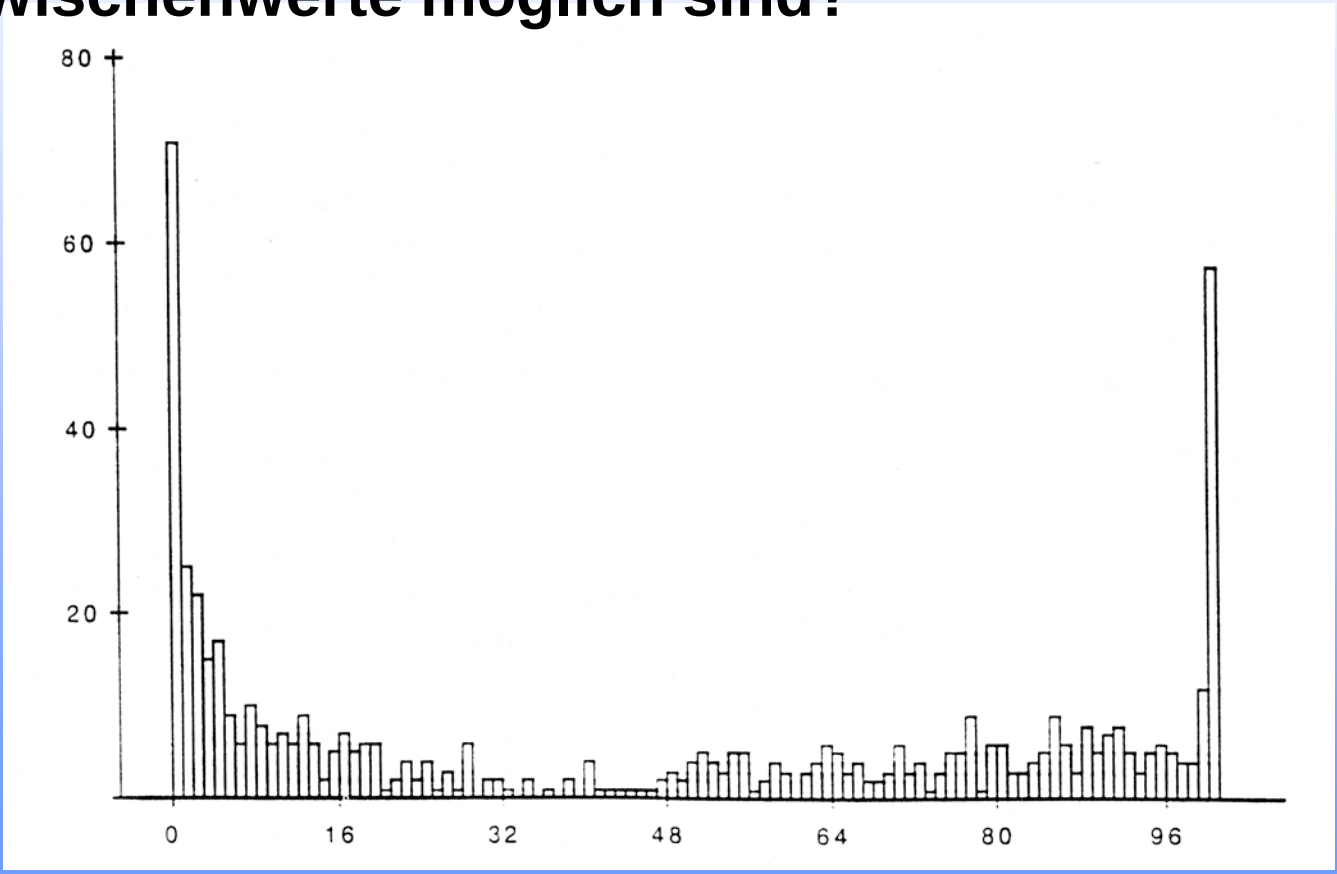
- **Wie werten *Endnutzer*, wenn Zwischenwerte möglich sind?**



Janes, J.W. (1993). On the distribution of relevance judgments.
In ASIS '93. Proceedings of the 56th ASIS Annual Meeting (pp. 104-114). Medford, NJ: Learned Information.

B.3 Relevanz und Pertinenz

- **Wie werten *Information Professionals*, wenn Zwischenwerte möglich sind?**



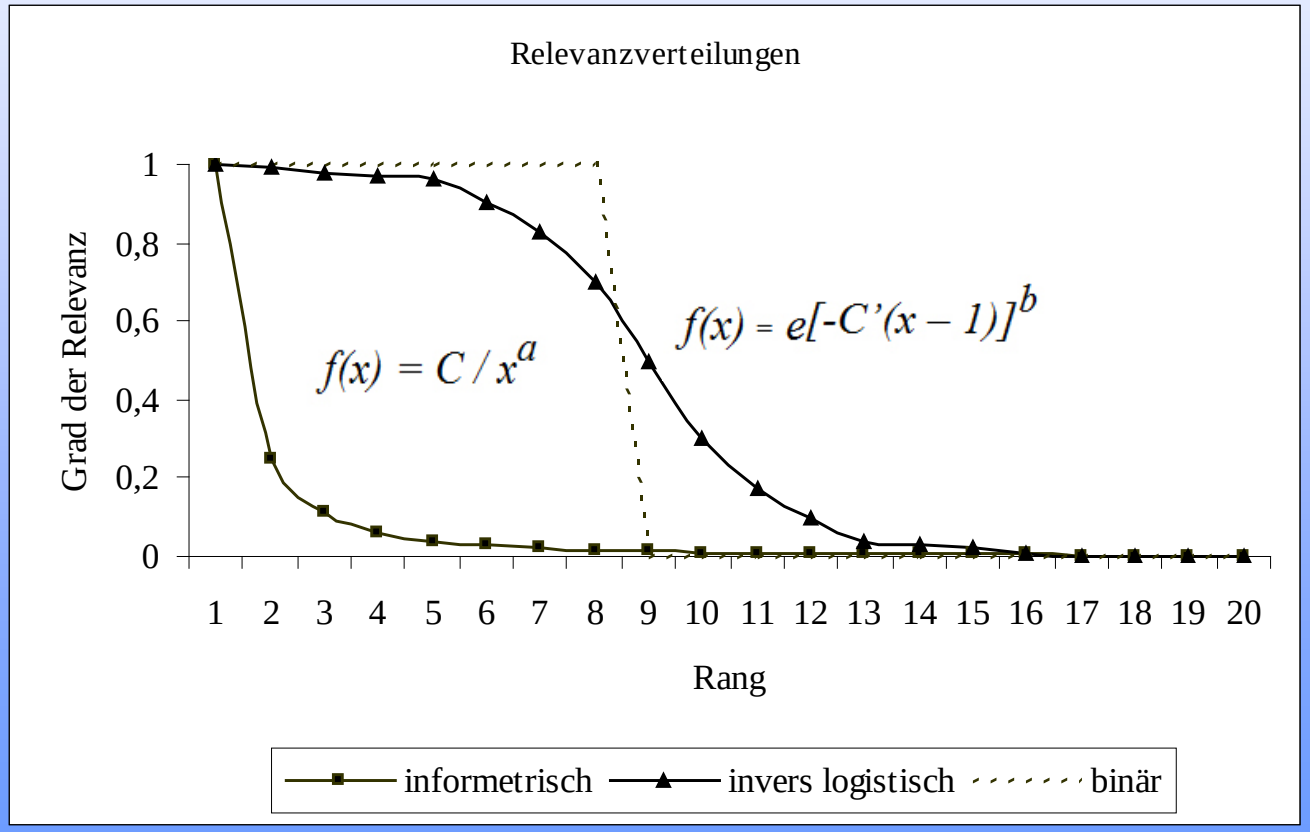
B.3 Relevanz und Pertinenz

- **Relevant oder nicht relevant?**
  - **Information Professionals tendieren weitaus stärker zur binären Sicht als die Endnutzer.**
  - **Aus Nutzersicht gibt es damit einen „Mittelbereich“.**
  - **Suchmaschinen (im Gegensatz zu frühen Retrievalsystemen) arbeiten mit Relevance Ranking, also mit Abstufungen der Relevanz.**
  - **Fazit: Der binäre Ansatz ist äußerst fragwürdig.**
  - **besser? Das gesamte Intervall [0,1] benutzen**

Della Mea, V., & Mizzaro, S. (2004). Measuring retrieval effectiveness.
A new proposal and a first experimental validation.
Journal of the American Society for Information Science and Technology, 55, 530-543.

B.3 Relevanz und Pertinenz

- **Relevanzverteilungen**
  - **Grenzen: Power Law oder invers logistisch**



Kapitel B.4

Crawler

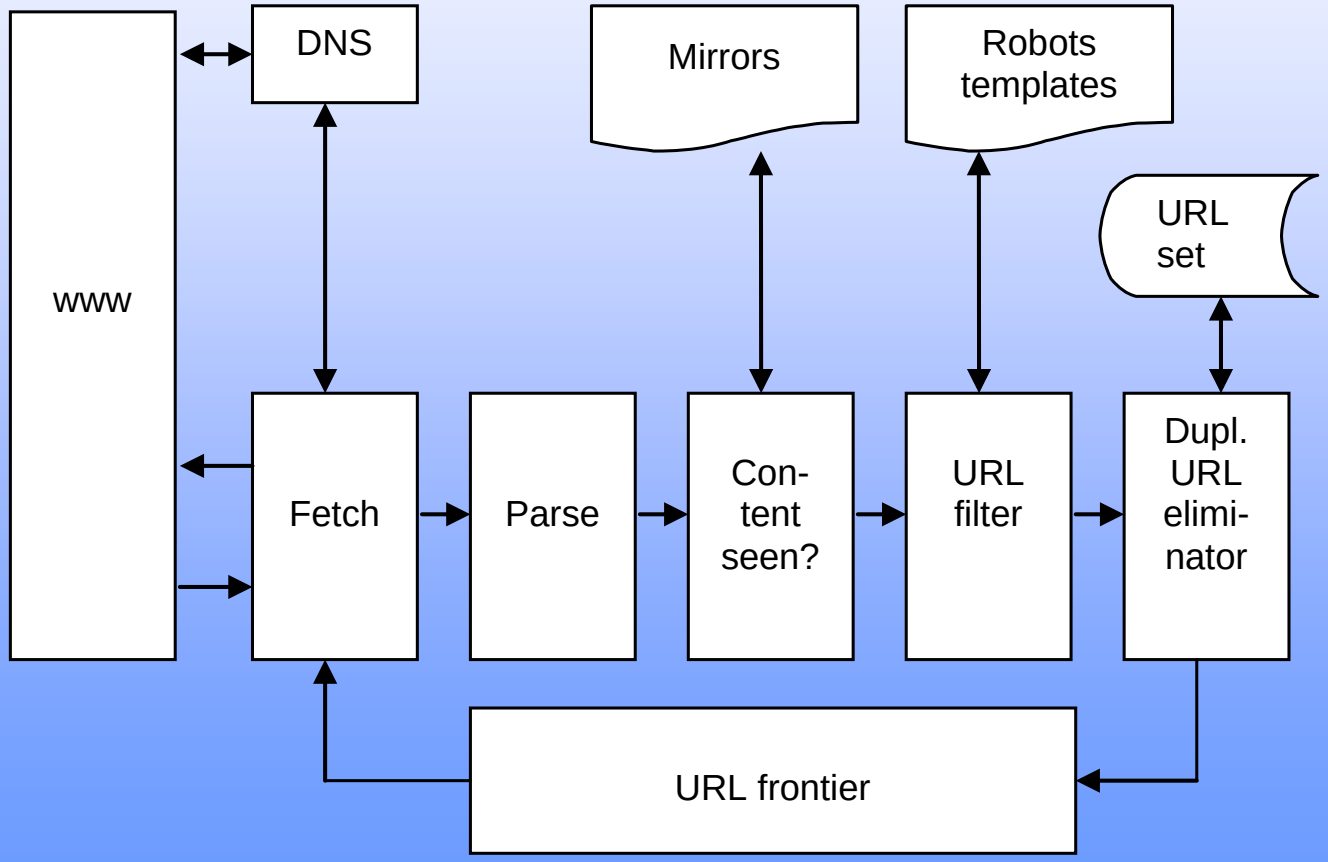
B.4 Crawler

Wie kommen die Datenbanken zu ihren Dokumenten?

- **(1) intellektuelle Auswahl nach Kriterien der Dokumentationswürdigkeit**
- **(2) automatisches Einsammeln durch Crawler (Spider, Robots)**
  - **Verfolgen der Links in bereits gesammelten Dokumenten**
  - **Beachtung von Robot Exclusion Standards**
  - **„Politeness“**
  - **alternativ: Feeds**
  - **Crawlertypen: allgemeine Crawler, thematische Crawler, Deep Web-Crawler**

B.4 Crawler

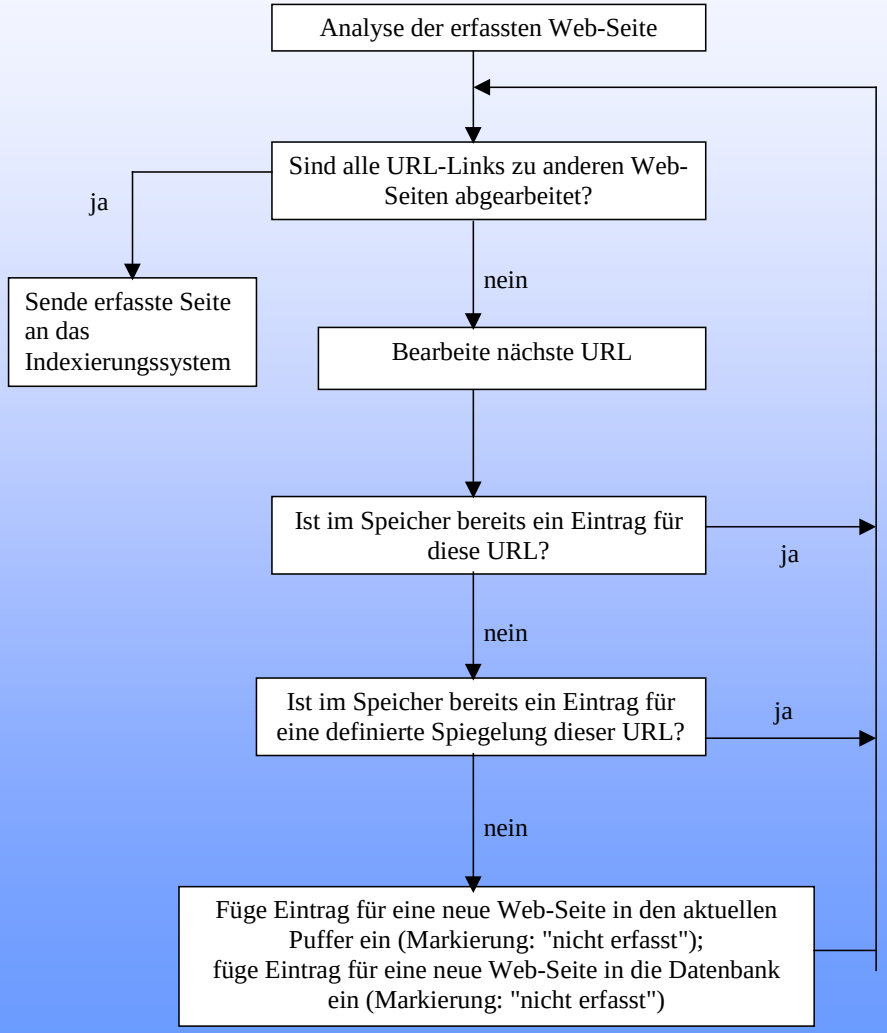
Typische Crawler-Architektur



B.4 Crawler

Crawler

Beispiel: Scooter von AltaVista

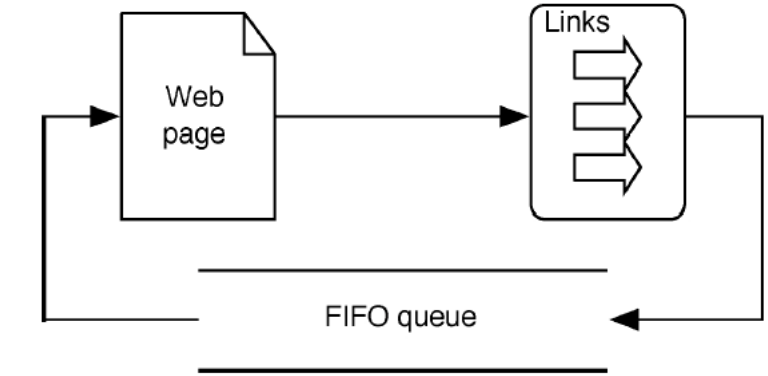


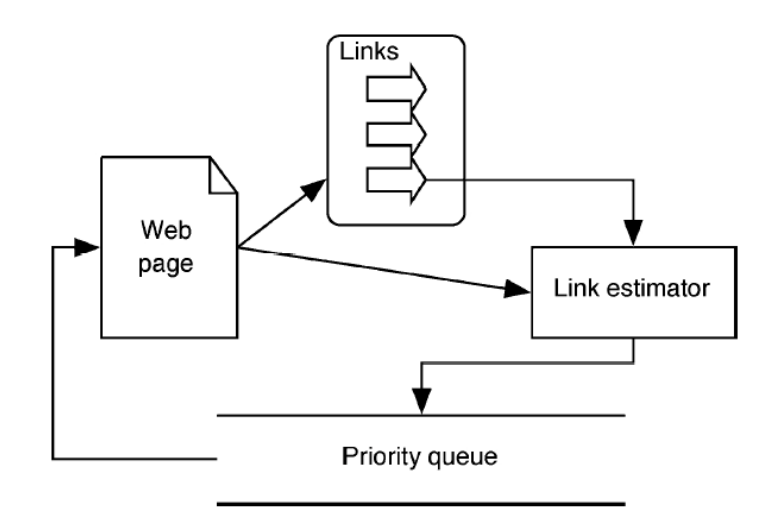
Monier, L.M. (1998). System for adding a new entry to a web page table upon receiving a web page including a link to another web page not having a corresponding entry in the web page table. Patent Nr. US 6.032.196. Assignee: Digital Equipment Corp. Priorität: 28. August 1998; erteilt: 29. Februar 2000.

B.4 Crawler

Allgemeiner Crawler

- FIFO-Crawler
  - Breadth First
  - Depth First
- Best First
  - PageRank
  - Fish-Search (Shark-Search)





B.4 Crawler

Spiegel (Dubletten) erkennen und übergehen

Erkennungskriterien:

- Pfadstruktur

- ausgehende Links

- nützlich: Spracherkennung (zur Identifikation übersetzter Sites)

Bharat, K. &, Broder, A. (1999). Mirror, mirror on the Web. A study of host pairs with replicated content. In Proceedings of the 8th International World Wide Web Conference (pp. 1579-1590). New York, NY: Elsevier.

B.4 Crawler

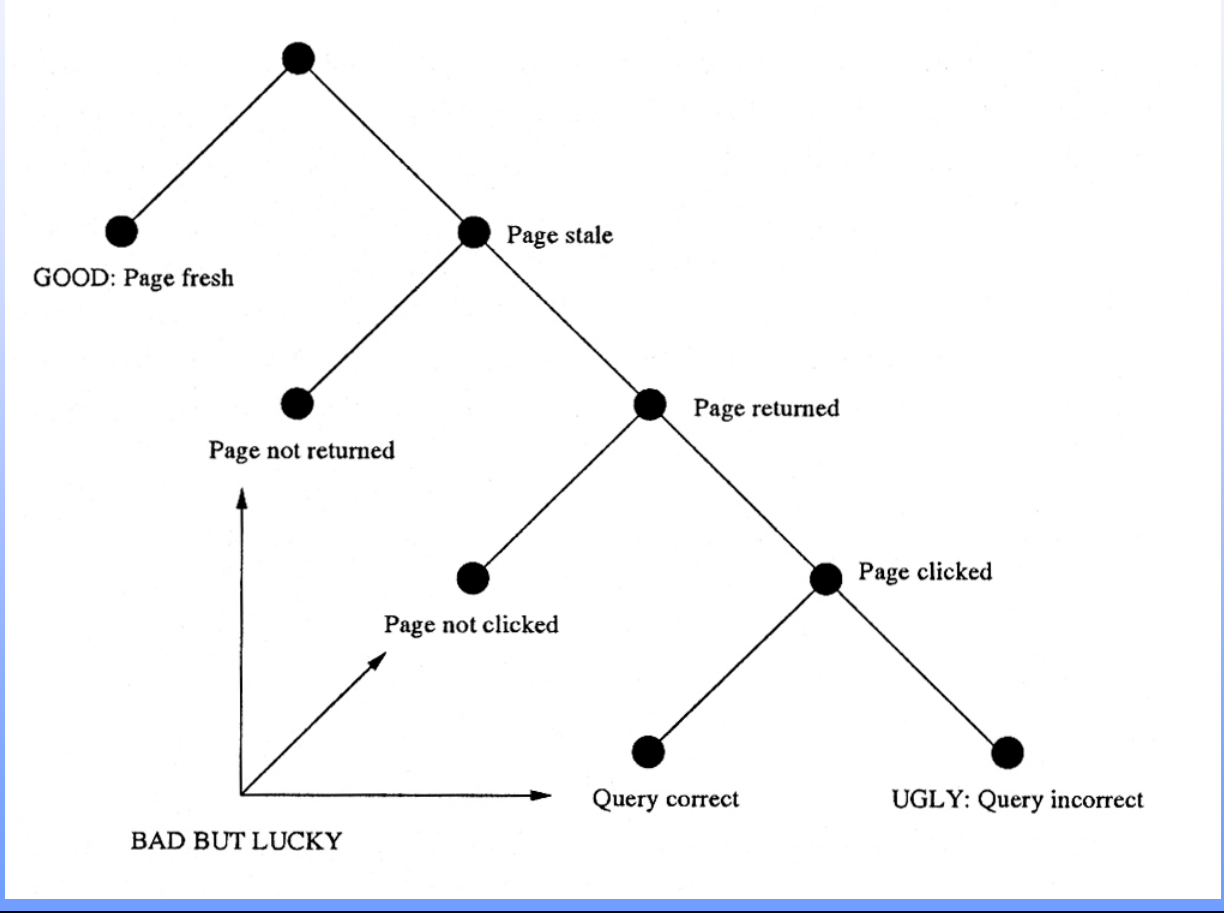
Aktualisierung der Datenbank

- **stabile dokumentarische Bezugseinheiten: Änderungen an der DBE nicht nötig, aber an der DE (z.B. neue Zitationen bei wissenschaftlichen Artikeln, neuer Rechtsstand bei Patenten)**
- **instabile dokumentarische Bezugseinheiten (viele Webdokumente): Crawler sorgt für Aktualisierung**
  - **Besuch der Seiten im selben Abstand**
  - **seitenzentrierte Aktualisierung**
  - **nutzerzentrierte Aktualisierung**

Cho, H., & Garcia-Molina, H. (2003). Effective page refresh policies for Web crawlers. ACM Transactions on Database Systems, 28(4), 390-426.

B.4 Crawler

Wahrscheinlichkeit, dass ein Nutzer eine überholte Seite findet



Wolf, J.L. et al. (2002). Optimal crawling strategies for Web search engines. In Proceedings of the 11th International World Wide Web Conference (pp. 136-147). New York, NY: ACM.

B.4 Crawler

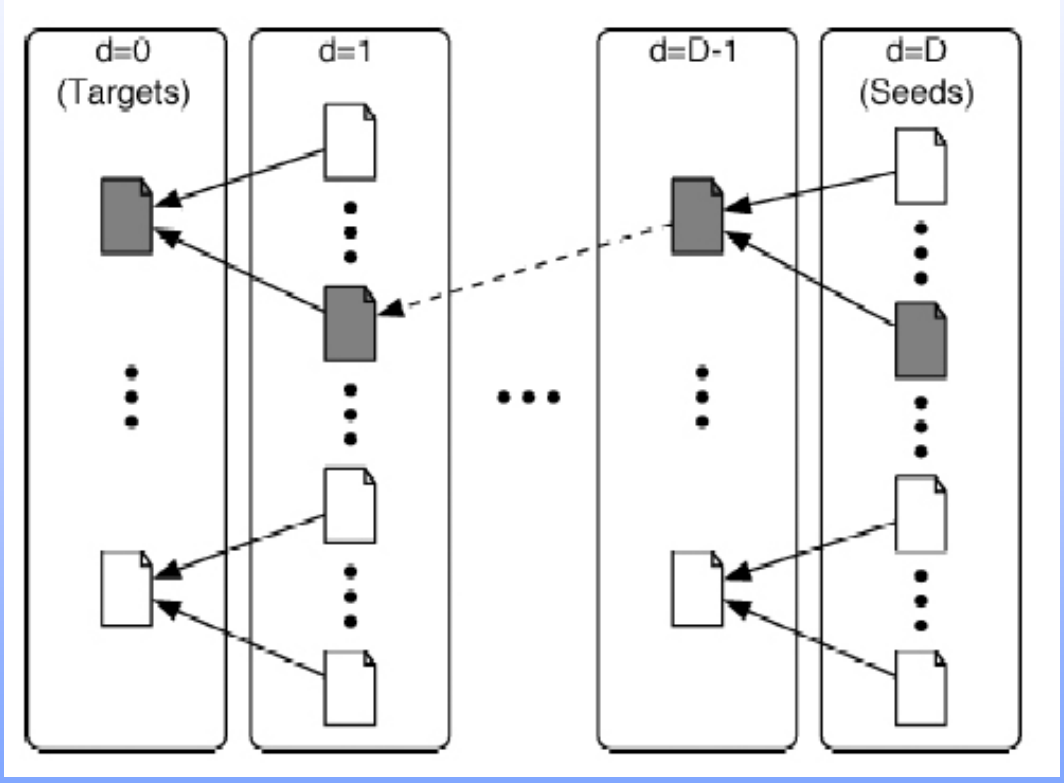
Politeness

- **Robot Exclusion Standards:**
  - **in den Meta-Tags**
  - **oder in Textdatei (robots.txt)**
- **Inhalt der Standards :**
  - **nicht indexieren („no index“)**
  - **ab hier keine Unterseiten („no follow“)**
- **Identifikation als Crawler (im Feld „user agent“)**
- **nicht zu viele Anfragen an einen Server in kurzer Zeit**

B.4 Crawler

Thematischer Crawler.

Aufgabe: Finden thematisch einschlägiger Seiten

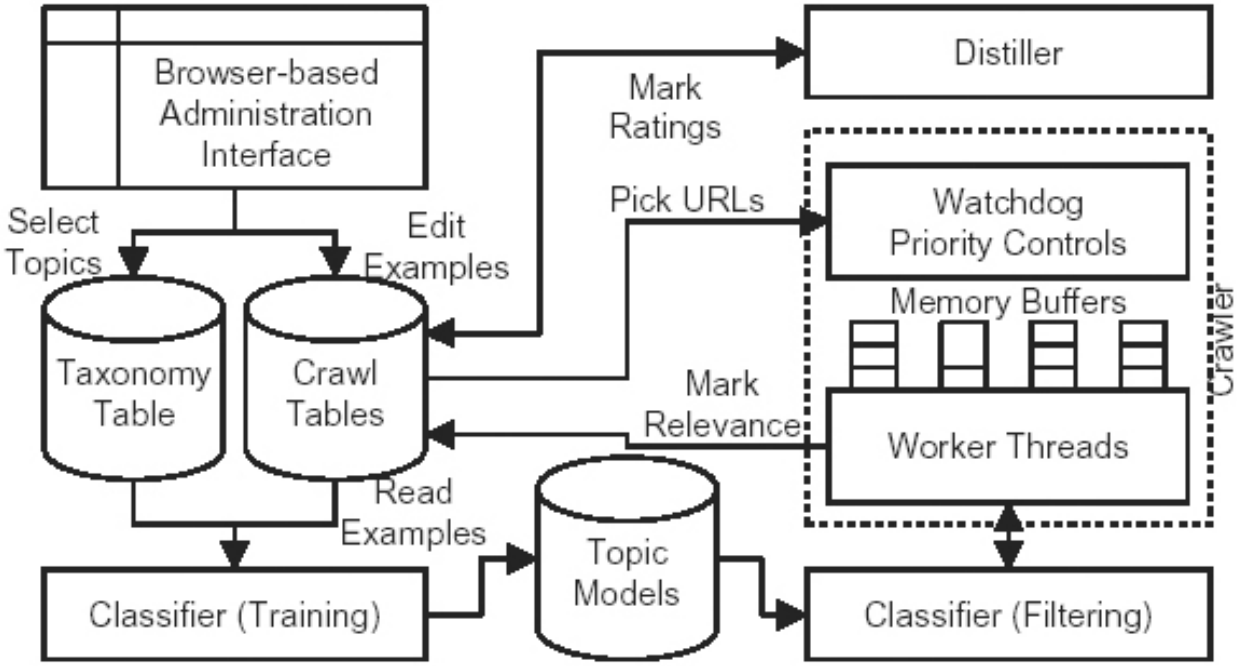


Menczer, F., Pant, G., & Srinivasan, P. (2004). Topical Web crawlers. Evaluating adaptive algorithms. ACM Transactions on Internet Technology, 4(4), 378-419.

B.4 Crawler

Thematischer Crawler.

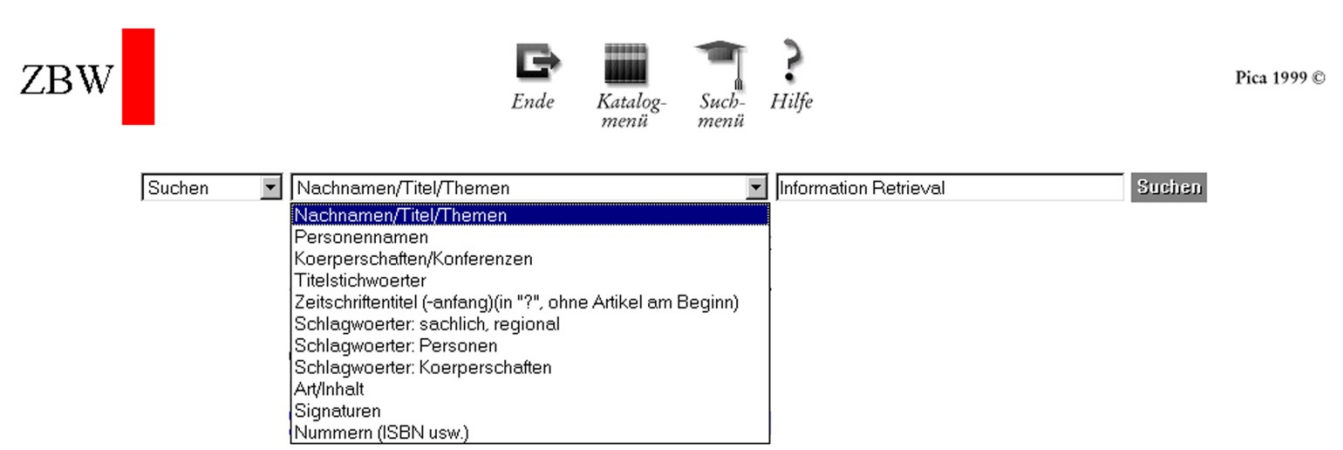
Architektur (Beispiel)



Chakrabarti, S., van den Berg, M., & Dom, B. (1999). Focused crawling. Computer Networks, 31, 1623-1640.

B.4 Crawler

Deep Web-Crawler.



Aufgaben:

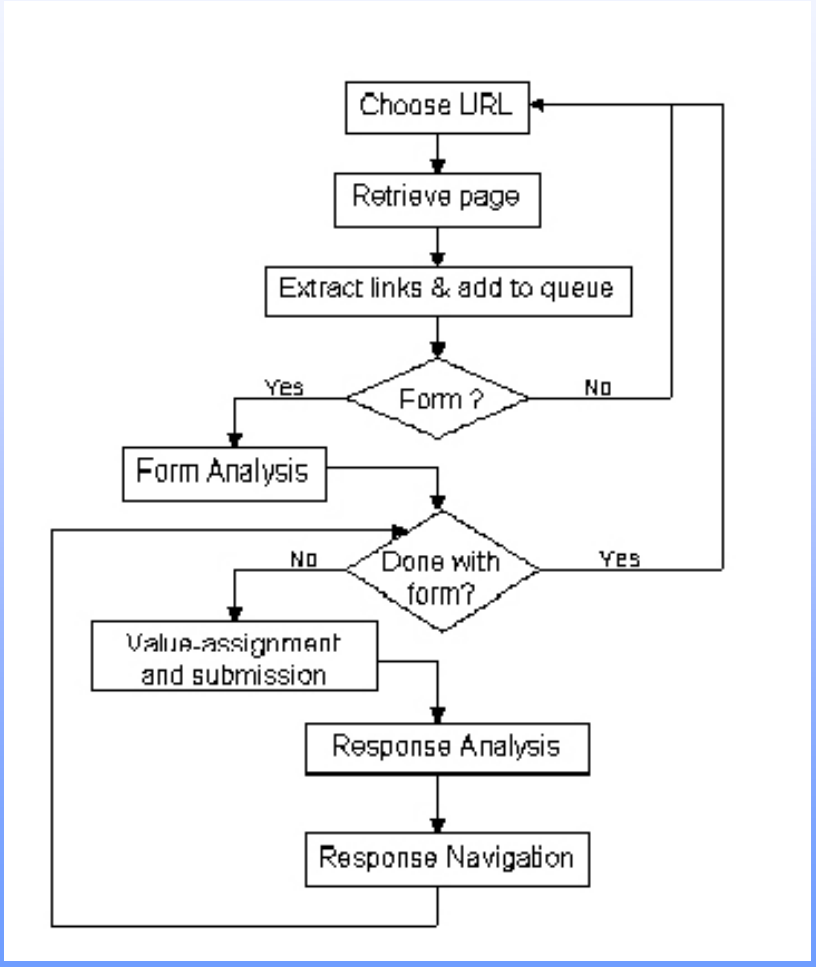
1. Suchmaske „verstehen“

2. optimale Suchargumente finden

3. Trefferliste und Dokumentanzeige „verstehen“

B.4 Crawler

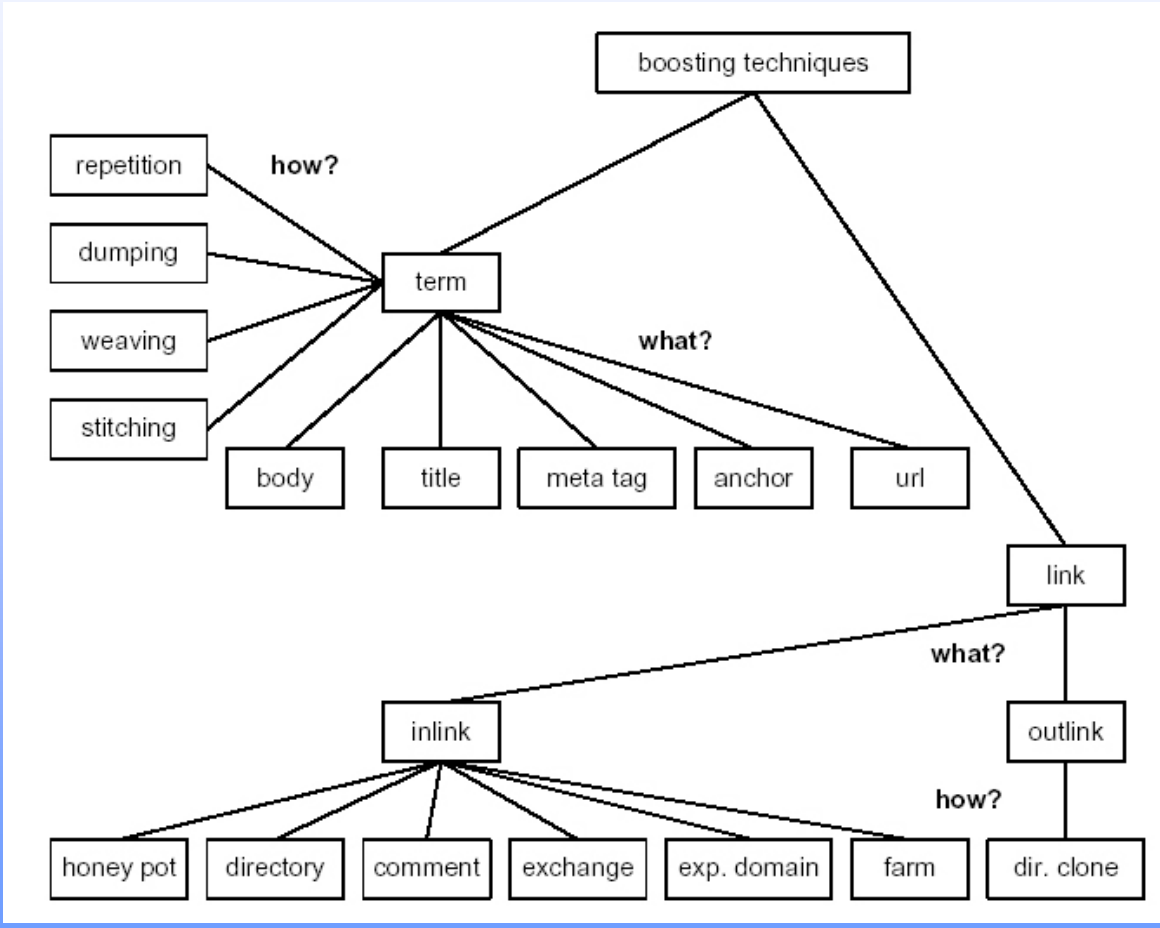
Deep Web-Crawler.



Raghavan, S., & Garcia-Molina, H. (2001). Crawling the hidden Web. In Proceedings of the 27th Conference on Very Large Databases (pp. 129-138). San Francisco: Morgan Kaufmann.

B.4 Crawler

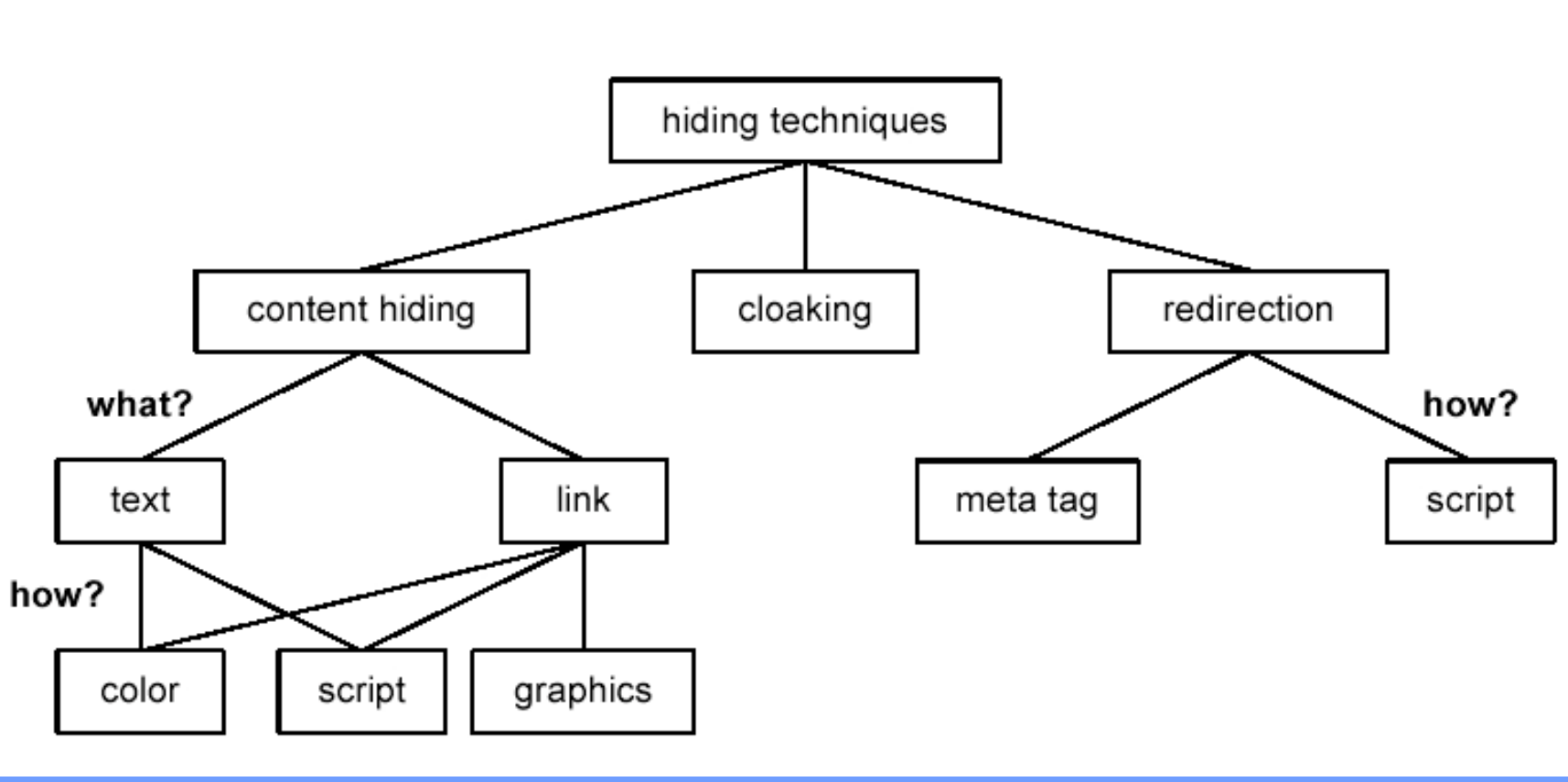
Spam



Gyöngyi, Z., & Garcia-Molina, H. (2005). Web spam taxonomy. In First International Workshop on Adversarial Information Retrieval on the Web.

B.4 Crawler

Spam



Kapitel B.5

Typologie von Retrievalsystemen

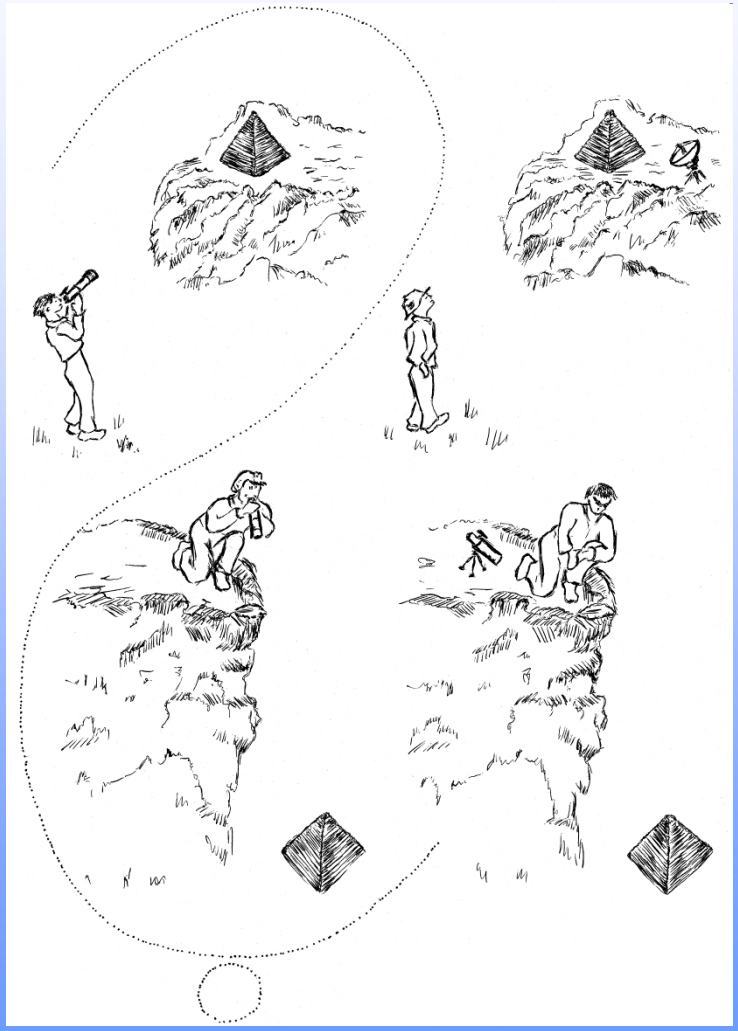
B.5 Typologie von Retrievalsystemen

- **nach Medienform der Dokumente:**
  - **textuelle Dokumente**
    - **strukturierte Dokumente**
    - **schwach strukturierte Dokumente: Gegenstand von Text-IR**
    - **nicht strukturierte Dokumente**
  - **nicht-textuelle Dokumente: Gegenstand von Multimedia-IR**
    - **gesprochene Dokumente**
    - **Musik**
    - **Bilder**
    - **Video**

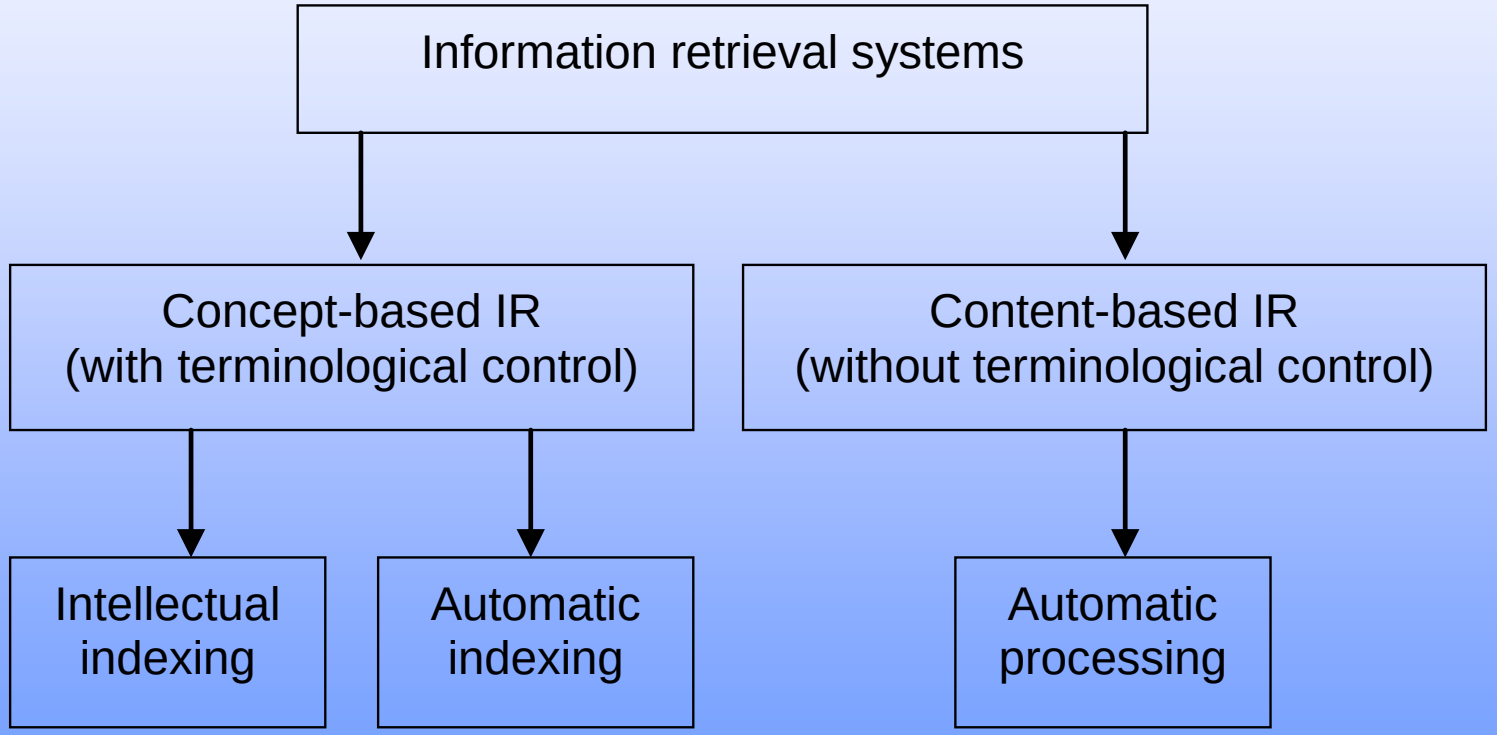
B.5 Typologie von Retrievalsystemen

Grenzen des Text-IR

- **"The man saw the pyramid on the hill with the telescope“**
- **vier mögliche Interpretationen, die IR-Systeme nicht unterscheiden können**
- **Text-IR: „man – see – pyramid – hill – telescope“**

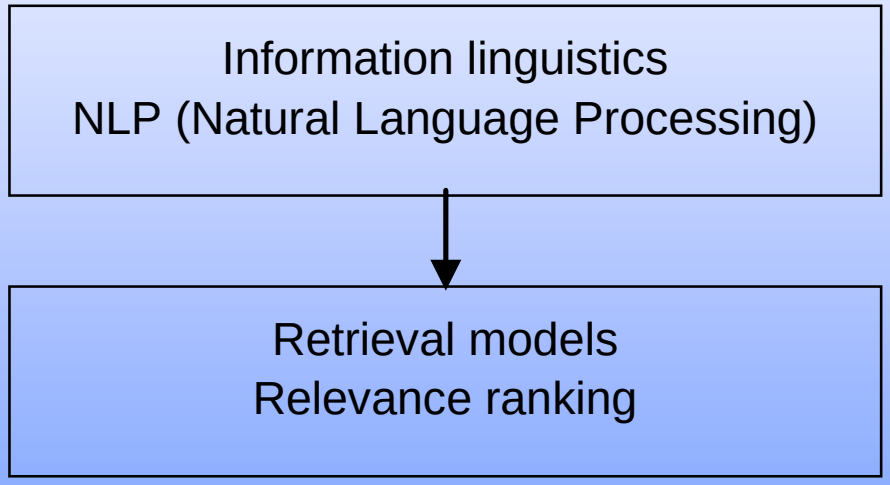


B.5 Typologie von Retrievalsystemen

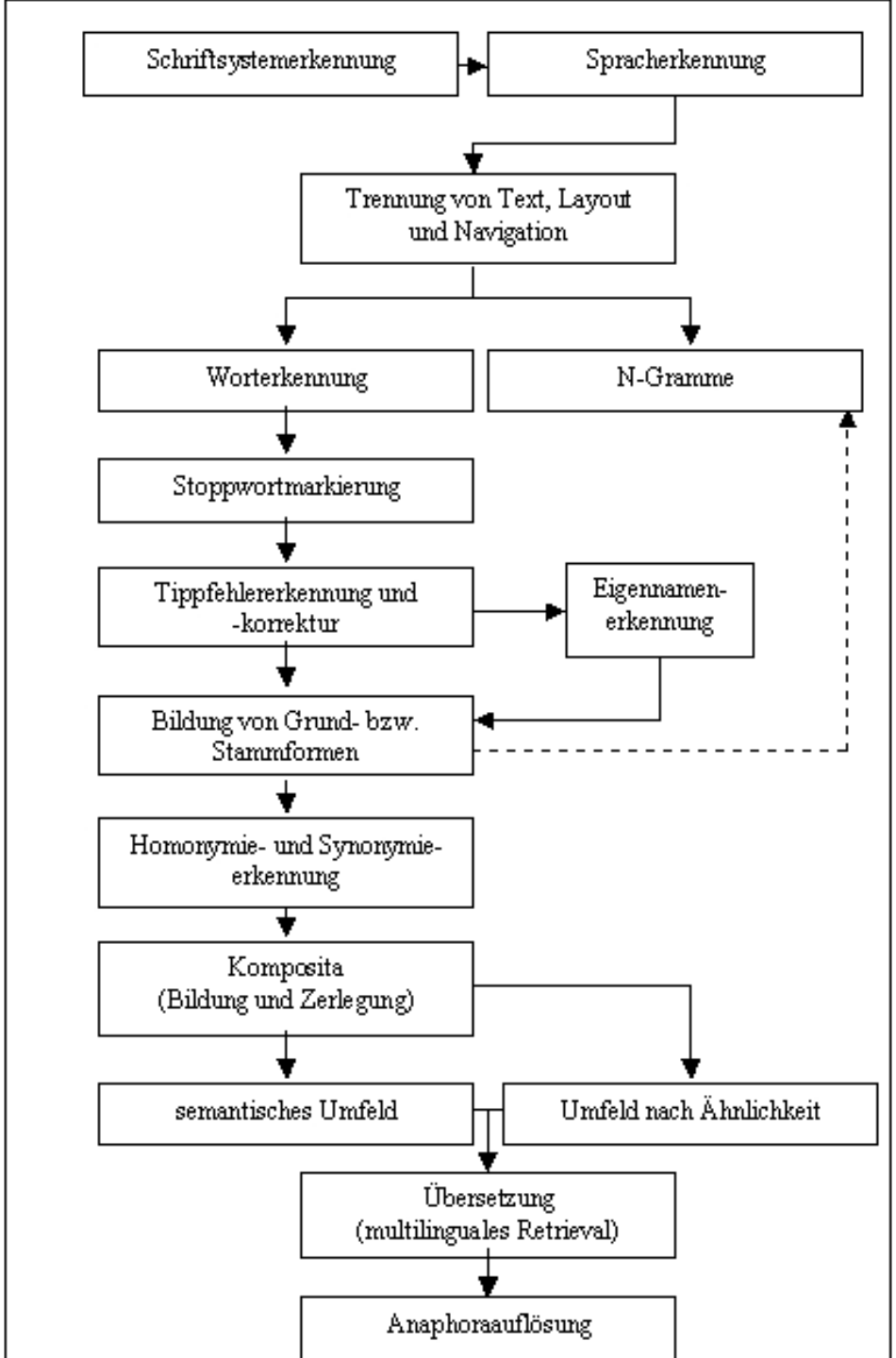


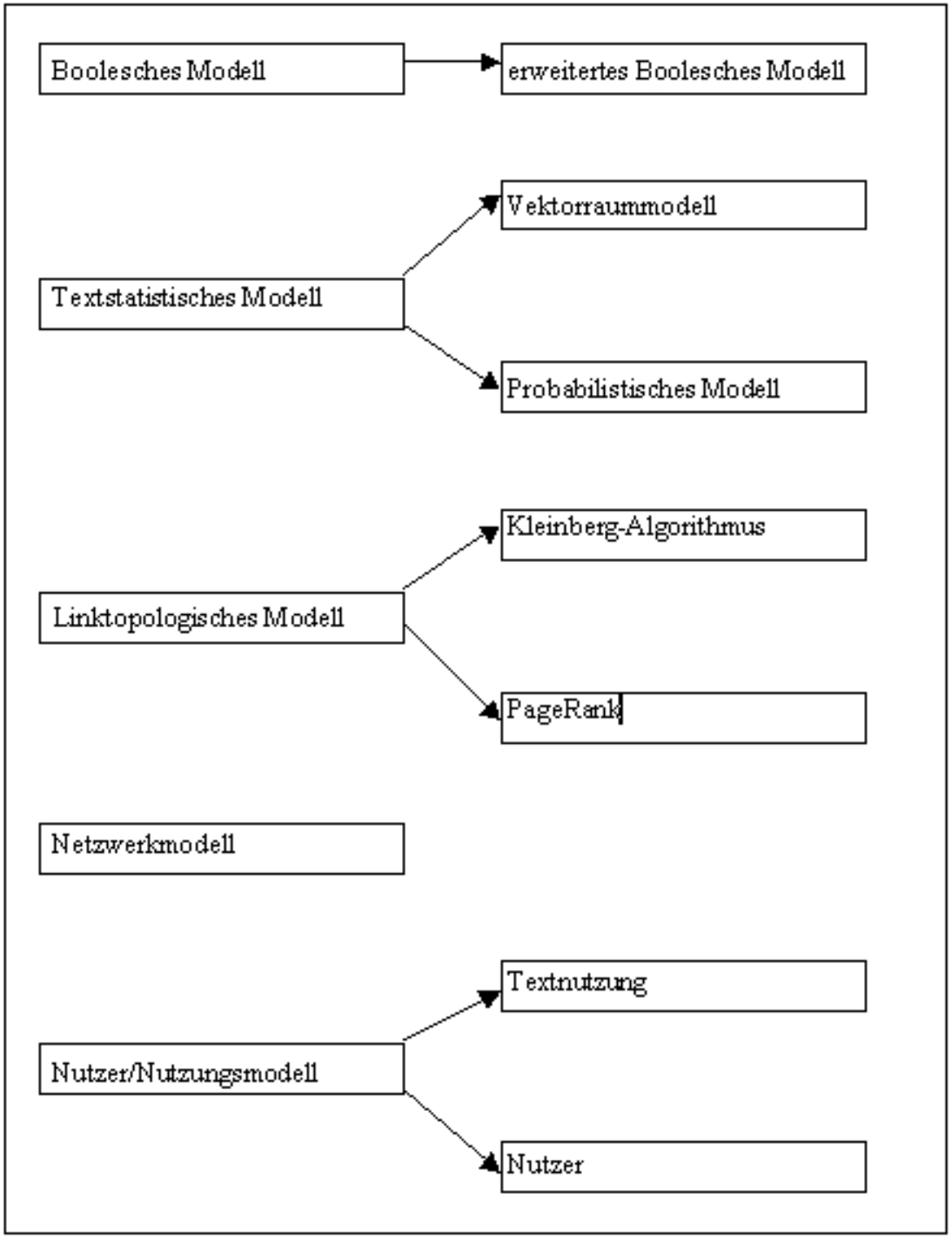
B.5 Typologie von Retrievalsystemen

Content-based IR



Informationslinguistik





Retrievalmodelle

B.5 Typologie von Retrievalsystemen

Retrievaldialog

- **Nutzer: initiale Suchanfrage**
- **System: Tippfehlernachfrage (-korrektur)**
- **System: Homonyme trennen, Synonyme zusammenfassen**
- **Nutzer: Tippfehler / Homonyme / Synonyme abarbeiten**
- **System: 1. Trefferliste**
- **Nutzer: Suchfrageerweiterung (Suchargumente streichen, neue hinzufügen, Verknüpfungen modifizieren, ...)**
- **System: 2. Trefferliste; top platzierte Dokumente vom Nutzer nach Relevanz einschätzen lassen**
- **Nutzer: Relevanzurteile angeben**
- **System: Relevance Feedback; neue Trefferliste**
- **usw. bis Nutzer endgültig zufrieden**

B.5 Typologie von Retrievalsystemen

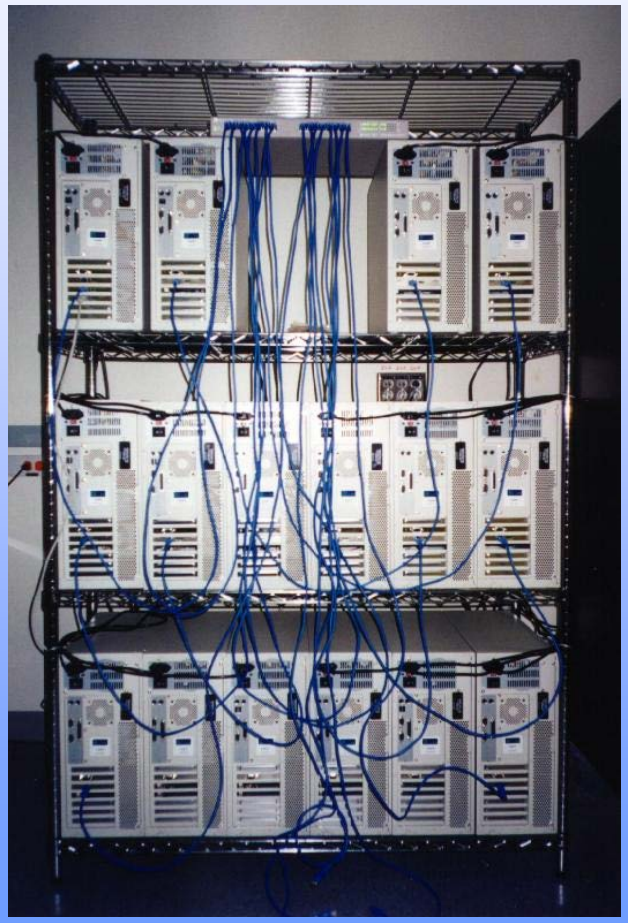
Weltregionen des Internet

1. Oberflächen-Web – 2. Deep Web

- **(1.) Die „Oberfläche“**
  - **digitale Dokumente *im* Web**
  - **(prinzipiell) auffindbar durch Suchwerkzeuge**
  - **Dokumente sind u.U. unerwünschter Ballast („Spam“)**
- **Suchwerkzeuge / Typen:**
  - **Suchmaschinen**
  - **Webkataloge**
  - **Meta-Suchmaschinen**
  - **(Social Media-Dienste)**

B.5 Typologie von Retrievalsystemen

Suchmaschinen (Search Engines)



bing

NAVER

Яндекс

Baidu 百度

Google™

B.5 Typologie von Retrievalsystemen

Suchmaschinen (Search Engines)

- **Gegenstand: Dokumente im WWW, gerichtet auf einzelne Webseiten**
- **automatisches Einsammeln der Dokumente mittels Crawler**
- **automatisches Aktualisieren der Datenbasis**
- **Kopieren der Dokumente (oder von Teilen) in die eigene Datenbank**
- **automatisches Indexieren der kopierten Dokumente**
- **eher große Datenbasis (mehrere Mrd. Dokumente)**
- **Suchsystem mit natürlichsprachiger Eingabe**

B.5 Typologie von Retrievalsystemen

Webkataloge (Web-Directories)

B.5 Typologie von Retrievalsystemen

Webkataloge (Web-Directories) – inzwischen nahezu„ausgestorben“

- **Gegenstand: Dokumente im WWW, gerichtet vor allem auf Einstiegsseiten in Websites**
- **intellektuelle Auswahl**
- **intellektuelles Indexieren (i.d.R. Klassifikation)**
- **Datenbasis: „Titel“ der Dokumente (vom Webkatalog oder vom Anmeldenden vergeben) und URL**
- **eher kleine Datenbasis (einige Mio. Dokumente)**
- **unregelmäßiges Update**
- **Suchsysteme mit Klassifikationshierarchien und natürlichsprachiger Suche (über die Klassenbezeichnungen und die Dokumenten„titel“)**

B.5 Typologie von Retrievalsystemen

Meta-Suchmaschinen

MetaGer SUMA-EV

B.5 Typologie von Retrievalsystemen

Meta-Suchmaschinen

- **keine Datenbasis; greifen auf die Datenbasen anderer Suchwerkzeuge zurück („Schmarotzer“)**
- **Relevance Ranking über mehrere Suchmaschinen hinweg**

B.5 Typologie von Retrievalsystemen

Weltregionen des Internet
1. Oberflächen-Web – 2. Deep Web

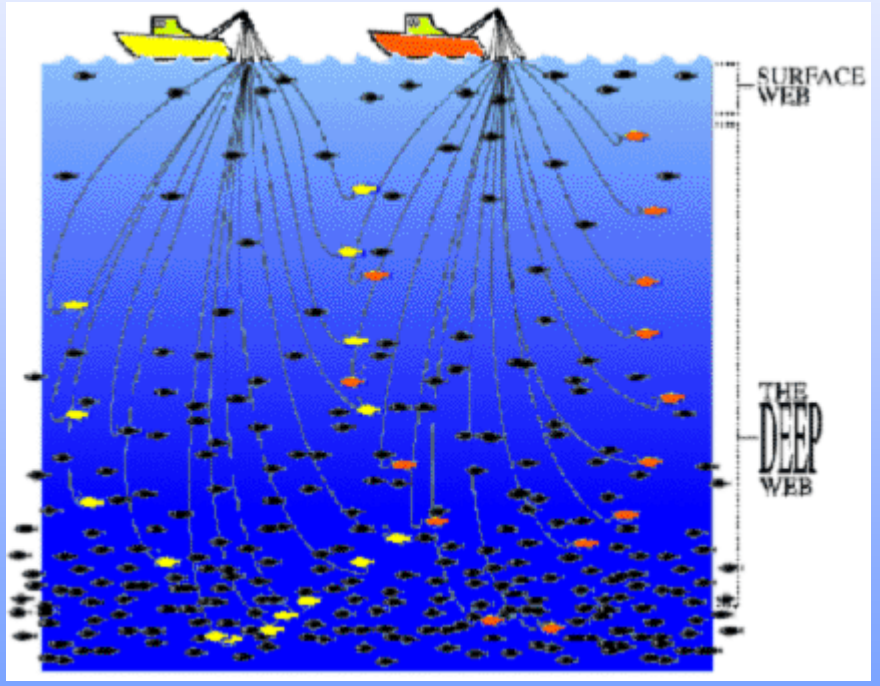
- **(2.) Das „Deep Web“ (oder „Invisible Web“)**
  - **digitale Dokumente, die nicht direkt im Web liegen, aber *via* Web erreichbar sind**
  - **derzeit nicht (alle) auffindbar durch Suchwerkzeuge**
  - **Dokumente sind (meist) qualitätsgeprüft**
  - **Terminologie:**
    - **„invisible Web“ – Sherman & Price**
    - **„Deep Web“ – Bergman (Schätzung: Deep Web ist 500mal größer als das Oberflächenweb – wahrscheinlich stark überschätzt)**

Sherman, C., & Price, G. (2001). The Invisible Web. Medford, NJ: Information Today.
Bergman, M.K. (2001). The Deep Web: Surfacing Hidden Value. The Journal of Electronic Publishing, 7(1).

B.5 Typologie von Retrievalsystemen

„Deep Web“

- **Typen:**
  - **Singuläre Datenbanken**
    - **Kostenfrei**
    - **Kommerziell**
  - **Content-Aggregatoren**
    - **Kostenfrei**
    - **Kommerziell**



Quelle: Bergman, 2001

B.5 Typologie von Retrievalsystemen

(Kostenfreie) Singuläre Datenbanken im Web

- **thematisch orientierte Datenbanken**
- **(i.d.R.) aufgebaut von öffentlichen Einrichtungen (durch öffentliche Mittel bereits finanziert)**
- **mehrere tausend Datenbanken via Web erreichbar**

ZBW

DPMA

PubMed

United States Patent and Trademark Office

B.5 Typologie von Retrievalsystemen

(Kommerzielle) Singuläre Datenbanken im Web – „Selbstvermarkter“

- **thematisch orientierte Datenbanken**
- **(i.d.R.) aufgebaut von Privatunternehmen mit dem Zweck der Erzielung von Gewinnen**

B.5 Typologie von Retrievalsystemen

(Kostenfreie) Content-Aggregatoren (Fachportale)

- **Bündelung von Datenbanken (zu einem Thema)**
- **Bibliograpische Informationen: kostenfrei**
- **Volltext: kostenpflichtig**
- **Beispiel: ZB MED**

B.5 Typologie von Retrievalsystemen

Kommerzielle Content-Aggregatoren („Online-Hosts“)

- Bündelung von Datenbanken (allgemein oder zu einem Thema): von mehreren hundert bis über 30.000 einzelnen Datenbanken
- kostenpflichtig
- Typen:
  - WTM-Hosts (STN, Ovid, DIMDI, Questel, Web of Knowledge, Springer Link, Science Direct, ...)
  - Wirtschaftshosts (Genios, Factiva, DIALOG, Profound, LexisNexis, ...)
  - Rechtshosts (Juris, Beck online, LexisNexis, Westlaw, ...)

B.5 Typologie von Retrievalsystemen

Weltregionen des Internet: Grenzüberschreitungen

- **Hybrid-Suchmaschine (Content-Aggregator *und* WWW-Suchmaschine)**

B.5 Typologie von Retrievalsystemen

Weltregionen des Internet und die Welt gedruckter Dokumente

... wenn die benötigten Dokumente nur in Printausgaben vorliegen:

Nutzung von Document Delivery Services

subito
Lieferdienst der Bibliotheken

home
ingenta

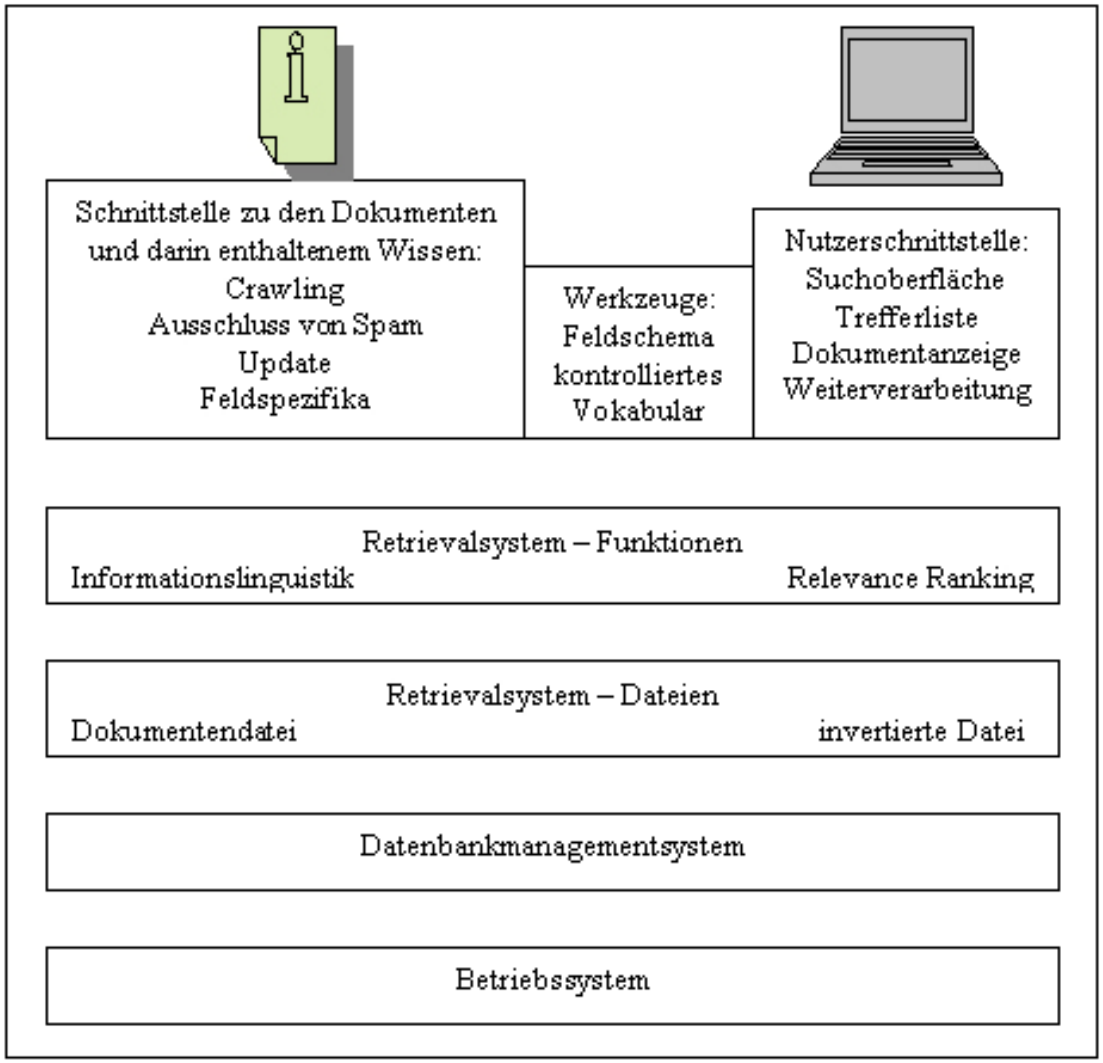
INFOTRIEVE

Oßwald, A. (2004). Document Delivery / Dokumentlieferung. In Grundlagen der praktischen Information und Dokumentation, 5.Aufl. (pp. 571-578). München: Saur.

Kapitel B.6

Architektur eines Retrievalsystems

B.6 Architektur eines Retrievalsystems



B.6 Architektur eines Retrievalsystems

Zeichensätze

- **ASCII 7-bit-Code (128 Zeichen)**

```
1000111111001011101011100101111001111001101000001000
1111101111111010011101000100001
```

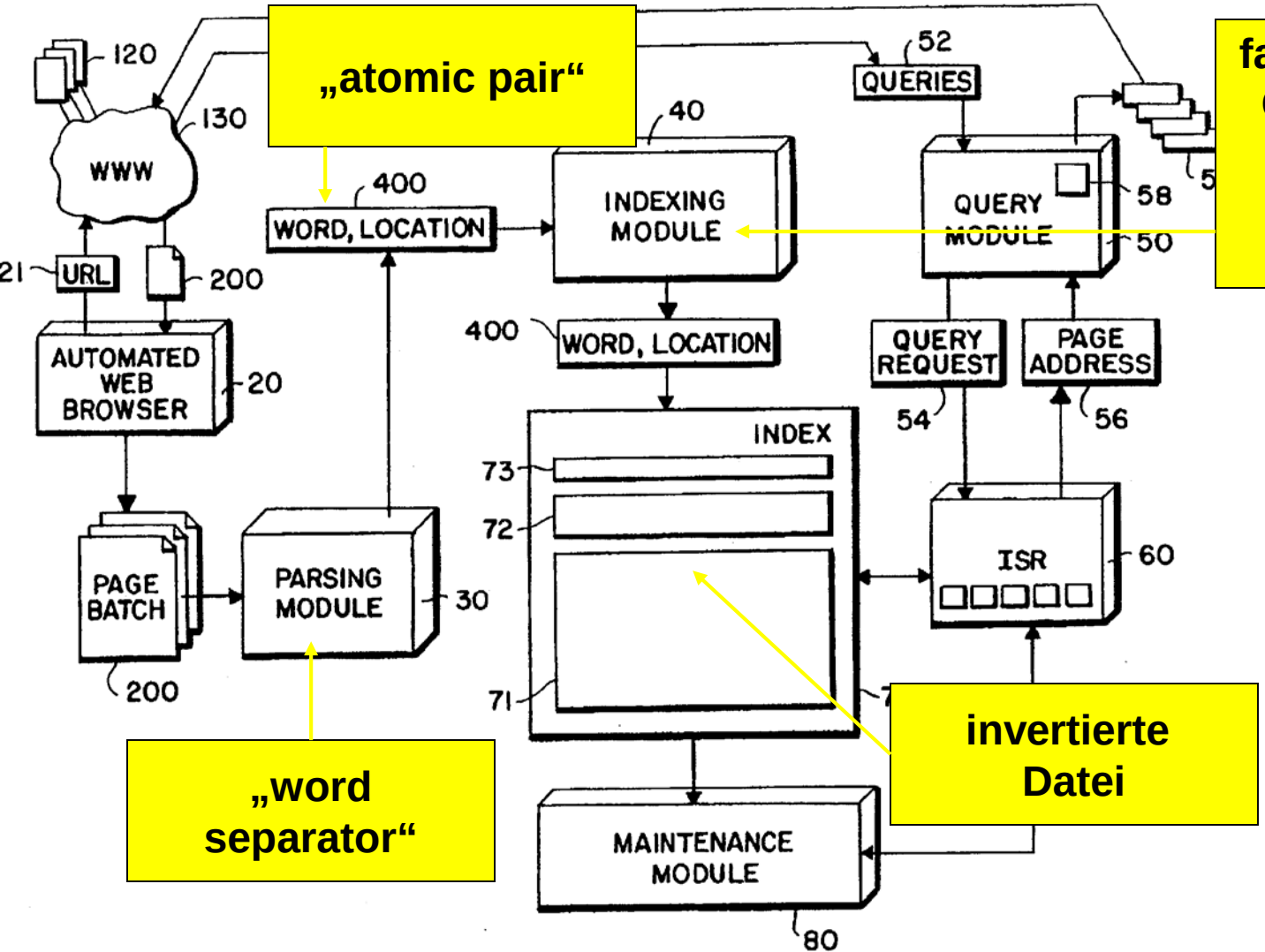
- **ASCII 8-bit-Code (256 Zeichen; die oberen 128 Zeichen werden sprachspezifisch vergeben)**

```
0100011101110010111111001101111100100000010001
1101101111011101000111010000100001
```

- **Unicode / UCS (Universal Multiple-Octet Coded Character Set): bis zu 4 Byte (32 bit). Ziel: alle Sprachsysteme der Welt abbilden**

B.6 Architektur eines Retrievalsystems

Speicherung von Wort und Ort. *Beispiel:* AltaVista



„atomic pair“

„word separator“

invertierte Datei

falls 1 Buchstabe ein Großbuchstabe ist: „Synonym“; *Bsp.:* [1, To] [1, to]

Burrows, M. (1996). Method for indexing information of a database. Patent Nr.: US 5,745,899. Priorität: 9.8.1996. Inhaber: Digital Equipment Corp.

B.6 Architektur eines Retrievalsystems

- **Vorgabe eines Feldschemas. Beispiel: Fachdatenbank (ifoDok)**

Eingabefeld	*Feldname bei Anzeige*
lfd. Nr.	
Kennung	
Dokumenttyp	PU
Verfasser	AU
Herausgeber	
Abteilung des Verfassers	
Titel Aufsatz	TI
Rubrik	
Zeitschrift	
Buchtitel	
Titel der Buchreihe	
Nr. in Reihe	
Band / Heft	
Jahrgang	YR
Ausgabedatum	
Verlagsort	CI
Verlag	CO
ISSN	IS
ISBN	IB
Seiten	
Sprache	LA
Deskriptoren	DE
Thesaurus-Identnummern	DN
weitere Themen	IW
Sachklassifikation	KL
Klassifikationscode	KC
Ländercode	LC
Abstract	AB
Signatur	SI
Kurzbezeichnung	
Endekriterium	
Quelle	SO (wird generiert)

B.6 Architektur eines Retrievalsystems

Dateien

- **Dokumentenspeicher (sequentielle Aufnahme aller Daten eines Dokumentes) – Zuordnung einer eindeutigen Dok.-Nr.**
- **Invertierte Dateien: feldspezifische (i.d.R. alphabetische) Listen aller Einträge eines Feldes aller Dokumente – unter Zuordnung der Dok.-Nr. und weiterer Angaben**
- **Basic Index: Invertierte Datei über bestimmte Felder (je nach System alle Felder oder Auswahl)**
- **Wortindex: jedes einzelne Wort ist Indexeintrag**

  Phrasenindex: zusammengehörige Phrasen bilden *einen* Indexeintrag

B.6 Architektur eines Retrievalsystems

Invertierte Datei

- eigene Adresse im Speicher
- Dokumentnummer(n) bzw. deren Adresse(n)
- Häufigkeit in Gesamtdatenbank
  - Anzahl der Dokumente, in denen der Eintrag (min. einmal) vorkommt
  - Gesamtanzahl des Vorkommens in der Datenbank
- Position(en) im Dokument
  - Wortnummer(n)
  - Vorkommen in Satz/Sätzen Nummer(n) X, X‘, ...
  - Vorkommen in Absatz/Absätzen Nummer(n) Y, Y‘, ...
  - beim Einsatz syntaktischen Indexierens: Vorkommen in Themenkette(n) T, T‘, ...

B.6 Architektur eines Retrievalsystems

Invertierte Datei

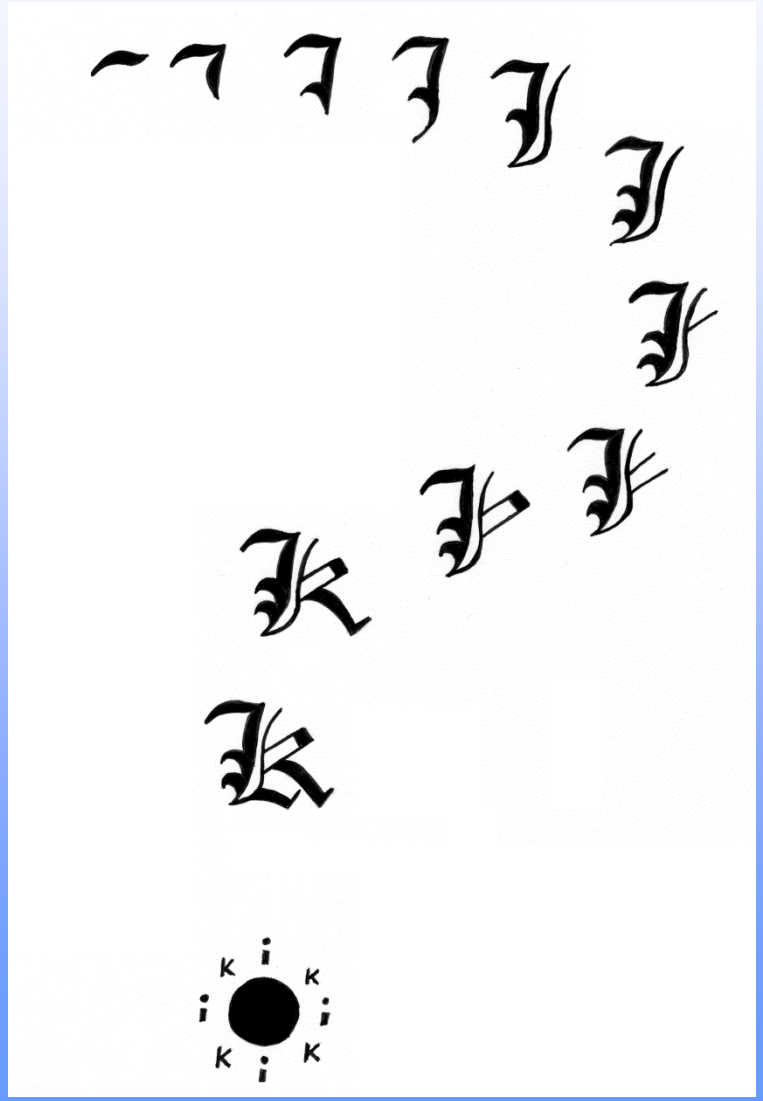
- **ggf.: Kennzeichen auf Position (z.B. Größe des Druckerfonts)**
- **ggf.: Gewichtungswert**
- **ggf. jeder Eintrag zweimal: normale Buchstabenfolge und zusätzlich rückläufig**

B.6 Architektur eines Retrievalsystems

Invertierte Dateien. *Beispiel* (Textbody)

Eintrag:	**Unternehmen / nemhenretnU**
Dok.-Nr.	**2, 23, 45, 56**
# Dok.	**4**
# insg.	**7**
Wort-Nr.	**(2: 4, 28), (23: 99), (45: 13, 17, 55), (56: 432)**
Satz-Nr.	**(2: 1, 3), (23: 15), (45: 9, 9, 15), (56: 58)**
Absatz-Nr.	**(2: 1, 1), (23: 1), (45: 1, 1, 2), (56: 4)**
Font	**(2.4: 28), (2.28: 10), (23.99: 12), (45.13: 72), (45.17: 12), (45.55: 12), (56.432: 20)**

Teil C: Verarbeitung natürlicher Sprache



Kapitel C.1

n-Gramme

C.1 *n*-Gramme

Worte

- **natürlichsprachiges Wort (steht zwischen zwei Leer- oder Satzzeichen)**
- **formales Wort der Länge n (n-Gramm)**
  - **Varianten von n-Grammen im Retrieval:**
  - **1. Zerlegung identifizierter natürlichsprachiger Worte in Zeichenfolgen zu n**
  - **2. wie 1., aber hier unter Auffüllen von Leerzeichen am Anfang und am Ende**
  - **3. gleitende n-Gramme über den Text (sinnvoll: Trennung bei Satz- oder Absatzende)**

C.1 *n*-Gramme

INFORMATION RETRIEVAL

in Variante 2 in Tetragrammen

*****I, **IN, *INF, INFO, NFOR, FORM, ORMA, RMAT, MATI, ATIO, TION, ION*, ON**, N*****

*****R, **RE, *RET, RETR, ETRI, TRIE, RIEV, IEVA, EVAL, VAL*, AL**, L*****

in Variante 3 mit gleitenden Tetragrammen

*****I, **IN, *INF, INFO, NFOR, FORM, ORMA, RMAT, MATI, ATIO, TION, ION*,**

ON*R, N*RE, ⟵ neu

RET, RETR, ETRI, TRIE, RIEV, IEVA, EVAL, VAL*, AL**, L**

C.1 *n*-Gramme

Anzahl der n-Gramme ist begrenzt: $|Alphabet|^n$

***Bsp.:* deutsches Alphabet: 26 Zeichen plus Leerzeichen**

bei n=3 → 27^3 = 19.683 3-Gramme

bei n=4 → 27^4 = 531.441 4-Gramme

bei n=5 → 27^5 = 14.348.907 5-Gramme

- **Vergleich: arbeitet man mit Worten, so ist deren Anzahl (zumindest prinzipiell) unendlich groß**
- **nicht alle möglichen n-Gramme sind auch besetzt**
- **englisch (3-Gramme): nur 16% faktisch vorhanden**

C.1 *n*-Gramme

Vorteile der n-Gramme gegenüber Wörtern:

- **überschaubares, endliches Material**
- **besonders geeignet bei Sprachen ohne Wortgrenzen (chinesisch, japanisch, koreanisch)**
- **keine weiteren Algorithmen wie bei der Wortbearbeitung (Morphologie, Zerlegung von Mehrwortausdrücken usw.)**
- **Relevance Ranking direkt anhand der n-Gramme**
- **kostengünstig**

C.1 *n*-Gramme

Nachteile:

- **semantische „Fallen“ sind möglich („Widerspruchsfreiheit“)**
- **kein semantisches Umfeld**
- **Präzision suboptimal (allerdings nur im Vergleich mit hochentwickelten Algorithmen der Wortbearbeitung)**
- **Probleme mit Flexionen**
  - **Umlautung (Fuchs – Füchsin)**
  - **Ablaute (singen – Gesang)**
  - **Zirkumfigierung (stöhnen – Gestöhne)**
  - **Infixe (wie im Arabischen)**

C.1 *n*-Gramme

Pentagramm-Register

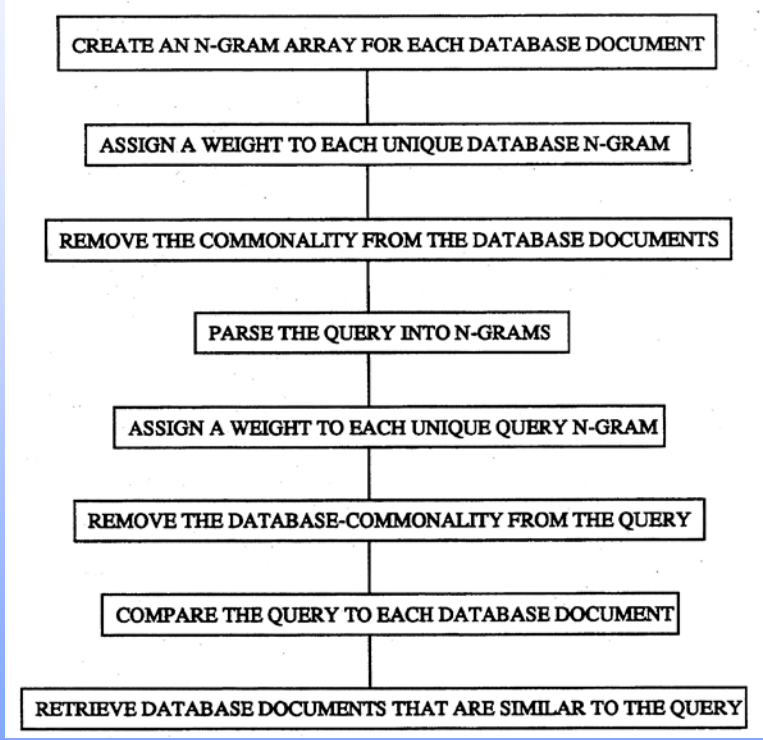
- **Identifikation von Registereinträgen innerhalb (auch langer) Worte**
- **Beispiel: WIDERSPRUCHSFREIHEITSBEWEIS**

  WIDER, IDERS, DERSP, ERSPR, RSPRU, **SPRUC**, PRUCH, RUCHS, UCHSF, CHSFR, HSFRE, SFREI, **FREIH**, **REIHE**, EIHEI, IHEIT, HEITS, EITSB, ITSBE, TSBEW, SBEWE, **BEWEI**, EWEIS
- **ins Register übernommen werden die sinnvollen Pentagramme (ein anderes Wort im Register beginnt mit derselben Zeichenfolge)**
- **so findet man „Widerspruchsfreiheitsbeweis“ auch unter „Beweis“, „Freiheit“, „Reihe“ und „Spruch“**

Henrichs, N. (1975). Sprachprobleme beim Einsatz von Dialog-Retrieval-Systemen. In Deutscher Dokumentartag 1974, Bd. 2 (pp. 219-232). München: Verl. Dokumentation.

C.1 *n*-Gramme

ACQUAINTANCE (Vektorraum-IR-System)



„commonality“:

Zentroid-Vektor

Damashek, M. (1994). Method or retrieving documents that concern the same topic. Patent-Nr. US 5418951. Patentanmelder: The United States of America as represented by the Director of National Security Agency, Washington, D.C. Erteilt am: 23.5.1995. (Eingereicht am 30.9.1994).

C.1 *n*-Gramme

ACQUAINTANCE

M =: Dokument mit m Dimensionen (d.h. unterschiedlichen n-Grammen)

N =: anderes Dokument (z.B. Suchanfrage) mit n Dimensionen

j =: Anzahl der Dimensionen

x(m,j) / y(n,j) =: Gewichtungswert eines n-Gramms aus M bzw. N (relative Häufigkeit)

***μ*(j) =: Gewichtungswert des Zentroiden in Dimension j**

Relevance Ranking nach Cosinus:

$$S_{mn} = \frac{\sum_{j=1}^{J}(x_{mj}-\mu_j)(y_{nj}-\mu_j)}{\left[\sum_{j=1}^{J}(x_{mj}-\mu_j)^2\sum_{j=1}^{J}(y_{nj}-\mu_j)^2\right]^{1/2}} = \cos\theta_{mn}, \quad m,=1,\ldots,M, \quad n=1,\ldots N$$

C.1 *n*-Gramme

HAIRCUT (Probabilistisches IR-System)

- **Berechnung eines Ähnlichkeitswertes zwischen Such-n-Grammen und n-Grammen in Texten**
- **P: Wahrscheinlichkeit für Relevanz (rechte Gleichungsseite: relative Häufigkeit)**
- **D: Dokument (Text)**
- **C: gesamte Datenbank (collection)**
- **Q: Suchanfrage; q: n-Gramm aus Suchanfrage**
- **α: Konstante (Glättungsparameter)**

$$P(D|Q) = [\alpha * P(q_1|D) + (1- \alpha) * P(q_1|C)] * \ldots * [\alpha * P(q_n|D) + (1- \alpha) * P(q_n|C)]$$

McNamee, P., & Mayfield, J. (2004). Character *n*-gram tokenization for European language text retrieval. Information Retrieval, 7, 73-97.

C.1 *n*-Gramme

HAIRCUT – Arbeitsschritte

- **Erkennung von Sätzen**
- **Löschen von Stoppworten (z.B. „the“) (nicht Stopp-n-Grammen; Bsp.: Trigramm „the“ in mathematics)**
- **gleitende n-Gramme über die Sätze**
- **Errechnung der relativen Häufigkeiten**
- **Relevance Ranking nach HAIRCUT-Formel**

C.1 *n*-Gramme

n-Gramme: Welches n für welche Sprache?

Average precision using 4-, 5-, and 6-grams for each language using the optimal value for smoothing.

	4-grams		5-grams		6-grams	
	MAP	α	MAP	α	MAP	α
Dutch	0.4284	0.7	**0.4323**	**0.6**	0.4163	0.5
English	**0.4917**	**0.9**	0.4718	0.8	0.4420	0.7
Finnish	0.3839	0.5	**0.3968**	**0.3**	0.3760	0.2
French	**0.4097**	**0.7**	0.3899	0.7	0.3605	0.5
German	**0.4055**	**0.9**	0.4013	0.6	0.3782	0.7
Italian	**0.4051**	**0.7**	0.4041	0.6	0.3805	0.6
Spanish	**0.4678**	**0.7**	0.4479	0.5	0.4069	0.7
Swedish	**0.4312**	**0.8**	0.4116	0.6	0.3827	0.2

Maximal values are highlighted. The best performance using 3- or 7-grams was worse. All differences between 5-grams and 6-grams are statistically significant at the 0.05 level; differences between 4-grams and 5-grams are only significant for Swedish.

C.1 *n*-Gramme

N-Gramme: Vorteile

- **IR-Systeme sind unabhängig von der Verarbeitung natürlicher Sprache**
- **sehr sinnvoll bei Sprachen ohne Wortgrenzen**
- **auch einsetzbar bei verstümmeltem Text**
- **n-Gramme sind im IR an diversen Stellen einsetzbar:**
  - **Identifikation der Sprache**
  - **Fehler-tolerantes Retrieval**
- **Nachteil**
- **(insb. beim Vektorraummodell) Probleme mit kurzen Anfragen**

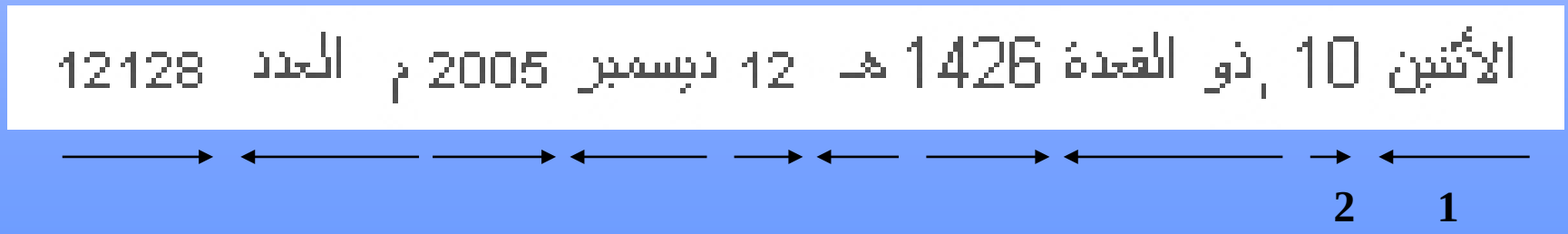
Kapitel C.2

Wörter

C.2 Wörter

Schriftsystemerkennung

- **Zeichensätze erkennen: falls Unicode eingesetzt wird, kein Problem**
- **wenn nicht: automatische Schriftsystemerkennung**
- **Vergleich der Zeichenverteilung eines Textes mit (bekannten) Zeichenverteilungen von Schriftsystemen**
- **Feststellung der Leserichtung (von links nach rechts und umgekehrt; Umkehr der Leserichtung im Text)**



C.2 Wörter

Spracherkennung

- **Ansatz 1: Mustertypen**

  typische Buchstabenkombinationen, typische Sonderzeichen

  - **ery_ : englisch**
  - **eux_ : französisch**
  - **_der_ : deutsch**
  - **lj : serbo-kroatisch**
  - **cchi : italienisch**
  - **¿ : spanisch**
  - **Å : schwedisch**

- **wenig sicher**

C.2 Wörter

– **Ansatz 2: Wortverteilungen**

Spracherkennung auf Satzebene

%	Danish	%	Dutch	%	English	%	German	%	Spanish
4.35	og	4.19	en	6.16	the	3.14	und	5.38	que
3.71	hun	3.11	t	4.31	and	2.31	die	4.75	de
1.82	i	2.34	de	2.87	a	2.62	sie	4.74	y
1.70	de	2.25	van	2.06	of	1.92	zu	2.70	la
1.59	det	1.84	een	2.02	to	1.91	der	2.57	a
1.45	var	1.62	in	1.61	was	1.59	sich	2.15	en
1.38	han	1.56	die	1.57	he	1.30	er	2.14	el
1.38	at	1.50	is	1.50	in	1.24	in	1.65	no
1.25	saa	1.23	niet	1.04	it	1.19	nicht	1.23	los
1.24	en	1.22	ick	0.86	his	1.12	das	1.22	se

- Erstellung von Wortlisten nebst Auftretenswahrscheinlichkeiten für Sprachen
- Satz: Zählen der Auftretenshäufigkeit der Worte im Satz; Multiplikation mit Auftretenswahrscheinlichkeiten aller Sprachen, Werte summieren
- „gewonnen“ hat die Sprache mit dem höchsten Wert

McNamee, P. (2005). Language identification: A solved problem suitable for undergraduate instruction. Journal of Computing Sciences in Colleges, 20(3), 94-101.

C.2 Wörter

– **Ansatz 3: n-Gramme**

Acquaintance

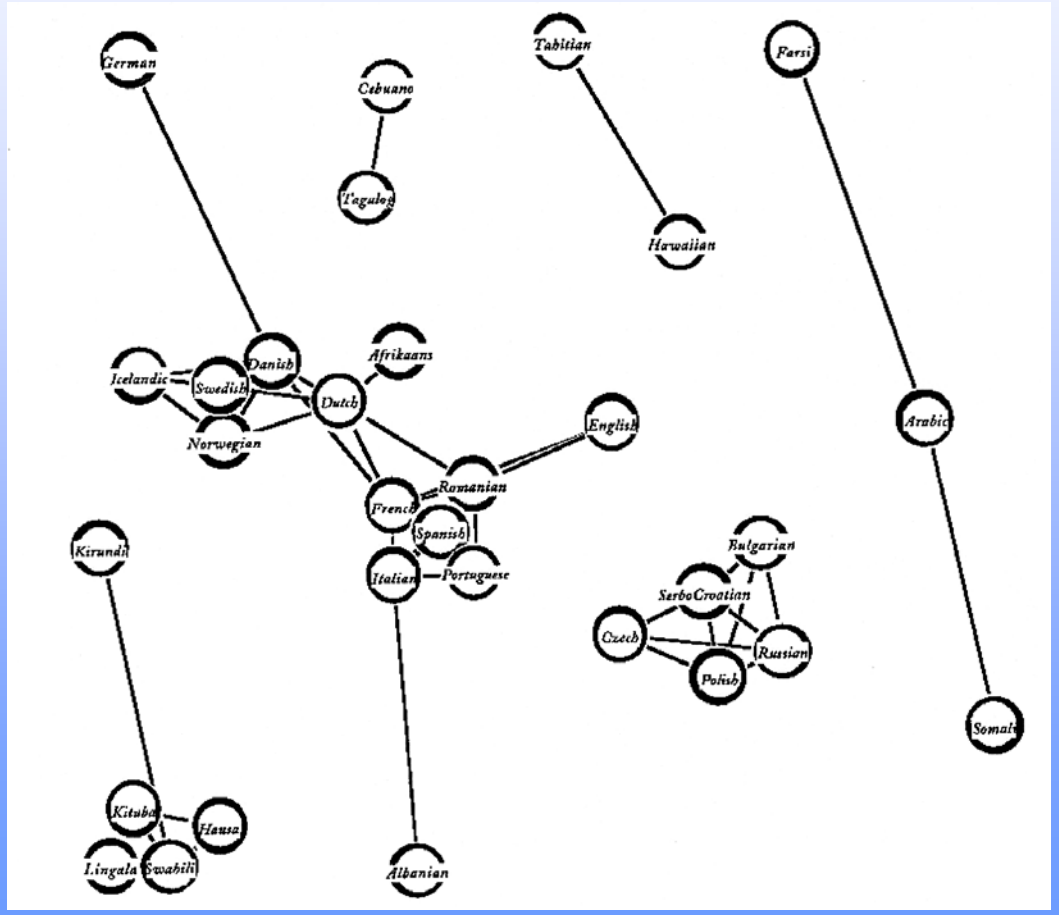
$$S_{mn} = \frac{\sum_{j=1}^{J} x_{mj} x_{nj}}{\left(\sum_{j=1}^{J} x_{mj}^2 \sum_{j=1}^{J} x_{nj}^2 \right)^{1/2}} = \cos \theta_{mn}$$

- erstellen: Zentroiden für Sprachen
- Dokument in n-Gramme parsen (diesmal Zentroiden nicht abziehen), Cosinus zwischen Dokumentvektor und Sprach-zentroiden errechnen
- auf Rang 1 liegt die wahrscheinlichste Sprache

Damashek, M. (1995). Gauging similarity with N-grams: Language-independent categorization of text. Science, 267, 843-848.

C.2 Wörter

Exkurs: Automatisches Gruppieren von Sprachen nach dem Cosinus der Sprachzentroiden (von Marc Damashek)



C.2 Wörter

Stoppworte:

- **Wort, das die gleiche Wahrscheinlichkeit hat, in einem relevanten sowie in einem nicht-relevanten Dokument vorzukommen, „non-content word“**
- **Elimination von Stoppworten ist wenig sinnvoll:**
  - **bei gewissen Phrasen werden sie benötigt: „to be or not to be“**
  - **u.U. wird gezielt nach Stoppworten gesucht: „Studien zum englischen Hilfsverb ‚to be‘“**
  - **Pronomina sind eigentlich Stoppworte. Sie werden aber (theoretisch) bei der Informationsstatistik benötigt.**
- **deshalb: Stoppworte markieren und von „normaler“ Suche ausschließen. Wenn Nutzer will, jedoch berücksichtigen**

C.2 Wörter

Stoppwortliste als Negativliste: 1. Allgemeine Stoppworte

- häufig in einer Sprache vorkommende Worte (Fox: mehr als 300mal im Brown-Corpus enthalten)
- Worte intellektuell aus Liste entfernen (z.B. business, family)
- weitere Worte intellektuell in Liste einfügen: „extra fluff words“ (z.B. above [296mal im Brown-Corpus])
- Zusatzliste „nearly free words“: Flexionsformen bereits in der Liste enthaltener Worte
- überlegenswert (nicht von Fox): gewisse Wortarten (Substantive, Adjektive, substantivierte Verben) bevorzugen; alle anderen in allgemeine Stoppwortliste

Fox, C. (1989). A stop list for general text. ACM SIGIR Forum, 24(1-2), 19-35.

C.2 Wörter

2. Domänspezifische Stoppworte

- **spezielle Stoppworte in bestimmten fachlichen Zusammenhängen**

  nur solche Worte zulassen, die facheinschlägig sind; alle anderen in Stoppwortliste

3. Dokumentspezifische Stoppworte

- **beim Suchen nach bestpassenden Stellen innerhalb eines Dokuments**
- **manche Terme sind durchaus geeignet, einen Artikel als Ganzes zu finden, aber nicht, die beste Stelle zu markieren**
- **Methode: Terme, die im Dokument häufig auftreten *und* die im Text gleichmäßig verteilt sind, sind (für genau dieses Dokument) Stoppworte**

C.2 Wörter

Conflation (Verschmelzung von Wortformen)

Reduktionsformen

1. Wortstamm

Stemming

2. Grundform

Lemmatisierung

Ausgangswort: RETRIEVED

RETRIEV

RETRIEVAL

Wortstammbildung durch Abtrennen (oder Zufügen) von Buchstaben nach gewissen Regeln

Grundformbildung durch Vergleich mit Wörterbuch (ggf. alternativ: nach gewissen Regeln)

C.2 Wörter

Grundformbildung / Lemmatisierung

Variante 1: regelgeleitet. Der S-Lemmatisierer für das Englische

- **Wortform hat 3 Buchstaben oder weniger: Verfahren beenden**
- **Wortform endet auf IES (aber nicht EIES oder AIES): ersetze IES durch Y**
- **Wortform endet auf ES (aber nicht AES, EES oder OES): ersetze ES durch E**
- **Wortform endet auf S (aber nicht US oder SS): lösche S**

Harman, D. (1991). How effective is suffixing?
Journal of the American Society for Information Science, 42, 7-15.

C.2 Wörter

Grundformbildung / Lemmatisierung

Variante 2: wörterbuchbasiert

- **Voraussetzung: Lexikon der jeweiligen Sprache ist vorhanden**
- **Oberfläche: Wortform (Bsp.: BÜCHERN)**
- **Kategorisierung: Wortart und morphosyntaktische Eigenschaft (Bsp.: Substantiv – Dativ Plural)**
- **Grundform nachschlagen (Bsp.: BUCH)**

C.2 Wörter

Wortstammbildung / Stemming

Befreien der Wortformen von ihren Suffixen (nicht von den Präfixen).

Variante 1: Longest-Match-Stemmer (Lovins-Stemmer)

- **Abtrennen der jeweils längsten zutreffenden Endung (Voraussetzung: Liste aller Endungen)**
- **Folgen spezifischer Regeln zur Abtrennung**
- **Folgen von Regeln zur Re-Kodierung**

Lovins, J.B. (1968). Development of a stemming algorithm.
Mechanical Translation and Computational Linguistics, 11(1-2), 22-31.

C.2 Wörter

Longest-Match-Stemmer (Lovins-Stemmer)

Regeln (Bsp.): **B : Stamm muss mind. 3 Zeichen haben**
C : Stamm muss mind. 4 Zeichen haben
E : nach E nicht abschneiden

.11.	.09.	.09.—*Cont.*	.08.
alistically B	allically C	entiality A	ableness A
arizability A	antaneous A	entialize A	arizable A
izationally B	antiality A	entiation A	entation A
	arisation A	ionalness A	entially A
.10.	arization A	istically A	eousness A
antialness A	ationally B	itousness A	ibleness A
arisations A	ativeness A	izability A	icalness A
arizations A	eableness E	izational A	ionalism A
entialness A	entations A		ionality A

C.2 Wörter

Longest-Match-Stemmer (Lovins-Stemmer)

Re-Kodierung

Regeln (Bsp.): doppelte Konsonanten auf einen reduzieren

RPT durch RB ersetzen

Input	*Longest-Match-Stamm*	*re-kodierter Stamm*
metal	**metal**	**metal**
metallic	**metall**	**metal**
absorbing	**absorb**	**absorb**
absorption	**absorpt**	**absorb**

C.2 Wörter

Wortstammbildung (stemming)

Variante 2: iterativer Stemmer (Porter-Algorithmus)

Abk.: C	**Konsonant: alles außer A, E, I, O, U; Y nur dann, wenn nicht nach Konsonant (wie in Toy)**
V	**Vokal**
CCC, ...	**sei C**
VVV, ...	**sei V**
[C]VCVC...[V]	**(allgemeine Form)**
(VC){m}(V)	**Anzahl der VC = m in einem Wort**
Bsp.:	**m=0 : tree, by**
	m=1 : trouble, trees
	m=2 : troubles, private

Porter, M.E. (1980). An algorithm for suffix stripping. Program, 14(3), 130-137.

C.2 Wörter

Porter-Algorithmus

Regel: (Bedingung) S1 → S2: falls ein Wort mit dem Suffix S1 endet und der Stamm vor S1 die Bedingung erfüllt, dann wird S1 durch S2 ersetzt

die Bedingung wird durch m definiert; etwa: (m > 1)

Bsp.: (m>1) EMENT → _

S1 = EMENT; S2 = Null

REPLACEMENT → REPLAC

***S** der Stamm endet mit „S“

V der Stamm enthält einen Vokal

***d** der Stamm endet mit einem Doppelkonsonant (etwa: -TT, -SS)

and, or, not : Kombinationen von Bedingungen

bei mehreren Regeln in einem Schritt: nur eine anwenden, und zwar die mit dem „longest match“

C.2 Wörter

Porter-Algorithmus

Beispiel:

Schritt 1

(insgesamt 5 Iterationsrunden

```
Step 1a

    SSES -> SS                      caresses  ->  caress
    IES  -> I                       ponies    ->  poni
                                    ties      ->  ti
    SS   -> SS                      caress    ->  caress
    S    ->                         cats      ->  cat

Step 1b

    (m>0) EED -> EE                 feed      ->  feed
                                    agreed    ->  agree
    (*v*) ED  ->                    plastered ->  plaster
                                    bled      ->  bled
    (*v*) ING ->                    motoring  ->  motor
                                    sing      ->  sing
```

```
Step 1c

    (*v*) Y -> I                    happy     ->  happi
                                    sky       ->  sky
```

Kapitel C.3

Phrasen, Eigennamen, Komposita, semantisches Umfeld

C.3 Phrasen, Eigennamen, Komposita

Zusammengesetzte Ausdrücke

- **Phrasen: Begriffe, die aus mehr als einem Wort bestehen**
  - **Allgemeinbegriffe**
    - **häufig im Englischen: "soft ice", "high school"**
    - **seltener im Deutschen: "juristische Person", eher üblich: Komposita: "Wellensittichfutter", "Staubecken"**
    - **in vielen Fachsprachen: "Pfeiffersches Drüsenfieber", "Frankfurter Schule", "Braggsche Kurve"**

C.3 Phrasen, Eigennamen, Komposita

Zusammengesetzte Ausdrücke

- **Phrasen: Begriffe, die aus mehr als einem Wort bestehen**
  - **Namen (named entities)**
    - **Personennamen: "Miranda Otto"**
    - **Institutionen: "Henkel KGaA"**
    - **Orte: "Düsseldorf-Bilk"**
    - **Produkte: "HP Laserjet 1300"**

Namen müssen stets aus Conflation- und Übersetzungsroutine herausgehalten werden (sonst: statt "Julia Roberts" --> "Julia Robert" und "Heath Ledger" --> "Heide Kantholzträger")

C.3 Phrasen, Eigennamen, Komposita

Phrasenbildung

- **(1.) statistische Methode:**
  - **Wahrscheinlichkeit des gemeinsamen Auftretens der Phrasenbestandteile**
  - **Parameter:**
    - **Phrasenlänge (Anzahl der Bestandteile, z.B. 2)**
    - **Umgebung (Textgebiet, z.B. Satz)**
    - **Abstand (max. Anzahl von Worten, die zwischen den Phrasengliedern - unter Abzug der Stoppworte - liegen dürfen, z.B. 1)**
    - **Dokumenthäufigkeit der Bestandteile (Anzahl der Dokumente, in denen das Wort min. einmal vorkommt); Definition eines Schwellenwertes, ab dem ein Wort erster Teil einer Phrase werden darf**
    - **Dokumenthäufigkeit der Phrase (Anzahl der Dokumente, in den die Phrase vorkommt); Definition von oberem und unterem Schwellenwert**

Fagan, J.L. (1987). Automatic phrase indexing for document retrieval: An examination of syntactic and non-syntactic methods. In Proceedings of the 10th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 91-101). New York, NY: ACM.

C.3 Phrasen, Eigennamen, Komposita

Phrasenbildung

- **(2.) Methode der "Verklumpung" von Textteilen:**
  - **massive Erweiterung der Stoppwortliste**
  - **Interpretation der Stoppworte als Textseparatoren**
  - **Beispieltext:**

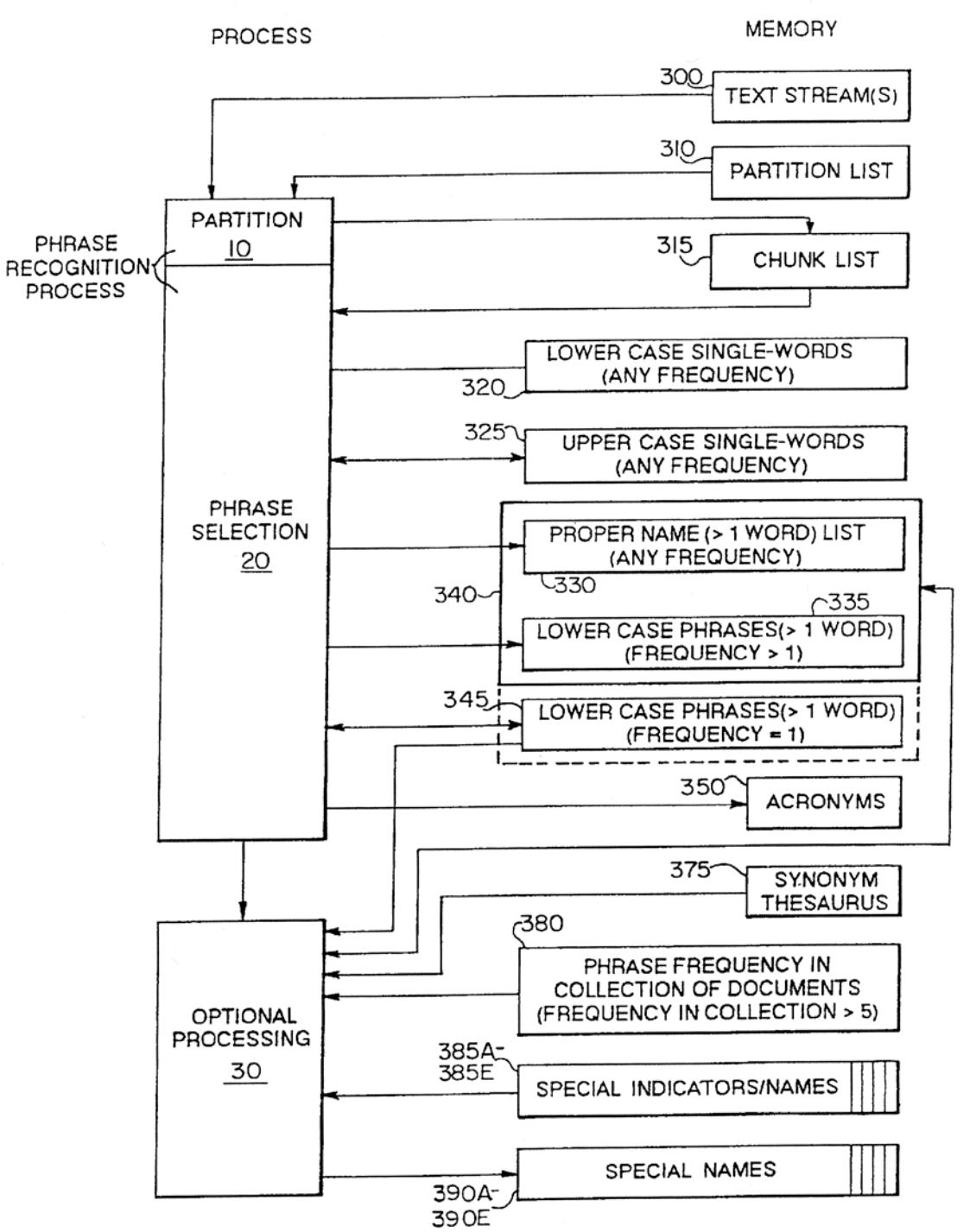
    Citing what is **called** newly **conciliatory comments** by the **leader** of the **Irish Republican Army**'s **political wing,** the **Clinton Administration** announced **today** that it would issue him a **visa** to attend a **conference** on **Northern Ireland** in **Manhattan** on **Tuesday.** The **Administration** had been **leaning** against **issuing** the **visa** to the **official**, **Gerry Adams**, the **head** of **Sinn Fein**, **leaving** the **White House** caught between the **British Government** and a **powerful bloc** of **Irish-American legislators** who favored the **visa.**

Lu, X.A., Miller, D.J., & Wassum, J.R. (1996). Phrase recognition method and apparatus. Patent-Nr. US 5.819.260. Patentinhaber: Lexis-Nexis. Priorität: 6.10.1998.

C.3 Phrasen, ...

Phrasenbildung

Textklumpenmethode



C.3 Phrasen, Eigennamen, Komposita

Phrasenbildung

- **(3.) Nutzung von Phrasenwörterbüchern**
  - **Vergleich von Wortfolgen im Text mit Wörterbucheinträgen**
  - **Voraussetzung: Wörterbuch ist vorhanden**
  - **günstig: sukzessiver Aufbau des Wörterbuchs durch die Methoden (1.) und (2.)**

C.3 Phrasen, Eigennamen, Komposita

Eigennamenerkennung

- **Normdatei (z.B. Personennamendatei der deutschen Bibliotheken)**
- **intellektuell aufgebaut und gepflegt**
- **Homonyme getrennt - Synonyme zusammengefügt**

NORMDATEN: *Personenname (11862444X)*

Tucholsky, Kurt
Lebensdaten: 1890-1935
Beruf/Funktion: Dt. Schriftsteller, Bühnenautor, Dichter und Journalist (Literatur- und Theaterkritiker); Emigration 1924 nach Frankreich, 1929 nach Schweden
Verweisungsformen: Grotius, Hugo [Pseud.]
Hauser, Kaspar [Pseud.] (Schriftsteller)
Panter, Peter [Pseud.]
Tiger, Theobald [Pseud.]
Wrobel, Ignaz [Pseud.]
Tucholsky, ...
Tuchol'skij, Kurt
Tûkôlsqî, Qûrt
Tukôlsqî, Qûrt
Bünzly, Paulus [Pseud.]
Körner, Theobald [Pseud.]
Old Shatterhand [Pseud.]

C.3 Phrasen, Eigennamen, Komposita

Eigennamenerkennung

- innere Namenscharakteristika
  - Namensbestandteile beginnen mit Großbuchstaben
  - (im Englischen): Mr., Mrs. bei Personennamen
  - Indikatorworte
    - Vornamen bei Personen
      - bei erkannter Phrase: Bestandteil vor Indikator löschen
    - Inc., GmbH usw. bei Unternehmen
      - bei erkannter Phrase: Bestandteil hinter Indikator löschen
- äußere Namenscharakteristika
  - wichtig insb. bei homonymen Namen
  - Personennamen mit "Typen", z.B. Künstler, Politiker
  - Definition typenspezifischer Worte, z.B. star or actor bei Künstlern

C.3 Phrasen, Eigennamen, Komposita

Kompositazerlegung

- **Fugenbildung im Deutschen unregelmäßig**
  - **Komposition aus Singular- und aus Pluralform (*Bsp.:* „Haus-meister“ – „Häuser-meer“)**
  - **Komposition mit und ohne Fugen-S (*Bsp.:* „Schwein-s-blase“ – „Schwein-e-bauch“ – „Schwein-kram“)**
  - **Verkürzung von Bestandteilen (*Bsp.*: "Schwimmverein" aus "Schwimm*en*" und "Verein")**
- **Zerlegung**
  - **Voraussetzung: Wörterbuch mit Grundformen nebst Flexionen und ggf. Zerlegungsverboten (Verb-rechen; Staat-sex-amen)**
  - **buchstabenweise Abtrennung von links oder rechts bis zum Longest Match**
  - **Problem: „Staubecken“ führt zu „Stau-becken“ oder „Staub-ecken“, je nachdem, ob die Analyse links oder rechts beginnt**

C.3 Phrasen, Eigennamen, Komposita

Kompositazerlegung

- **„Staubecken“ (von rechts nach links)**
- **Staubecke-n** — **kein Eintrag vorhanden**
- **Staubeck-en** — **kein Eintrag vorhanden**
- **Staubec-ken** — **kein Eintrag vorhanden**
- **Staube-cken** — **"staube": Imperativ von "Stauben"; "cken": kein Eintrag vorhanden**
- **Staub-ecken** — **Eintrag „Staub“ vorhanden; "ecken": Nominativ Plural von "Ecke"**
- **Stau-b** — **"Stau"; Grundform; "b": kein Entrag**
- **Sta-ub** — **kein Eintrag vorhanden**
- **St-aub** — **kein Eintrag vorhanden**
- **S-taub** — **kein Eintrag vorhanden**
- **ERGEBNIS: Staub - Ecke**

C.3 Phrasen, Eigennamen, Komposita

Kompositazerlegung

- **„Staubecken“ (von links nach rechts)**
- **S-taubecken** — **kein Eintrag vorhanden**
- **St-aubecken** — **kein Eintrag vorhanden**
- **Sta-ubecken** — **kein Eintrag vorhanden**
- **Stau-becken** — **Eintrag „Becken“ vorhanden**
  Eintrag „Stau“ vorhanden
- **b-ecken** — **kein Eintrag vorhanden**
- **be-cken** — **kein Eintrag vorhanden**
- **...** — **kein Eintrag vorhanden**
- **ERGEBNIS: Stau - Becken**

- **Stau-Becken oder Staub-Ecken? Kontext statistisch auswerten!**

C.3 Phrasen - Eigennamen - Komposita

Semantisches Umfeld

- **"Find what I mean, not what I say" (Susan Feldman)**
- **natürlichsprachiges Umfeld**

  Werkzeug: natürlichsprachiger Thesaurus

  Beispiel: WordNet
- **fachsprachliches Umfeld**

  Werkzeug: fachsprachlicher Thesaurus

  Beispiel: Standard-Thesaurus Wirtschaft
- **Begrifforientierte IR-Systeme:**
  - **Klärung der Homonyme und Synonyme**
  - **Einbeziehen des semantischen Umfelds in die Suchanfrage (z.B. hierarchisches Retrieval, Suchanfrageerweiterung)**

C.3 Phrasen - Eigennamen - Komposita

Kleinste semantische Einheiten

- **Synsets (natürlichsprachiger Thesaurus)**
- **Beispiel: WordNet**

Enter a word to search for: car Search WordNet

KEY: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Noun

- S: (n) car, auto, automobile, machine, motorcar (a motor vehicle with four wheels; usually propelled by an internal combustion engine) *"he needs a car to get to work"*
- S: (n) car, railcar, railway car, railroad car (a wheeled vehicle adapted to the rails of railroad) *"three cars had jumped the rails"*
- S: (n) cable car, car (a conveyance for passengers or freight on a cable railway) *"they took a cable car to the top of the mountain"*
- S: (n) car, gondola (the compartment that is suspended from an airship and that carries personnel and the cargo and the power plant)
- S: (n) car, elevator car (where passengers ride up and down) *"the car was on the top floor"*

C.3 Phrasen - Eigennamen - Komposita

Kleinste semantische Einheiten

- **Deskriptoren (fachsprachiger Thesaurus)**
- **Beispiel: Standard-Thesaurus Wirtschaft**

Such-Anfrage	1. Satz von 1 aus Liste zu **Schlagworte aus Suchanfrage** ‹Titel u. Themen = Zahnpasta› Schlagwortsatz	merken

		Oberbegriffe	Verwandte Begriffe
Benennung	**Mundpflegemittel**		
englischer Term	Oral hygiene product	Körperpflegemittel	Keine
Klassifikation	P.12.05 Waschmittel und Körperpflegemittel		
benutze für	Gebißreiniger; Gebissreiniger; Mundwasser; Zahnpasta; Zahnpulver	Unterbegriffe	
		Keine	
Identnummern	5665 TIN: 14736-2		
zum Schlagwort	29 Titel		

C.3 Phrasen - Eigennamen - Komposita

WordNet

- **Voraussetzung: morphologische Analyse erfolgreich abgeschlossen**
- **Wortklassen**
  - **Substantiv**
  - **Verb**
  - **Adjektiv**
  - **Adverb**
  - **(Funktionsworte)**
- **klassenspezifische Zuordnung von Relationen zu den Worten**
- grundgelegt von George A. Miller; weitergeführt von Christiane Fellbaum

Fellbaum, C., Ed. (1998). WordNet – An Electronic Lexical Database. Cambridge, MA; London: MIT Press.

C.3 Phrasen - Eigennamen - Komposita

Wort-Begriff-Matrix

Word Meanings	Word Forms				
	F_1	F_2	F_3	. . .	F_n
M_1	$E_{1,1}$	$E_{1,2}$			
M_2		$E_{2,2}$			
M_3			$E_{3,3}$		
⋮				. . .	
M_m					$E_{m,n}$

George A. Miller

Erkennen von Synonymen (hier: $E_{1,1}$ und $E_{1,2}$)

Erkennen von Homonymen (hier: $E_{1,2}$ und $E_{2,2}$)

C.3 Phrasen - Eigennamen - Komposita

Relationen von Substantiven

- **Hyponymie (Abstraktionsrelation; Hyponym - Hypernym): „kind of ...“**
  - **transitiv, asymmetrisch**
- **Meronymie (Teil-Ganzes-Relation; Meronym - Holonym): „part of ...“**
  - **asymmetrisch**
  - **Problem: Meronymie ist z.T. transitiv oder z.T. nicht transitiv. [Die Rinde ist Teil des Baumes. Der Baum ist Teil des Waldes. Kann man schließen: Die Rinde ist Teil des Waldes?]**
- **Antonymie (Gegensatz)**
  - **symmetrisch**
- **Ähnlichkeit (symmetrisch; nicht transitiv)**

Miller, G.A. (1998). Nouns in WordNet. In C. Fellbaum (Ed.),
WordNet. An Electronic Lexical Database (pp. 23-46). Cambridge, MA; London: MIT Press.

C.3 Phrasen - Eigennamen - Komposita

Relationen von Verben

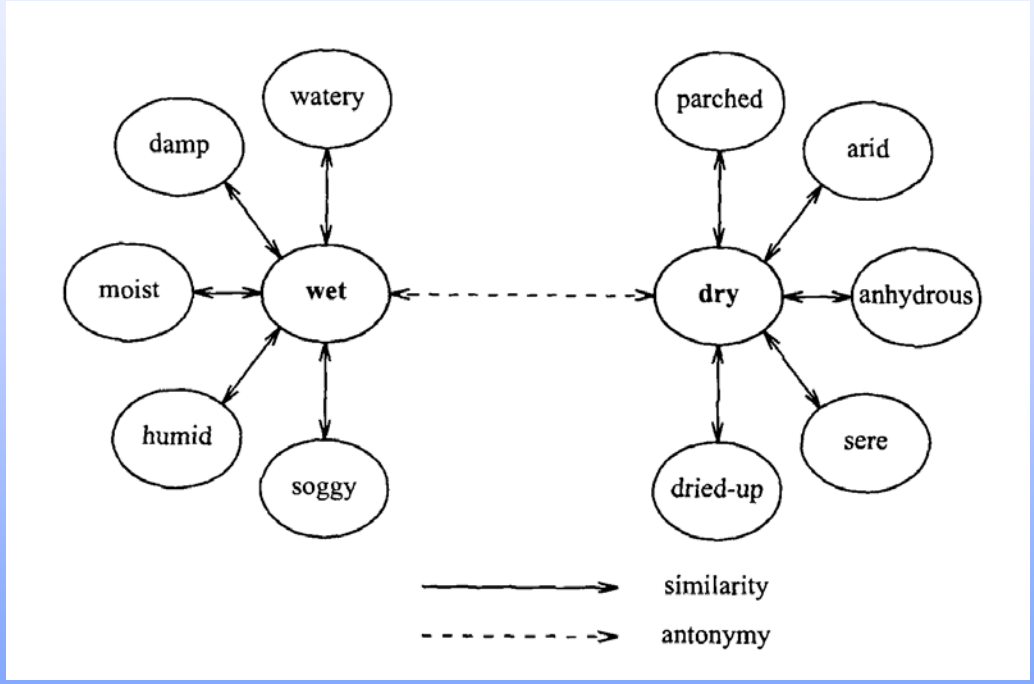
- **Verb-Hyponymie**
  - **+ Temporal Inclusion (Enthaltensein bei Gleichzeitigkeit)**
    - **+ Troponymie („echte“ Abstraktion: talk – lisp)**
    - **- Troponymie (abhängige Tätigkeit: sleep – snore)**
  - **- Temporal Inclusion**
    - **Backward Presupposition (try – succeed; auch: Gegensätze: untie – tie)**
    - **Cause (have – give; show – see)**

Fellbaum, C. (1998). A semantic network of English verbs.
In C. Fellbaum (Ed.), WordNet. An Electronic Lexical Database (pp. 69-104).
Cambridge, MA; London: MIT Press.

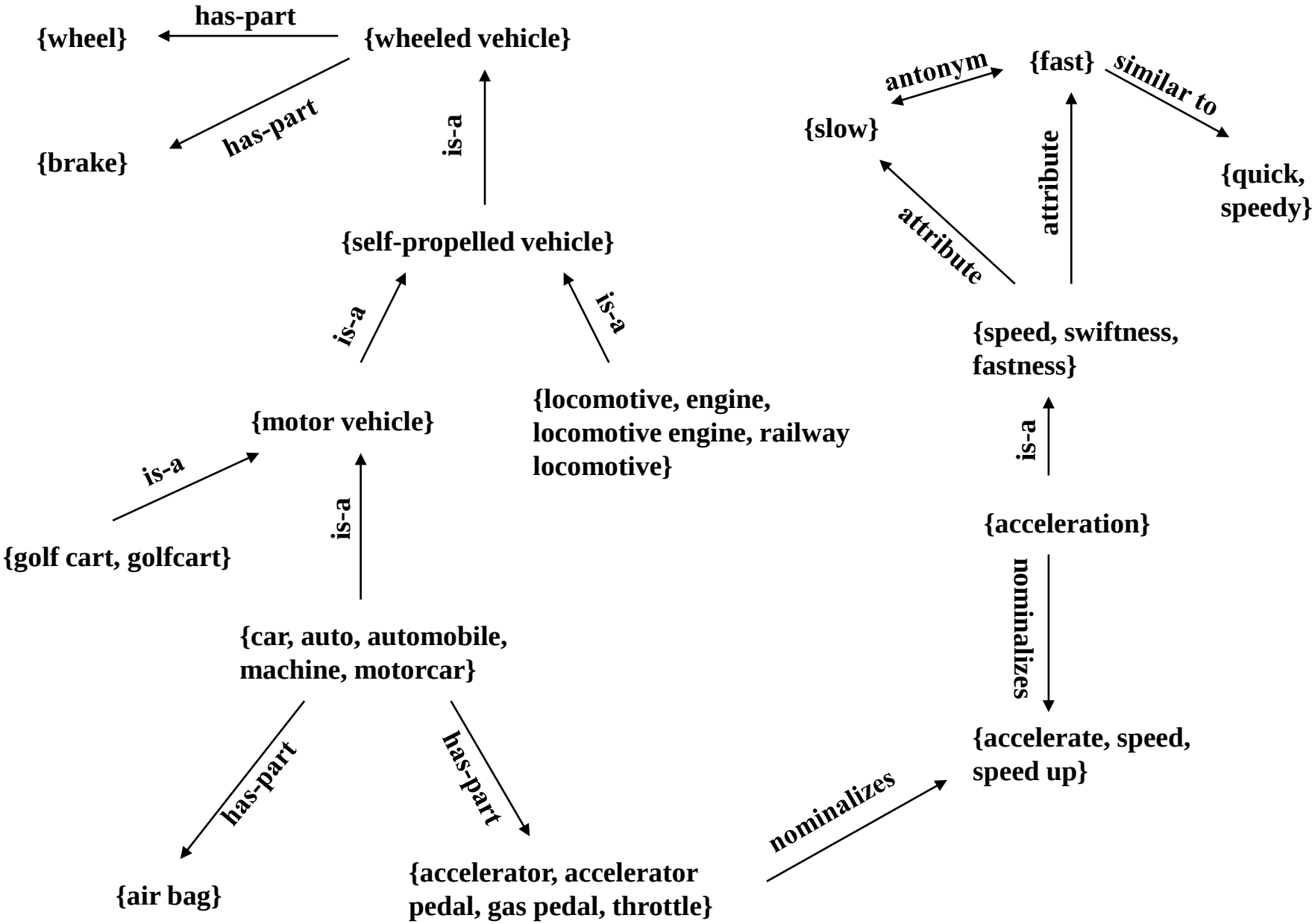
C.3 Phrasen - Eigennamen - Komposita

Relationen von Adjektiven

- Ähnlichkeit
- abgeleitet von ...
- Antonymie
  - kontradiktorisch (tot - lebendig)
  - konträr (Abstufungen: fett - normalgewichtig - schlank - dünn)
  - im Retrieval insb. bei kontradiktorischen Antonymen mit "nicht" oder "un-" arbeiten



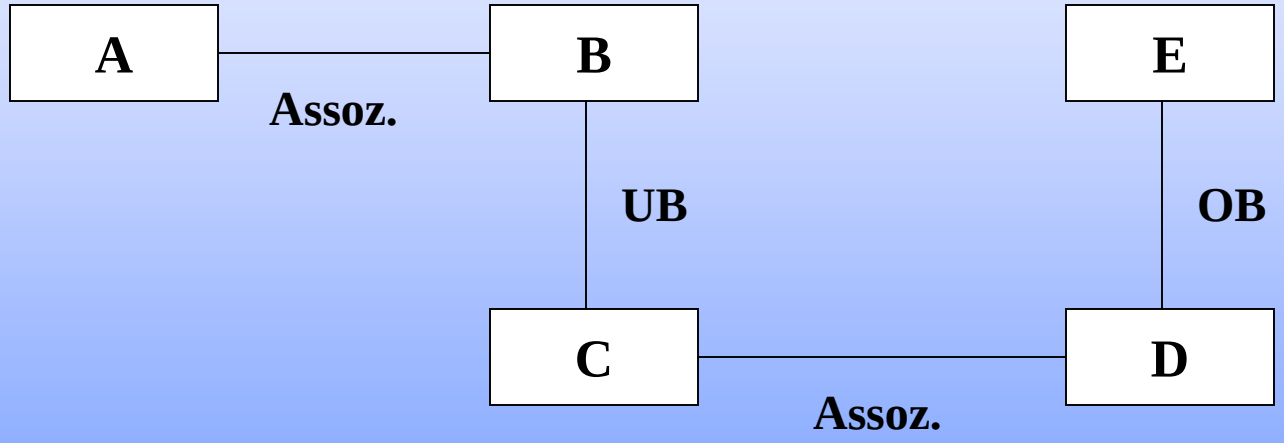
Miller, K.J. (1998). Modifiers in WordNet. In C. Fellbaum (Ed.),
WordNet. An Electronic Lexical Database (pp. 47-67). Cambridge, MA; London: MIT Press.


{wheel}
has-part
{wheeled vehicle}
{brake}
has-part
is-a
{self-propelled vehicle}
is-a
is-a
{motor vehicle}
{locomotive, engine, locomotive engine, railway locomotive}
is-a
is-a
{golf cart, golfcart}
{car, auto, automobile, machine, motorcar}
has-part
has-part
{air bag}
{accelerator, accelerator pedal, gas pedal, throttle}
nominalizes
{accelerate, speed, speed up}
nominalizes
{acceleration}
is-a
{speed, swiftness, fastness}
attribute
attribute
{slow}
antonym
{fast}
similar to
{quick, speedy}

C.3 Phrasen - Eigennamen - Komposita

Semantische Ähnlichkeit zweier Begriffe

- **(1.) Anzahl der Pfade auf dem kürzesten Weg zwischen den Begriffen im Thesaurus**

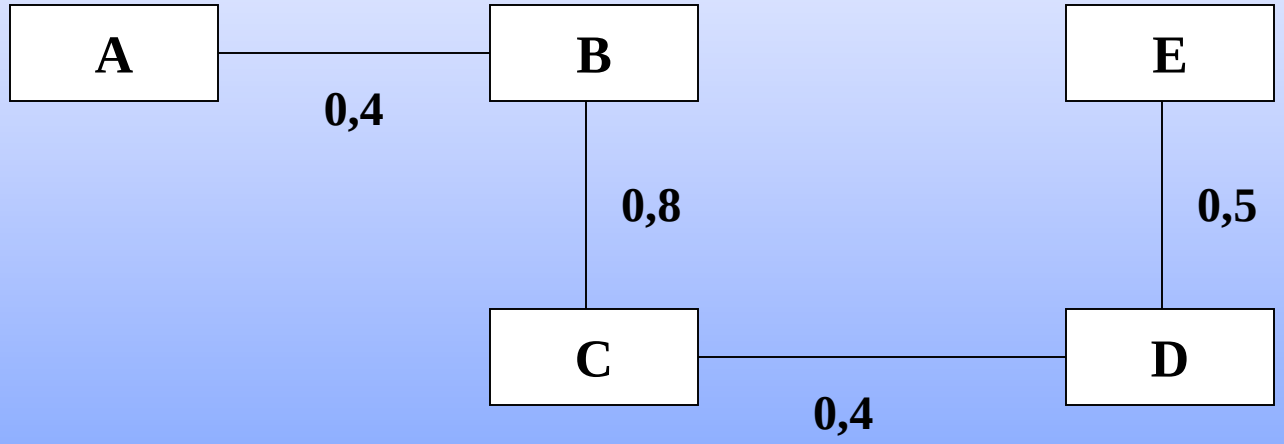


- **semantischer Abstand (A, E) = 4**

C.3 Phrasen - Eigennamen - Komposita

Semantische Ähnlichkeit zweier Begriffe

- **(2.) gewichtete Relationen (etwa: Synonymie: 1,0; Unterbegriff: 0,8; Oberbegriff: 0,5; assoz. Begriff: 0,4)**



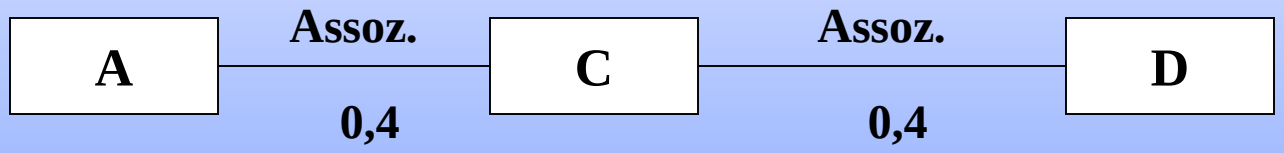
- **semantischer Abstand (A, E) = 1/0,4 + 1/0,8 + 1/0,4 + 1/0,5 = 8,25**

Vogel, C. (2002). Quality Metrics for Taxonomies. Vienna, VA: Convera.

C.3 Phrasen - Eigennamen - Komposita

Semantische Ähnlichkeit zweier Begriffe

- **bei nicht-transitiven Relationen darf stets nur mit einem Pfad gerechnet werden**



- **semantische Ähnlichkeit (A, D) = 0**

C.3 Phrasen - Eigennamen - Komposita

Disambiguierung homonymer Worte

- **Schritt 1: Nutzer gibt Suchatom ein, zu dem mehrere Begriffe vorliegen**
  - **entweder: Auswahlliste vorlegen**
  - **oder: bei mehreren Suchatomen ("Java" "Perl") semantischen Abstand bestimmen; bei nur einem Suchatom: Probleme, wenn keine (Nutzer-)Daten vorliegen**
- **Schritt 2: nach Klärung der Homonymie: die richtigen Dokumentationseinheiten finden**
  - **bei Systemen mit Knowledge Organization Systems: trivial**
  - **bei automatischer Indexierung: in einem Textfenster um das Suchatom herum im Dokument weitere Worte gewinnen; semantischen Abstand zum passenden Synset/Deskriptor errechnen**

Kapitel C.4

Anaphora

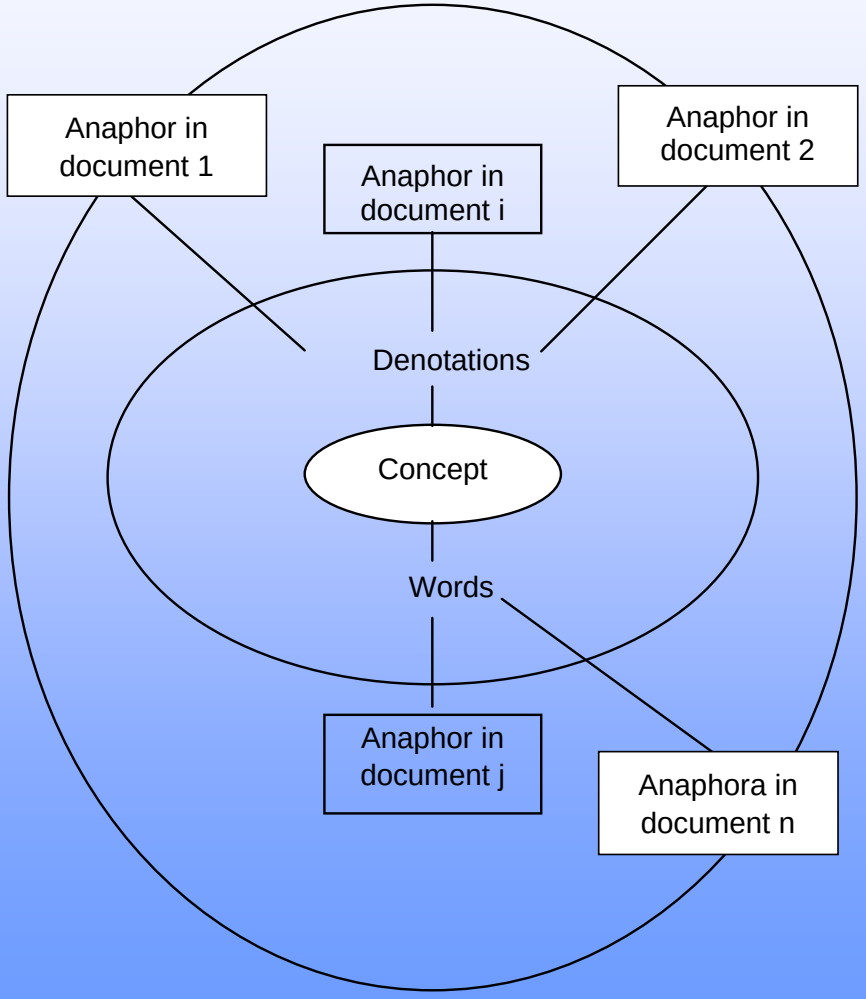
C.4 Anaphora

- **Anaphora beziehen sich auf frühere Textelemente (Referenten/Antecedenten) und haben deren Bedeutung**
- **im gleichen Satz oder auch über Satzgrenzen hinaus**
- **Wortklassen:**
  - **Pronomen (vor allem Personalpronomen: er, sie, es, wir, ...)**
  - **Nomen (Ersetzung, Abkürzung, Metapher, ...)**
  - **Zahlwörter**
  - **Asyndeton (Auslassung von Nomen und Pronomen: er kam, sah, siegte)**

C.4 Anaphora

- **Ellipsen**
  - **Referenzausdruck wird ausgelassen ("Egon hackte Holz, genau wie Rolf [Holz hackte]")**
- **Warum für IR wichtig?**
  - **Textstatistik**
  - **Abstandsoperatoren**

C.4 Anaphora


Anaphor in document 1
Anaphor in document i
Anaphor in document 2
Denotations
Concept
Words
Anaphor in document j
Anaphora in document n

C.4 Anaphora

***Bsp.*:**

(1) Max versteckt Moritz' Autoschlüssel.
Er war betrunken.

(2) Max versteckt Moritz' Autoschlüssel.
Er spielte ihm einen Streich.

Wer ist "er"?

C.4 Anaphora

- **Anaphora – Ellipsen beim Einsatz von Abstandsoperatoren**
  - **Recall-Zugewinn bei Eigennamen durch Anaphoraauflösung:**
    - **Ellipsen (Satzebene): 38,2%**
    - **Anaphora (Satzebene): 17,6%**
  - **Personen- vs. Organisationsnamen; Zugewinn:**
    - **Personennamen (Anaphora; Satzebene): 23,3%**
    - **Institutionen (Anaphora; Satzebene): 14,8%**
  - **die häufigsten Formen anaphorischer Namensnennungen:**
    - **Personalpronomen**
    - **Auslassungen vom kompletten Namen**

Pirkola, A., & Järvelin, K. (1996). The effect of anaphora and ellipsis resolution on proximity searching in a text database. Information Processing & Management, 32, 199-216.

C.4 Anaphora

- **Anaphora – Ellipsen bei Textstatistik**
  - **zentrale Themen eines Dokuments werden häufig anaphorisch beschrieben**
  - **Auftretenshäufigkeit der Referenzformulierung bei zentralen Themen eher selten (Bsp.: "Immanuel Kant"); entsprechend: geringe Gewichtung**
  - **Auftretenshäufigkeit ihrer Referenzausdrücke hoch (Bsp.: "unser Königsberger Philosoph", "Mentor des Kritizismus", "er")**
  - **erst die Anaphoraauflösung führt zu einer zutreffenden Gewichtung**

DuRoss Liddy, E. (1990). Anaphora in natural language processing and information retrieval. Information Processing & Management, 26, 39-52.

C.4 Anaphora

Anaphora-Resolution. Beispiel: MARS („Mitkov's Anaphora Resolution System")

- Schritt 1: Wörter identifizieren (einschließlich Geschlecht und Numerus)
- Schritt 2: Identifikation von Anaphora; Ausschluss von pleonastischem Gebrauch von Pronomen (wie „es regnet")
- Schritt 3: mögliche Antecedenzen des anaphorischen Ausdrucks sammeln (Nomen mit gleichem Geschlecht und Numerus)
- Schritt 4: Score für alle möglichen Antecedenzen berechnen
- Schritt 5: Antecedenz mit höchstem Score auswählen

Mitkov R. (2002). Anaphora Resolution. London: Longman.

Kapitel C.5

Fehlertolerantes Retrieval

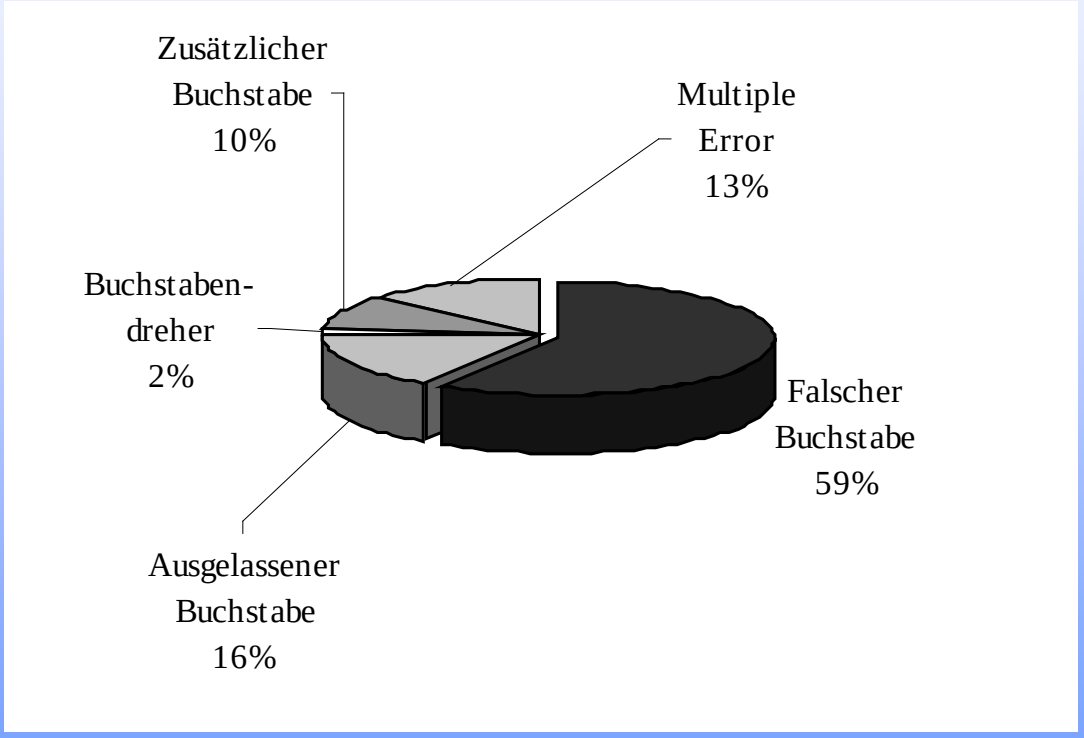
C.5 Fehlertolerantes Retrieval

- **Eingabefehler**
  - **in den Dokumenten**
  - **in den Suchanfragen**
- **Formen**
  - **Leerzeichenfehler ("...ofthe..."; "th_ebook")**
  - **Fehler an Worten, die isoliert erkannt werden**
    - **typographische Fehler ("teh" statt "the")**
    - **orthographische Fehler ("recieve" statt "receive")**
    - **phonetische Fehler ("4u" statt "for you")**
  - **Fehler an Worten, die erst im Kontext erkannt werden**
    - **syntaktische Fehler ("the study was conducted be XY")**
    - **semantische Fehler ("they are leaving in about 15 minuets to go ...")**

Kukich K. (1992). Techniques for automatically correcting words in texts.
ACM Computing Surveys, 24, 377-439.

C.5 Fehlertolerantes Retrieval

Fehler an Worten, die isoliert erkannt werden



Damerau, F.J. (1964). A technique for computer detection and correction of spelling errors. Communications of the ACM, 7, 171-176.

C.5 Fehlertolerantes Retrieval

Fehlertolerantes Retrieval. Ansatz 1: Phonetik
(1a) Der Soundex-Algorithmus

- **Verschmelzung von Wortformen anhand ihres Klanges**
- **Vorgehen:**
  - **erster Buchstabe bleibt erhalten**
  - **Vokale a, e, i, o, u, y** **1**
  - **labiale und labiodentale Laute b, f, p, v** **2**
  - **Kehl- und Zischlaute c, g (übergehen: gh), k ,q, x, s, z (ohne Schluss-s und -z)** **3**
  - **Dentallaute d, t** **4**
  - **palataler Reibelaut l** **5**
  - **labionasaler Laut: m** **6**
  - **dento- oder linguanasaler Laut n** **7**
  - **dentaler Reibelaut r** **8**

Russell, R.C. (1917). Index. Patent-Nr. US 1.261.167. Priorität: 25.10.1917.

C.5 Fehlertolerantes Retrieval

Fehlertolerantes Retrieval. Ansatz 1: Phonetik
Der Soundex-Algorithmus

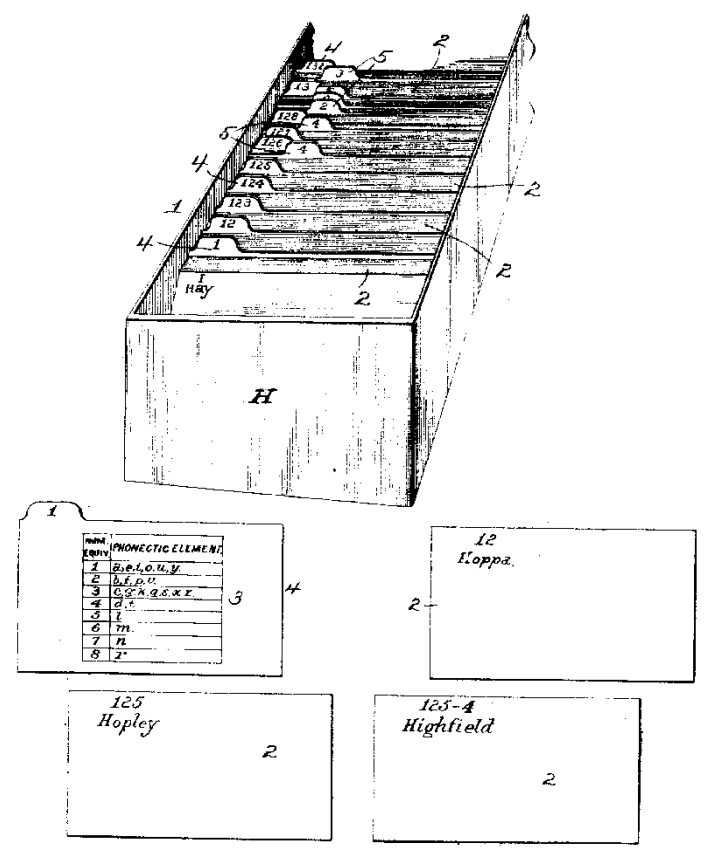
- **Regeln:**
  - **aufeinander folgende Buchstaben derselben Lautklasse: nur den ersten berücksichtigen ("Ball" wird zu "Bal")**
  - **mehrere Vokale im Wort: nur den ersten berücksichtigen ("Carter" wird zu "Catr")**

- **Heutiger Stand:**
  - **H ist Vokal**
  - **m und n: nur eine Klasse**

Jacobs, J.R. (1982). Finding words that sound alike. The SOUNDEX algorithm. Byte, 7, 473-474.

C.5 Fehlertolerantes Retrieval

Fehlertolerantes Retrieval. Ansatz 1: Phonetik
Der Soundex-Algorithmus



Hoppa

H	**oppa**
H	**opa (doppelte Belegung)**
H	**op (nur 1 Vokal)**
H	**12**

Highfield

H	**ighfield**
H	**ifield (gh wird übergangen)**
H	**ifld (nur 1 Vokal)**
H	**1254**

C.5 Fehlertolerantes Retrieval

Fehlertolerantes Retrieval. Ansatz 1: Phonetik (1b): Phonix

- **Verfeinerungen von Soundex**
  - **phonetische Regeln werden auch auf den 1. Buchstaben angewandt (night - knight)**
  - **"phonetische Ersetzung":**
    - **gleiche Buchstabenfolgen (z.B. "ough") klingen in unterschiedlichen Worten unterschiedlich ("plough" - "cough")**
    - **Regeln beziehen sich auf die Stellung der Zeichenfolge im Wort: am Anfang (z.B. "kn" zu "n"), in der Mitte und am Ende (dort "kn" nicht ändern)**

Gadd, T.N. (1988). 'Fisching fore werds': Phonetic retrieval of written text in information systems. Program, 22, 222-237.
Gadd, T.N. (1990). PHONIX: The algorithm. Program, 24, 363-366.

18. Fehlertoleranz

phonetische Ersetzung

(SUB replaces letters at START, MIDDLE or END of word)

SUB	START	MIDDLE	END
G	DG	DG	DG
KO	CO	CO	CO
KA	CA	CA	CA
KU	CU	CU	CU
SI	CY	CY	CY
SI	CI	CI	CI
SE	CE	CE	CE
KL	CL if CLv		
K	CK	CK	CK
K			GC
K			JC
KR	CHR if CHRv		
KR	CR if CRv		
R	WR		
NK	NC	NC	NC
KT	CT	CT	CT
F	PH	PH	PH
AR	AA	AA	AA
SH	SCH	SCH	SCH
TL	BTL	BTL	BTL
T	GHT	GHT	GHT
ARF	AUGH	AUGH	AUGH
LB	LJ if vLJv		
LOW	LOUGH	LOUGH	LOUGH
KW	Q		
N	KN		
N			GN
N	GHN	GHN	GHN
N			GNE
NE	GHNE	GHNE	GHNE
NS			GNES
N	GN		
N		GN if GNc	GN if GNc
S	PS		
T	PT		
C	CZ		
Z		WZ if vWZ	
CH		CZ	
LSH	LZ	LZ	LZ
RSH	RZ	RZ	RZ
S		Z if Zv	
TS	ZZ	ZZ	ZZ
TS		Z if cZ	
REW	HROUG	HROUG	HROUG
OF	OUGH	OUGH	OUGH
KW		Q if vQv	
Y		J if vJv	
Y	YJ if YJv		
G	GH		
E			GH if vGH

Where v = vowel and c = consonant

C.5 Fehlertolerantes Retrieval

Fehlertolerantes Retrieval. Ansatz 2: Damerau-Methode

- **benötigt Wörterbuch**
- **Schritt 1: Fehleridentifikation (Vergleich: Wort - Wörterbuch)**
- **Schritt 2: Identifikation des Fehlertyps (die Damerau-Methode bearbeitet nur Einzelfehler, keine multiple errors)**
- **Schritt 3: Fehlerkorrektur**

Damerau, F.J. (1964). A technique for computer detection and correction of spelling errors. Communications of the ACM, 7, 171-176.

C.5 Fehlertolerantes Retrieval

Fehlerkorrektur nach Damerau-Methode

FALSCHER BUCHSTABE:

```
                12345678
Eingabe:        ALPHIBET
Lexikon:        ALPHABET
                einzige Differenz bei Stelle 5
Ergebnis        Korrigiere Alphibet zu Alphabet!
```

C.5 Fehlertolerantes Retrieval

Fehlerkorrektur nach Damerau-Methode

BUCHSTABENDREHER:

```
                12345678
Eingabe:        ALHPABET
Lexikon:        ALPHABET
                Differenzen bei Stellen 3 und 4. HP
                in der Eingabe entspricht
                umgekehrter Reihenfolge PH im Lexikon
Ergebnis        Korrigiere Alhpabet zu Alphabet!
```

C.5 Fehlertolerantes Retrieval

Fehlerkorrektur nach Damerau-Methode

ZUSÄTZLICHER BUCHSTABE

```
                123456789
Eingabe:        ALLPHABET
Lexikon:        ALPHABET
                Erste Differenz bei Stelle 3 –
                Löschen des L bei Eingabe
Ergebnis        ALPHABET Übereinstimmung mit Lexikon:
                Korrigiere Allphabet zu Alphabet!
```

C.5 Fehlertolerantes Retrieval

Fehlerkorrektur nach Damerau-Methode

AUSGELASSENER BUCHSTABE

```
                12345678
Eingabe:        ALPABET
Lexikon:        ALPHABET
                Erste Differenz bei Stelle 4 –
                Löschen des H bei Lexikon
Ergebnis        ALPABET Übereinstimmung mit Eingabe:
                Korrigiere Alpabet zu Alphabet!
```

C.5 Fehlertolerantes Retrieval

Fehlertolerantes Retrieval. Ansatz 3: Levenshtein-Distanz

- **zählt die Edierschritte zwischen zwei Wörtern**
- **„Edieren“:**
  - **Buchstabe löschen**
  - **Buchstabe einfügen**
  - **Buchstabe vertauschen**
- **Levenshtein-Distanz: minimale Anzahl der Edierschritte, um zwei Wörter gleich zu machen**

Levenshtein, V.I. (1966). Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics – Doklady, 10(8), 707-710.

C.5 Fehlertolerantes Retrieval

Levenshtein-Distanz

Distanz „surgery“ – „survey“ = 2

		s	**u**	**r**	**g**	**e**	**r**	**y**
	0	1	2	3	4	5	6	7
s	1	**0**	1	2	3	4	5	6
u	2	1	**0**	1	2	3	4	5
r	3	2	1	**0**	1	2	3	4
v	4	3	2	1	**1**	2	3	4
e	5	4	3	2	2	**1**	**2**	3
y	6	5	4	3	3	2	2	**2**

C.5 Fehlertolerantes Retrieval

Fehlertolerantes Retrieval. Ansatz 4: *n*-Gramme

- **benötigt Wörterbuch; Zerlegung der Lexeme in n-Gramme**
- **Schritt 1: Fehleridentifikation (wenn Wort ein n-Gramm enthält, das nicht im Wörterbuch vorkommt)**
- **Schritt 2: Fehlerkorrektur (Ähnlichkeit nach Dice) (m, m': Anzahl der Buchstaben):**
  - **# n-Gramme des Wortes: m' + n - 1**
  - **# n-Gramme des Lexems: m + n - 1**
  - **# gemeinsamer n-Gramme: g**
  - **Ähnlichkeit(Wort-Lexem) = 2g / (m + n -1 + m' + n - 1)**

Angell, R.C., Freund, G.E., & Willett, P. (1983). Automatic spelling correction using a trigram similarity measure. Information Processing & Management, 19, 255-261.

C.5 Fehlertolerantes Retrieval

Eingabewort: CONSUMMING; Lexem: CONSUMING; N=3

CONSUMMING hat zehn Buchstaben und wird demnach durch zwölf Trigramme (m'+n-1 = 12) ausgedrückt:

****C, *CO, CON, ONS, NSU, SUM, UMM, MMI, MIN, ING, NG*, G**.**

Die Zerlegung von CONSUMING ergibt elf Trigramme (m+n-1 = 11):

****C, *CO, CON, ONS, NSU, SUM, UMI, MIN, ING, NG*, G**.**

Gemeinsam haben die beiden Zeichenketten zehn Trigramme (g = 10):

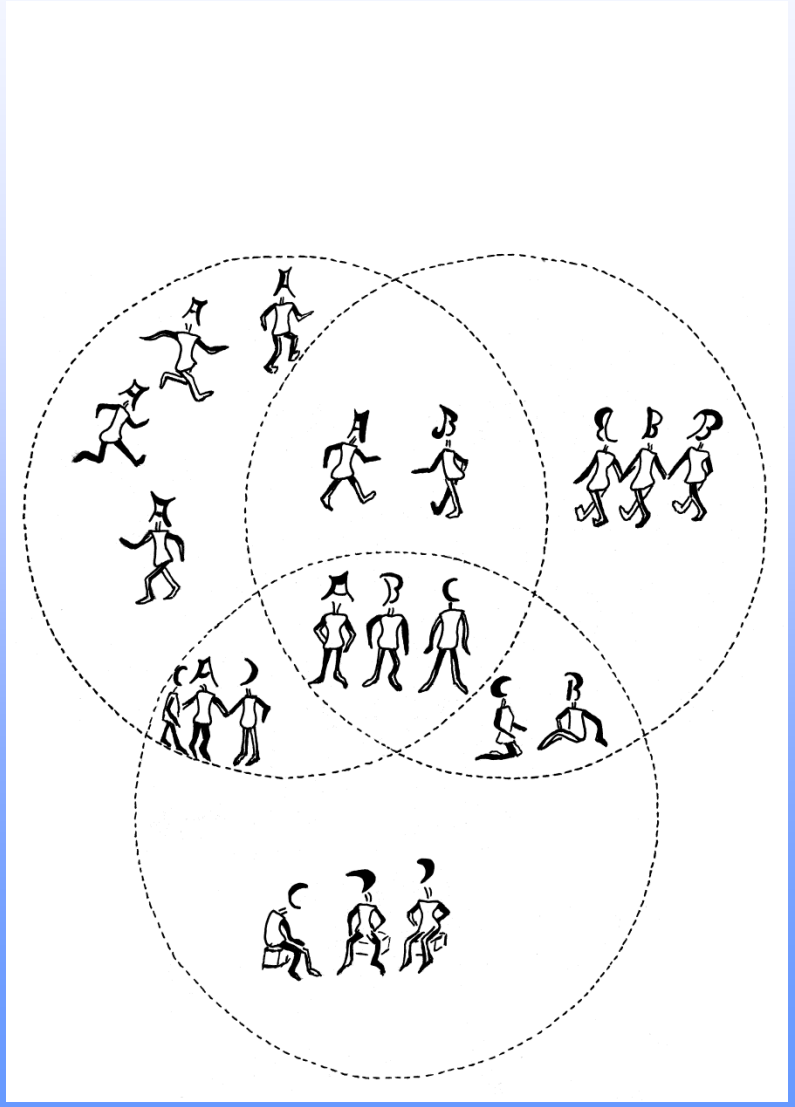
****C, *CO, CON, ONS, NSU, SUM, MIN, ING, NG*, G**.**

Die Ähnlichkeit von CONSUMMING und CONSUMING beträgt also:

2 * 10 / (12 + 11)

= 20 / 23 = 0,87.

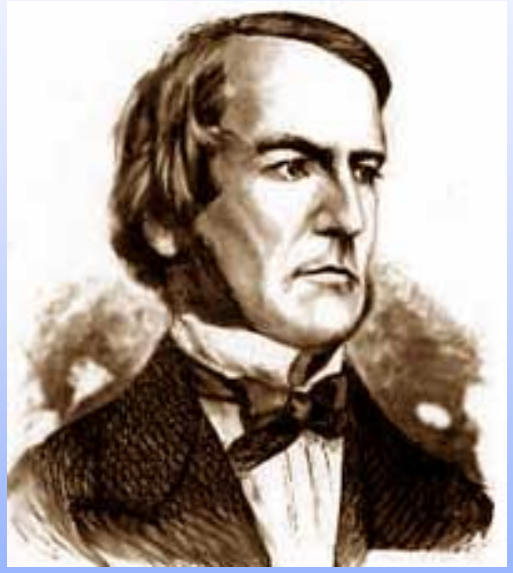
Teil D: Boolesche Retrievalsysteme



Kapitel D.1
Boolesches Retrieval

D.1 Boolesches Retrieval

George Boole (1815 – 1864)



- **„The Laws of Thought“ (1854)**
- **Wahrheitswerte (0 und 1)**
- **Funktoren:**
- **UND**
- **ENTWEDER – ODER**
- **UND NICHT**

- **Wirkung auf:**
- **Informatik (Boolesche Algebra) und**
- **Informationswissenschaft (Boolesche Retrievalsysteme)**

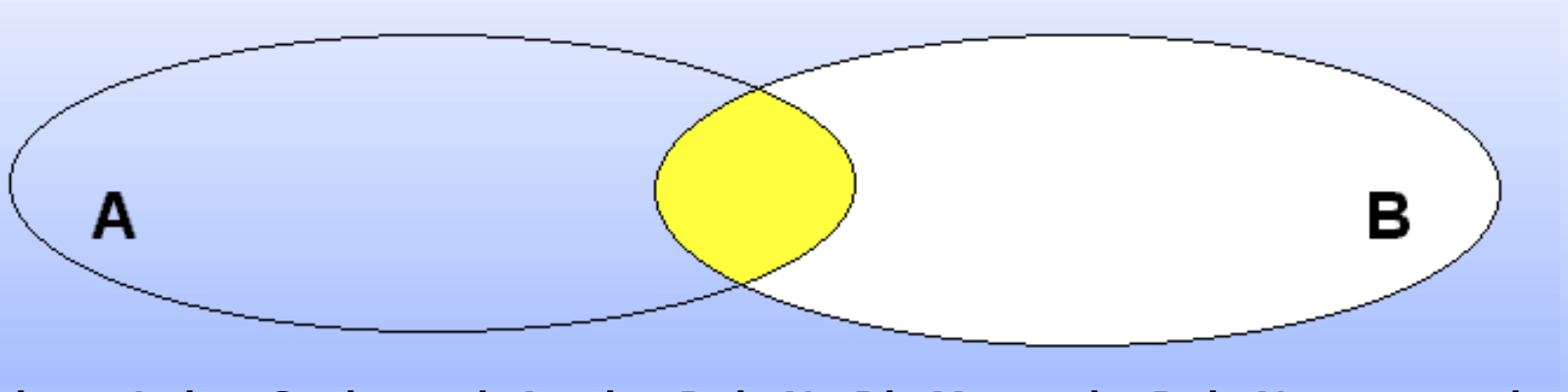
D.1 Boolesches Retrieval

Atomare Suchargumente

- **kleinste Einheiten eines Sucharguments: je nach Indexform Wort oder Phrase**
- **Trunkierungen:**
  - **offene Rechts- / Linkstrunkierung (0 bis ∞ viele Zeichen)**
  - **begrenzte Rechts- / Linkstrunkierung (ersetzt genau n Zeichen)**
  - **Binnentrunkierung (begrenzt)**
  - **spezifische Optionen (Ersetzung genau eines alphabetischen / numerischen Zeichen, eines Zeichens in einem gegebenen Intervall)**
- **Groß- / Kleinschreibung**
- **Umgang mit Sonderzeichen (z.B. @) oder mit Funktoren (z.B. und) – i.d.R. mittels Anführungszeichen**
- **feldspezifisch (AU=Marx, Karl) oder „querbeet“ (basic index)**

D.1 Boolesches Retrieval

Schnittmenge
A UND B

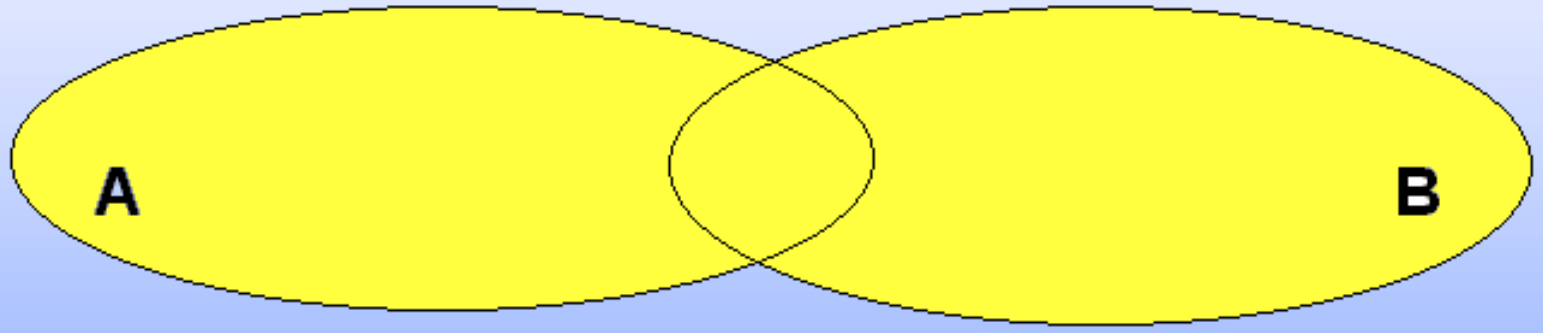


1. **Invertierter Index: Suche nach A nebst Dok.-Nr. Die Menge der Dok.-Nummern sei „Menge 1“.**
2. **Invertierter Index: Suche nach B nebst Dok.-Nr. Die Menge der Dok.-Nummern sei „Menge 2“.**
3. **Bestimme Schnittmenge aus „Menge1“ und „Menge 2“. Entstehende Menge sei „Menge 3“.**
4. **Folge den Verweisen aus „Menge 3“ zu den Dokumenten, kopiere diese zur Ausgabe!**

Salton, G., & McGill, M. (1983). Information Retrieval – Grundlegendes für Informationswissenschaftler. Hamburg: McGraw-Hill.

D.1 Boolesches Retrieval

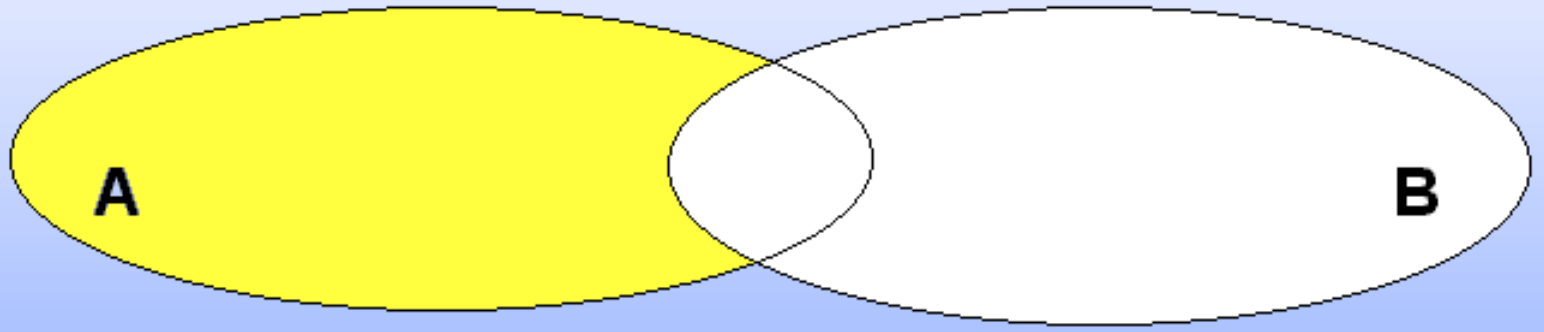
Vereinigungsmenge
A ODER B



1. **Invertierter Index: Suche nach A nebst Dok.-Nr. Die Menge der Dok.-Nummern sei „Menge 1“.**
2. **Invertierter Index: Suche nach B nebst Dok.-Nr. Die Menge der Dok.-Nummern sei „Menge 2“.**
3. **Bestimme Vereinigungsmenge aus „Menge1“ und „Menge 2“. Entstehende Menge sei „Menge 3“.**
4. **Folge den Verweisen aus „Menge 3“ zu den Dokumenten, kopiere diese zur Ausgabe!**

D.1 Boolesches Retrieval

Exklusionsmenge
A UND NICHT B

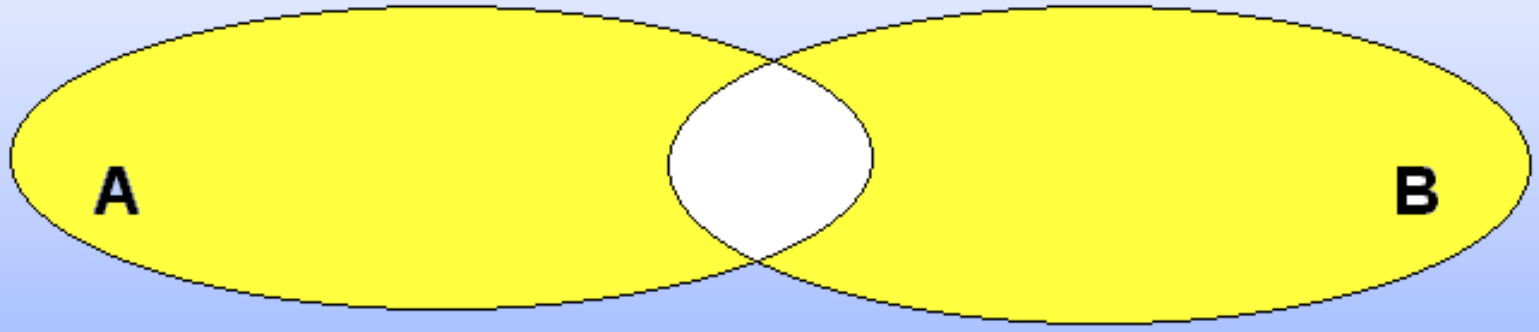


1. **Invertierter Index: Suche nach A nebst Dok.-Nr. Die Menge der Dok.-Nummern sei „Menge 1“.**
2. **Invertierter Index: Suche nach B nebst Dok.-Nr. Die Menge der Dok.-Nummern sei „Menge 2“.**
3. **Lösche sämtliche Elemente aus „Menge 1“, die auch Elemente aus „Menge 2“ sind!**
4. **Folge den Verweisen aus der verbleibenden „Menge 1“ zu den Dokumenten, kopiere diese zur Ausgabe!**

D.1 Boolesches Retrieval

Ausschließende Exklusionsmenge

A XOR B



1. Invertierter Index: Suche nach A nebst Dok.-Nr. Die Menge der Dok.-Nummern sei „Menge 1“.
2. Invertierter Index: Suche nach B nebst Dok.-Nr. Die Menge der Dok.-Nummern sei „Menge 2“.
3. Bestimme Vereinigungsmenge aus „Menge1“ und „Menge 2“. Entstehende Menge sei „Menge 3“.
4. Bestimme Schnittmenge aus „Menge1“ und „Menge 2“. Entstehende Menge sei „Menge 4“.
5. Lösche alle Elemente aus „Menge 3“, die auch Element von „Menge 4“ sind!
6. Folge den Verweisen aus der verbleibenden „Menge 3“ zu den Dokumenten, kopiere diese zur Ausgabe!

D.1 Boolesches Retrieval

Boolesche Funktoren in aussagenlogischer Deutung

A	B	A UND B	A ODER B	A UND NICHT B	A XOR B
w	w	w	w	f	f
w	f	f	w	w	w
f	w	f	w	f	w
f	f	f	f	f	f
		Konjunktion	***Disjunktion***	***Postsektion***	***Kontravalenz***
		„beides“	***„mindestens eines“***	***„das eine ohne das andere“***	***„entweder das eine oder das andere“***

Bochenski, I.M., & Menne, A. (1973). Grundriß der Logistik. Paderborn: Schöningh.

D.1 Boolesches Retrieval

Regeln

(A, B, C: Suchatome; BO: beliebiger Boolescher Operator)

- **Assoziativgesetz: (A BO B) BO C = A BO (B BO C) (für UND, ODER, XOR)**
- **Kommutativgesetz: A BO B = B BO A (für UND, ODER, XOR)**
- **konjunktives Distributivgesetz:**
  (A UND B) ODER (A UND C) = A UND (B ODER C)
- **disjunktives Distributivgesetz:**
  (A ODER B) UND (A ODER C) = A ODER (B UND C)
- **Distributivgesetze der Postsektion (DeMorgansche Gesetze):**
  (A UND NICHT B) UND (A UND NICHT C) = A UND NICHT (B ODER C)
  (A UND NICHT B) ODER (A UND NICHT C) = A UND NICHT (B UND C)

D.1 Boolesches Retrieval

Abstandsoperatoren (Verschärfung des Booleschen UND)

- **(1) direkte Nachbarschaft:**
  - **Phrasen: *„Miranda Otto“***
  - **benachbarte Worte in Reihenfolge: *Miranda ADJ Otto* findet Miranda Otto**
  - **benachbarte Worte ohne Beachtung der Reihenfolge: *Miranda (N) Otto* findet Miranda Otto und Otto, Miranda**
- **(2) numerische Abstandsoperatoren:**
  - **Suche nach Worten im Abstand von n Worten (n frei wählbar): *Miranda (N) Otto W/25 Eowyn* findet alle Texte, in denen die Namen im Abstand von max. 25 Worten vorkommen**
    - **mehrfache Anwendung von W/n findet (bei geschickt gewähltem n) hochrelevante Texte: *Auenland W/25 Auenland W/25 Auenland***
  - **Suche nach Worten im Abstand von n Worten (n fest, i.d.R. 10): *Eowyn NEAR Aragorn* findet Texte, in denen die Namen im Abstand von max. 10 Worten vorkommen**

D.1 Boolesches Retrieval

Abstandsoperatoren (Verschärfung des Booleschen UND)

- **(3) grammatische Nachbarschaft:**
  - **(nicht) im gleichen Satz (auch bei thematischen Ketten des syntaktischen Indexierens)**
    - **A UND.S B**
    - **A NICHT.S B**
  - **(nicht) im gleichen Absatz**
    - **A UND.P B**
    - **A NICHT.P B**
  - **(nicht) im gleichen Feld**
    - **A UND.F B**
    - **A NICHT.F B**
  - **Satzanfang: #A (A steht am Satzanfang)**
- **Probleme mit Anaphora (... Miranda Otto ... . Sie spielt die Eowyn.)**

D.1 Boolesches Retrieval

Hierarchische Suche

Beispiel:

STN

```
=> s Haushalt/CTDE
L11         1442 HAUSHALT/CTDE

=> s Haushalt+NT/CTDE
L12         4275 HAUSHALT+NT/CTDE   (22 TERMS)

=> s Haushalt+NT1/CTDE
L13         4263 HAUSHALT+NT1/CTDE   (20 TERMS)

=> s Haushalt+BT1/CTDE
L14        37067 HAUSHALT+BT1/CTDE   (4 TERMS)

=> s Haushalt+RT/CTDE
L15         2510 HAUSHALT+RT/CTDE   (6 TERMS)

=> s Haushalt+NT1,RT/CTDE
L16         5232 HAUSHALT+NT1,RT/CTDE   (24 TERMS)
```

D.1 Boolesches Retrieval

Algebraische Operatoren

Z: numerischer Wert

- **A GLEICH [Z]**
- **A GRÖSSER ALS [Z]**
- **A KLEINER ALS [Z]**
- **A IM INTERVALL [Z, Z`]**

Häufigkeitsoperator (Angabe der Minimalhäufigkeit)

- **ATLEAST 20 (A): A muss min. 20mal vorkommen**

Kapitel D.2
Suchstrategien

D.2 Suchstrategien

Menügeführtes Boolesches Retrievalsystem

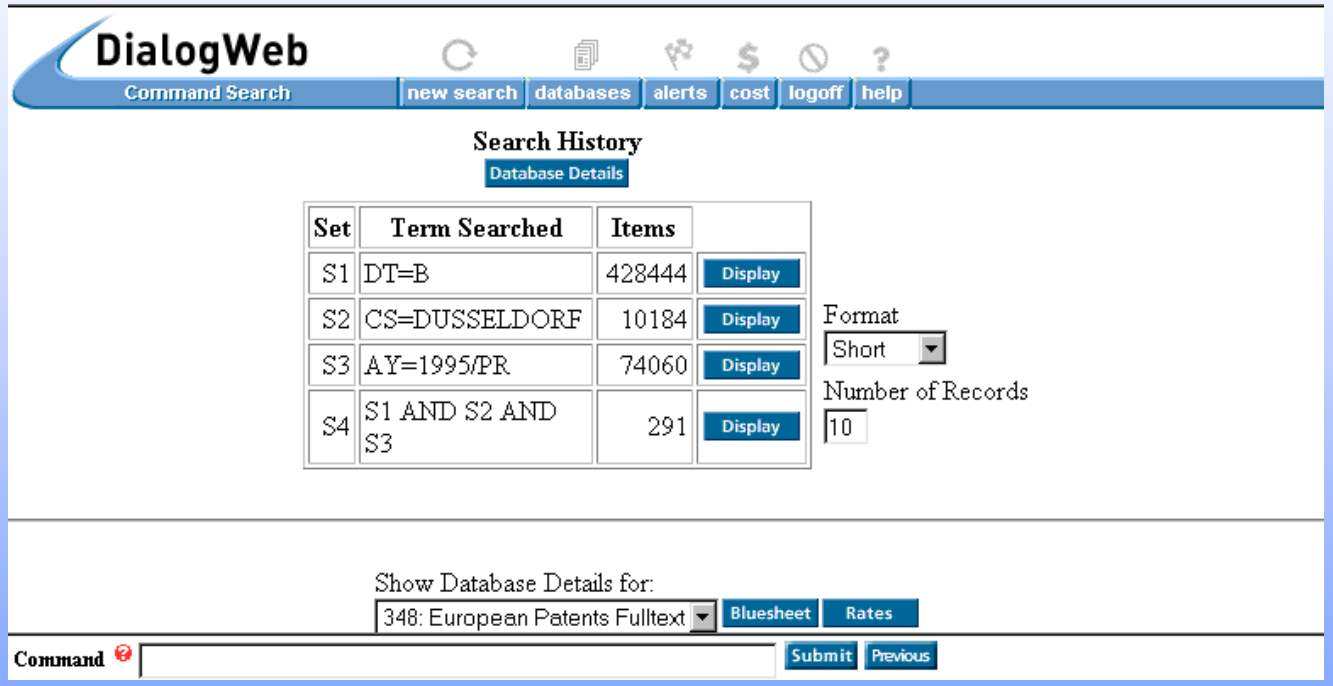
Beispiel:

MarketResearch.com Profound

D.2 Suchstrategien

Befehls-orientiertes Boolesches Retrievalsystem

Beispiel:
DialogWeb



D.2 Suchstrategien

Endnutzersysteme (Suchmaschinen)

- **bis ca. 2000: automatische Verknüpfung der Suchatome mit ODER**
- **danach: automatisch mit UND verknüpft**

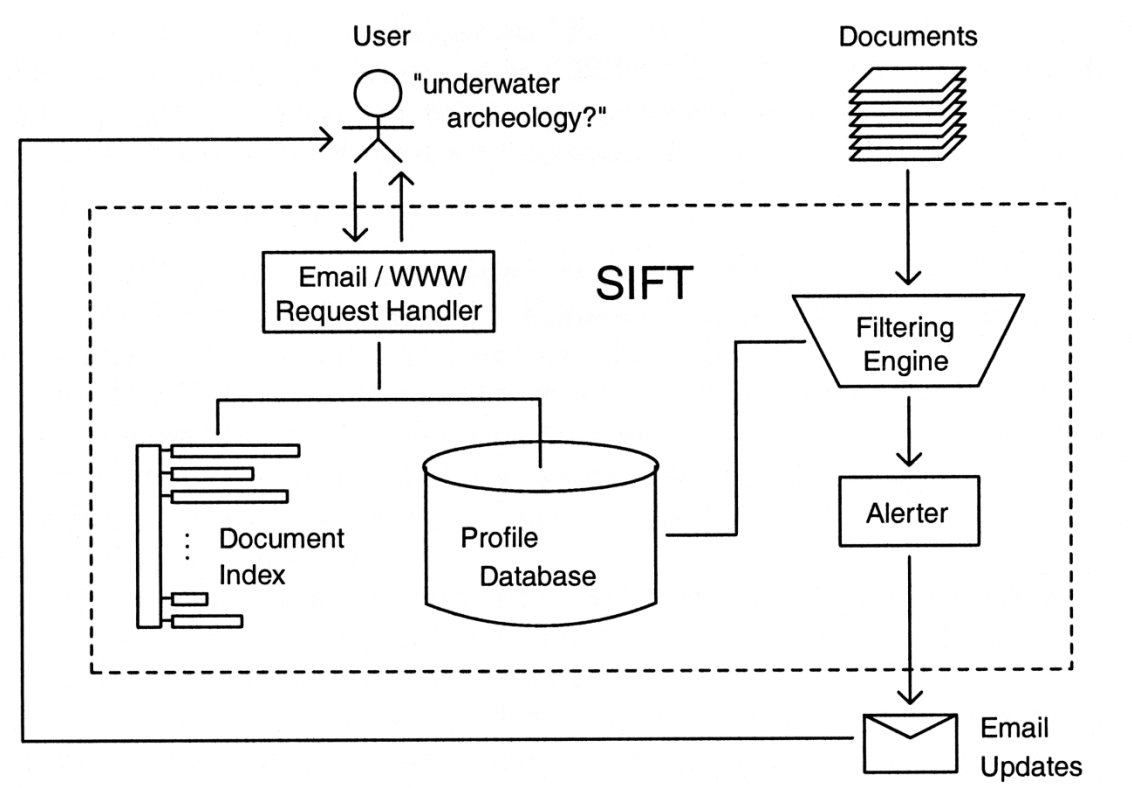
Ist dies sinnvoll?

- **Suchatome mit semantischer Ähnlichkeit (z.B. einem gemeinsamen Oberbegriff): mit ODER verknüpfen**
- **bei semantischer Unähnlichkeit: UND**

Das-Gupta, P. (1987). Boolean interpretation of conjunctions for document retrieval. Journal of the American Society for Information Science, 38, 245-254.

D.2 Suchstrategien

Einrichten eines Pushdienstes



Yan, T.W., & Garcia-Molina, H. (1994). Index structures for selective dissemination of information under the Boolean model. ACM Transactions on Database Systems, 19, 332-364.
Yan, T.W., Garcia-Molina, H. (1999). The SIFT information dissemination system. ACM Transactions on Database Systems, 24, 529-565.

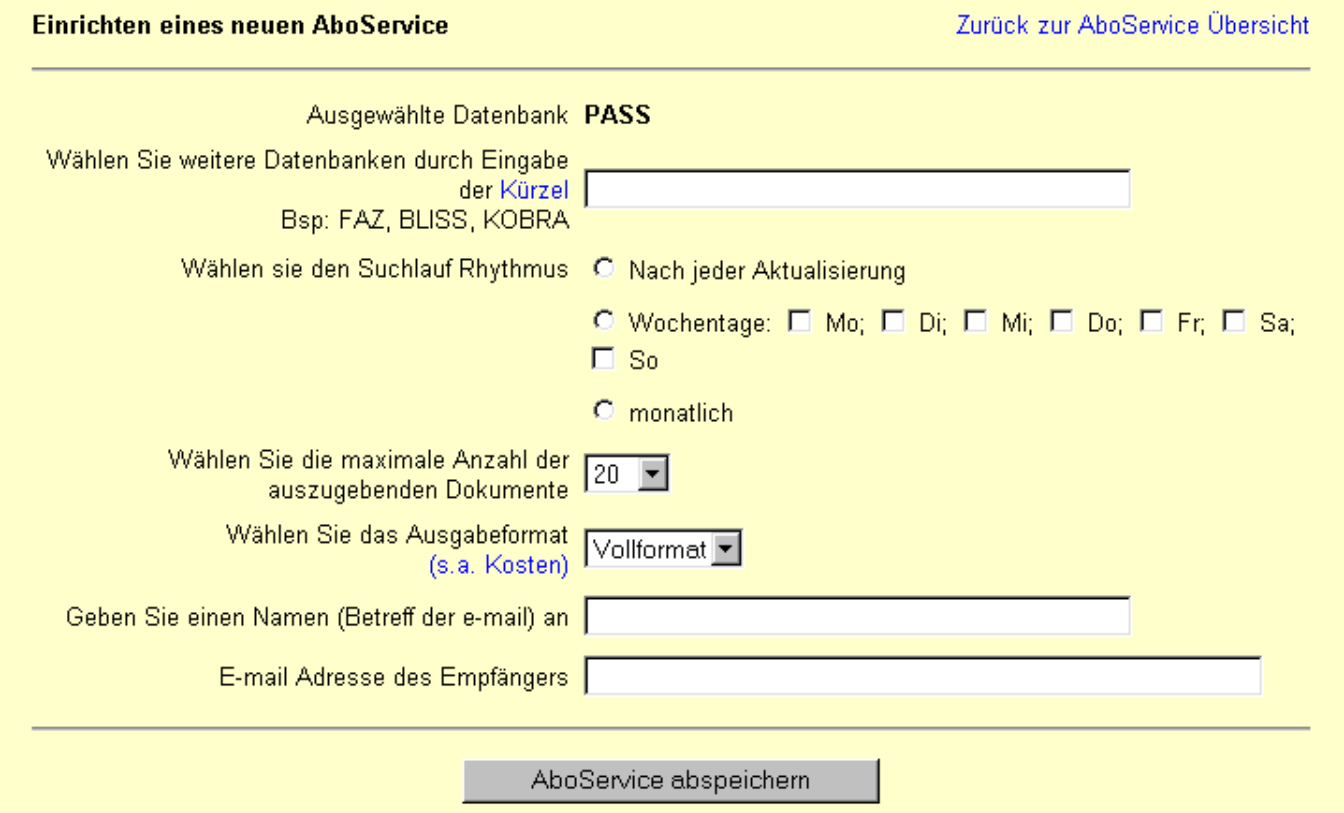
D.2 Suchstrategien

Einrichten eines Pushdienstes

- **Suchargument ist vorhanden; Treffermenge zufriedenstellend**
- **Name des Suchprofils definieren**
- **Periodizität festlegen**
- **Lieferanschrift eingeben**

  hier:

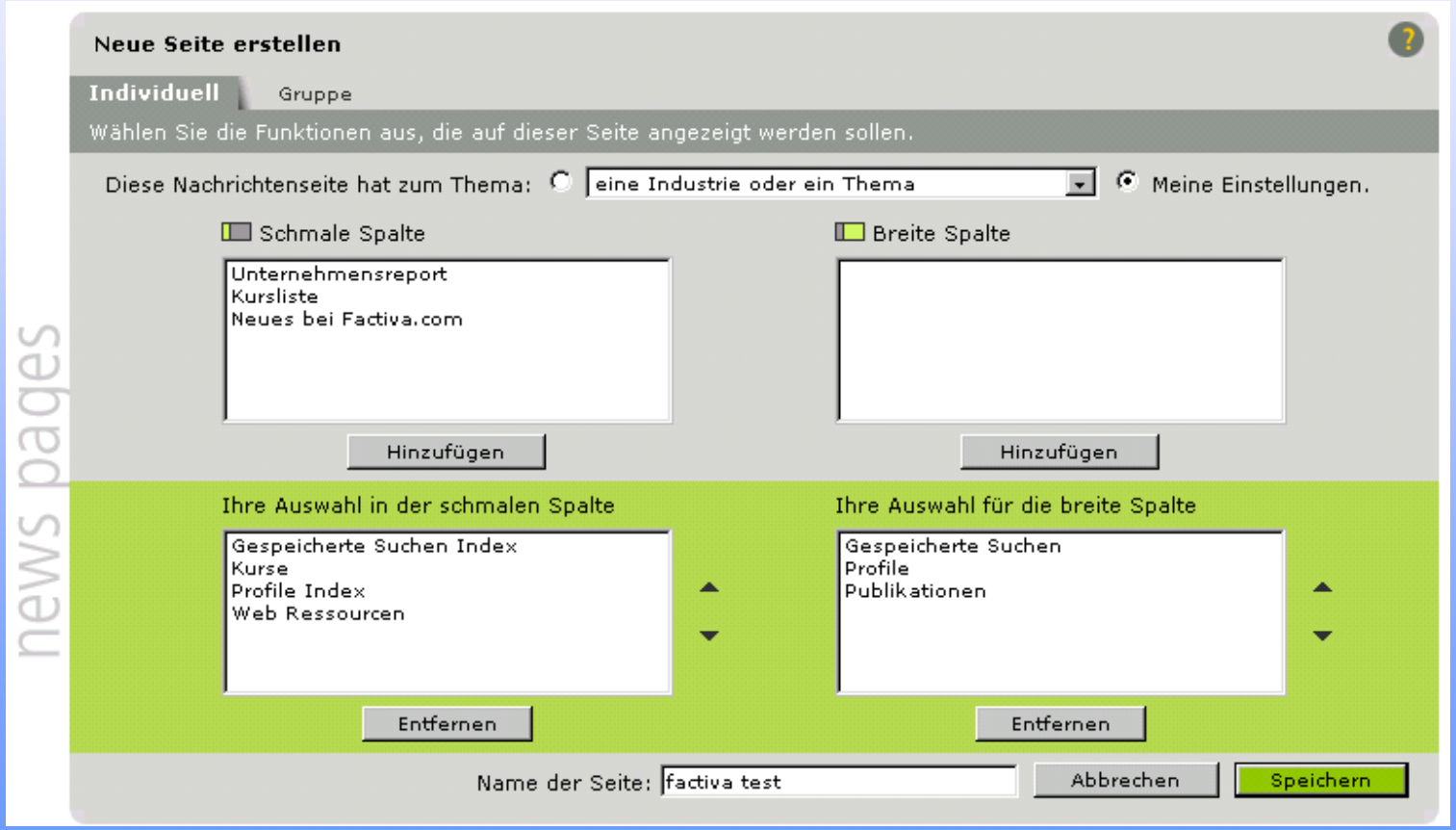
  E-Mail (Genios)



D.2 Suchstrategien

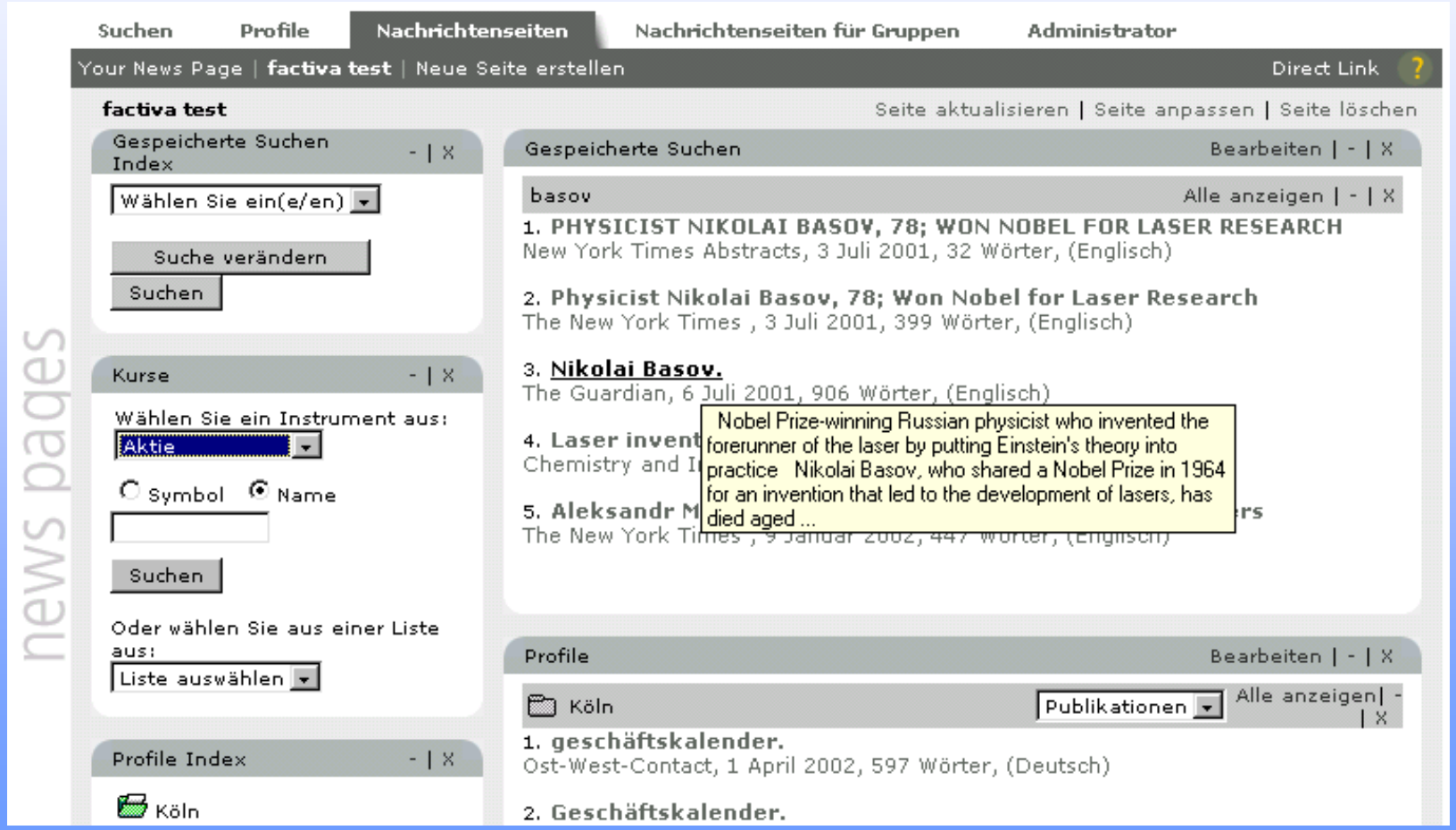
Einrichten eines Pushdienstes

– **auf Homepage ausliefern (*Beispiel:* Factiva)**



D.2 Suchstrategien

Auslieferung des Pushdienstes auf Homepage



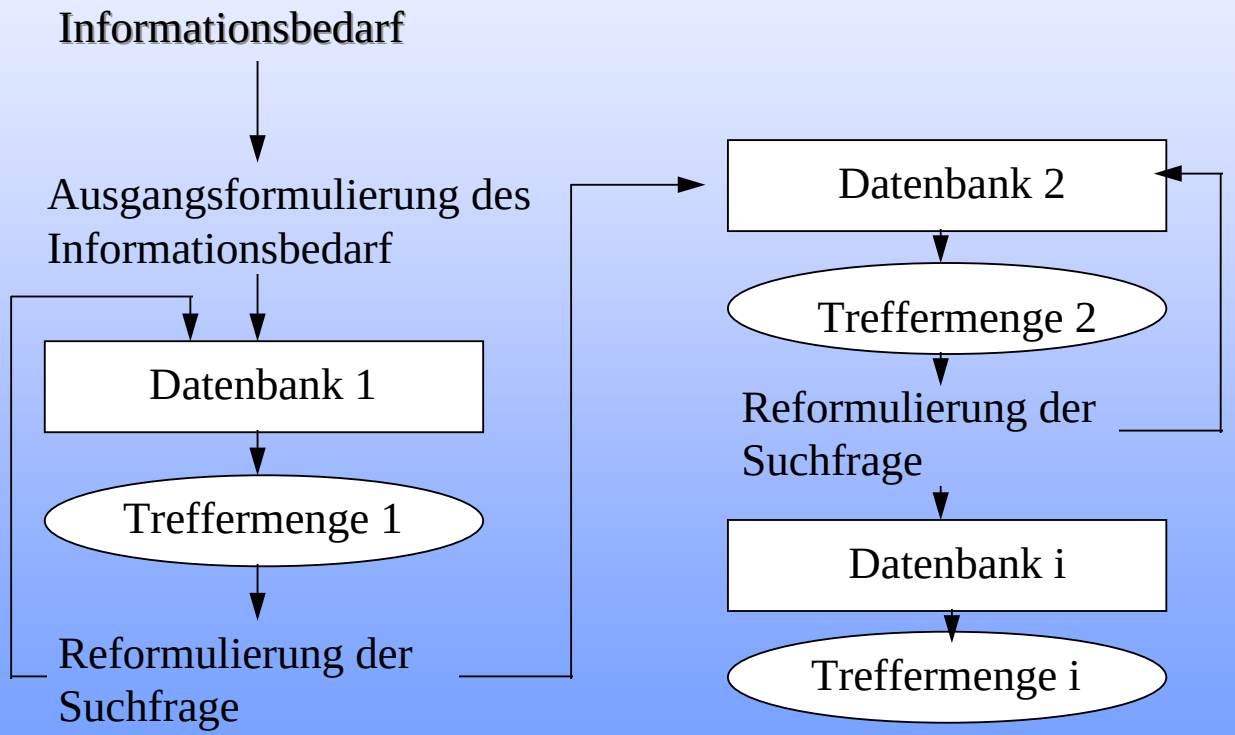
D.2 Suchstrategien

Recherchestrategien („Nadel-im-Heuhaufen-Syndrom“)

- **Phase 1: Suchen und Finden der relevanten Datenbanken**
  derzeit: Web-Suchwerkzeuge; mehr als 30.000 kommerzielle Datenbanken; weitere tausende von freien Deep-Web-Datenbanken (Typen: bibliographische Datenbanken, Volltext-datenbanken, Faktendatenbanken)
- **Phase 2: Suchen und Finden der relevanten Dokumentationseinheiten**
  derzeit enthalten große Anbieter mehrere Milliarden von Dokumentationseinheiten (Google: ca. 10 Mrd.; LexisNexis: ca. 5 Mrd.) – Struktur: B-E-S-T (begin – expand – select – type)
- **Phase 3: Modifikation der Suchargumente**
  recall- / precision-steigernde Maßnahmen

D.2 Suchstrategien

Strategie beim problemorientierten Informationsbedarf: Berrypicking



Bates, M.J. (1989). The design of browsing and berrypicking techniques for the online search interface. Online Review, 13, 407-424.

D.2 Suchstrategien

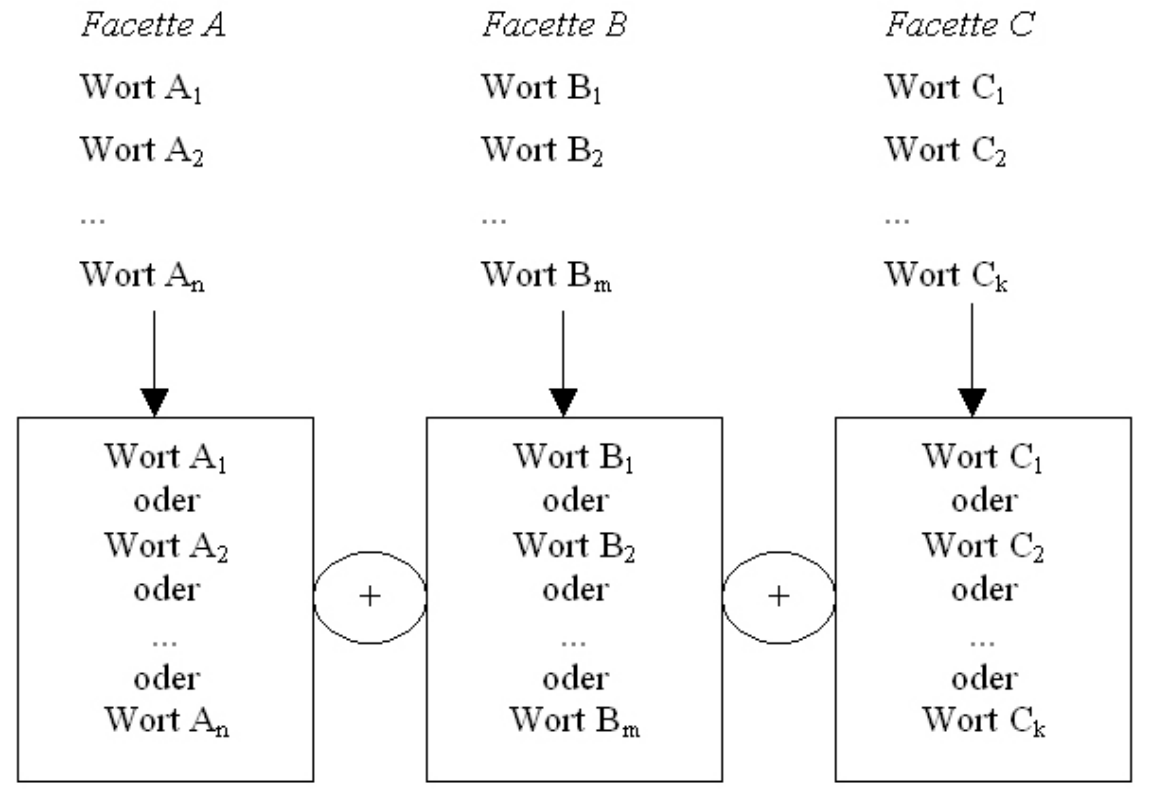
Basisstrategien intellektueller Anfragemodifikation

- **Blockstrategie**
  "building blocks strategy"
- **Perlensuche**
  "citation pearl growing strategy"

Efthimiadis, E.N. (1996). Query expansion.
Annual Review of Information Science and Technology, 31, 121-187.

D.2 Suchstrategien

Blockstrategie



D.2 Suchstrategien

Blockstrategie

- **jede Facette einzeln bearbeiten und optimieren**
  - **Wissensordnung**
    - **Synonyme**
    - **Quasi-Synonyme**
    - **Ober- / Unterbegriffe**
    - **verwandte Begriffe**
  - **Titelterme**
  - **Volltext-Terme**
  - **Verknüpfung: i.d.R. ODER**
- **Facetten über Suchschrittnummern verbinden**
  - **Verknüpfung: UND / Abstandsoperator, ggf. UND NICHT**

D.2 Suchstrategien

Rolle des Vorwissens des Anfragenden

- **Beispiel:**
- **"Abgasreinigung von Dieselmotoren in Blockheizkraftwerken"**
- **Auftraggeber: Mitarbeiter eines Blockheizkraftwerkes**

Blockstrategie

Abgasreinigung	***Dieselmotor***	***Blockheizkraftwerk***
Abgasrückführung	***Vorkammer-***	***Heizkraftwerk***
Katalysator	***dieselmotor***	***Kraft-Wärme-Kopplung***
Filter		***Blockkraftwerk***

vom Kolke, E.G. (1994). Online-Datenbanken. Systematische Einführung in die Nutzung elektronischer Fachinformation. München, Wien: Oldenbourg.

D.2 Suchstrategien

- **"Abgasreinigung von Dieselmotoren in Blockheizkraftwerken"**

- **nötiges Vorwissen:**
- ***Funktioniert die Abgasreinigung von Dieselmotoren bei Blockheizkraftwerken genauso wie auch sonst?***

- **falls ja: nur zwei Blöcke**
- **korrekte Anfrage: "Abgasreinigung von Dieselmotoren"**
- **ggf. (bei großer Treffermenge): im letzten Schritt mit "Block-heizkraftwerk" (und Synonymen) im Abstract oder Volltext suchen; Treffer *zusätzlich* ausgeben**

- **falls nein: alle drei Blöcke**

D.2 Suchstrategien

"Perlensuche": optimal zutreffende Dokumente suchen

- **Strategie 1: äußerst eng verbal suchen (mit maximal diskriminierenden Termen recherchieren)**

- **Strategie 2: Sortieren einer größeren (thematisch einschlägigen) Treffermenge nach Zitationshäufigkeit (Achtung: problematisch bei neuer Literatur); u.U. alle Dokumente bis zur h-Index-Grenze berücksichtigen**
- **anschließend: Dokument(e) terminologisch oder zitatenanalytisch ausschlachten**

D.2 Suchstrategien

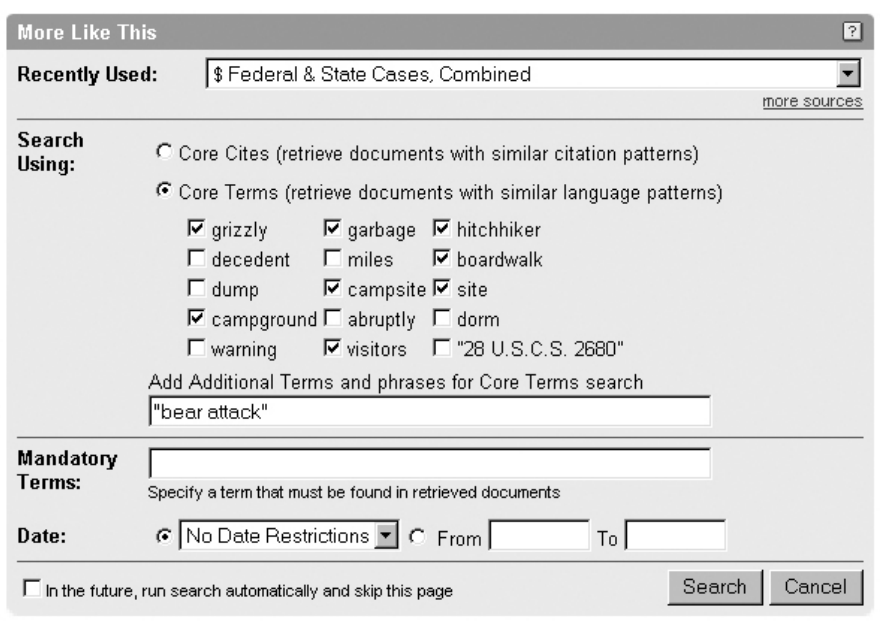
- **Ausgang: Perlen**

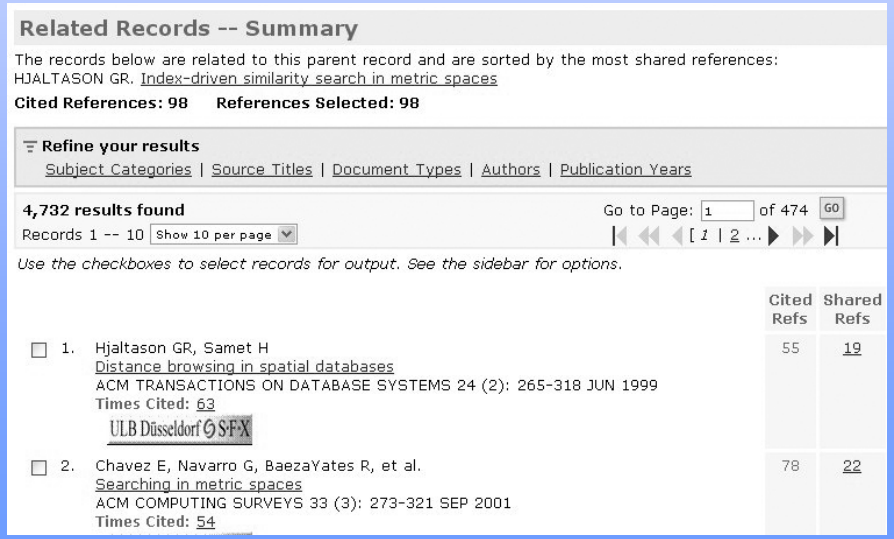
- **bibliographischer Nachweis**
  - **Terminologie entlehnen (Sachtitel, kontrolliertes Vokabular, ggf. Abstract)**
- **Terme für Blockstrategie übernehmen**

- **bei fehlendem Vorwissen (hier stets: 1. Schritt in der Suchstrategie)**
- **Volltext**
  - **Lesen und verstehen**
  - **ggf. Suchstrategie anpassen**

D.2 Suchstrategien

- **Ausgang: "Perle"**
- **"finding nearest neighbors"**
  - **über gemeinsame Terme**





- **"finding nearest neighbors"**
  - **über gemeinsame Referenzen (bibliographic coupling)**

D.2 Suchstrategien

Vorschlag neuer Suchargumente

- **Nutzung paradigmatischer Relationen (Unterbegriffe, Oberbegriffe, verwandte Begriffe)**
- **Nutzung syntagmatischer Relationen der Gesamtdatenbank**
  - **im Web Information Retrieval: Beschränkung auf Ankertexte**
- **Nutzung syntagmatischer Relationen der Dokumente in der Treffermenge (Bsp.: AltaVista "Prisma")**

Kapitel D.3

Gewichtetes Boolesches Retrieval

D.3 Gewichtetes Boolesches Retrieval

Problem Boolescher Systeme: binärer Ansatz der Relevanz – keine Sortierung nach Relevanz möglich

Suchanfrage 1: A UND B UND C UND D UND E

Dokument enthält: A, B, C, D (relevant?)

Suchanfrage 2: A ODER B ODER C ODER D ODER E

Dokument 1 (Treffer) enthält A

Dokument 2 (Treffer) enthält, A, B, C, D, E

Dokument 2 relevanter als Dokument 1?

D.3 Gewichtetes Boolesches Retrieval

Boolesches Retrieval mit Gewichtung

- **1. Gewichtung der Terme in den Dokumenten und damit des Dokuments selbst (Retrievalstatuswert) und**
- **ggf. 2. der Terme aus der Suchanfrage**
- **damit: Neuinterpretation der Booleschen Funktoren**

D.3 Gewichtetes Boolesches Retrieval

Waller-Kraft „Wish List“ für legitime Boolesche Ausdrücke

- **Der Retrievalstatuswert eines Dokuments ist eine Funktion der Gewichtungswerte der Suchterme im Dokument.**
- **Diese Funktion ergibt bei binärer Gewichtung (0/1) das klassische Boolesche Retrieval.**
- **Logisch äquivalente Suchargumente führen zu identischen Retrievalstatuswerten der Dokumente.**
- **Der Retrievalstatuswert steigt mit dem Anstieg der Termgewichte monoton an.**
- **Eine Gewichtung eines Suchterms mit 0 ist zugelassen.**

Cater, S.C., & Kraft, D.H. (1989). A generalization and clarification of the Waller-Kraft wish list. Information Processing & Management, 25, 15-25.

D.3 Gewichtetes Boolesches Retrieval

Waller-Kraft „Wish List“ für legitime Boolesche Ausdrücke

- **Senkt ein Nutzer den Gewichtungswert für genau ein Suchatom ab, so werden (soweit vorhanden) weitere Dokumente gefunden.**
- **Werden, mit UND verknüpft, neue Suchatome einer Anfrage zugefügt, so findet man keine neuen Dokumente. Werden, mit ODER verknüpft, neue Suchatome einer Anfrage zugefügt, so bleiben die ursprünglichen Dokumente in der Treffermenge erhalten.**
- **Es gelten bei UND-NICHT-Verknüpfungen die Gesetze von DeMorgan.**

D.3 Gewichtetes Boolesches Retrieval

Fuzzy Boolesches Retrieval

Grundlagen: Mehrwertige Logik – Fuzzy Logic

- **Minimum-Maximum-Modell (MM-Modell):**
- **Negation:**

  NICHT(mehrwertig) A = 1 – A
- **Konjunktion:**

  A UND(mehrwertig) B = MIN (A, B)
- **Disjunktion:**

  A ODER(mehrwertig) B = MAX (A, B)

Zadeh, L. (1965). Fuzzy sets. Information and Control, 8, 338-353.

D.3 Gewichtetes Boolesches Retrieval

Mixed Minimum-Maximum-Modell (MMM-Modell):

- **Konjunktion:**

  $e(d_k, j \text{ UND } j' \text{ UND } j'' \text{ UND } ...) = \gamma * MAX(j_k, j'_k, j''_k, ...)$

  $+ (1 - \gamma) * MIN(j_k, j'_k, j''_k, ...)$ für $0 \leq \gamma < 0{,}5$

- **Disjunktion:**

  $e(d_k, j \text{ ODER } j' \text{ ODER } j'' \text{ ODER } ...) = \gamma * MAX(j_k, j'_k, j''_k, ...)$

  $+ (1 - \gamma) * MIN(j_k, j'_k, j''_k, ...)$ für $0{,}5 \leq \gamma < 1$

e =: Retrievalstatuswert des Dokuments

d_k =: Dokument k

Fox, E.A., Betrabet, S., Koushik, M., & Lee, W. (1992). Extended Boolean models. In W.B. Frakes & R. Baeza-Yates (Eds.), Information Retrieval. Data Structures & Algorithms (pp. 393-418). Englewood Cliffs, NJ: Prentice Hall.

D.3 Gewichtetes Boolesches Retrieval

Gewichtete Query

Gewichtung des Query-Terms: j(q)

Gewichtung dieses Terms in einem Dokument: j(k)

Statuswert des Terms im Dokument:

d(t) = j(q) * j(k)

(weiter – je nach Modell – mit MM bzw. MMM)

D.3 Gewichtetes Boolesches Retrieval

Ein unscharfer Funktor: UNDODER (ANDOR)

$e(d_k, j \text{ UNDODER } j' \text{ UNDODER } j'' \text{ UNDODER } ...) =$

$\gamma * MAX (j_k, j'_k, j''_k, ...)$

$+ (1 - \gamma) * MIN (j_k, j'_k, j''_k, ...)$ für $0 \leq \gamma < 1$

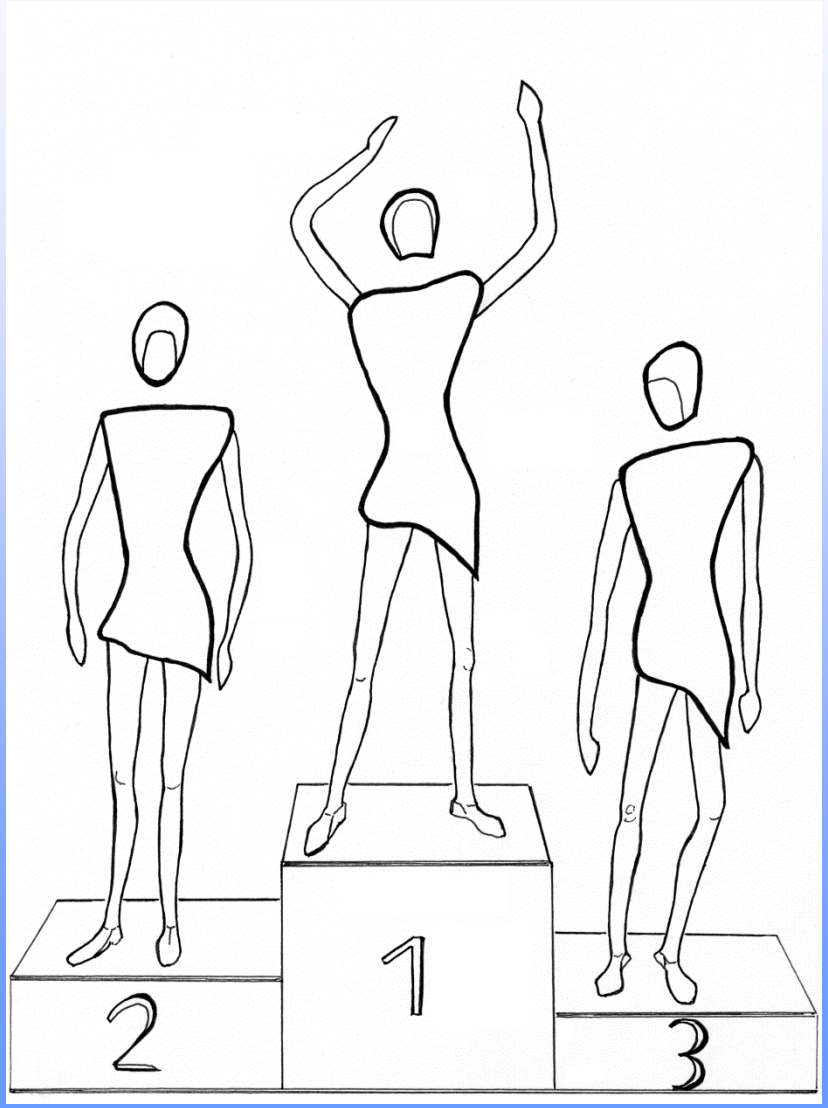
γ ist frei einstellbar (Schieberegler)

Werte größer 0,5 tendieren zu einem fuzzy ODER

Werte kleiner 0,5 tendieren zu einem fuzzy UND

Waller, W.G., & Kraft, D.H. (1979). A mathematical model of a weighted Boolean retrieval system. Information Processing & Management, 15, 235-245.

Teil E: Klassische Retrievalmodelle

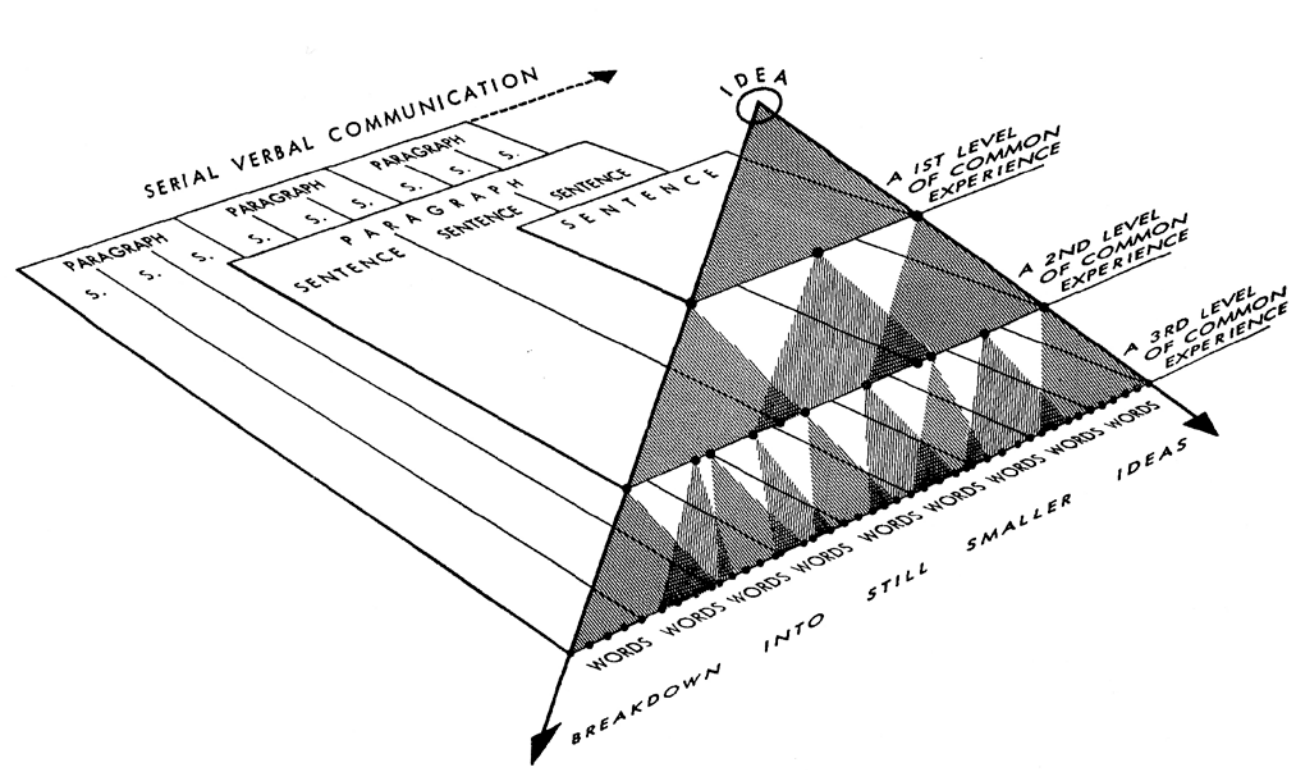


Kapitel E.1

Textstatistik

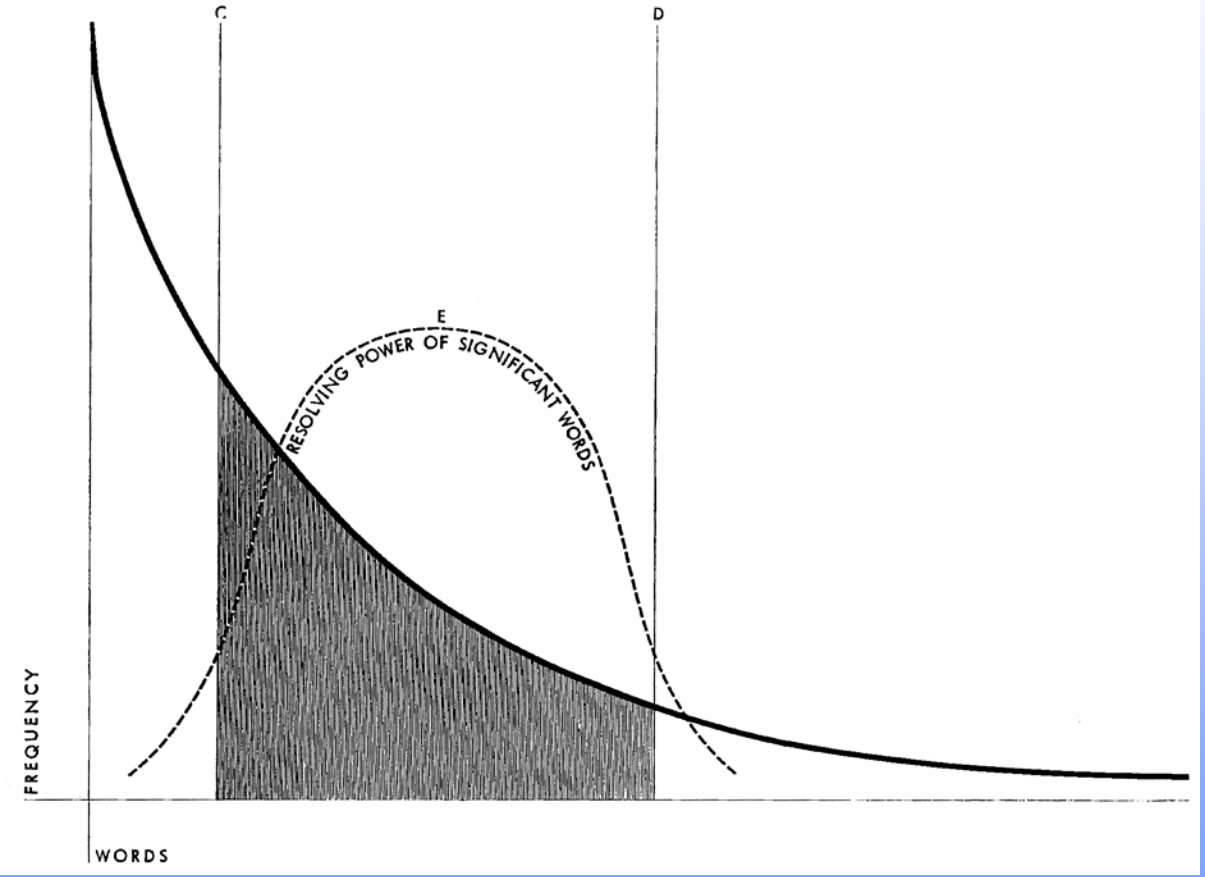
E.1 Textstatistik

Die These von Luhn: Termhäufigkeit als Signifikanzfaktor



Luhn, H.P. (1957). A statistical approach to mechanized encoding and searching of literary information. IBM Journal, 1(4), 309-317.

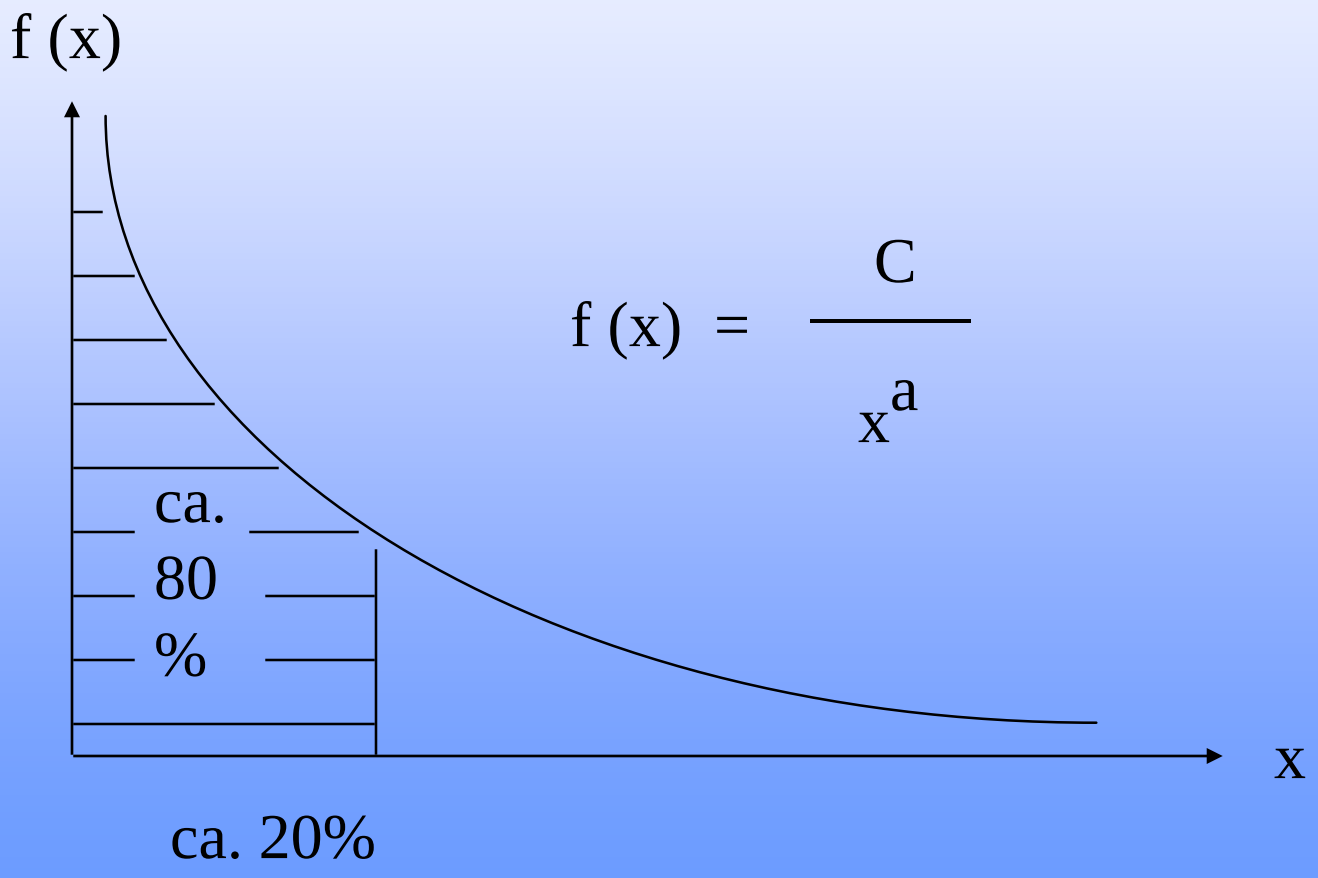
E.1 Textstatistik



Luhn, H.P. (1958). The automatic creation of literature abstracts. IBM Journal, 2(2), 159-165.

E.1 Textstatistik

Das Power-Law Verteilungsgesetz

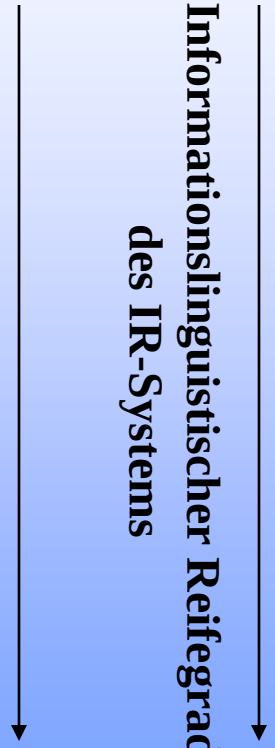


$$f(x) = \frac{C}{x^a}$$

E.1 Textstatistik

Termgewichtung: Was bedeutet "Term"?

- **1. unbearbeitete (flektierte) Wortform**
- **2. Grundform / Wortstamm**
- **3. Phrase**
- **4. Kompositum und (sinnvolle) Teile**
- **5. "named entities" erkannt**
- **6. Stufen 1 bis 6 mit Eingabefehlerkorrektur**
- **7. Begriff (Synonyme zugefügt - Homonyme getrennt)**
- **8. Begriff inkl. der Anaphora seiner Wörter**



E.1 Textstatistik

Termgewichtung: Aspekte

1. **Dokumentspezifische Termgewichtung (TF)**
2. **Kollektionsspezifische Termgewichtung (IDF)**
3. **üblich: TF * IDF**

E.1 Textstatistik

Termgewichtung: Zählbasis

- **Häufigkeit eines Terms**

  Term t im Dokument d: Freq(t,d)
- **Position des Terms im Dokument (P)**
  - **Term und Ort**
  - **Term in spezifischem Feld (bzw. Meta-Tag)**
- **Anzahl aller Terme in einem Dokument (L)**
- **Häufigkeit desjenigen Terms, der im Dokument d am häufigsten vorkommt: maxfreq(d)**
- **Anzahl aller Dokumente in einer Datenbank, in denen ein bestimmter Term (mindestens einmal) vorkommt (n)**
- **Gesamtanzahl der Dokumente in einer Datenbank (N)**

Harman, D. (1992). Ranking algorithms. In W.B. Frakes & R. Baeza-Yates (Hrsg.): Information Retrieval. Data Structures & Algorithms (pp. 363-392). Upper Saddle River, NJ: Prentice Hall.

E.1 Textstatistik

Termhäufigkeit (TF)

- 0,1-Verfahren (0: Term kommt nicht vor; 1: Term kommt vor)
  untauglich
- absolute Häufigkeit; freq(t,d)
  untauglich
- relative Häufigkeit bezogen auf Dokumentlänge (Vorschlag von Salton)
  TF-relH-Salton(t,d) = freq(t,d) / L
- relative Häufigkeit bezogen auf auf den häufigsten Term; zusätzlich Gewichtungsfaktor K (Vorschlag von Croft)
  TF-relH-Croft(t,d) = K + (1 - K) * freq(t,d) / maxfreq(d)
- WDF (logarithmische Werte) gemäß Harman
  WDF(t,d) = (ld [Freq(t,d) + 1]) / ld L
  Frage: Warum im Zähler "+1"?

E.1 Textstatistik

WDF - Beispiele

Beispiel (1) nach Harman-Formel: Dokument 1 hat 1024 Terme; Term A kommt 7mal vor

WDF (A,1) = (ld [7+1]) / ld 1024 = 3 / 10 = 0,3

Vergleich: relative Häufigkeit(A): 7/1024 = 0,7%

Beispiel (2) nach Harman-Formel: Dokument 1 hat 1024 Terme; Term B kommt 15mal vor

WDF (B,1) = (ld [15+1]) / ld 1024 = 4 / 10 = 0,4

Vergleich: relative Häufigkeit(B): 15/1024 = 1,5%

Der Wertebereich nach Harman-Formel ist „gestauchter“ als die relative Häufigkeit in %.

E.1 Textstatistik

Feld- oder positionsspezifische Termhäufigkeit (PWDF)

Variante der WDF-Formel

$$PWDF(t,d) = \{ld\ (2 * [freq(t\text{-}Titel,d)] + 1{,}7 * [freq(t\text{-}Sw,d)] + 1{,}3 * [freq(t\text{-}Abs,d)] + [freq(t\text{-}Text,d)] + 1)\} / \{ld\ (2 * L(Titel) + 1{,}7 * L(Sw) + 1{,}3 * L(Abs) + L(Text))\}$$

Die Gewichtungswerte (hier: Titelterm: 2, Schlagwort Sw: 1,7; Term im Abstract Abs: 1,3; Term im Fließtext: 1) sind frei einstellbar.

E.1 Textstatistik

Inverse Dokumenthäufigkeit (inverted document frequency weight) IDF

- **relative Häufigkeit des Vorkommens eines Terms in Dokumenten der gesamten Datenbank (je häufiger, desto kleiner IDF); je seltener ein Term ist, desto diskriminierender ist er**

Berechnungsformel nach Karen Sparck Jones:

$$IDF(t) = (ld\ N / n) + 1$$

Variante: $IDF'(t) = (ld\ N / n)$ (Einsatz, wenn keine Stoppwortliste vorhanden ist)

- **Beispiel: Datenbank hat 3584 Datensätze; Term A kommt in 7 Datensätzen vor**
- **$IDF(A) = (ld\ 3584 / 7) + 1 = (ld\ 512) + 1 = 9+1 = 10$**

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, 28, 11-21.

E.1 Textstatistik

Gewichtung G eines Terms t im Dokument d:

$$G(t,d) = TF(t,d) * IDF(t)$$

Der Retrievalstatuswert e des Dokuments ergibt sich durch Akkumulation der Gewichtungswerte

- **bei Booleschen Systemen: Zuordnung der Gewichtungswerte zu den Termen; Beachtung der Regeln erweiterter Boolescher Systeme (--> Kapitel D.3)**
- **bei natürlichsprachigen Systemen: Zuordnung der Gewichtungswerte zu den Termen**
  - **Variante 1 ("ODER")**

    $e(d) = G(t_1,d) + G(t_2,d) + ... + G(t_n,d)$ **für alle Dokumente d**
  - **Variante 2 ("UND")**

    Schritt 1: $e(d) = G(t_1,d) + G(t_2,d) + ... + G(t_n,d)$ **für solche Dokumente, die alle Suchterme enthalten**

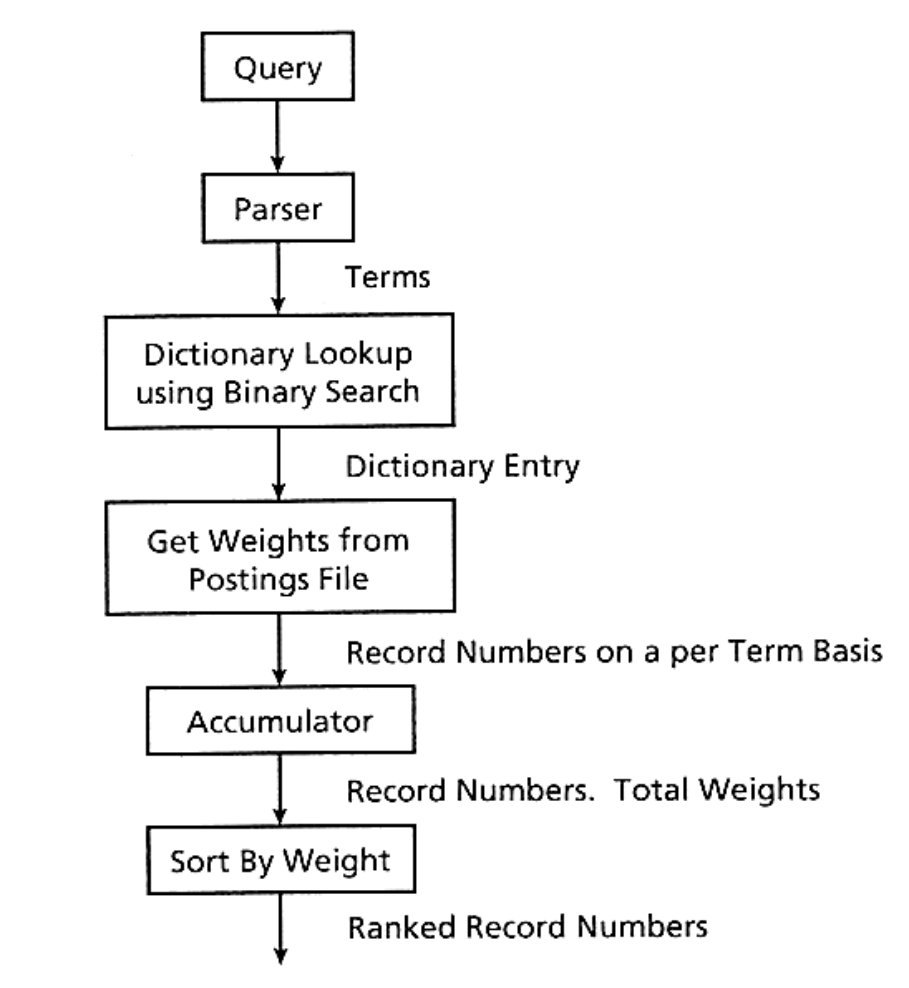
    Schritt 2: $e(d) = G(t_1,d) + G(t_2,d) + ... + G(t_{n-1},d)$ **für solche Dokumente, die alle Suchterme bis auf einen enthalten**

    usw.

E.1 Textstatistik

Sortierung nach Gewichtung.

Ablauf



E.1 Textstatistik

Sortiervariante 1: „ODER“. *Beispiel:* Freestyle

- **Errechnung der Gewichtungen durch WDF und IDF; Elimination von Stoppworten und von Hochfrequenzworten**
- **Abbildung der Gewichtungswerte für die Dokumente auf eine Skala von >0 bis 100**
- **Schaffen von Transparenz für den Nutzer**
  - **Angabe von Dokumentenanzahl und (normiertem) IDF für die Suchworte (.WHY)**
  - **Angabe der top-gerankten Dokumente mit dem Vorkommen bzw. Nichtvorkommen der Suchworte (.WHERE)**

Stock, W.G. (1998). Lexis-Nexis‘ Freestyle. Natürlichsprachige Suche – More like this! Password, Nr. 11, 21-28.

E.1 Textstatistik

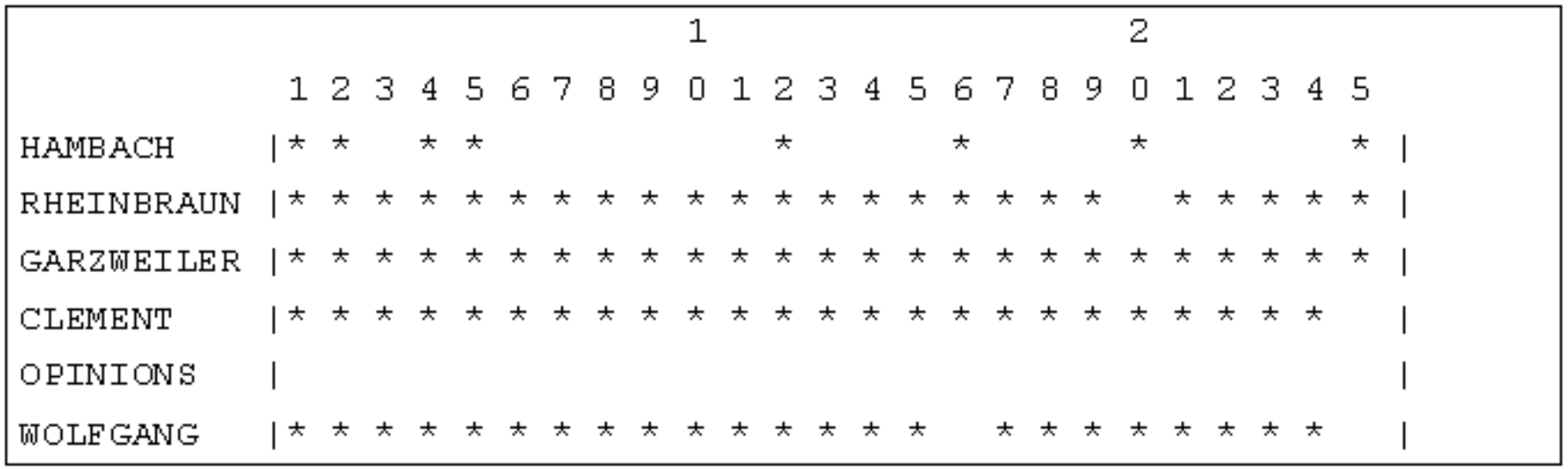
- **Anfragebeispiel: "Wolfgang Clement‘s opinions for the future of Rheinbraun in Hambach and Garzweiler" (max. 50 Dokumente)**

	Documents Retrieved	Documents Matched	Term Importance (0-100)
HAMBACH	8	373	29
RHEINBRAUN	49	675	25
GARZWEILER	50	1080	23
CLEMENT	48	4013	15
OPINIONS	1	13003	9
WOLFGANG	46	89679	1
FUTURE	--	--	--
FOR	--	--	--

Quelle: Lexis-Nexis

.WHY: Welche Gewichtungswerte erhalten die Suchargumente?

E.1 Textstatistik



	1										2															
	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	
HAMBACH	*	*		*	*							*				*				*					*	
RHEINBRAUN	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*		*	*	*	*	*	
GARZWEILER	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	
CLEMENT	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*		
OPINIONS																										
WOLFGANG	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*		*	*	*	*	*	*	*	*		

Quelle: Lexis-Nexis

.WHERE: Wo kommen die Suchargumente vor?

E.1 Textstatistik

Sortiervariante 2: „UND“

- **Einsatz bei diversen Suchmaschinen (u.a. Google, Alltheweb/FAST) – Variante: Abbruch nach Schritt 1**

1 - 3 of 3 Results for kerpen sindorf visteon Offensive content filter: **On** - Off

Visteon Corporation: **Visteon** Expands with New Technical Center in **Kerpen** / Germany
... Visteonstrasse" in **Kerpen-Sindorf**'s industrial park ... Center reinforces **Visteon**'s strong commitment ... **Visteon** Technical Center **Kerpen**: Image ... Technical Center **Kerpen Visteon** Interior ...
Description: **Visteon** Expands with New Technical Center in **Kerpen** / Germany
http://www.visteon.com/newsroom/press/2001/01story174.shtml - 29 KB

Visteon Corporation: Groundbreaking Ceremony For **Visteon** European Headquarters In...
... Groundbreaking Ceremony For **Visteon** European Headquarters In **Kerpen KERPEN**, Germany ... headquarters in **Kerpen**, Germany, **Visteon** Corporation (NYSE ... of the existing **Visteon** Technical Centre ... industrial park at **Kerpen-Sindorf**. The building covers ...
Description: Groundbreaking Ceremony For **Visteon** European Headquarters In **Kerpen**
http://www.visteon.com/newsroom/press/2002/02story63.shtml - 28 KB

Visteon Corporation: **Visteon** Corporation Annou
... **Visteon** Corporation Announces ... Headquarters In **Kerpen** Corporation (NYSE ... will be located at **Kerpen-Sindorf**, 20 kr technical ...
Description: **Visteon** Corporation Announces Plans To Open N
http://www.visteon.com/newsroom/press/2002/02story47.shtml

Risiko: Informations-verlust (Boolesches UND zu restriktiv)

alltheweb find it all
advanced search :: customize preferences :: submit site :: help
kerpen sindorf visteon pumuckl SEARCH
Results in: Any Language English
Web | News | Pictures | Video | Audio | FTP files
Offensive content filter: **On** - Off

No Web pages found that match your query.

Suggestions:

Kapitel E.2

Vektorraummodell

E.2 Vektorraummodell

- **Dokumente wie Anfragen werden als Vektoren in einem n-dimensionalen Raum verstanden, wobei die Dimensionen Termen entsprechen**
  - **Relevance Ranking geschieht nach der „Nähe“ der Vektoren (genauer: nach dem Winkel, der zwischen dem Anfragevektor und den Dokumentvektoren liegt)**
  - **entwickelt von Gerard Salton im Kontext des SMART-Systems (ab 1961 an der Harvard-University und ab 1965 Cornell-University, Ithaca)**
    SMART: System for the Mechanical Analysis and Retrieval of Text

Salton, G. (1968). Automatic Information Organization and Retrieval. New York, NY: McGraw-Hill.
Salton, G., Hrsg. (1971). The SMART Retrieval System – Experiments in Automatic Document Processing. Englewood Cliffs, N.J.: Prentice Hall.
Salton, G., & Lesk, M.E. (1965). The SMART automatic document retrieval system – An illustration. Communications of the ACM, 8, 391-398.
Salton, G., & McGill, M.J. (1983). Information Retrieval – Grundlegendes für Informationswissenschaftler. Hamburg [u.a.]: McGraw-Hill.
Salton, G., Wong, A., & Yang, C.S. (1975). A vector space model for automatic indexing. Communications of the ACM, 18, 613-620.

E.2 Vektorraummodell

- **gegeben: Anfrage mit m Termen**

  ATERM(j,k) ist der Gewichtungswert eines Terms j in Anfrage k
- **gegeben: Dokumentenmenge mit n Dokumenten und t unterschiedlichen Termen; TERM(i,j) ist der Gewichtungswert des Terms j im Dokument i**
- **die m Terme aus den Anfragen und die t Terme aus den Dokumenten werden jeweils als eine Dimension in einem (m+t)-dimensionalen Vektorraum angesehen**

	$TERM_1$	$TERM_2$	...	$TERM_t$
DOK_1	$TERM_{11}$	$TERM_{12}$	...	$TERM_{1t}$
DOK_2	$TERM_{21}$	$TERM_{22}$	...	$TERM_{2t}$
⋮	⋮			
DOK_n	$TERM_{n1}$	$TERM_{n2}$	...	$TERM_{nt}$

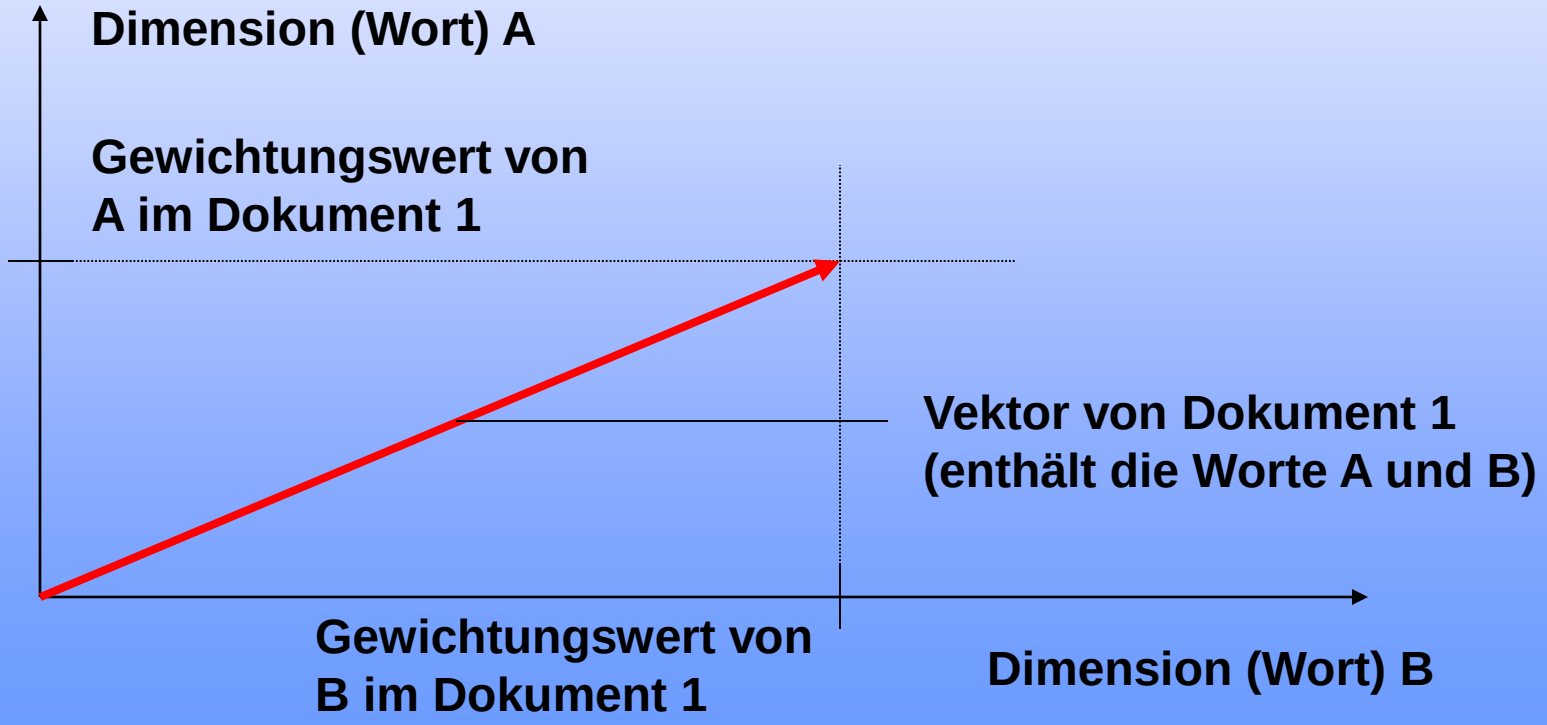
Dokument-Term (Dimension)-Matrix

- TERM(i,j) sind numerische Gewichtungswerte

- kommt ein Wort in einem Dokument nicht vor, ist TERM = 0

E.2 Vektorraummodell

- **ein Dokument wird durch den Vektor seiner Dimensionen (Terme) unter Berücksichtigung der jeweiligen Gewichtung (TF*IDF) repräsentiert**



E.2 Vektorraummodell

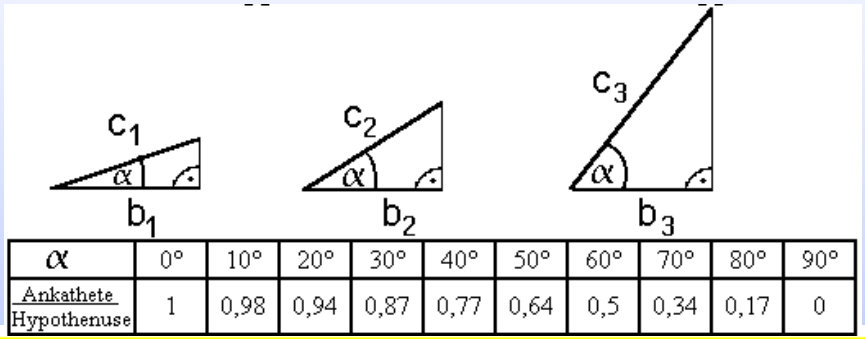
Ähnlichkeit zwischen Anfrage (genauer: Anfragevektor) und Dokumenten (genauer: Dokumentvektoren)

$$\text{COSINUS}(\text{DOK}_i, \text{ANFRAGE}_j) = \frac{\sum_{k=1}^{t} (\text{TERM}_{ik} \cdot \text{ATERM}_{jk})}{\sqrt{\sum_{k=1}^{t} (\text{TERM}_{ik})^2 \cdot \sum_{k=1}^{t} (\text{ATERM}_{jk})^2}}$$

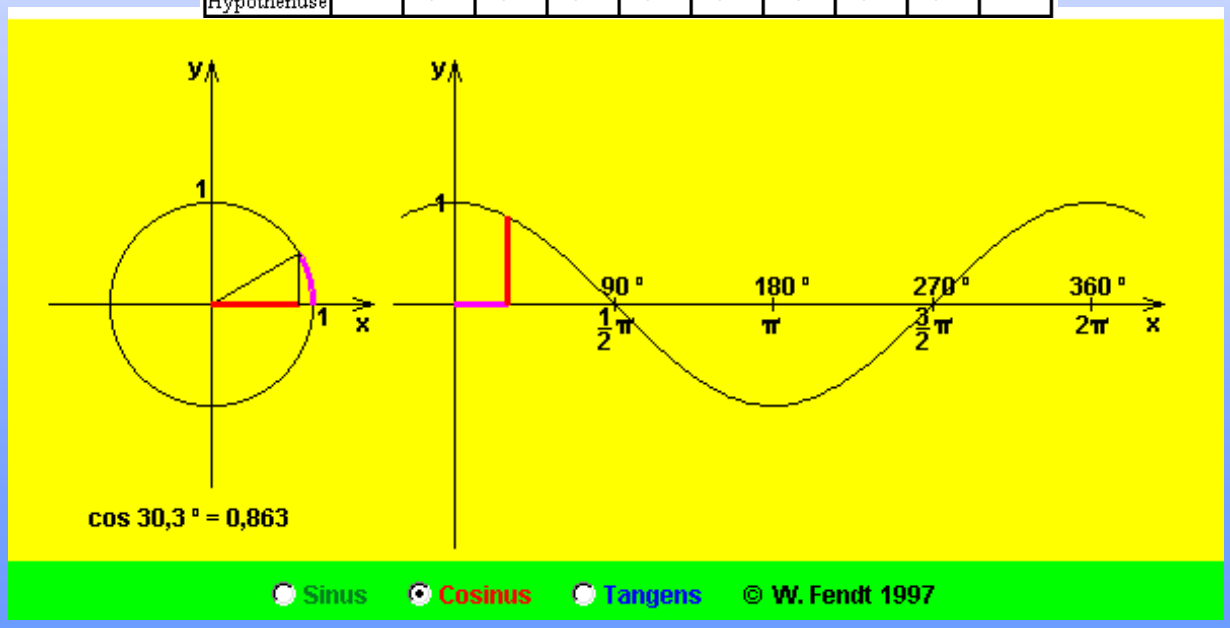
- **wenn die Vektoren übereinander liegen: Winkel: 0°**
- **Cosinus 0° = 1**

E.2 Vektorraummodell

Cosinus α = b : c (Ankathete : Hypothenuse)

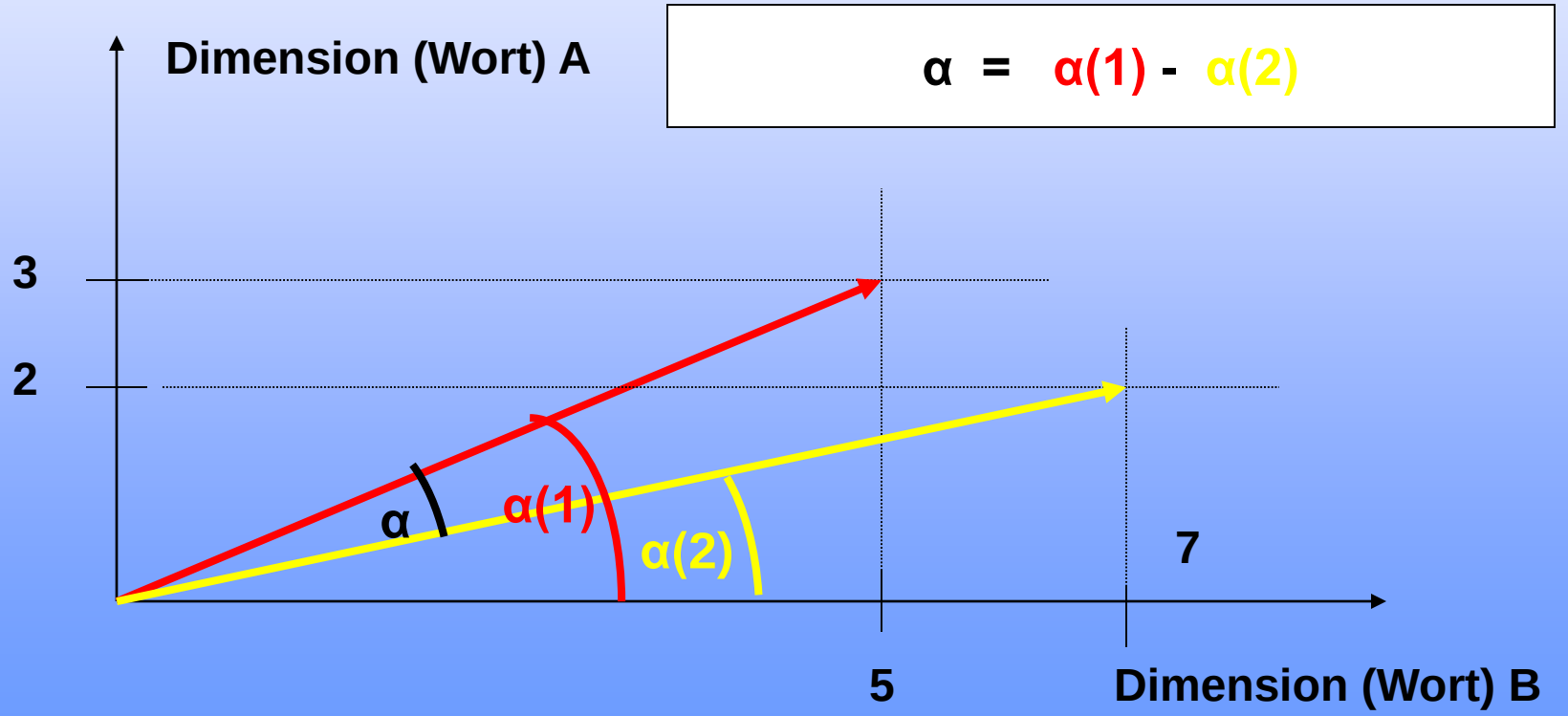


α	0°	10°	20°	30°	40°	50°	60°	70°	80°	90°
Ankathete / Hypothenuse	1	0,98	0,94	0,87	0,77	0,64	0,5	0,34	0,17	0



E.2 Vektorraummodell

Vektorraummodell: Berechnungsbeispiel Nähe von Dokument 1 (Wort A mit Gewicht 3 und Wort B mit Gewicht 5) und Dokument 2 (Wort A mit Gewicht 2 und Wort B mit Gewicht 7)



E.2 Vektorraummodell

Dok1 (5 | 3); Dok2 (7 | 2); cos α(n) = b : c (da c unbekannt, nach Pythagoras berechnen: $a^2 + b^2 = c^2$)

$\cos \alpha(n) = b : (a^2 + b^2)^{1/2}$

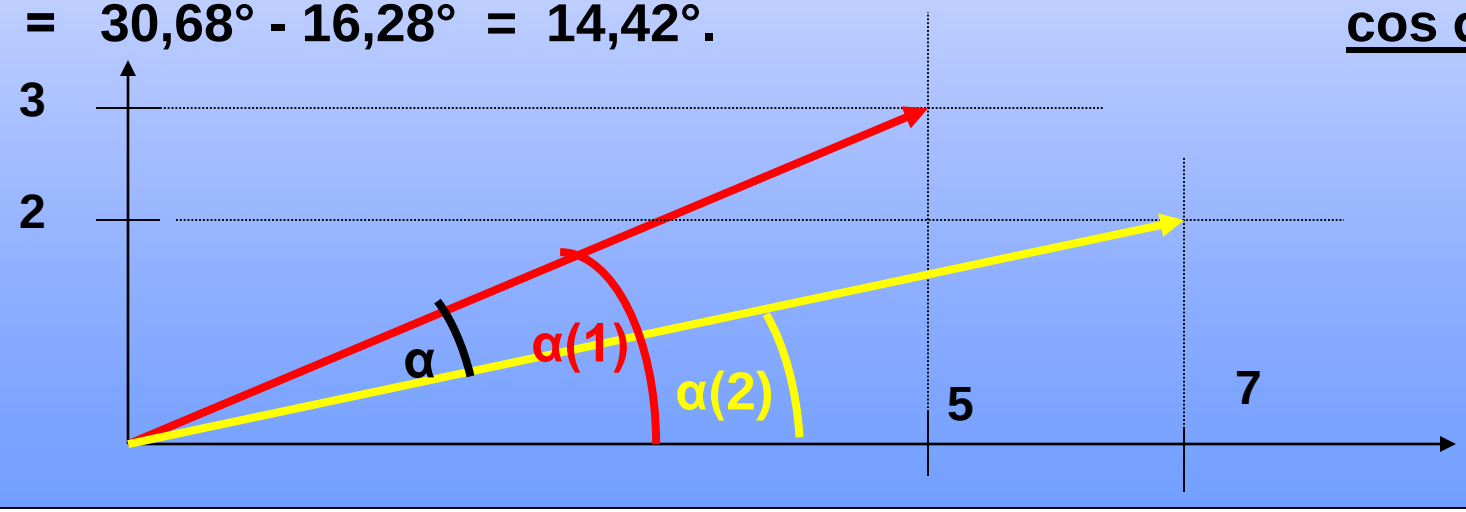
$\cos \alpha(1) = 5 : (5^2 + 3^2)^{1/2} = 5 : (25+9)^{1/2} = 5 : 34^{1/2} = 0{,}86 \sim \alpha(1) = 30{,}68°$

$\cos \alpha(2) = 7 : (7^2 + 2^2)^{1/2} = 7 : (49+4)^{1/2} = 7 : 53^{1/2} = 0{,}96 \sim \alpha(2) = 16{,}28°$

α = α(1) - α(2).

α = 30,68° - 16,28° = 14,42°.

cos α = 0,97



oder direkt nach dem Additionstheorem berechnen:
cos (α(1) – α(2)) = cos α(1) * cos α(2) + sin α(1) * sin α(2)

E.2 Vektorraummodell

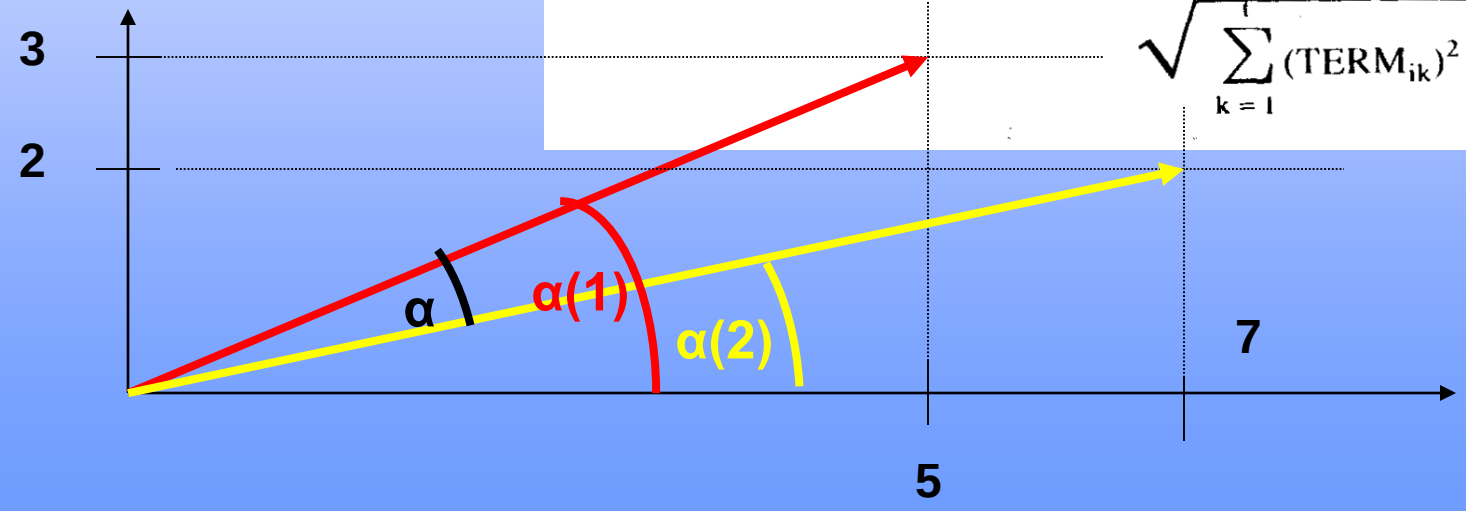
Dok1 (5 | 3); Dok2 (7 | 2); Berechnung nach Cosinus-Formel (Salton-Vorschlag)

Zähler: (5 * 7) + (3 * 2) = 35 + 6 = 41

Nenner: $((5^2 + 3^2) * (7^2 + 2^2))^{1/2} = ((25 + 9) * (49 + 4))^{1/2} = (34*53)^{1/2} = 1.802^{1/2} = 42,45$

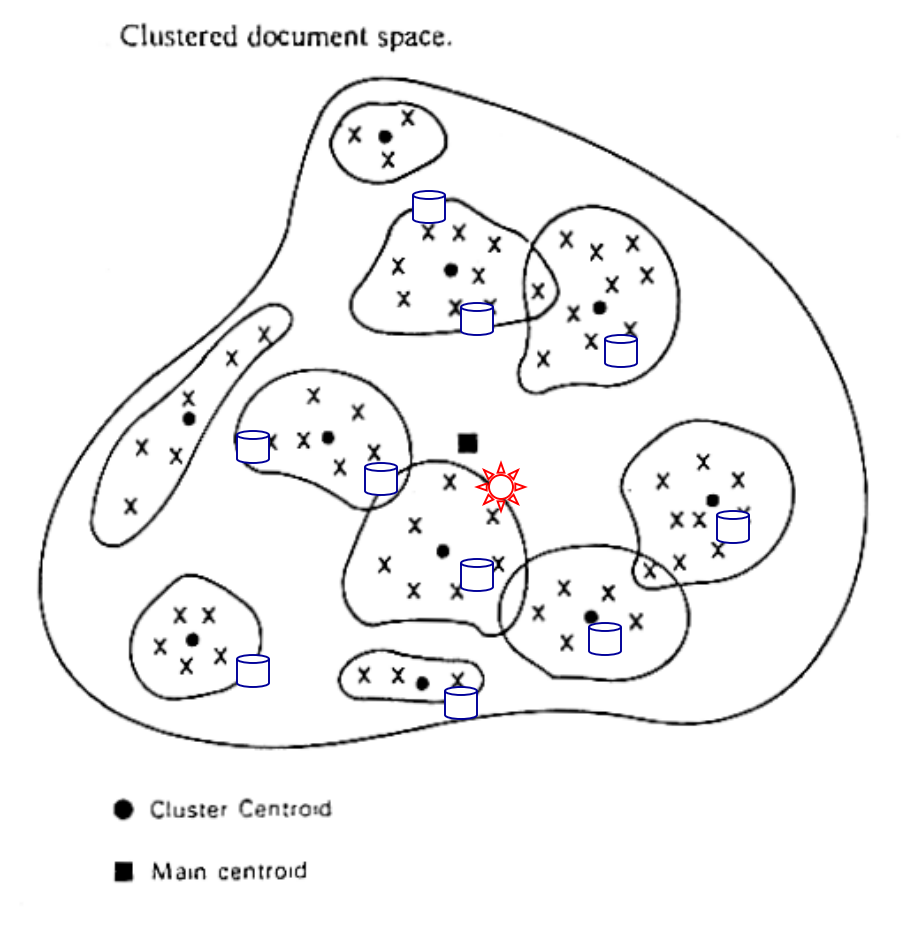
cos α = 41 : 42,45 = 0,97

$$\text{COSINUS}(\text{DOK}_i, \text{ANFRAGE}_j) = \frac{\sum_{k=1}^{t} (\text{TERM}_{ik} \cdot \text{ATERM}_{jk})}{\sqrt{\sum_{k=1}^{t} (\text{TERM}_{ik})^2 \cdot \sum_{k=1}^{t} (\text{ATERM}_{jk})^2}}$$



E.2 Vektorraummodell

Dokumentenraum mit Clustern



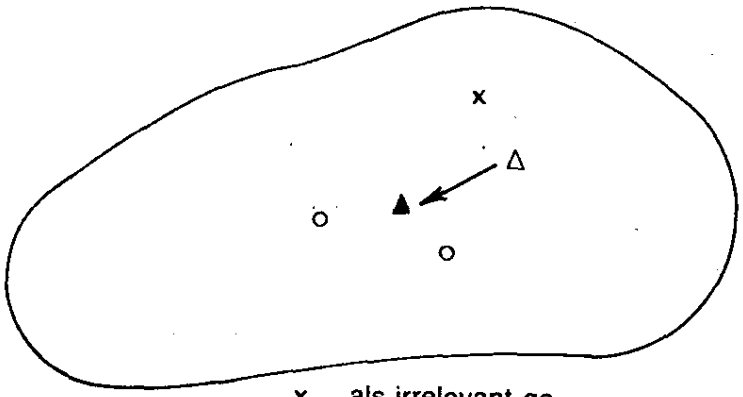
Zentroid: imaginäres Dokument, dessen Dimensionen den arithmetischen Mittelwert der Dokumentwerte beinhalten

Superzentroid: imaginäres Dokument, dessen Dimensionen den arithmetischen Mittelwerten der zugehörigen Zentroid-Vektoren beinhalten

E.2 Vektorraummodell

Relevance Feedback

- **Dimensionen (Worte) der relevanten Dokumente werden der Suchfrage hinzugefügt (oder höher gewichtet)**
- **Dimensionen (Worte) der irrelevanten Dokumente werden aus der ursprünglichen Such-frage entfernt (oder niedriger gewichtet)**
- **iteratives Verfahren: kann mehrfach wiederholt werden**



E.2 Vektorraummodell

Relevance Feedback

- **Verfahren nach Rocchio**

$$\vec{q}_m = \alpha\,\vec{q} + \frac{\beta}{|D_r|}\sum_{\forall \vec{d}_j \in D_r} \vec{d}_j - \frac{\gamma}{|D_n|}\sum_{\forall \vec{d}_j \in D_n} \vec{d}_j$$

- **D(r) : als relevant markierte Dokumente**
- **D(n) : nicht relevante Dokumente**
- **α, β, γ : frei einstellbare Gewichtungsfaktoren**

Rocchio, J.J. (1971). Relevance feedback in information retrieval.
In G. Salton (Ed.), The SMART Retrieval System – Experiments in Automatic Document Processing (pp. 313-323).
Englewood Cliffs, N.J.: Prentice Hall.

E.2 Vektorraummodell

Vektorraummodell – Problem

- **Modell nimmt an, dass die Terme unabhängig voneinander sind**
- **dies ist falsch: Terme sind vielmehr vom Auftreten anderer Terme abhängig – kommen also verstärkt gemeinsam vor (*Bsp.* Synonyme oder Homonyme in unterschiedlichen Kontexten)**
- **Lösungsmöglichkeiten:**
  - **semantisches Vektorraummodell (ab Ebene t7: Begriffe)**
  - **Latent Semantic Indexing (Reduktion der Dimensionen durch Faktorenanalyse)**

Link zur Petition: Change.org/saveiws

Kapitel E.3

Probabilistisches Modell

E.3 Probabilistisches Modell

- **„klassische“ Variante entwickelt von Maron und Kuhns 1960**
- **Ansatz: Wie wahrscheinlich ist es, dass ein Dokument für eine Suchanfrage relevant ist? Ausgabe der Dokumente im Relevance Ranking absteigend nach der Wahrscheinlichkeit**
- **Wahrscheinlichkeitstheorie: P(B |A) „bedingte Wahrscheinlich-keit“: Wie groß ist die Wahrscheinlichkeit, dass B eintritt, wenn A gegeben ist? Hier: A: Anfrage; B: Relevanz eines Dokuments**
- **Bayessches Theorem:**

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Maron, M.E. & Kuhns, J.L. (1960). On relevance, probabilistic indexing and information retrieval. Journal of the ACM, 7, 216-244.

E.3 Probabilistisches Modell

$$P(D \mid Q) = P(Q \mid D) * P(D) \; / \; P(Q)$$

- **P(D | Q) : Retrievalstatuswert von D unter Q**

- **P(Q) = 1**
- **P(D) = TF * IDF**
- **P(Q | D) = ??? Wie relevant sind die Terme Q im Dokument?**
  - **hierzu nötig: Relevanzinformationen**
    - **entweder Relevance Feedback durch Nutzer**
    - **oder Pseudo-Relevance-Feedback durch System**

E.3 Probabilistisches Modell

- **Das Modell geht davon aus, dass die Relevanz der einzelnen Dokumente unabhängig voneinander ist (dies erfordert das Bayessche Theorem)**
- **„Relevanz“ muss unter Ausschluss des Nutzers definiert werden, also ohne „Pertinenz“ (oder „Usefulness“ nach Robertson)**
- **Sonst wäre das Modell nämlich faktisch falsch: Ein Nutzer, der bereits n Dokumente erhalten hat, orientiert sich bei der Relevanzeinschätzung des n+1sten Dokuments an den n vorangegangenen (ein Dokument, das weit oben platziert ist, hat größere Chancen, als relevant eingeschätzt zu werden, als eines, das weiter unten gereiht ist)**

Robertson, S.E. (1977). The probability ranking principle in IR. Journal of Documentation, 33, 294-304.

E.3 Probabilistisches Modell

- **Relevanzinformationen aus Stichprobe. Ableitung von Wahrscheinlichkeitswerten:**
  - **P_1 (Vorkommen von x_{mi} ist wesentlich für die Relevanz des Dokuments) = r / R**
  - **P_2 (Nicht-Vorkommen von x_{mi} ist wesentlich für die Relevanz des Dokuments) = (R – r) / R**
  - **P_3 (Vorkommen von x_{mi} ist wesentlich für die Nicht-Relevanz des Dokuments) = (n – r) / (N – R)**
  - **P_4 (Nicht-Vorkommen von x_{mi} ist wesentlich für die Nicht-Relevanz des Dokuments) = (N – n – R + r) / (N – R)**

    R: # rel. Dok.; r: # rel. Dok. mit Term x; N: # alle Dok; n: # aller Dok. mit Term x; P: Wahrscheinlichkeit
  - **$P_2 = 1 - P_1$**
  - **$P_4 = 1 - P_3$**

E.3 Probabilistisches Modell

	relevant	nicht relevant	
„Miranda“ = 1	r 10	n-r 1	n 11
„Miranda“ = 0	R-r 2	N-R-n+r 7	N-n 9
	R 12	N-R 8	N 20

wichtig: (1.) Verhältnis der relevanten Dok, die den Term enthalten zu den relevanten, die ihn nicht enthalten: r / R-r

(2.) Verhältnis der nicht relevanten Dok, die den Term enthalten zu den nicht relevanten, die ihn nicht enthalten: n-r / N-R-n+r

E.3 Probabilistisches Modell

- **Robertson / Sparck-Jones - Formel**

$$w_i = \log \frac{r_i/(R - r_i)}{(n_i - r_i)/(N - n_i - R + r_i)}$$

$$w_{(\text{„Miranda“})} = \log \frac{10 / (12 - 10)}{(11 - 10) / (20 - 11 - 12 + 10)} = \log \frac{10 / 2}{1 / 7} = \log 35 = 1{,}54$$

Robertson, S.E., & Sparck-Jones, K. (1976). Relevance weighting of search terms. Journal of the American Society für Information Science, 27, 129-146.

E.3 Probabilistisches Modell

	relevant	nicht relevant	
„Miranda“ = 1	r 11	n-r 0	n 11
„Miranda“ = 0	R-r 0	N-R-n+r 9	N-n 9
	R 11	N-R 9	N 20

Problem: Was tun, wenn ein Term in allen relevanten Dokumenten vorkommt und in allen nicht-relevanten nicht vorkommt?

E.3 Probabilistisches Modell

- **Robertson / Sparck-Jones - Formel**

$$w_i = \log \frac{r_i/(R - r_i)}{(n_i - r_i)/(N - n_i - R + r_i)}$$

$$w_{(\text{„Miranda“})} = \log \frac{11 / (11 - 11)}{(11 - 11) / (20 - 11 - 11 + 11)} = \log \frac{11 / 0}{0 / 9} = \text{nicht def.}$$

E.3 Probabilistisches Modell

- **Modifizierte Robertson – Sparck-Jones – Formel (bereinigt um Probleme bei der Division durch Null):**

$$w = \log \frac{(r + 0{,}5) / (R - r + 0{,}5)}{(n - r + 0{,}5) / (N - n - R + r + 0{,}5)}$$

E.3 Probabilistisches Modell

$$w = \log \frac{(r + 0{,}5) / (R - r + 0{,}5)}{(n - r + 0{,}5) / (N - n - R + r + 0{,}5)}$$

$$w(\text{„Miranda“}) = \log \frac{11+0{,}5 / (11-11+0{,}5)}{(11-11+0{,}5) / (20-11-11+11+0{,}5)} = \log \frac{11{,}5 / 0{,}5}{0{,}5 / 9{,}5} = \log \frac{23}{0{,}053}$$

$$= \log 437 = 2{,}64$$

E.3 Probabilistisches Modell

Pseudo-Relevance-Feedback

- **Was tun, wenn keine Relevanzinformationen vorhanden sind?**
- **Durchführen einer Suche nach einem anderen Modell (etwa dem Vektorraummodell oder einfach nach TF*IDF)**
- **Annahme: die top gerankten Dokumente sind relevant (Nutzer wird nicht gefragt)**
- **Problem: Welche sind die top Dokumente? Welche Verteilungsform liegt vor?**
- **Schätzung des Zusammenhangs der Dokumente (mit deren Worten) und der Relevanz: Welche Worte kommen dort vor? „Verdecktes Relevance Feedback-Verfahren“**
- **neue Suche unter Einbezug der Worte und der (nunmehr errechneten) Gewichtungen aus den top gerankten Dokumenten**
- **$w = \log [(r + 0{,}5) / (R - r + 0{,}5)]$**

Croft, W.B., & Harper, D.J. (1979). Using probabilistic models of document retrieval without relevance information. Journal of Documentation, 35, 285-295.

E.3 Probabilistisches Modell

Vorgehen nach dem Berechnen des w-Wertes

- **Formel: G(t,d) = w * TF(t,d) * IDF(t) (t: Term; d: Dokument)**
- **(zur Erinnerung: P(D |Q) = P(Q |D) * P(D) / P(Q))**

- **für alle ursprünglichen Query-Terme und für neugewonnene Terme aus den Dokumenten**
  - **etwa: alle aus denjenigen Sätzen (oder Absätzen), in denen die Query-Terme vorkommen**
  - **oder: alle aus einem frei wählbaren Textfenster um die Query-Terme herum (z.B. alle im Abstand bis zu 25 Termen)**
- **Berechnung des Retrievalstatuswertes e: Aggregation (i.d.R.: Summe) der Gewichtungswerte aller (neuen wie alten) Query-Terme**

Kapitel E.4

Retrieval nicht-textueller Dokumente

E.4 Retrieval nicht-textueller Dokumente

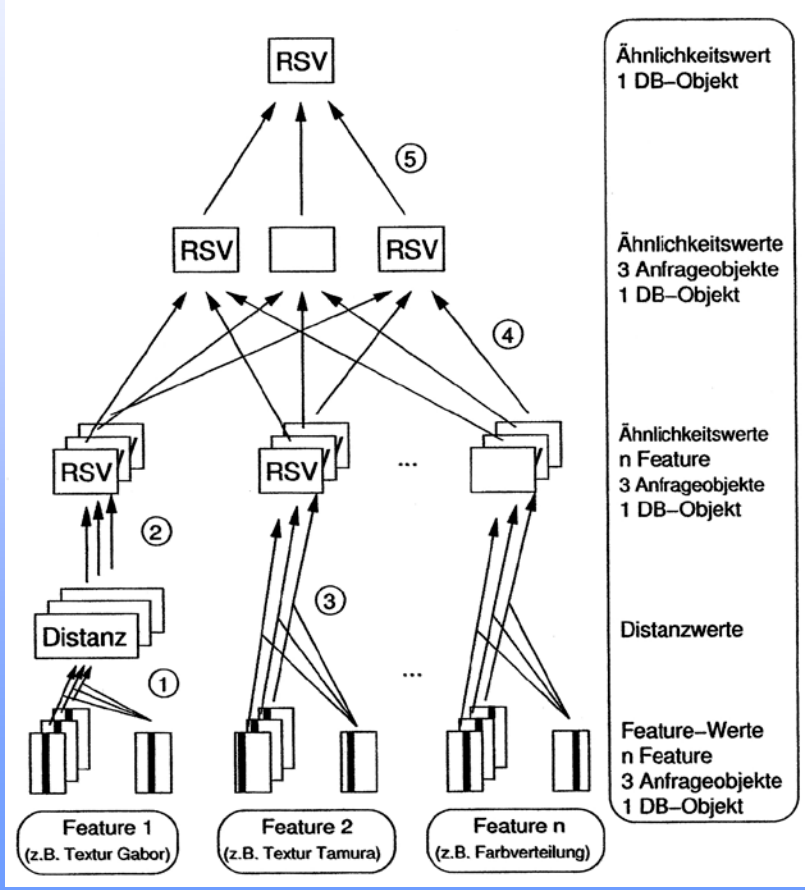
Multimedia Information Retrieval

- **Content-based Information Retrieval**
  - **gesprochene Sprache**
  - **Musik und weitere Audio-Dokumente (z.B. Geräusche)**
  - **Bilder**
  - **Videos (Bewegtbilder plus Audio-Elemente)**

Schmitt, I. (2006). Ähnlichkeitssuche in Multimedia-Datenbanken. Retrieval, Suchalgorithmen und Anfragebehandlung. München, Wien: Oldenbourg.

E.4 Retrieval nicht-textueller Dokumente

Ranking nach Content über spezifische Dimensionen



E.4 Retrieval nicht-textueller Dokumente

Aufgaben

- **Aufspüren der jeweils bestpassenden Dimensionen**
- **Finden von Metriken zum Berechnen der Ähnlichkeiten auf der jeweiligen Dimension**
- **Errechnen der Distanzen zwischen Anfrage- und Datenbankobjekt anhand der gewählten Metrik**
- **Ableiten eines dimensionsspezifischen Retrievalstatuswertes für jede Dimension**
- **Kombination der dimensionsspezifischen Retrievalstatuswerte zu einem Retrievalstatuswert für das gesamte Objekt (Bild-, Ton- oder Videodokument)**

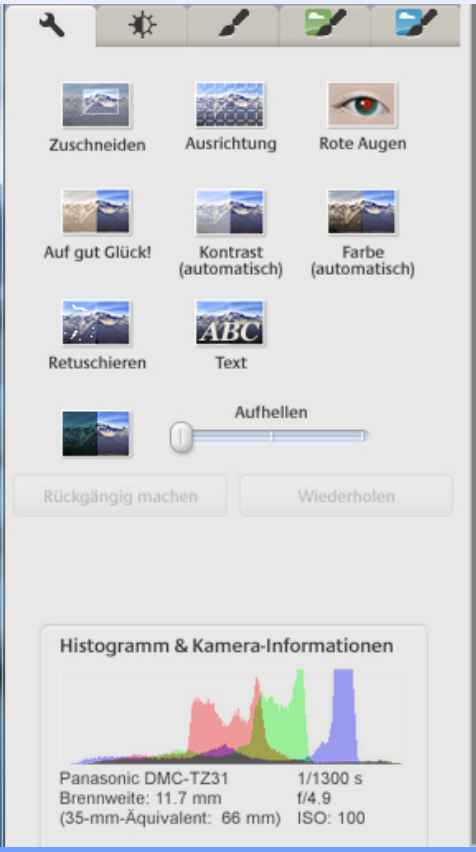
E.4 Retrieval nicht-textueller Dokumente

Dimensionen des Bildretrieval

- **Farbe**
  - **Spektralwerte (elektromagnetische Wellen zwischen 380 und 780 nm)**
  - **Farbräume**
    - **RGB: rot, grün, blau; Anwendung: Fernsehen, Displays**
    - **CMY: cyan, magenta, yellow; Anwendung: Farbdrucker**
    - **HSI: hue (Farbton), saturation (Farbsättigung), intensity (Farbintensität)**
  - **Farbindexierung**
    - **1. Schritt: Zuordnung der Farbinformation zu jedem Pixel**
    - **2. Schritt: Erstellung von Histogrammen**
    - **3. Schritt: Gewinnung beleuchtungsinvarianter Farbwerte**

E.4 Retrieval nicht-textueller Dokumente

Farbhistogramm



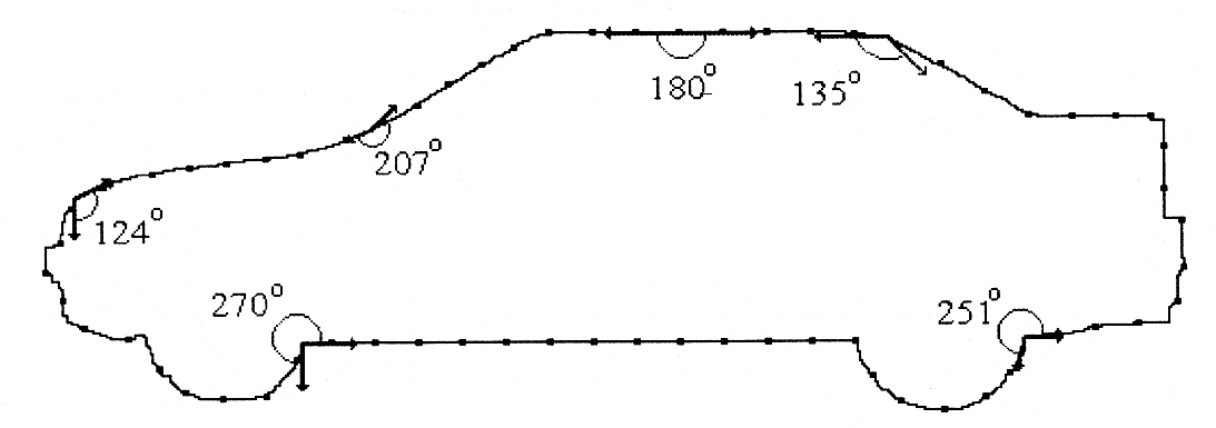
E.4 Retrieval nicht-textueller Dokumente

Dimensionen des Bildretrieval

- **Textur**
  - **1.: innerhalb Textur: homogene Variationen der Intensitäten von benachbarten Pixeln; außerhalb Textur: nicht homogen**
  - **2.: homogene Eigenschaft über das gesamte Bild hinweg**
  - **Texturindexierung**
    - **Gaborfilter: Ermittlung von Grauwert- oder Farbänderungen**
    - **Tamurafilter: Granularität, Kontrast, Richtung, Linienähnlichkeit, Regularität (Zerlegung eines Bildes in Teilbilder), Grobheit (Granularität *und* Kontrast)**
- **Gestalt**
  - **Normalisierung der Position**
    - **Rand, Bildmitte**
    - **Größe (Vordergrund, Hintergrund)**
    - **Blickwinkel (rotationsunabhängige Sicht auf die Gestalt)**

E.4 Retrieval nicht-textueller Dokumente

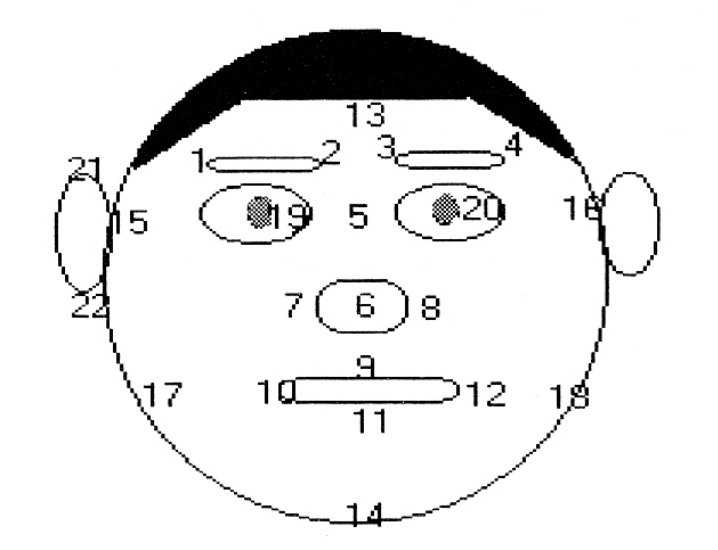
Normalisierung mittels Winkelberechnung bei Eckpunkten



Wang, C.S., & Shih, T.K. (2004). An efficient content-based retrieval system for large image database. In S. Deb (Ed.), Multimedia Systems and Content-Based Image Retrieval (pp. 249-276). Hershey, PA: Idea Group Publ.

E.4 Retrieval nicht-textueller Dokumente

Dimensionen der Gestalt bei der Erkennung von Gesichtern

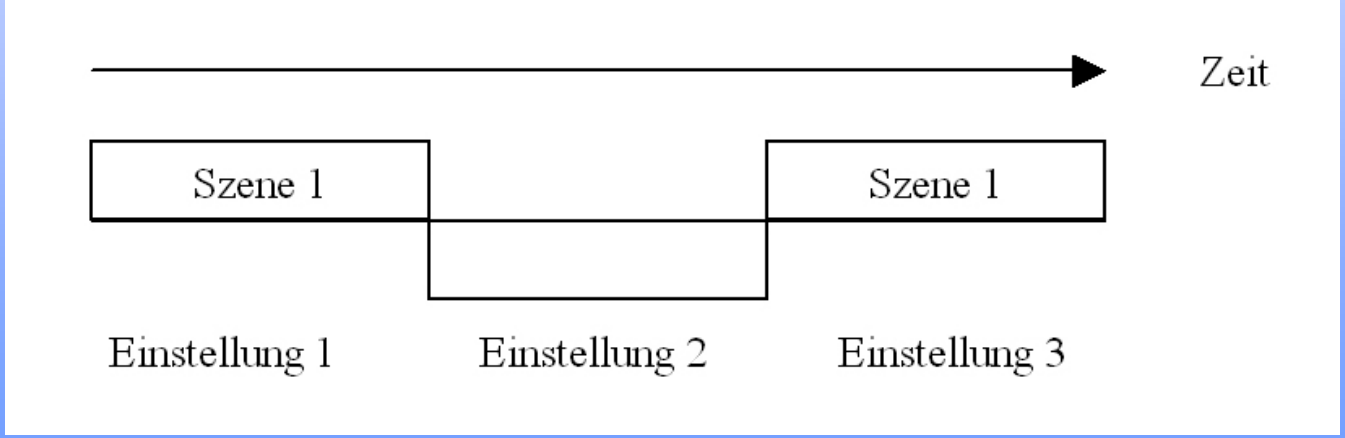


Singh, S.K., Vatsa, M., Singh, R., Shukla, K.K., & Boregowda, L.R. (2004). Face recognition technology: A biometric solution to security problems. In S. Deb (Ed.), Multimedia Systems and Content-Based Image Retrieval (pp. 62-99). Hershey, PA: Idea Group Publ.

E.4 Retrieval nicht-textueller Dokumente

Videoretrieval

- **dokumentarische Bezugseinheit**
  - **Einzelbild (insbesondere "key frames")**
  - **Sequenz: Einstellung (technische Sicht)**
  - **Sequenz: Szene (inhaltliche Sicht)**



E.4 Retrieval nicht-textueller Dokumente

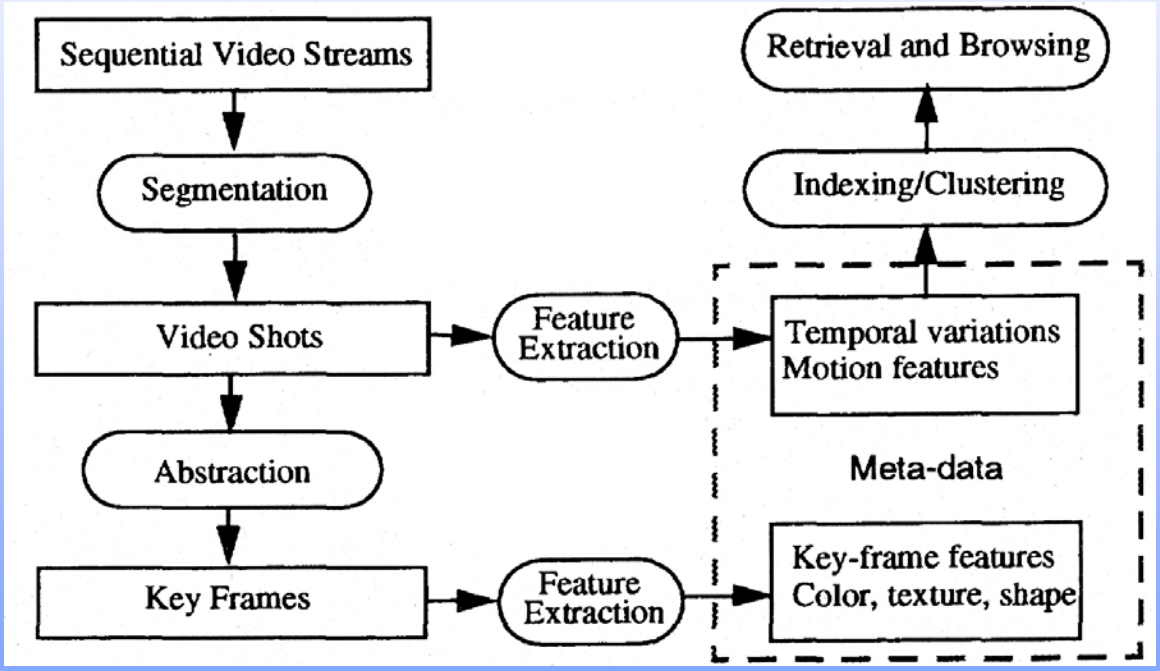
Sequenz

- **Segmentierung**
  - **harter Schnitt: problemlos erkennbar**
  - **weicher Übergang (z.B. Überblenden): Vergleiche vor und hinter Zeitfenster (ca. 4 sec)**
  - **Zusatzproblem: Auseinanderklaffen von Bild und Ton**

- **Bewegung**
  - **Kamerabewegung (Schwenken, Zoomen)**
  - **"echte" Bewegung**
    - **relative Konstanz einer Gestalt**
    - **dadurch: Erkennen der Gestalt**

E.4 Retrieval nicht-textueller Dokumente

Videoretrieval



Zhang, H.J., Wu, J., Zhong, D., & Smoliar, S.W. (1997). An integrated system for content-based video retrieval and browsing. Pattern Recognition, 30, 643-658.

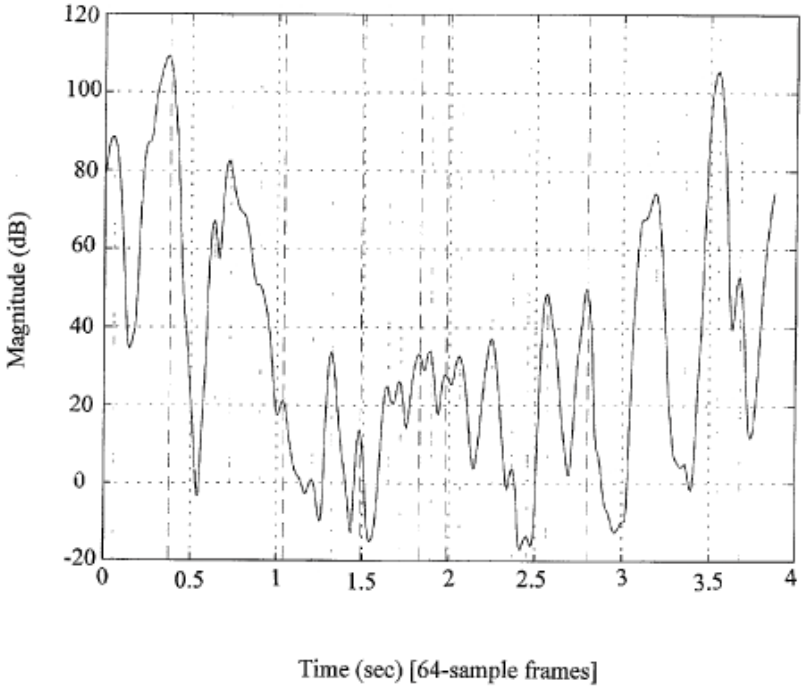
E.4 Retrieval nicht-textueller Dokumente

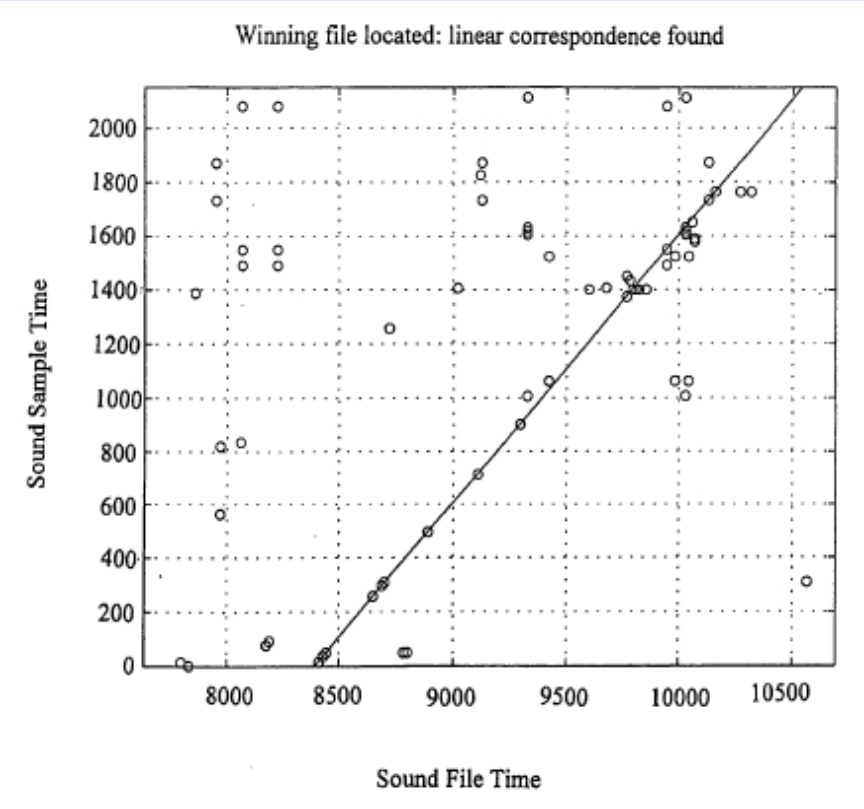
Dimensionen des Musikretrieval

- **Tonhöhe**
- **Dauer / Rhythmus**
- **Melodie**
- **Harmonie (Polyphonie)**
- **Klangfarbe (Instrumente)**

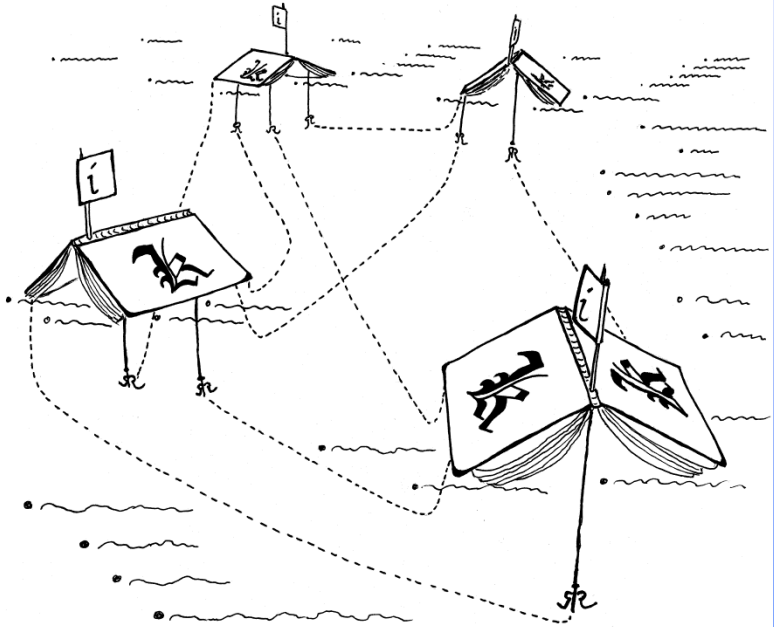
E.4 Retrieval nicht-textueller Dokumente

Anwendungsfall: Query by Humming (z.B. Shazam)





Teil F: Web Information Retrieval



Kapitel F.1

Linktopologie

F.1 Linktopologie

	Oberflächenweb IR	*Deep Web IR*
Dokumente		
Sprachen	**viele Sprachen**	**i.d.R. einheitliche Indexierungssprache**
Formate	**viele Formate**	**i.d.R. ein Format**
Länge	**unterschiedlich**	**bei bibliographischen Datenbanken: in etwa gleich**
Teile	**unterschiedliche Teile (Bilder, Sprungmarken, ...)**	**genau ein Datensatz**
Verlinkung	**Hyperlinks**	**ggf. Referenzen und Zitationen**
Spam	**ja**	**nein**
Struktur	**schwach**	**Feldstruktur**
Inhalt	**heterogen**	**homogen**

Lewandowski, D. (2005). Web Information Retrieval. Technologien zur Informationssuche im Internet. Frankfurt: DGI. (DGI-Schrift Informationswissenschaft; 7).

F.1 Linktopologie

Grundgesamtheit		
Größe	**Webgröße unbekannt**	**(in etwa) bekannt**
Abdeckung	**nicht messbar**	**(in etwa) messbar**
Duplikate	**ja: Spiegel**	**nein**
Nutzer		
Zielgruppe	**alle Web-User**	**i.d.R. Fachexperten**
Bedarf	**unterschiedlich**	**fachbezogen**
Kenntnisse	**gering**	**professionelle Endnutzer: hoch; Information Professionals: sehr hoch**
IR-System		
Interface	**einfach**	**oft (sehr) komplex**
Funktionalität	**gering**	**hoch**
Relevance Ranking	**ja**	**nein (ggf. zusätzlich angeboten)**
Sortierung nach Datum	**nein**	**ja (FILO)**

F.1 Linktopologie

- **Algorithmische Web-Suchmaschinen**
  - **Erste Generation**
    - **vor allem Textstatistik, Vektorraummodell und probabilistisches Modell**
    - **scheitert an der negativen Kreativität der Spammer**
  - **Zweite Generation**
    - **anfrageabhängige Rankingfaktoren**
      - **PWDF*IDF**
      - **Wortabstand**
      - **Reihenfolge der Suchatome**
      - **Strukturinformationen in Dokumenten (z.B. Größe des Fonts)**
      - **Ankertexte (die auf ein Dokument verweisen)**
      - **Sprache des Nutzers**
      - **räumliche Nähe zum Nutzer**

F.1 Linktopologie

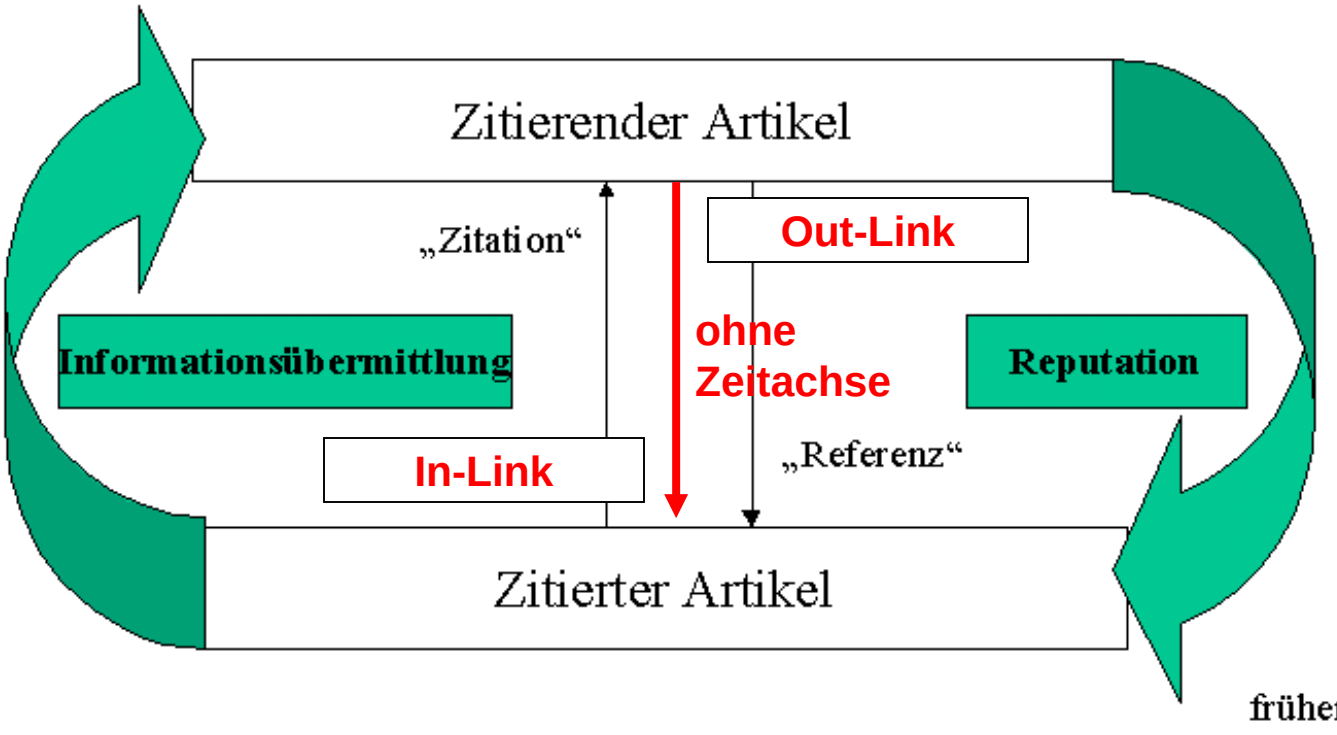
- **anfrageunabhängige Rankingfaktoren**
  - **eingehende und ausgehende Links (Linktopologie)**
  - **Stellung der Webseite innerhalb eines Graphen (Zentralitätsmaße)**
  - **Stellung der Webseite in der Verzeichnishierarchie**
  - **Klick-Popularität**
  - **Aktualität**

F.1 Linktopologie

- **Grundthese der Linktopologie: Die Linkstruktur in einer Hypertext-umgebung (wie dem WWW) sagt etwas über die Relevanz eines Dokuments aus**
  - **zwei grundlegende Ansätze**
    - **Hubs und Authorities – Kleinberg-Algorithmus**
      (Aspekte davon früher eingesetzt bei: AltaVista und Teoma)
    - **PageRank – Algorithmus von Brin und Page**
      (eingesetzt bei: Google)
  - **theoretische Basis: Zitationsanalyse (Garfield)**
    - **Bibliographic Coupling**
    - **Co-Citations**

F.1 Linktopologie

Zitationen und Referenzen:
Analogie: eingehende und ausgehende Links



F.1 Linktopologie

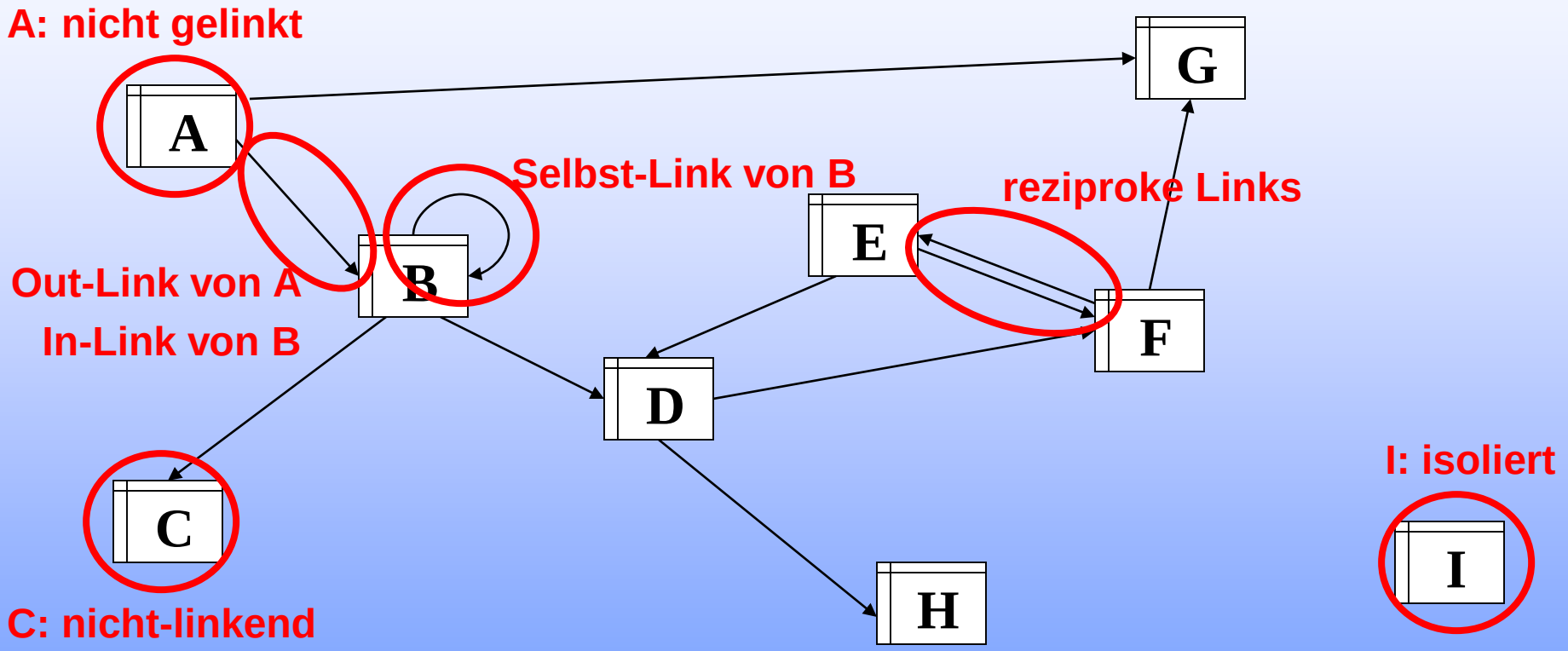
- **Theoretische Basis: Zitationsanalyse (Garfield)**

Garfield, E. (1979). Citation Indexing. New York, NY: Wiley.

- **Sind Web-Links wirklich analog zu Zitationen und Referenzen?**
- **Stichprobe: rund 10% aller Links sind gleich wie akademische Zitationen; weitere rund 20% können zumindest als analog gelten**
- **D.h.: ca. 70% aller Links sind nicht wie Zitationen benutzt, sondern dienen der Navigation, sind Linklisten „verwandter Themen“ oder Werbung**
- **also: Es ist Vorsicht geboten, Ergebnisse der Zitationsanalysen auf die Link-Topologie unkritisch zu übertragen.**

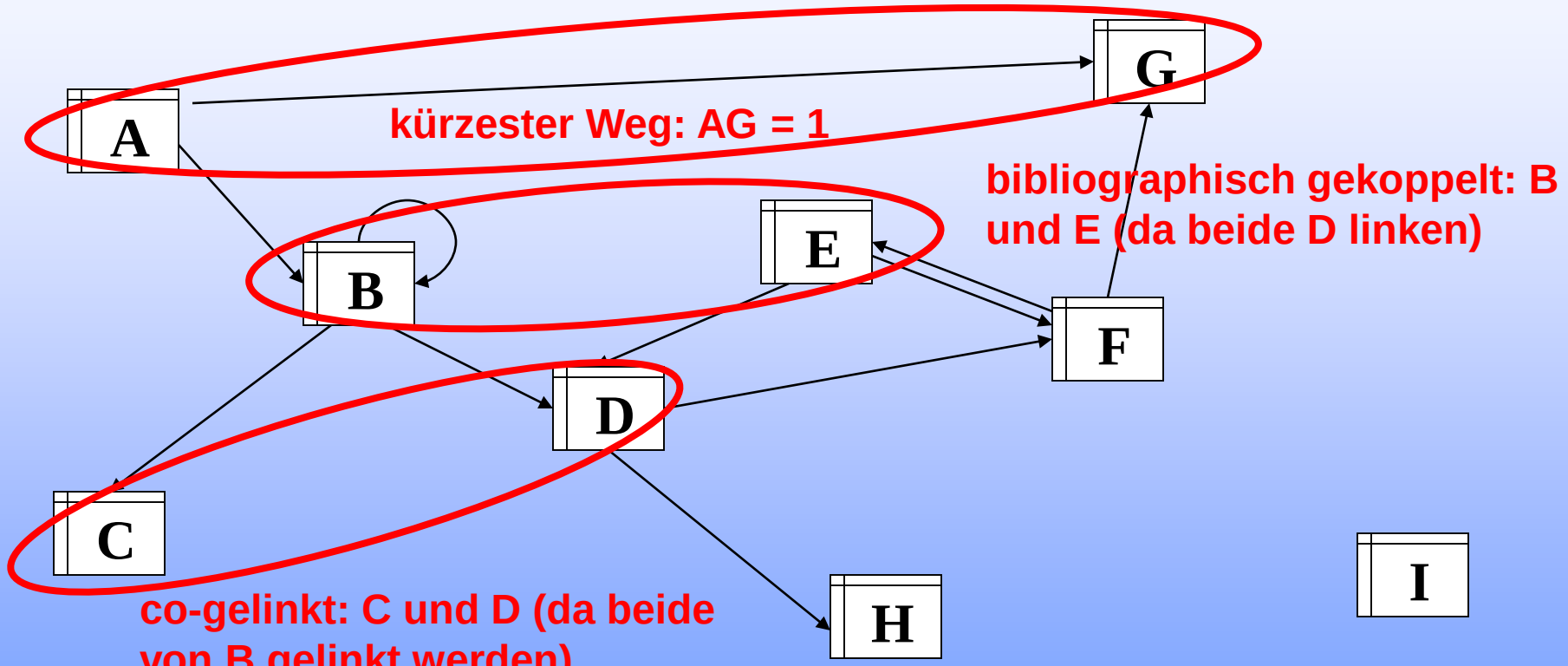
Smith, A.S. (2004). Web links as analogues of citations. Information Research, 9(4), paper 188.

F.1 Linktopologie



Björneborn, L., & Ingwersen, P. (2004). Towards a basic framework for webometrics. Journal of the American Society for Information Science and Technology, 55, 1216-1227.

F.1 Linktopologie



F.1 Linktopologie

Hubs und Authorities. Der Kleinberg-Algorithmus

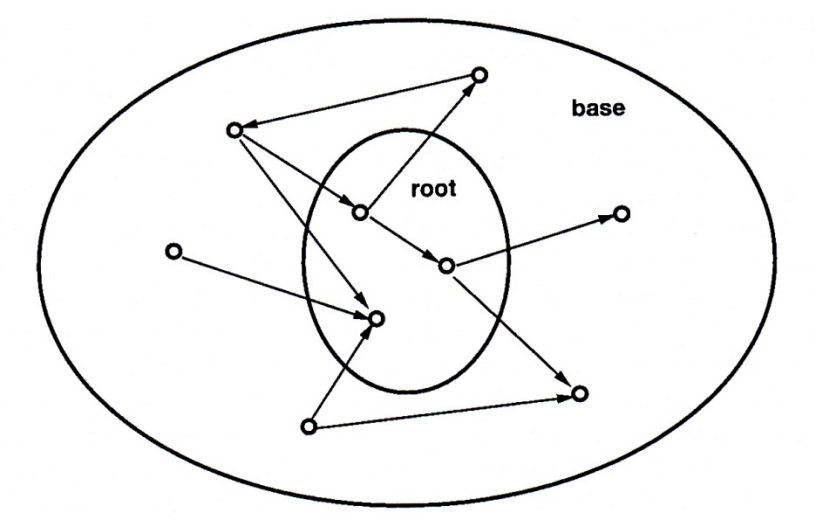
HITS (Hyperlink-Induced Topic Search)

- **„hub“ (Radnabe; Mittelpunkt): ein Dokument, von dem viele Links ausgehen**
- **„authority“: ein Dokument, auf das viele andere Dokumente (aus anderen Websites) linken**
- **„good hub“: hub, der auf viele gute authorities linkt (das „good“ ist ein Unterschied zur Zitatenanalyse, in der alle Zitierungen gleich sind)**
- **„good authority“: authority, die von vielen guten hubs gelinkt wird**
- **Ziel: kleine Menge einer „community“ von hubs und authorities: „high-quality pages“**

Kleinberg, J.M. (1999). Authoritative sources in a hyperlinked environment. Journal of the ACM, 46, 604-632.

F.1 Linktopologie

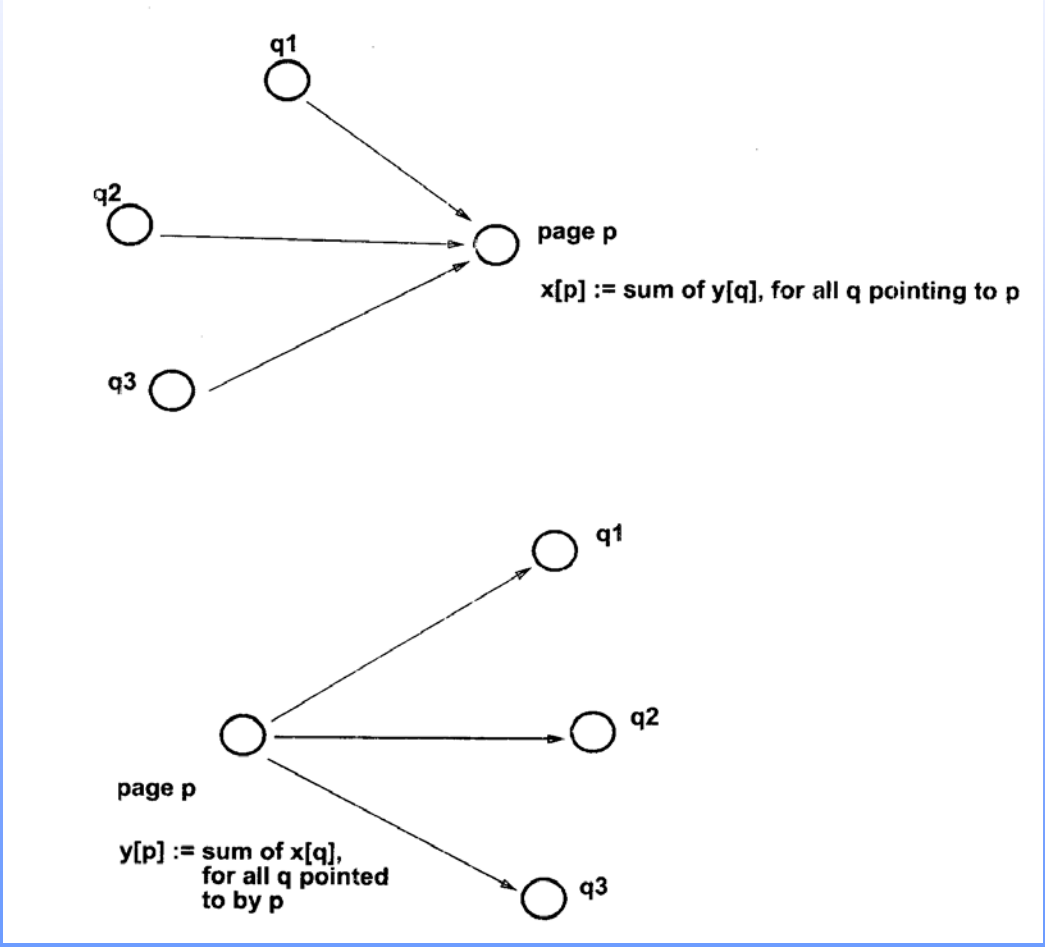
- **Pseudo-Relevance-Feedback: initiale Suche**
  - **Suchergebnis: "root set" (max. 200 Dokumente)**
  - **Hinzufügen von Seiten, die auf ein Dokument im root set linken (max. 50 pro Dokument)**
  - **Hinzufügen aller Seiten, auf die ein Dokument im root set linkt**



- **für "base set": Berechnung nach Kleinberg-Algorithmus**

F.1 Linktopologie

Hubs und Authorities. Der Kleinberg-Algorithmus



„Zitationen“ (eingehende Links): Zählbasis für authorities (x)

„Referenzen“ (ausgehende Links): Zählbasis für hubs (y)

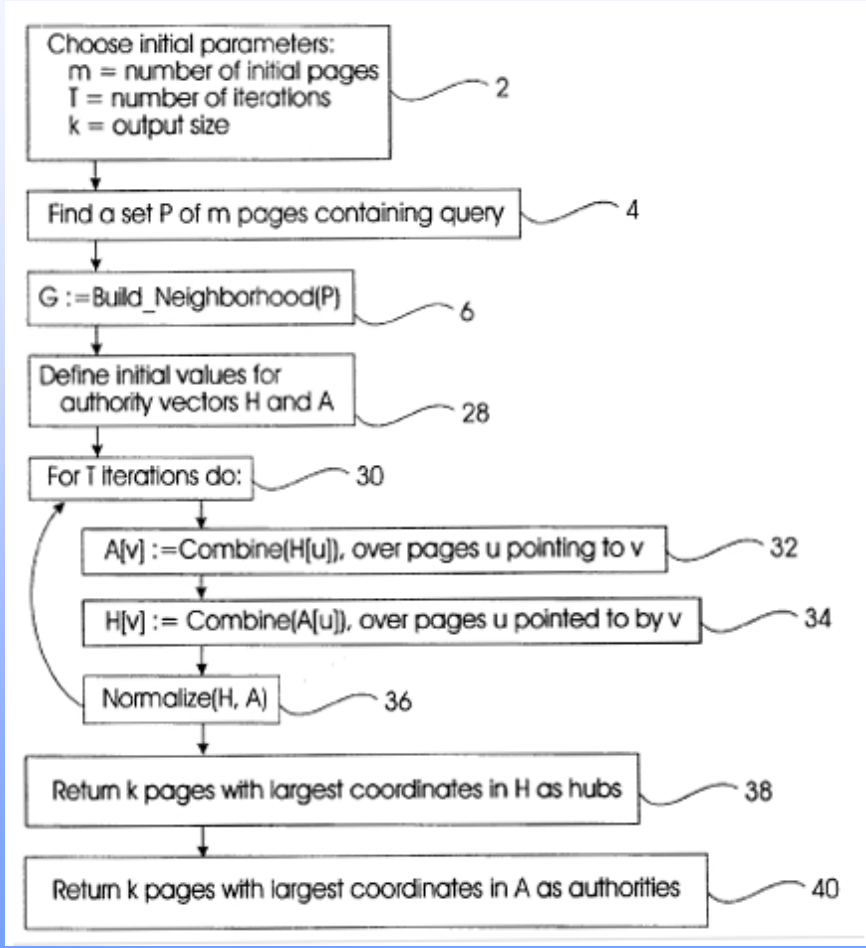
F.1 Linktopologie

- **Werte für q (eingehende Links) und p (ausgehende Links) werden für alle Dokumente einer initialen Treffermenge gezählt**
- **Authority-Gewicht:** $A(p) = x(p) = q_1 + q_1 + \ldots + q_n$
- **Hub-Gewicht:** $H(p) = y(p) = q_1 + q_1 + \ldots + q_n$
- **erste Berechnungsrunde: alle** $q_i = 1$
- **Abbildung der Gewichtungswerte auf das Intervall [0, .., 1], d.h. Multiplikation von jedem Wert mit:**

$$1 \;/\; (q_1^2 + q_2^2 + \ldots + q_n^2)^{1/2}$$

- **iteratives Vorgehen**
- **Abbruch, wenn die entstehenden (x,y)-Paare einen Grenzwert (x*, y*) erreicht haben (ca. 20 Runden)**
- **die jeweils ersten zehn (nach Hub bzw. Authority sortierten) Seiten sind der "Kern" der thematischen Community.**

F.1 Linktopologie



Kleinberg, J. (1997). Method and system for identifying authoritative information resources in an environment with content-based links between information resources. Patent-Nr. US 6.112.202.

F.1 Linktopologie

PageRank

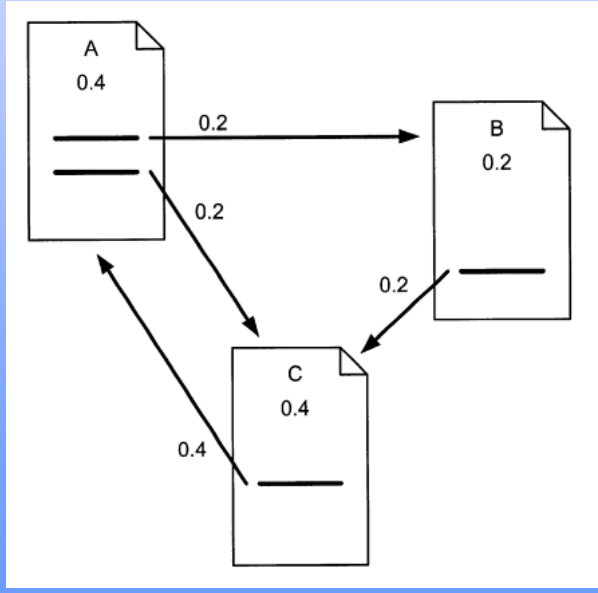
- **das WWW wird als „citation graph“ (link graph) aufgefasst**
- **nach der Stellung in diesem Graphen bekommt jede Webseite ihr spezifisches Gewicht: den PageRank (nach Larry *Page*; Wort-spiel mit Web-*Page*)**
- **intuitive Begründung: „Zufallssurfer“ („random surfer“) klickt ausgehend von einer zufällig gewählten Seite durch das Web; er nutzt nie den Back-Button, kann aber irgendwann erneut eine Seite nach Zufall auswählen. Die Wahrscheinlichkeit, dass der Zufallssurfer eine Webseite besucht, ist deren PageRank.**

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30, 107-117.
Page, L. (1998). Method for node ranking in a linked database. Patent Nr. US 6,285,999.

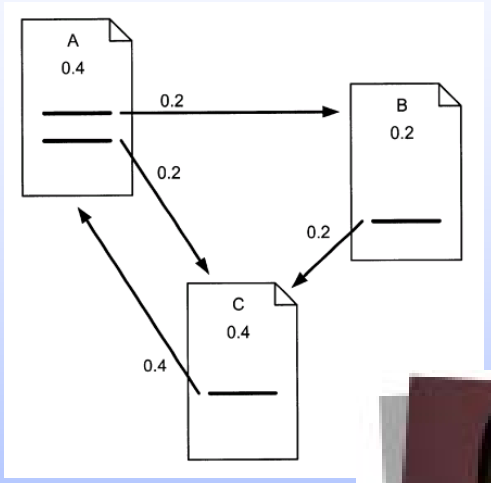
F.1 Linktopologie

- **(1) eine Webseite hat einen hohen PageRank *r*, wenn viele andere Seiten auf sie linken (analog zur Zitationsrate bei Garfield)**
- **(2) eine Webseite hat einen hohen PageRank *r*, wenn Seiten mit ihrerseits hohem PageRank auf sie linken (keine Analogie zur Zitationsrate)**

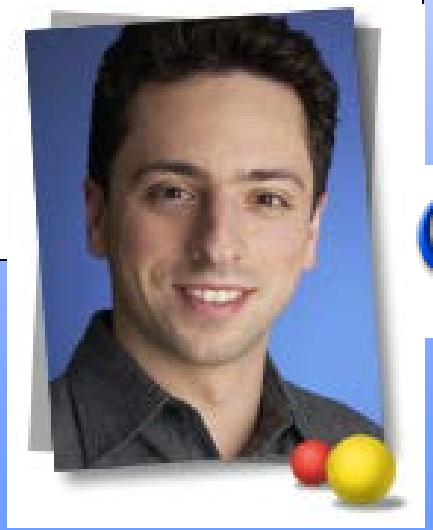
Pfeile: Links
Schritt 1. r(A) = r(C) (C linkt A und sonst nichts)
Schritt 2. r(B) = r(A) / 2 (A linkt B und C, also insgesamt 2 Seiten)
Schritt 3. R(C) = r(B) + r(A)/2
Wahrscheinlichkeit des Zufallssurfers, auf A, B oder C zu stoßen, ist 1, also: r(A) + r(B) + r(C) = 1



F.1 Linktopologie

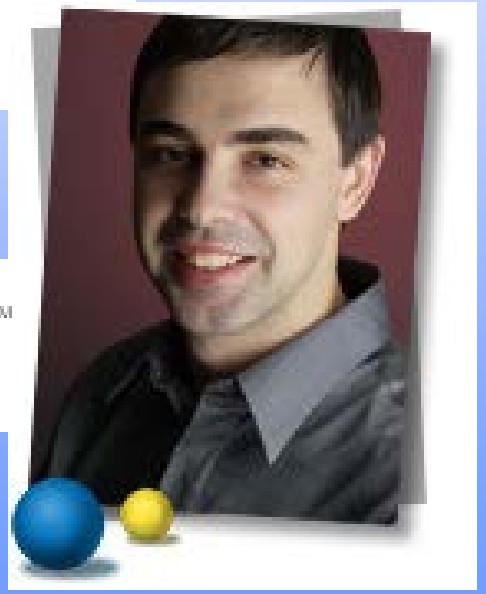


r(A) + r(B) + r(C) = 1
da r(A) = r(C), folgt:
2 r(C) + r(B) = 1
da r(B) = r(A)/2 = r(C)/2, folgt:
2 r(C) + ½ r(C) = 1
5/2 r(C) = 1
r(C) = 2/5 = 0,4
also: r(A) = 0,4
r(B) = 0,2



Sergey Brin

Google™



Larry Page

F.1 Linktopologie

Berechnungsformel für PageRank

$$PR(A) = (1 - d) + d\,[PR(T_1) / C(T_1) + \ldots + PR(T_n) / C(T_n)]$$

PR(A) : PageRank der Webseite A

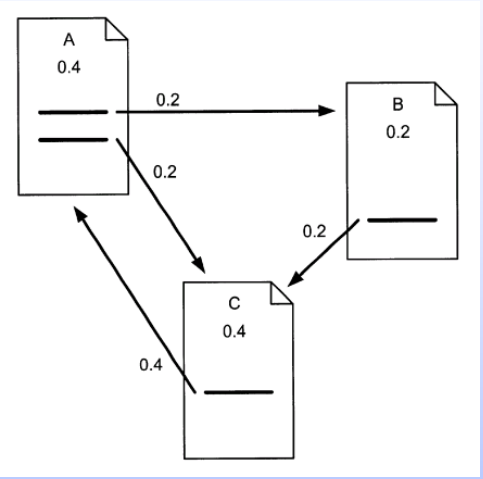
d : Faktor (Wahrscheinlichkeit, mit der ein „Zufallsnutzer“ die Links einer Seite weiterverfolgt) – Wert im Intervall [0,1] - derzeit wird bei Google mit d = 0,85 gearbeitet

T : Dokumente, die einen Link auf A enthalten

C : Anzahl der Links, die von einer Seite T ausgehen

Von der Berechnung ausgenommen werden Seiten ohne Out-Links ("dangling links").

F.1 Linktopologie



- **iterative Berechnung**
- **zunächst: alle PR = 1**
- **PR(A) = 0,15 + 0,85*(1/1) = 1**
- **PR(B) = 0,15 + 0,85*(1/2) = 0,575**
- **PR(C) = 0,15 + 0,85*(1/1 + 1/2) = 1,425**
- **2. Iterationsrunde: Werte einsetzen**
- **PR(A) = 0,15 + 0,85*(1,425/1) = 1,36125**
- **PR(B) = 0,15 + 0,85*(1/2) = 0,575**
- **PR(C) = 0,15 + 0,85*(0,575/1 + 1/2) = 1,06375**
- **3. Iterationsrunde:**
- **PR(A) = 0,15 + 0,85*(1,06375/1) = 1,0541875**
- **PR(B) = 0,15 + 0,85*(1,36125/2) = 0,72853125**
- **PR(C) = 0,15 + 0,85*(0,575/1 + 1,36125/2) = 1,21728125**

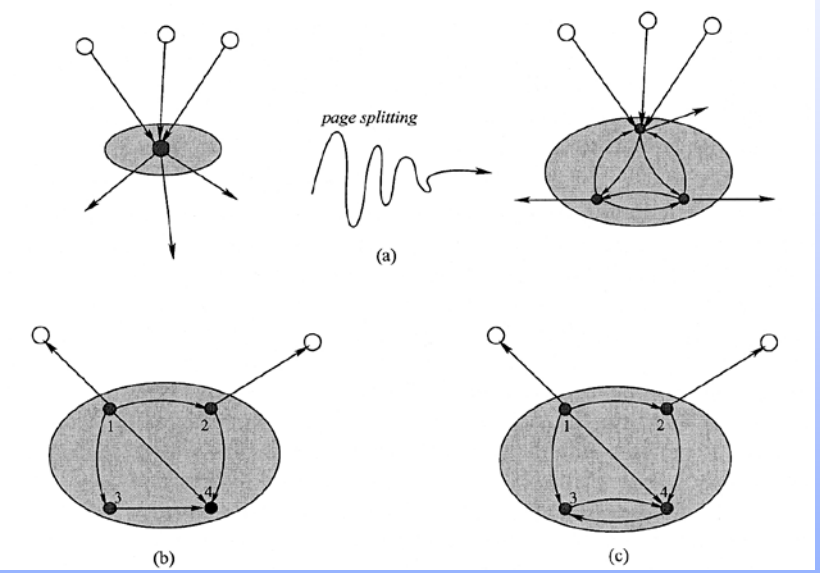
F.1 Linktopologie

- **weitere Iterationsrunden ...**
- **...**
- **12. Iterationsrunde:**
- **PR(A) = 0,15 + 0,85*(1,1927/1) = 1,163795** **~ 1,16**
- **PR(B) = 0,15 + 0,85*(1,1616/2) = 0,64368** **~ 0,64**
- **PR(C) = 0,15 + 0,85*(0,6459/1 + 1,1616/2) = 1,192695** **~ 1,19**

- **nach dem Page-Patent erreicht man nach 100 Iterationsrunden zufriedenstellende Ergebnisse**

F.1 Linktopologie

- **"Goldene Regeln" für Webmaster**
- **1. Inhalt einer größeren Seite auf mehrere kleinere Seiten aufteilen (Fall a)**
- **2. Sackgassen so platzieren, dass an der selben Stelle viele andere (interne) Links stehen (b)**
- **3. Externe Links (die ja PageRank "abgeben") so platzieren, dass an der selben Stelle viele interne Links abzweigen (c)**



Bianchini, M., Gori, M., & Scarselli, F. (2005). Inside PageRank.
ACM Transactions on Internet Technology, 5(1), 92-128.

Kapitel F.2

Rankingfaktoren

F.2 Rankingfaktoren

Learning to Rank

- **Aufgabe: Finden der „richtigen“ Faktoren beim Sortieren für das Relevance Ranking**
- **Bei Suchmaschinen u.a.: TF*IDF, Vektorraum, prob. Modell, PageRank, Aktualität, Pfadlänge usw.**
- **Bei Microblogging-Services u.a.: Anzahl der Follower eines Nutzers, Anzahl der Re-Tweets, Aktualität usw.**
- **Bei Social Network Services (z.B. Facebooks „EdgeRank“): Präferenzen des lesenden Nutzers, Präferenzen des schreibenden Nutzers, Gewicht (je nach „Edge“: Text, Bild, Video, …), Aktualität**
- **Wenn Evaluationsergebnisse (wie bei TReC) vorliegen, kann eine „richtige“ Kombination von Faktoren automatisiert ermittelt werden**

F.2 Rankingfaktoren

Erkennung von Strukturinformationen in Web-Dokumenten

- **Aufgaben:**
  - **Kreation von Metadaten**
    - **optimal: Eintragen von (normiertem) Vokabular in Felder**
  - **Kreation von Kriterien für Relevance Ranking**
    - **optimal: Markieren von Themen (aboutness) und deren Wichtigkeit im Dokument**

F.2 Rankingfaktoren

Regelgeleitete Metadatenkreation

- **Erkennen einer feldspezifischen Information (z.B. Autor)**
  - **falls By-Line vorhanden: Werte hinter "by_"**
  - **falls Autorenname hinter Titelzeile: Werte bis zum nächsten CR**
  - **usw.**
- **Erkennen der Werte (z.B. Separation der unterschiedlichen Autoren)**
  - **falls "und" vorhanden: Trenner (aber: "Thurn und Taxis")**
  - **falls ",_&" vorhanden: Trenner**
  - **usw.**
- **Erkennen der Begriffe (z.B. Reihenfolge von Vor- und Nachnamen)**
  - **in normierte Reihenfolge bringen**
  - **Homonyme trennen**
  - **Synonyme zusammenführen**
- **Vorzugsbenennung des erkannten Begriffs speichern**

F.2 Rankingfaktoren

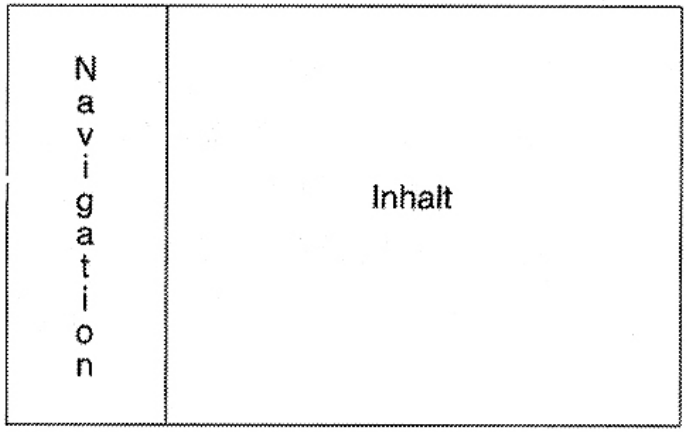
Trennung von Aboutness und Rest

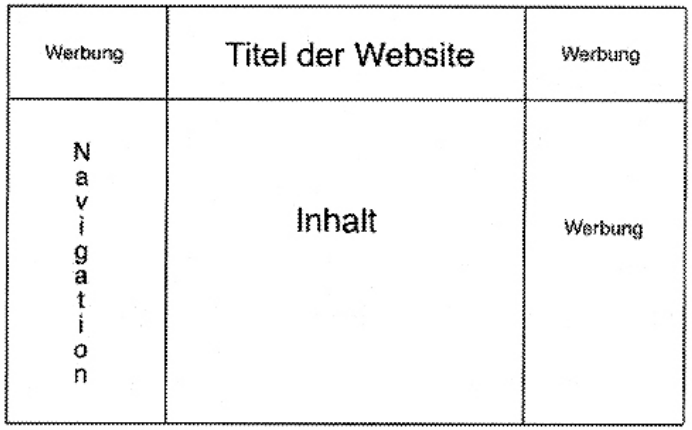
- **Ziel: Inhaltserschließung**
  - **nur der Inhaltsteil eines Dokuments eignet sich für textstatistische Verfahren**
  - **entfernen: reine Navigationstexte (einschließlich Werbung)**
  - **entfernen: formalbibliographische Texte ("Isness") - sofern über Metadatenkreation erkannt (z.B. Entfernen des Autornamens)**

F.2 Rankingfaktoren

Entfernen von Navigationstexten

- **Tabelle links (wahrscheinlich Navigation)**
  - **Seite in Ordnerhierarchie ganz oben: Text der Aboutness zuordnen**
  - **Seite tiefer in Ordnerhierarchie und Inhalt ähnlich mit Top-Seite: Text nicht erschließen**
- **Tabelle rechts (wahrscheinlich Werbung: nicht erschließen) - Risiko!**
- **Tabelle in der Mitte (sehr wahrscheinlich Inhalt: erschließen)**





F.2 Rankingfaktoren

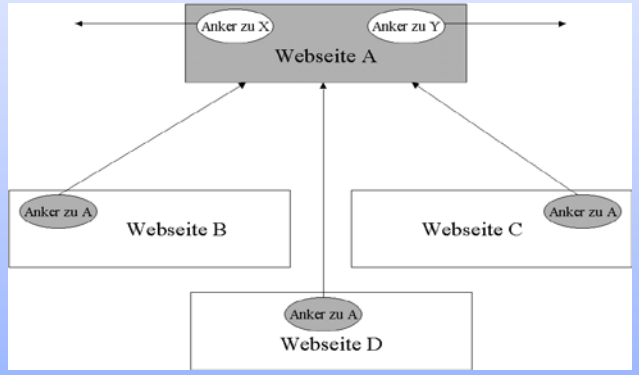
Nutzen von Layout-Informationen für Relevance Ranking

- **<h1> bis <h6>: Überschriften-Hierarchie**
  - **alle <h> Texte höher als Body-Text bewerten**
  - **<h1> höher als <h2> bewerten; usw.**
- **<b>, <i>: fett / kursiv ausgezeichnete Textteile höher bewerten**
- **Schriftgröße (oder Angabe: größer/kleiner als Standard): je größer, desto höher bewerten**
- **Zeilenumbrüche (und damit Absätze) erfassen; Text im ersten Absatz (ggf. auch im letzten) höher bewerten - Nutzen ist nur für News belegt**
- **URL: bei "sprechender" URL: Texte höher bewerten**
- **title-tag (sowie Meta-Tags wie description oder keywords): wegen häufigen Missbrauchs ggf. nicht höher bewerten**
- **table-tag (nicht bei Navigation): Text ggf. höher bewerten**

F.2 Rankingfaktoren

Anker

- **Karneval in Köln**
  Text "Karneval in Köln" wird (auch) der dokumentarischen Bezugseinheit xyz.de zugeschrieben



- **Pseudodokument: alle Ankertexte, die auf genau eine Seite verlinken**

- **Gewichtung der Terme im Pseudodokument: WDF * IDF * G**

  G: willkürlich gewählter Gewichtungsfaktor
  - **linkende und verlinkte Seite auf derselben Site (interner Link): kleines G**
  - **externer Link: großes G**
  - **ggf. Links von [intellektuell ausgezeichneten] "Qualitäts-Sites" höher gewichten**

Kraft, R., & Zien, J. (2004). Mining anchor text for query refinement.
In Proceedings of the 13th International World Wide Web Conference (pp. 666-674). New York, NY: ACM.

F.2 Rankingfaktoren

Pfadlänge

- **Idee (Dean et al. / Google): je länger der Pfad, desto weniger relevant das Dokument**
- **"Pfadlänge": Anzahl der "/" bzw. "." in der URL über Minimum**

- **Beispiel 1: http://www.phil-fak.uni-duesseldorf.de/infowiss**
  Pfadlänge: 2
- **Beispiel 2: http://www.phil-fak.uni-duesseldorf.de/infowiss/mitarbeiter/stock/publ/2008/dok1.pdf**
  Pfadlänge: 7

Dean, J.A., Gomes, B., Bharat, K., Harik, G., & Henzinger, M.R. (2001).
Methods and apparatus for employing usage statistics in document retrieval.
Patent Nr. US 8,156,100 B2.

F.2 Rankingfaktoren

Pfadlänge

- **Gew(Pfadlänge) (d) = log (20 – PL) / log(20)**

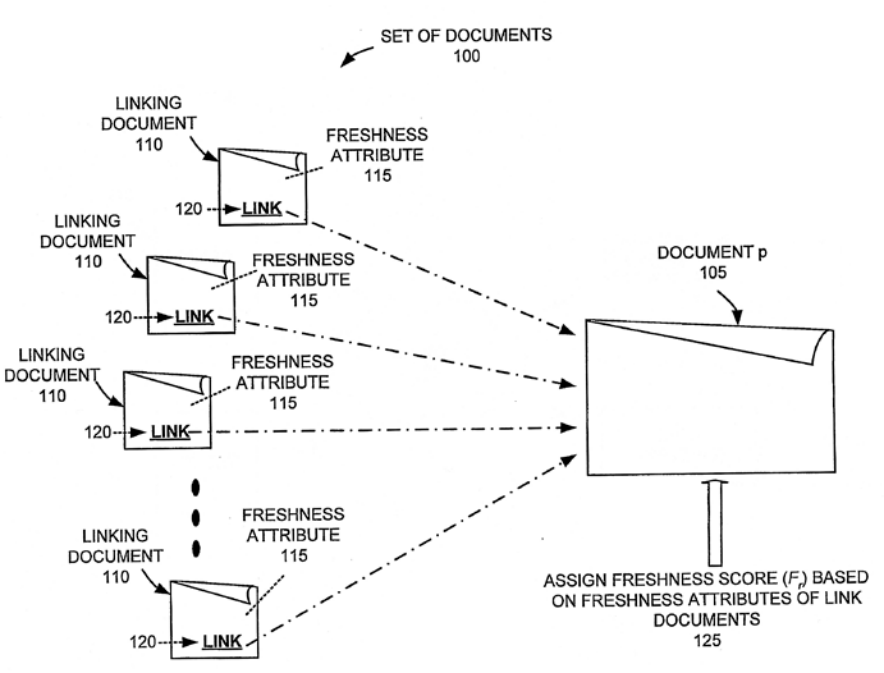
- **d : Webseite**
- **PL: Pfadlänge (über Minimum)**

- **Pfadlänge = 0 : log(20) / log(20) = 1**
- **Pfadlänge = 1 : log(19) / log(20) = 0,98**
- **Pfadlänge = 2 : log(18) / log(20) = 0,96**
- **...**
- **Pfadlänge = 18 : log(2) / log(20) = 0,23**
- **Pfadlänge = 19 : log(1) / log (20) = 0**

F.2 Rankingfaktoren

Aktualität

- **Idee (Henzinger / Google): je "frischer" desto relevanter**
- **"last modified" der betreffenden Seite: nicht genutzt (zu fehleranfällig)**
- **sondern: Nutzung der "last modified"-Daten derjenigen Dokumente, die auf das Dokument d linken**
- **oder (soweit dort kein Datum vorhanden): letztes Crawl-Datum des auf d linkenden Dokuments (danach: Link auf d entfernt)**



Henzinger, M.R. (2004). Systems and methods for determining document freshness. Patentanmeldung WO 2005/033977 A1.

F.2 Rankingfaktoren

Aktualität

- **"Frische" : Schwellenwert (z.B. 2 Jahre)**
- **"alte Seite" : Seite linkt(e) auf d *und* Seite hat Last-modified-Angabe oder letzte Version der Seite mit Link auf d > Schwellen-wert**
- **"neue Seite" : Seite linkt(e) auf d *und* Seite hat Last-modified-Angabe oder letzte Version der Seite mit Link auf d < Schwellen-wert**

- **Freshness Score(d) = Anzahl frischer Seiten / Anzahl alter Seiten**
  - **überwiegen alte linkende Seiten: Freshness Score < 1**
  - **überwiegen neue linkende Seiten: Freshness Score > 1**

F.2 Rankingfaktoren

Typen von Webanfragen

- **Broder: Navigation - Information - Transaktion**
- **Rose/Levinson: Navigation - Information - Ressourcen (z.B. Downloads)**
- **Wissenschaftssuchmaschine: Homepage Wissenschaftler / Institution - Tagung - Artikel - Patent**

- **jeweilige Anfragetypen im Suchbildschirm zur Auswahl anbieten**

- **je nach Anfragetyp zusätzliche Merkmale beim Dokument speichern**

Broder, A. (2002). A taxonomy of Web search. ACM SIGIR Forum, 36(2), 3-10.

F.2 Rankingfaktoren

Nutzungsstatistiken als Rankingkriterium

- **Idee: je häufiger eine Webseite aufgerufen wird, desto wichtiger ist sie**
- **Methoden**
  - **Statistik der angeklickten Links in Trefferlisten einer Suchmaschine**
    - **Einsatz bei: Direct Hit**
    - **"Abfallprodukt" bei Suchmaschinen**
    - **Nachteil: nur Seiten, die schon einmal in einer Trefferliste vorkamen und auch angeklickt worden sind**
  - **Statistik aller besuchter Seiten ausgewählter Rechner (früher: Toolbar; heute: Smartphone-Nutzer)**
    - **Einsatz bei: Google (und vielen anderen Suchmaschinen)**
    - **alle Seitenaufrufe werden gespeichert und periodisch an die Suchmaschine gemeldet**
    - **Vorteil: nicht nur Trefferlisten, Zusatzinformationen: Ort (via GPS), Zeit**

F.2 Rankingfaktoren

Dwell Time

- **Verbleibdauer auf Zielseite**
- **Kurze Zeit (also schnell zur Trefferliste zurück): Seite für Nutzer wenig interessant**
- **Lange Zeit: für Nutzer interessant (???) – muss nicht sein**

- **Also: offen, ob überhaupt als Rankingfaktor brauchbar**

F.2 Rankingfaktoren

Ranking nach Sprache

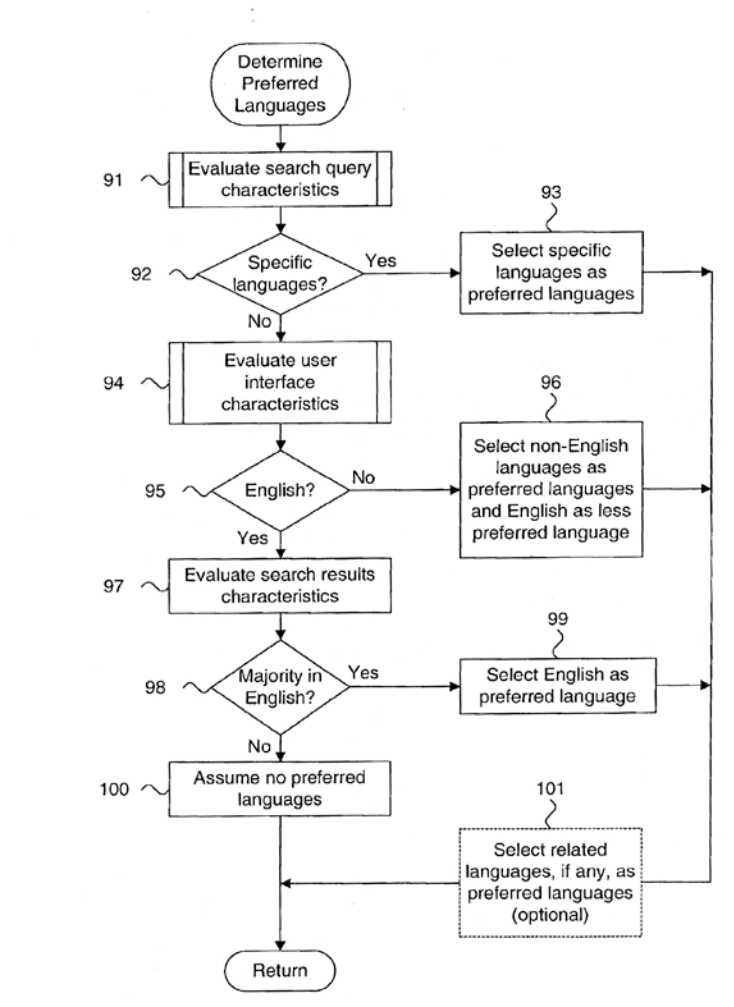
- **Voraussetzungen**
  - **Sprache der Dokumente erkannt**
  - **bevorzugte Nutzersprache(n) erkannt**
    - **Terme der Suchanfrage (da nur wenige, recht ungenau)**
    - **Sprache der angeklickten Dokumente**
    - **aufgerufenes Suchinterface (Google.de, Google.fr, ...)**
    - **Spracheinstellung im Interface**
    - **IP-Adresse (z.B. .de-Domain)**
    - **nächstgelegener Einwählknoten**

Lamping, J., Gomes, B., McGrath, M., & Singhal, A. (2003).
System and method for providing preferred language ordering of search results.
Patent Nr. US 7,451,129.

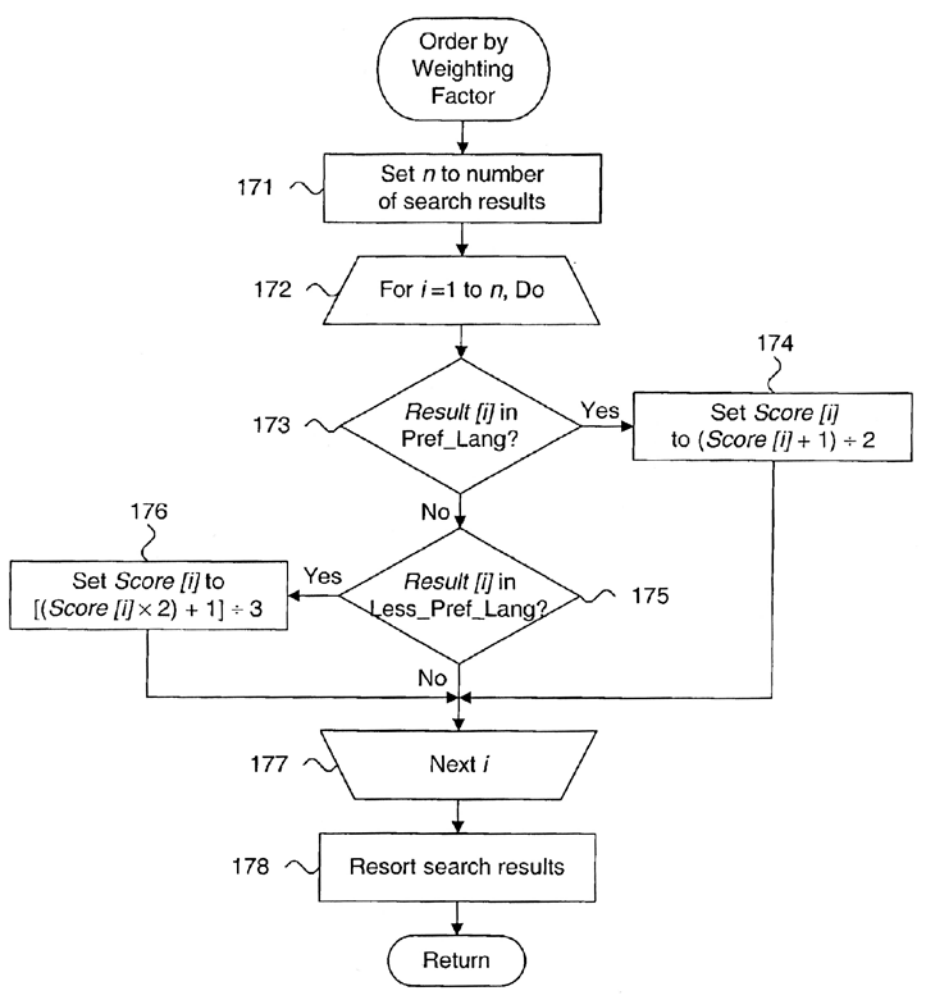
F.2 Rankingfaktoren

Ranking nach Sprache

- **Google: Englisch als Standard**



F.2 Rankingfaktoren



Ausgang: Treffermenge

initialer Score der Dokumente: $s_{(i)}$

Re-Ranking nach Sprache (Score e)

keine bevorzugte Sprache: $e_{(i)} = s_{(i)}$

Dokument in bevorzugter Sprache: $e_{(i)} = [s_{(i)} + 1] / 2$

Dokument in weniger bevorzugter Sprache: $e_{(i)} = [s_{(i)} * 2 + 1] / 3$

Beispiel: s(Dok) = 0,8

bevorzugte Sprache: (0,8 + 1) / 2 = 0,9

weniger bevorzugte Sprache: (0,8 * 2 + 1) / 3 = 2,6 / 3 = 0,87

alle anderen Sprachen: 0,8

F.2 Rankingfaktoren

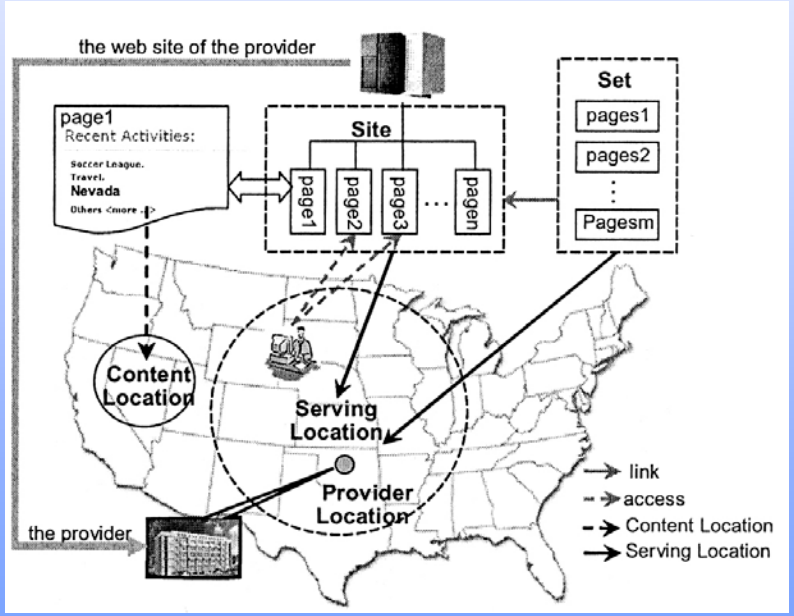
Ranking nach Entfernung: Geographisches Information Retrieval (GIR)

- **Voraussetzungen**
  - **Ortsbezug einer Webseite (z.B. Pizzaservice)**
    - **Ort des Providers (i.d.R. irrelevant)**
    - **Ort des Dienstleisters (z.B. Kerpen-Sindorf)**
    - **Ort der Dienstleistung (z.B. Kerpen und Bergheim)**
  - **Ortsbezug der Anfrage**
    - **hypothetischer Standort (z.B. "Hotel in Chicago", Nutzer recherchiert in Kerpen)**
    - **aktueller Standort (entweder beim Nutzerprofil hinterlegt, in Suchfeld eingegeben oder via GPS ermittelt)**
  - **geographische Begriffsordnung ("Gazetteer": enthält Koordinaten der Orte)**
  - **Erfassen geographischer Suchargumente ("in der Nähe von", "5 Autominuten von", "nördlich von", "maximal 25km entfernt", ...)**

F.2 Rankingfaktoren

GIR: Ortsbezüge innerhalb einer Webseite

- **extrahieren: Ort des Dienstleisters und der Dienstleistung**
  - **Orts- und Straßennamen**
  - **Postleitzahlen**
  - **Telefonvorwahlnummern**
- **so exakt wie möglich:**
  optimal: genaue Adresse



Wang, C., Xie, X., Wang, L., Lu, Y., & Ma, W.Y. (2005). Detecting geographic locations from Web resources. In Proceedings of the 2005 Workshop in Geographic Information Retrieval (pp. 17-24). New York, NY: ACM.

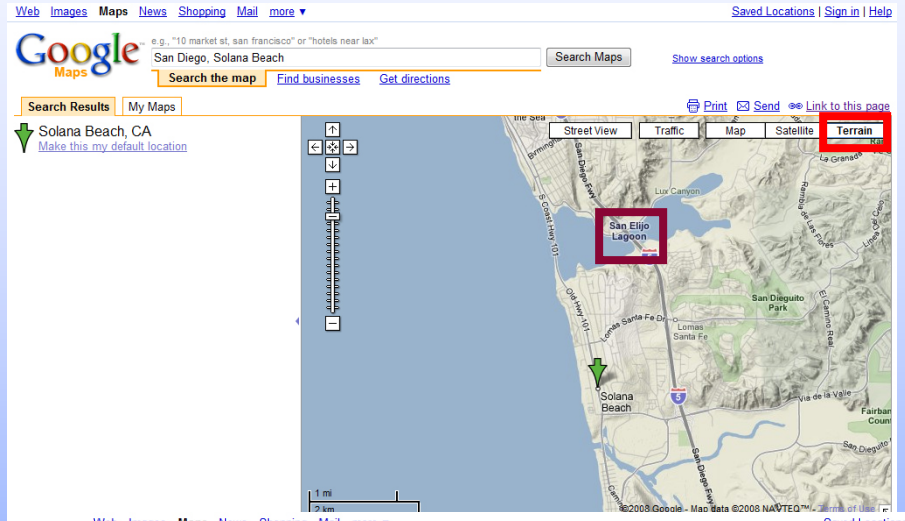
F.2 Rankingfaktoren

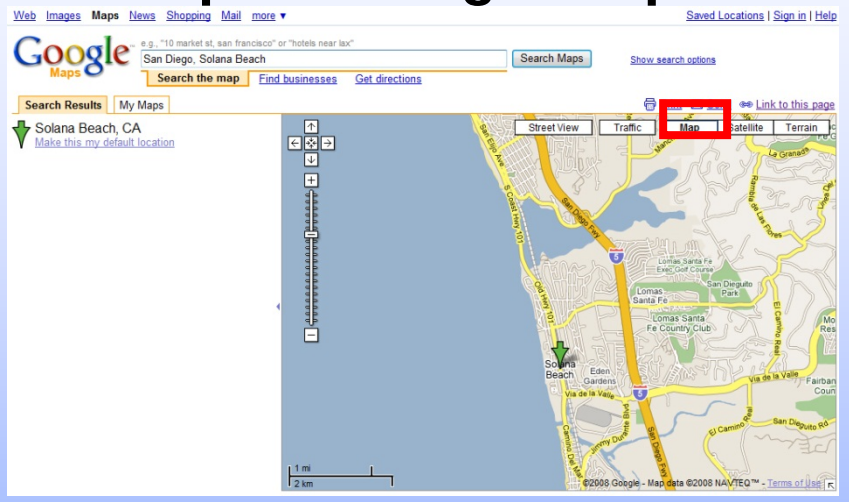
GIR: Suchradien

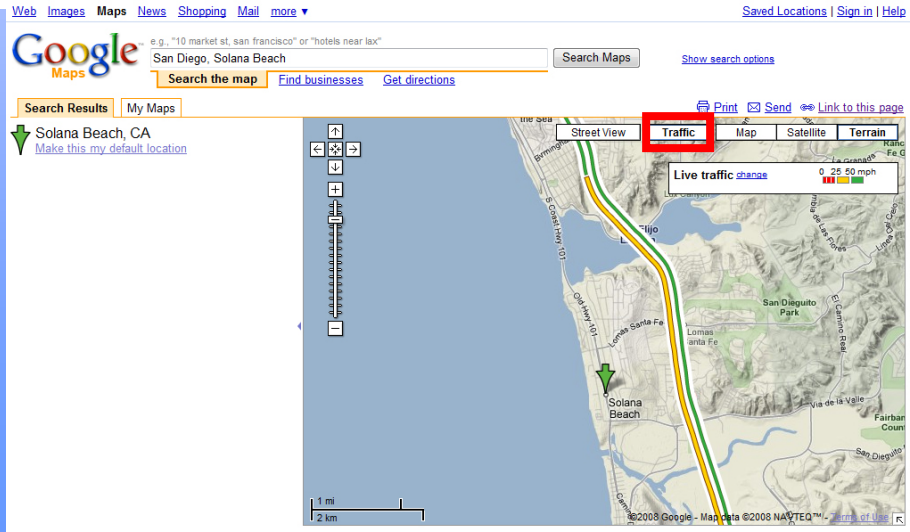
- **Suchfrageerweiterung durch geographische Begriffsordnung**
- **Umrechnen aller geographischer Angaben in Koordinaten**
- **Erkennen des Sucharguments (ggf. in Entfernung und Richtung umdefinieren), Extraktion von Suchradien (und ggf. Ausschnitten; bei Richtungsanfragen)**
- **Match: anhand der Koordinaten**
- **Ranking: nach Entfernung**
  - **Luftlinie**
  - **Berücksichtigung von Wegen**

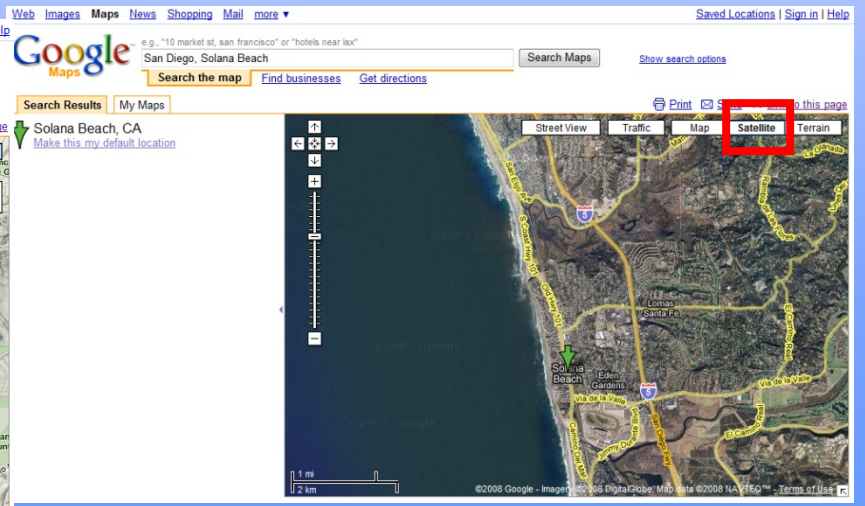
F.2 Rankingfaktoren

GIR: Ausgabe als Landkarte / Beispiel "Google Maps"





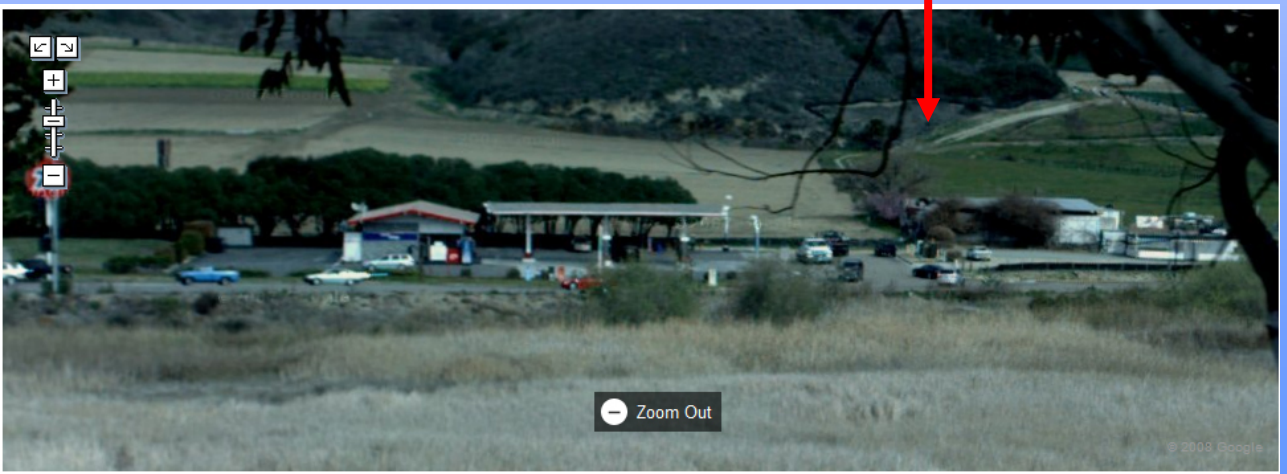




F.2 Rankingfaktoren

GIR:

Ausgabe als Landkarte und Photoserie



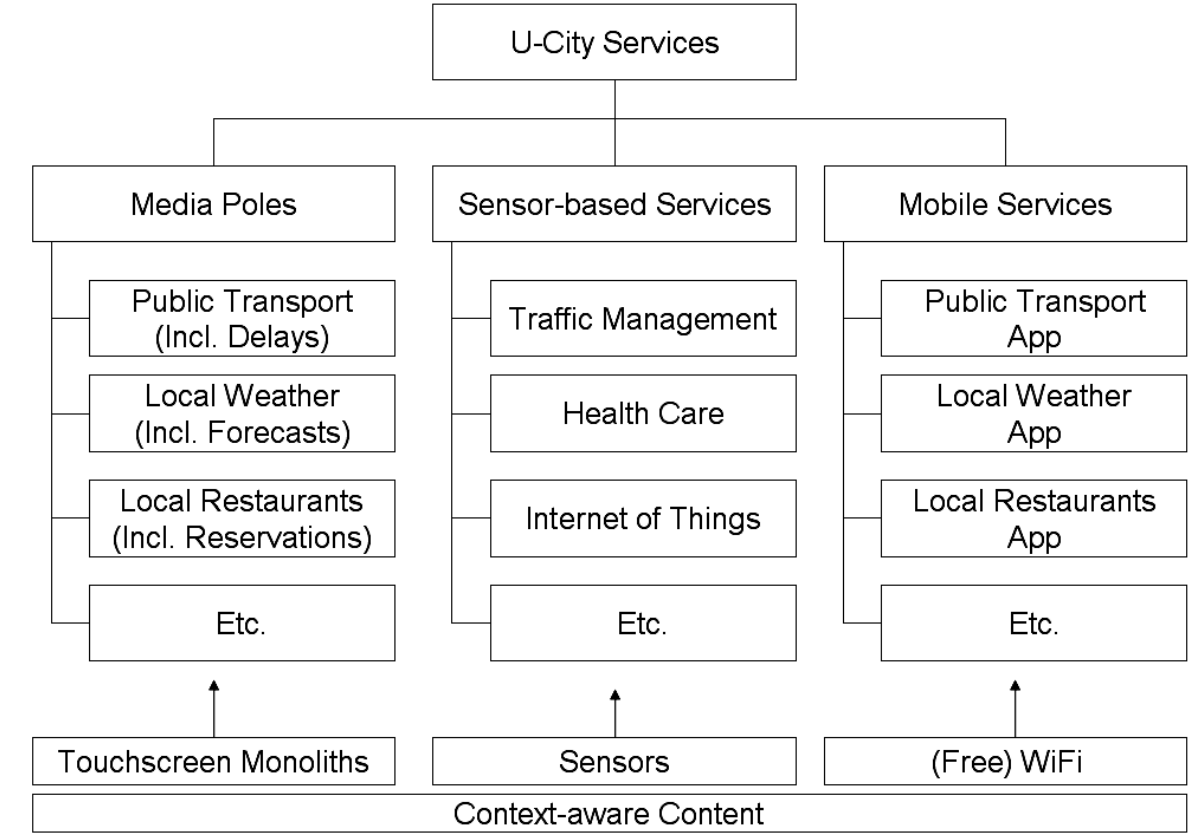
F.2 Rankingfaktoren

Ubiquitäres Retrieval

- **im Rahmen von „ubiquitären Städten“ (U-Cities)**
- **kontext-spezifisches Retrieval (i.d.R.: regionaler bzw. städtischer Kontext und privater Kontext; auch: „Internet der Dinge“)**
- **Dienste**
  - **Mediensäulen**
  - **sensor-basierte Dienste**
  - **Apps / Webdienste**

F.2 Rankingfaktoren

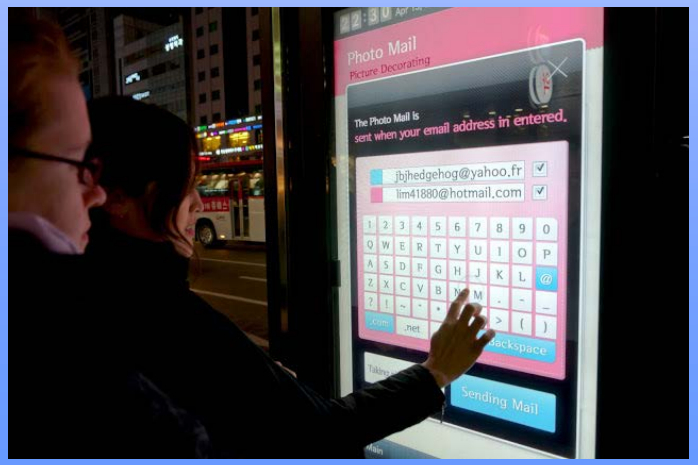
Ubiquitäres Retrieval



F.2 Rankingfaktoren

Ubiquitäres Retrieval: Mediensäulen

Beispiel: Seoul (Gangnam)



F.2 Rankingfaktoren

Ubiquitäres Retrieval: Sensor-basierte Services

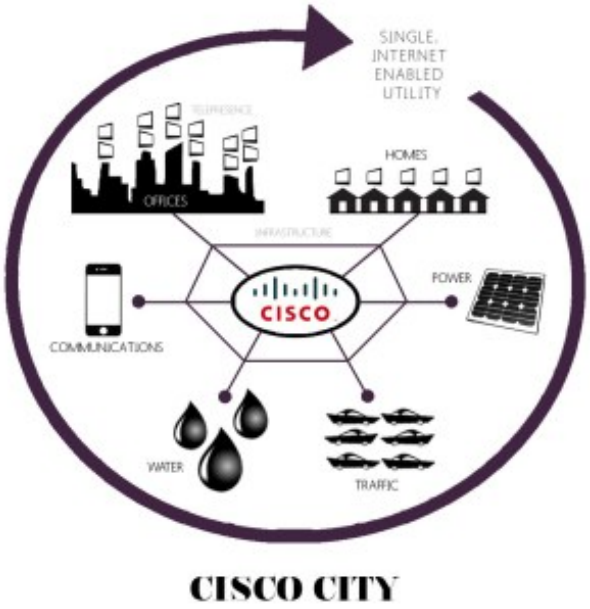
Basiert (u.a.) auf RFID (radio frequency identification)

F.2 Rankingfaktoren

Ubiquitäres Retrieval: Sensor-basierte Services

Beispiel: Songdo

Straßenverkehrssteuerung, U-Bahn-Taktung, Straßenbeleuchtung, …



CISCO CITY

F.2 Rankingfaktoren

Ranking nach Gebot: Sponsored Links

- **RealNames: Suchargumente werden verkauft (Festpreis; pro Suchargument nur ein Käufer)**
- **GoTo (heute Yahoo!): Suchargumente werden versteigert, Ranking nach gebotenem Cost-per-Click (CPC)-Preis**

F.2 Rankingfaktoren

Rangordnung der Ads bei Google AdWords

- **Auktion / Versteigerung des Rangplatzes (nicht nur des 1. Platzes)**
- **Variante einer Second-Price Auktion (Vickrey-Auktion)**
- **Retrievalstatuswert (e) des Ads berücksichtigt neben dem gebotenen max. Preis (max CPC) auch den Quality Score (SC; Klickrate, „Qualität“ der Zielseite) des Werbetextes**

$$e_{Ad} = \max CPC_{Ad} * QC_{Ad}$$

F.2 Rankingfaktoren

Berechnung von Rangposition und Preis bei Google AdWords

- **Preis: Retrievalstatuswert e des direkten Nachbarn / QS + 0,01 $**

Rang	*Kunde*	*Max CPC*	*QS*	*e*	*Preisberechnung*	*Preis*
1	A	$0,40	1,8	0,40 * 1,8 = **0,72**	0,65 / 1,8 + 0,01 = 0,3711	**$0,37**
2	B	$0,65	1,0	0,65 * 1,0 = **0,65**	0,375 / 1,0 + 0,01 = 0,385	**$0,39**
3	C	$0,25	1,5	0,25 * 1,5 = **0,375**	Mindestgebot: 0,04	**$0,04**

Kapitel F.3

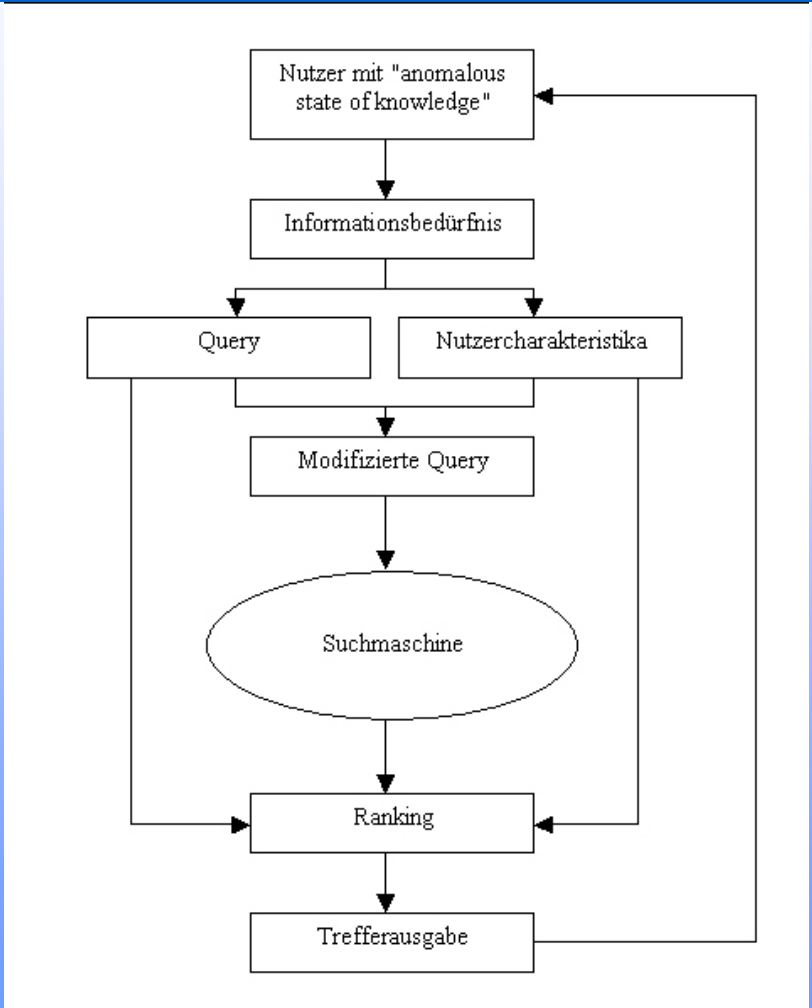
Personalisiertes Retrieval

F.3 Personalisiertes Retrieval

Personalisierung

- **kein *Relevance* Ranking, sondern stets *Pertinence* Ranking**
- **Quellen für Personalisierung**
  - **Eingabe eines grundlegenden Anfragetyps**
  - **Anlegen (oder "Ausspähen") eines Nutzerprofils**
  - **Speichern und Auswerten früherer Suchanfragen eines Nutzers**
  - **Speichern und Auswerten ähnlicher Suchanfragen anderer Nutzer**
  - **Berücksichtigung des Einwählknotens oder einer bevorzugten Sprache des Nutzers**
  - **Berücksichtigung aufgerufener URL**
  - **Berücksichtigung des Standortes bei entfernungskritischen Anfragen**
- **Ansatzpunkte**
  - **Query Expansion**
  - **Sortierreihenfolge**
  - **personalisierte Werbung (z.B. Facebook)**

F.3 Personalisiertes Retrieval



F.3 Personalisiertes Retrieval

Simulation eines Rechercheinterviews

- **nicht nur: *WAS* wird gesucht?**
- **sondern auch: *WARUM* wird gesucht?**

- **Beispiel: Suchargument "Keramik"**
  - **Nutzer 1: sucht Hochzeitsgeschenk**
  - **Nutzer 2: sucht VHS-Kurs in der Nähe**
  - **Nutzer 3: sucht Buch**
  - **...**
- **jeder benötigt völlig andere Dokumente**
- **"adaptives Suchwerkzeug"**

Brusilovsky, P. (2001). Adaptive hypermedia. User Modeling and User-Adapted Interaction, 11, 87-110.

F.3 Personalisiertes Retrieval

Der gläserne Nutzer

- **Culliss (Ask Jeeves): Nutzerprofil**
  - **demographische Angaben:**
    - **Alter, Geschlecht, Wohnort, Land, Einkommen, Größe, Gewicht, Rasse, Religion, sexuelle Orientierung, politische Orientierung, Herkunftsland, Schulbildung, bisherige Verurteilungen, Gesundheit, ...**
  - **persönliche Interessen:**
    - **Hobbys, Sport, Beruf, Expertise, Behinderungen, Gewohnheiten, ...**
  - **persönliche Aktivitäten:**
    - **Suchterme, Suchstrategien, angesehene Dokumente, ... (einschließlich Datum und Uhrzeit der Aktivitäten)**

Culliss, G.A. (2003). Personalized search methods including combining index entries for categories of personal data. Patent-Nr. US 6,816,850.

F.3 Personalisiertes Retrieval

Der gläserne Nutzer

- **Gross/McGovern/Colwell: Nutzerprofil**
  - **Idee: Persönlichen Rechner des Nutzers durchsuchen und dabei die zentralen Themen erfassen**
    - **E-Mail,**
    - **Dokumente aus Textverarbeitung und Tabellenkalkulation**
    - **Präsentationen**
    - **Graphikdateien**
    - **PDF-Dokumente**
  - **"scanning the words" - relative Häufigkeiten der erkannten Worte berechnen - Rangordnung der Worte erstellen**

Gross, W., McGovern, T., & Colwell, S. (2005). Personalized search engine. Patentanmeldung US 2005/0278317 A1.

F.3 Personalisiertes Retrieval

Der gläserne Nutzer

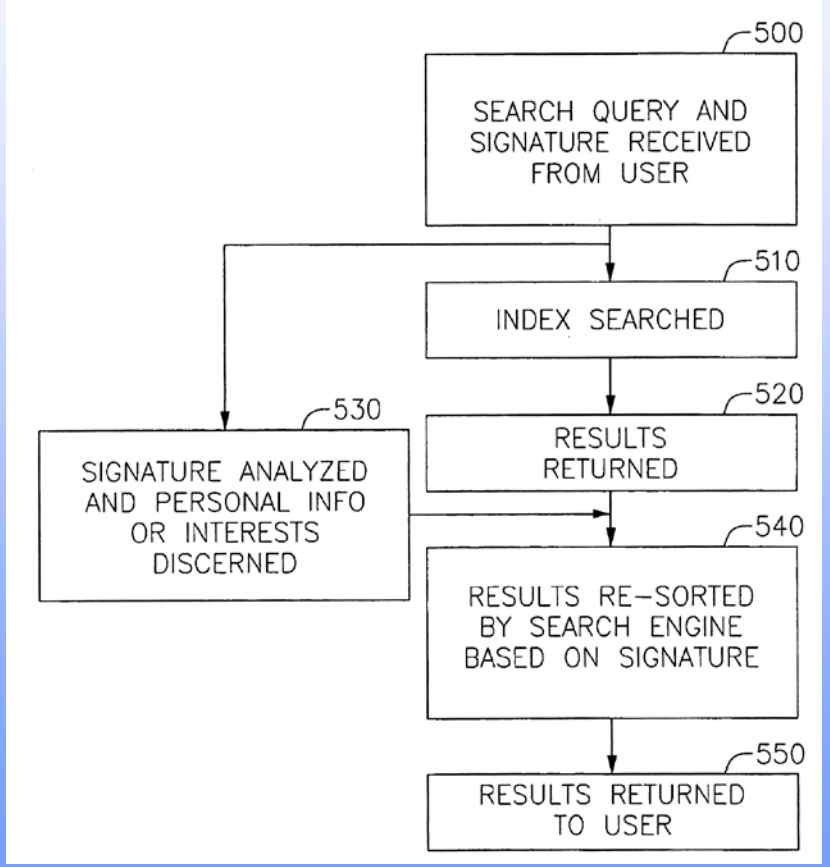
- **Teevan/Dumais/Horwitz (Microsoft): Nutzerprofil**
  - **Idee: alle Bewegungen des Nutzers speichern**
  - **Voraussetzung: Vorliegen von GPS-Daten und Standorten stationärer Systeme**
    - **besuchte Orte im Zeitablauf**
    - **bei GPS: Übersetzung der Positionsangaben in textuelle (Städte)Namen und in Postleitzahlen**
    - **Zuordnung Ort - Wochentag**
    - **Zuordnung Ort - Tageszeit**

Teevan, J.B., Dumais, S.T., & Horvitz, E.J. (2004). Systems, methods, and interfaces for providing personalized search and information access. Patentanmeldung Nr. 2006/0074883 A1.

F.3 Personalisiertes Retrieval

Suche mit thematischem Nutzerprofil

- **Re-Ranking nach Nutzer"signatur" (Gross/McGovern/Colwell)**
- **Signatur**
  - **Wortverteilung über alle Queries (oder typische Wortverteilung einer Sprache)**
  - **Wortverteilung der Themen des Nutzers**
  - **relative Häufigkeit Wort(Nutzer) > relative Häufigkeit(alle Queries) --> zentrales Interesse des Nutzers**
- **bei allen Anfragen des Nutzers sein "zentrales Interesse" in einem zweiten Rankingschritt mitberücksichtigen**
- **Lösung des Homonymproblems ("java")?**



Kapitel F.4

Themenentdeckung und -verfolgung

F.4 Themenentdeckung und -verfolgung

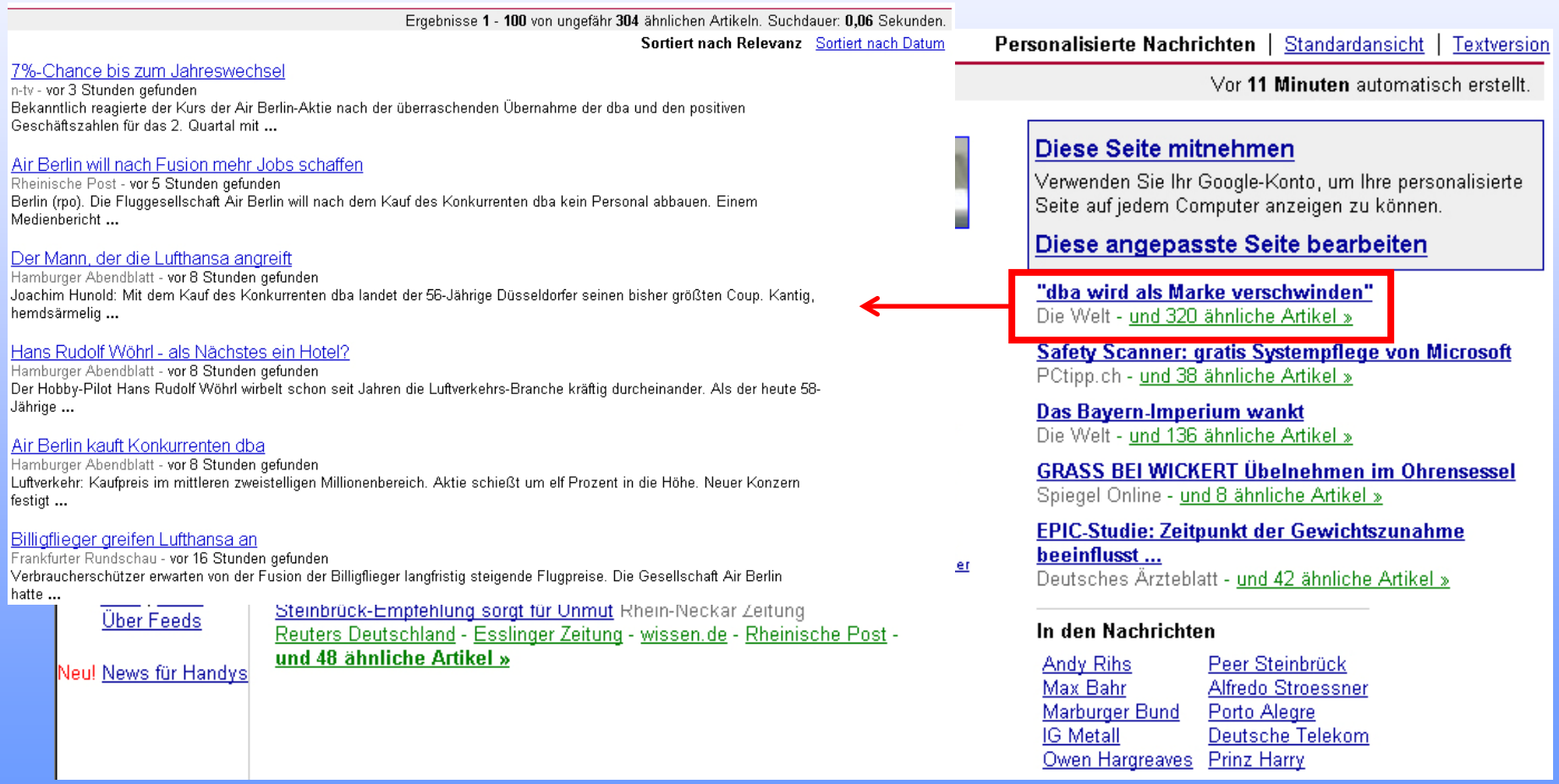
TDT

- **dokumentarische Bezugseinheit: Thema, nicht Dokument**
- **Zusammenfassen gleicher oder ähnlicher Dokumente zu einem Thema (Pseudodokument)**
  - **News im Web (z.B. Google News)**
  - **Blogs**
  - **E-Mails**
  - **Artikel aus Zeitungen, Zeitschriften und von Agenturen**
  - **Rundfunksendungen (Hörfunk, Fernsehen)**
- **Arten**
  - **TDT ohne Zeitbezug (anwendbar auf alle Dokumente)**
  - **Event-Tracking: Entdeckung neuer Ereignisse**

Allan, J. (Ed.) (2002). Topic Detection and Tracking: Event-based Information Organization. Boston, MA: Kluwer.
Allan, J. (2002). Detection as multi-topic tracking. Information Retrieval, 5, 139-157.

F.4 Themenentdeckung und -verfolgung

Google News: Themen

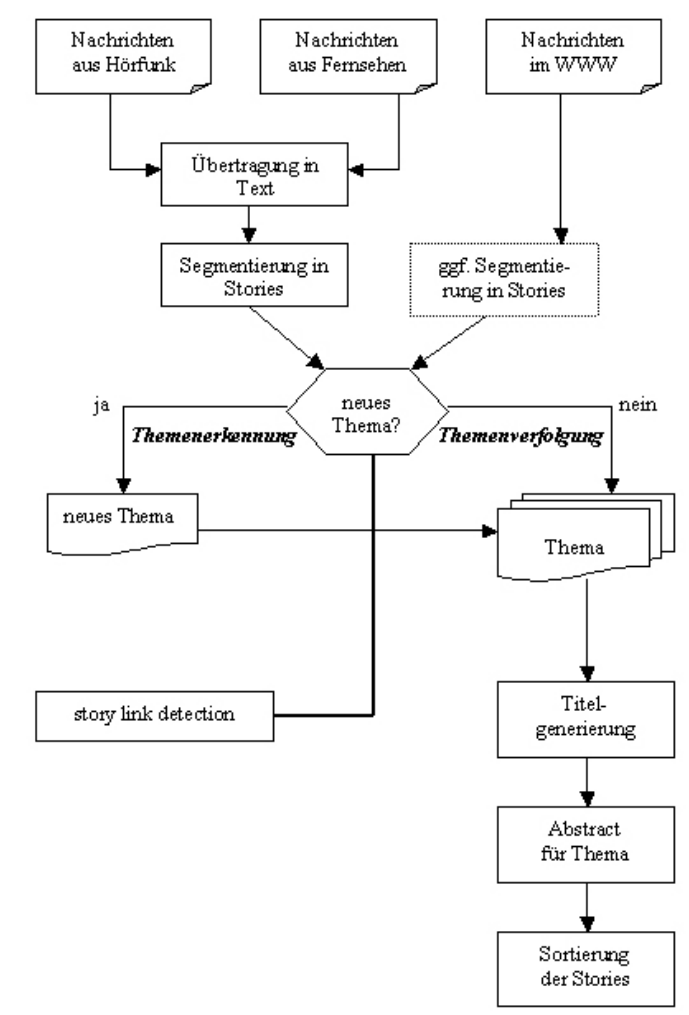


F.4 Themenentdeckung und -verfolgung

TDT - Grundbegriffe

- **"Story": abgrenzbare Textstelle (oder ein ganzes Dokument), in der (oder in dem) ein Ereignis besprochen wird**
- **"Thema" (topic): Beschreibung der Aboutness der jeweiligen Stories**
- **"Themenentdeckung" (topic detection): Gruppieren aktueller Stories zu einem *neuen* Thema**
- **"Themenverfolgung" (topic tracking): Hinzufügen von Stories zu einem *bekannten* Thema**

F.4 Themenentdeckung und -verfolgung



F.4 Themenentdeckung und -verfolgung

- **aktuelle Story: neues oder bekanntes Thema?**
- **Story s: dargestellt im Vektorraum**
- **Dimensionen: Terme t (Conflation, Phrasen, named entities), Werte: TF*IDF**
- **$tf_{t,s}$ = absolute Auftretenshäufigkeit von Term t in Story s**
- **df_t = Anzahl der Stories in der Datenbank, die t enthalten**
- **N = Anzahl aller Stories in der Datenbank**
- **Termgewicht:**

$$w_{t,s} = [tf_{t,s} * \log((0{,}5 + N) / df_t)] / [\log(N + 1)]$$

- **Story-Vektor: Top 1.000 Terme**
- **Vergleich von s mit allen vorhandenen Themen s_i: Cosinus (Sim)**
- **$Sim(s,s_i) > 0{,}21$: s ist Story des bekannten Themas s_i**
- **$Sim(s,s_i) \leq 0{,}21$: s ist Story eines neuen Themas**

Allan, J., Harding, S., Fisher, D., Bolivar, A., Guzman-Lara, S., & Amstutz, P. (2005).
Taking topic detection from evaluation to practice.
In Proceedings of the 38th Annual Hawaii International Conference on System Sciences.

F.4 Themenentdeckung und -verfolgung

- **praktische Erfahrung: Cosinus allein trennt nicht exakt**

Topic not seen before

Turkey has sent 10,000 troops to its southern border with Syria amid growing tensions between the two neighbors, newspapers reported Thursday. Defense Minister **Ismet Sezgin** denied any troop movement along the border, but said **Turkey's** patience was running out. **Turkey** accuses Syria of harboring Turkish Kurdish rebels fighting for autonomy in **Turkey's** southeast; it says rebel leader Abdullah Ocalan lives in Damascus.

Closest Story due to Named Entities

A senior **Turkish** government official called Monday for closer military cooperation with neighboring Bulgaria. After talks with President Petar Stoyanov at the end of his four-day visit, Turkish Deputy Premier and National Defense Minister **Ismet Sezgin** expressed satisfaction with the progress of bilateral relations and the hope that Bulgarian-**Turkish** military cooperation will be promoted.

gemäß Cosinus gehört die obere Story zum unten angegebenen Thema

dies ist falsch

Grund: Übereinstimmung in "Turkey/ Turkish" (hohe TF) und in "Ismet Sezgin" (sehr hoher IDF)

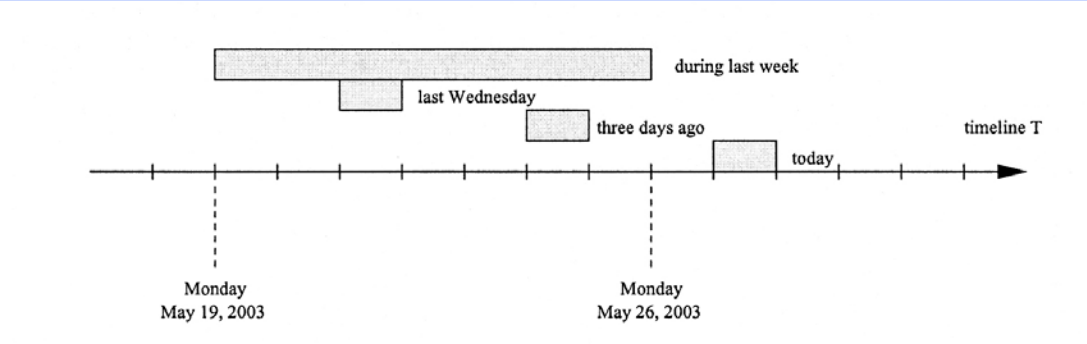
Vorschlag: Ähnlichkeiten getrennt für "named entities" und restliche "topic terms" (mit bewährter Grenze bei 0,21) berechnen

Kumaran, G., & Allan, J. (2005).
Using names and topics for new event detection. In Proceedings of Human Language Technology Conference / Conference on Empirical Methods in Natural Language Processing, Vancouver (pp. 121-128).

F.4 Themenentdeckung und -verfolgung

Zeitliche Nähe

- **bei Event-Detection: Zeit(intervall) des Ereignisses erfassen**
- **zeitliche Ausdrücken in den Stories ("gestern", "letzte Woche") auflösen**



- **"gleiches Event": zusätzlich zu Cosinus: Story und Thema liegen in einem gemeinsamen Zeitfenster**

Makkonen, J., Ahonen-Myka, H., & Salmenkivi, M. (2004). Simple semantics in topic detection and tracking. Information Retrieval, 7, 347-368.

F.4 Themenentdeckung und -verfolgung

Themenverfolgung

- **bei erster Story eines Themas: Story-Vektor = Themenvektor**
- **ab zweiter Story: jeweils Errechnen des Story-Zentroiden**

- **Titelgenerierung: Top n (z.B. 10) Terme des Zentroiden als Titelersatz**

- **Abstractgenerierung: automatisches Extracting der zentralen Sätze aus allen Stories des Themas - Nutzung der top n Zentroid-Terme**

F.4 Themenentdeckung und -verfolgung

Ranking der Stories

- **an erster Stelle: die (zeitlich gesehen) erste Story**

oder

- **Google News-Sortierkriterien (Qualitätskriterien für die Quellen)**
  - **Anzahl der Artikel der Quelle in einer Zeiteinheit**
  - **durchschnittliche Länge der Artikel**
  - **"breaking news score"**
  - **Anzahl der Mitarbeiter**
  - **Anzahl der Büros**
  - **usw.**

Curtiss, M., Bharat, K., & Schmitt, M. (2003).
Systems and methods for improving the ranking of news articles. Patent Nr. US 7,577,655 B2.

Teil G:
Spezialprobleme des Information Retrieval

Kapitel G.1

Soziale Netzwerke und „Kleine Welten"

G.1 Soziale Netzwerke und "small worlds"

Soziale Netzwerke
Grundbegriffe

- **Struktur: Graph**
- **Akteur: Knoten** **(im IR z.B.: Dokumente)**
- **Verbindung zwischen Akteuren:**
  Linien **(Links)**

- **gerichteter Graph: alle Linien haben Richtung (In-Link, Out-Link)**
- **ungerichteter Graph: ohne Richtung (Link-bibliographic coupling, Co-Link)**

Wasserman, S., & Faust, K. (1994). Social Network Analysis: Methods and Applications. Cambridge: Cambridge University Press.

G.1 Soziale Netzwerke und "small worlds"

Graphendichte und -durchmesser

- **Dichte (D)**

$$D(\text{G-ungerichtet}) = L / [n * (n - 1) / 2)]$$
$$D(\text{G-gerichtet}) = L / [n * (n - 1)]$$

- **n : Anzahl der Knoten im Graphen**
- **L : Anzahl der Linien im Graphen**
- **n * (n - 1) : maximale Anzahl der Linien im (gerichteten) Graphen**

- **Durchmesser: größte vorhandene Pfadlänge im Graphen**

G.1 Soziale Netzwerke und "small worlds"

Zentralität (1): Degree

- **Anzahl der Linien eines Knotens in Relation zur Anzahl aller Knoten im Netz**

$$CD(N) = m / (n - 1)$$

- **CD(N) : Degree des Knotens N**
- **m : Anzahl der vom Knoten N ausgehenden Linien**
- **n : Anzahl aller Knoten im Netz**

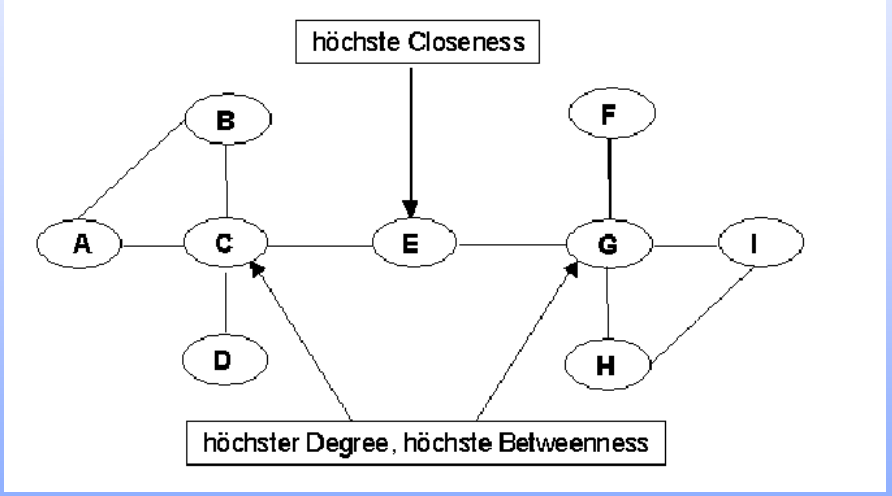
- **In-Degree: m : Anzahl der Linien, die auf den Knoten N zeigen**
- **Out-Degree: m : Anzahl der Linien, die vom Knoten N ausgehen**

- **PageRank: orientiert am In-Degree**
- **Kleinberg: orientiert an In- und Out-Degree**

G.1 Soziale Netzwerke und "small worlds"

Zentralität (1): Degree

CD(N) = m / (n - 1)



Degree von C: **CD(C) = 4 / (9 - 1) = 4 / 8 = 0,5**

Mutschke, P. (2004). Autorennetzwerke: Verfahren der Netzwerkanalyse als Mehrwertdienste für Informationssysteme. Bonn: InformationsZentrum Sozialwissenschaften. (IZ-Arbeitsbericht; 32).

G.1 Soziale Netzwerke und "small worlds"

Zentralität (2): Closeness

- **Distanz eines Knotens zu allen anderen Knoten im Netz (auf dem jeweils kürzesten Pfad)**

$$CC(N) = (n - 1) / (Pf_1 + ... + Pf_i)$$

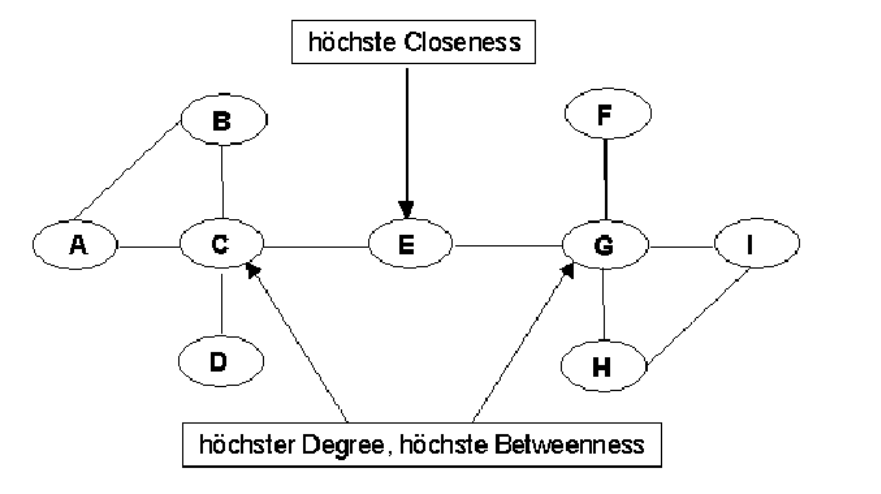
- **n : Anzahl aller Knoten im Netz**
- **Pf_i : Pfadlänge: Anzahl der Linien zwischen dem Knoten N und dem Knoten I (auf dem kürzesten Weg)**

- **kein Unterschied zwischen gerichteten und ungerichteten Graphen**

G.1 Soziale Netzwerke und "small worlds"

Zentralität (2): Closeness

$$CC(N) = (n - 1) / (Pf_1 + ... + Pf_i)$$



Closeness von A: $CC(A) = 8 / (1+1+2+2+3+4+4+4) = 8 / 21 = 0,38$

Closeness von E: $CC(E) = 8 / (1+2+2+2+1+2+2+2) = 8 / 14 = 0,57$

Closeness von D: $CC(D) = 8 / (2+2+1+2+3+4+4+4) = 8 / 22 = 0,36$

G.1 Soziale Netzwerke und "small worlds"

Zentralität (3): Betweenness

- **Kontrolle eines Knotens über Pfade im Netz**

$$CB(N) = \frac{(g_{nA} / g_A) + (g_{nB} / g_B) + \ldots + (g_{nI} / g_i)}{(n - 1) * (n - 2) / 2}$$

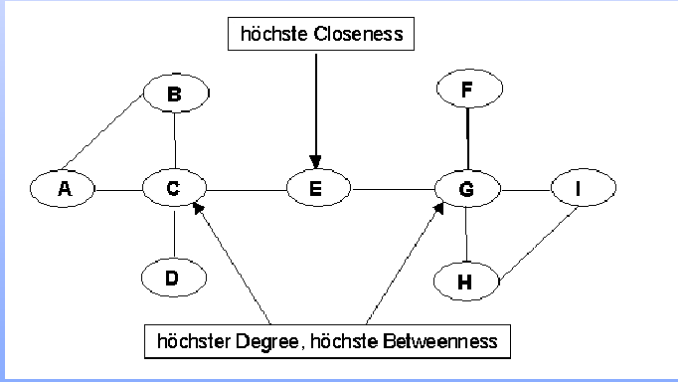
- **g_A : Anzahl aller Pfade, die zwischen A und allen anderen Knoten im Netz existieren**
- **g_{nA} : Anzahl aller Pfade, die zwischen A und allen anderen Knoten im Netz existieren *und* die über N führen**
- **n : Anzahl aller Knoten im Netz**

- **gilt (in dieser Form) nur in ungerichteten Graphen**

G.1 Soziale Netzwerke und "small worlds"

Zentralität (3): Betweenness

$$CB(N) = \frac{(g_{nA} / g_A) + (g_{nB} / g_B) + \ldots + (g_{nI} / g_i)}{(n - 1) * (n - 2) / 2}$$



Betweenness von E: CB(E) = (8 * 4/7) / [(9-1)*(9-2)/2] = 4,57 / 28 = 0,163

Betweenness von C: CB(C) = (2*6/7+7/7+5*3/7) / 28 = 4,86 / 28 = 0,174

Betweenness von D: CB(D) = 0

G.1 Soziale Netzwerke und "small worlds"

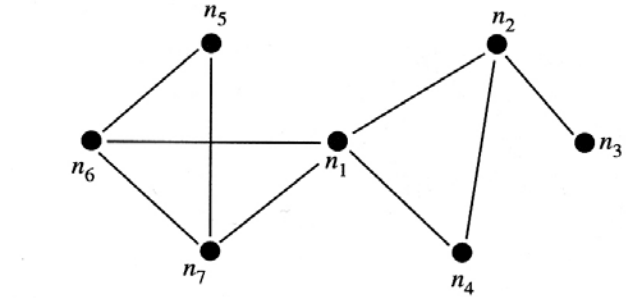
Retrievalstatuswert einer Webseite nach Zentralität

$$\alpha * CD(N) + \beta * CC(N) + \gamma * CB(N)$$
$$\alpha, \beta, \gamma \geq 1$$

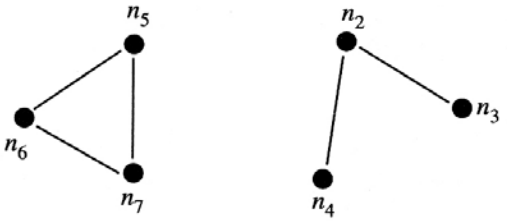
- **bei Experimenten (teilweise) signifikante Verbesserung der Retrievalqualität**

G.1 Soziale Netzwerke und "small worlds"

Exponierte Stellung im Netz (1): Trennpunkt



Node n_1 is a node cut, or cutpoint

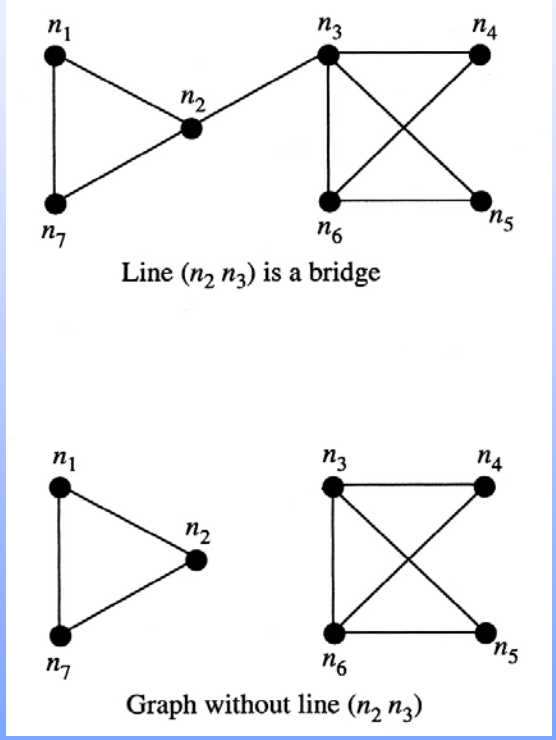


The graph without node n_1

- **Trennpunkt-Dokument (n_1) höher gewichten oder als Trennpunkt anzeigen?**

G.1 Soziale Netzwerke und "small worlds"

Exponierte Stellung im Netz (2): Brücken

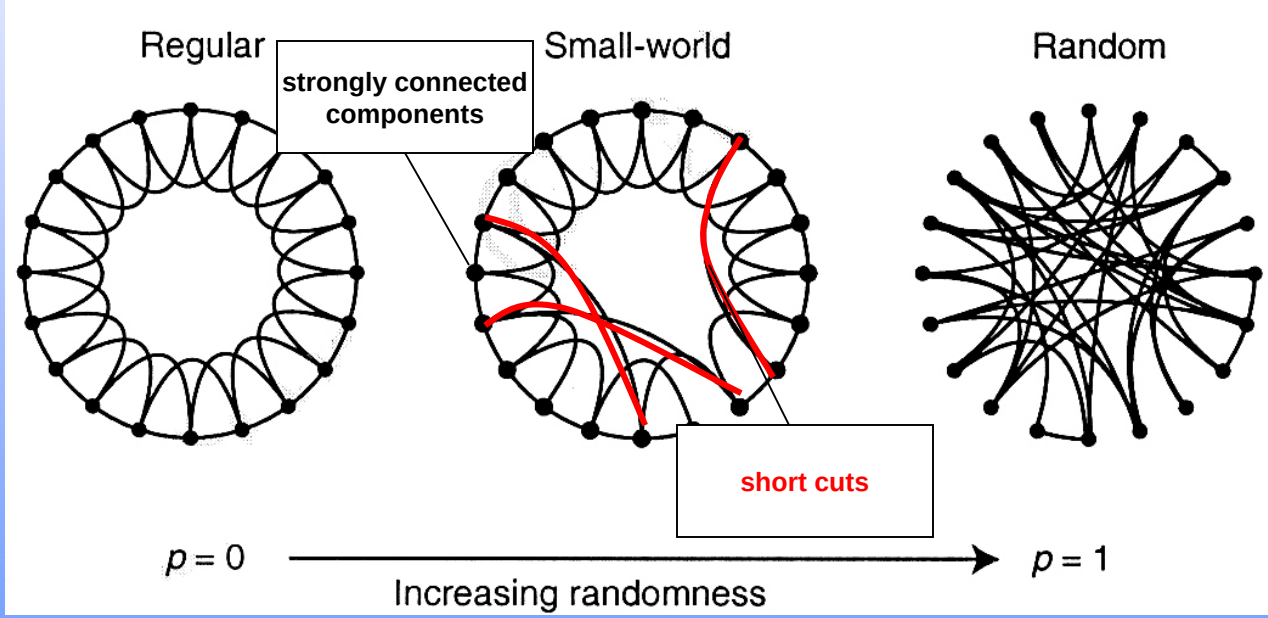


- **Brückendokumente (n_2 und n_3) höher gewichten oder als Brücken anzeigen?**

G.1 Soziale Netzwerke und "small worlds"

"Small worlds"

- **Graphendichte hoch**
- **Graphendurchmesser klein**



Watts, D.J., & Strogatz, S.H. (1998). Collective dynamics of 'small-world' networks. Nature, 393, 440-442.

G.1 Soziale Netzwerke und "small worlds"

"Small Worlds"

- **Beispiele**
  - **Milgram-Experiment (1967)**
  - **Erdös-Zahlen (Kleine Welt der Mathematiker)**
  - **"Six Degrees of Kevin Bacon" (Kleine Welt der Schauspieler)**

- **Wie erfassen?**
  - **strongly connected components (SCC): Netzwerkanalyse (Graphendichte)**
  - **short cuts ("weak ties" - schwache Verbindungen / Granovetter)**
    - **hohe Betweenness**
    - **Zugehörigkeit zu unterschiedlichen SCC, also: Brücken oder Trennpunkte**

G.1 Soziale Netzwerke und "small worlds"

"Small Worlds" im IR

- **SCC**
  - **Analyse der initialen Treffermenge nach SCC anhand der Graphendichte**
  - **Sortierung der SCC nach der Anzahl der Dokumente (Adamic): der größte Graph steht am Anfang**
  - **alternativ: Sortierung der SCC nach Graphendichte: der am engsten verknüpfte Graph steht am Anfang**
  - **beim Anklicken eines SCC: Ranking der Dokumente im Graphen nach Degree, ggf. nach Closeness**
- **short cuts**
  - **gesondert anzeigen**
  - **als Dokumentenpaar (bei Brücken)**
  - **Einzeldokument (bei Trennpunkt)**
  - **jeweils: Angabe der verbundenen SCC**

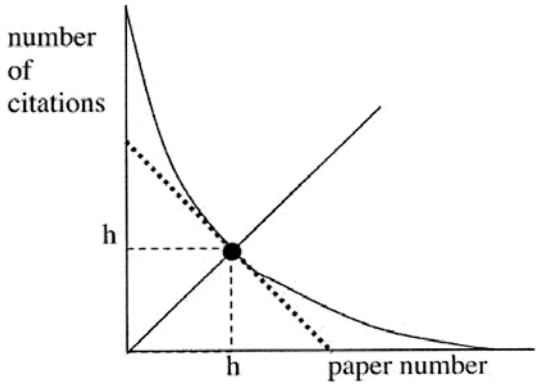
optimal:

Trefferanzeige nicht als Liste, sondern als Netzwerk

G.1 Soziale Netzwerke und "small worlds"

Ranking nach h-Index?

- **h-Index: h ist die Anzahl der Artikel eines Wissenschaftlers, die mindestens h-mal zitiert worden sind**
- **für Wissenschaftssuchmaschinen (z.B. Web of Science, Scopus, Google Scholar): h-Index des Autors als Sortierkriterium?**



Hirsch, J.E. (2005). An index to quantify an individual's scientific research output.
Proceedings of the National Academy of Sciences of the United States of America, 102(46), 16569-16572.

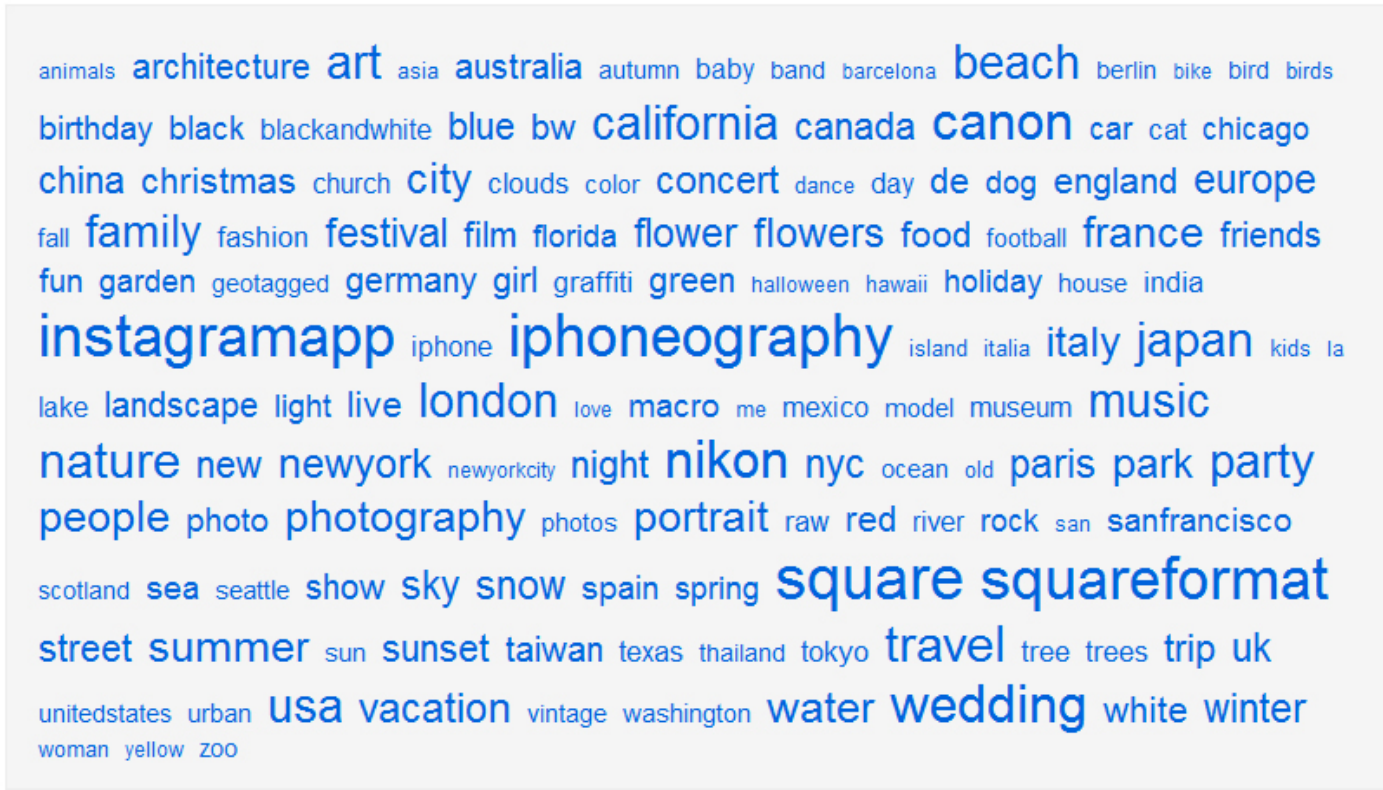
Kapitel G.2

Visuelle Retrievalhilfsmittel

G.2 Visuelle Retrievalhilfsmittel

Tag Cloud

Beispiel: Die häufigsten Tags bei Flickr



G.2 Visuelle Retrievalhilfsmittel

Term Clouds (Tag Clouds)

graphische Darstellung der Terme

- **eines ganzen Informationsdienstes**
- **der aktuellen Treffermenge**
- **eines Dokumentes**

Terme

- **Alphabetisch sortiert**
- **je nach Häufigkeit mit größerem Druckerfonts**

$$\text{TermSize}(t_i) = 1 + C * [\log (f_i - f_{min} + 1)] / [\log (f_{max} - f_{min} + 1)]$$

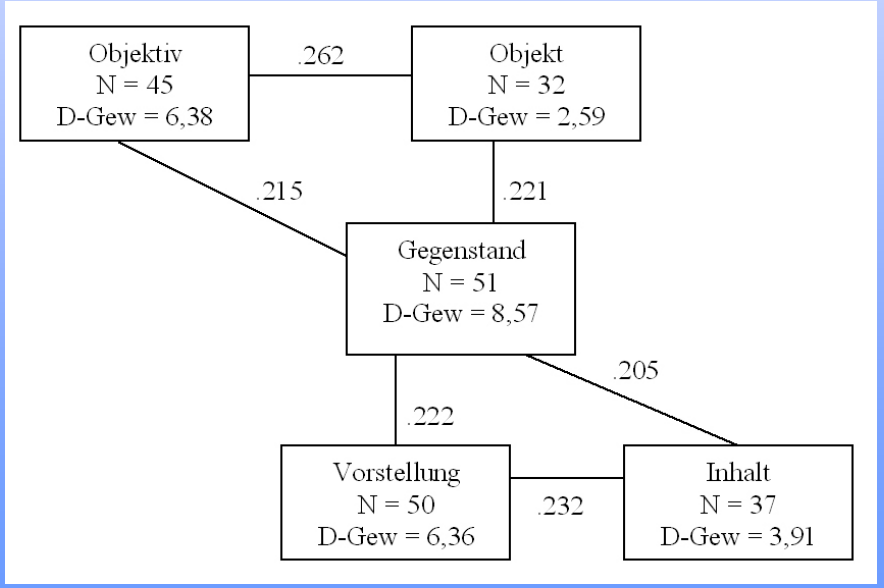
f_i : Häufigkeit des Terms i in der Dokumentenmenge

G.2 Visuelle Retrievalhilfsmittel

Einstellbare syntagmatische Netze

- "Statistischer Thesaurus"
- Errechnung der Ähnlichkeit (Jaccard-Sneath, Dice, Cosinus)
- Darstellung als Cluster (k-Nearest Neighbors, Single Linkage, Complete Linkage, Group Average Linking)
- Auflösung des Clusters ist abhängig von der Einstellung des Ähnlichkeitswertes (ψ)

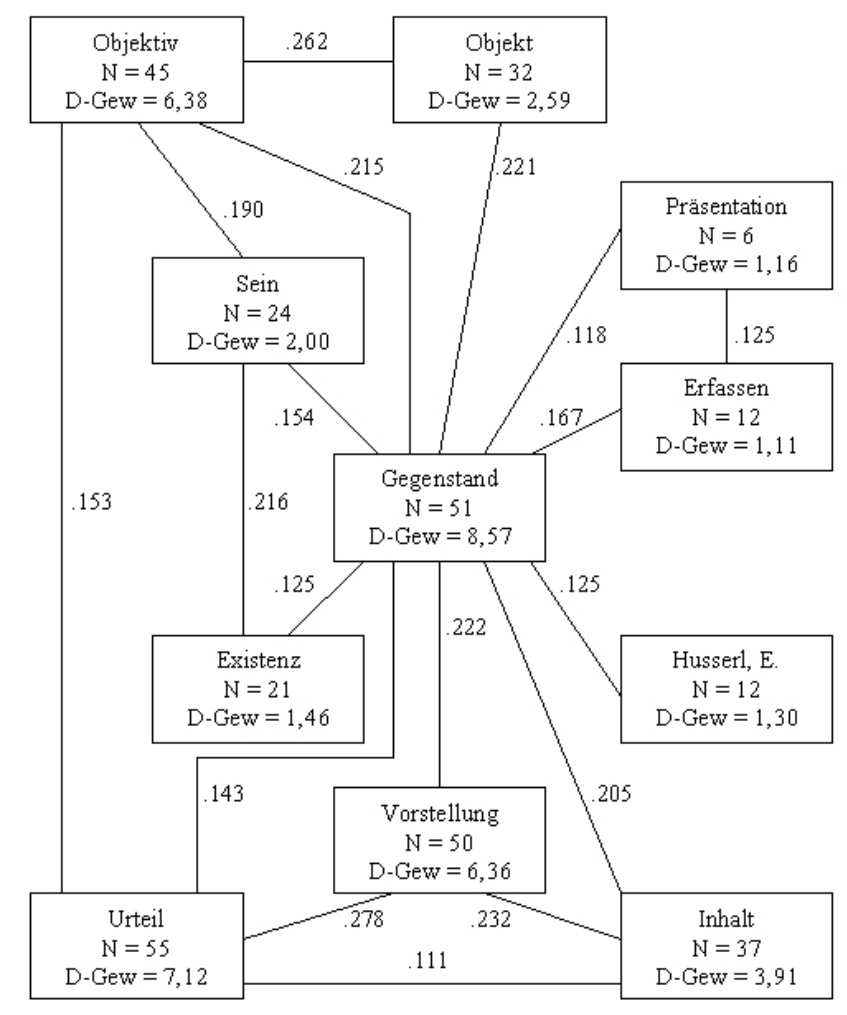
ψ > 0,2



G.2 Visuelle Retrievalhilfsmittel

ψ > 0,11

- **Absenken von ψ:**
  Netz wird reichhaltiger, aber unübersichtlicher
- **Erhöhen von ψ:**
  Netz wird übersichtlicher, aber Details fallen weg



G.2 Visuelle Retrievalhilfsmittel

"Dynamische" Klassierung

- **Indexieren der Dokumente mittels Notationen (oder Deskriptoren, ...) - wie gewohnt**
- **Anzeige der Trefferhäufigkeiten bei der Notation (oder beim Deskriptor, ...) - auch nach bereits erfolgter Suche**
- **ermöglicht Browsen**

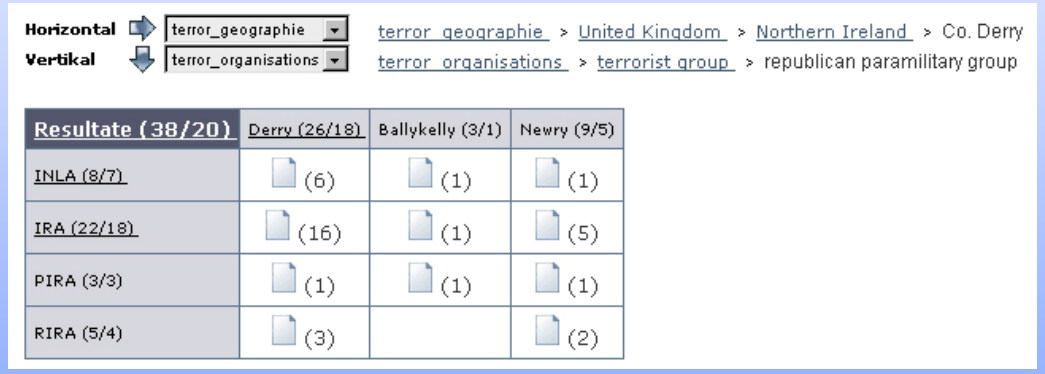
Beispielsuche:
"car bomb"

Ausgabe: Geographie-Facette des eingesetzten Thesaurus

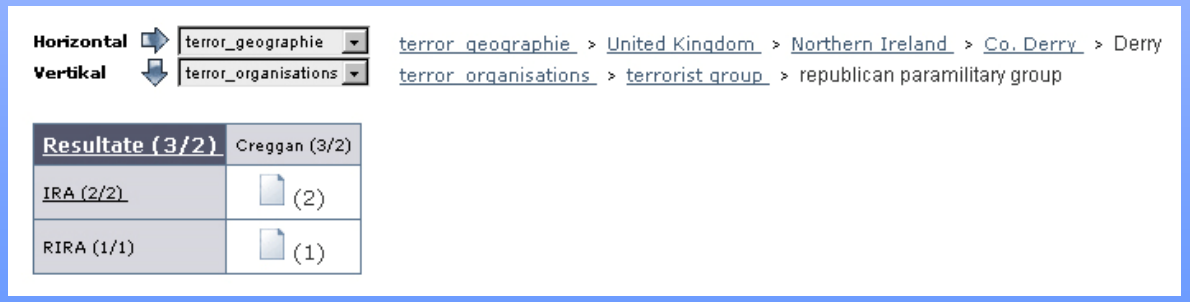
G.2 Visuelle Retrievalhilfsmittel

"Dynamische" Klassierung beim Einsatz mehrerer Begriffsordnungen (oder mehrerer Facetten)

Beispielsuche "car bomb"



anfragesensitives Stöbern



Kapitel G.3

Sprachübergreifendes Retrieval

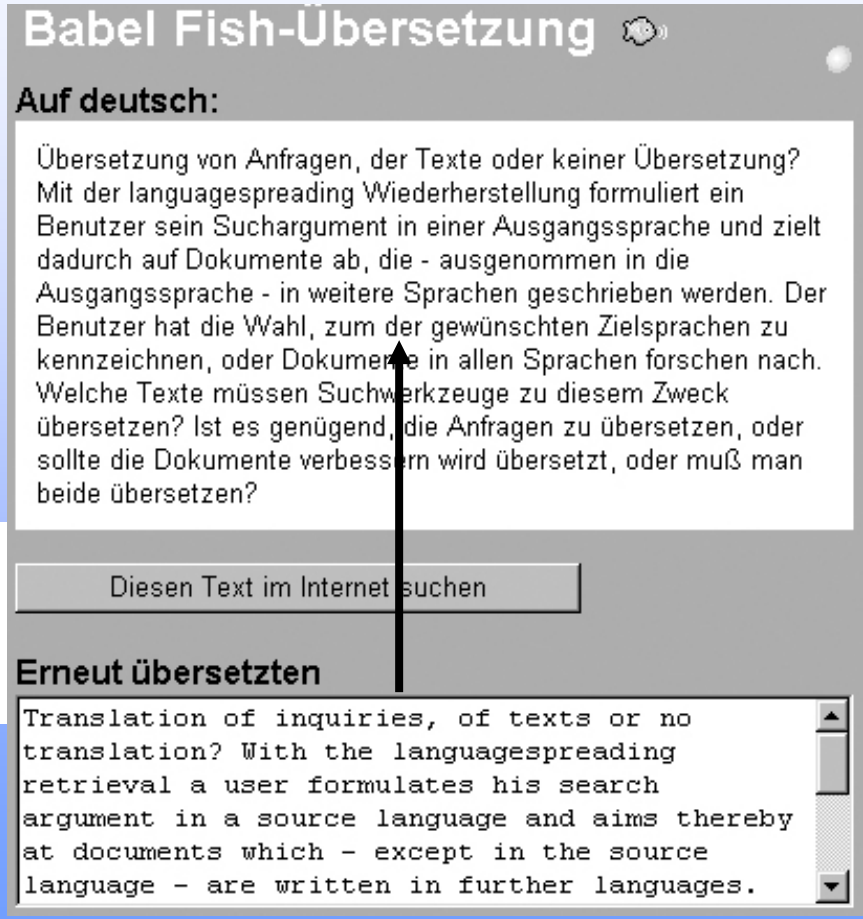
G.3 Sprachübergreifendes Retrieval

Derzeitiger Stand automatischer Übersetzung: suboptimal

- **Test: Übersetzung in Zielsprache - anschließende Rückübersetzung des Ergebnisses in Ausgangssprache**

Übersetzung von Anfragen, von Texten oder gar keine Übersetzung?

Beim sprachübergreifenden Retrieval formuliert ein Nutzer sein Suchargument in einer Ausgangssprache und zielt dabei auf Dokumente, die – außer in der Ausgangssprache – in weiteren Sprachen verfasst sind.



G.3 Sprachübergreifendes Retrieval

Sprachen

meistgesprochen

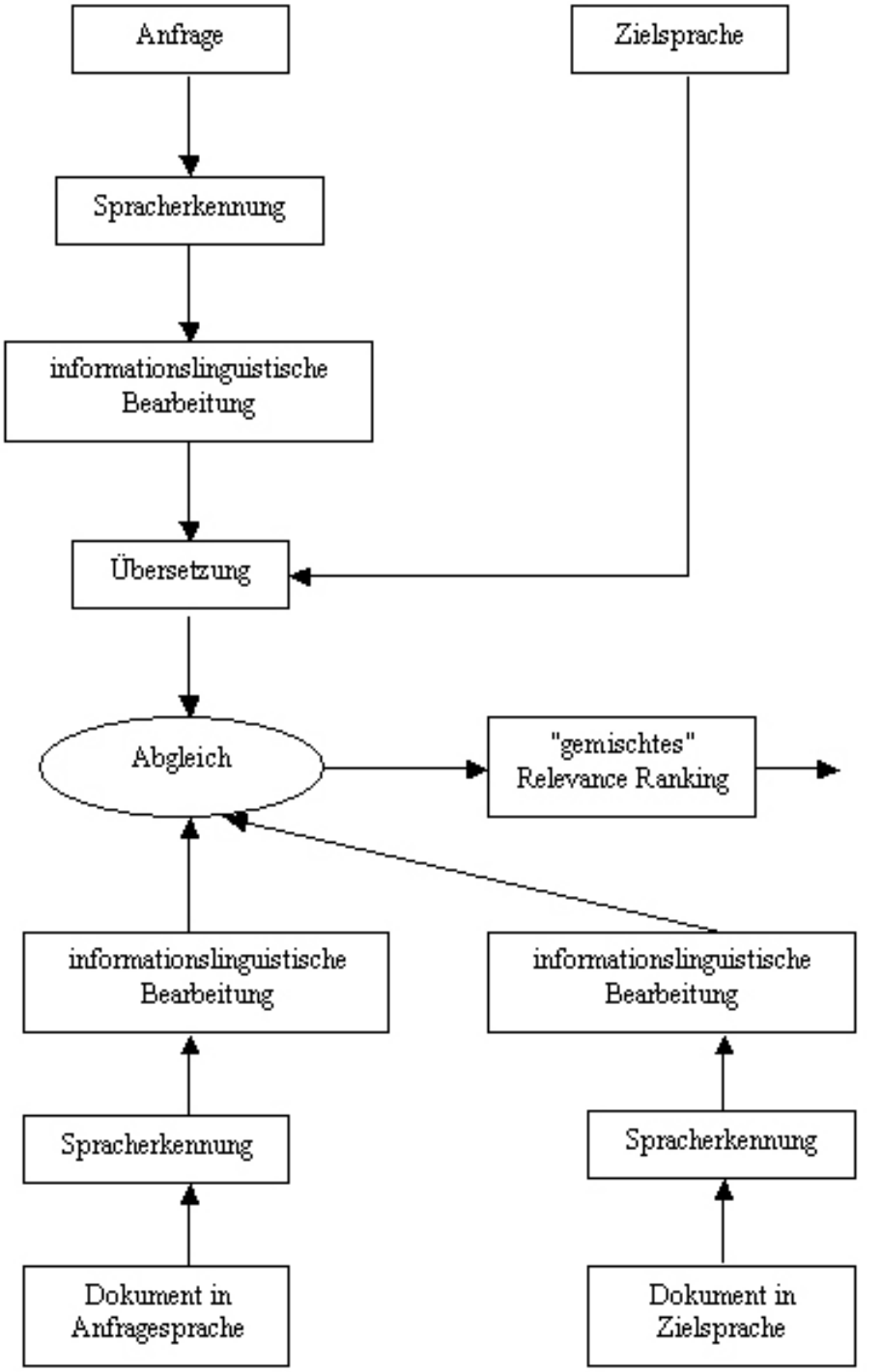
Rang	*Sprache*	*Anzahl Sprecher*
1	chinesisch (Mandarin)	885 Mio.
2	spanisch	332 Mio.
3	englisch	322 Mio.
4	bengalisch	189 Mio.
5	Hindi	182 Mio.
6	portugiesisch	170 Mio.
7	russisch	170 Mio.
8	japanisch	125 Mio.
9	deutsch	98 Mio.
10	chinesisch (Wu)	77 Mio.

mit meisten Webseiten (2003)

Rang	*Sprache*	*Anzahl Webseiten*
1	englisch	1.690 Mio.
2	deutsch	257 Mio.
3	französisch	121 Mio.
4	russisch	98 Mio.
5	japanisch	95 Mio.
6	chinesisch (vereinfacht)	93 Mio.
7	spanisch	89 Mio.
8	italienisch	78 Mio.
9	koreanisch	66 Mio.
10	niederländisch	65 Mio.

Gey, F.C., Kando, N., & Peters, C. (2005). Cross-language information retrieval: The way ahead. Information Processing & Management, 41, 415-431.

Arbeitsschritte beim CLIR



G.3 Sprachübergreifendes Retrieval

CLIR/MLIR - Ansätze

- **Übersetzung der Anfrage mittels Wörterbuch**
  - **ausschließlich Sprachwissen**
- **Übersetzung der Anfrage mittels Thesaurus**
  - **Sprach- und Weltwissen**
- **MLIR unter Nutzung paralleler Korpora**

G.3 Sprachübergreifendes Retrieval

Maschinenlesbares Wörterbuch

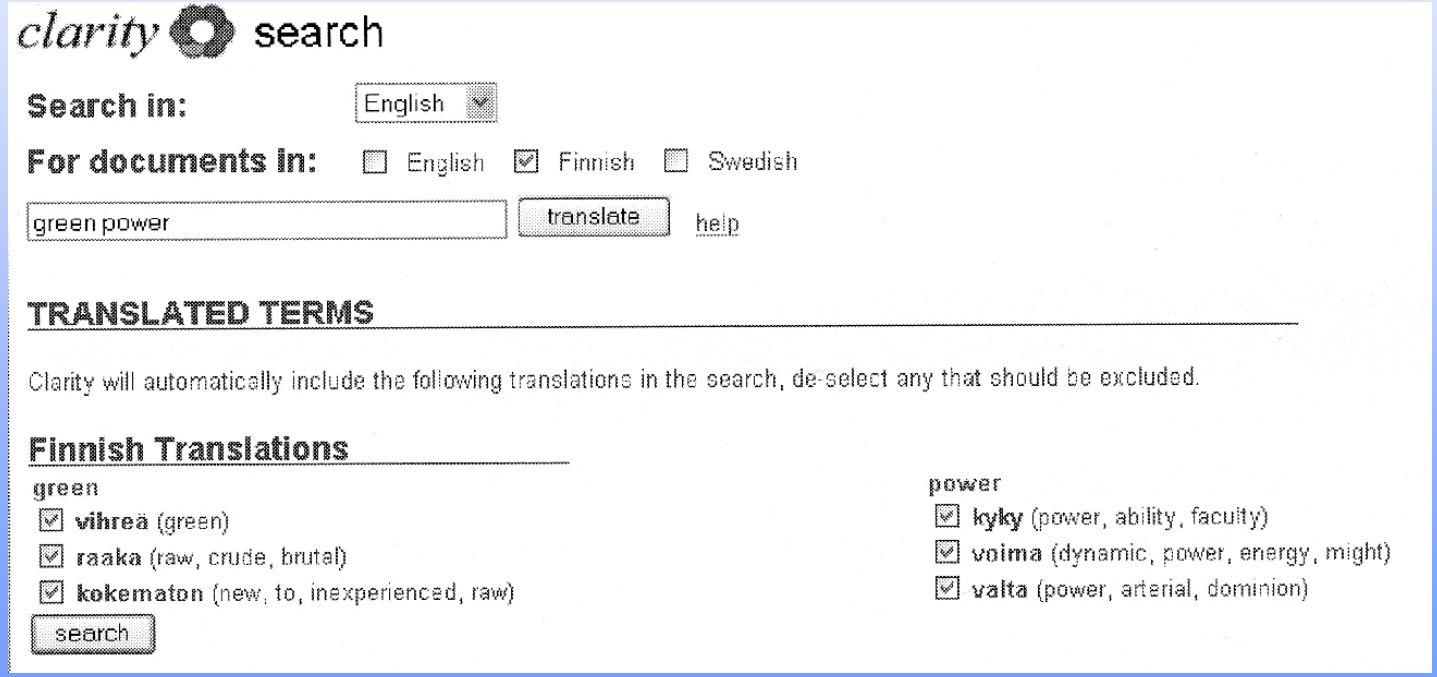
- Problem: Mehrdeutigkeit der meisten Begriffe
- Lösungsoption: Relevance Feedback
  - vor der Übersetzung (also in Originalsprache)
    - zum Finden von Musterdokumenten
    - Markierung der wichtigsten Terme
    - Hinzufügen zu den ursprünglichen Suchtermen
  - Übersetzung der Suchterme
  - nach der Übersetzung
    - zum Finden relevanter Dokumente ("richtige" Übersetzung)
    - Termmarkierung und Nutzung in der nächsten Retrievalrunde
    - zum Finden nicht relevanter Dokumente ("falsche" Übersetzung)
    - Termmarkierung und Ausschluss in der nächsten Runde

Ballesteros, L., & Croft, W.B. (1998). Statistical methods for cross-language information retrieval. In G. Grefenstette (Ed.), Cross-Language Information Retrieval (pp. 23-40). Boston, MA: Kluwer.

G.3 Sprachübergreifendes Retrieval

Maschinenlesbares Wörterbuch

- **von Vorteil: Dialog mit Nutzer**



G.3 Sprachübergreifendes Retrieval

Multilinguale Fachthesauri

- **Begriffe: sprachunabhängig definiert**
- **Übersetzungen in alle Zielsprachen**
- **Problem: erfordert 1. Multilingualen Thesaurus und 2. Datenbasis, die damit indexiert worden ist**

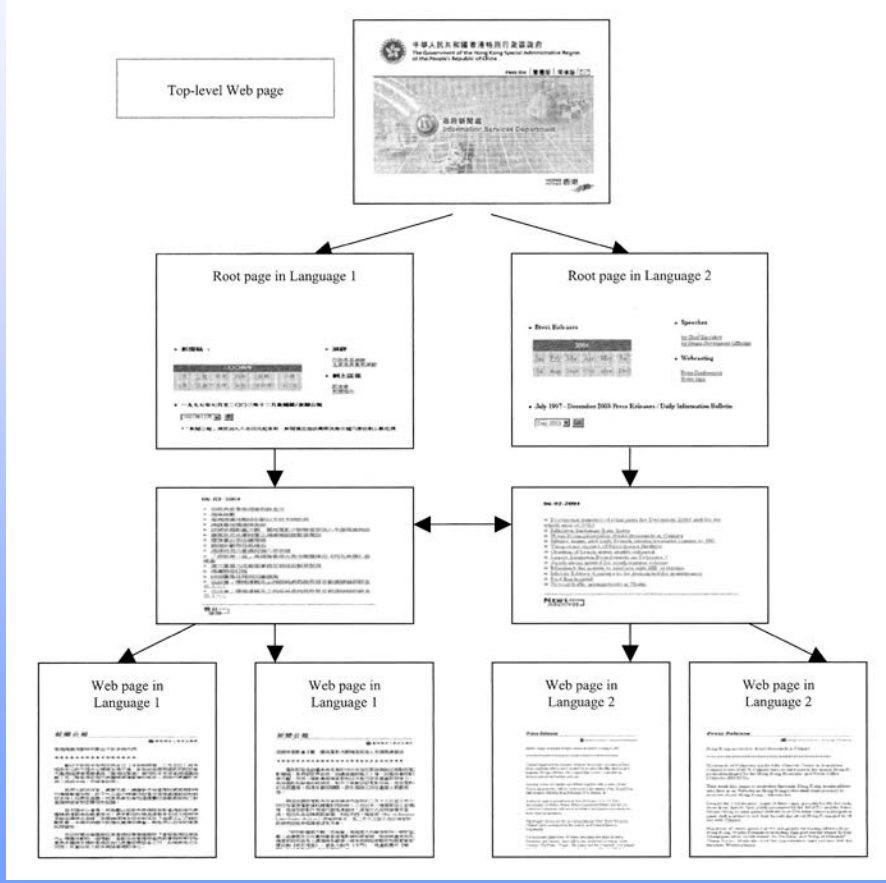
Information retrieval
MT 5.35 Documentary information processing
FR Récupération de l'information
SP Recuperación de información
UF Bibliographic searches, Literature searches, Retrospective searches
NT1 Online searching
NT2 Search strategies
RT Bibliography compilation
RT Databases
RT Indexing languages
RT Information processing
RT Information systems
RT Information systems evaluation
RT Information/library networks
RT Reference services

Quelle:
UNESCO Thesaurus

G.3 Sprachübergreifendes Retrieval

Korpus-basierte Methode

- **Vor.: parallele Korpora (z.B. EU-Dokumente)**
- **initiale Suche in Sprache des Nutzers**
- **Relevance Ranking der Dokumente; Nutzung der ersten n (z.B. 10 Dokumente)**
- **(ggf.) innerhalb der Dokumente: Aufspüren der bestpassenden Textstelle**
- **Aufsuchen der n Dokumente in der Zielsprache (und darin der jeweiligen Textstellen)**
- **Pseudo-Relevance Feedback: neue Suche (nunmehr in der Zielsprache) mit den Termen der Textstellen (Croft-Harper-Formel)**

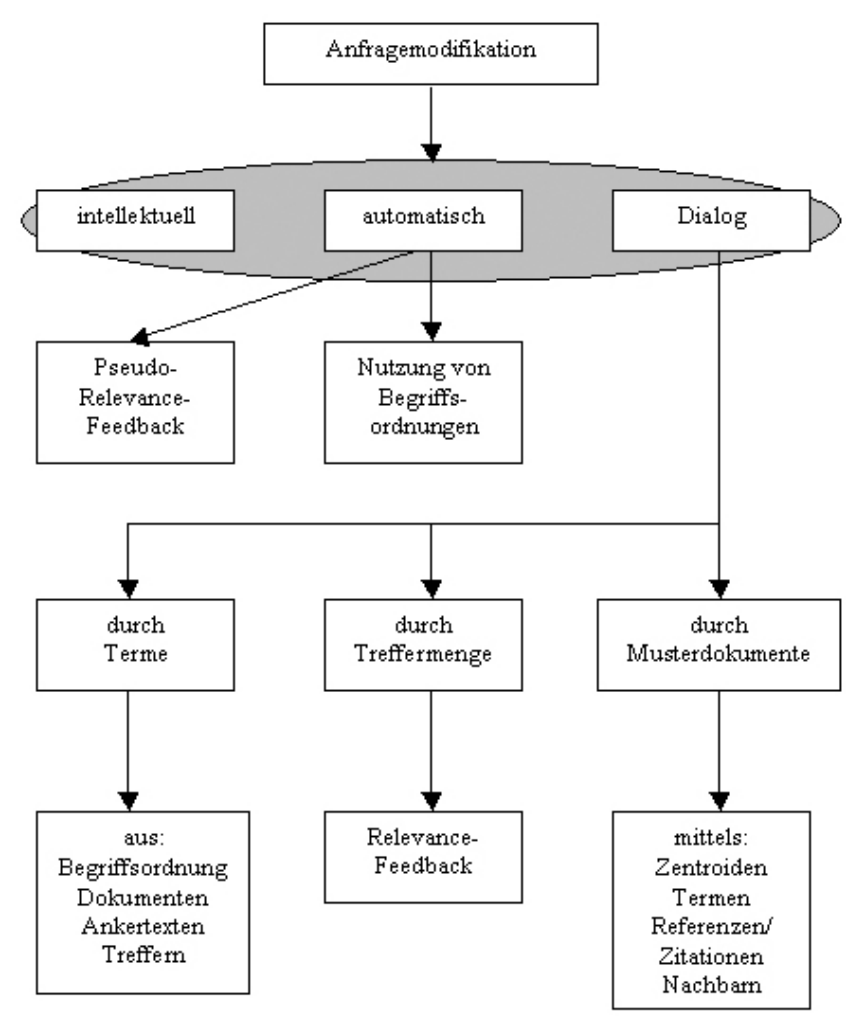


Li, K.W., & Yang, C.C. (2006). Conceptual analysis of parallel corpus collected from the Web. Journal of the American Society for Information Science and Technology ,57, 632-644.

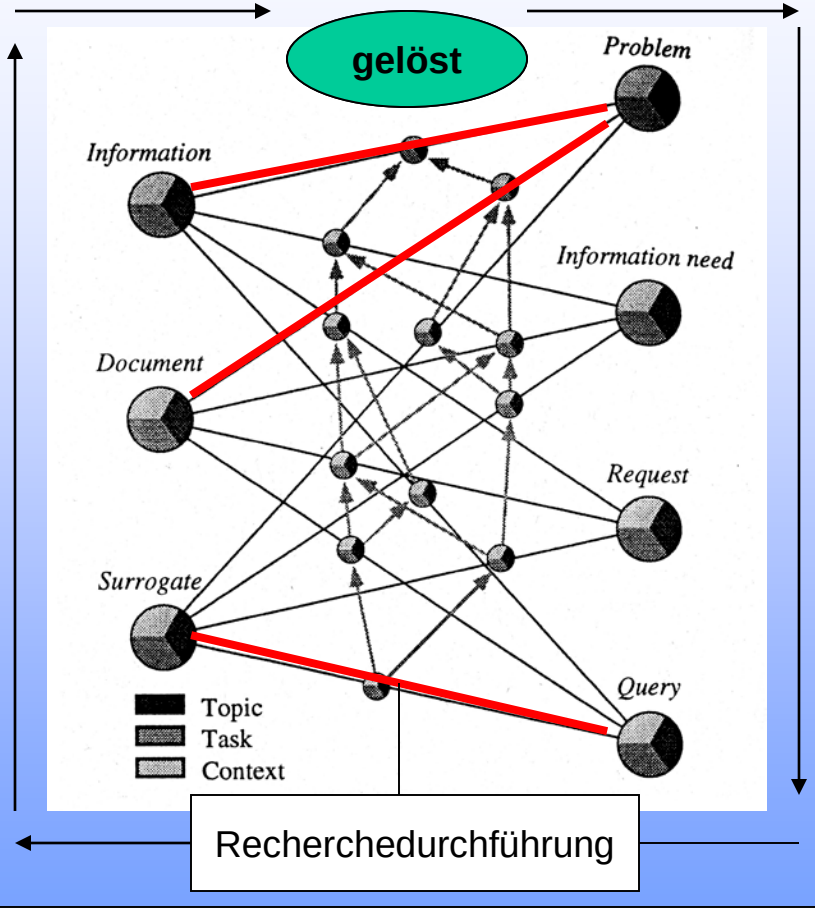
Kapitel G.4

(Halb-)Automatische Anfrageerweiterung

G.4 (Halb-)Automatische Suchanfrageerweiterung



G.4 (Halb-)Automatische Suchanfrageerweiterung



Rechercheinterview

Rechercheaufbereitung

Recherchedurchführung

Verändert nach: Mizzaro, S. (1997). Relevance: The whole story.
Journal of the American Society for Information Science, 48, 810-832.

G.4 (Halb-)Automatische Suchanfrageerweiterung

Rechercheinterview

- **Phasen des Interviews**
  - **Kontaktphase (positive Gesprächsatmosphäre, Vertrauensbildung, Kennenlernen: insbesondere Stellung des Auftraggebers im Unternehmen)**
  - **Problemdefinition (Auftraggeber beschreibt Informationsbedürfnis, Vermittler hört zu und umschreibt Verstandenes mit eigenen Worten)**
  - **Bedingungsanalyse (Für was braucht der Auftraggeber das Wissen? Was kann er: Sprachen, Fachkenntnisse? Wie viel Zeit hat er, beschaffte Dokumente durchzuarbeiten? Geld für nicht-abonnierte Datenbanken vorhanden? - Achtung: Auftraggeber darf sich nicht "ausgefragt" fühlen!)**
  - **Zielanalyse (genaue Formulierung des Informationsbedarfs; Auftraggeber muss nunmehr wissen, ob und inwieweit eine Recherche bei seinem Problem helfen kann)**
  - **Vorgehen (iterativ: Rückkopplung zwischen Vermittler und Auftraggeber, oder einmalig: Vermittler liefert die Ergebnisse)**

Schmidt, R. (1995). Kuck mal, wer da fragt! Nutzeranalyse und Empathie in der Informationsvermittlung. In 17. Online-Tagung der DGD. Proceedings (pp. 111-126). Frankfurt/M.: DGD.

G.4 (Halb-)Automatische Suchanfrageerweiterung

Fixpunkte der Informationsvermittlung

• **das richtige Wissen**	**pertinente Dokumente**
• **zum richtigen Zeitpunkt**	**wann braucht es Nutzer?**
• **am richtigen Ort**	**wohin geliefert?**
• **im richtigen Umfang**	**wie viel?**
• **in der richtigen Form**	**Print? Digital?**
• **mit der richtigen Qualität**	**Aufbereitungsgrad / Bearbeitung der Rechercheergebnisse**

G.4 (Halb-)Automatische Suchanfrageerweiterung

Aufbereitungsgrad bei der Informationsvermittlung

- **einfache Recherche (Lieferung bibliographischer Nachweise; Volltexte entweder nach Bestellung beschaffen, oder Auftraggeber bedient sich selbst) - eher selten**
- **Volltexte - unstrukturiert (Lieferung der einschlägigen Dokumente)**
- **synthetische Informationsvermittlung (strukturierte Zusammenstellung der Dokumente in einem Reader, mit Inhaltsverzeichnis und ggf. Register)**
- **synoptische Informationsvermittlung (synthetische Informationsvermittlung, die den ermittelten Stand des Wissens in einem eigenen kurzen Text beschreibt: Literaturbericht, State-of-the-Art, Neuheitsrecherche, "Weltstandsvergleich")**
- **analytische Informationsvermittlung (Verdichtung des recherchierten Wissens in einen kurzen Text; Dokumente werden gar nicht mitgeliefert oder - wenn doch - dann nur in einem Anhang)**

Kapitel G.5

Recommendersysteme

G.5 Recommendersysteme

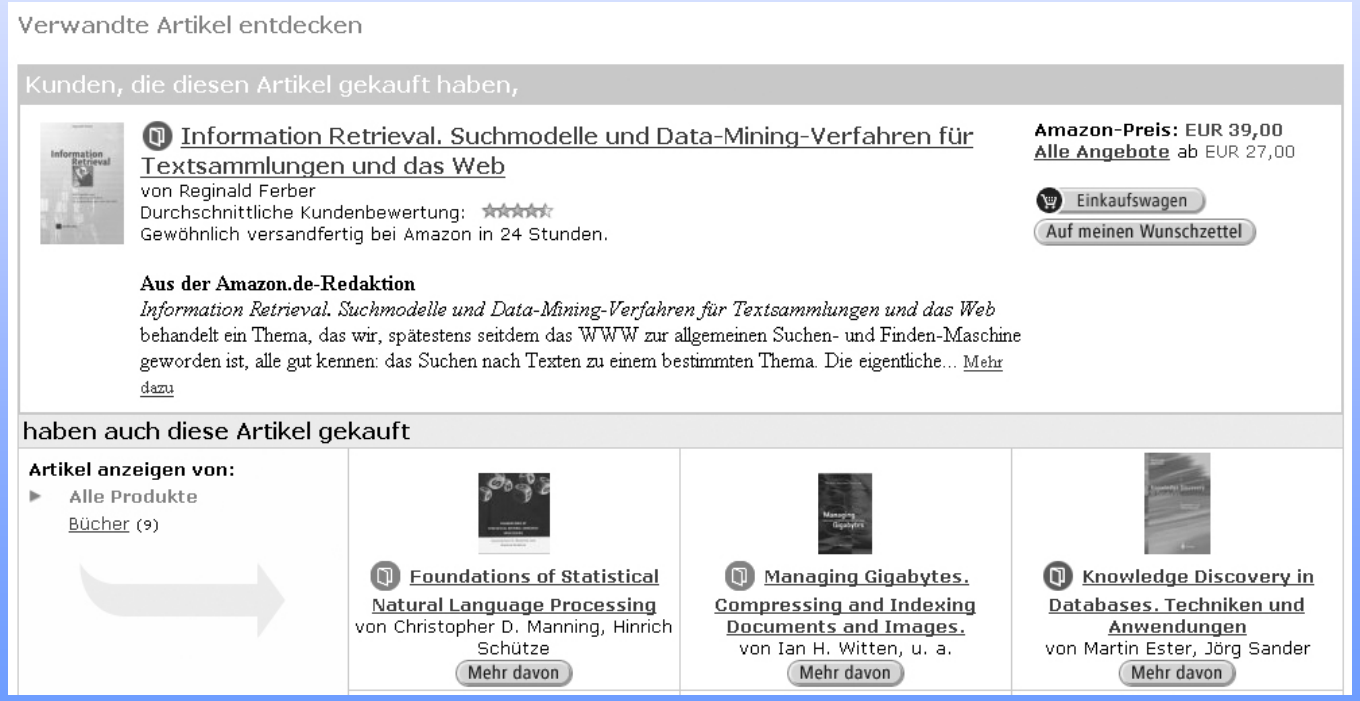
Empfehlungssysteme / Recommendersysteme

- **"collaborative recommender systems": Vergleich von Nutzer zu Nutzer**
  - **explizit (Bewertungen)**
  - **implizit (Nutzerverhalten)**
- **"content-based recommendation": Vergleich von Nutzer(vorlieben) und Dokumentinhalten**
- **Hybridsysteme (Vereinigung beider Ansätze)**
  - ***Beispiel:* Amazon**

G.5 Recommendersysteme

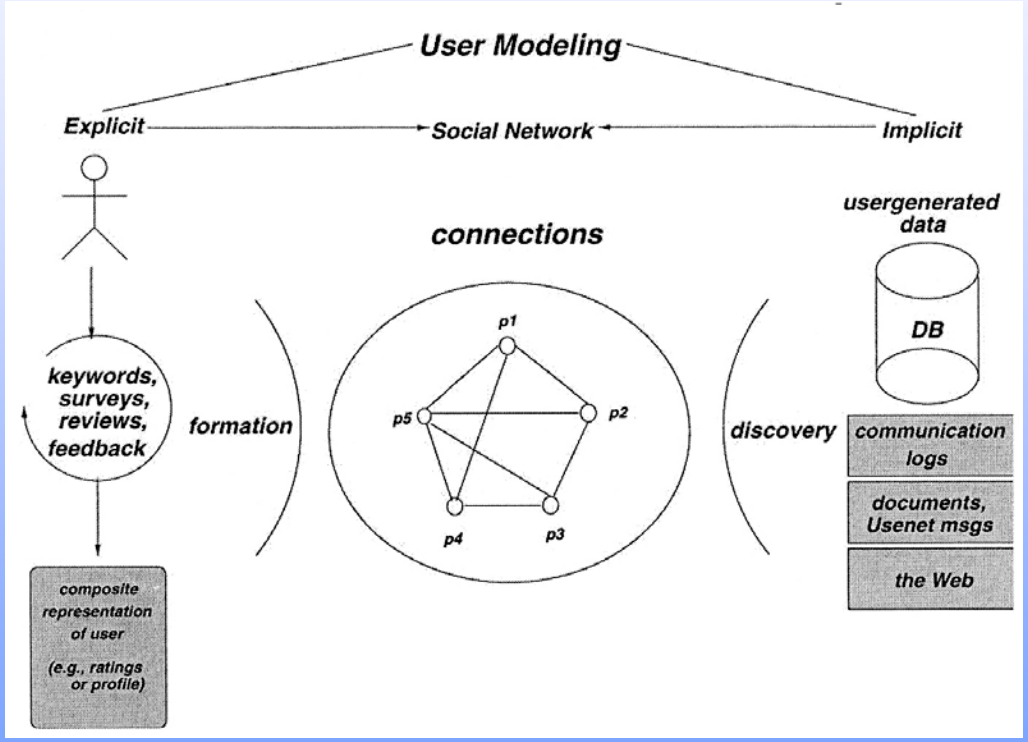
Kooperative Recommendersysteme

- **Feststellung der Ähnlichkeit zwischen Nutzern**
- **Übertragung von Gewohnheiten / Bewertungen ähnlicher Nutzer**



G.5 Recommendersysteme

Explizite und implizite kooperative Systeme



Perugini, S., Goncalves, M.A., & Fox, E.A. (2004). Recommender systems research: A connection-centric study. Journal of Intelligent Information Systems, 23(2), 107-143.

G.5 Recommendersysteme

(Implizite) Kooperative Systeme

Vorgehen: Einsatz des Vektorraummodells

- **Dokumente: Dimensionen des Vektorraums**
- **Nutzer: Vektoren**
- **Ähnlichkeit zwischen Nutzern: Cosinus der Winkel zwischen den Nutzer-Vektoren**
- **all das, was die ähnlichsten Nutzer angesehen, gekauft oder positiv bewertet haben (und was der Ausgangsnutzer nicht kennt), wird dem aktuellen Nutzer vorgeschlagen**

$$\text{COSINUS}(\text{DOK}_i, \text{ANFRAGE}_j) = \frac{\sum_{k=1}^{t} (\text{TERM}_{ik} \cdot \text{ATERM}_{jk})}{\sqrt{\sum_{k=1}^{t} (\text{TERM}_{ik})^2 \cdot \sum_{k=1}^{t} (\text{ATERM}_{jk})^2}}$$

G.5 Recommendersysteme

Inhaltsbasierte, nutzerorientierte Systeme

- **Feststellung eines (inhaltlichen) Nutzerprofils**
- **Vorschläge, die dem Profil entsprechen**

G.5 Recommendersysteme

Inhaltsbasierte, nutzerorientierte Systeme

Vorgehen: analog zu SDIs

- **Nutzer hinterlegt Informationsprofil**
  - **durch explizite Bewertungen**

  - **oder durch Aktionen (Anschauen, Käufe)**

G.5 Recommendersysteme

Inhaltsbasierte, nutzerorientierte Systeme

- **Analyse des Informationsprofils**
  - **Titelterme**
  - **Sachgebiet**
  - **Autor**
  - **usw.**
- **oder Vektorraummodell: Nutzer als Dimensionen und Dokumente als Vektoren**
- **periodisch Retrievalläufe zum Profil (mit Relevance Ranking)**
- **oder: ausgehend von einem aktuell angeschauten Musterdokument**
- **Ausgabe der ähnlichsten Dokumente**

G.5 Recommendersysteme

Hybridsysteme
Vereinigung beider Ansätze

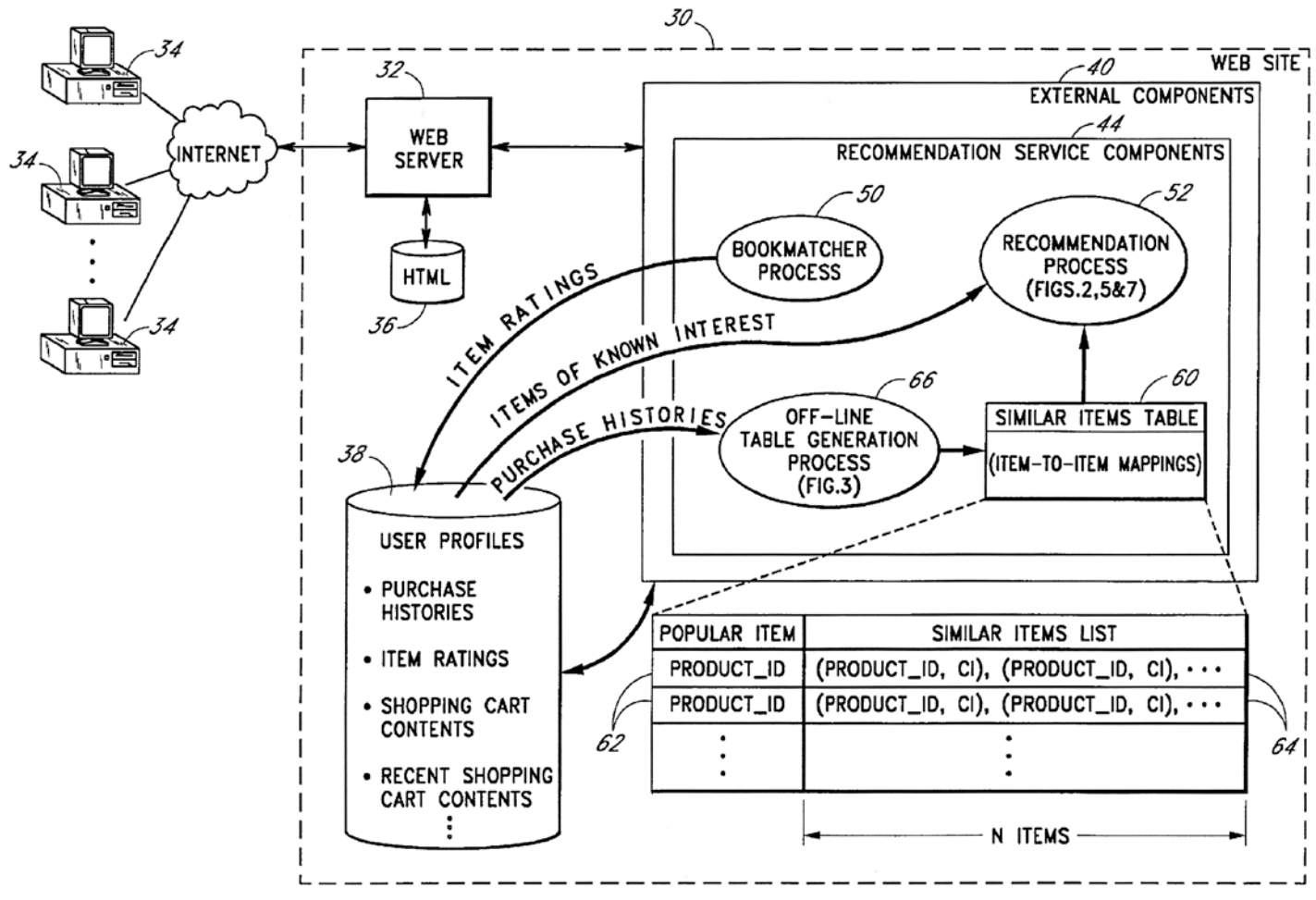
Beispiel: **Amazon**

- **"item-to-item collaborative filtering"**
- **Nutzerprofil: Liste bereits gekaufter Produkte (items) - Nutzer kann sein Profil bearbeiten**
- **Sortierung der Vorschläge: jeweils nach Cosinus**

Linden, G.D., Jacobi, J.A., & Benson, E.A. (1998). Collaborative recommendations using item-to-item similarity mappings. Patent-Nr. US 6,266,649.

Linden, G.D., Smith, B., & York, J. (2003). Amazon.com recommendations. Item-to-item collaborative filtering. IEEE Internet Computing, 7(1), 76-80.

G.5 Recommendersysteme



Beispiel:
Amazon

G.5 Recommendersysteme

Probleme von Recommendersystemen

- **Missbrauch bei expliziten Bewertungen**
  - **eigene Produkte positiv bewerten**
  - **Produkte der Wettbewerber abwerten**
- **Sind die aktiven Bewerter wirklich ähnlich dem "Durchschnittsnutzer"?**
- **Privatsphäre des Nutzers**

Kapitel G.6

Retrieval von Textstellen und Frage-Antwort-Systeme

G.6 Retrieval von Textstellen

- **"Globaler" Ansatz des IR**
  - **Ranking von Dokumenten als Ganzes**
  - **aber: manche Dokumente sind sehr lang (z.B. Patent-schriften, Dissertationen)**
- **"Lokaler" Ansatz des IR**
  - **Ranking von Dokumenten nach "passages"**
- **Textstellen / "passages"**
  - **Subeinheiten dokumentarischer Bezugseinheiten**
    - **Diskursstellen (linguistisch orientiert: Sätze, Absätze, Kapitel)**
    - **semantische Textstellen (am Gegenstand orientiert)**
    - **Textausschnitt (Textfenster gewisser Größe)**

Callan, J.P. (1994). Passage-level evidence in document retrieval.
In Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 302-310). New York, NY: ACM.
Kaszkiel, M., Zobel, J., & Sacks-Davis, R. (1999). Efficient passage ranking for document databases.
ACM Transactions on Information Systems, 17(4), 406-439.

G.6 Retrieval von Textstellen

Ranking von Dokumenten nach ihren jeweils bestpassenden Stellen

- **Basis: Textfenster ("arbitrary passages")**

- **Ansatz 1: Textfenster fester Länge (100 ... 350 Worte)**
- **Ansatz 2: theoretisch optimal: Textfenster beliebiger Länge, aber viel zu rechenintensiv**
- **nur genau ein Textfenster oder mehrere?**

Kaszkiel, M., & Zobel, J. (1997). Passage retrieval revisited.
In Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 178-185). New York, NY: ACM.
Kaszkiel, M., & Zobel, J. (2001). Effective ranking with arbitrary passages.
Journal of the American Society for Information Science and Technology, 52, 344-364.

G.6 Retrieval von Textstellen

Ranking von Textstellen innerhalb eines (langen) Dokuments

- Ausgang: Dokument gefunden
- Aufgabe: Anzeige der besten Textstelle(n) zur Suchanfrage ("within-document retrieval")

- Textfenster fester Länge (etwa 200 Worte)
- Retrievalmodell: statistisches Sprachmodell

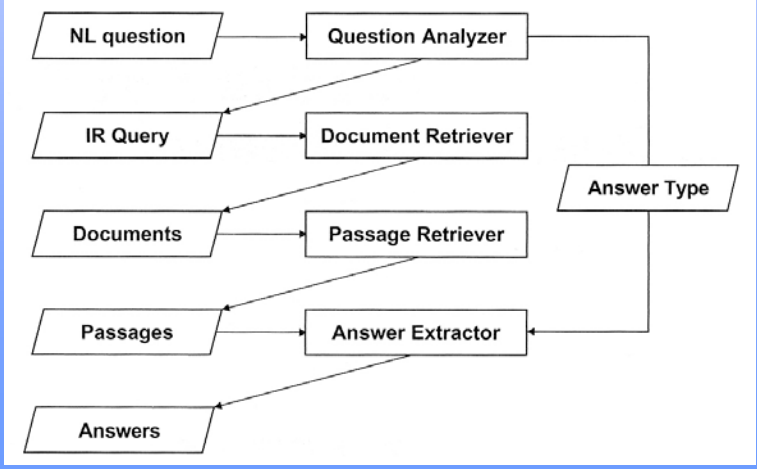
- Option: bestes Textfenster auf Absatzebene erweitern (und gesamten Absatz ausgeben)
- bei sich überlappenden Textfenstern: nur das beste ausgeben

Harper, D.J., Koychev, I., Sun, Y., & Pirie, I. (2004). Within-document retrieval: A user-centred evaluation of relevance profiling. Information Retrieval, 7, 265-290.

G.6 Retrieval von Textstellen

Frage-Antwort-Systeme

- **gegeben: konkreter Informationsbedarf / Faktenfrage**
- **Ziel: Ausgabe der besten Stelle aller Dokumente der Datenbasis als Antwort**
- **Verbindung aus Passage Retrieval und Within-Document Retrieval**
- **im "besten" Textfenster: Extraktion desjenigen Satzes (ggf. Absatzes), in dem die Suchargument möglichst dicht nebeneinander vorkommen**



Lin, J., & Katz, B. (2006).
Building a reusable test collection for question answering.
Journal of the American Society for Information Science and Technology, 57, 851-861.