Visualisierung von Informationsräumen

Martha Argentina Barberena Najarro

Technische Universität Ilmenau,
Institut Praktische Informatik und Medieninformatik
Fachgebiet Telematik
Martha.Najarro@tu-ilmenau.de

Zusammenfassung Im vorliegenden Beitrag werden neue Methoden zur effizienten Untersuchung der aus Suchergebnissen resultierenden Informationsräume vorgestellt. Die Präsentation des Ergebnisses einer Suchmaschine ist im Allgemeinen auf eine textbasierte Liste beschränkt, von der jeweils nur ein kleiner Ausschnitt für den Benutzer sichtbar ist. Die Liste ist kaum in ihrer Gesamheit zu ergreifen und unübersichtlich. Es wird ein Ansatz beschrieben, der aufgrund einer Kombination aus Filtermechanismen, grafischer Darstellung und Selektionsmechanismen die Analyse von Informationsmengen in ihrer Gesamtheit zuläßt und so zu einer effizienten Informationsextraktion führt.

1 Einleitung

Für den Benutzer von Suchmaschinen ist es schwer, in der textbasierten Informationsliste, die durch einee Suche im World Wide Web geliefert wird, zu erkennen, welche Dokumente für ihn tatsächlich relevant sind. Die Aufbereitung der Suchergebnisse unzureichend, da in der Regel tausende Dokumente aufgelistet werden. Obwohl Suchmaschinen die Ergebnisliste mittels Ranking-Algorithmen prioritätsgesteuert aufbauen, steht der Benutzer vor der Aufgabe, eine kleine Menge von für ihn interessanten Dokumenten aus einer großen Menge von uninteressanten Dokumenten herauszufinden. Er untersucht anhand von Verweisen ein Dokument nach dem anderen und nicht in jedem Fall führt ein Weg zum Ausgangspunkt zurück. Dieser Vorgang ist sehr zeitaufwendig und da es keine Möglichkeit zur sicheren Navigation in der Dokumentenmenge (Informationsraum) gibt geht der Überblick über die gefundenen Dokumente häufig verloren. Während das Ranking bei einer Suche mit nur einem Schlüsselwort noch relativ klar ist, treten bei der Suche nach Informationen mit mehreren Suchbegriffen Probleme auf. Die Nähe der gefundenen Dokumente zu den einzelnen Suchbegriffen wird aus der Ergebnisliste nicht ersichtlich. Dies ist insbesondere dann nachteilig, wenn eine unscharfe Informationssuche durchgeführt wurde.

Es existiert eine Anzahl von Prototypen für die Visualisierung von Informationsmengen sowie Forschungsprojekte auf diesem Gebiet. Sie wurden in erster Linie für die reine grafische Darstellung von Suchergebnissen entwickelt und enthalten daher oft auch entsprechend hoch entwickelte und komplexe Visualisierungsverfahren. Bei diesen Verfahren wird versucht, strukturelle Beziehungen

und Zusammenhänge aufzuzeigen, ohne eine Analyse und Interpretation des Suchergebnisses zu ermöglichen. Bisher konnte sich keiner dieser Ansätze in der Praxis durchsetzen. Gründe dafür sind u. a. Defizite beim Umgang mit großen Dokumentenmengen, komplizierte Handhabung für den Benutzer sowie die Fokussierung auf jeweils nur ein Darstellungsverfahren. Projekte, die vor allem zum Zweck der Technologieentwicklung durchgeführt wurden, sind in [1, 3–7] beschrieben. Aus den genannten Gründen steht die Forderung nach neuen Techniken zur Informationsaufbereitung und Recherche. Hier wird ein Ansatz vorgestellt, der den Benutzer bei der Darstellung und der Analyse der Ergebnismenge unterstützt, ihm Mittel zum effektiven Eingrenzen der tatsächlich interessanten Dokumentenmenge gibt und mehrere Varianten der grafischen Darstellung des Informationsraumes und Interaktionen anbietet. Dieser Ansatz wird hier in Form eines Architektur-Entwurfs beschrieben. Die Grundidee dieses Entwurfs ist ein modulares, erweiterbares System, das auf der Kombination von grafischer Darstellung, Interaktion und Filtern zur Reduktion der Ergebnismenge beruht.

2 Das Visualisierungssystem

Das Visualisierungssystem besteht aus der Suchmaschine und den Visualisierungsmodulen, die Struktur ist in der Abbildung 1 gezeigt. Die Suchmaschine oder Wissensbasis liefert als Ergebnis die Dokumentenmenge, die den angegebenen Suchbegriffen und Kriterien entspricht. Es ist wichtig, daß die Suchmaschine zu jedem Dokument die berechneten Relevanz-Werte pro Suchbegriff bereitstellt, die sie intern für das Ranking verwendet. Diese Werte bilden die Grundlage für die grafische Darstellung und einige Filter. Die Übergabe der Daten an das Visualisierungsmodul erfolgt durch direkte Ankopplung an die Suchmaschine über eine definierte XML-basierte *Schnittstelle* oder eine *Datenbasis*. Weiterhin kann über diese Schnittstelle zu jedem Zeitpunkt der aktuelle Systemzustand gespeichert und restauriert werden. Das Modul *Control* baut die notwendigen internen Datenstrukturen auf und steuert den Datenfluß zwischen den weiteren Modulen. Das Visualisierungssystem generiert eine grafische Darstellung des ermittelten Informationsraums mit den darin enthaltenen Dokumenten. Für die Erzeugung einer visuellen Repräsentation von Informationen wurde die aus der wissenschaftlich-technische Visualierungs bekannte Visualisierungspipeline adaptiert und angewendet. Dieses Vorgehen wurde in [1] veröffentlicht, daher soll hier nicht näher darauf eingegangen werden.
Die Menge der Daten für die Darstellung des Informationsraums liegt in einer unaufbereiteten, abstrakten Form ohne Topologie vor. Die Anzahl der Dokumente ist häufig so groß, daß eine sinnvolle Darstellung (besonders bei der Anwendung von dreidimensionalen Verfahren) nicht möglich ist. Das Modul *Filter* hat die Aufgabe der Reduktion der Datenmenge. Die folgenden Varianten der Datenfilterung sind im Entwurf vorgesehen, das System kann jedoch um weitere Filter erweitert werden (siehe [2]):

- Filterung nach den Namen der Dokumentautoren (nur Dokumente der ausgewählten Autoren anzeigen/verbergen),

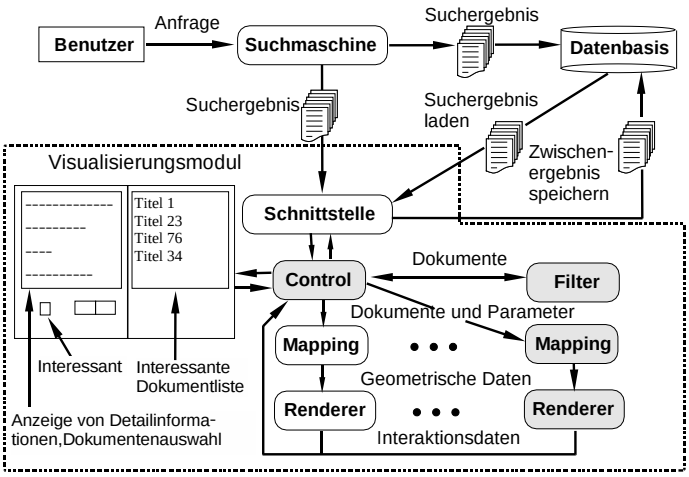


Abbildung 1. Systemarchitektur

- Filterung nach dem Erscheinungsdatum (es kann ein Zeitraum eingestellt werden),
- Filterung nach der Größe des Dokumentes,
- Filterung nach der Erscheinungsart (z. B. anzeigen/verbergen aller Dokumente, die zu einer bestimmten Kollektion gehören),
- Filterung nach Relevanzen. Fr jedes Schlüsselwort wird eine Relevanzschwelle eingestellt. Dokumente, bei denen mindestens ein Relevanzwert die eingestellt Schwelle unterschreitet, werden ausgeschlossen. Auf diese Weise ist eine sehr feine Einstellung der tatsächlichen Interessen des Benutzers möglich.
- Beseitigung von Redundanzen (mehrfach vohandene Dokumente werden von der Weiterverarbeitung ausgeschlossen.)

Durch den Modul *Mapping* werden die gefilterten Daten in einen virtuellen Informationsraum mit geometrischen Informationen abgebildet. Die Auswahl des Verfahrens zur Darstellung des Informationsraums hängt von dessen Eigenschaften (z.B. zwei- oder dreidimensional) ab. Das wiederum ist die Entscheidungsgrundlage fr die Auswahl des Renderers. Es ist z. B. sinnvoll, ein zweidimensionales Suchergebnis auch mit einem zweidimensionalen Renderer anzuzeigen. Das *Mapping-Modul* generiert eine interne Beschreibung (geometrisches Modell) des Informationsraumes 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.
Im *Renderer* hat der Benutzer die Möglichkeit, direkt mit den Dokumenten (bzw. ihren grafischen Darstellungen) zu interagieren. Dazu gehören die Navigation und die Selektion von Dokumenten. Die Resultate dieser Interaktionen werden zentral ausgewertet und die entsprechenden Reaktionen ausgelöst. Bei der Selektion eines Dokuments werden alle verfügbaren Detailinformationen angezeigt.

Der Benutzer hat die Möglichkeit, ein Dokument als „interessant“ oder „irrelevant“ einzustufen. Das entsprechende Programmelement ist es auf der linken Seite der Abbildung zu sehen.
Da das Wissen über die Zusammenhänge von Mapping-Verfahren und Renderern beim Benutzer nicht vorausgesetzt werden kann, bilden diese beiden Komponenten im vorliegenden Entwurf grundsätzlich eine Einheit. Entsprechend des zu untersuchenden Informationsraums schlägt das System passende Mapping-Verfahren vor. Der Benutzer wählt lediglich das gewünschte Abbildungsverfahren und erhält automatisch den dazugehörigen Renderer.
Auf diese Weise wird durch wiederholte Anwendung von Filtern, grafischer Darstellung und die Betrachtung detaillierter Informationen die Zahl der Dokumente eingegrenzt. Die Dokumente, die durch den Benutzer als „interessant“ klassifiziert wurden, werden in einer speziellen Datenstruktur gesammelt. Diese Liste von Dokumenten (Abbildung 1) ist das Ergebnis des hier beschriebenen Prozesses.

3 Test mit Suchergebnissen

3.1 Die Simulation von Suchergebnissen

Der Entwurf des Visualisierunssystems wurde fortlaufend durch die begleitende Implementierung eines Prototyps verifiziert. In diesen Prototyp wurde ein Simulationsmodus integriert. Somit können schnell und einfach Informationsräume mit einer beliebigen Anzahl von Dokumenten generiert werden. Die möglichen Dimensionen dieser Informationsräume liegen zwischen Eins und Vier. Die folgenden Informationen werden zu jedem Dokument mit Hilfe eines Zufallsgenerators erzeugt: Titel des Dokuments, seine URL, das Erstellungsdatum und die Größe in Bytes sowie die Relevanzen zu jedem Suchbegriff. Die Anzahl der Autoren kann vorgegeben werden, jedem Dokument wird ein Autor aus der generierten Menge zugewiesen. Ein auf diese Weise generiertes Dokument kann beispielsweise mit folgenden Informationen bestückt sein:

```
<document>
    <title>Title 3</title>
    <author>Author 12</author>
    <url>http://www.url2</url>
    <date>17-6-1992</date>
    <size>59663</size>
    <relevances>
        <rvalue>0.8667968</rvalue>
        <rvalue>0.49134594</rvalue>
        <rvalue>0.3243801</rvalue>
    </relevances>
</document>
```

Dieses fiktive Dokument ist Bestandteil eines dreidimensionalen Informationsraumes, wobei die XML-Notation der realen im Entwurf definierten Schnittstelle

entspricht. Die Schlüsselworte der generierten Suchergebnisse tragen die Bezeichnungen *Keyword_1 ... Keyword_n*, wobei *n* der Dimension des Informationsraumes entspricht. Um möglichst realistische Bedingungen zu simulieren, wurde der Datumsbereich auf 1990 bis 2002 eingegrenzt.

3.2 Das Auffinden relevanter Dokumente

Mit Hilfe der simulierten Informationsräume ist es möglich, die Effizienz der Verwendung einer grafischen Darstellung in Kombination mit verschiedenen Filtern zu prüfen. Dabei sind Informationsräume mit verschiedenen Dimensionen zu untersuchen.
Die Repräsentation eines eindimensionalen Informationsraumes wird hier nicht betrachtet, da sie kaum Vorteile gegenüber der Verwendung einer Textliste bringt. Daher wurde im Prototyp auch kein geeigneter Renderer für diesen Fall implementiert.
Experimentell werden jeweils 10000 Dokumente in Informationsräumen verschiedener Dimensionen simuliert. Im Prototyp sind die Schwellwertfilter auf den Wert 0.2 voreingestellt, dies ist auch gleichzeitig der einzige aktive Filter. Bei der Simulation eines zweidimensionalen Informationsraumes werden nach dem Durchlaufen der Filterstufe noch 6394 Dokumente vom System als relevant eingestuft. Nach einer Einstellung der Schwellwerte auf 0.8 existieren noch 372 relevante Dokumente. Wird die Schwelle für eines der beiden Schlüsselworte auf 0.9 erhöht, reduziert sich die Zahl der relevanten Dokumente auf 173. Liegt der Schwellwert bei beiden Suchbegriffen bei 0.9, werden nur noch 87 Dokumente als relevant klassifiziert. Diese Menge kann in der grafischen Darstellung gut übersehen und mit den weiteren zur Verfügung stehenden Filtern und Werkzeugen analysiert werden. Die Tabelle 1 zeigt eine Meßreihe, bei der der Schwellwert schrittweise von 0.5 auf 0.9 erhöht wurde. Dabei ist zu beachten, daß der Wert jeweils für *alle* vorhandenen Suchbegriffe eingestellt wurde. Eine solche Einstellung wird in der Praxis kaum vorgenommen, anhand der Zahlen der verbliebenen Objekte wird aber deutlich, wie fein die Menge der relevanten Dokumente durch gutes Dosieren der Schwellwerte eingegrenzt werden kann.

Tabelle 1. Simulation verschiedener Informationsräume

Schlüsselworte	0.5	0.6	0.7	0.8	0.9
2	2490	1534	895	372	87
3	1219	604	233	67	11
4	622	240	70	10	0
5	331	110	30	2	0

Aus diesem Experiment wird ersichtlich, daß die Stärke der Filtermechanismen genau dort zum Tragen kommt, wo im Umgang mit der Textliste die größten

Schwierigkeiten auftreten – bei der Auswertung von Ergebnissen mehrerer Suchbegriffe.
In der Praxis dürfte der Effekt noch stärker sein. Im Experiment sind durch die Verwendung von Zufallszahlen die Relevanzwerte gleichmäßig über die Dokumentenmenge verteilt. Bei realistischen Suchergebnissen steht eine große Menge von Dokumenten mit geringer Relevanz gegen eine sehr geringe Menge mit hoher Relevanz. Daher wird die Reduktion der reellen Dokumentenmenge noch effizienter sein.

4 Zusammenfassung

Dieser Ansatz bietet sowohl eine visuell erfassbare bedeutungstragende Form der grafische Darstellung von Informationen als auch ein System zur Informationsaufbereitung, das über offene Schnittstelle nach außen verfügt, erweiterbar, anpassbar und auch plattformunabhängig ist. Es ist deutlich geworden, daß die Suche nach relevanten Informationen mit einer Kombination aus grafischer Darstellung und Filtermechanismen erheblich vereinfacht und beschleunigt werden kann und dadurch sowohl die „zielgerichtete Suche " nach einem bestimmten Dokument als auch die „unscharfe Suche" – also die Informationssammlung zu einem Themenkomplex – gut unterstützt.
Der Benutzer wird in die Lage versetzt, binnen kürzester Zeit mit einigen wenigen unkomplizierten Einstellungen sein Suchergebnis auf eine Menge von Dokumenten zu reduzieren, die mit hoher Wahrscheinlichkeit seinen Erwartungen entsprechen. Diese kleine Menge kann er komfortabel mit mehreren Werkzeugen analysieren und beliebig zwischenspeichern.
Weiterhin kann dieses System an jede Suchmaschine gekoppelt werden, die in der Lage ist, die offene Schnittstelle zur Übergabe der Suchergebnisse zu bedienen. Die Ranking-Algorithmen der Suchmaschinen verwenden intern berechnete Relevanz-Werte der Dokumente zu den gegebenen Schlüsselwörtern. Eine Suchmaschine, die diese Relevanz-Werte über die Schnittstelle zur Verfügung stellt, kann als Datenlieferant für das System dienen. Aufgrund der Erweiterbarkeit kann das System perspektivisch jedem Anwender eine Visualisierung anbieten, die seinen persönlichen Präferenzen entspricht.

Literatur

1. Barberena Najarro M.: Darstellung von Suchergebnissen in 3D-Dimensionen. STJA 2000 MIK00 6, Workshop „Multimediale Informations- und Kommunikationssysteme" 09-12. Okt 2000, Kongreßzentrum Messe Erfurt AG.
2. Barberena Najarro M.: Die Aufbereitung von Suchergebnissen mittels Visualisierungstechniken 7. Workshop „Multimediale Informations- und Kommunikationssysteme", Erfurt, 12. September 2001, Tagungsband net.objectdays 2001, pp. 267-272
3. 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

4. Young P.: Three Dimensional Information Visualisation. Technischer Bericht, Computer Science Technical Report, No. 12/96, Visualisation Research Group, Departament of Computer Science University of Durham, 1996
5. Schwartz D.: Mehr Information durch Visualisierung von Daten? Konventionelle und innovative Visualisierungstechniken. Zeitschrift fr Bibliothek-Information-Technoligie-Online (B.I.T. Online); Ausgabe 4/1999.
6. Stuart K. Card, Jock D. Mackinlay, Ben Shneiderman, Jock D. Mckinley.: Readings in Information Visualization: Using Vision To Think. Morgan Kaufmann Publishers, San Francisco, CA, ISBN 1-55860-533-9, January 1999
7. Däßler R. Informationsvisualisierung: Stand,Kritik, Perspektiven. In Methoden/Strategien der Visualisierung in Medien, Wissenschaft und Kunst Wissenschaftlicher Verlag Trier (WVT), 1999