

# Data Mining im Internet

Helga Walter, Wuppertal

Data Mining im Internet nimmt für die Informationsbeschaffung eine immer größere Bedeutung ein. Als Recherche-Hilfsmittel für das Data Mining im Internet stehen Internet-Suchmaschinen zur Verfügung. Die Funktionalität konventioneller Suchmaschinen ist begrenzt. Intelligente Suchmaschinen sind klassischen / themenspezifischen Suchmaschinen deutlich überlegen. Ein Test der intelligenten Suchmaschine InsumaScout zeigt folgende Vorteile: lernfähiger Crawler (Internetsuche und Selektionsprozess werden durch Bewertung des Benutzers kontinuierlich verfeinert), hohe Aktualität durch erhöhte Suchfrequenz, relevante Informationen durch aufwendige Selektion, Bildung suchbarer Hit-Kollektionen, automatische Dubletteneliminierung, einfacher und schneller durchführbar als manuelle Suche.

*Data Mining in the Internet is of increasing importance for information retrieval. Internet search engines are used as search tools for Data Mining. Conventional search engines offer only a limited functionality for information retrieval. Intelligent search engines are advantageous to classical / specific Internet search engines. A test of the intelligent search engine InsumaScout reveals the following advanced functionality: adaptive crawler (Internet search and selection process are continuously refined by user-rating), high timeliness by increased search frequency, providing relevant information by sophisticated selection procedures, creation of searchable hit collections, automatic elimination of duplicates, easier and faster to perform than manual retrieval.*

## Strukturierte und nicht-strukturierte Informationen

Wissenschaftliche Informationen lassen sich in strukturierte und nicht-strukturierte Informationen aufteilen. Zu den strukturierten Informationen zählen z.B. bibliographische Datenbanken (Medline, Embase, etc.). Charakteristisch für bibliographische Datenbanken ist, dass die Dokumente einem logischen Aufbau folgen. Die einzelnen Dokumente sind in bestimmte Felder, wie z.B. AUTOR / TITEL / QUELLE / ABSTRACT, aufgeteilt. Mittels einer bestimmten Suchfunktion ist jedes Dokument recherchierbar. Als Recherchehilfsmittel kann ein Thesaurus (z.B. MeSH, Emtree) eingesetzt werden.

Nicht-strukturierte Informationen sind z.B. Textdateien, Multimediadateien und Internetdokumente. Diese liegen in den Formaten WORD, PDF, HTML, etc. vor. Bei der Recherche handelt es sich um eine Volltextsuche. Problematisch erweist sich das Auffinden relevanter Informationen aus nicht-strukturierten Quellen.

Für die Informationsbeschaffung gewinnen neben den strukturierten Informationsquellen auch die nicht-strukturierten Informationsquellen immer mehr an Bedeutung.

## Data Mining im Internet - Bedeutung für die Informationsbeschaffung

Bibliographische Datenbanken sind für die Suche nach wissenschaftlicher Information nach wie vor unabdingbar. Dennoch nimmt das Internet als Informationsquelle einen immer größeren Stellenwert ein. Dies gilt vor allem bei der Recherche nach Informationen, die nicht über die herkömmlichen Informationsquellen zu finden sind. Die Suche nach der „Stecknadel im Heuhaufen“ kann zu ei-

nem bedeutenden Wissensvorsprung führen. Data Mining erschließt das Internet als nicht-strukturierte Informationsquelle. Im Internet kann so nach frühen Hinweisen auf Forschungsergebnisse (neue Ansätze, Methoden), noch nicht publizierten Ideen, Expertenforen, Meinungsbildnern, aktuellen Übersichten und Vorträgen zu bestimmten Themen, etc. recherchiert werden.

## Konventionelle Internet-Suchmaschinen

Als Recherche-Hilfsmittel für das Data Mining im Internet stehen eine Reihe von Internet-Suchmaschinen zur Verfügung. Zu den allgemeinen Suchmaschinen zählen u.a. Google, AltaVista und Metasuchmaschinen. Northern Light, ChemGuide (FIZ Chemie) und MedPharmGuide sind spezialisierte, themenspezifische Suchmaschinen. ChemGuide konzentriert sich beispielsweise auf chemiebezogene Internetseiten.

Die konventionellen Suchmaschinen arbeiten nach folgendem Prinzip. Zunächst erfolgt eine Eingabe von einem oder mehreren Suchbegriffen. Diese werden in den indizierten Seiten gesucht. Das Suchergebnis wird als Trefferliste angezeigt. Die Funktionalität herkömmlicher Internet-Suchmaschinen ist begrenzt. Bei komplexen Suchanfragen ist die Grenze der Suchmaschine schnell erreicht. Viele Internet-Suchmaschinen bieten weder eine Speicher- und Editierfunktion für die Suchstrategie noch eine „Selective Dissemination of Information“ (SDI)-Funktion. Als Ergebnis werden große Treffermengen ausgegeben, die überwiegend irrelevant sein können. Die Durchführung der Recherche und das Sichten der Treffer erfordern einen hohen Zeitaufwand.

Die genannten Schwierigkeiten haben die Informationsabteilung der Bayer Pharma

Forschung dazu veranlasst, den Nutzen einer intelligenten Suchmaschine für das Data Mining im Internet zu testen.

## Intelligente Suchmaschine InsumaScout

Der InsumaScout wurde von der Insuma GmbH in Tübingen (INSUMA=intelligente Suchmaschinen) entwickelt. Diese Suchmaschine bietet folgende Vorteile:

Die Suche läuft automatisiert ab, der Nutzer muss nicht – wie bei den konventionellen Suchmaschinen – die Recherche manuell anstoßen. Dies führt zu einer erheblichen Arbeitserleichterung, da hier auch komplexe Suchanfragen gelöst werden können. Die Anzahl der Suchbegriffe ist nicht begrenzt. Die Suchstrategie kann neben einzelnen Suchbegriffen auch Textblöcke bzw. ganze Textseiten und Internetadressen enthalten. Bei der Recherche handelt es sich um individualisierte Suchprozeduren. Das Programm ist, im Gegensatz zu klassischen Internet-Suchmaschinen, auf den einzelnen Nutzer zugeschnitten. Das Prinzip von InsumaScout zeichnet sich durch eine gewichtete Suche über einen lernfähigen Crawler aus. Ein Crawler ist ein Informationsagent, der im Internet „auf der Jagd“ nach relevanten Internetlinks ist. Die Arbeitsweise des intelligenten Crawlers lässt sich wie folgt beschreiben:

Der themenspezifische, lernfähige Crawler durchsucht das Internet und sammelt Primärhits. Diese werden in einem zweiten Arbeitsschritt gefiltert. Um den Filter passieren zu können, müssen bestimmte Voraussetzungen erfüllt werden. Das Ergebnis sind selektierte Hits, die automatisch vorsortiert werden. Der Anteil der relevanten Treffer ist, verglichen mit den Ergebnissen aus den konventionellen Suchmaschinen, deutlich erhöht. Dem Nutzer stehen die Ergebnisse aus

den einzelnen Recherche-Durchläufen in suchbaren Kollektionen zur Verfügung (Abb.1).

**Aufbauphase der intelligenten Suchmaschine**

Als erster Schritt muss der Filter aufgebaut werden, dann erfolgt der Start des Crawlers. Ein Filter versteht sich als Schlagwortliste mit dazugehöriger Gewichtung. Zum Generieren eines Filters ist eine Ausgangsinformation nötig. Diese setzt sich zusammen aus:

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

Die Art und Menge der Startinformation beeinflusst die Ausgangs-Qualität des Filters. Der Crawler durchsucht das Internet in einem ersten Durchlauf. Im Selektionsprozess erfolgt das Auffinden themenspezifischer Dokumente. Duplikate (identische URLs)

werden automatisch eliminiert. Es kommt zum Aufbau einer themenspezifischen Kollektion (Trefferliste). Die Treffer werden in Ähnlichkeits-Clustern nach Relevanz oder nach URLs sortiert. Dem Nutzer wird das Suchergebnis in einem „Control Center“ angezeigt.

**Routinephase der intelligenten Suchmaschine**

Die Routinephase zeichnet sich durch kontinuierliches Lernen des Filters aus. Der Filter verändert sich durch das Beurteilen der Treffer (Rating) oder durch Hinzufügen bzw. Entfernen von URLs, Schlagwörtern und Textblöcken. Je mehr Dokumente als relevant beurteilt werden, desto höher ist die Filterqualität und die Qualität der Treffer im nächsten Durchlauf.

Für das Beurteilen stehen fünf Relevanzstufen von „+2“ bis „-2“ zur Verfügung.

In der Routinephase werden die Dokumente durch den Nutzer bewertet, was anschließend zur Anpassung des Filters führt.

**InsumaScout – Recherchethemen**

In einem Pilotprojekt wurden drei sehr unterschiedliche Recherchethemen getestet.

Alzheimer'sche Erkrankung  
 —> *Terminologie eindeutig*

Kardiovaskuläre Erkrankung  
 —> *Terminologie nicht immer eindeutig*

Naturstoffe  
 —> *Terminologie nicht eindeutig (im Sinne pharmazeutischer Anwendung)*

Anhand der unterschiedlichen Fragestellungen sollte festgestellt werden, wie eine intelligente Suchmaschine mit Themen, die klar definierbar sind, aber auch mit Themen, für die es nicht immer eine genau definierbare und eindeutige Terminologie gibt, umgeht. Diese drei Themen wurden über einen längeren Zeitraum getestet. Die daraus resultierenden Ergebnisse wurden mit den Treffern aus den konventionellen Suchmaschinen verglichen. Dabei wurde festgestellt, dass die

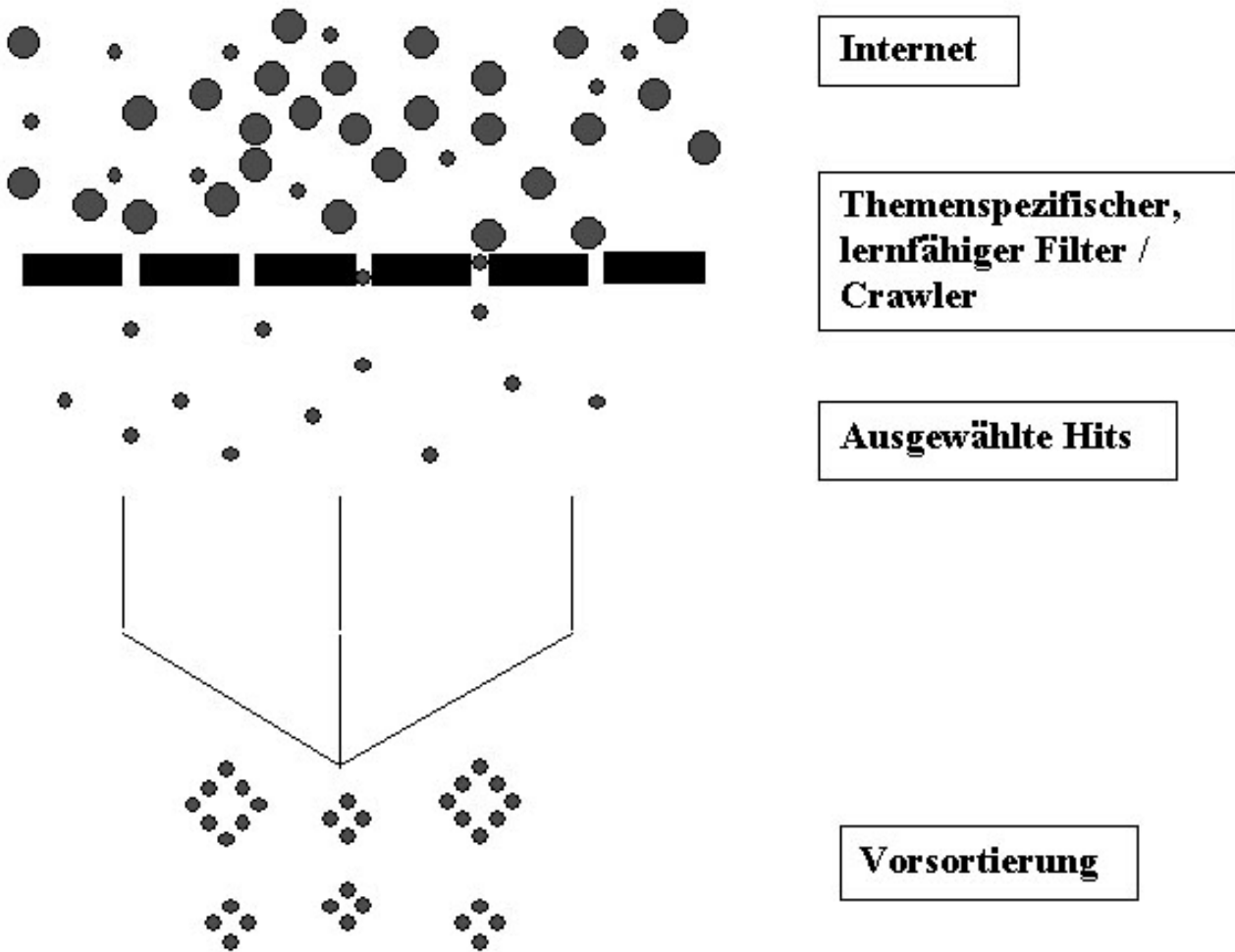


Abb. 1: Selektionsprinzip des InsumaScout

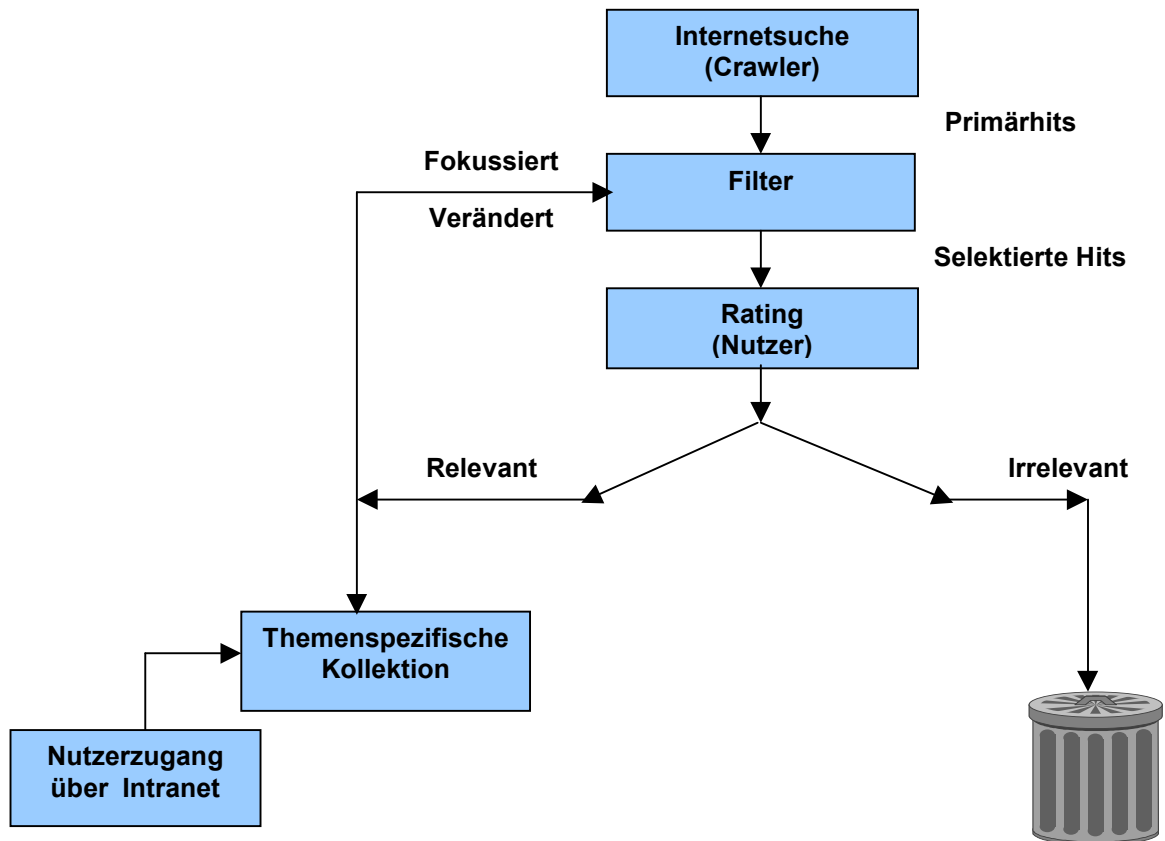


Abb. 2: Workflow InsumaScout

intelligente Suchmaschine in der Lage ist, eine hohe Bandbreite an Fragestellungen zu bearbeiten.

Eine Untersuchung der Qualität der Treffer ergab folgendes Ergebnis:

Zu den Themen „Alzheimer’sche Erkrankung“ und „Naturstoffe“ wurden jeweils 2600 Dokumente (100 Dokumente pro Woche) durch Endnutzer bewertet.

Als relevant (Relevanz +1 bzw. +2) beurteilt wurden bei „Alzheimer“ 19% und bei „Naturstoffe“ 28%. Bei herkömmlichen Suchmaschinen wurde dagegen nur eine Relevanz von unter 5% beobachtet.

#### Intelligente Suchmaschine – Rechercheergebnisse und Fazit

Eine intelligente Suchmaschine kann nicht die gesamte intellektuelle Arbeit eines Nutzers ersetzen. Sie nimmt dem Nutzer jedoch einen großen Teil der Routine-Arbeit ab und analysiert die Suchergebnisse.

Es gibt Schwierigkeiten und Grenzen, an die auch eine solche Suchmaschine stoßen kann. Dazu zählen von der Terminologie her relevante, aber für den Nutzer bereits bekannte Informationen. Internet-Links mit wissenschaftlich niedrigem Anspruch sind für einen Wissenschaftler uninteressant. Dazu zählen z.B. Patientenforen oder Übersichten zu bestimmten Krankheiten für Laien. Diese Internetseiten bieten selten einen Neuheits-

wert und überwiegend keinen wissenschaftlichen Anspruch.

Publikationen aus wissenschaftlichen Zeitschriften werden besser durch die Literaturrecherche in bibliographischen Datenbanken abgedeckt.

In der Trefferliste können Links enthalten sein, die ein aktuelles Datum haben, deren Inhalt jedoch veraltet ist. Hierbei handelt es sich um ein generelles Problem der Suchmaschinen.

Die Vorteile der intelligenten Suchmaschine übertreffen die Nachteile bei weitem. Zu den wichtigsten Vorzügen zählt der intelligente, lernfähige Crawler, der in einem aufwendigen Selektionsprozess relevante Internetsites ermittelt. Die Anzahl irrelevanter Hits ist stark reduziert. Der Nutzer muss nicht umfangreiche Trefferlisten sichten, um ein paar wenige relevante Dokumente zu finden. Der Arbeitsaufwand, um relevante Treffer zu finden, ist demnach deutlich reduziert.

Die erhöhte Suchfrequenz erlaubt das Auffinden hochaktueller Informationen. Der Nutzer hat die Möglichkeit, die Suchfrequenz individuell festzulegen.

Ein weiterer Vorteil ist die automatische Dubletteneliminierung. Das Programm erkennt identische URLs und zeigt diese nicht mehrfach an. Dadurch bleibt dem Nutzer unnötiger Ballast erspart.

Die Rechercheergebnisse werden für den Nutzer weiter aufbereitet. Es werden

themenspezifische Kollektionen aufgebaut. Die Dokumente können in Ähnlichkeits-Clustern nach Relevanz oder nach URLs angezeigt werden. Die einzelnen Kollektionen sind anschließend für den Nutzer recherchierbar.

Die automatisierte Suche bietet dem Anwender großen Nutzen. Die Recherche kann einfacher und schneller durchgeführt werden als die manuelle Suche mit herkömmlichen Suchmaschinen.

Bei dem Vergleich zwischen konventionellen und intelligenten Suchmaschinen stellt sich heraus, dass sie sich in der Funktionalität deutlich unterscheiden. Intelligente Suchmaschinen sind klar überlegen. Um im Internet komplexe Suchanfragen mit möglichst geringem Aufwand und intelligenter Analyse bewältigen zu können, sollten die Vorzüge einer intelligenten Suchmaschine genutzt werden.

Dipl.-Dok. Helga Walter  
Bayer HealthCare AG  
Pharma Forschung  
Wissenschaftliche Information und Dokumentation  
D-42096 Wuppertal  
Tel.: +49 (0) 202 36 8241  
Fax: +49 (0) 202 36 4200  
E-Mail: helga.walter@bayerhealthcare.com