

Wer sucht, der findet?

Bessere Informationsrecherche im World Wide Web

Wer vor 20 Jahren als Schüler, Student, Wissenschaftler oder Journalist Informationen suchte, brauchte Geduld und flinke Zeigefinger: Millionen von Karteikärtchen in den Schubkästen der Autoren- und Schlagwort-Kataloge von Bibliotheken warteten darauf, durchgeblättert zu werden. Informationsbeschaffung kostete Zeit und Nerven. Das World Wide Web stellt heute eine immens große Anzahl von Informationsquellen online zur Verfügung. Suchmaschinen vermitteln dabei vor allem ungeübten Nutzern die Illusion, dass sie alle im Internet vorhandenen Informationen finden und dadurch herkömmliche Quellen wie Bibliothekskataloge oder Literaturdatenbanken überflüssig machen. Ein Trugschluss.

Genauere Analysen zeigen nämlich, dass ein großer Teil der wissenschaftlichen Literatur nicht auf normalen Web-Seiten zu finden ist, sondern bestenfalls in zugangsbeschränkten Datenbanken, die durch die Web-Suchmaschinen nicht erfasst werden. Darüber hinaus sind Web-Suchmaschinen grundsätzlich nur bedingt für die Literaturrecherche geeignet.

Wissenschaftler der Duisburg-Essener Fakultät für Ingenieurwissenschaften haben nun ein System entwickelt, das die wissenschaftliche Recherche im Internet deutlich verbessert. „Daffodil“ bietet nicht nur vielfältige, komfortable Möglichkeiten zur Suche, sondern unterstützt auch die weitere Arbeit mit den gefundenen Dokumenten.

Zunächst sollen jedoch die Probleme bei der bisherigen Literatursuche am Beispiel der erfolgreichen Suchmaschine Google und des speziell für die wissenschaftliche Literaturrecherche entwickelten Dienstes Google Scholar demonstriert werden.

Eine Eins und 100 Nullen: Google

Für viele Nutzer – insbesondere für Studenten – ist Google (www.google.com) die wichtigste Quelle bei der Suche nach wissenschaftlicher Information. Der Name „Google“ ist ein Wortspiel mit dem mathematischen Begriff „Googol“, den der Amerikaner Edward Kasner geprägt hat. Er bezeichnet eine 1 mit 100 Nullen. Ganz so viel bietet Google zwar noch nicht, aber immerhin: Mitte 2006 indizierte der Dienst mehr als 25 Milliarden Web-Seiten. Damit deckt Google den frei zugänglichen Teil des WWW sehr gut ab.

Die besondere Stärke von Google liegt in der Sortierung der Suchergebnisse mit Hilfe des so genannten Page-Rank-Algorithmus. Dadurch werden besonders gute Antworten an den Anfang der Trefferliste gestellt. Dieses Verfahren berücksichtigt die Links zwischen den Web-Seiten: Eine Seite erhält ein umso höheres Gewicht, je mehr Links auf sie verweisen. Zusätzlich wird auch das Gewicht der Seiten berücksichtigt, die auf sie verweisen. Vertiefende Analysen und Experimente haben gezeigt, dass dieses Verfahren populäre Web-Seiten bevorzugt und zwar insbesondere Homepages und Linkverzeichnisse auf den Eingangsseiten einer Website. Seiten, die in der Hierarchie einer Website tiefer liegen, tauchen dagegen in der Trefferliste erst später auf. Bei Anfragen, die nicht auf das Finden von Homepages abzielen, wird daher die Qualität der Antworten durch das Page-Rank-Verfahren eher verschlechtert.

Neben dem Qualitätsproblem zeigt sich auch, dass sich Google bei der Suche nach wissenschaftlicher Literatur nur für sehr einfache Informationsprobleme eignet.

Der dänische Informationswissenschaftler Peter Ingwersen vom Department for Information Studies der Royal School of Library and Information Science in Kopenhagen schlägt dagegen eine Typisierung von Informationsproblemen vor, die durch die Aufgabenkomplexität und die Art des Informationsbedürfnisses bestimmt wird.

Wie Abbildung 1 zeigt, wird das Informationsbedürfnis dabei durch Art und Anzahl der gesuchten Objekte charakterisiert, während die Aufgabenkomplexität durch das Wissen über den Suchgegenstand beschrieben werden kann. Das einfachste Informationsproblem ist demnach die Suche nach einem bekannten Objekt, um zum Beispiel den Volltext eines Dokumentes zu finden oder die Literaturangaben zu vervollständigen. Schwieriger wird es schon, wenn wir nur partielles Wissen über ein Objekt haben – bei-

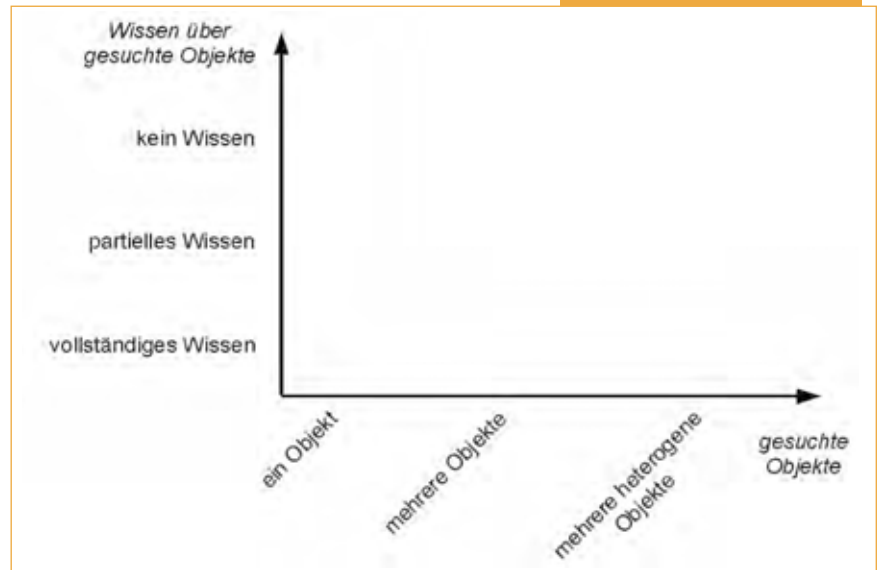


Abbildung 1:
Typisierung von
Suchproblemen

spielsweise kennen wir den Autor und das Thema eines Dokumentes, aber nicht seinen Titel – oder wenn wir mehrere bekannte Objekte, etwa alle Publikationen eines Autors aus einem bestimmten Zeitraum, suchen. Die komplexesten Informationsprobleme liegen vor, wenn man kaum Wissen über das gesuchte Thema hat und zugleich mehrere heterogene Objekte sucht – etwa um sich in ein neues Gebiet einzuarbeiten.

Basierend auf dieser Systematik haben die Forscher um Norbert Fuhr Experimente durchgeführt, bei denen die Nutzer Informationsprobleme unterschiedlicher Komplexität bearbeiten sollten. Dabei arbeiteten sie zum einen mit Google und zum anderen mit dem an der Universität Duisburg-Essen entwickelten System Daffodil. Hierbei zeigte sich, dass Google nur für die einfachste Problemstellung geeignet ist, also die Suche nach einem bekannten Objekt. Schon bei geringfügig komplexeren Aufgaben wird die Suche mühselig oder verläuft gänzlich ohne Ergebnis.

Angebot für Akademiker: Google Scholar

Eine wesentliche Beschränkung von Google besteht darin, dass hier nur frei zugängliche Web-Seiten erfasst werden. Ein großer Teil der wissenschaftlichen Veröffentlichungen liegt aber auf zugangsbeschränkten Datenbanken – insbesondere bei Verlagen, die Inhalte nur gegen Gebühr zur Verfügung stellen. Um auch diese Dokumente zu erfassen, wurde im November 2004 Google Scholar gestartet (scholar.google.com). Die Verlage stellen Google für diesen Dienst auch ihre zugangsbeschränkten Inhalte zur Indexierung zur Verfügung. Zusätzlich zeigt Google Scholar wissenschaftliche Dokumente auf frei zugänglichen Servern an. Im Juni 2006 erfasste Google Scholar schätzungsweise 1 Milliarde Dokumente.





Abbildung 2: Formular für die erweiterte Suche in Google Scholar

Abbildung 2 zeigt das Suchformular von Google Scholar. Als Erweiterung der normalen Web-Suche kann man hier zusätzlich noch Autoren, Erscheinungsjahr und Publikationsorgan angeben, um dadurch präziser zu suchen. Ein entsprechendes Suchergebnis ist in Abbildung 3 dargestellt. Die Ergebnisse werden normalerweise nach Anzahl der Zitationen geordnet, so dass man wieder populäre Dokumente am Beginn der Trefferliste findet.

Im Normalfall liefert Google Scholar entweder einen direkten Verweis auf das Originaldokument oder ein Formular, mit dem ein gesuchtes Dokument gegen Gebühr von einem Verlag bezogen werden kann. Allerdings ist zu beachten, dass ein großer Teil der Antworten nur aus Zitationen besteht, also keinen direkten Verweis auf die Volltexte der gesuchten Publikationen liefern.

Ein weiteres Manko von Google Scholar sind Aktualität und Vollständigkeit: Da Google Scholar die Verlagsserver jeweils neu durchsuchen muss, um neue Dokumente aufzuspüren, sind diese so genannten Crawls relativ zeitaufwändig und werden daher weitaus seltener als im Falle von Google Web Search durchgeführt, so dass neue Inhalte oft erst mit einigen Monaten Verzögerung sichtbar sind. Zudem werden viele frei zugängliche Bestände insbesondere aus dem Open-Access-Bereich bisher nur wenig berücksichtigt.

Aktualität und Vollständigkeit leiden aber auch darunter, dass Google Scholar nur Volltexte erfasst. Publikationen von Verlagen, mit denen Google kein entsprechendes Abkommen getroffen hat, werden in Google Scholar nur über den Umweg von Zitati-

onen sichtbar – wenn also andere Autoren von diesen Publikationen Notiz genommen haben, sie zitieren und der entsprechende Artikel publiziert und von Google Scholar indexiert ist. Dies bedeutet erhebliche Nachteile in punkto Aktualität. Im Gegensatz dazu sind Datenbanken, die in der Regel frei zugängliche Metadaten von Publikationen erfassen, wesentlich aktueller.

Google Scholar stellt zwar ein neues interessantes Angebot zur Literaturrecherche dar, Aktualität und Vollständigkeit lassen aber bisher zu wünschen übrig. In vielen Fachgebieten gibt es – auch frei zugängliche – Datenbanken, die hier bessere Angebote darstellen.

Lasst Blumen sprechen: Daffodil

Ein Suchender, der Datenbanken oder digitale Bibliotheken zur Literaturrecherche nutzen möchte, wird schnell feststellen, dass sich diese oft erheblich in Funktionalität, Anfragesprache, Eingabemaske und dargebotener Benutzeroberfläche voneinander unterscheiden. Eine reine Meta-Suchmaschine böte zwar für Literaturdatenbanken einen einheitlichen Zugriff, reicht aber für viele Anwendungen nicht aus.

Das an der Universität Duisburg-Essen entwickelte System Daffodil bietet dagegen eine integrierte Suche in heterogenen Datenbeständen eines Fachgebiets und führt die Ergebnisse zusammen. Darüber hinaus werden die Benutzer bei ihrer Suche durch eine Reihe von integrierten Werkzeugen strategisch unterstützt.

Daffodil (Akronym für: „Distributed Agents for User-Friendly Access of Digital Libraries“, auch Englisch für: Osterglocke) verbindet in seiner graphischen Benutzeroberfläche in natürlicher Weise Browsing- und Suchstrategien. Die angebotenen Werkzeuge unterstützen den Anwender durch eine Reihe höherer Suchfunktionalitäten. Zusätzlich bietet das System bereits bei der Anfrageformulierung durch Fehlerkorrektur und Vorschläge Hilfestellungen an.



Abbildung 3: Suchergebnisse in Google Scholar

Für Nadeln im Heuhaufen: Zentrales Suchwerkzeug

Das zentrale Arbeitsmittel des Systems ist das Suchwerkzeug, das in Abbildung 4 links zu sehen ist. Es ist der gebräuchlichste Ausgangspunkt für eine Literaturrecherche in Daffodil. Das Suchwerkzeug bietet eine aus vielen anderen Suchmaschinen vertraute Eingabemaske, die es erlaubt, einheitliche Anfragen an die verteilten Datenbanken zu formulieren. Die Suchdomäne kann durch die Auswahl der angebotenen Literaturdatenbanken eingeschränkt werden.

Vor der Verarbeitung der Anfrage und der Weiterleitung an die Datenbanken setzt die aktive Anfrageunterstützung an. Benutzer sind bei der Bearbeitung von Suchaufgaben oft unsicher, haben nur unklar formulierte Ziele und können oder wollen der Aufgabe oft nicht ihre volle Aufmerksamkeit widmen. Die proaktive Unterstützung bei der Anfrageformulierung soll den Nutzer hier entlasten.

So sollen Suchende auf mögliche Anfragefehler oder -probleme, aber auch auf Alternativen aufmerksam gemacht werden. Ist das Themengebiet der Suche unzulänglich bekannt, stellt sich oft das Vokabularproblem: Benutzer wissen ihr Suchproblem nicht mit geeigneten Termen oder Synonymen zu beschreiben, um die erwünschten Ergebnisse zu erzielen. Daffodil zeigt mögliche Rechtschreibfehler auf (Abb. 5), weist auf überspezifizierte Suchbedingungen hin, für die keine Resultate zu erwarten sind (Abb. 6), und bietet Synonyme und verwandte Begriffe zu Suchtermen an (Abb. 7).

Ziel der Unterstützung ist die Vermeidung oder Verringerung typischer, grundlegender Fehler bei der Anfrageformulierung, die Entlastung des Benutzers, der schnelle Rückmeldung zu eventuellen Problemen der Anfrage erhält, und ein größeres Vertrauen des Benutzers in die Angemessenheit und Korrektheit der formulierten Anfrage.

Die Suchanfragen werden von Daffodil über so genannte Wrapper an die Datenbanken der Anbieter weitergeleitet und parallel bearbeitet. Die Ergebnisse werden anschließend zusammengeführt und in einheitlicher Weise zur Betrachtung und Navigation als gewichtete Resultatliste präsentiert. Bereits betrachtete, gespeicherte oder bearbeitete Dokumente werden dabei auf graphische Weise kenntlich gemacht. Mit Hilfe von Extraktionsfunktionen kann aus den Resultaten zusätzlich ein Überblick über besonders bedeutende Autoren, Konferenzen, Journale oder Schlagworte eines Suchergebnisses gewonnen werden. So entstehen neue Ansatzpunkte für eine weiterführende Suche.

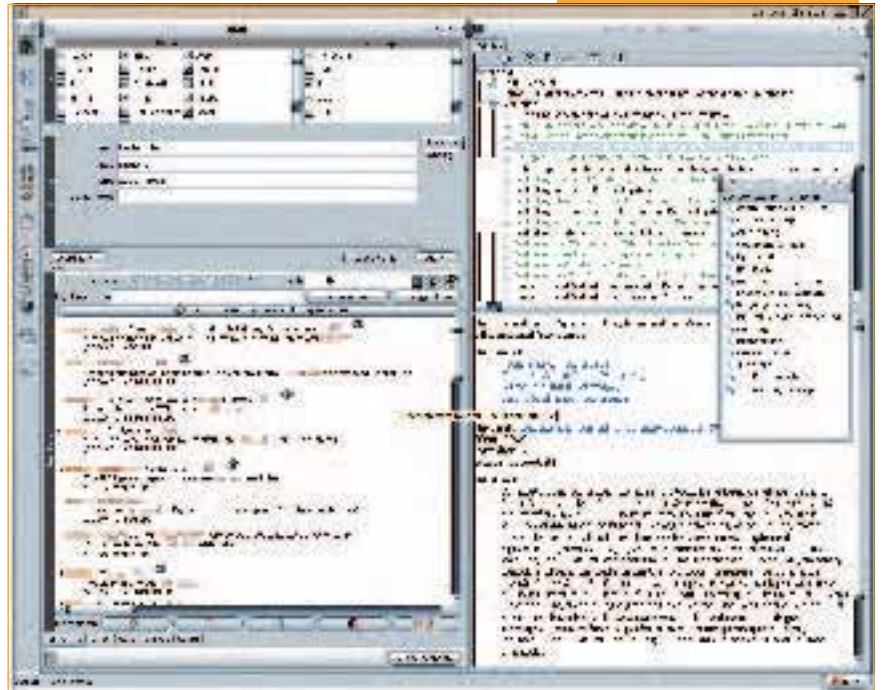


Abbildung 4: Der Daffodil-Desktop (Suchwerkzeug links, Handbibliothek rechts)

Wissensspeicher: Persönliche Handbibliothek

In der persönlichen Handbibliothek (Abb. 4, rechts) ist das Speichern von Anfragen und Suchergebnissen, Volltexten, Stichworten oder Suchtermen, wichtigen Autoren, Konferenzen oder interessanten Zeitschriften über den Kontext einer Suche hinaus möglich. In persönlichen Ordnern können Suchende ihre Ergebnisse strukturiert ablegen und so über mehrere Suchsituationen hinweg ein Archiv ihrer Literaturrecherchen aufbauen. In gemeinsam genutzten Gruppenordnern können Recherchierende zusammenarbeiten, ihre Ergebnisse teilen und Anmerkungen zu gefundenen Dokumenten notieren.

Alle abgelegten Objekte, seien es Dokumente oder Anfragen, können zu einem späteren Zeitpunkt mit den anderen Werkzeugen des Daffodil-Desktops weiterverwendet oder bearbeitet werden. Für die Nutzung der Resultate außerhalb des Systems stehen Exportfunktionen für eine Reihe gebräuchlicher Formate zur Verfügung.

Auf Grundlage der in Ordnern gespeicherten Dokumente ist es möglich, sich durch das System Empfehlungen geben oder über passende Neuerscheinungen informieren zu lassen.

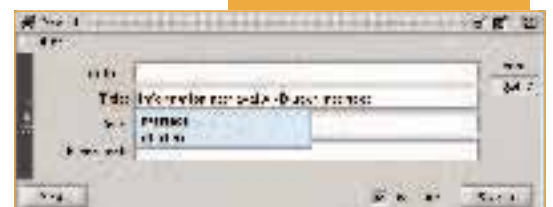


Abbildung 5: Korrektur von Schreibfehlern



Abbildung 6: Überspezifizierte Anfragen

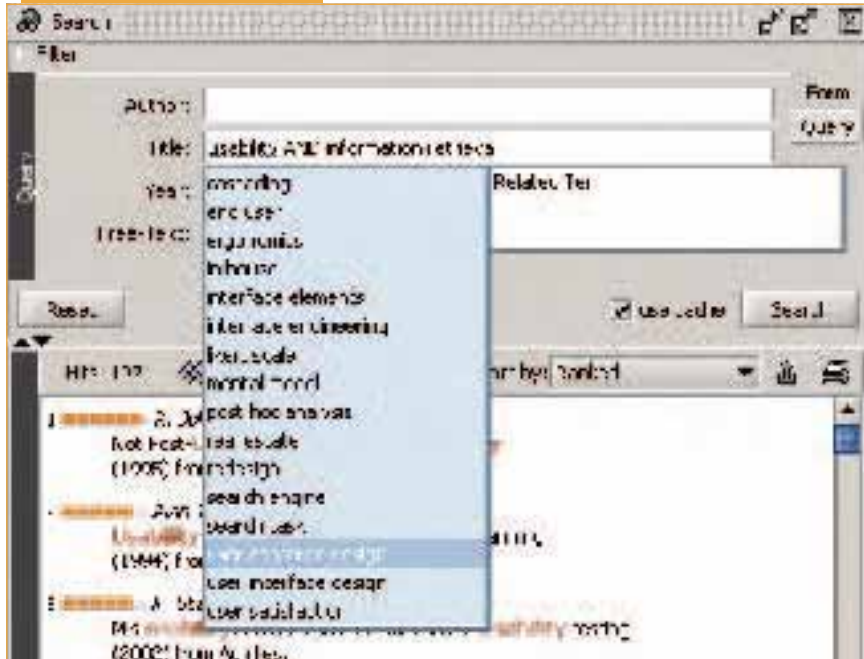


Abbildung 7:
Verwandte Begriffe

Effektive Recherche durch strategische Unterstützung

Während herkömmliche, insbesondere webbasierte Suchmaschinen oft nur einfache, grundlegende Suchoperationen unterstützen, soll Daffodil durch die enge Integration von höheren Suchfunktionen gerade Nichtexperten eine effektivere Literaturrecherche ermöglichen.

Marcia Bates, Professorin am „Department of Library & Information Science“ der Universität von Los Angeles, hat in diesem Zusammenhang drei höhere Abstraktionsebenen für die Beschreibung von Suchaktivitäten definiert:

- Unter *Taktiken* sind einzelne grundlegende Suchoperationen oder einfache Kombinationen solcher zu verstehen. Hierunter fasst man etwa die Schritte, die notwendig sind, um eine Suchanfrage zu verfeinern oder einen Suchbegriff zu generalisieren.
- *Strategeme* stellen eine komplexe Abfolge von Aktionen aus Grundoperationen und Taktiken dar, welche die Struktur einer Domäne zur Suche ausnutzen. Hierbei kann es sich zum Beispiel um die Ausnutzung von Autorenbeziehungen für eine autorenbasierte Suche handeln.
- Einer *Strategie* schließlich liegt ein vollständiger Plan zur Ausführung einer Recherche zugrunde. Strategien setzen sich im Allgemeinen aus einer Reihe von Grundoperationen, Taktiken und Strategemen zusammen.

Daffodil setzt auf der Ebene der Strategeme an und bietet durch strategische Unterstützung Benutzern die Möglichkeit, komplexe Suchstrategien

zu verfolgen. Durch diese höheren Suchfunktionen soll der Benutzer zusätzlich in allen Phasen der Informationsnutzung begleitet werden:

- In der Phase des Entdeckens (*Discover*) erlaubt Daffodil die Suche und Auswahl von Datenquellen durch den Benutzer. Die verschiedenen Werkzeuge gestatten unterschiedliche Startpunkte für eine Suche.
- In der Phase des Findens (*Retrieve*) durchsuchen Benutzer die ausgewählten Datenquellen mit Unterstützung des Systems nach Dokumenten, die ihr Informationsbedürfnis befriedigen.
- In der Phase des Zusammenführens (*Collate*) erlaubt Daffodil die strukturierte und geordnete Ablage von Ergebnissen und gefundenen Dokumenten.
- In der Interpretationsphase (*Interpret*) ermöglicht Daffodil die Annotation von Dokumenten und unterstützt auch die kollaborative Arbeit durch den gemeinsamen Zugriff auf Dokumente und Annotationen.
- Schließlich erfolgt in der abschließenden Phase (*Re-Present*) die Aufbereitung des Gefundenen zu einer neuen Darstellung – sei es in einem Aufsatz, einem Artikel oder einem längeren wissenschaftlichen Text.

Die Unterstützung in allen Phasen bietet dem Recherchierenden eine umfassende Begleitung seines Suchprozesses. Auf einige der Werkzeuge, die zur Literatursuche höhere Suchfunktionen bereitstellen, soll im Folgenden kurz eingegangen werden.

Referenzen und Zitationen

Das Referenzwerkzeug erlaubt ausgehend von einem bekannten Dokument – etwa aus der Resultatliste oder der persönlichen Handbibliothek – die Suche nach anderen Dokumenten, die von diesem zitiert worden sind oder dieses selbst zitieren. Das Werkzeug kann intuitiv durch Drag&Drop eines Dokuments aus anderen Werkzeugen heraus aktiviert werden. Die Ergebnisse lassen sich direkt weiter verwenden und können wiederum in der Persönlichen Handbibliothek abgespeichert werden.

Autorennetzwerke

Ein weiteres, oft benutztes und von Daffodil unterstütztes Suchstrategem ist die Autorensuche. Ausgehend von einem Autoren, dessen Relevanz für das Suchinteresse bekannt ist, kann zum Beispiel aus einem konkreten Dokument des Autors oder einer extrahierten Autorenliste nach weiteren Publikationen dieses Autors gesucht oder die Ko-Autorenbeziehungen des Autors für den Aufbau eines Beziehungsnetzwerks ausgenutzt werden.

In einem solchen Ko-Autorennetzwerk, das auch als Beziehungsgraph visualisiert werden kann (Abb. 8), lassen sich leicht sowohl zentrale Autoren erkennen als auch solche finden, die häufig gemeinsam publizieren.

Journalne und Konferenzen

Zum Blättern und Suchen in den Jahrgangsbänden wissenschaftlicher Zeitschriften oder in den Berichten wichtiger Fachkonferenzen steht das Journal- und Konferenzwerkzeug zur Verfügung. Hier kann nach Titeln von Journalen oder Konferenzen gesucht werden, um anschließend innerhalb der Ergebnisse zu browsen – oft mit direktem Zugriff auf Metadaten oder Volltextlinks.

Der Einsatz des Werkzeuges kann Ausgangspunkt oder Zwischenschritt eines umfangreicheren Suchplans sein. Explizite Verknüpfungen in den Detailansichten von Suchergebnissen weisen auf eine Zeitschrift oder einen Konferenzband hin, in dem ein Dokument ursprünglich veröffentlicht wurde, und erlauben den direkten Aufruf des Journal- und Konferenzwerkzeuges aus einem Ergebnisdokument heraus.

Klassifikationen

Mit Hilfe eines Klassifikationswerkzeuges erhalten Benutzer des Systems Zugriff auf eine hierarchische, themenorientierte Darstellung des Suchraumes. Das Werkzeug erlaubt das Browsen in Klassifikationsschemata wie dem ACM Computing Classification Scheme und die Übernahme von Klassifikationstermen in eine Suchanfrage.

Thesauri

Über den Thesaurus können zu Schlagworten allgemeine oder spezifischere Begriffe gefunden werden. Auch Erklärungen zu Begriffen lassen sich abfragen. Fachspezifische oder webbasierte Thesauri werden für das Auffinden verwandter Begriffe benutzt. Die so gefundenen Begriffe können dann leicht in anderen Werkzeugen zur weiteren Bearbeitung oder für Anfragen genutzt werden.

Erfolgsmodell für die Universität

Web-Suchmaschinen wie zum Beispiel Google indexieren zwar riesige Datenbestände, haben aber keinen Zugriff auf zugangsbeschränkte Inhalte; außerdem bieten sie keine adäquaten Funktionen zur Literaturrecherche. Google Scholar überwindet diese Nachteile teilweise, leidet aber unter mangelnder Aktualität. Das an der Universität Duisburg-Essen entwickelte System Daffodil stellt dagegen eine Meta-Suchmaschine für Literaturdatenbanken dar und bietet zusätzlich zahlreiche weitergehende Funktionen. Das Konzept der

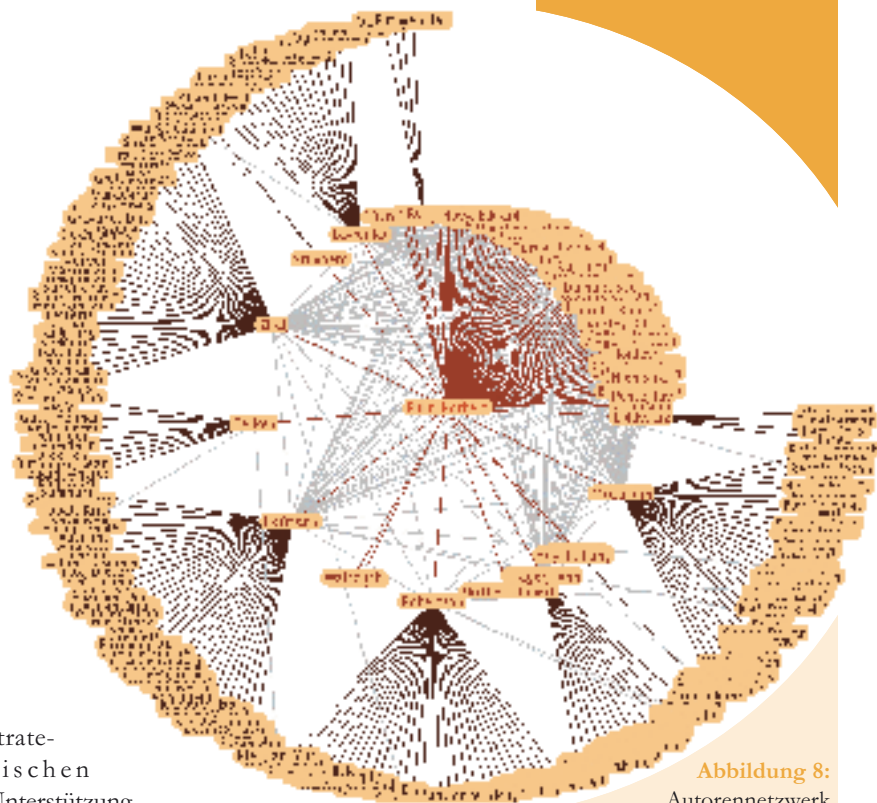


Abbildung 8:
Autorennetzwerk

strategischen Unterstützung wird in Daffodil auf allen Ebenen des Systems realisiert; eine Handbibliothek bietet dabei Personalisierung und Gruppenunterstützung. Benutzer werden durch das System in die Lage versetzt, komplexe Suchstrategien über mehrere, heterogene Literaturdatenbanken auf natürliche Weise zu verfolgen. Die Aufbereitung der Suchresultate durch Daffodil hilft bei der kognitiven Erfassung der Information. Proaktive Funktionen bieten zusätzliche Hilfestellungen an.

Für die Literaturrecherche steht somit eine Meta-Suchmaschine zur Verfügung, die nicht allein die gleichzeitige Suche in unterschiedlichen Datenbeständen erlaubt, sondern den Benutzer in allen Phasen des Suchprozesses strategisch unterstützt.

Daffodil ist derzeit nicht nur mit Erfolg in der Abteilung für Informatik im Einsatz, sondern wird von verschiedenen Instituten und Universitäten in Forschung und Lehre eingesetzt. Der Prototyp kann unter <http://www.daffodil.de> via „Java Web-Start“ genutzt werden. Die Quellen zu Daffodil unterliegen der „Apache License“.

Kontakt

Prof. Dr.-Ing. Norbert Fuhr
Dipl.-Inform. Sascha Kriewel

Informationssysteme

fuhr@is.informatik.uni-duisburg.de
kriewel@is.informatik.uni-duisburg.de

<http://www.is.informatik.uni-duisburg.de>

