Zürcher Hochschule
für Angewandte Wissenschaften

zh
aw

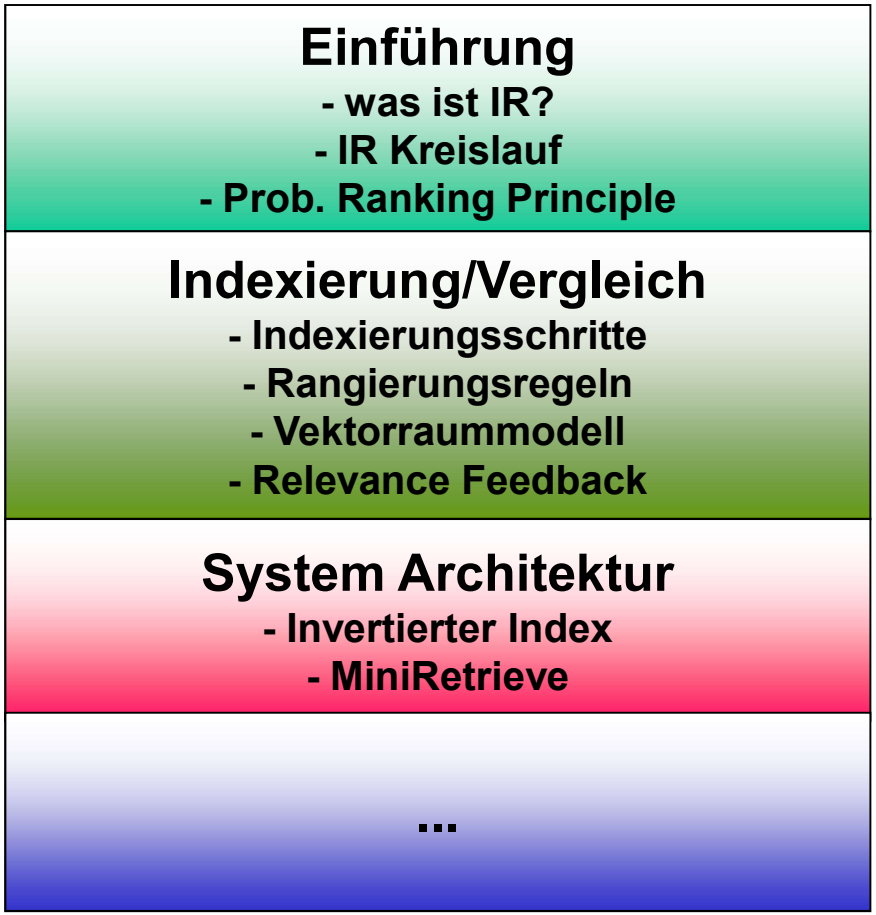
Information Engineering 1: Information Retrieval

Einführung

Kapitel 1

Martin Braschler

"Coming Up Next"



"You are here"

Inhalt

- Definition „Information Retrieval“
- Daten, Information, Wissen
- Das „Retrievalproblem“
- IR-Prozess
- Information Retrieval Paradigmen
- IR im Vergleich zu Datenbanksuche
- IR im Vergleich zu „Textsuche“
- Retrievaleffektivität
- „Probability Ranking Principle“
- Geschichte

Frage:

- Was ist Information Retrieval?

Definition „Information Retrieval“

- „Das akademische Fachgebiet, welches Methoden untersucht, um grosse Mengen an unstrukturierter und strukturierter Information zu organisieren und bedürfnisgerecht aufzufinden.“
- Zugriff erfolgt im Allgemeinen in Form einer „Anfrage“ (drückt Informationsbedürfnis mehr oder weniger treffend aus).
- Resultat im Allgemeinen in Form einer Rangliste von Dokumenten, (die die gesuchte Information potentiell enthält).

Definition „Information Retrieval“

- Information Retrieval wird im oft mit Retrieval auf unstrukturiertem Volltext in der Form von natürlichsprachigen Dokumenten gleichgesetzt. Es werden fortgeschrittene Indexierungs- und Gewichtungsmethoden verwendet.
- IR als akademisches Feld befasst sich auch mit verwandten Problemen, wie:
  - Kategorisierung
  - automatisches Zusammenfassen
  - u. a.

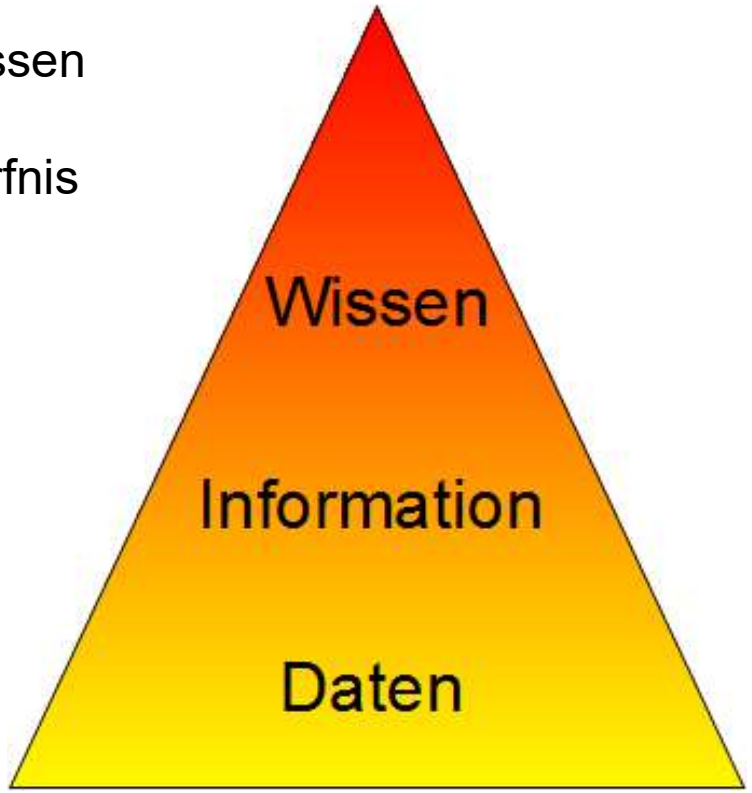
Frage:

- Was ist der Unterschied zwischen Daten, Information und Wissen?

Daten-Information-Wissen

Definitionen im Rahmen von Information Retrieval:

- Daten:
  - Fakten in einer codierten Form (z. B. Zeichenketten). Daten tragen per se keine Bedeutung.
- Information:
  - Daten plus die Bedeutung, die ihnen beigemessen wird. Information ist relevant oder irrelevant. Information ist immer an ein Informationsbedürfnis gebunden.
- Wissen:
  - das Resultat der Verarbeitung von Information (Schlüsse, Erkenntnisse)



Was sind Daten?

- Daten bestehen aus Zeichen- bzw. Symbolketten.
- Daten sind in einem bestimmten Format (z.B. RTF, XML, JPEG).
- Daten können falsch oder korrekt sein (Die Angaben in einer Adressdatenbank können beispielsweise korrekt oder falsch sein).

Frage:

- Wie viele Daten gibt es?
- Wo fallen diese Daten an?

Wieviele Daten?

- Die Frage ist natürlich etwas heikel, u. a. wegen der Frage, wie mit Kopien umgegangen wird
- Auch Fragen der betrachteten Datenträgertypen können mitspielen

- → Es geht aber darum, eine Idee zu bekommen, was die Grössenordnung ist, und damit ein Gefühl für die Problemgrösse

Was sind Daten?

- **How much Information 2003** von Peter Lyman und Hal Varian von der University of California's Berkeley School of Information Management Sciences
  - http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/
- Im Jahr 2002 wurden 5 Exabyte Daten produziert

  (d.h 5 Milliarden Gigabyte) oder 800MB pro Person.
- Wachstum zwischen 1999 bis 2002 ca. 30% pro Jahr

Frage:

- Merkt jemand was den Dozenten am Titel der Studie stört?

Zettabytes….

- 2002 ist schon lange her...

- IDC schätzte für 2006 einen globalen Daten-Ouptut von 161 Exabytes
- Für 2007 musste die Schätzung nach oben revidiert werden: 281 Exabytes
- Für diejenigen von uns, die immer noch in Gigabytes rechnen: das sind 281,000,000,000 Gigabytes.

- Das IDC schätzte 2006 eine Versechsfachung dieser Zahl bis 2010.
- Für 2011 ging man dann von einer eine Verzehnfachung dieser Zahl ggü. 2006 aus

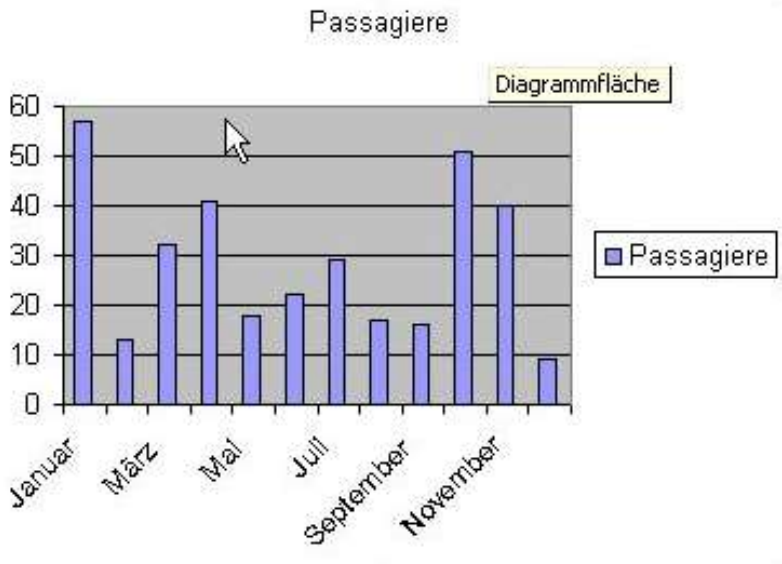
- EMC/IDC 2014: « [..] by 2020 containing nearly as many digital bits as there are stars in the universe»
- The data we create and copy annually by 2020: 44 Zettabyes = 44,000,000,000,000 GB (Volumen 2020 = 10 * Volumen 2013)
- Statista: 163 ZB by 2025

Was ist Information?

- Information benötigt man, um Aufgaben zu erledigen.
- Information wird mit Hilfe von Daten dargestellt.
- Information ist bezüglich einer Aufgabe mehr oder weniger relevant.
- Information ist bezüglich einer Aufgabe mehr oder weniger vollständig.

Was ist Information?

- Gleiche Information, unterschiedlicher Speicherbedarf



15.0 KByte

Monat	Passagiere
Januar	57
Februar	13
März	32
April	41
Mai	18
Juni	22
Juli	29
August	17
September	16
Oktober	51
November	40
Dezember	9

148 Byte

Was ist Wissen?

- Wissen ist vernetzte Information, z.B. über einen Geschäftsprozess.
- Beispiele:
  - Bestellung eines Kunden abwickeln
  - Flugzeug warten (primär explizites Wissen)
  - Marketing-Strategie entwickeln (primär: implizites Wissen)
- Häufig müssen interne und externe Informationen vernetzt werden.

Frage:

- Wir beschäftigen uns ja nun nicht mit DATENbanken, sondern mit INFORMATION Retrieval-Systemen. Was sollte also unser Ziel an ein solches System sein?

Das „Retrievalproblem“

- Retrievalproblem: „Das Auffinden von möglichst viel relevanter Information bei gleichzeitigem Minimieren der ebenfalls gelieferten irrelevanten Information.“
- Wir wollen nicht nur Informationen wieder finden, sondern vor allem neue Information finden. Die Information wird indirekt geliefert, in Form von „relevanten“ Dokumenten.

Das „Retrievalproblem“

- Sprache ist nicht „eindeutig“:
  - Synonyme (eine Bedeutung – mehrere Wörter)
  - Homonyme (mehrere Bedeutungen – ein Wort)
  - Umschreibungen
  - Metaphern
  - Wortformen (Singular, Plural, Verbformen, etc.) → Anfrage und Dokument „passen nicht zusammen“.
  - Schreibfehler
- Informationsbedürfnisse werden ungenügend verbalisiert
- Informationsbedürfnisse werden ungenügend formuliert
- Die Dokumente/Informationen im System sind unstrukturiert oder inhomogen
- Irreführender Inhalt
- Autorität, Quelle, Aktualität, Urheberrecht. Auch: Einsammeln der Dokumente
- Widersprüchliche Ziele: Ausbeute versus Präzision

Das „Retrievalproblem“

- Einige typische Annahmen für Volltextsuche:
  - Benutzer sucht „relevante Elemente“
  - Benutzer weiss wenig oder nichts über die gesuchten Elemente
  - Die Anzahl der relevanten Elemente ist unbekannt

- Was halten Sie von diesen Annahmen?

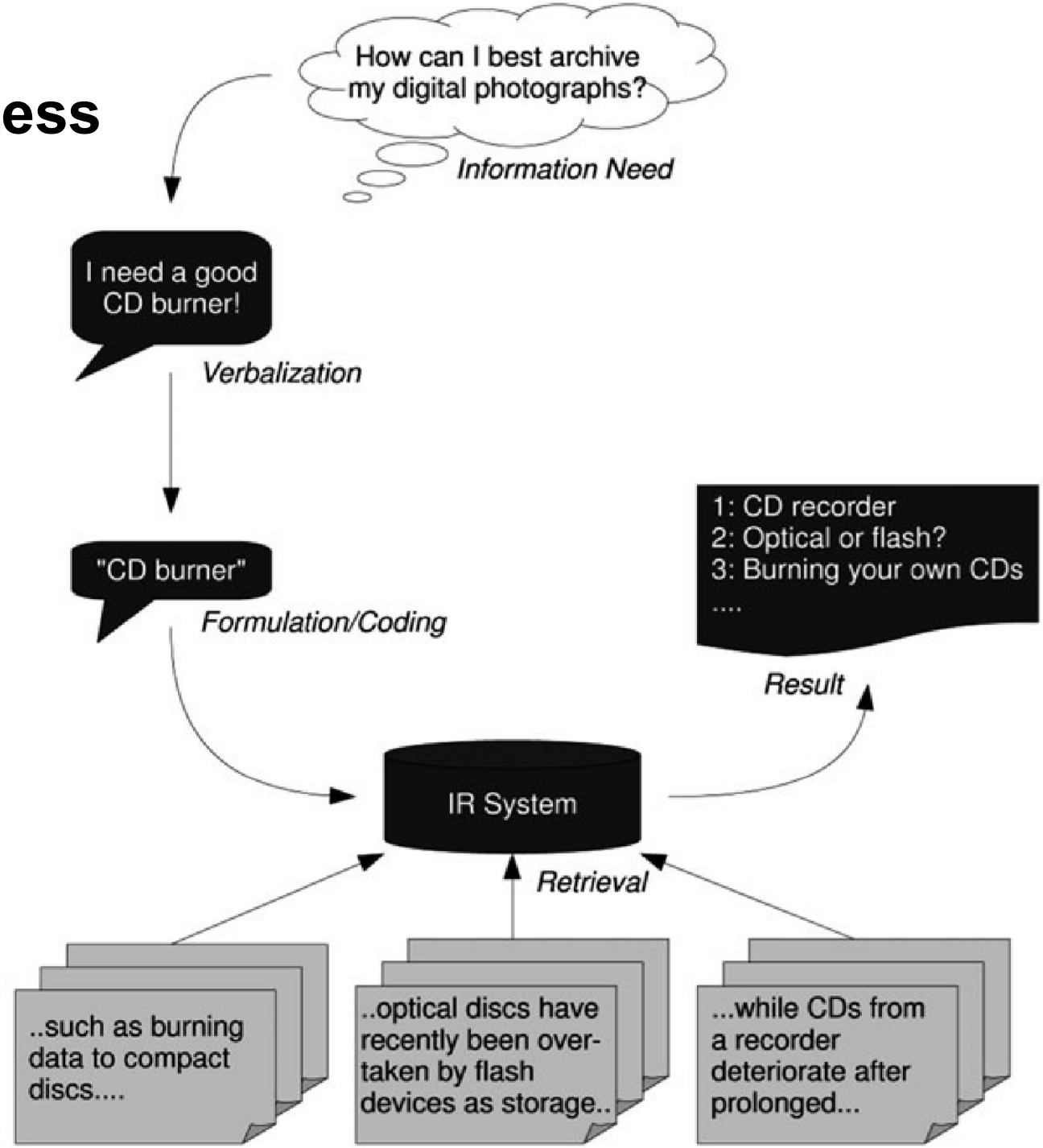
Das „Retrievalproblem“

- Der Begriff der Relevanz
- Das Verständnis einer Anfrage und eines Dokuments hängt immer auch vom konkreten Benutzer ab:
  - Vor-/Hintergrundwissen
  - Reihenfolge des Auffindens
  - Wandelnde Informationsbedürfnisse
  - Persönliche Präferenzen
  - Vollständigkeit der Antwort
- Folge:
  - Ein perfektes Retrievalresultat losgelöst von Benutzer und Kontext gibt es nicht.

Demo

- Schweizerisches Bundesgericht

IR-Prozess



Aus Peters et al., 2012

Verbalisierung/Codierung

- **Verbalisierung**: „Verstehen des Problems“: richtige Begriffe, Vollständigkeit
  - Paradox: muss Problem verstanden haben, um es richtig zu verbalisieren
- **Codierung**: „Verstehen des Systems“: richtige Operatoren, etc.
  - Paradox: muss die Resultate kennen, um Anfrage richtig zu codieren
- Das System kann a priori nur die **explizite** Information, welche in der „codierten“ Anfrage enthalten ist, auswerten.
- Das Resultat bevorzugt also Entscheide, welche bestmöglich zu dieser „codierten Anfrage“ passen. Nicht explizit formulierte Präferenzen oder Hintergrundwissen des Benutzers können nicht in die Resultatfindung einfliessen.
- → Das Resultat ist immer nur so gut wie die Anfrage!

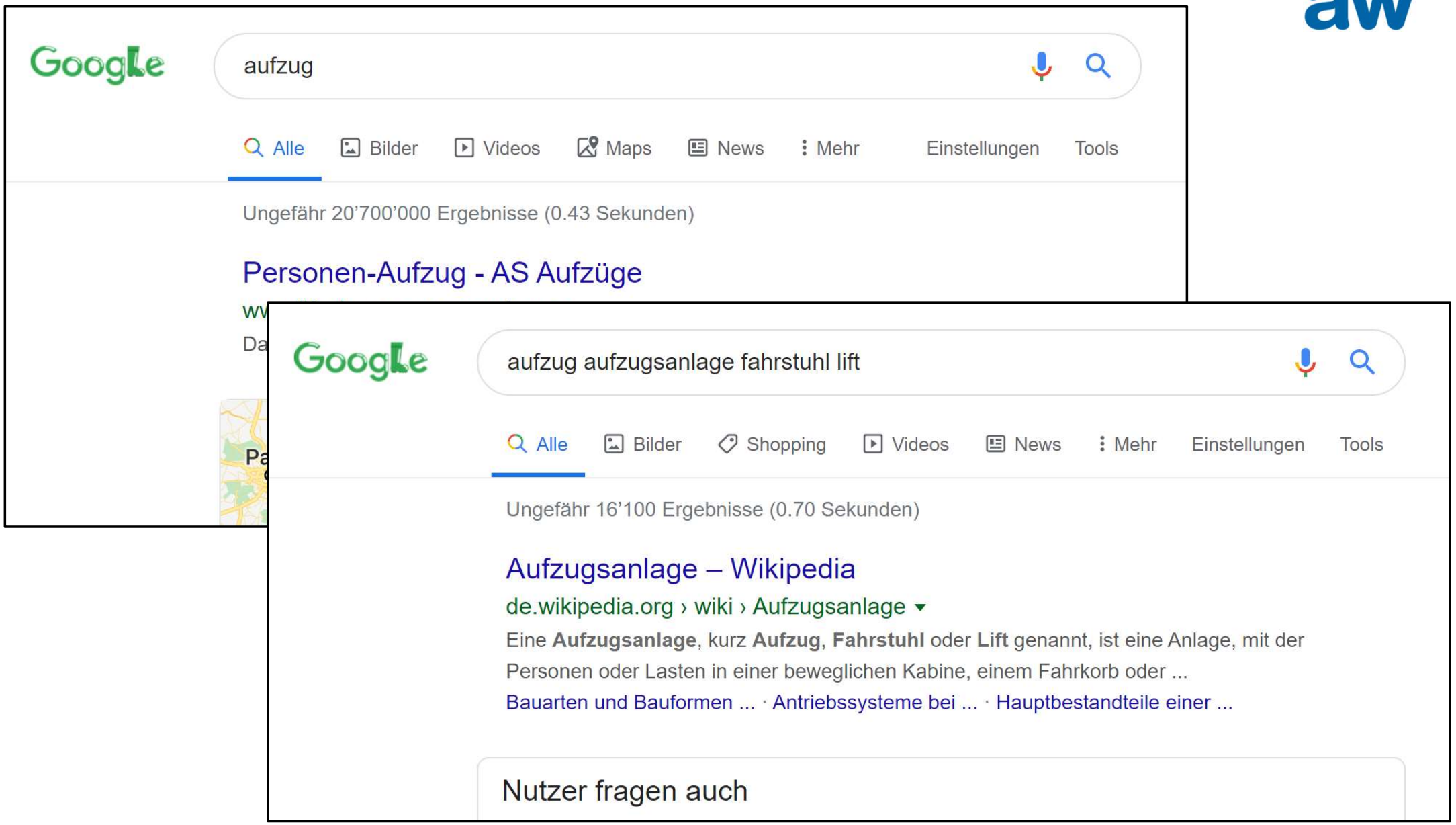
Die „Konsequenzen“ des Retrievalproblems

- Der “unscharfe” Begriff der Relevanz und die vielfältigen Darstellungsformen der Information, welche eine Übereinstimmung von Anfrage und Dokument erschweren, führen zu einer wahrscheinlichkeitsbasierten Lösung:
  - Es werden diejenigen Dokumente geliefert, für welche die Wahrscheinlichkeit, dass sie vom Benutzer als relevant zur Anfrage beurteilt werden, am höchsten sind.
  - Ein Retrievalresultat ist fast nie vollständig “korrekt”: es fehlen relevante Dokumente, oder irrelevante Dokumente werden zusätzlich gefunden.
  - Merke: “scharfe Kriterien” (ja/nein) sind ungeeignet, da der Benutzer die Anzahl und die Form der gesuchten Dokumente kennen müsste, um eine Anfrage zu formulieren, welche das gewünschte Resultat liefert → Paradox
  - Gute Retrieval-Systeme erlauben dem Benutzer alles zu formulieren, was er weiss, ohne die Gefahr, zuviel oder zuwenig zu finden.

Frage:

- Kennt jemand ein Gegenbeispiel?

Gegenbeispiel



Provokative Frage:

- Inwiefern verletzt Google unsere Forderungen?
- Heisst das, Google ist ein *schlechtes* Suchsystem?
- Falls ja, warum wird Google so gelobt?

Das Suchparadox

- Google kann sich "Vereinfachungen" erlauben dank dem Suchparadox
- Es ist einfacher, in mehreren Milliarden Dokumenten zu suchen als in mehreren Tausend.
  - In grossen Datenmengen ist tendenziell die Redundanz massiv höher
  - Information kann also mit "beliebigen" Verbalisierungen gefunden werden
  - Benutzer von Google sind häufig präzisionsorientiert → wenige gute Treffer reichen
- Wird auf "kleinen" Datenmengen gesucht, so ist eine Übereinstimmung zwischen Informationsbedürfnis und Dokumenten schwerer nachzuweisen

Kontrollfragen

- Was ist Information Retrieval?
- Was ist das Retrievalproblem?
- Was ist die Bedeutung des IR-Kreislaufs?
- Ist Google ein schlechtes IR-System? Was war der Knackpunkt?
- Warum sollten wir bei (Web-)suche NICHT "Die Nadel im Heuhaufen" bemühen?

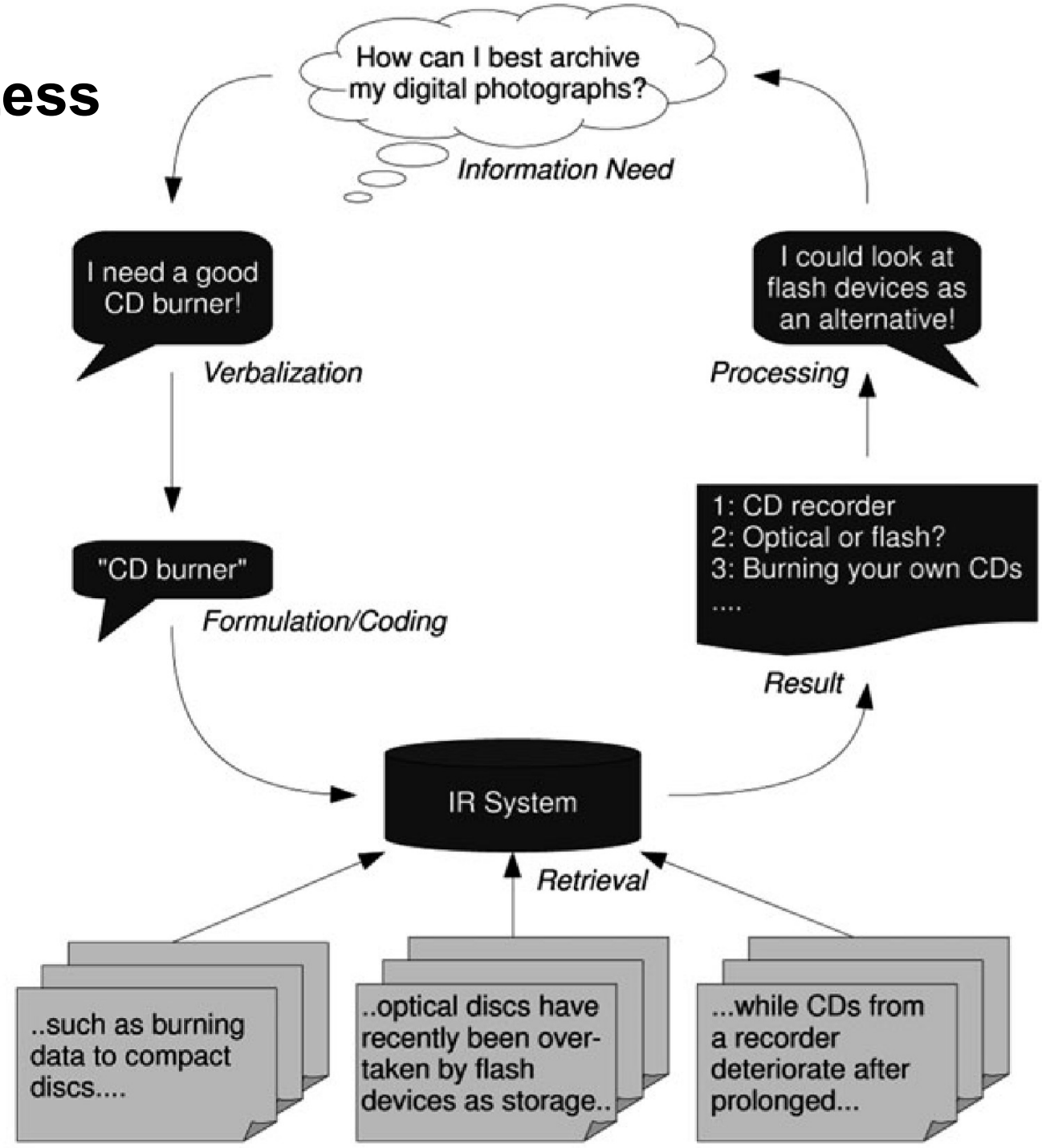
Vom Informationsbedürfnis zur Information

- Vergleiche das Schema (→Folie 24) auch mit der klassischen Bibliotheksuche:
- Der Bibliothekar hilft dem Benutzer das Informationsbedürfnis zu verbalisieren und in eine Form zu bringen, um nach der gewünschten Information suchen zu können.
- Ein Bibliothekar hilft gegebenenfalls auch, das Informationsbedürfnis besser zu verstehen. Information Retrieval-Systeme können ähnliche Funktionen übernehmen (Suchassistent).
- Tatsächlich war die Suche in Bibliotheken ursprünglich die zentrale Anwendung von Information Retrieval.

Iterative Suche – Weiterentwicklung eines Informationsbedürfnisses

- Das Verständnis des Benutzers für sein Informationsbedürfnis ändert sich mit der Information, welche er sammelt. Die Suche ist oft ein iterativer Prozess (kann von Information Retrieval-Systemen aktiv unterstützt resp. gefördert werden):
  - Werden irrelevante, oder zu viele resp. zu wenige relevante Dokumente gefunden, formuliert der Benutzer die Anfrage um (ändert die Verbalisierung resp. Codierung)
  - Der Benutzer kann vom System bei der Umformulierung unterstützt werden (automatische Erweiterung der Anfrage, Suche nach ähnlichen Dokumenten) → was ist der spezielle Vorteil, den das System dabei gegenüber dem Nutzer hat?
  - Werden relevante Dokumente gefunden, so ändert sich das Verständnis des Informationsbedürfnisses.
  - Der Benutzer kann vom System beim Verständnis unterstützt werden (z. B. automatische Kontextanalyse)
- Beispiel: Erkenntnis, dass andere Speichermedien eine Alternative sein könnten

IR-Prozess



Aus Peters et al., 2012

Beispiele für Informationsbedürfnisse

- Beispiele:
  - Übersicht über Architektur in Berlin
  - Alles über die „Elektroschwachtheorie“
  - Liste von Überschwemmungen in Europa
  - Wie hoch ist der Eiffelturm?
  - Wo kann ich eine Pizza bestellen?
  - Biographie von Mozart
  - Was ist die Adresse der Homepage der Coca Cola Corporation?
  - Wo befinden sich unerschlossene Ölvorkommen?
  - Wo überall wird ein bestimmtes Dokument referenziert?
  - Ist ein Patent gültig (prior art?)
- → Informationelle, navigationale, transaktionale, örtlich gebundene Bedürfnisse.
- “Klassisches IR” konzentriert sich auf informationelle Bedürfnisse (“Bibliothekszenario”).

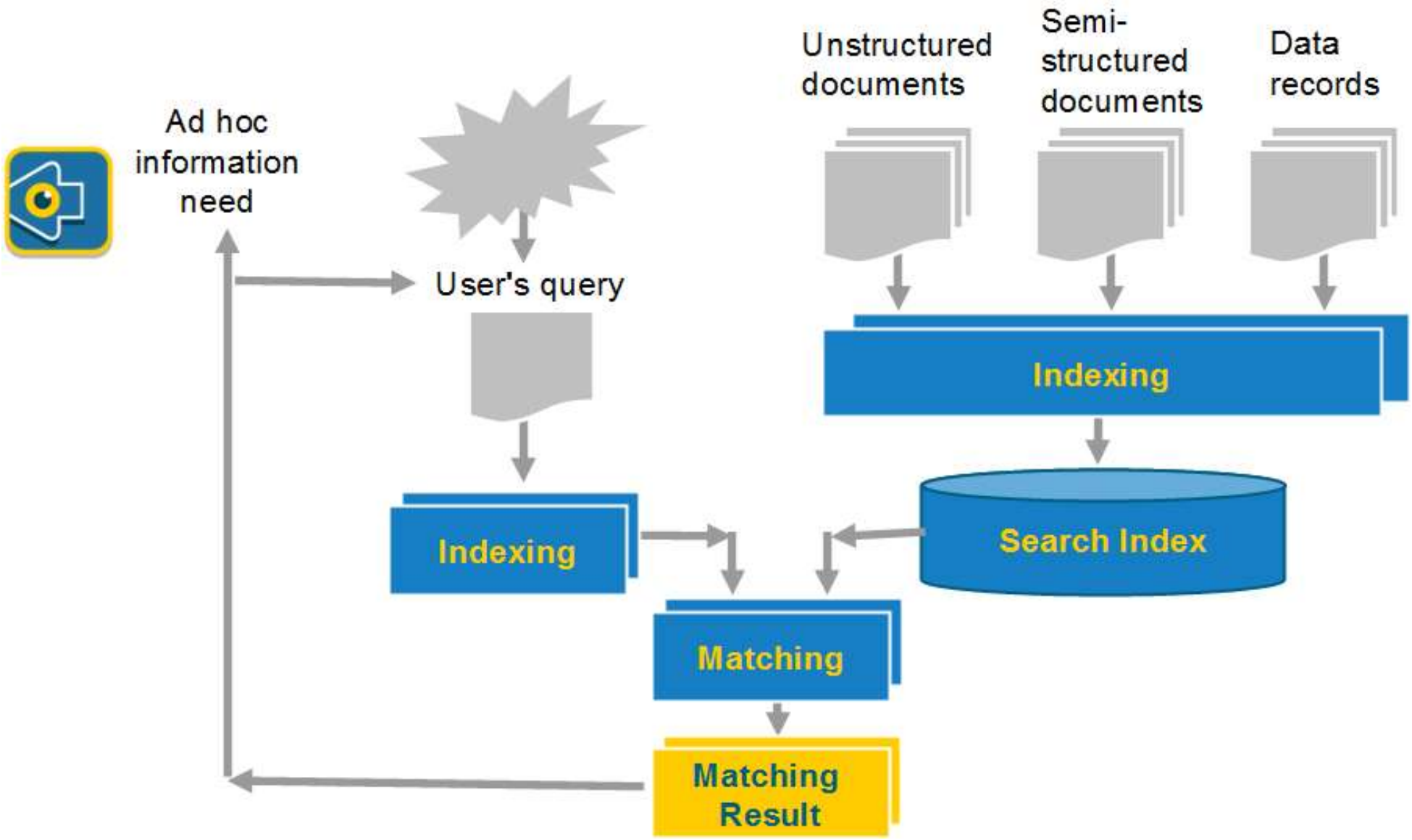
Information Retrieval Paradigmen

- Aufgrund eines Ad hoc-Informationsbedürfnises Dokumente mit relevanten Informationen suchen (pull).

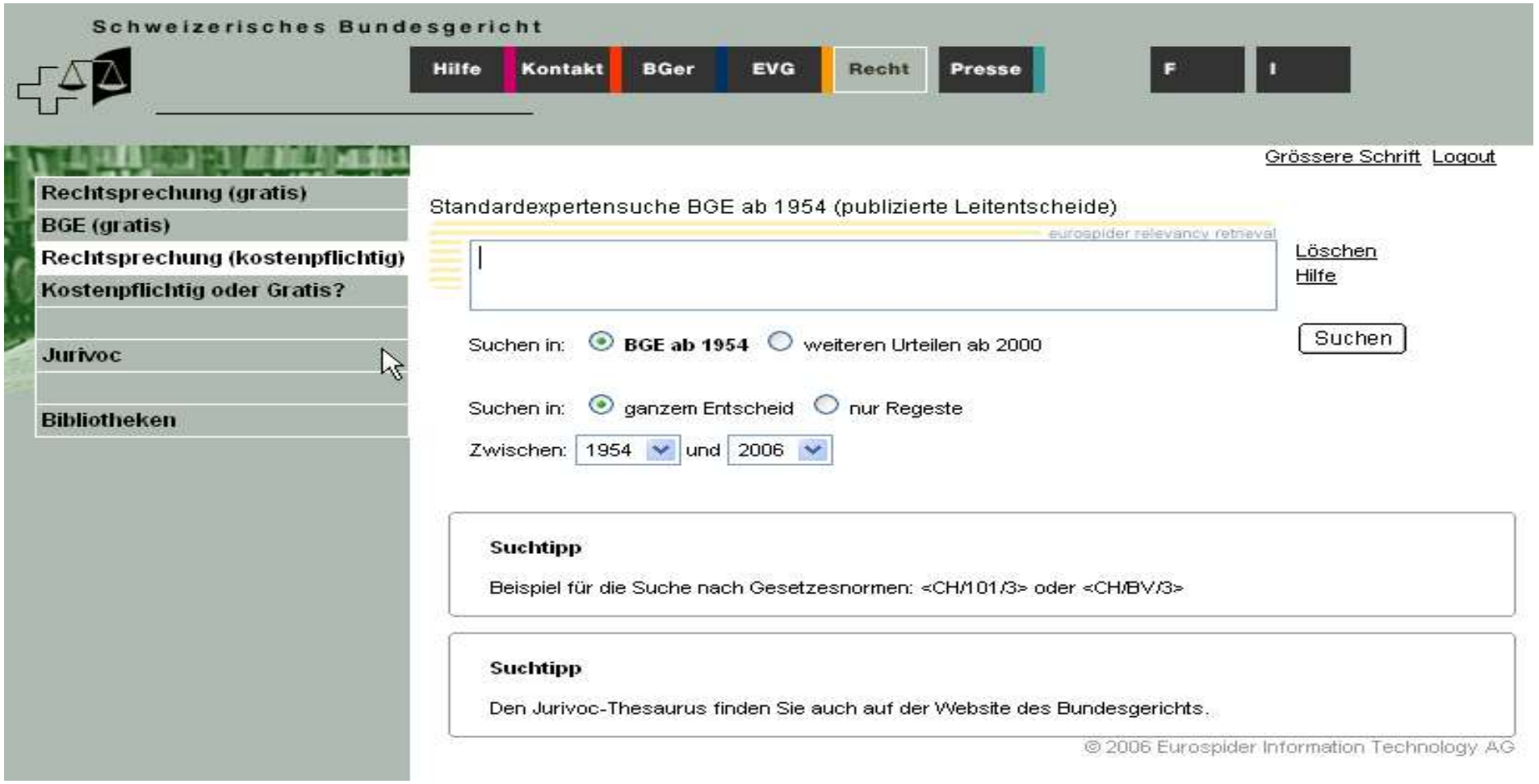
- Dokumente mit relevanten Informationen aus einem Dokumentenstrom herausfiltern und weiterleiten (push).

- Neue Dokumente kategorisieren und in eine Informationsstruktur einordnen (browse).

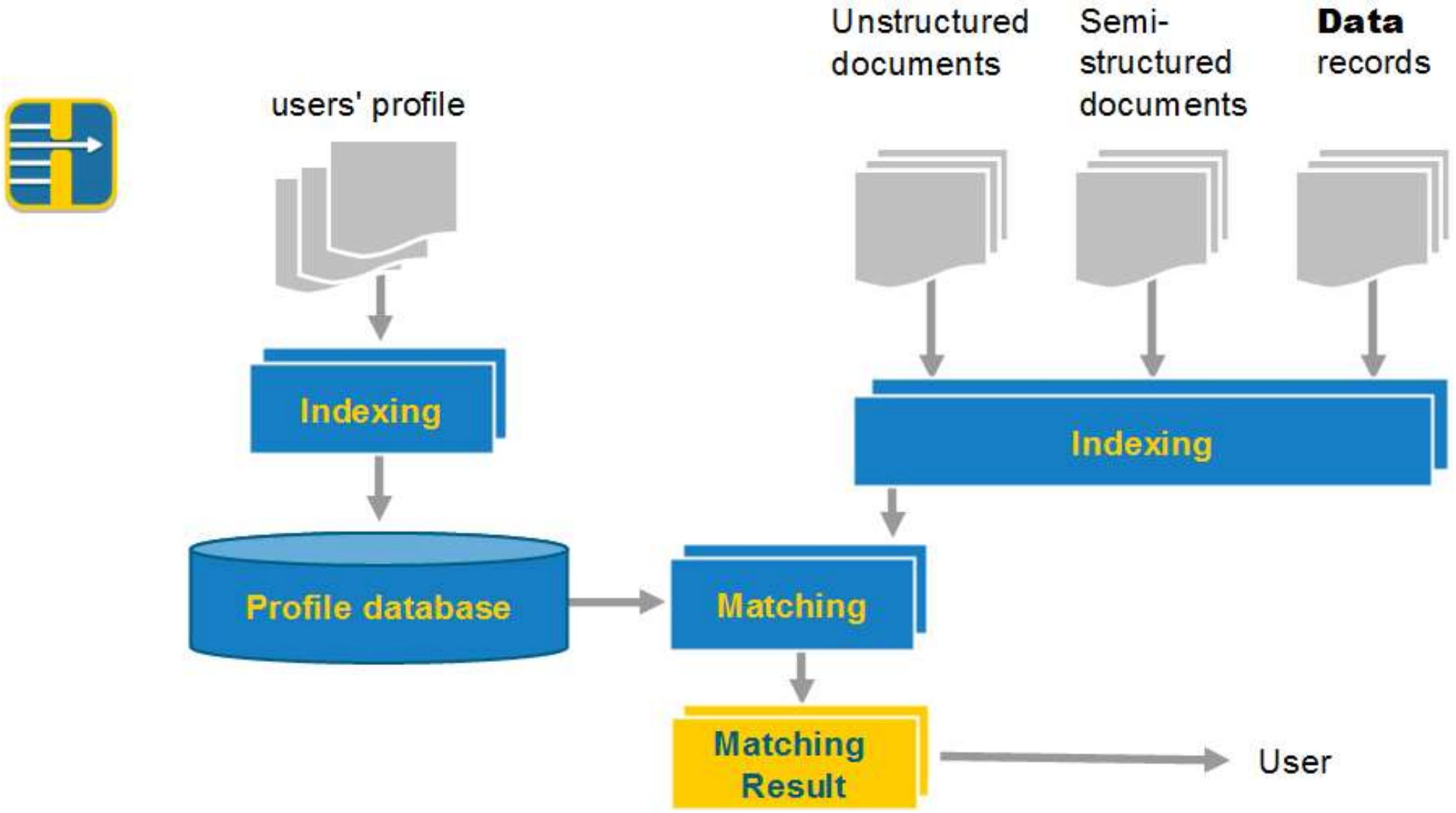
Ad-hoc Suche (PULL)



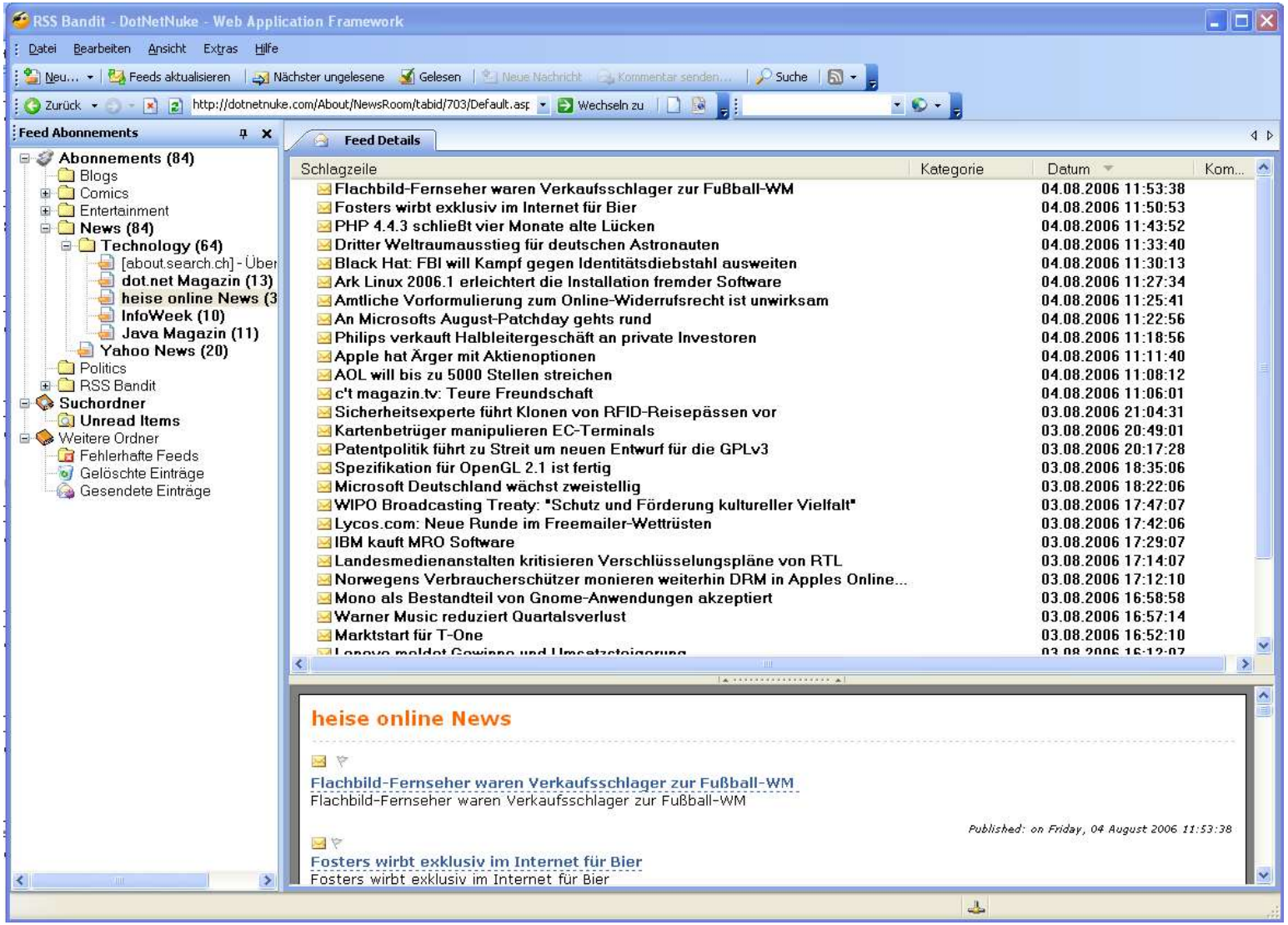
Ad-hoc Suche (PULL) Beispiel



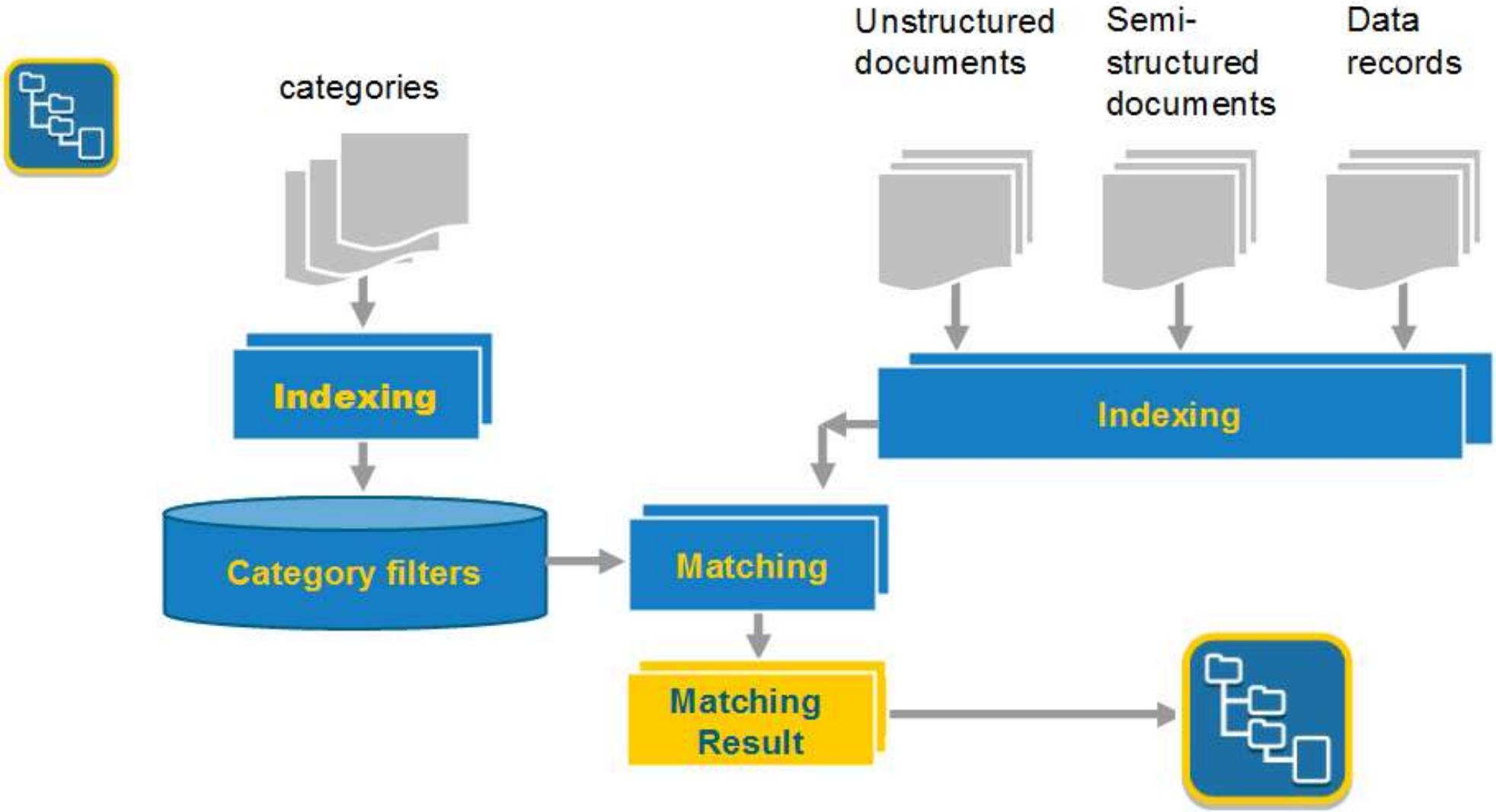
Bringdienste (PUSH)



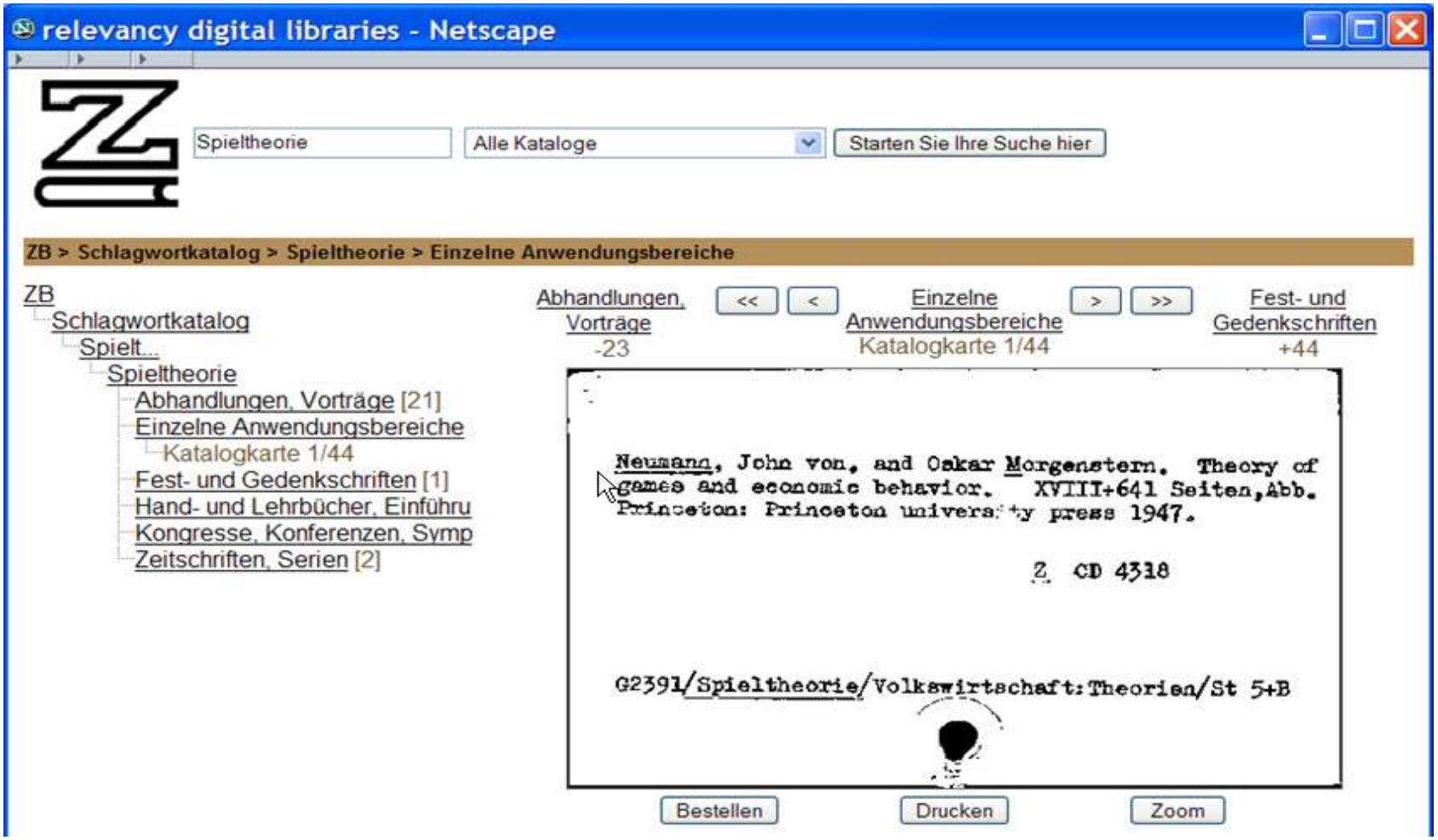
Bringdienste (PUSH) Beispiel



Dokumente in Themenhierarchie einordnen

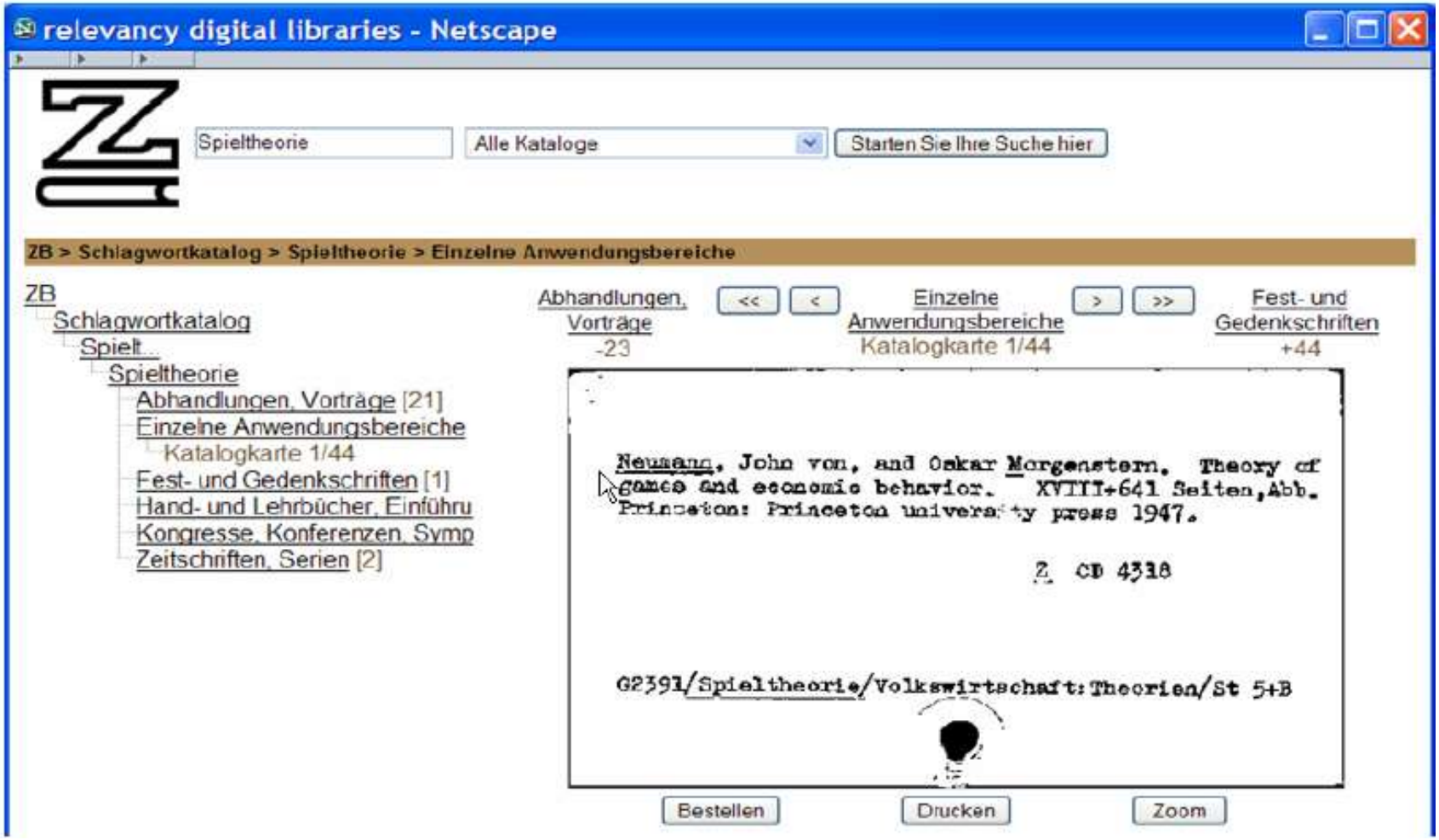


Dokumente in Themenhierarchie einordnen



Dokumente in Themenhierarchie einordnen

Dokumente in Themenhierarchie einordnen



Frage:

- Was ist der Unterschied zwischen Information Retrieval und einer Datenbanksuche?

IR im Vergleich zu Datenbanksuche

- Datenbanken liefern für strukturierte Information mit kontrolliertem Vokabular perfekte Resultate.
- Daten in Datenbanken sollten unabhängig sein von der Applikation (erlaubt zukünftige Nutzung). Redundanz wird vermieden.
- Elemente sind entweder Teil der Resultatmenge oder nicht (binäre Unterscheidung).
- Boole'sche Kriterien zur Selektion
- Geeignet für die Suche in (hoch-)struktierter Information mit kontrolliertem Vokabular.

IR im Vergleich zu Datenbanksuche

	Data Retrieval	Information Retrieval
Matching	Exact match	Partial match, best match
Inference	Deduction	Induction
Model	Deterministic	Probabilistic
Classification	Monothetic	Polythetic
Query language	Artificial	Natural
Query specification	Complete	Incomplete
Items wanted	Matching	Relevant
Error response	Sensitive	Insensitive

Quelle: C. J. van RIJSBERGEN: INFORMATION RETRIEVAL (http://www.dcs.gla.ac.uk/Keith/Chapter.1/Ch.1.html)

Frage:

- Was ist der Unterschied zwischen Information Retrieval und einer Suche in MS-Word (Textsuche)?

IR im Vergleich zu „Textsuche“

- Suchfunktionen in Editoren/Textverarbeitungen basieren auf Mustervergleichen.
- Skaliert nicht - Linearer Scan über Dokument(e)
- Binär: erfüllt Suchkriterium oder nicht
- Ermöglicht nur einfache Anfragen
- Geeignet für die Suche innerhalb eines Dokuments (vor allem zur Bearbeitung), nicht für die Suche nach Dokumenten.

Frage:

- Wann ist ein Suchresultat gut?
- Wann ist ein DB-Resultat gut?

Masse für Retrievaleffektivität

- Retrievalresultate sind fast nie perfekt. Es sind Masse nötig, um die Qualität eines Retrievalresultats zu beurteilen.
- Zwei allgemein anerkannte Masse für Retrievalqualität sind **Ausbeute** und **Präzision**
- Diese Eigenschaften dieser Masse sind gut analysiert und verstanden
- Beide Masse modellieren die Annahme, dass möglichst viel relevante, und möglichst wenig irrelevante Information gefunden werden soll.

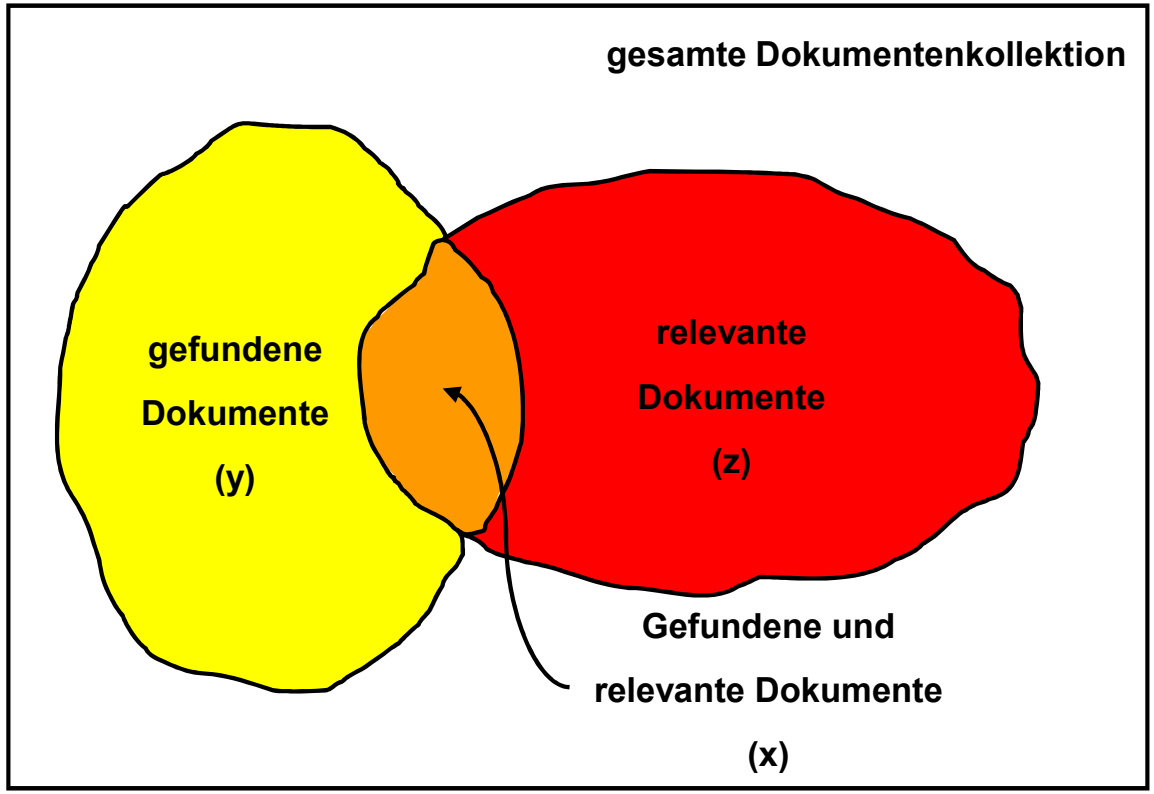
Masse für Retrievaleffektivität

$$\text{Präzision} := \frac{\#\text{relevante Dokumente im Resultat}}{\#\text{Dokumente im Resultat}}$$

$$\text{Ausbeute} := \frac{\#\text{ relevante Dokumente im Resultat}}{\#\text{relevante Dokumente in der Kollektion}}$$

- Ziel: Optimierung eines oder beider Kriterien!
- Was fällt Ihnen auf?

Ausbeute und Präzision

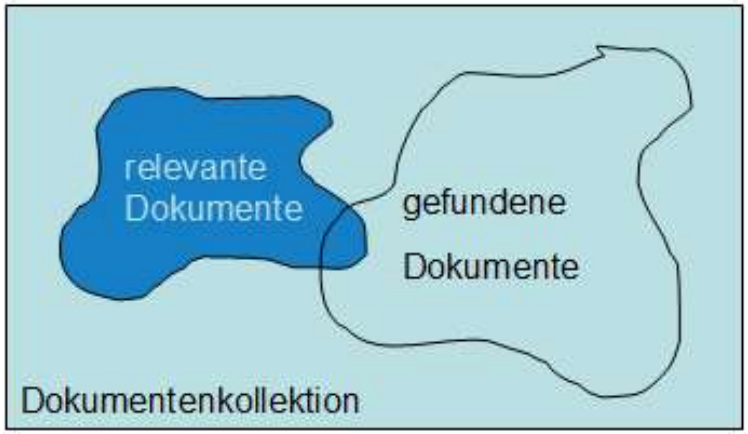


Präzision = x / y

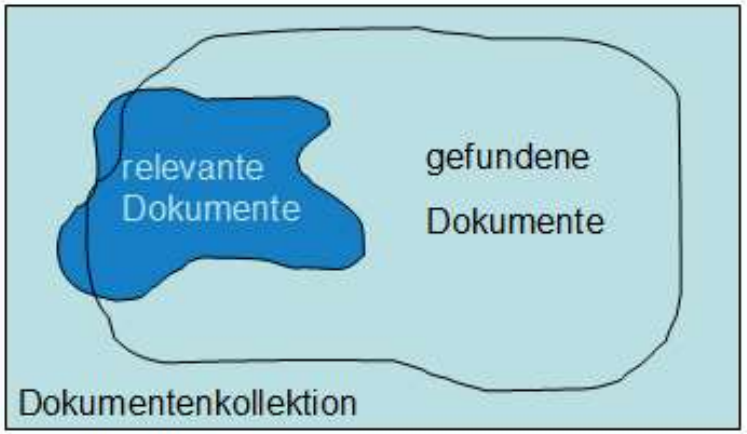
Ausbeute = x / z

Quelle: http://www.ir.iit.edu/~nazli/CS495-Slides/01Introduction_05.pdf

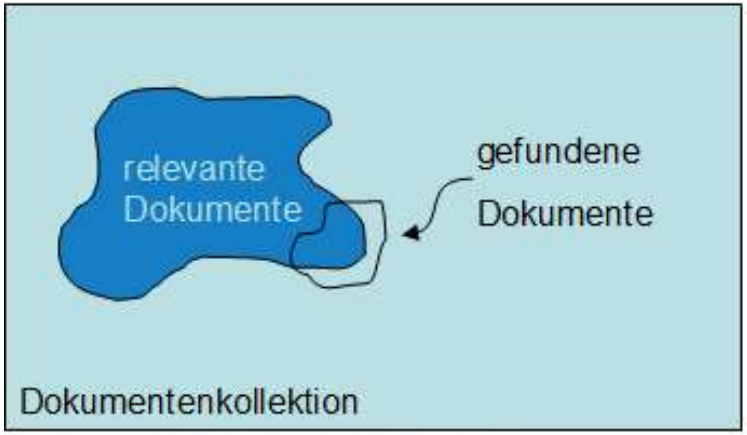
Ausbeute versus Präzision



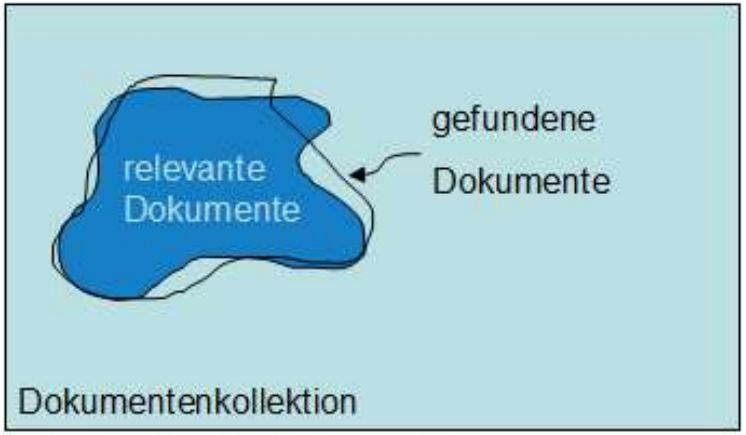
Geringe Ausbeute, geringe Präzision



Hohe Ausbeute, geringe Präzision



Geringe Ausbeute, hohe Präzision

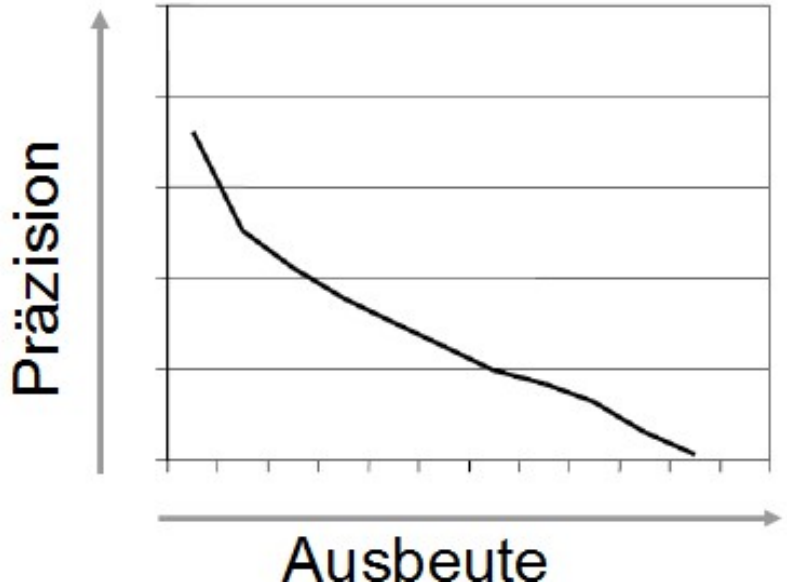


Hohe Ausbeute, hohe Präzision

Quelle: http://www.swisseduc.ch/informatik/soekia/docs/2005-03-issep-suchmaschinen.pdf

Masse für Retrievaleffektivität

- Die beiden Masse sind mengenbasiert.
- Die beiden Masse widersprechen sich oft:
  - hohe Ausbeute → niedrige Präzision
  - hohe Präzision → niedrige Ausbeute



„Probability Ranking Principle“

- Eine Anordnung der „gefundenen“ Dokumente in der Reihenfolge ihrer Wahrscheinlichkeit, dass sie zur Anfrage relevant sind, unter Berücksichtigung sämtlicher zur Verfügung stehender Informationen und geeigneter Annahmen, ist optimal in mehrerer Hinsicht.
- Falls das so genannte „Probability Ranking Principle“ befolgt wird, lassen sich die folgenden Eigenschaften mathematisch beweisen:
  - Die Präzision an einem beliebigen “cut-off point” wird optimiert.
  - Die Ausbeute an einem beliebigen “cut-off point” wird optimiert.
  - Die Kosten der Auswertung (relevant = positiv, irrelevant = negativ) werden optimiert.
  - und andere…
- Folgerung: die Verwendung von wahrscheinlichkeitsbasierten Ranglisten ist theoretisch fundiert.
- → Dieses Konzept wird uns im weiteren Verlauf noch begleiten...

Geschichte

- 1945: Bush, „As We May Think“: Aufforderung, Information zugreifbar zu machen
- 1952: Moers: Der Begriff „Information Retrieval“
- 1958: „International Conference on Scientific Information“ in Washington: Der Beginn des akademischen Felds „Information Retrieval“
- 1958: Joyce & Needham: Thesauri ermöglichen bessere Übereinstimmung zwischen Anfrage und Dokument
- 1960: Maron & Kuhns: Begriffe wie „Dokumentenrelevanz“, „wahrscheinlichkeitsbasiertes Retrieval“, „sortierte Rangliste“ tauchen auf
- 1961: Luhn: vollautomatische Indizierung mittels Begriffen aus dem Volltext der Dokumente
- 1962 & 1967: Cleverdon: Cranfield-Experimente. Die erste „moderne“ Evaluation von Retrieval; automatische Ansätze zeigen viel versprechende Resultate

Geschichte

- 1960er: Arbeiten von Salton, SMART System, Vector Space Model
- 1965: Rocchio: Relevance Feedback
- 1977: Probability Ranking Principle
- 1973: erste SIGIR Konferenz, ab 1978 regelmässig
- 1992: Harman, TREC-1 Conference: Erste grosse Evaluationskampagne
- 1990er: moderne sprachübergreifende Suche
- 1990er: Multimedia Retrieval
- Ca. 1996: Web Retrieval
- 2000: CLEF: Europäische Evaluationskampagne
- Im Allgemeinen von kürzerer Lebensdauer: Arbeiten hinsichtlich Effizienz
- Aktuell: viele fachspezifische Probleme, Probleme, welche die Grenzen zw. NLP/IR und ML/IR verwischen: Filtern komplexer Sachverhalte (z.B. gesundheitlicher Risiken), Erkennung von Multimediainhalten, …

Recherchierkompetenz Lernkontrolle

- Wir sollten nun in der Lage sein, die folgenden Probleme besser zu verstehen:
  - Warum finde ich zu meiner Anfrage 1000 Resultate? Wer soll die alle lesen?
  - Warum finde ich die falschen Resultate? Unvollständige Resultate?
  - Warum finde ich mein persönliches Dokument nicht?