Band 20 – Heft 1 – Jahrgang 2005 – ISSN 0175-1336

Zeitschrift für Computerlinguistik und Sprachtechnologie
GLDV-Journal for Computational Linguistics and Language Technology

LDV Forum

Themenschwerpunkt

Text Mining

Herausgegeben von
Alexander Mehler und Christian Wolff

Gesellschaft für linguistische Datenverarbeitung www.gldv.org

Impressum

LDV-Forum Zeitschrift für Computerlinguistik und Sprachtechnologie
ISSN 0175-1336 GLDV-Journal for Computational Linguistics and Language
Band 20 - 2005 - Heft 1 Technology – Offizielles Organ der GLDV

Herausgeber Gesellschaft für Linguistische Datenverarbeitung e. V. (GLDV)

Juniorprofessor Dr. Alexander Mehler, Universität Bielefeld,
alexander.mehler@uni-bielefeld.de
Prof. Dr. Christian Wolff, Universität Regensburg
christian.wolff@sprachlit.uni-regensburg.de

Anschrift der Redaktion Prof. Dr. Christian Wolff,
Universität Regensburg
Institut für Medien-, Informations- und Kulturwissenschaft
D-93040 Regensburg

Wissenschaftlicher Beirat Vorstand, Beirat und Arbeitskreisleiter der GLDV
http://www.gldv.org/cms/vorstand.php,
http://www.gldv.org/cms/topics.php

Erscheinungsweise 2 Hefte im Jahr, halbjährlich zum 31. Mai und 31. Oktober. Preprints und redaktionelle Planungen sind über die Website der GLDV einsehbar (*http://www.gldv.org*).

Einreichung von Beiträgen Unaufgefordert eingesandte Fachbeiträge werden vor Veröffentlichung von mindestens zwei ReferentInnen begutachtet. Manuskripte sollten deshalb möglichst frühzeitig eingereicht werden und bei Annahme zur Veröffentlichung in jedem Fall elektronisch und zusätzlich auf Papier übermittelt werden. Die namentlich gezeichneten Beiträge geben ausschließlich die Meinung der AutorInnen wieder. Einreichungen sind an die Herausgeber zu übermitteln.

Bezugsbedingungen Für Mitglieder der GLDV ist der Bezugspreis des LDV-Forums im Jahresbeitrag mit eingeschlossen. Jahresabonnements können zum Preis von 25,- € (inkl. Versand), Einzelexemplare zum Preis von 15,- € (zzgl. Versandkosten) bei der Redaktion bestellt werden.

Satz und Druck Christoph Pfeiffer, Regensburg, mit *LaTeX* (*pdfeTeX / MiKTeX*) und *Adobe InDesign CS 3.0.1* , Druck: Druck TEAM KG, Regensburg

Editorial

Christian Wolff und Alexander Mehler

Editorial

Liebe GLDV-Mitglieder, liebe Leserinnen und Leser des LDV-Forums,

nach langer, bis vor das Jahr 2003 zurückreichender Vorbereitungszeit können wir Ihnen nun Heft 1 des 20. Jahrgangs des LDV-Forums vorlegen. In bewährter Tradition handelt es sich dabei um ein Themenheft, dessen Beiträge dem Thema *Text Mining* gewidmet sind.

Die inhaltliche Abrundung und der Abschluß dieses Heftes ist nicht zuletzt der Tatsache geschuldet, dass mit Alexander Mehler ein weiterer Herausgeber für das LDV-Forum gefunden werden konnte. Die Herausgeber hoffen, dass mit der nun erreichten Zusammenstellung von Beiträgen eine fruchtbare und interessante Darstellung eines noch jungen Arbeitsgebiets erreicht werden konnte. Den Autorinnen und Autoren – und selbstverständlich in gleicher Weise auch den Leserinnen und Lesern – sei jedenfalls für die lange Geduld bis zum Erscheinen des Hefts sehr herzlich gedankt.

Der im vergangenen Jahr angekündigte Ausbau der Website der GLDV (*http://www.gldv.org*) zu einem computerlinguistischen und sprachtechnologischen Informationsportal ist mittlerweile vollzogen. Unter der Regie von Bernhard Schröder (Bonn) konnte ein Content Management-System erfolgreich für die Belange der GLDV adaptiert werden. Der Inhalt des LDV-Forums kann sich daher wie geplant auf Fachbeiträge konzentrieren.

Da – dem Gegenstand angemessen – die Beiträge dieses Hefts über erhebliche formale Anteile verfügen, war es nicht möglich, die für das Vorgängerheft entwickelte gestalterische und publikationstechnische Lösung vollständig beizubehalten. Wir haben uns aber bemüht, beim Satz der Beiträge in TEX eine behutsame Annäherung an die bisherige Gestaltung vorzunehmen. Dabei ist erneut Herrn cand. phil. Christoph Pfeiffer großer Dank geschuldet, der wieder den Satz übernommen hat und dabei seine umfangreichen TEX-Kentnisse erfolgreich einbringen konnte.

Mit Erscheinen dieses Heftes zum 31. Mai 2005 wird der reguläre Publikationstakt des LDV-Forum endlich wieder erreicht. Ihm wird im Herbst diesen Jahres als zweites Heft des 20. Bandes ein Themenheft mit dem Schwerpunkt Corpuslinguistik folgen.

Regensburg und Bielefeld, im Mai 2005

Christian Wolff und Alexander Mehler

Matthias Dehmer

Data Mining-Konzepte und graphentheoretische Methoden zur Analyse hypertextueller Daten

1 Einleitung

Der vorliegende Artikel hat das Hauptziel, eine verständliche Übersicht bezüglich der Einsetzbarkeit von *Data Mining*-Konzepten auf hypertextuellen Daten zu geben, wobei insbesondere graphentheoretische Methoden fokussiert werden. Die Anwendung von klassischen Data Mining-Konzepten, wie z.B. die Cluster- und die Klassifikationsanalyse, auf webbasierte Daten wird als *Web Mining* bezeichnet. Ein Teilbereich des Web Mining, der in dieser Arbeit besonders im Vordergrund steht, ist das *Web Structure Mining*, welches die Aufdeckung und die Erforschung von strukturellen Aspekten webbasierter Hypertextstrukturen zum Ziel hat. Die strukturelle Untersuchung von Hypertexten und speziell deren *graphentheoretische Analyse* hat sich besonders durch die Entwicklung des *World Wide Web* (WWW) zu einem eigenständigen Forschungsbereich im Hypertextumfeld entwickelt. Vergleicht man den aktuellen Forschungsstand dieses Bereiches jedoch aus der Sicht der Informationssysteme im Hypertextumfeld – den Hypertextsystemen – so fällt auf, dass die Entwicklung und Erforschung der Hypertextsysteme deutlich stärker und schneller fortgeschritten ist als die der strukturellen Analyse. Mit der Bedeutung der multimedialen Kommunikation stellen aber gerade graphentheoretische Methoden ein hohes Analysepotenzial zur Verfügung. Es besteht jedoch noch eine Herausforderung in der Entwicklung aussagekräftigerer, graphbasierter Modelle und graphentheoretischer Analysealgorithmen, die webbasierte Dokumentstrukturen ohne großen Strukturverlust verarbeiten können.

Dieser Artikel ist wie folgt strukturiert: In Kapitel (2) wird zunächst eine kurze Zusammenfassung der Grundlagen bezüglich Hypertext und Hypermedia gegeben. Während in Kapitel (3) Data Mining-Konzepte und die Teilgebiete des Web Mining vorgestellt werden, gibt Kapitel (4) einen Überblick über bestehende Arbeiten der graphentheoretischen Analyse von Hypertexten. Kapitel (5) stellt *Struktur entdeckende* Verfahren, die Clusteringverfahren, vor, die hier insbesondere als Motivation zur Anwendung auf Ergebnisse zu sehen sind, welche mit graphbasierten Methoden des Web Structure Mining erzielt werden.

2 Grundlagen von Hypertext und Hypermedia

Ausgehend vom klassischen Buchmedium ist die Struktur, und in der Regel auch die Lesereihenfolge, eines Buches sequentiell. Dagegen ist die Kerneigenschaft von *Hypertext*, dass die textuellen Informationseinheiten, die so genannten *Knoten*, auf der Basis von *Verweisen*, oder auch *Links* genannt, in Form eines gerichteten Graphen, also *nicht linear*, miteinander verknüpft sind (Kuhlen 1991). Die einfachste graphentheoretische Modellierung einer Hypertextstruktur ist die Darstellung als unmarkierter, gerichteter Graph $\mathcal{H} := (V, E)$, $E \subseteq V \times V$. Dabei heißen die Elemente $v \in V$ Knoten von $\mathcal{H}$ und $e \in E$ wird als gerichtete Kante bezeichnet.

Der Hypertext-Begriff lässt eine unterschiedliche Interpretationsform zwischen den Geisteswissenschaften und der modernen Informatik erahnen. So kann man abhängig von der Fachdisziplin und vom Autor durchaus auf unterschiedliche Definitionen des Hypertextbegriffs stoßen, und Hypertext wird somit oft als Technologie, Methode oder Metapher bezeichnet. Tatsächlich wurden in der Literatur unzählige Definitionen und Ausprägungen von Hypertext gegeben, siehe z.B. Charney (1987), Halasz (1987), Oren (1987). Aus diesen Definitionen – wobei die Autoren unterschiedliche Aspekte betonen – stellt Hofmann (1991) dennoch vier wichtige Kernpunkte heraus, die er für eine vollständige Charakterisierung von Hypertext in der Informatik als notwendig ansieht:

- Hypertexte haben die Gestalt von gerichteten Graphen (Netzwerke). Die Knoten enthalten bzw. repräsentieren die Informationen, die durch Verweise, die Links, miteinander verknüpft sind.
- Das Lesen als auch das Schreiben von Hypertext sind nichtlineare Tätigkeiten. Eine Datenstruktur, die diese Vernetzung unterstützt, ist dabei die Voraussetzung.
- Hypertexte sind nur in einem *medialen* Kontext, also maschinenunterstützt denkbar. Direkte Anwendungen davon sind klassische Hypertext- und Onlinesysteme.
- Hypertexte besitzen einen *visuellen* Aspekt. Das bedeutet, dass Hypertext nicht nur ein Konzept der Informationsstrukturierung, sondern auch eine Darstellungs- und Zugriffsform von textuellen Informationen ist.

Auch in der Sprachwissenschaft und in der Linguistik wurde Hypertext als eine neue Form der schriftlichen Sprachverwendung studiert, z.B. Lobin (1999), Storrer (2004). Dabei wurden insbesondere linguistische Aspekte, wie *Kohärenz*- und

Kohäsionsbeziehungen, in Hypertext untersucht. Eine bekannte Studie in diesem Problemkreis wurde von Storrer (1999) durchgeführt. In dieser Arbeit geht es im Wesentlichen um die Fragestellung, ob die Ergebnisse über Untersuchungen von Kohärenzbildungsprozessen in linear organisierten Texten für den Entwurf von Hypertexten übertragbar sind. Weiterhin wurde die interessante Problemstellung der *automatischen Generierung* von Hypertext aus natürlich sprachigem Text untersucht, insbesondere, wie und unter welchen Kriterien Hypertext automatisiert konstruierbar ist. Ein linguistisches Kriterium, welches als Grundlage zur Generierung von Hypertext aus Texten dient, ist von Mehler (2001) angegeben worden, wobei hier weitere bekannte Arbeiten zur automatischen Generierung von Hypertexten, aus informationswissenschaftlicher Sicht, beschrieben worden sind.

3 Problemstellungen des Web Mining

Durch die Entstehung des World Wide Web ist die Popularität von Hypertext in den neunziger Jahren deutlich gestiegen. Ein sehr bekanntes, modernes Forschungsfeld, das hypertextuelle Einheiten nach vielen Gesichtspunkten untersucht, ist das *Web Mining* (Chakrabarti 2002). Unter dem Begriff Web Mining versteht man genauer die Anwendung von *Data Mining*-Verfahren (Han & Kamber 2001) auf webbasierte, hypertextuelle Daten mit dem Ziel der automatischen *Informationsextraktion* und der Datenanalyse. Daher werden im Folgenden die Bereiche des Data Mining und deren Kernaufgaben vorgestellt. Data Mining Verfahren wurden entwickelt, um die gigantischen Datenmengen in vielen industriellen und wissenschaftlichen Bereichen zu analysieren und damit neues *Wissen* zu gewinnen. Beispielsweise liegen in vielen Unternehmen riesige Mengen von Kundendaten vor, jedoch ist das Wissen über die Anforderungen und über das Verhalten der Kunden oft nicht besonders ausgeprägt. Solche Datenbestände werden dann in *Data Warehouse* Systemen gespeichert und mit Methoden des Data Mining untersucht. Das Ziel einer solchen Untersuchung ist die Entdeckung von statistischen Besonderheiten und Regeln innerhalb der Daten, die beispielsweise für Studien des Kunden- oder Kaufverhaltens eingesetzt werden können. Die Schwerpunkte der Teilbereiche des Data Minings, lassen sich mit der Hilfe der folgenden Übersicht erläutern:

- Die Suche nach *Assoziationsregeln* (Hastie et al. 2001): Ein bekanntes Beispiel ist die so genannte *Warenkorbanalyse*, die zum Ziel hat, aus dem aktuellen Kaufverhalten Assoziationsregeln für zukünftiges Kaufverhalten abzuleiten.

- Die *Clusteranalyse* (Everitt et al. 2001): Der entscheidende Unterschied zwischen der Clusteranalyse und der *Kategorisierung* ist, dass bei der Clusteranalyse das Klassensystem von vorneherein unbekannt ist. Das Ziel ist die Gruppierung der Datenobjekte in Gruppen (Cluster), so dass sich die Objekte innerhalb eines Clusters möglichst ähnlich und zwischen den Clustern möglichst unähnlich sind. Dabei basiert die Ähnlichkeit zwischen den Objekten auf einem jeweils problemspezifischen Ähnlichkeitsmaß.
- Die *Kategorisierung* (Duda et al. 2000): Sie stellt Verfahren für die Einordnung von Objekten in Kategoriensysteme bereit. Die Kategorisierung stellt mit Hilfe von Zusammenhängen zwischen gemeinsamen Mustern und Merkmalen ein Kategoriensystem für die vorhandenen Objekte her, um dann auf der Basis eines statistischen Kategorisierungsmodells unbekannte Objekte in das Kategoriensystem einzuordnen. Bekannte Kategorisierungsverfahren stammen aus dem Bereich des *Machine Learning* oder basieren z.B. auf *Entscheidungsbäumen*.
- Die *Regressionsanalyse* (Hastie et al. 2001): Die Regressionsanalyse ist ein Verfahren aus der mathematischen Statistik, welches auf Grund von gegebenen Daten einen mathematischen Zusammenhang in Gestalt einer Funktion zwischen zwei oder mehreren Merkmalen herstellt.

Durch die äußerst starke Entwicklung des World Wide Web gewinnt die Anwendung von Data Mining Verfahren auf webbasierte Daten immer mehr an Bedeutung. Während das Allgemeinziel des Web Mining die Informationsgewinnung und die Analyse der Webdaten ist, werden drei bekannte Teilbereiche detailliert unterschieden (Kosala & Blockeel 2000):

- *Web Content Mining*: Das World Wide Web enthält mittlerweile viele Milliarden von Webseiten. Täglich kommen hunderttausende dazu. Das Web Content Mining stellt Methoden und Verfahren bereit, mit deren Hilfe Informationen, und damit neues Wissen, aus dieser Datenflut automatisch extrahiert werden können. Diese Verfahren finden beispielsweise bei der Informationssuche mit *Suchmaschinen* im World Wide Web eine Anwendung. Während bekannte Suchmaschinen wie z.B. `Yahoo` auf einer einfachen textuellen Schlagwortsuche basieren, stellt die Konzeption neuer, besserer Verfahren für die Informationssuche im Bereich des Web Content Mining immer noch eine große Herausforderung dar. Die aktuellen Suchmaschinen sind nicht in der Lage, *semantische Zusammenhänge* zwischen webbasierten Dokumenten zu detektieren bzw. die Dokumente nach semantischen Gesichtspunkten zu kategorisieren.

- *Web Structure Mining*: Die Aufgabe des Web Structure Mining ist es, strukturelle Informationen von Websites zu erforschen und zu nutzen, um inhaltliche Informationen zu gewinnen, wobei die *interne und externe Linkstruktur* eine wichtige Rolle spielt. Interne Linkstrukturen können mit Auszeichnungssprachen wie `HTML` oder `XML` abgebildet werden und beschreiben innerhalb eines Knotens eingebettete Graphstrukturen. Die externe Linkstruktur beschreibt die Verlinkung der Webseiten untereinander und lässt sich in Form eines hierarchisierten und gerichteten Graphen darstellen. Die Graphstruktur des World Wide Web ist in den letzten Jahren in vielen Arbeiten intensiv untersucht worden (Deo & Gupta 2001), wobei diese Studien zur Entwicklung und Verbesserung von Suchalgorithmen im World Wide Web geführt haben (Kleinberg 1998). Weiterhin sind *Ausgangsgrad-* und *Eingangsgradverteilungen* von Knoten, *Zusammenhangskomponenten* und der *Durchmesser* des WWW-Graphen untersucht worden. Detaillierte Ergebnisse solcher Untersuchungen sind z.B. in Broder et al. (2000), Deo & Gupta (2001) zu finden. Eine sehr bekannte Arbeit, die im Bereich des Web Structure Mining eine wichtige Anwendung innerhalb der bekannten Suchmaschine `Google` gefunden hat, stammt von Kleinberg. In Kleinberg (1998) führte er die Begriffe *Hubs* und *Authorities* ein. Kleinberg bezeichnet Authorities als Webseiten, die aktuelle und „inhaltlich brauchbare" Informationen enthalten, wobei sich diese graphentheoretisch durch hohe Knoten-Eingangsgrade auszeichnen. Dagegen werden Hubs als solche Webseiten bezeichnet, die viele „gute Links" zu gewissen Themengebieten offerieren. Ein guter graphentheoretischer Indikator für potentielle Hubs ist nach Kleinberg ein hoher Knoten-Ausgangsgrad der betrachteten Webseite.

- *Web Usage Mining*: Unter dem Web Usage Mining (Rahm 2000) versteht man die Suche und Analyse von Mustern, die auf das Nutzungsverhalten eines Users schließen lässt. Üblich ist dabei, die Anwendung von Data Mining Verfahren mit dem Ziel, das Zugriffsverhalten mit Hilfe von *Web-Logs* zu protokollieren. Die Ergebnisse solcher Analysen sind für Unternehmen, besonders aber für Online-Versandhäuser aller Art interessant, weil aus ihnen Aussagen zur Effektivität, zur Qualität und zum Optimierungsbedarf der Websites abgeleitet werden können. Da bei vielbesuchten Websites täglich riesige Datenmengen allein von Web-Logs anfallen, kann der Einsatz von Data Warehouse Systemen notwendig werden, um diese großen Datenmengen zielgerecht und effizient zu verarbeiten.

Die Bedeutung und die Vertiefung des Web Structure Mining, soll in diesem Artikel anhand von zwei weiteren Problemstellungen hervorgehoben werden und zwar 1.) zum Einen im Hinblick auf geplante Arbeiten im Bereich der strukturellen Analyse von webbasierten Hypertexten und 2.) zum Anderen als Motivation für das Kapitel (5):

1. Das Allgemeinziel des Web Structure Mining ist die Erforschung der strukturellen Eigenschaften von webbasierten Dokumentstrukturen und den daraus resultierenden Informationen. An diesem Ziel orientierend, soll an dieser Stelle auf ein Problem aufmerksam gemacht werden, welches bei der inhaltsorientierten Kategorisierung von webbasierten Hypertexten auftritt. Mehler et al. (2004) stellen die Hypothese auf, dass die beiden Phänomene *funktionale Äquivalenz* und *Polymorphie* charakteristisch für webbasierte Hypertextstrukturen sind. Dabei bezieht sich der Begriff der funktionalen Äquivalenz auf das Phänomen, dass dieselbe Funktions- oder Inhaltskategorie durch völlig verschiedene Bausteine webbasierter Dokumente manifestiert werden kann. Der Begriff der Polymorphie bezieht sich auf das Phänomen, dass dasselbe Dokument zugleich mehrere Funktions- oder Inhaltskategorien manifestieren kann. Nach Definition ist die Hypertextkategorisierung (Fürnkranz 2001) aber funktional, das heißt, jede webbasierte Einheit, z.B. eine Webseite, wird höchstens einer Kategorie zugeordnet. Die Ergebnisse der praktischen Kategorisierungsstudie (Dehmer et al. 2004, Mehler et al. 2004) untermauern jedoch die aufgestellte Hypothese, da es zu einer fehlerhaften Kategorisierung im Sinne von extremen Mehrfachkategorisierungen der Webseiten kam. Letztendlich folgt aber aus der auf der Basis des bekannten *Vektorraummodells* (Ferber 2003) durchgeführten Studie, dass diese Modellierung unzureichend ist. Das Ziel bleibt daher eine verstärkte strukturelle Analyse und eine adäquate Modellierung webbasierter Dokumente.
2. Im Hinblick auf die Bestimmung der Ähnlichkeit webbasierter Dokumente fassen *Document Retrieval* Anwendungen die Dokumente als die Mengen ihrer Wörter auf und berechnen auf der Basis des Vektorraummodells deren Ähnlichkeit. Als Motivation für eine graphorientierte Problemstellung innerhalb des Web Structure Mining und für Kapitel (5), wird eine Methode von (Dehmer et al. 2004, Emmert-Streib et al. 2005) zur Bestimmung der strukturellen Ähnlichkeit skizziert, die nicht auf der vektorraumbasierten Repräsentation beruht, sondern auf der Graphdarstellung von webbasierten Hypertexten. Ausgehend von der automatisierten Extraktion der Hypertexte und einer GXL-Modellierung (Winter 2002) der

Graphen, werden hierarchisierte und gerichtete Graphen erzeugt, die komplexe Linkstrukturen berücksichtigen (Mehler et al. 2004). Diese Graphen werden in eindimensionale Knotensequenzen abgebildet. Das Problem der strukturellen Ähnlichkeit von zwei Graphen ist dann gleichbedeutend mit der Suche eines optimalen Alignments dieser Sequenzen (bezüglich einer Kostenfunktion α). Da es sich um hierarchisierte Graphstrukturen handelt, erfolgt die Bewertung der Alignments ebenenorientiert durch die induzierten Ausgangsgrad- und Eingangsgradsequenzen auf einem Level $i, 0 \leq i \leq h$, wobei h die Höhe der Hypertextstruktur bezeichnet. Die Berechnung der Ähnlichkeit erfolgt schließlich über ein Maß, in das die Werte der Ähnlichkeiten von Ausgangsgrad- und Eingangsgradalignments eingehen. Da diese Methode in seiner algorithmischen Umsetzung effizient ist, verspricht sie im Hinblick auf die Verarbeitung von Massendaten ein für das Web Structure Mining hohes Anwendungspotenzial, z.B.:

- Die Bestimmung der strukturellen Ähnlichkeit von webbasierten Dokumentstrukturen, wie z.B. graphbasierte Websitestrukturen in Form von hierarchisierten und gerichteten Graphen oder DOM-Trees (Chakrabarti 2001).
- Suche und struktureller Vergleich von Graphpatterns in webbasierten Hypertextstrukturen bei Fragen der Interpretation von Hypertext-Navigationsmustern.
- Besseres Verständnis der graphentheoretischen Struktur webbasierter Hypertexte.

4 Graphentheoretische Analyse von Hypertextstrukturen

Wie in Kapitel (2) bereits dargestellt, lässt sich die auszeichnende strukturelle Eigenschaft von Hypertext, die Nichtlinearität, in Form eines Netzwerks mit Hilfe einer graphentheoretischen Modellierung beschreiben. Damit liegt die Frage nach der Einsetzbarkeit von graphentheoretischen Analysemethoden auf der Hand. Das vorliegende Kapitel soll einen Eindruck über die Realisierbarkeit graphbasierter Modellierungen und über die Tragfähigkeit der Aussagen geben, die man mit einfachen graphentheoretischen Modellen, angewendet auf die Hypertextstruktur, erzielen kann. Als erste Motivation für graphorientierte Methoden sei die Analyse des oft zitierten *„lost in hyperspace"*-Problems (Unz 2000) genannt. Aus der Natur der graphbasierten Modellierung, einer hohen Komplexität der vorliegenden Hypertextstruktur, einem fehlenden kontextuellen Zu-

sammenhang der Links und der Tatsache, dass der Navigierende nur einen eingeschränkten Bereich im Hypertextgraph rezipiert, folgt, dass der *Hypertextuser* die Orientierung verlieren kann. Graphentheoretische Analysemethoden, die als Abstraktionswerkzeug zu verstehen sind, werden oft eingesetzt, um das „lost in hyperspace"-Problem besser unter Kontrolle zu halten. Dazu werden graphentheoretische Kenngrößen definiert, die beispielsweise Aussagen über die Erreichbarkeit von Knoten und deren Einfluss im Hypertextgraph treffen (Botafogo & Shneiderman 1991, Botafogo et al. 1992, Ehud et al. 1994). Die Definition von *Indizes* zur Beschreibung typischer Ausprägungen von Hypertextgraphen kann als weitere Motivation für den Einsatz graphentheoretischer Methoden angesehen werden. Beispielsweise können solche Maße von *Hypertextautoren* eingesetzt werden, um den *Vernetztheitsgrad* und die *Linearität* einer Hypertextstruktur zu bestimmen (Botafogo et al. 1992). Eine weitaus tiefer gehende Fragestellung wäre an dieser Stelle, ob man auf der Basis von graphentheoretischen Indizes eine Gruppierung von ähnlichen Strukturen vornehmen könnte, um dann auf ähnliche Funktionen und Qualitätsmerkmale zu schließen. In jedem Fall müssen aber Fragen nach der *Einsetzbarkeit* und der *Interpretierbarkeit* solcher Maßzahlen gestellt werden, die in Kapitel (4.1) kurz diskutiert werden.

Dieses Kapitel gibt im Wesentlichen einen Überblick über die bekannten Arbeiten der graphentheoretischen Analyse von Hypertextstrukturen, wobei es nicht den Anspruch auf Vollständigkeit erhebt. Einerseits werden damit Möglichkeiten vorgestellt wie man mit einfachen graphentheoretischen Mitteln Hypertexte auf Grund charakteristischer Eigenschaften beschreiben und solche Maße auf Probleme der Hypertextnavigation anwenden kann. Andererseits zeigen einige der nachfolgenden Arbeiten die Grenzen von graphentheoretischen Maßzahlen auf, die sich z.B. in der Allgemeingültigkeit ihrer Aussagekraft und in der Interpretierbarkeit ihrer Wertebereiche äußern.

Die in der Fachliteratur existierenden Ansätze und Arbeiten, die sich mit der graphentheoretischen Analyse und Beschreibung von Hypertextstrukturen beschäftigen, verfolgen im Wesentlichen die folgenden Ziele:

- Die strukturelle Beschreibung und Charakterisierung von Hypertexten durch *globale* graphentheoretische Maße. Sie heißen global, weil sie auf der gesamten Hypertextstruktur definiert sind. Sehr bekannte Beispiele sind die Hypertextmetriken *Compactness* und *Stratum* von Botafogo et al. (1992).

- Die Suche, die Bestimmung und die graphentheoretische Interpretation von Graphmustern in Hypertexten. Solche spezifischen Graphmuster werden oft bei der Beschreibung von Hypertext-Navigationsproblemen

(McEneaney 2000, Unz 2000) und im Zusammenhang von Lernproblemen (Noller et al. 2002, Winne et al. 1994) mit Hypertext analysiert und interpretiert.

Die ersten einschneidenden Arbeiten im Bereich der strukturellen Analyse von Hypertexten stammen von Botafogo & Shneiderman (1991), Botafogo et al. (1992), Ehud et al. (1994). In Botafogo et al. (1992) wurden die bekannten Hypertextmetriken Compactness und Stratum definiert, wobei in dieser Untersuchung Hypertextgraphen als unmarkierte, gerichtete Graphen $\mathcal{H} = (V, E)$, $E \subseteq V \times V$, aufgefasst werden. Mit Hilfe der *konvertierten Distanzmatrix*

$$(\mathcal{KDM}_{ij})_{ij} := \begin{cases} w_{ij} & : \text{falls } w_{ij} \text{ existiert} \\ \mathcal{K} & : \text{sonst,} \end{cases} \tag{1}$$

wobei w_{ij} den kürzesten Weg von v_i nach v_j und $\mathcal{K}$ die Konvertierungskonstante[1] bezeichnet, wird Compactness definiert als

$$\mathcal{C} := \frac{(|V|^2 - |V|) \cdot \mathcal{K} - \sum_{i=1}^{|V|} \sum_{j=1}^{|V|} \mathcal{KDM}_{ij}}{(|V|^2 - |V|) \cdot \mathcal{K} - (|V|^2 - |V|)}. \tag{2}$$

$|V|$ bezeichnet die Ordnung (Anzahl der Knoten) des Hypertextgraphs und nach Definition gilt $\mathcal{C} \in [0,1]$. Es ist $\mathcal{C} = 0 \iff \mathcal{H} = (V, \{\})$. Weiterhin gilt $\mathcal{C} = 1 \iff |E| = |V \times V| - |V|$. $(|V|^2 - |V|) \cdot \mathcal{K}$ ist der Maximalwert der Matrixelemente aus der konvertierten Distanzmatrix. Er wird erreicht, falls $E = \{\}$. $(|V|^2 - |V|)$ ist der minimale Wert der Summe der Matrixelemente und wird erreicht, wenn $\mathcal{H}$ der *vollständige Graph* ist.

Informell ausgedrückt, gibt der Wert für das Gütemaß Compactness bezüglich einer bestimmten Hypertextstruktur Aufschluss darüber, wie „dicht" die Hypertextstruktur vernetzt ist. Ein hoher Compactness-Wert im Sinne von Botafogo sagt aus, dass von jedem Knoten aus jeder andere Knoten leicht erreicht werden kann.

Als Beispiel betrachte man die Graphen aus Abbildung (1). Der erste Graph ist der vollständige[2] gerichtete Graph K_4 und nach Gleichung (2) folgt $\mathcal{C} = 1$. Der zweite Graph besitzt die leere Kantenmenge, deshalb $\mathcal{C} = 0$. In Botafogo et al. (1992) wurde von einigen Hypertexten der Compactness-Wert bestimmt und näher untersucht. So besaß beispielsweise die hypertextuelle Beschreibung des Fachbereichs Informatik der Universität Maryland CMSC (Computer Science De-

1 Botafogo et al. (1992) setzen in ihren Untersuchungen $\mathcal{K} = |V|$.

2 Allgemein wird der vollständige Graph mit n Knoten in der Graphentheorie mit K_n bezeichnet.

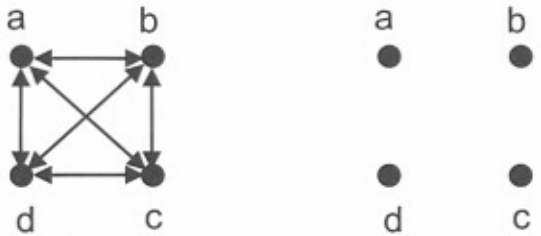

Abbildung 1: Der vollständige gerichtete Graph K_4 und der entsprechende Graph mit der leeren Kantenmenge

partment at the University Maryland) einen Compactness-Wert von C=0.53. Für das Buch in Hypertextform HHO (Hypertext Hands On!, Shneiderman & Kearsley (1989)) wurde der Wert C=0.55 ermittelt. Da es sich bei diesen Hypertexten um hierarchische, baumähnliche Graphen handelte lag die Vermutung nahe, dass ein Compactness-Wert von ca. 0.5 typisch für solch strukturierte Hypertexte ist. Die Bildung eines Intervalls, in das man die Compactness-Werte von Hypertexten einordnen kann, um dann aus dem Wert innerhalb dieses Intervalls auf Gütemerkmale wie z.B. „gutes Navigationsverhalten" zu schließen, ist jedoch aus Gründen der nicht eindeutigen Interpretierbarkeit dieser Hypertextmetrik nicht möglich.

Für die Definition von Stratum betrachte man die Distanzmatrix von $\mathcal{H}$

$$(\mathcal{D}_{ij})_{ij} := \begin{cases} w_{ij} & : \text{falls } w_{ij} \text{ existiert} \\ \infty & : \text{sonst.} \end{cases}$$

$(\hat{\mathcal{D}}_{ij})_{ij}$ sei die Matrix, die man durch Ersetzung der Matrixelemente ∞ durch 0 in $(\mathcal{D}_{ij})_{ij}$ erhält. Botafogo zeigt in Botafogo et al. (1992), dass damit für Stratum die Gleichungen

$$\mathcal{S} = \begin{cases} \frac{4\sum_{i=1}^{|V|}\left(\left|\sum_{j=1}^{|V|}\hat{\mathcal{D}}_{ji}-\sum_{j=1}^{|V|}\hat{\mathcal{D}}_{ij}\right|\right)}{|V|^3} & : \text{falls } |V| \text{ gerade} \\ \frac{4\sum_{i=1}^{|V|}\left(\left|\sum_{j=1}^{|V|}\hat{\mathcal{D}}_{ji}-\sum_{j=1}^{|V|}\hat{\mathcal{D}}_{ij}\right|\right)}{|V|^3-|V|} & : \text{falls } |V| \text{ ungerade,} \end{cases}$$

bestehen. Nach Definition von $\mathcal{S}$ gilt $\mathcal{S} \in [0,1]$. $\mathcal{S} = 0$ bedeutet, dass die Hypertextstruktur in sich geschlossen und beispielsweise kreisförmig angeordnet ist. $\mathcal{S} = 1$ beschreibt $\mathcal{H}$ in Form einer vollständig linearen Graphstruktur. Wenn man zur gegebenen Hypertextstruktur die zugehörige Hierarchisierung

betrachtet, drückt Stratum aus wie tief und linear die hierarchische Struktur ist. Beide Maße, Compactness und Stratum, sind auf unmarkierten gerichteten Graphen definiert und beinhalten keinerlei semantische Relationen des vorgelegten Hypertextes. Botafogo et al. führten diese Untersuchungen durch, in dem sie von allen semantischen, pragmatischen und syntaktischen Typmerkmalen der hypertextuellen Träger abstrahierten. Ein bekanntes Phänomen von *quantitativen* Maßen zur strukturellen Charakterisierung von Hypertexten und zur Beschreibung von Hypertextnavigationsproblemen ist, dass die Ergebnisse solcher Maße oft vom konkret betrachteten Hypertext abhängen und somit mit anderen Messungen schlecht vergleichbar sind. Um diesem Problem entgegen zu wirken, stellte Horney (1993) eine weitere Untersuchung zur Messung von Hypertextlinearität, in Bezug auf die Hypertextnavigation, an. Dabei untersuchte Horney Pfadmuster, die durch bestimmte Aktionen der *User* im Hypertext erzeugt wurden, indem er Pfadlängen ausgehend von Knoten und *Vaterknoten* bestimmte und mittelte. Dieses Prinzip wandte er auf das gesamte Hypertextdokument an und erhielt somit lineare Funktionen für diese Sachverhalte, die er als ein Maß für die Linearität eines Hypertextes definierte.

Abgesehen von Botafogo et al. untersuchten und evaluierten De Bra & Houben (1997) ebenfalls Compactness und Stratum. Da in Botafogo et al. (1992) Compactness und Stratum unter der Annahme definiert worden sind, dass im Hypertextgraph lediglich Vorwärtsbewegungen[3] ausgeführt werden, definierten sie Compactness und Stratum neu, und zwar unter dem Aspekt, Backtracking-Bewegungen[4] im Hypertextgraph durchzuführen. Somit werden durch die modifizierten Maße *navigational Compactness* und *navigational Stratum* von De Bra et al. die Navigationsstrategien von Usern in Hypertextstrukturen besser abgebildet.

Ebenfalls wurden die Auswirkungen von Compactness und Stratum auf die Hypertext-Navigation in McEneaney (2000) untersucht, indem aus den schon bekannten Maßen Pfadmetriken definiert und diese empirisch evaluiert wurden. Anstatt der in Botafogo et al. (1992) definierten Matrizen, verwendete McEneaney Pfadmatrizen für die analoge Anwendung dieser Hypertextmetriken. In einer Pfadmatrix repräsentiert ein Matrixelement die Häufigkeit von Knotenübergängen von einem Knoten zu jedem anderen Knoten im Navigationspfad. Diese Pfadmetriken ermöglichen aus Navigationsmustern, dargestellt durch Navigationspfade, die Navigationsstrategien von Hypertextusern zu erkennen.

3 Im Sinne von Botafogo et al. (1992) heißt das: Falls der Weg von v_i zu v_j nicht existiert, wird er mit der Konvertierungskonstante K bewertet.

4 Das heißt, man folgt der gerichteten Kante (v_j, v_i), falls man vorher die Bewegung (v_i, v_j) ausgeführt hat.

Außer Compactness, Stratum und den bisher vorgestellten Maßen gibt es noch weitere graphentheoretische Maße im Hypertextumfeld, die jetzt vorgestellt werden. Unz (2000) beschreibt außer Compactness und Stratum die zwei weiteren Maße *Density* und *Kohäsion*. Hauptsächlich gibt Unz (2000) aber einen umfassenden Überblick über das Thema „Lernen mit Hypertext", insbesondere bezogen auf Navigationsprobleme und die Informationssuche in Hypertexten. Density und Kohäsion wurden ursprünglich von Winne et al. (1994) eingeführt, um das Verhalten von Hypertextusern im Zusammenwirken mit bestimmten Lernaktionen, wie z.B. „Einen Text markieren", „Einen Text unterstreichen" und „Eine Notiz machen" im Hypertextsystem STUDY graphentheoretisch zu analysieren. Um die spezifischen Graphmuster der Hypertextuser zu gewinnen, bilden WINNE et al. formale Sequenzen von ausgeführten Lernaktionen in *Adjazenzmatrizen* ab und erhalten so Graphmuster, die das Benutzerverhalten wiedergeben. Dabei hat eine gewöhnliche Adjazenzmatrix die Gestalt

$$\mathcal{A} := \begin{cases} 1 & : (v_i, v_j) \in E \\ 0 & : \text{sonst}. \end{cases}$$

Um dann messen zu können, welche Aktionen bei den Hypertextusern welche Auswirkungen hatten, definierten WINNE et al. die graphentheoretischen Maßzahlen

$$\mathcal{D} := \frac{\sum_{i=1}^{|V|} \sum_{j=1}^{|V|} a_{ij}}{|V|^2}, \qquad \text{(Density)} \tag{3}$$

und

$$\mathcal{COH} := \frac{\sum_{i=1}^{|V|} \sum_{j=1}^{|V|} a_{ij} \cdot a_{ji}}{\frac{|V|^2-|V|}{2}}. \qquad \text{(Kohäsion)} \tag{4}$$

In den Gleichungen (3), (4) bezeichnet a_{ij} den Eintrag in der Adjazenzmatrix in der i-ten Zeile und der j-ten Spalte. $\mathcal{D}$ gibt das Verhältnis von der Anzahl der tatsächlich vorkommenden Kanten, zur Anzahl aller möglichen Kanten inklusive *Schlingen* an und nach Definition gilt $\mathcal{D} \in [0,1]$. $\mathcal{COH}$ misst den Anteil von zweifach-gerichteten Kanten – das sind Kanten der Form $(v_i, v_j), (v_j, v_i)$ für zwei Knoten $v_i, v_j \in V$ – ohne Schlingen. Der Ausdruck $\frac{|V|^2-|V|}{2}$ gibt die Anzahl aller möglichen Knotenpaare an und es gilt ebenfalls $\mathcal{COH} \in [0,1]$. Aus der Definition der Kohäsion schlossen Winne et al. (1994) nun: Je höher der Wert für die Kohäsion eines betrachteten Graphmusters ist, desto weniger schränkten die Lernaktionen den Hypertextuser ein. Allgemeiner betrachtet kann man diese Maße als benutzerspezifische Präferenzen innerhalb des Graphmusters interpretieren. Weitergehender untersuchten Noller et al. (2002) diese Problematik und entwickelten eine automatisierte Lösung zur Analyse von Navigationsverläufen. Die Navigationsmuster analysierten sie mit graphentheoretischen Mitteln und interpretierten sie ebenfalls als psychologische Merkmale wie z.B. gewisse Verarbeitungsstrategien, konditionales Vorwissen und benutzerspezifische Präferenzen.

Bisher sind hauptsächlich graphentheoretische Maße vorgestellt worden, die zur strukturellen Charakterisierung von Hypertext und zur Interpretation von Graphmustern dienen. Bekannt sind aber auch solche graphentheoretischen Maße, die zur Charakterisierung von Graphelementen konstruiert wurden, vor allem für Knoten in einem Graph. Solche Maße sind in der Fachliteratur allgemeiner als *Zentralitätsmaße* bekannt und finden starke Anwendung in der *Theorie der sozialen Netzwerke*. Sehr bekannte und grundlegende Arbeiten in diesem Bereich findet man bei Harary (1965). *Knotenzentralitätsmaße*, die etwas über die „Wichtigkeit" und „Bedeutsamkeit" von Knoten im Graph aussagen, wurden auch von Botafogo et al. (1992) definiert, bzw. bekannte Maße in einem neuen Kontext angewendet. So definierten sie die Maße

$$\text{ROC}_v := \frac{\sum_{i=1}^{|V|} \sum_{j=1}^{|V|} \mathcal{KDM}_{ij}}{\sum_{j=1}^{|V|} \mathcal{KDM}_{vj}}, \qquad \text{(Relative Out Centrality)}$$

$$\text{RIC}_v := \frac{\sum_{i=1}^{|V|} \sum_{j=1}^{|V|} \mathcal{KDM}_{ij}}{\sum_{j=1}^{|V|} \mathcal{KDM}_{jv}}. \qquad \text{(Relative In Centrality)}$$

Dabei bedeuten $\mathcal{KDM}_{ij}$ wieder die Einträge in der konvertierten Distanzmatrix, die durch die Definitionsgleichung (1) bereits angegeben wurde. BOTAFOGO et al. wandten das ROC-Maß an, um beispielsweise so genannte *Landmarks* zu kennzeichnen. So werden identifizierbare Orientierungspunkte im Hypertext bezeichnet, weil Landmarks die Eigenschaft besitzen, mit mehr Knoten verbunden zu sein als andere Knoten im Hypertext. BOTAFOGO et al. kennzeichneten damit Knoten mit einem hohen ROC-Wert als Kandidaten für Landmarks. Dagegen sind Knoten mit niedrigem RIC-Wert im Hypertextgraph schwer zu erreichen. Letztlich dienen aber diese beiden Maße zur Analyse von Navigationsproblemen und damit wieder zum besseren Umgang mit dem „lost in hyperspace"-Problem. Als Abschluss dieser Übersicht wird noch eine Arbeit genannt, die ein graphentheoretisches Maß für den Vergleich von Hypertextgraphen liefert.

Dafür definierten Winne et al. (1994) das Maß *Multiplicity* für zwei gerichtete Graphen $\mathcal{H}_1$ und $\mathcal{H}_2$ als,

$$\mathcal{M} := \frac{\sum_{i=1}^{|V|}\sum_{j=1}^{|V|} a_{ij} \cdot b_{ij}}{|V|^2} \qquad i \neq j\,. \tag{5}$$

Nach Definition gilt $\mathcal{M} \in [0,1]$ und a_{ij} bzw. b_{ij} bezeichnen in Gleichung (5) die Einträge in der Adjazenzmatrix von $\mathcal{H}_1$ bzw. $\mathcal{H}_2$. Dabei wird hier die Knotenmenge V als gemeinsame Knotenmenge der beiden Graphen angesehen und Multiplicity misst damit die Anzahl der gemeinsamen Kanten beider Graphen, relativ zur Anzahl aller möglichen Kanten. Die Motivation zur Definition von Multiplicity war, individuelle Taktiken und Strategien, die sich in zwei Graphen niederschlagen, vergleichbarer zu machen.

4.1 Kritik und Ausblick

Die Darstellungen in Kapitel (4) zeigen, dass die Wirkung und die Aussagekraft von globalen Maßen zur strukturellen Charakterisierung von Hypertexten und zur Beschreibung von Graphmustern, z.B. Navigationsverläufe, beschränkt ist. Das liegt zum einen daran, dass einige der vorgestellten Maße für speziellere Problemstellungen entwickelt worden sind oder in einer speziellen Studie entstanden sind, z.B. bei Winne et al. (1994). Auf der anderen Seite erlauben quantitativ definierte Maße, z.B. Compactness (Botafogo et al. 1992), keine allgemeingültigen Aussagen über eine verlässliche strukturelle Klassifikation von Hypertextgraphen bzw. über die Güte und Verwendbarkeit solcher Strukturen. Eine aussagekräftige Evaluierung der Maße und die Interpretation einer solchen Auswertung ist in vielen Fällen nicht erfolgt. Ein positiver Aspekt ist die durchgängig klare, einfache mathematische Modellierung und die leichte Implementierbarkeit, indem von komplexeren Typmerkmalen der Knoten und Links abstrahiert wird. Der negative Aspekt, der daraus unmittelbar resultiert, ist die fehlende semantische Information über solche Typmerkmale, die sich in insbesondere in der mangelnden Interpretierbarkeit von Werteintervallen innerhalb des ausgeschöpften Wertebereichs äußert.

Für den Vergleich von Hypertextgraphen, im Hinblick auf lernpsychologische Implikationen, ist das Maß Multiplicity von Winne et al. (1994), welches über der Kantenschnittmenge definiert ist, vorgestellt worden. Beispielsweise ist mit Multiplicity kein ganzheitlich struktureller Vergleich komplexer Hypertextgraphen möglich, da dieses Maß zu wenig der „gemeinsamen Graphstruktur" erfasst. Wünschenswert wäre für den strukturellen Vergleich solcher Hypertextgraphen

ein Modell, welches (i) möglichst viel der gemeinsamen Graphstruktur erfasst und (ii) parametrisierbar ist, d.h. die Gewichtung spezifischer Grapheigenschaften. An dieser Stelle sei nun als Ausblick und Motivation für weitere Arbeiten die automatisierte Aufdeckung und die verstärkte Erforschung der graphentheoretischen Struktur, gerade für webbasierte Hypertexte, genannt, weil (i) bisher wenig über deren charakteristische graphentheoretische Struktur und deren Verteilungen bekannt ist (Schlobinski & Tewes 1999) und (ii) im Hinblick auf anwendungsorientierte Problemstellungen die Graphstruktur ganz besonders als Quelle zur Informationsgewinnung dienen kann. Das bedeutet mit stetig steigender Anzahl der hypertextuellen Dokumente im WWW werden Aufgaben wie die gezielte Informationsextraktion, das automatisierte *webbasierte Graphmatching* und die Gruppierung ähnlicher Graphstrukturen (s. Kapitel (5)) für ein effizientes *Web Information Retrieval* immer wichtiger. In Bezug auf das webbasierte Graphmatching wurde bereits das am Ende des Kapitel (3) skizzierte Verfahren von Dehmer et al. (2004), Emmert-Streib et al. (2005) erwähnt.

5 Verfahren zur Clusterung von Daten

In Kapitel (4) sind bekannte Arbeiten zur graphentheoretischen Analyse von Hypertextstrukturen vorgestellt worden. Dabei kamen auch Maße zur Beschreibung typischer Ausprägungen von Hypertextstrukturen und deren Anwendungen zur Besprechung. Im Hinblick auf die Entwicklung weiterführender graphentheoretischer Methoden im Bereich des Web Structure Mining werden in diesem Kapitel eine Gruppe von *multivariaten Analysemethoden*, die *Clusteringverfahren*, vorgestellt. Bei den im Kapitel (4) dargestellten Verfahren, stand die Charakterisierung typischer Ausprägungen graphbasierter Hypertexte auf der Basis numerischer Maßzahlen im Vordergrund. Im Gegensatz dazu gehören die Clusteringverfahren zur Gruppe der Struktur entdeckenden Verfahren, weil deren Ziel die Aufdeckung von strukturellen Zusammenhängen zwischen den betrachteten Objekten ist. Dabei ist die Einbeziehung mehrerer vorliegender Objektausprägungen die stark auszeichnende Eigenschaft von Clusteringverfahren (Backhaus et al. 2003). Als Motivation zum vorliegenden Kapitel können Clusteringverfahren, als Bindeglied des webbasierten Graphmatching, beispielsweise (i) zur Aufdeckung von *Typklassen*[5] webbasierter Hypertexte eingesetzt werden oder (ii) zur Trennung von strukturell signifikant unterschiedlichen Webseiten.

Clusteringverfahren (Everitt et al. 2001) werden zur Gruppierung (Clusterung) von Objekten angewendet, um möglichst *homogene* Cluster zu erzeugen.

5 Z.B. die Klasse der Mitarbeiterseiten innerhalb eines akademischen Webauftritts

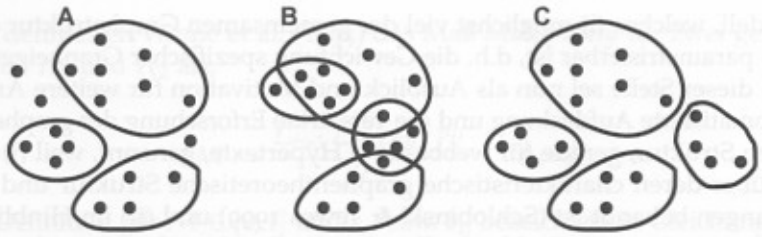

Abbildung 2: A: Disjunkte, aber nicht partitionierende Clusterung mit *nicht gruppierbaren* Objekten. B: Überlappende Clusterung. C: Partitionierende Clusterung

In der Regel ist bei Beginn der Clusterung die Anzahl der Cluster und die Clusterverteilung unbekannt, somit auch die Zuordnung der Objekte innerhalb der einzelnen Cluster. Clusteringverfahren sind deshalb im Bereich des *Unsupervised Learning* (Hastie et al. 2001) angesiedelt, weil sie „unüberwacht", also ohne Lernregeln, eine möglichst optimale Clusterung erzeugen sollen. Die Clusterung soll die Kerneigenschaft besitzen, dass ähnliche Objekte in Clustern zusammengeschlossen werden, so dass die Objekte der gefundenen Cluster eine ganz bestimmte *Charakteristik* aufweisen, bzw. jedes Cluster einen eigenen *Typ* repräsentiert. Die Abbildung (2) zeigt verschiedene Varianten von Clusterungen, die entweder je nach Anwendungsfall gewünscht sind oder deren Effekte, z.B. die Überlappung der Cluster, verfahrensbedingt auftreten.

Formeller ausgedrückt lässt sich diese Aufgabe für das Web Mining folgendermaßen beschreiben: Es sei $D := \{d_1, d_2, \ldots, d_n\}$, $\mathbb{N} \ni n > 1$ die Menge der zu clusternden Dokumente. Will man die Clusteraufgabe in voller Allgemeinheit beschreiben, so fasst man die Dokumentenmenge als eine Menge $O := \{O_1, O_2, \ldots, O_n\}$ von unspezifizierten Objekten $O_i, 1 \leq i \leq n$ auf. Eine *Clusterung* C_{fin} ist nun eine k-elementige *disjunkte Zerlegung* von D, also $C_{fin} := \{C_i \subseteq D | 1 \leq i \leq k\}$. Die Cluster C_i sollen dabei die Eigenschaft besitzen, dass basierend auf einem problemspezifischen Ähnlichkeitsmaß $s : D \times D \longrightarrow [0,1]$ (oder Abstandsmaß $d : D \times D \longrightarrow [0,1]$), die Elemente $d \in C_i$ eine hohe Ähnlichkeit zueinander besitzen, wohingegen die Elemente $d, \tilde{d}$ mit $d \in C_i \wedge \tilde{d} \in C_j, i \neq j$ eine geringe Ähnlichkeit zueinander besitzen sollen. Falls die Ähnlichkeits- oder Abstandsmaße bei webbasierten Dokumentstrukturen auf *inneren* (strukturellen) Eigenschaften des Dokuments basieren, ist z.B. die Darstellung gemäß Vektorraummodell oder eine graphentheoretisch basierte Modellierung gemeint.

In der Praxis des Web Mining finden oft *partitionierende-* und *hierarchische*

Clusteringverfahren Anwendung, wobei es noch eine Vielzahl anderer Verfahren gibt, z.B. *graphentheoretische*, *probabilistische* und *Fuzzy* Clusteringverfahren (Everitt et al. 2001). Bevor ein Clusteringverfahren angewendet wird, ist es wichtig, die Ausprägungen der Beschreibungsmerkmale zu analysieren, um dann entscheiden zu können, ob zur Beschreibung der Unterschiede zwischen den Dokumenten ein Ähnlichkeits- oder ein Abstandsmaß gewählt wird. Die Frage nach der Lösung einer Clusteraufgabe stellt in der Regel ein Problem dar, da sie von der jeweiligen Anwendung und vom Verwendungszweck der Clusterung abhängt. Oft wählt man eine überschaubare Anzahl der gewonnenen Cluster aus, um sie entweder (i) aus der jeweiligen Anwendungsperspektive zu interpretieren oder (ii) sie mit statistischen Mitteln auf ihre Aussagekraft hin zu überprüfen. Generell sind die Anforderungen an moderne Clusteringverfahren hoch, da sie auf der Basis ihrer Konzeption möglichst viele Eigenschaften besitzen sollen, z.B.:

- geringe Parameteranzahl
- einfache Interpretierbarkeit der Cluster
- gute Eigenschaften bei *hochdimensionalen* und *verrauschten* Daten
- die Verarbeitung von möglichst vielen Datentypen.

Jedoch ist nicht jedes Verfahren, das diese Eigenschaften besitzt, für eine Clusteraufgabe geeignet, weil die Verfahren gewisse Vor- und Nachteile besitzen, die in der Regel von den Daten, dem zugrundeliegenden Ähnlichkeits- oder Abstandsmaß und der Konstruktion des Verfahrens abhängen. Dennoch sind die meisten bekannten Clusteringverfahren theoretisch und praktisch intensiv untersucht worden, so dass sie gut voneinander abgrenzbar sind und somit die Auswahl eines Verfahrens für eine Clusteraufgabe leichter fällt.

5.1 Interpretation von Clusterlösungen

Um die Wirkungsweise von Clusteringverfahren besser zu verstehen, wird zunächst allgemein die Forderung der *Homogenität*, die bereits in Kapitel (5) kurz erwähnt wurde, erläutert. Eine anschauliche Interpretation dieses Maßes, bezüglich eines Clusters C, liefert Bock (1974), indem er die Homogenität als numerische Größe $h(C) \geq 0$ beschreibt, die angibt, wie ähnlich sich die Objekte in C sind, oder anders formuliert, wie gut sich diese Objekte durch ihre charakteristischen Eigenschaften beschreiben lassen. Ausgehend von einer Objektmenge $O = \{O_1, O_2, \ldots, O_n\}$, einem Cluster $C \subseteq O$ und einer Ähnlichkeitsma-

trix $(s_{ij})_{ij}$, $1 \leq i \leq n$, $1 \leq j \leq n$, $s_{ij} \in [0,1]$ gibt Bock (1974) ein Maß für die Homogenität von C durch

$$h(C) := \frac{1}{|C| \cdot (|C|-1)} \sum_{\mu \in I_C} \sum_{\nu \in I_C} s_{\mu\nu} \in [0,1] \tag{6}$$

an, wobei I_C die entsprechende Indexmenge von C bezeichnet. Je größer $h(C)$ ist, desto homogener ist C und umgekehrt. Ist anstatt der Ähnlichkeitsmatrix eine Distanzmatrix $(d_{ij})_{ij}$, $1 \leq i \leq n$, $1 \leq j \leq n$ gegeben, so sind

$$\begin{aligned} h_1^*(C) &:= \frac{1}{|C| \cdot (|C|-1)} \sum_{\mu \in I_C} \sum_{\nu \in I_C} d_{\mu\nu}, \\ h_2^*(C) &:= \frac{1}{2|C|} \sum_{\mu \in I_C} \sum_{\nu \in I_C} d_{\mu\nu} \end{aligned}$$

Maße für die *Inhomogenität* und es gilt hier: je kleiner die Werte von $h_i^*(C), i \in \{1,2\}$ sind, desto homogener ist C und umgekehrt.

Insgesamt gesehen kann oftmals das Ergebnis einer Clusterung als der erste Schritt betrachtet werden, um detailliertes Wissen über die betrachteten Objekte zu erlangen und um darüber hinaus eventuell neue Eigenschaften der Objekttypen zu erkennen. Weiterhin ist es notwendig die Interpretation einer Clusterlösung vor einem speziellen Anwendungshintergrund zu sehen oder das Ergebnis der Clusterung stellt die Grundlage für eine weitergehende praktische Anwendung dar, da eine Clusterlösung für sich isoliert betrachtet, keine weitreichende Aussagekraft besitzt.

5.2 Hierarchische Clusteringverfahren

Um nun die grundlegende Funktionsweise von hierarchischen Clusteringverfahren für das Web Mining zu beschreiben, sei wieder die Dokumentenmenge $D := \{d_1, d_2, \ldots, d_n\}$ mit einem problemspezifischen Ähnlichkeitsmaß $s : D \times D \longrightarrow [0,1]$ (oder Abstandsmaß) betrachtet. Bock motiviert in Bock (1974) hierarchische Clusteringverfahren mit Eigenschaften der Homogenität in Bezug auf partitionierende Clusteringverfahren, bei denen $C_{fin} := (C_1, C_2, \ldots, C_k)$ die Eigenschaften einer *Partition* (siehe Kapitel (5.3)) von D erfüllt. Dabei ist es offensichtlich, dass bei partitionierenden Verfahren (i) größere Homogenitätswerte der Cluster C_i durch eine größere Kardinalität der Menge C_{fin} erreicht werden können, und umgekehrt (ii) sich hohe Homogenitätswerte nur bei hinreichend großer Kardinalität von C_{fin} erreichen lassen. Prinzipiell kann man zwei Arten von partitionierenden Verfahren unterscheiden: (i) die Kardinalität der Menge C_{fin} ist vorgegeben oder (ii) die Homogenitätswerte der Cluster C_i werden von Anfang an durch Schranken gefordert. Dann ergibt sich im ersten Fall die Homogenität der Cluster durch das Verfahren selbst und im zweiten Fall ist k von der geforderten Ähnlichkeit innerhalb der Cluster abhängig. Da aber bei Clusteraufgaben die Zahl k und die Werte der Homogenitätsschranken in der Regel nicht bekannt sind, gelten beide der eben vorgestellten Möglichkeiten als nicht optimal. Hierarchische Clusteringverfahren versuchen dieses Problem dadurch zu lösen, dass sie eine Sequenz von Clusterungen erzeugen mit dem Ziel, dass die Homogenitätswerte der Cluster mit wachsendem k steigt. Weiterhin gilt nach Konstruktion dieser Verfahren, dass immer homogenere Cluster dadurch gebildet werden, dass größere Cluster in kleinere unterteilt werden und dass dieses Prinzip beliebig nach unten fortgesetzt wird. Generell werden bei hierarchischen Clusteringverfahren *divisive* (top-down) oder *agglomerative* (bottom-up) Clusteringverfahren unterschieden, wobei sich in der Praxis die agglomerativen Verfahren durchgesetzt haben. Chakrabarti (2002) gibt eine Vorschrift in *Pseudocode* an, aus der die wesentlichen Konstruktionsschritte von agglomerativen Verfahren leicht zu erkennen sind:

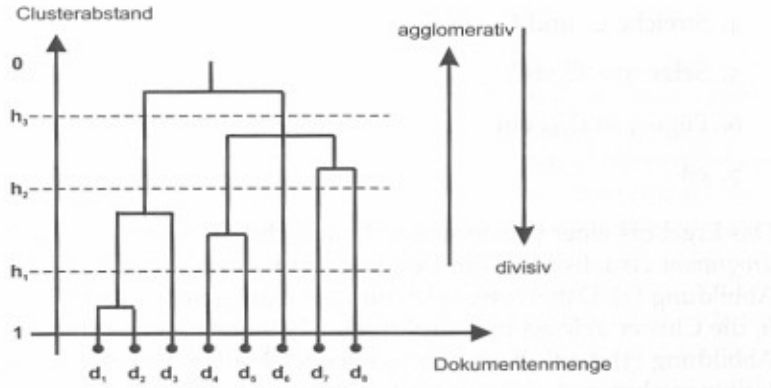

Abbildung 3: Dendrogramm für eine Clusteraufgabe mit acht Dokumenten. Die gestrichelten Linien deuten die gewählten Homogenitätsstufen an.

1. Die initiale und damit die feinste Partition von D ist $C_{fin} = \{C_1, C_2, \ldots, C_n\}$, wobei $C_i = \{d_i\}$.
2. `while` $|C_{fin}| > 1$ `do`
3. Wähle $C_i, C_j \in C_{fin}$ und berechne den Abstand $\alpha(C_i, C_j)$

4. Streiche C_i und C_j aus C_{fin}
5. Setze $\gamma = C_i \cup C_j$
6. Füge γ in C_{fin} ein
7. od

Das Ergebnis einer Clusterung mit hierarchischen Verfahren lässt sich als *Dendrogramm* visualisieren. Ein Dendrogramm einer fiktiven Clusterung zeigt die Abbildung (3). Dabei lassen sich nun auf jeder gewünschten Homogenitätsstufe h_i die Cluster ablesen und strukturell miteinander vergleichen. Man erkennt in Abbildung (3) deutlich ein auszeichnendes Merkmal eines agglomerativen Clusteringverfahrens: Auf der untersten Ebene stellen die Dokumente einelementige Cluster $\{d_1\}, \{d_2\}, \ldots, \{d_8\}$ dar; mit fallender Homogenität werden die Cluster auf den Ebenen immer gröber, bis sie zu einem einzigen verschmolzen werden, welches alle Dokumente enthält. Ein weiteres wichtiges Merkmal eines hierarchischen Clusteringverfahrens liegt darin, dass Dokumente, die auf der Basis eines Ähnlichkeitsmaßes als sehr ähnlich gelten, sehr früh zu einem Cluster verschmolzen werden. Das ist aber gleichbedeutend damit, dass der dazugehörige Homogenitätswert h_i im Dendrogramm nahe bei eins liegt. Weiterhin sind die Cluster auf den jeweiligen Homogenitätsstufen im Dendrogramm bezüglich ihrer inneren Struktur interpretierbar, da ein Cluster, das im Dendrogramm über mehrere Homogenitätsstufen in sich geschlossen bleibt, als sehr homogen angesehen werden kann. Wird dagegen ein Dokument erst im letzten oder vorletzten Schritt mit einem Cluster verschmolzen, so muss es auf Grund seiner Merkmale weniger ähnlich sein, als die Dokumente in einem sehr homogenen Cluster. Für das Ergebnis einer Clusteraufgabe, die mit einem hierarchischen Verfahren gelöst werden soll, ist aber auch die Güte der Daten, die Aussagekraft des zugrundeliegenden Ähnlichkeits- oder Abstandsmaßes und vor allen Dingen die Wahl des Maßes α entscheidend, um die Abstände $\alpha(C_i, C_j)$ zweier Cluster zu berechnen. Ausgehend von einem Ähnlichkeitsmaß $s : D \times D \longrightarrow [0,1]$ und den Clustern C_i und C_j, sind

$$\alpha_{SL}(C_i, C_j) := \min_{d,\tilde{d}} \{s(d,\tilde{d}) \mid d \in C_i, \tilde{d} \in C_j\} \text{ (Single Linkage)},$$

$$\alpha_{AL}(C_i, C_j) := \frac{1}{|C_i||C_j|} \sum_{\tilde{d} \in C_i} \sum_{d \in C_j} s(d,\tilde{d}) \text{ (Average Linkage)},$$

$$\alpha_{CL}(C_i, C_j) := \max_{d,\tilde{d}} \{s(d,\tilde{d}) \mid d \in C_i, \tilde{d} \in C_j\} \text{ (Complete Linkage)},$$

gängige Clusterabstände.

Zusammenfassend formuliert ist die übersichtliche und anschauliche Darstellbarkeit des Ergebnisses in Form eines Dendrogramms als positive Eigenschaft von hierarchischen Clusteringverfahren zu sehen. Das Dendrogramm, welches auch als *Baumstruktur* visualisiert werden kann, verlangt dabei nicht eine Clusteranzahl als Vorgabe, sondern auf jeder Ebene entsteht eine Anzahl von Clustern in natürlicher Weise. Weiterhin sind die einfache Implementation und die gute Interpretierbarkeit der entstehenden Cluster als Vorteile von hierarchischen Verfahren zu werten. Für Daten, bei denen eine hierarchische Struktur zu erwarten ist, sind hierarchische Clusteringverfahren besonders sinnvoll. Da in der Regel diese Kenntnis nicht vorhanden ist, muss das Dendrogramm für den jeweiligen Anwendungsfall interpretiert werden, da die hierarchische Struktur durch den Algorithmus erzwungen wird. Als Nachteil ist die Komplexität von hierarchischen Clusteringverfahren zu sehen, weil die Erzeugung der Ähnlichkeitsmatrix bereits quadratische Laufzeit besitzt und somit für Massendaten problematisch wird. Die Verwendung von verschiedenen Clusterabständen ist ebenfalls ein kritischer Aspekt, da Clusterabstände wie Single Linkage bzw. Complete Linkage oft die Tendenz zur Entartung haben, z.B. die Bildung von besonders großen bzw. kleinen Clustern.

5.3 Partitionierende Clusteringverfahren

In diesem Kapitel werden die Ziele und die grundlegende Wirkungsweise von partitionierenden Clusteringverfahren erläutert. Wieder ausgehend von der Dokumentenmenge D und einem Ähnlichkeitsmaß $s : D \times D \longrightarrow [0,1]$, bildet die Menge $C_{fin} := (C_1, C_2, \ldots C_k)$ eine partitionierende Clusterung von D, falls die Eigenschaften $C_i \cap C_j, i \neq j$ (Disjunktheit) und $\bigcup_{1 \leq i \leq k} C_i = D$ (volle Überdeckung der Menge D) erfüllt sind. Basierend auf der vorgegebenen Menge D, formulierte Bock Bock (1974) die Hauptaufgabe der partitionierenden Clusteringverfahren als die Suche nach einer disjunkten, also nicht überlappenden, Clusterung, die die obigen Eigenschaften einer Partition besitzt und die auszeichnenden Merkmale der Dokumente optimal wiederspiegelt. Weiterhin schlägt Bock (1974) Ansätze zur Lösung dieses Problems vor, z.B.:

- Bereitstellung von statistischen oder entscheidungstheoretischen Modellen, mit denen die noch unbekannten Cluster und deren Objekteigenschaften als Parameter behandelt und abgeschätzt werden können
- Einführung eines *Optimalitätskriteriums*, auf dem die *lokal optimale* Clusterung maßgeblich basiert

- Initiale Festlegung von Startclustern und anschließende Konstruktion der gesuchten Cluster
- Zuhilfenahme von daten- und anwendungsspezifischen Heuristiken

Bei partitionierenden Verfahren ist die finale Clusteranzahl k bei Beginn der Clusterung nicht bekannt und die Dokumente $d \in D$ werden ausgehend von gewählten Startclustern solange ausgetauscht, bis sich auf Grund eines Abbruchkriteriums eine möglichst lokal optimale Clusterung ergibt. Dagegen liegt bei der hierarchischen Clusterung auf jeder Hierarchiestufe eine eindeutige Menge von Clustern verfahrensbedingt vor, wobei diese Cluster nicht mehr aufgebrochen werden. Das in Theorie und Praxis bekannteste partitionierende Clusteringverfahren ist das $k-means$ Verfahren (Hastie et al. 2001), wobei es in verschiedenen Ausprägungen existiert, die sich meistens in der Art und Formulierung des Optimalitätskriteriums unterscheiden. Da k-means nur für quantitative Eingabedaten konzipiert ist, deren Abstände oft über die quadrierte *Euklidische Distanz* berechnet werden, eignet sich für das *Dokumentenclustering* eine Abwandlung von k-means, das $k-medoids$ Verfahren (PAM=Partitioning Around Medoids, cf. Han & Kamber (2001)). Anstatt von numerischen Startobjekten, die bei Beginn die Clusterzentren repräsentieren, wählt man in k-medoids Objekte (*Medoide*) aus D als Clusterzentren. Im weiteren Verlauf des Verfahrens werden lediglich die Ähnlichkeiten bzw. die Distanzen benötigt, um das Optimalitätskriterium, in Form einer Zielfunktion, und die neuen Medoids zu berechnen. Die wesentlichen Schritte von k-medoids, lassen wie sich nachfolgend formulieren, wobei davon ausgegangen wird, dass die Dokumente $d \in D$ in einer für das Clustering geeigneten Repräsentation vorliegen (Han & Kamber 2001):

1. Wähle zufällig k Dokumente als initiale Medoide und definiere damit die Menge M ($|M| = k$)
2. `while` (no change) `do`
3. Ordne jedes verbleibende Dokument dem nächsten Medoid zu (minimalem Abstand)
4. Wähle zufällig ein Dokument $d_r \in D$, das kein Medoid ist
5. Berechne auf der Basis eines Kostenkriteriums c die Gesamtkosten S des Austauschs von d_r mit dem aktuellen Medoid d_{act}
6. `if` c `then` tausche d_{act} mit d_r um eine neue Menge M von Medoiden zu bilden

7. `od`

Vorteile von partitionierenden Clusteringverfahren wie k-means und k-medoids sind ihr intuitiver Aufbau und die einfache Implementierbarkeit. Als Lösungen liefern solche Verfahren aber nur *lokale Optima*, da mit einer anderen Startkombination eventuell eine bessere Clusterlösung berechnet werden könnte. Um diesem Problem entgegenzuwirken, bietet sich entweder eine Kombination mit anderen Clusteringverfahren oder eine iterierte Anwendung an. Ein Nachteil von beiden Verfahren, k-means und k-medoids, ist offensichtlich die Vorgabe der initialen Clusterzahl k, da diese in der Regel unbekannt ist. Eine weitere Schwäche von k-means ist die mangelnde *Robustheit* des Verfahrens, das heißt das Verhalten bezüglich „Ausreißern", da bei der Berechnung der quadrierten Euklidischen Distanzen offensichtlich hohe Distanzwerte ermittelt werden und diese die Clusterbildung stark beeinflussen. Dagegen besitzt k-medoids eine schlechtere Komplexität in Bezug auf Massendaten, aber eine bessere Robustheit (Hastie et al. 2001).

5.4 Sonstige Clusteringverfahren

Bisher wurden die hierarchischen und partitionierenden Clusteringverfahren detaillierter vorgestellt, da diese Verfahren aus praktischen Gründen und auf Grund ihrer recht guten Interpretationsmöglichkeiten im Umfeld des Web Mining oft eingesetzt werden. In der Fachliteratur werden jedoch noch viele andere Clusteringverfahren behandelt, siehe z.B. Everitt et al. (2001), Fasulo (1999). Zwei werden im folgenden noch skizziert:

- *Graphentheoretische* Clusteringverfahren: Ausgehend von der Dokumentenmenge D und einem problemspezifischen Abstandsmaß (ein Ähnlichkeitsmaß kann leicht in ein Abstandsmaß umgewandelt werden) $d : D \times D \longrightarrow [0,1]$, wird eine Abstandsmatrix $(d_{ij})_{ij}$, $1 \leq i \leq n$, $1 \leq j \leq n$ induziert, wobei $d_{ij} \in [0,1]$. Diese Struktur kann, graphentheoretisch interpretiert, als ein kanten-markierter, vollständiger und ungerichteter Graph $G_D = (V_D, E_D, f_{E_D}, A_{E_D}), f_{E_D} : E_D \longrightarrow A_{E_D} := \{(d_{ij})_{ij}\}_{1 \leq i \leq n, 1 \leq j \leq n}$ betrachtet werden. Nun interessiert man sich für *Umgebungen* in denen, auf Grund der Abstandswerte d_{ij}, ähnliche Dokumente gruppiert werden und die Menge D somit auf diese Weise geclustert werden kann. Bock (1974) charakterisiert dieses Problem mit dem Begriff der d-Umgebung. Er versteht unter der d-Umgebung des Dokuments $d_k \in D$ die Menge der Dokumente $d_i \in D$, deren Abstandswerte die Ungleichung $d_{ik} \leq d$, $d > 0$ erfüllen. Genauer formuliert, definierte Bock ein Cluster $C \subseteq D$ als d-Cluster

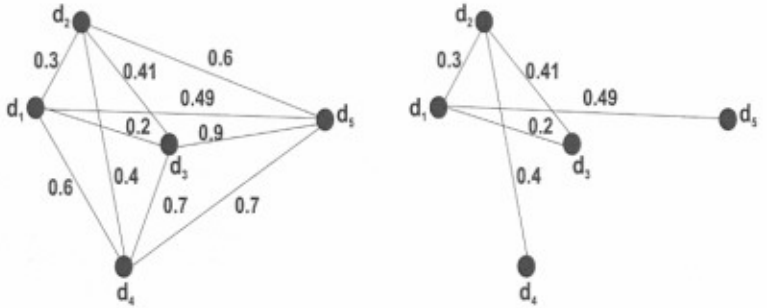

Abbildung 4: $|D| = 5$. Der vollständige Graph G_D und sein Teilgraph $G_D^{0.5}$

falls (i) $C \neq \{\}$, (ii) $\forall\ d_k \in C$ gehört auch die d-Umgebung von d_k zum d-Cluster dazu und (iii) kein Cluster $\tilde{C}$ mit $\tilde{C} \subseteq C$ darf die Eigenschaften (i) und (ii) erfüllen. Man betrachte nun denjenigen Teilgraph $G_D^d = (V_D, E_D^d)$, $E_D^d = E_D \backslash \{e = \{d_i, d_j\} | f_{E_D}(e) > d, \forall\, d_i, d_j \in V_D\}$ von G_D, für dessen Kantenmarkierungen die Ungleichungen $f_{E_D}(e) \leq d, \forall e \in E_D^d$ gelten. Bock bewies, dass die d-Cluster gerade die *Zusammenhangskomponenten* (Harary 1974) des Teilgraphen G_D^d von G_D sind. Die Abbildung (4) zeigt beispielhaft für eine Menge $D = \{d_1, d_2, \ldots, d_5\}$ mit gegebener Distanzmatrix den vollständigen Graph G_D und den Teilgraph $G_D^{0.5}$. Ein wichtiges und einfaches graphentheoretisches Konstruktionsmittel für die d-Cluster ergibt sich sofort aus dem *minimalen Spannbaum* von G_D. Dabei ist der minimale Spannbaum gerade der Teilgraph B_D mit den Eigenschaften: (i) B_D ist ein Baum (Harary 1974), (ii) B_D enthält alle Knoten aus G_D und (iii) die Summe seiner Kantenmarkierungen fällt minimal aus. Die Konstruktionsmethode des minimalen Spannbaums und die anschließende Gewinnung der d-Cluster wird ausführlich in Bock (1974) beschrieben. Weitere graphentheoretische Clusteringverfahren werden in Fasulo (1999) vorgestellt. Je nach Anwendungsfall werden auch *Dichte-basierte* Clusteringverfahren verwendet, die auf Grund ihrer Konstruktionsweise sehr verwandt zu graphentheoretischen Verfahren sind. Sie werden in Fasulo (1999), Han & Kamber (2001) näher beschrieben. Mehler (2002) stellt einen Algorithmus zur *perspektivischen* Clusterung ausgehend von so genannten Kohäsionsbäumen vor, die insbesondere der automatischen Textverlinkung dienen.

- *Probabilistische* Clusteringverfahren: Chakrabarti (2002) beschreibt Probleme des Clustering für webbasierte Dokumente in Bezug auf das Vektorraummodell. Algorithmen im *Web Information Retrieval* setzen voraus, dass die Elemente im Dokumentraum zufälligen Prozessen unterliegen, wobei die Verteilungen innerhalb der Dokumente zunächst nicht bekannt sind. Probabilistische Clusteringverfahren ordnen die Objekte mit einer bestimmten Wahrscheinlichkeit einem Cluster zu, dabei ist aber in der Regel die Verteilung der Objekte und die Anzahl der Cluster unbekannt. Ein sehr bekannter Algorithmus im Bereich der probabilistischen Clusteringverfahren ist der EM-Algorithmus (*Expectation Maximization*), der im Wesentlichen auf zwei Schritten beruht: (i) die Bestimmung der Clusterwahrscheinlichkeiten (Expectation) und (ii) die Parameterabschätzung der Verteilung mit dem Ziel, die Wahrscheinlichkeiten zu maximieren (Maximization). Der EM-Algorithmus wird, bezogen auf das Web Information Retrieval, ausführlich in Chakrabarti (2002) erklärt, wobei man weitere Überblicke in Everitt et al. (2001), Fasulo (1999) findet.

6 Ausblick

In diesem Artikel wurden Data Mining-Konzepte besprochen mit dem Ziel, sie auf bestehende und zukünftige Problemstellungen des Web Mining anzuwenden. Hierbei lag die besondere Betonung auf dem Web Structure Mining. Weiterhin wurden bestehende Arbeiten in der graphentheoretischen Analyse von Hypertextstrukturen besprochen.
Im Zuge der webbasierten Kommunikation wäre es für die zukünftige Entwicklung des Web Structure Mining sehr interessant, neuere Ergebnisse in den Bereichen

- Aufdeckung und bessere Beschreibung bestehender webbasierter Graphstrukturen,
- Fortschritte in der adäquaten und aussagekräftigen Modellierung Webbasierter Hypertexte, besonders in Hinsicht auf eine bessere Möglichkeit der inhaltsbasierten Kategorisierung sowie
- neuere und damit leistungsfähigere graphentheoretische Analysealgorithmen für hypertextuelle Graphstrukturen

zu gewinnen. Gerade in dem Umfeld des Web Structure Mining, wo mit graphentheoretischen Methoden und Data Mining-Verfahren Eigenschaften, Ausprägungen und sogar strukturelle Vergleiche hypertextueller Graphstrukturen

bestimmt werden, besteht besonderer Bedarf. Insbesondere sind damit graphentheoretische Methoden angesprochen, mit denen eine aussagekräftige Ähnlichkeitsgruppierung, z.B. auf der Basis spezifischer Eigenschaften oder auf der Graphstruktur selbst, möglich ist. Darauf basierend könnten einige anwendungsorientierte Problemstellungen, z.B. die strukturorientierte Filterung und Fragen bezüglich zeitlich bedingter struktureller Veränderungen webbasierter Hypertextstrukturen, besser gelöst werden. Dabei werden einige der Clusteringverfahren, die im Kapitel (5) vorgestellt wurden, zur Lösung solcher Aufgaben beitragen. Betrachtet man aber die Anzahl der heute vorliegenden Clusteringverfahren so erscheint die Auswahl eines geeigneten Verfahrens für den gewünschten Anwendungsfall jedoch nicht leicht. Die Auswahl sollte sich auf jeden Fall an den vorliegenden Daten, am zugrundeliegenden Ähnlichkeitsmaß und an der geplanten Weiterverwendung einer Clusterlösung orientieren. Zur Interpretation einer Clusterlösung sind in Kapitel (5.1) mathematische Verfahren vorgestellt worden. In Hinsicht auf die Clusterung strukturell ähnlicher webbasierter Hypertextstrukturen ist es denkbar, z.B. auch visuelle oder anwendungsbezogene Kriterien als zusätzliche Gütekennzeichen einer Clusterlösung zu definieren. Somit stellt eine Clusterlösung dann kein isoliert betrachtetes Ergebnis dar, sondern dient als Grundlage für die oben skizzierten Anwendungen im Web Structure Mining.

Literatur

Backhaus, K., Erichson, B., Plinke, W., & Weiber, R. (2003). *Multivariate Analysemethoden.* Springer.

Bock, H. H. (1974). *Automatische Klassifikation. Theoretische und praktische Methoden zur Gruppierung und Strukturierung von Daten.* Studia Mathematica - Mathematische Lehrbücher, Vandenhoeck & Ruprecht Verlag.

Botafogo, R., Rivlin, E., & Shneiderman, B. (1992). Structural analysis of hypertexts: Identifying hierarchies and useful metrics. *ACM Transactions on Information Systems, 10*(2), 142–180.

Botafogo, R. A. & Shneiderman, B. (1991). Identifying aggregates in hypertext structures. In *Proc. of the 3th annual ACM conference on Hypertext*, (pp. 63–74).

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., & Wiener, J. (2000). Graph structure in the web: Experiments and models. In *Proc. of the 9th World Wide Web Conference.*

Chakrabarti, S. (2001). Integrating the document object model with hyperlinks for enhanced topic distillation and information extraction. In *Proc. of the 10th International World Wide Web Conference*, (pp. 211–220).

Chakrabarti, S. (2002). *Mining the Web: Discovering Knowledge from Hypertext Data.* San Francisco: Morgan Kaufmann.

Charney, D. (1987). Comprehending non-linear text: The role of discourse cues and reading strategies. In *Proc. of the ACM conference on Hypertext, Hypertext'87*, (pp. 109–120).

De Bra, P. & Houben, G. J. (1997). Hypertext metrics revisited: Navigational metrics for static and adaptive link structures. `http://citeseer.ist.psu.edu/139855.html` (seen 05/2005).

Dehmer, M., Gleim, R., & Mehler, A. (2004). A new method of measuring similarity for a special class of directed graphs. Tatra Mountains Mathematical Publications, Submitted for publication.

Dehmer, M., Mehler, A., & Gleim, R. (2004). Aspekte der Kategorisierung von Webseiten. In *GI-Edition - Lecture Notes in Informatics (LNI) - Proceedings, Jahrestagung der Gesellschaft für Informatik*, (pp. 39–43).

Deo, N. & Gupta, P. (2001). World Wide Web: A graph-theoretic perspective. Technical report, Computer Science Technical report, University of Central Florida.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern Classification.* Wiley - Interscience.

Ehud, R., Botafogo, R. A., & Shneiderman, B. (1994). Navigating in hyperspace: Designing a structure-based toolbox. *Commun. ACM, 37*(2), 87–96.

Emmert-Streib, F., Dehmer, M., & Kilian, J. (2005). Classification of large graphs by a local tree decomposition. In *to appear in in: Proceedings of DMIN'05, International Conference on Data Mining, In conjunction with: World Congress in Applied Computing 2005, Las Vegas/USA.*

Everitt, B. S., Landau, S., & Leese, M. (2001). *Cluster Analysis.* Arnold Publishers.

Fasulo, D. (1999). An analysis of recent work on clustering algorithms. Technical report, Technical Report 01-03-02, University of Washington, Seatle/USA.

Ferber, R. (2003). *Information Retrieval.* dpunkt.Verlag.

Fürnkranz, J. (2001). Hyperlink ensembles: A case study in hypertext classification. Technical report, University Vienna, Technical Report No. OEFAI-TR-2001-30.

Halasz, F. G. (1987). Reflections on notecards: Seven issues for the next generation of hypermedia systems. In *Proc. of the ACM conference on Hypertext, Hypertext'87*, (pp. 345–366).

Han, J. & Kamber, M. (2001). *Data Mining: Concepts and Techniques.* Morgan and Kaufmann Publishers.

Harary, F. (1965). *Structural models. An introduction to the theory of directed graphs.* Wiley, New York.

Harary, F. (1974). *Graphentheorie.* Oldenbourg Verlag.

Hastie, R., Tibshirani, R., & Friedman, J. H. (2001). *The Elements of Statistical Learning*. Springer.

Hofmann, M. (1991). *Benutzerunterstützung in Hypertextsystemen durch private Kontexte*. PhD thesis, Springer.

Horney, M. (1993). A measure of hypertext linearity. *Journal of Educational Multimedia and Hypermedia*, 2(1), 67–82.

Kleinberg, J. M. (1998). Authoritative sources in a hyperlinked environment. In *Proceedings of the 9th annual ACM-SIAM Symposium on Discrete Algorithms*, (pp. 668–677).

Kosala, R. & Blockeel, H. (2000). Web Mining Research: A survey. *SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining*, 2.

Kuhlen, R. (1991). *Hypertext - Ein nicht-lineares Medium zwischen Buch und Wissensbank*. Springer.

Lobin, H. (1999). *Text im digitalen Medium. Linguistische Aspekte von Textdesign, Texttechnologie und Hypertext Engineering*. Westdeutscher Verlag.

McEneaney, J. E. (2000). Navigational correlates of comprehension in hypertext. In *Proc. of the ACM conference on Hypertext*, (pp. 251–255).

Mehler, A. (2001). *Textbedeutung. Zur prozeduralen Analyse und Repräsentation struktureller Ähnlichkeiten von Texten*. Peter Lang, Europäischer Verlag der Wissenschaften.

Mehler, A. (2002). Hierarchical orderings of textual units. In *Proc. of COLING'02, Taipeh/Taiwan*.

Mehler, A., Dehmer, M., & Gleim, R. (2004). Towards logical hypertext structure. a graph-theoretic perspective. In *Proc. of I2CS'04, Guadalajara/Mexico*.

Noller, S., Naumann, J., & Richter, T. (2002). Logpat - Ein webbasiertes Tool zur Analyse von Navigationsverläufen in Hypertexten. `http://www.psych.uni-goettingen.de/congress/gor-2001` (seen 05/2005).

Oren, T. (1987). The architecture of static hypertext. In *Proc. of the ACM conference on Hypertext, Hypertext'87*, (pp. 291–306).

Rahm, E. (2000). Web Usage Mining. *Datenbank-Spektrum*, 2(2), 75–76.

Schlobinski, P. & Tewes, M. (1999). Graphentheoretische Analyse von Hypertexten. `http://www.websprache.uni-hannover.de/networx/docs/networx-8.pdf` (seen 05/2005).

Shneiderman, B. & Kearsley, G. (1989). *Hypertext Hands On!: An introduction to a new way of organizing and accessing information*. Addison Wesley.

Storrer, A. (1999). Kohärenz in Text und Hypertext. In L. H. (Ed.), *Text im digitalen Medium. Linguistische Aspekte von Textdesign, Texttechnologie und Hypertext Engineering* (pp. 33–65). Wiesbaden/Germany: Westdeutscher Verlag.

Storrer, A. (2004). Text und Hypertext. In L. H. (Ed.), *Texttechnologie. Perspektiven und Anwendungen*. Wiesbaden/Germany: Stauffenburg Verlag.

Unz, D. (2000). *Lernen mit Hypertext. Informationsuche und Navigation*. Waxmann Verlag.

Winne, P. H., Gupta, L., & Nesbit, L. (1994). Exploring individual differences in studying strategies using graph theoretic statistics. *The Alberta Journal of Educational Research*, 40, 177–193.

Winter, A. (2002). Exchanging Graphs with GXL. `http://www.gupro.de/GXL` (seen 05/2005).