Der MLHTBrowser: Interaktive Exploration von Kollektionen auf verschiedenen Abstraktionsebenen

Michael Chojnacki[*], Gudrun Fischer[†]
Universität Duisburg–Essen

1. März 2005

Interaktive Exploration unbekannter Dokumentenkollektionen

Angesichts von ad hoc nicht mehr überschaubaren und sich ändernden Dokumentenmengen, z.B. Kollektionen aus dem Deep Web, benötigen Anwender unterstützende, hochinteraktive Werkzeuge, um sich dennoch schnell einen Überblick über eine solche Kollektion zu verschaffen. Als Prinzip eignet sich dafür Scatter/-Gather–Browsing ([4]). Im Vergleich zu anderen Cluster–basierten Browsing–Verfahren kann der Anwender hier in jeder Gather–Phase die zu clusternde Teilmenge selbst zusammenstellen, anstatt sich für genau einen Ast einer Cluster–Hierarchie entscheiden zu müssen. Das so entstehende Clustering richtet sich dadurch unmittelbar und flexibel nach dem Informationsbedürfnis des Anwenders.

In der vorliegenden Arbeit wird ein Modell für Browsing auf verschiedenen Abstraktionsebenen einer Dokumentenkollektion (Multi–Level–Hypertext, [1], [5]) vorgestellt. Für die Umsetzung dieses Modells wird das Scatter/Gather–Prinzip auf semistrukturierte Daten, wie sie z.B. in Deep–Web–Dokumenten vorkommen, erweitert und um mehrere Aspekte ergänzt. So wird der Scatter–Schritt um Möglichkeiten zum Ebenen-, Teilansicht- und Anordnungswechsel erweitert, und ein gezielteres Suchen wird durch Berrypicking–Funktionalität ([2]) und gleichzeitige dynamische Relevanzabschätzung unterstützt. Der entstandene Prototyp vereint erstmals alle hier genannten Aspekte in einem homogenen Werkzeug, dem MLHTBrowser.

Im Folgenden werden die verschiedenen Aspekte und ihre Umsetzung im MLHTBrowser kurz beschrieben.

Abstraktionsebenen: Multi–Level–Hypertext

Wenn zu Dokumenten zusätzlich beschreibende Daten (Metadaten) vorhanden sind, lässt sich die Kollektion als Multi–Level–Hypertext auf verschiedenen Abstraktionsebenen betrachten: Neben der Dokumentenebene gibt es die

[*]ThyssenKrupp Stahl AG, michael.chojnacki@tks-cs.thyssenkrupp.com
[†]Universität Duisburg–Essen, fischer@is.informatik.uni-duisburg.de

Ebene der Metadaten, auf der sich der Anwender mit kurzen Charakterisierungen anstelle der eigentlichen Dokumente befasst. Geht es eher um die Existenz und Verteilung von Eigenschaften in der Kollektion (Autoren, Erscheinungszeiträume), so bietet sich als weitere Abstraktion die Ebene der Attributwerte an. Etliche bestehende Portale bieten bereits ein Browsing zwischen einzelnen Aspekten dieser Ebenen. Beispielsweise kann man im ACM–Portal[1] von Autoren (Attributwertebene) zu Metadaten und zurück gelangen. Der MLHTBrowser verwirklicht dagegen den allgemeineren Fall, ein Browsing auf potentiell allen Objekten aller Abstraktionsebenen des Multi–Level–Hypertext–Modells.

Semistrukturierte Daten:
Aspekte, Datentypen und Anordnungsmöglichkeiten

Semistrukturierte Daten können auf verschiedenen Ebenen eines Multi–Level–Hypertextes auftreten: Volltexte können in Kapitel, Abschnitte und weitere Einheiten aufgeteilt sein, Metadaten liegen oft in einem semistrukturierten Schema vor, und auch bei Attributwerten sind semistrukturierte Werte denkbar. Sind solche Daten innerhalb der Kollektion zumindest strukturell homogen, so lassen sich ausgewählte Pfade auch als Felder mit Werten betrachten, die jeweils eigene Datentypen haben können (Text, Jahreszahlen usw.). Während des Browsings kann es nun sinnvoll sein, auf einer Ebene nur noch den Inhalt eines Feldes zu betrachten, sich also auf einen speziellen Aspekt der Ebene zu konzentrieren. Je nach Datentyp eines Feldes bietet der MLHTBrowser dabei verschiedene Anordnungsmöglichkeiten und ein dem Datentyp angepasstes Clustering an. So kann sich der Anwender die Ebene der Metadaten beispielsweise nach ähnlichen Kommentaren clustern oder nach Titeln alphabetisch anordnen lassen. Auf der Ebene der Attributwerte können beispielsweise Erscheinungsjahre in Intervalle aufgeteilt oder Autoren nach Ähnlichkeit gruppiert werden.

Insgesamt kann der Anwender im vorgestellten Browsing–Werkzeug zwischen den verschiedenen Anordnungen, Aspekten (betrachteten Feldern, Teilansichten) und Abstraktionsebenen frei wechseln. Dazu wurde gegenüber dem klassischen Scatter/Gather der Scatter–Schritt um den Ebenen- und Aspektwechsel, sowie um verschiedene Anordnungsmöglichkeiten für die jeweils ausgewählte Teilmenge erweitert.

Dynamische Relevanzschätzung mit Berrypicking

Um gezielter interessante Teilmengen zu identifizieren, kann der Anwender zu jedem Zeitpunkt des Browsings interessante Objekte (Dokumente, Wörter, Werte) markieren und in einen Korb legen. Alle Objekte und Gruppierungen von Objekten in jeder Darstellung werden daraufhin automatisch anhand ihrer Ähnlichkeit zum Korbinhalt beurteilt und in verschiedenen Abstufungen von Grün bis Rot mit ihrer geschätzten Relevanz gekennzeichnet. So erhält der Anwender einen zusätzlichen Anhaltspunkt für die weitere Navigation.

Clusteralgorithmen und Ähnlichkeitsmaße

Für das datentypspezifische Clustering wurden im MLHTBrowser bislang verschiedene Varianten von Buckshot (Variante von K–Means) getestet, wobei sich

[1] `http://www.acm.org/portal`

das ursprüngliche Buckshot aus [4] für die Testkollektionen bewährte. Textfelder wurden dabei als Term–Gewicht–Vektoren repräsentiert, Terme wiederum als Dokument–Gewicht–Vektoren und jeweils über das Kosinusmaß verglichen. Für spezielle Typen von Attributwerten wie Autoren und Verlage, deren Einzelwerte selten oder nie gemeinsam in Dokumenten zu finden waren, wurde die Ähnlichkeit dagegen über die Gesamtähnlichkeit der Dokumente definiert, in denen sie auftraten. Die einzelnen Experimente sind in [3] beschrieben, weitere Vergleiche mit anderen Clusteringalgorithmen, Ähnlichkeitsmaßen und auf weiteren Testkollektionen sind geplant.

Anwendungen

Der Prototyp wurde bislang als eigenständiges Werkzeug für HTML– und XML–Dokumentenkollektionen, sowie als Komponente im föderierten digitalen Bibliothekssystem DAFFODIL eingesetzt. Abbildung 1 zeigt die GUI beim Browsing in einer CompuScience[2]–Teilkollektion mit den Abstraktionsebenen der Metadaten und Attributwerte. Weitere Clusteringverfahren und neue Datentypen lassen sich leicht integrieren (Framework), und in einer laufenden Arbeit ist eine Verwendung als Browsing–Werkzeug in Peer–to–Peer–Netzen geplant. Aufgrund der hohen Konfigurierbarkeit bietet sich der MLHTBrowser außerdem als Oberfläche für Benutzerexperimente zu Clustering– und Präsentationsverfahren für semistrukturierte Daten an.

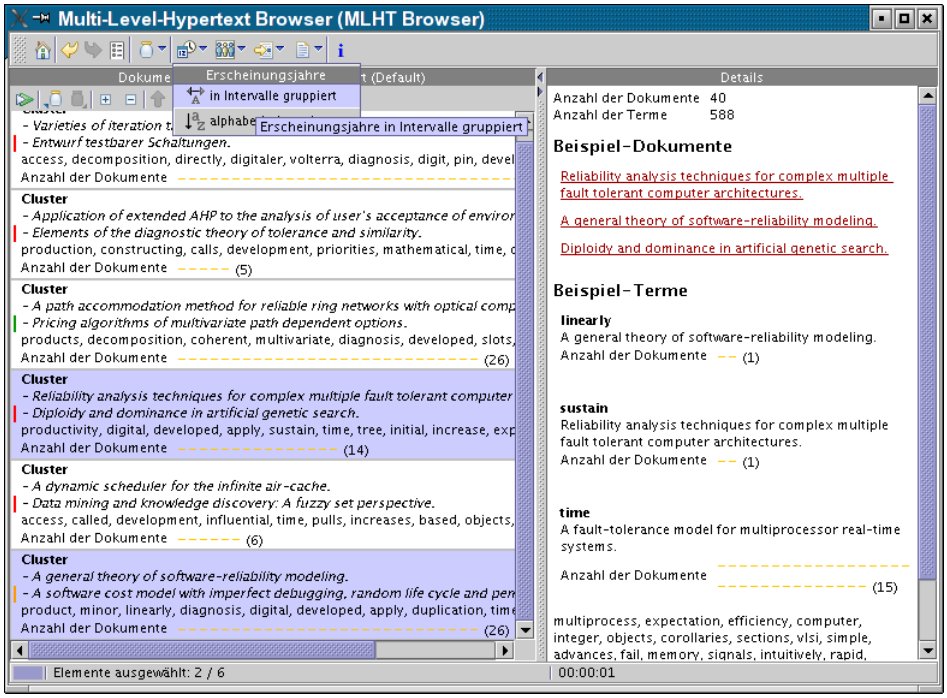

Figure 1: MLHTBrowser: GUI

[2]http://www.zblmath.fiz-karlsruhe.de

Literatur

[1] M. Agosti, R. Colotti, and G. Gradenigo. A two-level hypertext retrieval model for legal data. In *Proceedings of the Fourteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 316–325, New York, 1991. ACM.

[2] M. J. Bates. The design of browsing and berrypicking techniques for the online search interface. *Online Review*, 13(5):407–424, 1989. `http://www.gseis.ucla.edu/faculty/bates/berrypicking.html`.

[3] M. Chojnacki. Browsing in Multi-Level-Hypertext. Master's thesis, Universität Duisburg-Essen, 2004.

[4] D. R. Cutting, J. O. Pedersen, D. Karger, and J. W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 318–329, New York, 1992. ACM. `http://citeseer.nj.nec.com/cutting92scattergather.html`.

[5] N. Fuhr. Information Retrieval in Digitalen Bibliotheken. In *21. DGI-Online-Tagung – Aufbruch ins Wissensmanagement.*, Frankfurt, 1999. DGI.