In: Bekavac, Bernard; Herget, Josef; Rittberger, Marc (Hg.): Informationen zwischen Kultur und Marktwirtschaft. Proceedings des 9. InternationalenSymposiums fur Informationswissenschaft (ISI 2004), Chur, 6.-8.Oktober2004. Konstanz: UVK Verlagsgesellschaft mbH, 2004. S. 317 – 336

Dokumentbezogenes Wissensmanagement in dynamischen Arbeitsgruppen

Text Mining, Clustering und Visualisierung

Tino Schmidt [1], *Christian Wolff* [2]

[1]TU Chemnitz
Fakultät für Informatik
D-09107 Chemnitz
tino.schmidt@informatik.tu-chemnitz.de

[2]Universität Regensburg
Institut für Medien-, Informations- und Kulturwissenschaft
D-93040 Regensburg
christian.wolff@sprachlit.uni-regensburg.de

Zusammenfassung

Der Aufsatz stellt den Prototyp eines Visualisierungssystems für von Arbeitsgruppen gemeinsam genutzten Dokumentmengen vor. In einer Analyse typischer Informationsbedürfnisse und Erschließungsstrategien in kleinen und mittleren Arbeitsgruppen werden Defizite des derzeitigen Umgangs mit Dokumentmengen identifiziert. Aufbauend auf einer Zusammenschau aktueller Ansätze der Dokumentvisualisierung wird ein System entwickelt, das Text Mining- und Clusteringverfahren für die Analyse von Dokumentmengen nutzt und eine interaktive Dokumentlandkarte für die Navigation des Dokumentbestands generiert. Die Interaktionsmöglichkeiten mit dieser Dokumentlandkarte werden vorgestellt. Abschließend werden weitere Entwicklungsmöglichkeiten und Ansätze für die Evaluierung dieses Ansatzes aufgezeigt.

Abstract

In this paper, we present the research prototype of a visualization system for document sets. Starting with an analysis of typical information needs in small to medium-sized work groups, disadvantages of the current strategies for exploring document sets are identified. Based on a review of recent work in information and document visualization we present the architecture of a document visualization system which employs text mining and clustering algorithms for the analysis and structuring of document sets and makes use of the *document map* visualization metaphor for its interactive browsing

interface. Different possibilities of navigating through the document space are given.

1 Einleitung

Ausgehend von einer empirischen Untersuchung zum Dokumentenbestand sowie den Informationserschließungsstrategien wie sie für dynamische Arbeitsgruppen in einem akademischen Umfeld typisch sind, stellt der Beitrag den Prototyp eines Clustering- und Visualisierungswerkzeuges als Zugangsweg zu umfangreichen Dokumentbeständen vor. Dabei werden zunächst relevante Klassen von Informationsbedürfnissen identifiziert, für die ein solches Werkzeug einsetzbar erscheint. Für die Auswahl eines geeigneten Visualisierungsverfahrens werden die aus der Forschung bekannten Ansätze zu Dokumentenlandkarten untersucht und verwertet. Den funktionalen Kern des Systems bildet eine Text-Mining- und Clustering-Komponente, die dokumentbezogen relevante Beschreibungsterme bestimmt und mit Hilfe des TF*IDF-Maßes gewichtet. Zur Erzeugung geeigneter Clustergruppen wird eine SVD-Analyse in Kombination mit einem hierarchischen Clusterverfahren durchgeführt. Für die Schnittstelle zwischen Dokumentanalyse und Visualisierung findet ein für diesen Zweck definiertes XML-Schema Anwendung. In der Visualisierungskomponente werden die Clusteringergebnisse als mehrstufige Dokumentenlandkarte visualisiert. Dem Benutzer eröffnen sich Zugriffsmöglichkeiten auf den geclusterten Dokumentenbestand sowohl über die Visualisierung selbst, wobei Cluster durch zentrale Begriffe gekennzeichnet und zusätzlich nach einem einfachen Farbmuster klassifiziert sind, als auch über die textuelle Ausgabe der die verschiedenen Cluster charakterisierenden Terme. Auf der untersten Ebene ist, falls sich ein Cluster nicht weiter in Teilcluster aufgliedern lässt, der direkte Zugriff auf die den Cluster konstituierenden Dokumente möglich. Der Visualisierungsprototyp ist als Flash-Anwendung webbasiert lauffähig.

2 Wissensmanagement durch Knowledge Mapping

Mit dem Begriff des Knowledge Mappings bzw. der Wissenslandkarten verbindet sich die Idee, „[…] die relevante Handlungsumgebung einer Person graphisch darzustellen, um die Orientierung und Handlungsmöglichkeit in dieser Umgebung zu verbessern“ (vgl. [Eppler 02:38]) und somit dazu beizutragen, Wissensmanagement

- auf verschiedenen Ebenen (Individuum, Team, Organisation) bzw.
- in verschiedenen Lernphasen (Identifikation, Generierung von Wissen)

zu unterstützen. Durch die Visualisierung des Wissens kann die Nutzung der Karten in Organisationen die interne Wissenstransparenz erheblich verbessern und ermöglicht den systematischen Zugriff auf die Wissensbasis einer Organisation [Haun 01:106].

Die in diesem Zusammenhang beschriebenen Wissensbestandskarten [Eppler 02, Eppler 04] und Dokumentenlandkarten [Haun 01] bilden Werkzeuge des Wissensmanagements, die das Wissen einer Organisation, welches sich zum Großteil „ […] in Dokumenten codiert, in Handbüchern verankert, in Datenbanken gespeichert […], aber auch im Besitz der Individuen“ [Haun 01:109] befindet, und die Beziehungen zwischen einzelnen Wissensressourcen in visualisierter Form den Mitarbeitern einer Organisation zur Verfügung stellen. In diesem Zusammenhang versteht man die Karten als eine Metaebene der Wissensbasis einer Organisation, die den Zugriffspfad auf Wissensressourcen abbildet, selbst aber keine Wissensressourcen enthält.

Neben der rein statischen Realisierung einer Wissenslandkarte haben sich auch dynamische Ansätze entwickelt, bei denen die Wissenslandkarten neben dem *Anwendungshintergrund* des Wissensmanagements eine Schlüsselfunktion in der Benutzerschnittstelle einnehmen. Durch die Kombination von Interaktions- und Visualisierungselementen wie *Overview, Details on Demand, Zoom/Pan* oder *Dynamic Queries* [Shneiderman 98, Haun 01] mit Anforderungen des Wissensmanagements wie der Ermittlung von Schlagworten und dem Komplettieren und Bewerten von Inhalten und der Verknüpfung von Konzepten der Informationsvisualisierung entstehen Systeme, die dem Anwender einerseits einen *Überblick* über die verfügbaren Wissensressourcen vermitteln, aber auch die schnellere *Identifizierung* benötigter Informationen unabhängig von speziellen Kenntnissen und ungeachtet bestehender Organisationsstrukturen ermöglichen.

3 Untersuchungen zum Anwendungsgebiet

Die Entwicklung eines Systems zur Visualisierung von Dokumentenbeständen wurde mit dem Ziel verfolgt, Mitarbeiter in kleinen bis mittleren Arbeitsgruppen im Umgang mit umfangreichen Dokumentbeständen zu unterstützen. Aus einer Analyse an einem Lehrstuhl der Fakultät für Wirtschaftswissenschaften an der TU Chemnitz mit insgesamt etwa 20 studentischen und wissenschaftlichen Mitarbeitern ging hervor, dass Mitarbeiter neben ihren eigenen Arbeitplatzrechnern Netzlaufwerke zum Austausch oder zur Archivierung unterschiedlichster Informationen nutzen. Die in diesem Zusammenhang erstellten Dokumente sind durch die

Mitarbeiter in aktuelle Arbeitsprozesse eingebunden und werden themen- oder projektbezogen abgelegt. Endet der Bezug (z. B. durch den Abschluss eines Projektes), werden die Dokumente entweder auf den Netzlaufwerken archiviert (gefiltert und sortiert) oder verbleiben in der bisher definierten und gewachsenen Ordnungsstruktur der gemeinsam genutzten Speicherbereiche.

3.1 Datenbestände

Im konkreten Fall ergab die Untersuchung der Datenbestände, dass mit einem Anteil von 61% (ca. 10.000 von insg. 16.600 Dateien) vor allem *Textdokumente* abgelegt wurden. Eine differenzierte Auswertung ergab weiter, dass sich innerhalb der Menge der Textdokumente vor allem Dateien des Typs *MS Word* (82%), *MS Powerpoint* (7%), *Adobe Acrobat* (5%) und *HTML* (4%) befanden. Interviews mit Mitarbeitern bestätigten die Vermutung, dass ein inhaltlicher Überblick über den über mehrere Jahre gewachsenen Dokumentenbestand nicht bzw. nur sehr schwer generiert werden kann. Vor allem neu angestellte Mitarbeiter konnten die abgelegten Information ihrer Vorgänger nicht bzw. nur im geringen Maß für die Generierung eigenen Wissens und die beschleunigte Abwicklung wiederkehrender Arbeitsprozesse nutzen, da ihnen wichtige Informationen in Form von Dokumenten verborgen blieben. Auch die Verwendung der Suchfunktionen des Betriebssystems verbesserte die Nutzung des Dokumentenbestandes nicht, da die Anwender in erster Linie den zu durchsuchenden Bereich auf einen Teilbaum des Dateisystems eingrenzten und die Suche auf der Basis des Dateinamens durchführten. Im Gegensatz dazu findet eine inhaltliche Suche im Volltext der Dokumente seltener Verwendung (vgl. Tab. 1).

	Rangliste der Suchoption	Punkte / (Anzahl an Ranglisten-plätzen)	
		PC am Arbeitsplatz	Netzlaufwerk
1.	in Unterordnern	22 (6)	21 (6)
2.	nach Dateinamen	13 (4)	14 (4)
3.	nach Inhalten	9 (3)	9 (3)
4.	nach Datum, Zeitintervall	8 (4)	8 (4)
5.	nach Dateiendung	6 (2)	6 (2)

Tabelle 1: Nutzung der Suchoptionen im Vergleich

3.2 Relevante Erschließungsstrategien

Aus der Untersuchung geht hervor, dass sich der Dokumentenbestand der gemeinsam genutzten Speicherbereiche als Wissensbasis der

Organisationseinheit auffassen lässt, wobei der Zugriff und die gezielte Identifikation benötigter Ressourcen ohne geeignete Werkzeuge von den Anwendern als schwierig und langwierig beschrieben wird. Für die effektive Arbeit mit der Wissensbasis ist es z. B. erforderlich, aber nur schwer möglich, sich im Dokumentenbestand zu orientieren und einen Überblick zu erarbeiten. Auch die Identifizierung von Teil- und Themengebieten, wie sie für die Arbeit mit Dokumentenbeständen benötigt werden, lassen sich im Fall großer Dokumentkollektionen mit einfachen Suchfunktionen nur schwer lösen. Gleichzeitig dürfte der hohe Fluktuationsgrad der Mitarbeiter, die überwiegend nur wenige Jahre in diesem Kontext arbeiten, nicht nur für ein akademisches, sondern auch ein industrielles Umfeld typisch sein (kurzfristige Zuordnung zu Projekten, häufige Umstrukturierungen).

Betrachtet man in diesem Zusammenhang die von [Bates 86, Bates 02] beschriebenen Suchstrategien *Browsing*, *Searching* und *Monitoring* stellt man fest, das die Anwendung der Dokumentenlandkarten vor allem erfolgreich eingesetzt werden kann, „(...) wenn ein Suchziel unklar oder kaum exakt formulierbar ist“ [Haun 01:310]. Durch das ungerichtete Erforschen („Browsing“) des Datenbestands werden neue Erkenntnisse gewonnen und Themengebiete identifiziert, die im Anschluss daran helfen, eine konkrete Suchanfrage zu formulieren. Gerade in Projektphasen, in denen die rasche Einarbeitung in neue Themenkomplexe gefordert ist, kann die visualisierte Abbildung eines Dokumentenbestandes die Anwender dabei unterstützen, wichtige Themengebiete und relevante Dokumente schnell und gezielt zu identifizieren. Auch die Funktion eines allgemeinen Überblicks trägt dazu bei, inhaltliche Zusammenhänge aufzudecken und zu verstehen und soll so zu einer gesteigerten Effizienz im Umgang mit Dokumentenbeständen führen. So wichtig diese Eigenschaften für die Anwender auch sind, unterstützen doch nur wenige Dokumentenmanagementsysteme den Analyseprozess zur Identifikation *inhaltlicher Beziehungen zwischen einzelnen Dokumenten* [Haun 01:310] oder das ungerichtete Suchen. Der Anwendungsfall *Dokumentenbestand erkunden* umfasst in diesem Zusammenhang zwei grundlegende Varianten, zum einen die automatische Generierung einer Dokumentenlandkarte um Anwender zu unterstützen, die sich in einen für sie *unbekannten* Dokumentenbestand orientieren und zum anderen die *personalisierte Erstellung* einer Dokumentenlandkarte, um die Inhalte an die individuellen Bedürfnisse des Anwenders anzupassen.

Neben der ungerichteten Analyse des Dokumentenbestandes muss die Möglichkeit geschaffen werden, benötigte Informationen im Dokumentenbestand durch eine gerichtete Suche zu identifizieren. In diesem

Zusammenhang bietet die graphisch aufbereitete Darstellung dem Anwender eine zusätzliche Information dahingehend, dass die Suchergebnisse entsprechend ihrer inhaltlichen Beziehungen abgebildet werden. Ein Beispiel für diesen Anwendungsfall stellt das System *Lighthouse* [Leuski 01] dar, welches in der Lage ist, die Suchergebnisse von Suchmaschinen wie Inquery oder Altavista graphisch abzubilden.

Als dritter Anwendungsfall bietet sich die graphische Darstellung zeitabhängiger Informationen an. Die unter dem Begriff „Monitoring“ vorgestellte Suchstrategie wird z. Bsp. in Systemen wie *SPIRE/IN-SPIRE* (Wise et. al. 1995) oder *VxInsight* (Davidson 1998) umgesetzt und bietet dem Anwender die Möglichkeit, bezogen auf einen Dokumentenbestand eine zeitbasierte Analyse vorzunehmen und Entwicklungen und Trends zu erkennen.

4 Aufbau des Visualisierungssystems

Ausgehend von oben dargestellten Annahmen über sinnvolle Informationsstrategien wurde ein Dokumentanalyse- und -visualisierungssystem entworfen, das einerseits Text Mining- und Clustering-Verfahren für die Informationsaufbereitung einsetzt und andererseits Visualisierungstechniken aus dem Bereich des Knowledge Mappings für die Ergebnisdarstellung nutzt.

4.1 Systemarchitektur

Der Aufbau des Systems ist nach dem three-tier-Modell [Wolff 04:171] in die Schichten *Datenverwaltung*, *Anwendungskern* und *Benutzerschnittstelle* aufgeteilt. Abbildung 1 zeigt schematisch diesen Systemaufbau. Da es sich bei der Applikation um ein System zur Visualisierung von Informationen handelt, wurde die von [Card et al. 99] in ihrem *Reference Model for Information Visualization* beschriebene Aufteilung erforderlicher Transformationsschritte auf *data transformations*, *visual mappings* und *visual transformations* angewendet.

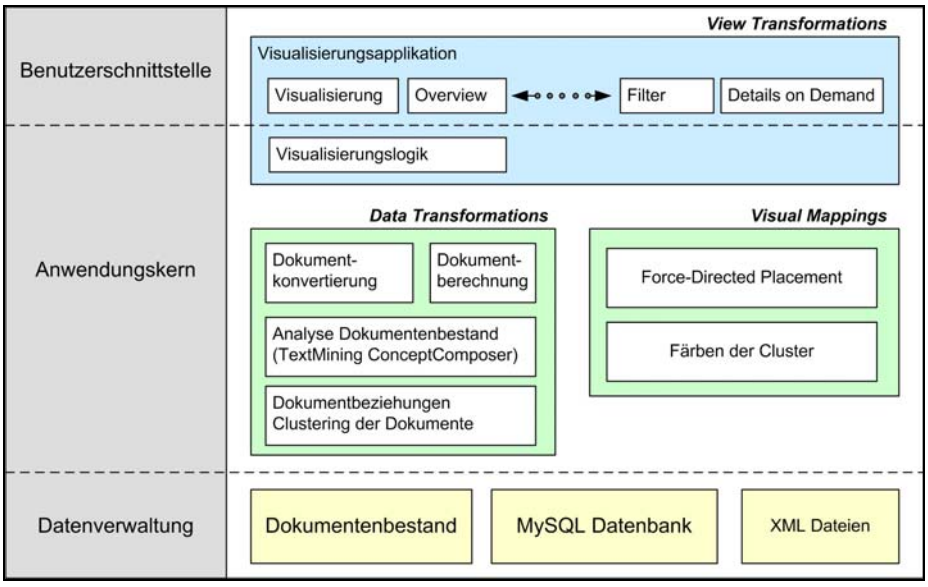

Abbildung 1: Schema der Systemarchitektur

Für die Implementierung und Erweiterung der Prozessschritte *data transformations* und *visual mappings* wurden Komponenten in Java realisiert, während die Visualisierungskomponente auf Grund der vielfältigen Möglichkeiten, interaktive Elemente zu entwickeln, als Macromedia Flash-Anwendung implementiert wurde. Die Verknüpfung der Ergebnisse aus den Transformations- und Abbildungsprozessen und der Visualisierungskomponente erfolgt durch einen XML-basierten Datenaustausch, für den ein geeignetes XML-Schema definiert wurde.

4.2 Data Transformations

Die Datentransformation als zugrunde liegender Dokumentanalyseprozess umfasst folgende Schritte:

- Dokumentkonvertierung (Extraktion von ASCII-Text aus z. T. proprietären Formaten wie PDF oder MS Word),
- Text Mining, Begriffsextraktion und Ermittlung von Dokument-Dokument-Beziehungen sowie
- das Clustern der Dokumente durch analysieren der extrahierten Terme.

4.2.1 Text Mining – Komponente

Die Analyse des Dokumentenbestandes wurde mit Hilfe der Text-Mining-Suite *Concept Composer* durchgeführt, die im Rahmen des Projekts „Deutscher Wortschatz“ entwickelt wurde [Quasthoff 98, Quasthoff & Wolff 00]. Das Ziel dieses Prozessschrittes ist es, möglichst charakterisierende Wörter in dem analysierten Corpus zu identifizieren, die als Indexwörter des Dokumentenbestandes verwendet werden können. In diesem Zusammenhang

bilden sie auch die Basis zur Berechnung der Dokumentencluster und zur Beschreibung der graphischen Symbole in der Visualisierung.

Um das Ziel zu erreichen, wird der Wortindex des Dokumentenbestandes gebildet und die Frequenz der Terme ermittelt. Diese Analyse ist der Ausgangspunkt für einen Vergleich zwischen dem untersuchten Fachcorpus und einem über einen längeren Zeitraum gewachsenen allgemeinsprachlichen Dokumentenbestand [Heyer et. al. 01]. Als Ergebnis dieses Prozesses lassen sich Terme identifizieren, deren Häufigkeit im untersuchten Dokumentenbestand stark gegenüber der Häufigkeit des allgemeinsprachlichen Corpus hervortritt. Auf der Basis dieses Verfahrens werden geeignete Fachterme ausgewählt, die als Indexwörter für die Beschreibung der Dokumente eingesetzt werden (vgl. [Faulstich et al. 02]).

4.2.2 Dokumentberechungen

Für die Bestimmung der Beziehungen zwischen Dokumenten wird auf das Verfahren des *Vector Space Model* [Salton & McGill 87] zurückgegriffen. Die in diesem Prozessschritt ermittelten TF-IDF-Gewichte der Term-Dokument Kombinationen (Termfrequenz x inverse Dokumentfrequenz [Salton & McGill 87:68]) werden berechnet und normalisiert. Damit lassen sich Dokumentencluster unabhängig von der Länge der Dokumente bilden.

4.2.3 Dokumentbeziehungen und Clustering

Die Berechnung der Clustergruppen für die verschiedenen Kartenebenen erfolgt zweistufig. Als Voraussetzung für die Anwendung eines agglomerativ hierarchischen Clusterverfahren wird auf die Term-Dokument-Matrix eine Singulärwertzerlegung (*singular value decomposition* (SVD) [Berry et al. 94]) angewendet, mit deren Hilfe die Dokument-Dokument-Beziehungen (vgl. [Boyack & Börner 03]) berechnet werden. Auf der Basis der errechneten Matrix werden Clustergruppen identifiziert, auf die das Clusterverfahren angewendet wird. Das Vorgehen gründet sich auf die Annahme, durch einen ersten Berechnungsschritt die Menge der Cluster im Vorfeld der Anwendung des hierarchischen Clusterprozesses zu begrenzen. Da der Einsatz eines partitionierenden Verfahrens eine Zielvorgabe für die Anzahl der Clustergruppen benötigt, wurde dieses Verfahren nicht angewendet. Für die Berechnung der Überblickskarte wird die Menge der Dokumentgruppen bis zu einer Ebene zusammengefasst, auf der sich anhand der Cluster die Themenbereiche des Dokumentenbestandes abzeichnen. Das Beispiel der berechneten Karte (s. u. Abbildung 3) mit 749 Dokumenten zeigt, dass die Segmentierung des Dokumentenbestandes durch rund 40 Cluster erfolgt. Die untergeordneten Dokumentenkarten der identifizierten Dokumentgruppen

verfolgen im Gegensatz zur Überblickskarte nicht das Ziel, möglichst viele Dokumente in einem Cluster zusammenzufassen, sondern die Dokumente nach Unterthemen anhand expliziter Beziehungsmerkmale zu gruppieren. Für die Berechnung dieser Karten wird die Obergrenze des Ähnlichkeitsmaßes in Abhängigkeit von der Kartenebene dynamisch verändert.

4.3 Visual Mappings

Für die Transformation der Beziehungen zwischen den Clustern in eine zweidimensionale Ebene wird das Prinzip des *Force-Directed Placement* [Fruchterman & Reingold 91] verwendet, wie es unter anderem in dem Visualisierungssystem *VxInsight* (s.u. Kap. 5.1) zum Einsatz kommt. Das Verfahren beruht auf der Idee, durch mehrere Iterationen die einzelnen Elemente neu zu platzieren und dabei die Energie des Elements in Abhängigkeit seiner Beziehungen zu anderen Clusterelementen zu berücksichtigen. Verringert sich die Energie des Elementes, wird es verschoben, ansonsten wird die Position nicht verändert. Die in Abbildung 2 wiedergegebene Formel bildet die Basis zur Berechnung der Energie eines Elements im Punkt $K_{(x,y)}$:

$$K_{i(x,y)} = \left[\sum_{j=1}^{n_i} (w_{i,j} \times l_{i,j}^2 \right] + D_{x,y} \text{ mit}$$

$K_{i(x,y)}$	Energie eines Knotens an der Position (x, y)
n_i	Anzahl der Kanten, die mit Knoten i verbunden sind
$w_{i,j}$	Kantengewicht zwischen Knoten i und dem Knoten, der mit i durch die Kante j verbunden ist
$I_{i,j}^2$	Quadrierte Distanz zwischen Knoten i und dem Knoten am anderen Ende der Kante j
$D_{x,y}$	Eine Kraft, die zur Dichte der Knoten in der Nähe der Position (x, y) proportional ist

Abbildung 2: Energiefunktion für Knoten K_i nach [Davidson et. al. 01:26]

Da im vorliegenden System – anders als in *VxInsight* – nicht jedes Element einzeln, sondern nur Clustergruppen abgebildet werden, musste die Berechung des Summanden $D_{x,y}$ an die geänderte Situation angepasst werden. Der in die Formel einbezogene Summand stellt einen Wert dar, der die Gesamtenergie des Clusters in Abhängigkeit der Dichte um den gewählten Punkt beeinflusst. Für die Berechnung des Wertes $D_{x,y}$ werden die Radien der Clusterobjekte im Umkreis des gewählten Punktes addiert und durch die Anzahl der Clusterobjekte dividiert. Clusterelemente, die auf Grund ihrer Struktur mit keinem anderen Cluster in Beziehung stehen, werden am Rand der Visualisierung angeordnet.

In der Visualisierung werden verschiedene Ebenen der Clusterstrukturen abgebildet. Da der Prototyp keine Funktionalität enthält, die eine interaktive Berechnung der Clusterebenen erlaubt, werden die Clusterebenen in dieser

Prozessphase berechnet. Dazu wurde ein Verfahren auf Basis der Ähnlichkeitsberechnung zwischen den Clustern implementiert, welches auf der obersten Clusterebene analog einem partitionierenden Clusterverfahren auf die Herausbildung abgeschlossener Cluster ausgelegt ist und daher Themengebiete im Dokumentenbestand identifiziert. Navigiert der Anwender in der Clusterhierarchie abwärts, treten die Beziehungen und inhaltlichen Zusammenhänge der Cluster in den Vordergrund.

5 Benutzerschnittstelle und Erschließungsmöglichkeiten

Bevor auf die Benutzerschnittstelle des Systemprototyps eingegangen wird, sollen zunächst verwandte Visualisierungsansätze in der Forschung knapp dargestellt werden.

5.1 Verwandte Systeme zur Visualisierung von Dokumentenbeständen

Für ähnlich gelagerte Informationsbedürfnisse wurde bereits eine Reihe von Visualisierungssystemen entwickelt; einen Überblick zu einschlägigen Systemen gibt [Schmidt 04:67ff]. Die Systeme verwenden dabei unterschiedliche Konzepte wie

- visuelle Metaphern, insbesondere Dokumentenlandkarten,
- den Self-organizing Map-Algorithmus [Kohonen et. al. 00] oder die
- Visualisierung von Graphstrukturen.

Zur Gruppe der auf (Landschafts-)Metaphern aufsetzenden Systeme zählen die Entwicklungen SPIRE/IN-SPIRE [Wise et. al. 95], VxInsight [Davidson 98, Boyack & Börner 03] oder die VisIsLands-Komponente im System xFind [Andrew 01]. Zu den Vertretern des Self-organizing Map-Algorithmus gehören Entwicklungen wie WebSom [Kohonen et al. 00], DocMiner [Becks 01] oder die Kombination mit FishEyeViews [Yang 99].

Visualisierungssysteme, die ausschließlich eine dreidimensionale Abbildung erlauben, können auf Grund der verschiedenen graphischen Gestaltungselemente nur schwer einer der Gruppen zugeordnet werden. Diese Problematik zeigt sich auch in den unterschiedlichen Klassifikationsansätzen für Visualisierungssysteme, die unter anderem von [Shneiderman 96], [Chi 00] und [Mann 02] vorgelegt wurden.

Das hier vorgestellte System setzt vor allem Technologien und Konzepte ein, wie sie im Fall der metaphernbasierten Visualisierungssysteme SPIRE und VxInsight genutzt werden. Technologische Kernpunkte des Bereiches sind die

Anwendung des Vector Space Models, der Einsatz von Clusteralgorithmen und die Verwendung von Dimensionsreduktionsverfahren (Singular Value Decomposition (SVD), Force-Directed Placement).

5.2 Struktur des Visualisierungssystems im Überblick

Die Visualisierungskomponente nutzt die durch den Analyseprozess erzeugte und in XML repräsentierte Karteninformation. Im ersten Schritt werden die Informationen der XML-Datei ausgewertet und der Visualisierungsapplikation zur Verfügung gestellt. Dem Anwender präsentiert die Applikation (Abbildung 3) eine Visualisierungsebene ([1] in Abbildung 3), in der die Cluster entsprechend ihrer Beziehungen zueinander gezeigt werden. Die Größe der Cluster gibt Aufschluss darüber, wie viele Dokumente in einem Cluster enthalten sind. Neben der Visualisierungsebene existiert im rechten Bereich eine der Größe nach absteigend sortierte Liste der Cluster ([2] in Abbildung 3). Für jeden identifizierten Cluster werden die beschreibenden Terme angezeigt, aber auch Interaktionssymbole, um weitere Aktionen in Verbindung mit dem Cluster durchzuführen zu können. Die unterschiedliche Färbung der Elemente dieser Liste soll auf einen Blick zeigen, welche Cluster miteinander in Beziehung stehen. Eine Overview-Komponente ([3] in Abbildung 3) zeigt dem Nutzer die aktuelle Position im visualisierten Dokumentenbestand in Abhängigkeit von der gewählten Zoom-Stufe.
Für die Interaktion mit dem Dokumentenbestandes wurden weitere Komponenten realisiert, die die Funktion der *Details on Demand* oder eine Filterung des Dokumentenbestandes (s. u. Abbildung 4) erlauben. Auch die Ebene der in einem Cluster enthaltenen Dokumentelemente lässt sich über die Auswahl *Dokumente* des Clustermenüs aktivieren. Der Anwender erhält mit dieser Visualisierungskomponente die Möglichkeit, die Dokumente in einer nach Eigenschaften sortierten Listendarstellung (s. u. Abbildung 7) anzuzeigen, oder eine Kreisdarstellung (s. u. Abbildung 8) zu wählen, bei der die Dokumente proportional zur ihrer Entfernung vom Clustermittelpunkt abgebildet werden.

Die Komponente eines *Search Memory* bietet dem Anwender die Gelegenheit, die während des Suchprozesses identifizierten Dokumente unterschiedlicher Cluster in einem Zwischenspeicher abzulegen und so im weiteren Verlauf der Suchaktivitäten, analog zur Theorie des *berry-picking*-Ansatzes von [Bates 89], eine Ergebnismenge von Dokumenten zu identifizieren, die dem Informationsbedürfnis des Anwenders entspricht.

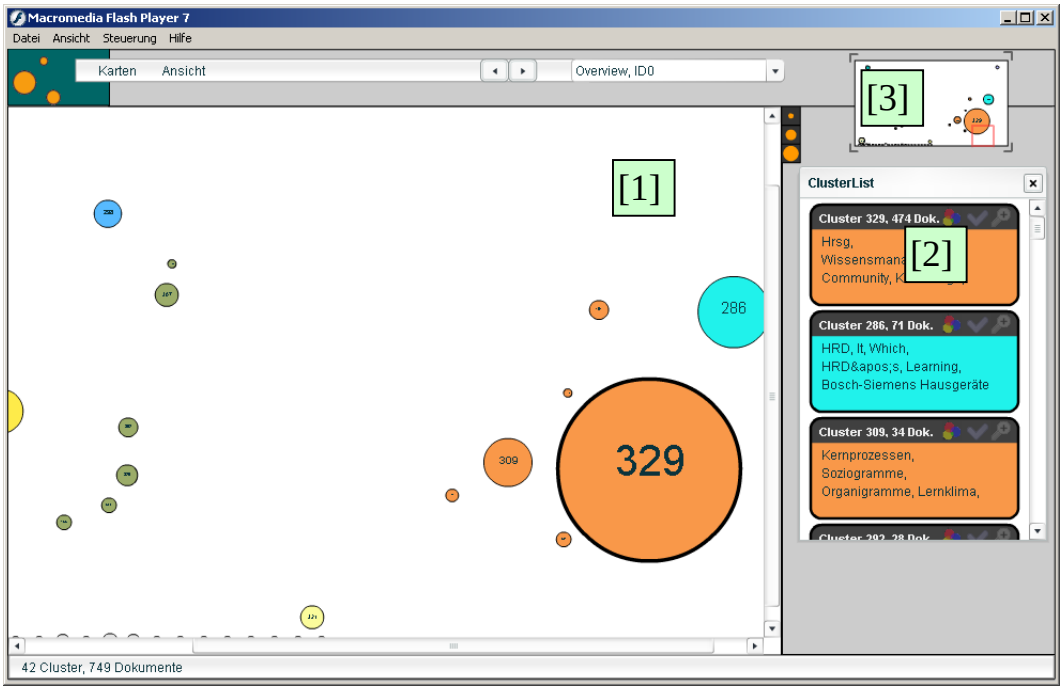

Abbildung 3: Visualisierungssystem

5.3 Anwendungsmöglichkeiten des Systems

Die Analyse des Dokumentenbestandes beginnt für den Anwender mit dem Starten der Applikation (als stand-alone-Programm oder als Webapplikation). Durch die Aktivierung des Menüpunktes *Dokumentenbestand - Überblick* wird eine Dokumentenlandkarte geladen. Startet man in diesem Punkt mit der Interpretation der abgebildeten Clusterelemente, erkennt der Anwender anhand der Größe der Cluster deren Einfluss im Dokumentenbestand.

Wie Abbildung 3 darstellt, lässt der durch die Terme *Wissensmanagement*, *Community*, *Knowledge* beschriebene *Cluster 329* auf eine wichtige Dokumentenmenge innerhalb des Dokumentenbestandes schließen. Zusätzlich werden weitere, mit dem *Cluster 329* verbundene Dokumentgruppen identifiziert, da sie in der gleichen Farbe um den Cluster angeordnet wurden, wobei ein Algorithmus eingesetzt wird, der inhaltlich verwandten Clustern auch ähnliche Farben zuordnet. Das Anklicken eines Clustersymbols führt zu einem Kontextmenü, welches verschiedene Aktionen in Abhängigkeit des Clusters ermöglicht. Über den Punkt „*Details*“ erhält der Anwender die Möglichkeit, sich die beschreibenden Terme des Clusters und die enthaltenen Dokumentelemente in einer Listendarstellung anzeigen zu lassen.

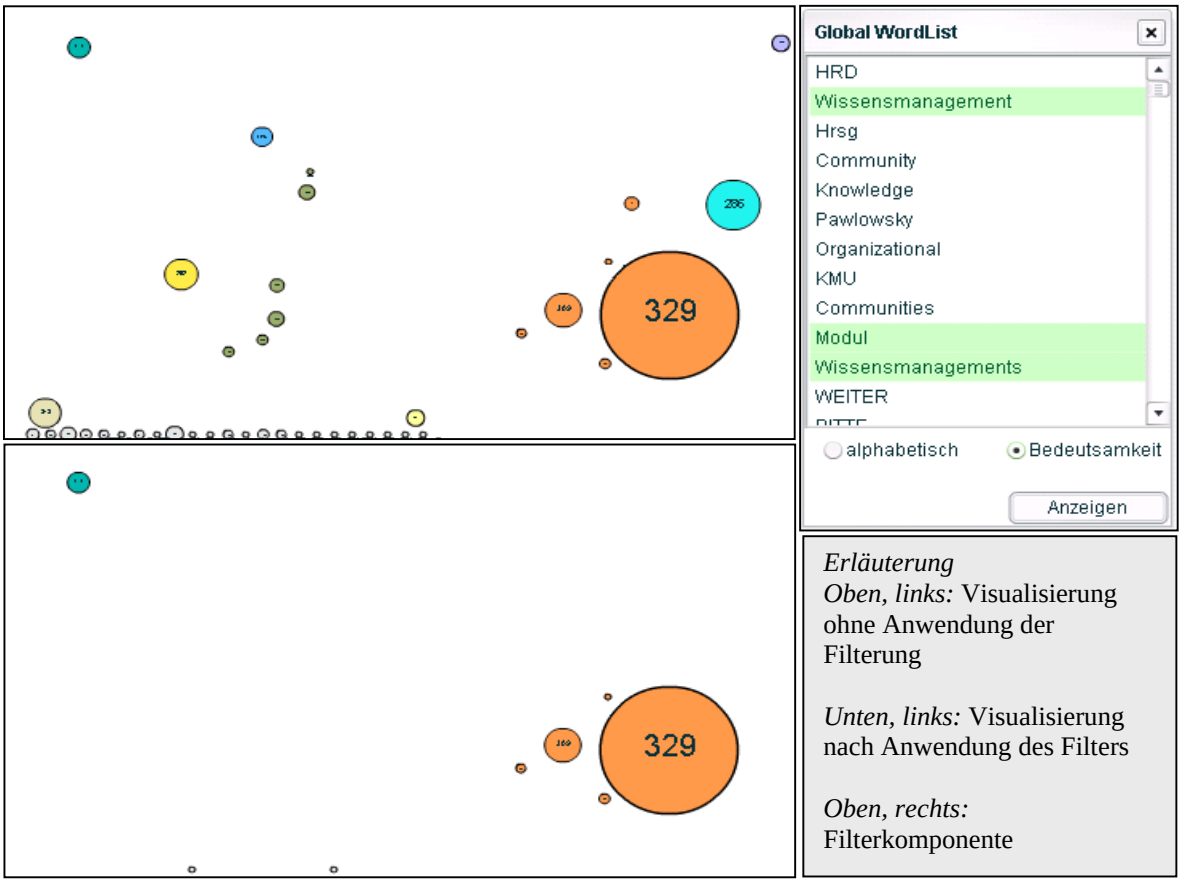

Abbildung 4: Filterkomponente

Während des Suchprozesses ist es möglich, einzelne Clustersymbole für einen später zu verfolgenden Analyseweg zu markieren und visuell hervorzuheben. Für die Entscheidung, welche Cluster für das Informationsbedürfnis und somit für eine weitere Analyse relevant sind, kann eine Filterkomponente verwendet werden. Durch die Anwendung der Filteroption (oben rechts in Abbildung 4) ist es möglich, die Visualisierung anhand einer Auswahl von Termen zu beeinflussen. Um im vorliegenden Fall den Dokumentenbestand weiter unter dem Gesichtspunkt *Wissensmanagement* zu untersuchen, wurden die Begriffe *Wissensmanagement* und *Modul* gewählt und anschließend die Filterung aktiviert (unten links in Abbildung 4). Abbildung 4 zeigt, dass unter anderem der 474 Dokumente enthaltende *Cluster 329* relevante Dokumente in Bezug zur getroffenen Auswahl beinhaltet.

Zur näheren Analyse der 474 Dokumente in *Cluster 329* ist es durch den Menüpunkt *untergeordnete Karte* des Kontextmenüs möglich, die im Cluster enthaltenen Dokumente zu strukturieren. Die in Abbildung 5 dargestellte Dokumentenkarte zeigt die dem *Cluster 329* untergeordnete Clusteringebene, die die enthaltenen Dokumente des Clusters nach demselben Visualisierungskonzept strukturiert abbildet.

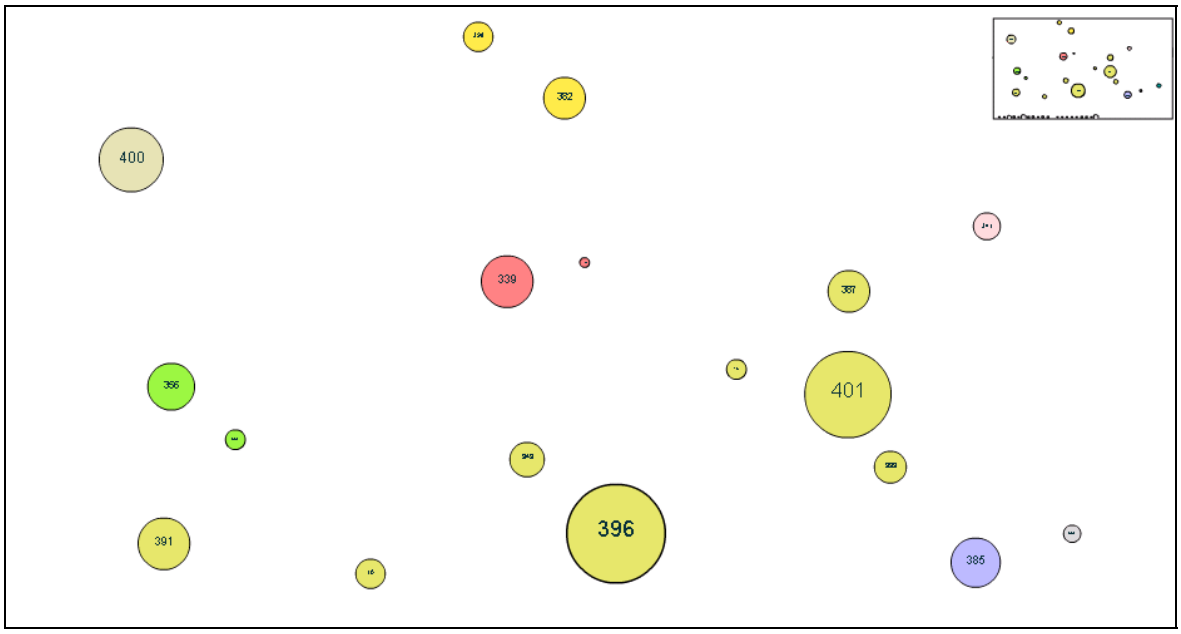

Abbildung 5: „Untergeordnete Karte“ des Clusters 329

Eine Analyse dieser Karte verdeutlicht, welche Themengebiete in diesem Teilbestand relevant sind. Abbildung 6 stellt zu diesem Zweck einen Auszug der Themenbereiche dar, die anhand der graphischen Clustersymbole und der interaktiven Komponenten identifiziert wurden. Auf der Basis der Gruppierung der Dokumente lassen sich diejenigen Gruppen finden, die dem Informationsbedürfnis des Anwenders entsprechen.

			
Gruppenarbeit, Gruppenleiter, Teammitglieder, Gruppenmitglieder, ...	Community, Communities, **Kernteam**, Lerngruppe, Modul, Practices, Knowledge ...	**Personalentwicklung**, **Personalmanagement**, Lernfähigkeit, Feedback, Organisations-entwicklung, Einflußfaktoren, Personalentwicklungs-abteilung, Führungskräfteent-wicklung, Mitarbeiterentwicklung, Führungskräfte-Entwicklung,	**Ontologien, GfWM, Taxonomie, Ontologie, Taxonomien, Wissensraum,** Repräsentieren, Wissensmanagement, Knowledge, Diskussionsraum, Identifizieren, ...
Gruppenarbeit, Gruppenführung, Gruppenmitglieder, Problemverschiebung, Feedback, Staehle, ...	Fallstudie, Mandl, **Kernteam**, Lehrstuhl, Wissenstransfer, KM-Tools, Verbundprojektes,...		

Abbildung 6: Unterthemen des *Clusters 329* (Clusterelemente aus Abbildung 5)

Durch die Aktivierung einer „Dokument-Funktion“ im Kontextmenü ist es möglich, die in jedem Cluster enthaltenen Dokumente darzustellen. Neben einer listenartigen Darstellung (*FileList*, Abbildung 7) können die Dokumente auch in Abhängigkeit ihres Abstandes vom Clustermittelpunkt visualisiert werden (*FileCircle*, Abbildung 8). Ähnliche Dokumente befinden sich in dieser Visualisierungsvariante gemeinsam auf den Kreisringen der Darstellung.

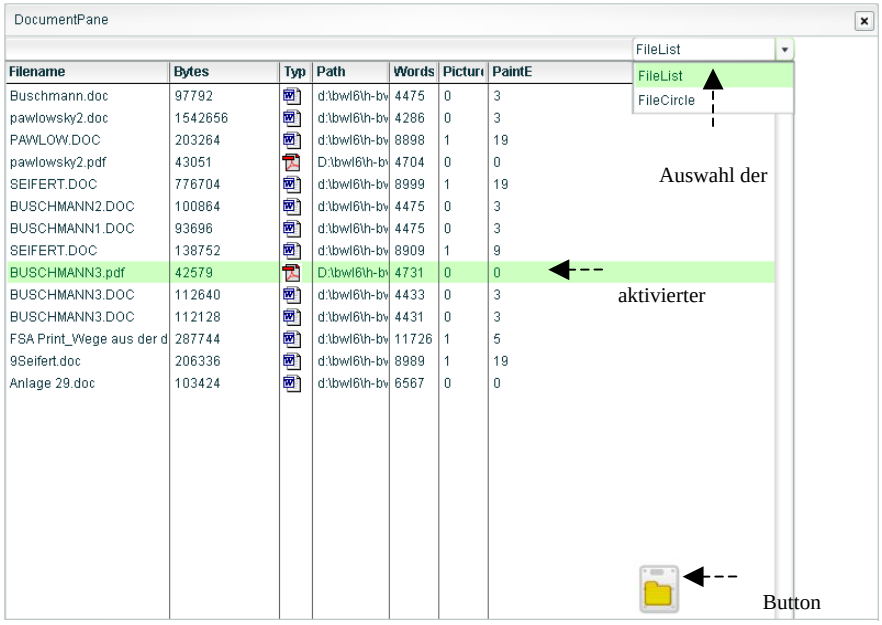

Abbildung 7: Dokumentvisualisierung *FileList*

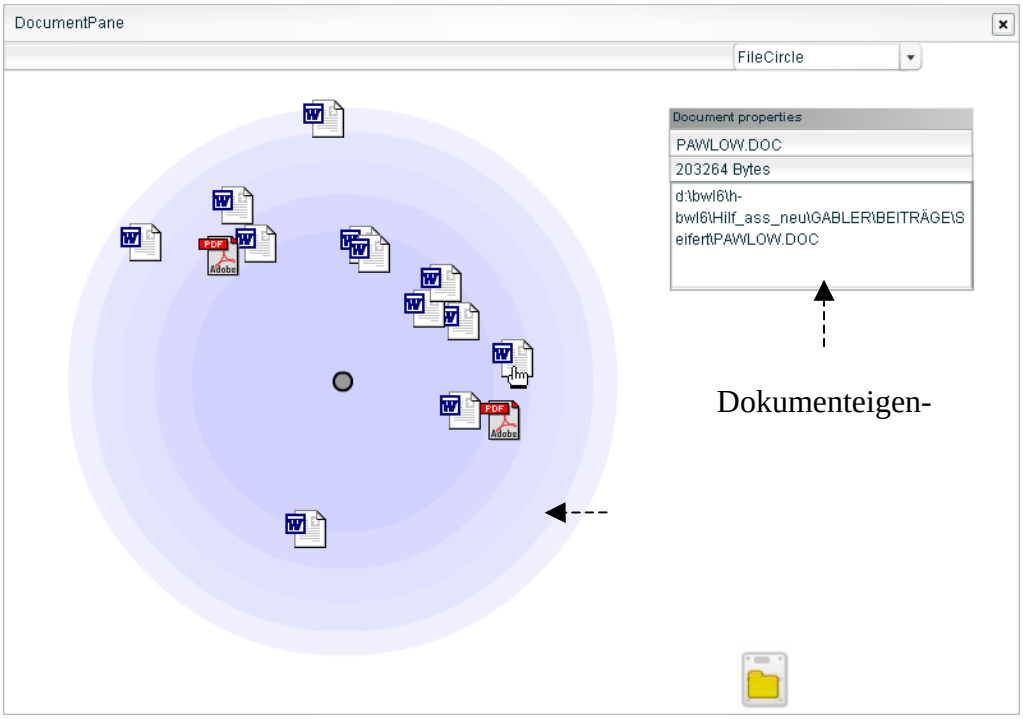

Abbildung 8: Dokumentvisualisierung *FileCircle*

Bei Überfahren (*rollover*) der Dokumentsymbole mit der Maus werden Metadaten wie Dateiname, Größe und Pfad angezeigt. Relevante Dokumente können durch *click & drag* in eine globale Liste (*SearchMemory*, Abbildung 9) eingetragen und unabhängig von weiteren Suchvorgängen jederzeit abgerufen werden.

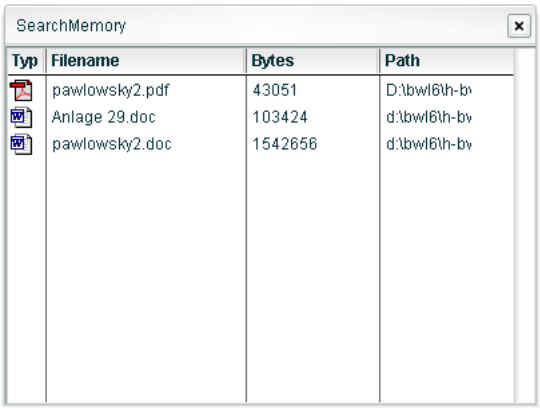

Abbildung 9: Komponente *SearchMemory*

6 Fazit und Ausblick

Mit Hilfe des hier vorgestellten Prototyps lassen sich die Beziehungen zwischen Dokumenten in eine interaktive Karte übertragen. Die Anwender sind in der Lage, wichtige Dokumentgruppen zu identifizieren und auf dieser Basis das in dem Dokumentenbestand verfügbare Wissen zu erschließen, indem sie die Visualisierung und die enthaltenen, interaktiven Komponenten nutzen.

Eine förmliche Evaluierung des Systems steht bisher noch aus, bei Vorstellung und Demonstration des Systems bei den Mitarbeitern der Anwendungsdomäne konnten aber immerhin erste Plausibilitätsvermutungen dafür gewonnen werden, dass die Visualisierung des Dokumentensbestands einer Arbeitsgruppe die Informationserschließung erleichtern könnte.

Ein wesentliches Argument ist hier sicherlich die von den durch z. T. idiosynkratisch geprägten Strukturierungsstrategien Einzelner unabhängige Visualisierung. Die Präsentation von Themenbereichen und interaktive Funktionen wie die Filterfunktion können zusätzlich zu einem effizienten Zugriff auf benötigte Informationen führen und bieten die Möglichkeit, „verborgene“ Dokumente hervorzuheben.

Für die Weiterentwicklung des Systems stellt sich die Frage, wie die Anwender die abstrakte graphische Darstellung wahrnehmen und wie diese

weiter entwickelt werden kann bzw. ob sich durch den zusätzlichen Einsatz einer graphischen *Metapher* Vorteile für die Anwendung des Systems ergeben. Bisher sind empirische Studien angesichts der Vielzahl innovativer Visualisierungssystem noch bedauerlich rar, erste Studien unterstreichen allerdings das Optimierungspotential durch den Einsatz von Visualisierungsverfahren [Reiterer et al. 03]. Entsprechende Studien zur Ergonomie von Informations- und Konzeptvisualisierungen im Bereich Information Retrieval sind in Vorbereitung.

Durch eine vergleichende empirische Studie könnte man im weiteren Entwicklungsprozess Aufschluss darüber erhalten, inwieweit sich aus Visualisierungssystemen wie dem oben vorgestellten Vorteile im Vergleich mit klassischen dateisystemorientierten Zugängen einerseits und Verfahren der gezielten Suche (Volltextindex, IR-System) andererseits ergeben.

Neben den Fragen zum User Interface wurden bei der Entwicklung des Systems auch Probleme der technischen Ebene identifiziert, die sich im Wesentlichen in drei Gruppen ordnen lassen:

1. Optimierung der zugrunde liegenden Analysealgorithmen, um eine bessere Skalierbarkeit auch für deutlich größere Dokumentmengen zu gewährleisten.
2. Darauf aufbauend Integration interaktiver Verfahren zur direkten Beeinflussung der Analyse- und Darstellungsprozesse („Online-Clustering“).
3. Verbesserung der linguistischen Analyse und des Text Mining, das im gegenwärtigen Zustand des Systems vollformenbasiert arbeitet (Ausschluss von Stoppwörtern, Grundformreduktion, vgl. o. Abbildung 6).

Zusätzlich zu technischen und gestalterischen Erweiterungen ist zu untersuchen, ob eine funktionale Ausweitung des beschriebenen Systems durch die Kombination der unterschiedlichen Suchstrategien *Browsing*, *Searching* und *Monitoring* aus Sicht der Anwender eine weitere Verbesserung ergibt.

Danksagungen

Die Autoren danken dem *Lehrstuhl für Personal und Führung* (Prof. P. Pawlowsky) der TU Chemnitz für Bereitstellung von Daten und die Unterstützung bei der empirischen Voruntersuchung und der *Abteilung*

Automatische Sprachverarbeitung (Prof. G. Heyer) am Institut für Informatik der Universität Leipzig sowie dem dortigen Projekt *Deutscher Wortschatz* (PD Dr. U. Quasthoff) für die Bereitstellung linguistischer Daten und der Text Mining-Software *Concept Composer*.

7 Literatur

[Andrews 01] Andrews, K. et. al. (2001). "Search Result Visualisation with xFIND." In: Proccedings of the 2nd Int. Workshop on User Interfaces to Data Intensive Systems (UIDIS). IEEE Computer Society Press, 50–58.

[Bates 86] Bates, M. J. (1986). "An Exploratory Paradigm for Online Information Retrieval." In: Intelligent Information Systems for the Information Society. Proceedings of the 6th int. Research Forum in Information Science (IRFIS 6), 91-99.

[Bates 89] Bates, M. J. (1989). "The design of Browsing and Berrypicking Techniques for the On-line Search Interface." In: Online Review 13(5) (1989), 407-424

[Bates 02] Bates, M. J. (2002). "Towards an Integrated Model of Information Seeking and Searching." Keynote Speech, Fourth International Conference on Information Needs, Seeking and Use in Different Contexts, Lisbon, Portugal, September 2002, *http://www.gseis.ucla.edu/ faculty/bates/articles/info_SeekSearch-i-030329.html* [Zugriff Mai 2004].

[Becks 01] Becks, A. (2001). Visual Knowledge Management with Adaptable Document Maps. GMD Research Series No. 15 / 2001. Sankt Augustin: GMD - Forschungszentrum Informationstechnik, [zug. Diss. RWTH Aachen], *http://www.bi.fraunhofer.de/ publications/research/2001/015/Text.pdf* [Zugriff Mai 2004].

[Berry et al. 94] Berry, M. W.; Dumais, S. T.; O'Brien, G. W. (1994). "Using Linear Algebra for Intelligent Information Retrieval." SIAM Review 37(4) (1995), 573-595 [zugl. Technical Report, University of Tennessee, Dept. of Computer Science, Doc. Nr. ut-cs-94-270], *http://www.cs.utk.edu/~library/TechReports/1994/ut-cs-94-270.ps.Z* [Zugriff Mai 2004].

[Boyack & Börner 03] Boyack, K. W.; Börner, K. (2003). "Indicator-assisted Evaluation and Funding of Research: Visualizing the Influence of Grants on the number and Quality of Research Papers." In: Journal of the American Society for Information Science and Technology 54(5) (2003), 447-461, *http://www.cs.sandia.gov/projects/VxInsight/ pubs/jasist03si.pdf* [Preprint, Zugriff Mai 2004].

[Card et al. 99] Card, S. K.; Mackinlay, J. D.; Shneiderman, B. (1999). Readings in Information Visualization: Using Vision To Think. San Francisco / CA: Morgan Kaufmann.

[Chi 00] Chi, E. H. (2000). "A Taxonomy of Visualization Techniques Using the Data State Reference Model." In: Proc. of the IEEE Symposium on Information Visualization 2000, Salt Lake City / UT, Ocotber 2000, 69-75.

[Davidson et al. 98] Davidson, G. S. et. al. (1998). "Knowledge Mining with VxInsight: Discovery through Interaction." In: Journal of Intelligent Information Systems, 11(3) (1998), 259-285.

[Davidson et al. 01] Davidson, G.S.; Wylie, B.N.; Boyack, K.W. (2001). „Cluster Stability and the Use of Noise in Interpretation of Clustering." In: Proc. IEEE Symposium on Information Visualization 2001 (INFOVIS'01), 23-30.

[Eppler 02] Eppler, M. J. (2002). „Wissen sichtbar machen: Erfahrungen mit Intranet-basierten Wissenskarten. Knowledge Mapping Methodik und Beispiele." In: Pawlowsky, P.; Reinhardt, R. (Hrsg.) (2002). Wissensmanagement für die Praxis – Methoden und Instrumente zur erfolgreichen Umsetzung. Neuwied, Kriftel: Luchterhand.

[Eppler 04] Eppler, M.J. (2004). „Making Knowledge Visible through Knowledge Maps: Concepts, Elements, Cases." In: Holsapple, C.W. (Ed.) (2004). Handbook on Knowledge Management. Vol. 1. Knowledge Matters. Berlin et al.: Springer, 199-206.

[Faulstich et. al 02] Faulstich, L. C.; Quasthoff, U.; Schmidt, F.; Wolff, Ch. (2002). "Concept Extractor - Ein flexibler und domänenspezifischer Web Service zur Beschlagwortung von Texten." In: Hammwöhner, R.; Wolff, Ch.; Womser-Hacker, Ch. (Hrsg.) (2002). Information und Mobilität, Proc. 8. International Symposion in Information Science, Regensburg, Oktober 2002, 165-180.

[Fruchterman & Reingold] Fruchterman, T.; Reingold, E. (1991). "Graph Drawing by Force-directed Placement." In: Software – Practice and Experience 21 (1991), 1129-1164.

[Haun 01] Haun, M. (2001). Handbuch Wissensmanagement. Grundlagen und Umsetzung, Systeme und Praxisbeispiele. Berlin et al.: Springer.

[Heyer et al. 01] Heyer, G.; Läuter, M.; Quasthoff, U.; Wolff, Ch. (2001). „Wissensextraktion durch linguistisches Postprocessing bei der Corpusanalyse." In: Lobin, H. (Ed.) (2001). Sprach- und Texttechnologie in digitalen Medien. Proc. GLDV-Jahrestagung 2001, Universität Gießen, März 2001, 71-83.

[Kohonen 00] Kohonen, T. et. al. (2000). "Self Organization of a Massive Document Collection." In: *IEEE Transactions on Neural Networks 11(3) (2000) [*Special Issue on Neural Networks for Data Mining and Knowledge Discovery*], 574-585.*

[Leuski 01] Leuski, A. (2001). Interactive Information Organization. Techniques and Evaluation. Ph.D. Thesis, University of Massachusetts, Center for Intelligent Information Retrieval, Amherst / MA, Mai 2001, *http://www-ciir.cs.umass.edu/~leouski/publications/papers/ir-232.pdf* [Zugriff Mai 2004].

[Mann 00] Mann, T. M. (2002): Visualization of Search Results from the World Wide Web. Dissertation: Universität Konstanz, *http://www.ub.uni-konstanz.de/kops/volltexte/2002/751/* [Zugriff Mai 2004].

[Quasthoff 98] Quasthoff, U. (1998). „Projekt deutscher Wortschatz." In: Heyer., G.; Wolff, Ch. (ed.) (1998). Linguistik und neue Medien. Wiesbaden: Dt. Universitätsverlag, 93-99.

[Quasthoff & Wolff 00] Quasthoff, U.; Wolff, Ch. (2000). "An Infrastructure for Corpus-Based Monolingual Dictionaries." In: Proc. Second International Conference on Language Resources and Evaluation (LREC-2000), Athens, May/June 2000, 241-246.

[Reiterer et al. 03] Reiterer, H. et al. (2003). „Ein visueller Metadaten Browser für die explorative Erkundung großer Datenmengen.“ In: Ziegler, J.; Szwillus, G. (Hrsg.) (2003). Proc. Mensch und Computer 2003 - Interaktion in Bewegung. Stuttgart et al.: B.G. Teubner, 165-176.

[Salton & McGill 87] Salton, G.; McGill, M. J. (1987). Information Retrieval - Grundlegendes für Informationswissenschaftler. Hamburg: McGraw-Hill.

[Schmidt 04] Schmidt, T. (2004). Visualisierung von Dokumentenbeständen auf Basis von Text-Mining-Verfahren. Diplomarbeit, Technische Universität Chemnitz, Fakultät für Informatik.

[Shneiderman 98] Shneiderman, B. (1998[3]): Designing the User Interface - Strategies for Effective Human-Computer Interaction. Reading / MA.: Addison-Wesley.

[Wise et al. 95] Wise J. A. et. al. (1995). “Visualizing the Non-Visual: Spatial Analysis and Interaction with Information from Text Documents.” In: Proc. IEEE Conference on Information Visualization 1995, Atlanta / GA, October 1995, 51-58.

[Wolff 04] Wolff, Ch. (2004). „Systemarchitekturen. Aufbau texttechnologischer Anwendungen.“ In: Lemnitzer, L.; Lobin, H. (Hrsg.) (2004). Texttechnologie. Perspektiven und Anwendungen. Tübingen: Stauffenburg, 165-192.