Die Aufbereitung von Suchergebnissen mittels Visualisierungstechniken

Martha Argentina Barberena Najarro
Technische Universität Ilmenau
Institut Praktische Informatik und Medieninformatik
Fachgebiet Telematik
Email: Martha.Najarro@PrakInf.tu-ilmenau.de

Abstract: Im Gegensatz zu den inzwischen hoch entwickelten Technologien zur Dokumentensuche ist die Präsentation der Ergebnisse von Suchmaschinen nach wie vor nahezu ausschließlich auf eine unübersichtliche, schlecht handhabbare textuelle Darstellung in Lis tenform begrenzt. Dieser Beitrag beschreibt einen Vorschlag zur Beseitigung dieses Mißverhältnisses mittels Methoden der Computergrafik, Filtertechniken und Dokumentenmanagement.

1 Einleitung

Die Suche nach der gewünschten Information, die über das Worl d Wide Web verfügbar ist, wird immer schwieriger und ist momentan sehr mühsam. Die Lage wird sich weiter verschlechtern, da die Menge der Dokumente und Informationsanbieter ständig wächst [May]. Suchmaschinen als Suchdienste im Internet bieten Suchmethoden für Hypertexte im Web an [Bek], so daß die Suche effektiver wird und zu einem gezielten Ergebnis führt und so wird auch die Relevanz der gefundenen Dokumente gesteigert und die Treffermenge recht schnell eingeschränkt [Phi]. Aber dennoch haben sich die Be reitstellung und die Präsentation der Suchergebnisse durch die textbasierte Dokumentenliste nicht wesentlich verbessert. Die Ergeb nisliste zeigt keinen Zusammenhang zwischen den Dokumenten. Besonders bei der Verwendung von mehreren Suchbegriffen ist sehr schwierig, die Beziehung der Dokumente zu den einzelnen Suchbegriffen zu erkennen. Es wird keine zusätzliche Hilfe zur Verbesserung der Anfrage gestellt . Eine Darstellung der gesamten Ergebnismenge muß in der Regel auf mehrere Bildschirmseiten verteilt wer den. Durch diesen Umstände geht die Übersicht über die Ergebnismenge schnell verloren.

Aufgrund dieser Problematik steht im Vordergrund die Forderung nach neuen Techniken zur Informationsaufbereitung und Recherche. Hier wird deswegen ein neuer Ansatz v orgestellt, der die Informationssuche effektiver gestaltet und das Navigieren im Informationsraum erleichtern soll.

Es gibt eine Vielzahl existierender Ansätze für die Visualisierung von Informationsmengen [Cha]. Ein Beispiel dafür sind das System *Hyperbolic Tree* [Inx, Cha] und der *Hyperspace* [Dae]. Diese und andere Ansätze wurden in erster Linie für die reine Darstellung von Suchergebnissen entwickelt und enthalten daher oft auch

entsprechend hoch entwickelte und komplexe Visualisierungsverfahren. Dabei wird versucht, strukturelle Beziehungen und Zusammenhänge aufzuzeigen.

Ziel des hier vorgestellten Ansatzes ist der Entwurf einer Architektur, die sowohl die Darstellung von Suchergebnissen präsentiert als auch das schnelle und zielgerichtete Auffinden der wirklich interessanten Information mittels grafischer Darstellung, Filtermechanismen und interaktiver Verwaltung der Dokumente gestattet. Ausgangspunkt ist auch dabei die von einer beliebigen Suchmaschine oder Wissensbasis ermittelte Menge von Dokumenten, die den angegebenen Suchbegriffen und -Kriterien entspricht.
Dieser Ansatz schließt auch die Auswertung vorhandener (aber zu wenig verwendeter) HTML-Features ein – den Meta-Tags. So können Deskriptoren, die im Bibliothekswesen erprobt sind, zur Klassifizierung von Dokumenten herangezogen werden. Die Verwendung solcher Möglichkeiten könnte vor allem im wissenschaftlichen Bereich zum Hervorheben von Publikationen gegenüber dem ständig stärker werdenden „Rauschen“ im Internet beitragen.

2 Das Visualisierungssystem

Als Visualisierungssystem wird hier die Kombination von Suchmaschine und Visualisierungsmodulen verstanden. Die grundlegenden Architekturbestandteile sind die Module Schnittstelle, Control, Filter, Mapping, Renderer und Informationsausgabe. Diese Struktur ist in Abbildung 1 gezeigt. Der zentrale Bestandteil des hier vorgestellten Systems ist die Komponente, die es ermöglicht, den ermittelten Informationsraum mit den darin enthaltenen Dokumenten grafisch darzustellen. Dazu wurde die aus der wissenschaftlich-technischen Visualisierung bekannte Visualisierungspipeline adaptiert [Bart]. Diese Methodik ist in [Bar] beschrieben, daher soll hier nicht näher darauf eingegangen werden.
Der Benutzer formuliert seine Suchanfrage an die Suchmaschine durch Angabe eines oder mehrerer Suchbegriffe, die logisch verknüpft sein können. Die Suchmaschine führt die Dokumentsuche durch und erzeugt eine interne Liste der gefundenen Dokumente. An dieser Stelle ist es ist sehr wichtig, daß die Suchmaschine zu jedem Dokument die berechneten Relevanz-Werte (Gewichtung) pro Suchbegriff bereitstellt, da darauf die Positionierung in der visuellen Repräsentation des Informationsraumes beruht. Das System generiert eine den Anforderungen des Nutzers entsprechende grafische Darstellung und erlaubt durch vielfältige Einstellungsmöglichkeiten die Reduktion der Dokumentenmenge bis auf eine Restmenge, die als wirklich „interessant“ eingestuft werden kann. Somit kombiniert das System in sinnvoller Weise die Vorteile einer direkten Suchanfrage und der Interaktion mit den Dokumenten zum Auffinden der relevanten Informationen. Die Module und ihre Features werden nun im Einzelnen vorgestellt:

Es gibt zwei Möglichkeiten, das Visualisierungsmodul mit Ausgangsdaten zu versorgen:

- direkt über die Schnittstelle, indem die Suchmaschine das Ergebnis weitergibt.

- indirekt, indem die Suchmaschine das Ergebnis in einer Datenbasis speichert und das System die Daten später einliest.

Das Modul Control:
Das Steuer-Modul enthält die Daten, die entweder vo n einer Suchmaschine oder von einer Datenbasis stammen und verwaltet sie. Es steuert außerdem Filter, Mapping und Renderer und reagiert auf die Interaktionen vom Renderer. Weiterhin enthält es die Datenstrukturen zur Verwaltung der Dokumente (z.B. Markier en von interessanten Dokumenten, Löschen irrelevanter Dokumente, ...),

Die Schnittstelle:
Das ist eine wohldefinierte Schnittstelle zum Informationstranfer von der Suchmaschine zu dem Modul Control. Die Übergabe der Suchergebnisse muß über eine definierte Form erfolgen, die durch die Schnittstelle (auf XML -Basis) definiert ist. Es ist festgelegt, welche Informationen pro Dokument verarbeitet werden können. Ebenso kann der gesamte aktuelle Systemzustand über diese Schnittstelle gespeichert werden, um Zwisch energebnisse der Informationsaufbereitung in einer **Datenbasis** festzuhalten.

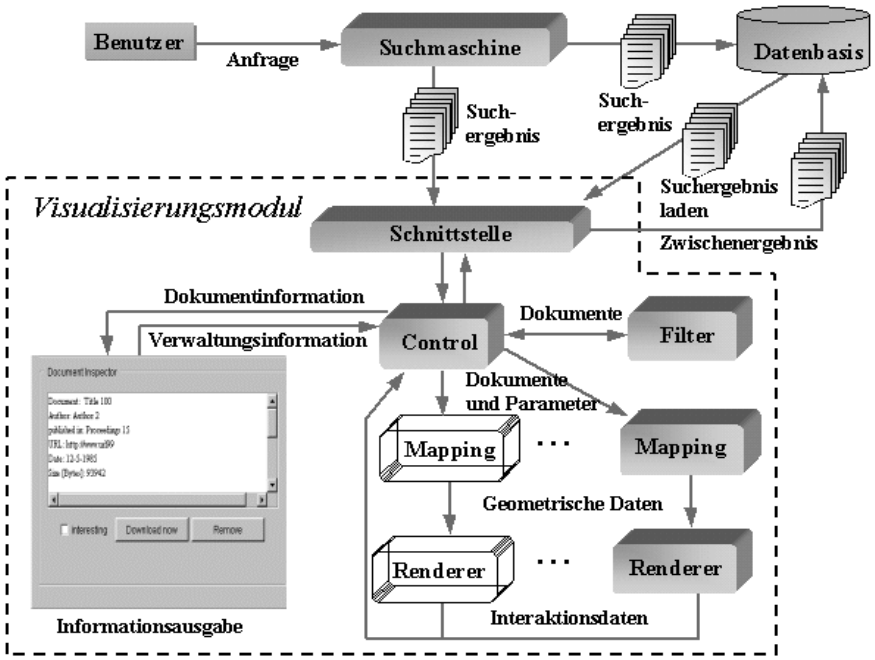


Abbildung 1: Systemarchitektur

Das Modul Filter:
Hier findet eine Reduktion der Dokumentmenge statt, die oft sehr große Ausmaße haben kann. Das ist notwendig, da diese Me nge dreidimensional dargestellt werden soll und diese Darstellung eine hohe Performance erfordert. Der zweite entscheidende Grund ist, daß ein großer Teil der Dokumente ohnehin irrelevant ist. Die Filterstufe kann verschiedene Filter enthalten z.B. für Aut or, Erscheinungsdatum, Dokumentgröße, Erscheinungsort, Relevanz-Schwellwerte und Redundanz.

- Oft sind Autoren bekannt. Der Benutzer möchte, daß nur die Dokumente von bestimmten Autoren angezeigt werden.
- Dokumente, deren Erscheinung vor oder nach einem best immten Datum liegt, können ausgeschlossen werden, so daß ein Erscheinungszeitraum eingestellt werden kann.
- Aus der Größe des Dokumentes kann in gewissen Grenzen auf die Art des Dokumentes geschlossen werden.
- Wenn ein interessantes Dokument gefunden ist, da s zu einer Sammlung gehört, können alle dazugehörigen Dokumente angezeigt werden.
- Pro Suchbegriff kann eine Relevanzschwelle eingestellt werden. Dokumente, bei denen mindestens ein Relevanzwert die entsprechende Schwelle unterschreitet, werden ausgeschlossen.
- Redundanzen in der Dokumentenmenge können auf zwei Arten zustande kommen:
  - identische Dokumente wurden auf verschiedenen Servern gefunden.
  - durch die Anwendung einer Meta -Suchmaschine, wobei mehrere eingebundene Suchdienste das gleiche Dokument gefunde n haben. Dieser Filter stellt sicher, daß die Dokumente nur einmal angezeigt werden.
- Alle Filter – mit Ausnahme des Redundanzfilters - können negiert werden.

Das Modul Mapping:
Es gibt verschiedene Möglichkeiten, den Informationsraum darstellbar zu mache n. Die Auswahl dieser Verfahren ist sehr stark von den Eigenschaften des Informationsraumes (z.B. zwei - oder dreidimensional) und vom verwendeten Renderer abhängig. Dieses Modul generiert eine interne Beschreibung (geometrisches Modell) des Informationsrau mes aus den zur Verfügung stehenden grafischen Ausgabeelementen (z.B. Quadrat, Kugel, Linie usw.). Das hier beschriebene System kann durch Hinzufügen neuer Mapping -Module um beliebig viele Darstellungsverfahren erweitert werden.

Das Modul Renderer:
Aus dem vom Mapping -Modul erzeugten Modell wird nun ein zweidimensionales Bild generiert und zur Anzeige gebracht. Diese Aufgabe übernimmt die Renderingstufe, die gleichzeitig für die Interaktion und Navigation verantwortlich ist. Das Bild zeigt den durch die Suchbegriffe aufgepannten Informationsraum entsprechend dem eingestellten Mapping-Verfahren und die darin enthaltene gefilterte

Dokumentenmenge. Es ist möglich, durch den Raum zu navigieren und Interaktionen mit den Dokumenten durchzuführen.
Präsentiert da s Bild nicht sofort das gewünschte Ergebnis, kann darauf durch Einstellungen an den beteiligten Modulen Einfluß genommen werden.

Die Informationsausgabe:
Dieses Modul erlaubt die detaillierte Ausgabe aller zum aktuell selektierten Dokument verfügbaren Inf ormationen sowie die Klassifikation als „interessant“ (Übernehmen in eine spezielle Liste) oder „uninteressant“ (aus der der Datenstruktur entfernen).

4 Beispielimplementierung

Ein Test mit der Beispiel -Implementierung (in Java) des Systems zeigt gegenwärtig folgendes Ergebnis: Ausgangsbasis ist ein per Zufallsgenerator erzeugtes Suchergebnis mit 5000 hinsichtlich ihrer Relevanz -Werte gleichverteilten Dokumenten, das durch drei angenommene Suchbegriffe ermittelt ist. Die Relevanz -Werte liegen grundsätzlich im Intervall [0.0, 1.0]. Nach der Einstellung des Relevanz -Filters auf den Wert “0.5” für jeden Suchbegriff ist die Menge noch 626 Dokumente groß. Der Erscheinungsdatum der Dokumente ist über einen Zeitraum von 20 Jahren verteilt. Nach der Einstellung des D atumsfilters auf einen Zeitraum von drei Jahren ergibt sich eine Menge von nur noch 64 Dokumenten, die sich mittels grafischer Interaktion leicht weiter evaluieren läßt. Damit ist ansatzweise gezeigt, daß die Zielmenge der relevanten Dokumente mit einem so lchen System innerhalb weniger Minuten ermittelbar ist..

5 Zusammenfassung

Das Ziel besteht also nicht nur in einer visuell erfaßbaren bedeutungstragenden Form der grafischen Darstellung von Ergebnissen einer Dokumentensuche, sondern in einem System zur Info rmationsaufbereitung, das über offene Schnittstellen nach außen verfügt, erweiterbar, anpaßbar und auch plattformunabhängig ist.
Noch dazu kann dieses System an eine beliebige Suchmaschine gekoppelt werden, die in der Lage ist, die Schnittstelle zur Übergabe der Suchergebnisse zu bedienen.
Aufgrund der Erweiterbarkeit kann das System perspektivisch jedem Anwender eine Visualisierung anbieten, die seinen persönlichen Preferenzen entspricht.

Durch die Kombination von Visualisierung und einer Reihe von Fil tern wird vom System sowohl die „zielgerichtete Suche“ auf der Suche nach einem ganz *bestimmten* Dokument als auch die „unscharfe Suche“ – also auf der Suche nach Themen oder Gebiet (Themenkomplex) – gut unterstützt.

Literatur:

[Bar] Barberena Najarro M. *Darstellung von Suchergebnissen in 3D-Dimensionen*. STJA 2000 – MIK00 6, pp. Workshop Multimediale Informations - und Kommunikationssysteme 09-12. Okt 2000, Kongreßzentrum Messe Erfurt AG.

[Bart] Barth R., Baier E., Pahnke B.; *Grafikprogrammierung mit Op enGL*, Bonn; Paris [u.a.] Addison-Wesley 1996; ISBN 3-89319-975-6.

[Bek] Bekavac B., Ritteberger M. *Kontextsensitive Visualisierung von Suchergebnissen.* Hypertext – Information Retrieval - Multimedia 1997 (HIM '97) Dortmund, 29. September - 2. Oktober 1997

[Cha] Cahomei Chen: *Information Visualisation and Virtual Environments;* Lybrary of Congress Cataloging-in-Publication Data; 1. Information display systems 2. Virtual reality 3. Computer Graphics Technoligical innovation; Springer-Verlag London Berlin Heidelberg, 1999; ISBN 1-85233-136-4

[Dae] Däßler R. *: Informationsvisualisierung: Stand, Kritik und Perspektiven.* Fachbereich Dokumentation; Fachshochschule Postdam, 1999.

[Inx] http://www.inxight.com/Content/279.html

[Phi] Philippus Thomas *: Informationssuche im Internet: Tips für Profis* ; VDE - Verlag GMBH, Berlin und Offenbach, 1997 ISBN 3-8007-2214-3.

[Sch] Schwartz D. *: Mehr Information durch Visualisierung von Daten? Konventionelle und innovative Visualisierungstechniken* . Zeitschrift für Bibliothek-Information-Technoligie-Online (B.I.T. Online); Ausgabe 4/1999.

[May] Mayer Matthias; *Kontextvisualisierung: BrowsingIcons und BrowsingGraphs zur Verbesserung der Orientierung und Navigation im World Wide Web,* In Proceedings of 6. Tagung der Deutschen Sektion der Internationalen Gesellschaft für Wissensorganisation Hamburg, Germany, 2000.