

# Data Mining im Internet

**Dipl.-Dok. Helga Walter**

**Bayer HealthCare, Wuppertal**  
**PH-R-EU Scientific Information and Documentation**

# Arten / Quellen wissenschaftlicher Information

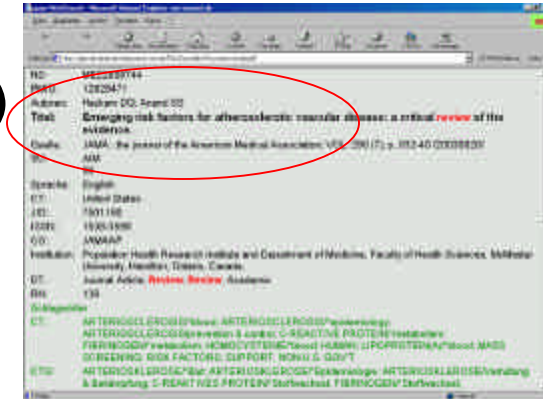
## Strukturierte Informationen:

- z.B. Bibliographische Datenbanken (Medline, ...)
- Dokumente sind logisch aufgebaut

↪ Felder: Autor/Titel/Quelle/Abstract/...

- Jedes Dokument recherchierbar (Suchfunktion)

↪ Thesaurus



## Nicht-strukturierte Informationen:

- z.B. Text- / Multimediadateien, Webdokumente
- Formate: Word, PDF, HTML, ...
- Keine Suche nach logischen Merkmalen
- Volltextsuche



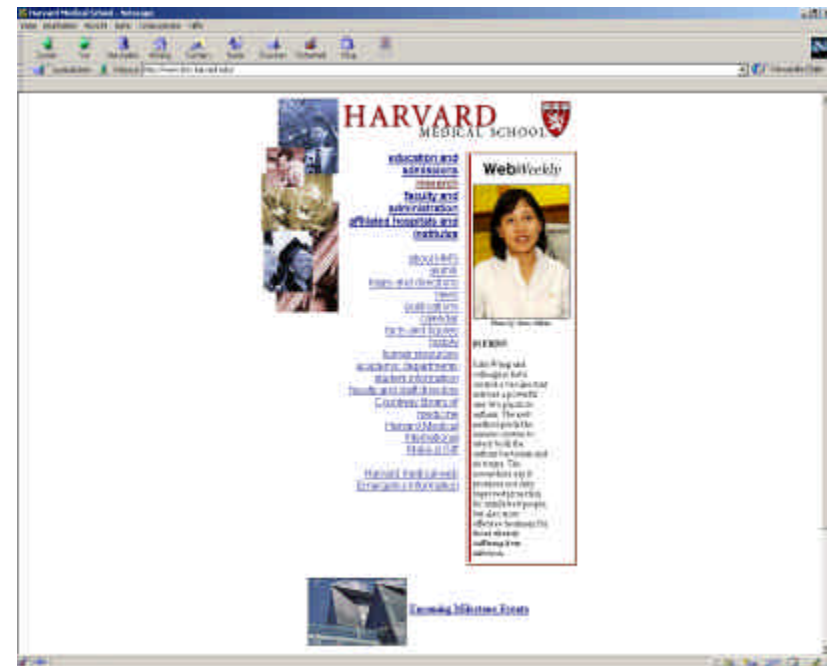
**Problem: Auffinden relevanter Informationen aus nicht-strukturierten Quellen**

# Data Mining im Internet

## Bedeutung für die Informationsbeschaffung

### Data Mining im Internet als nicht-strukturierte Informationsquelle

- Frühe Hinweise auf Forschungsergebnisse
- Noch nicht publizierte Ideen
- Expertenforen / Meinungsbildner
- Aktuelle Übersichten / Vorträge



**Internet als Informationsquelle immer bedeutender**

# Data Mining im Internet

## Recherche-Hilfsmittel

**Internet-Suchmaschinen:**

**Allgemein: Google, AltaVista, Metasuchmaschinen**

**Spezialisiert: Northern Light, ChemGuide, MedPharmGuide**

**Prinzip:**

- **Eingabe von Suchbegriffen**
- **Suche in indizierten Seiten**

**Nachteile:**

- **Manuelle Suche**
- **Bei komplexen Suchanfragen**
- **Bei Speicher- und Editierfunktion für Suchstrategie**
- **Fehlende SDI-Funktion**
- **Große Treffermengen - überwiegend irrelevant**



**Funktionalität herkömmlicher Internet-Suchmaschinen  
begrenzt**

# Data Mining im Internet

## InsumaScout - Intelligente Suchmaschine



**Prinzip: Gewichtete Suche über lernfähigen Crawler**

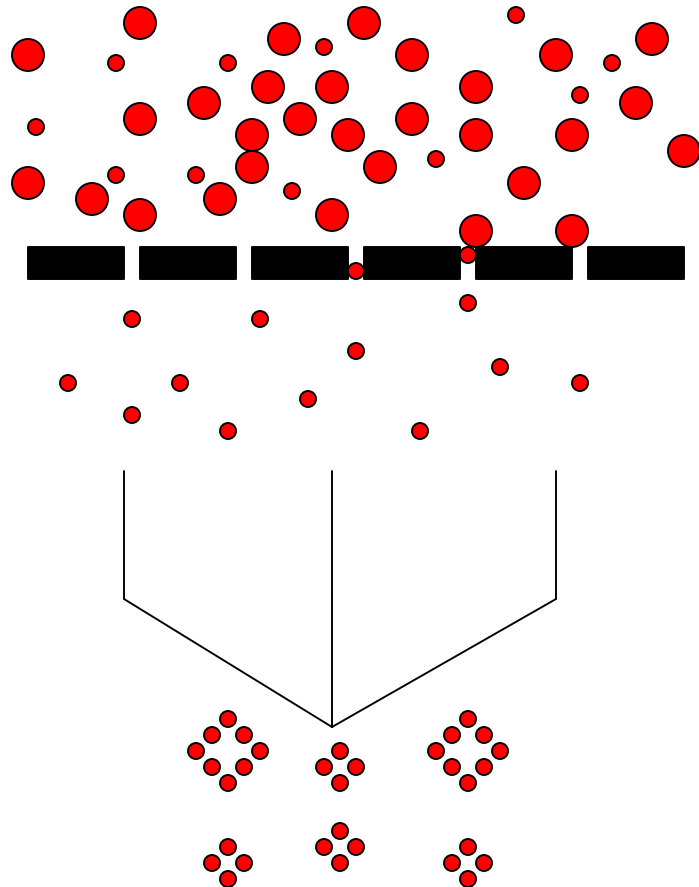
**Vorteile:**

- **Automatisierte Suche**
- **Für komplexe Suchen geeignet**
- **Lernfähigkeit**
- **Signifikante Reduktion irrelevanter Links**
- **Vorsortierung relevanter Informationen**
- **Suchbare Kollektionen**

**Individualisierte Suchprozeduren, verbesserte Selektion  
relevanter Ergebnisse**

# InsumaScout

## Arbeitsweise des intelligenten Crawlers



Internet

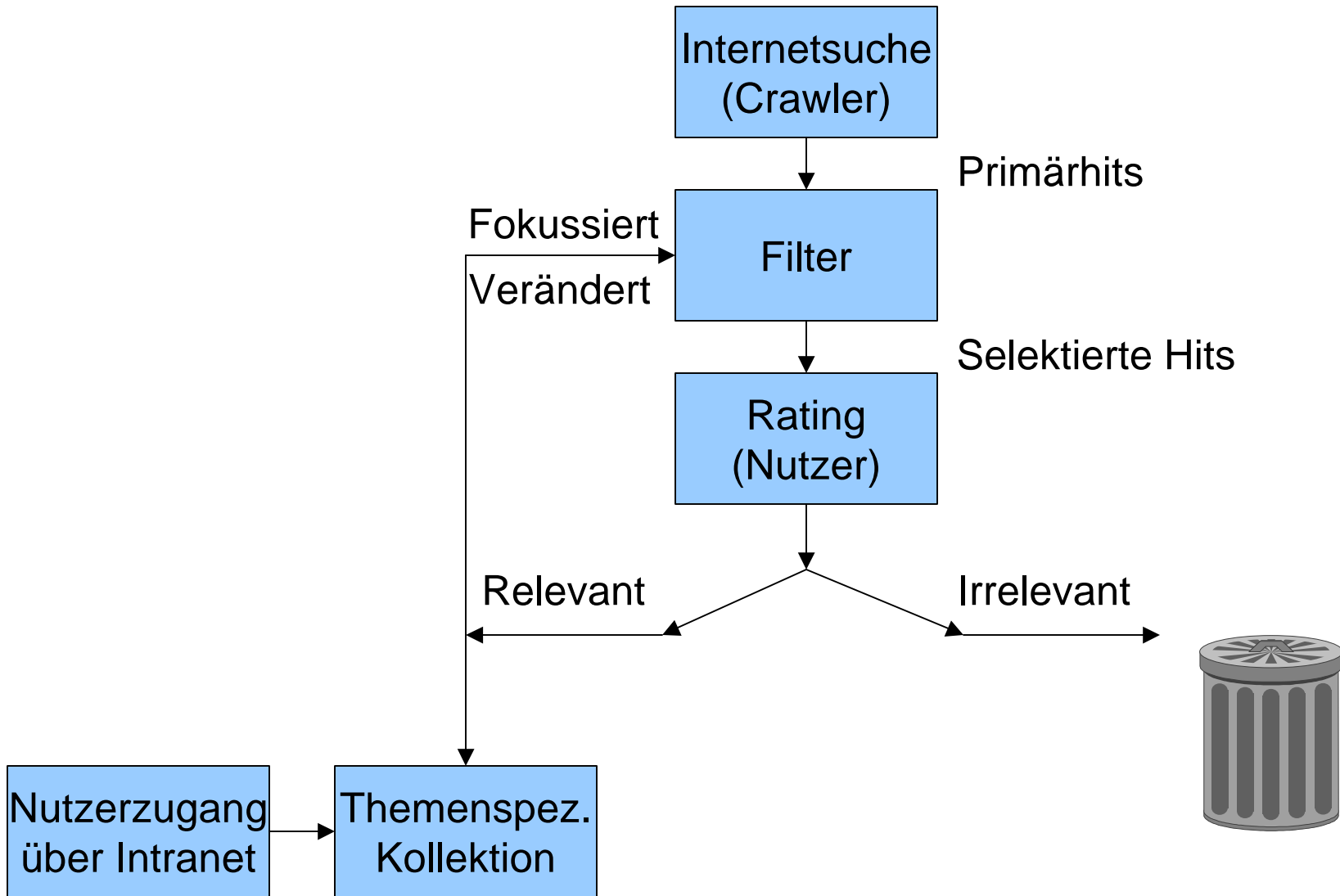
Themenspezifischer, lernfähiger  
Filter / Crawler

Ausgewählte Hits

Vorsortierung

Relevante Internetseiten durch intelligenten, lernfähigen  
Crawler

# Arbeitszyklen



# Aufbauphase

## Aufbau des Filters und Start des Crawlers

**FILTER = Schlagwortliste mit dazugehöriger Gewichtung**

**Ausgangsinformation zur Generierung eines Filters:**

- Liste relevanter URLs (Startadressen)
- Schlagwortliste
- Textblöcke aus Präsentationen, Publikationen, etc.
- Textblöcke aus Internetseiten



**Art und Menge der Startinformation beeinflusst Ausgangs-Qualität des Filters**

# Aufbau einer themenspezifischen Kollektion

- Crawler durchsucht das Internet im ersten Durchlauf
- Auffinden themenspezifischer Dokumente (Selektionsprozeß)
- Eliminieren von Duplikaten
- Aufbau einer Kollektion (Trefferliste)
- Sortieren der Treffer
  - ↳ in Ähnlichkeits-Clustern (nach Relevanz) bzw. nach URLs

**Suchergebnis (Kollektion) wird dem Nutzer im sog. Control Center angezeigt**

# Routinephase

## Kontinuierliches Lernen des Filters

**Verändern des Filters durch:**

- Beurteilen der Treffer (Rating)
- Hinzufügen / Entfernen von URLs, Schlagwörtern, Textblöcken

**Je mehr Dokumente als relevant beurteilt werden, desto höher die Filterqualität und die Qualität der Treffer im nächsten Durchlauf**

<b>5 Relevanzstufen:</b>	<b>+2</b>	<b>Sehr relevant</b>
	<b>+1</b>	<b>Relevant</b>
	<b>0</b>	<b>Keine Meinung</b>
	<b>-1</b>	<b>Nicht sehr relevant</b>
	<b>-2</b>	<b>Irrelevant</b>

**Bewerten der Dokumente durch Nutzer,  
entsprechendes Anpassen des Filters**

# InsumaScout Recherchethemen

## Pilotprojekt:

- ◆ Alzheimer'sche Erkrankung

 Terminologie eindeutig

- ◆ Kardiovaskuläre Erkrankungen

 Terminologie nicht immer eindeutig

- ◆ Naturstoffe

 Terminologie nicht eindeutig

**Hohe Bandbreite an Fragestellungen möglich**

# InsumaScout

## Qualität der Ergebnisse

**Alzheimer'sche Erkrankung / Naturstoffe:  
2600 Dokumente / Thema (100 Dokumente / Woche)**

**Auswertung:**

<i>Status Dokumente</i>	<i>Alzheimer</i>	<i>Naturstoffe</i>
Relevant (Relevanz +1 / +2)	19%	28%

**Herkömmliche Suchmaschinen:  $\leq$  5%**

**Suchergebnisse enthalten hohe Relevanz und  
Qualität**

# InsumaScout

## Schwierigkeiten / Grenzen

- ◆ **Relevante, jedoch bereits bekannte Informationen**
- ◆ **Links mit wissenschaftlich niedrigem Anspruch**
- ◆ **Publikationen aus Zeitschriften**
- ◆ **Patientenforen**
- ◆ **Veraltete Informationen**

**Ersetzt nicht die gesamte intellektuelle Arbeit des Nutzers**

# InsumaScout

## Vorteile der intelligenten Suchmaschine

- ◆ **Lernfähiger Crawler**
- ◆ **Hohe Aktualität durch erhöhte Suchfrequenz**
- ◆ **Relevante Informationen durch aufwendige Selektion**
- ◆ **Weitere Aufbereitung für Endnutzer**
- ◆ **Bildung suchbarer Hit-Kollektionen**
- ◆ **Automatische Dubletteneliminierung**
- ◆ **Einfacher / schneller durchführbar als manuelle Suche**

**InsumaScout ist konventionellen Suchmaschinen  
deutlich überlegen**