UNIVERSITÄT KOBLENZ · LANDAU
Fachbereich 4: Informatik

<is web>
Arbeitsgruppe ISWeb

Warum Wer Wen kennt. Eine themenspezifische Auswertung sozialer Netzwerke.

Diplomarbeit

zur Erlangung des Grades eines

Diplom-Informatikers
im Studiengang Informatik

vorgelegt von

René Henkes

Betreuer:	Dipl.-Inform. Klaas Dellschaft, Institut für Informatik, Fachbereich Informatik
Erstgutachter:	Dipl.-Inform. Klaas Dellschaft, Institut für Informatik, Fachbereich Informatik
Zweitgutachter:	Prof. Dr. Steffen Staab, Institut für Computervisualistik, Fachbereich Informatik

Koblenz, im März 2008

Erklärung

Ich versichere, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe und dass die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen hat und von dieser als Teil einer Prüfungsleistung angenommen wurde. Alle Ausführungen, die wörtlich oder sinngemäß übernommen wurden, sind als solche gekennzeichnet.

Mit der Einstellung dieser Arbeit in die Bibliothek bin ich einverstanden. ja ☐ nein ☐

Der Veröffentlichung dieser Arbeit im Internet stimme ich zu. ja ☐ nein ☐

Koblenz, den 18. März 2008 Unterschrift

Für

Nadja

Inhaltsverzeichnis

Verzeichnis der Bilder

Verzeichnis der Tabellen

Kapitel 1

Einleitung

1.1 Abstract

In unserer heutigen Welt spielen soziale Netzwerke eine immer größere werdende Rolle. Im Internet entsteht fast täglich eine neue Anwendung in der Kategorie *Web 2.0*. Aufgrund dieser Tatsache wird es immer wichtiger die Abläufe in sozialen Netzwerken zu verstehen und diese für Forschungszwecke auch simulieren zu können.

Da alle gängigen sozialen Netzwerke heute nur im eindimensionalen Bereich arbeiten, beschäftigt sich diese Diplomarbeit mit mehrdimensionalen sozialen Netzwerken. Mehrdimensionale soziale Netzwerke bieten die Möglichkeit verschiedene Beziehungsarten zu definieren. Beispielsweise können zwei Akteure nicht nur in einer *'kennt'*-Beziehung stehen, sondern diese Beziehungsart könnte auch in diverse Unterbeziehungsarten, wie z.B. Akteur A *'ist Arbeitskollege von'* Akteur B oder Akteur C *'ist Ehepartner von'* Akteur D, unterteilt werden. Auf diese Art und Weise können beliebig viele, völlig verschiedene Beziehungsarten nebeneinander existieren.

Die Arbeit beschäftigt sich mit der Frage, in welchem Grad die Eigenschaften von eindimensionalen auch bei mehrdimensionalen sozialen Netzwerken gelten. Um das herauszufinden werden bereits bestehende *Metriken* weiterentwickelt. Diese Metriken wurden für eindimensionale soziale Netzwerke entwickelt und können nun auch für die Bewer-

tung mehrdimensionaler sozialer Netzwerke benutzt werden. Eine zentrale Fragestellung ist hierbei wie gut sich Menschen finden, die sich etwas zu sagen haben.

Um möglichst exakte Ergebnisse zu erhalten, ist es notwendig reale Daten zu verwenden. Diese werden aus einem Web 2.0-Projekt, in das Benutzer Links zu verschiedenen Themen einstellen, gewonnen (siehe Kapitel 4). Der erste praktische Schritte dieser Arbeit besteht daher darin, das soziale Netzwerk einzulesen und auf diesem Netzwerk eine Kommunikation, zwischen zwei Personen mit ähnlichen Themengebieten, zu simulieren. Die Ergebnisse der Simulation werden dann mit Hilfe der zuvor entwicklelten Metriken ausgewertet.

1.2 Aufbau der Arbeit

In **Kapitel 2** wird der aktuelle Stand der Forschung im Themenbereich eindimensionale soziale Netzwerke dargestellt, soweit er für diese Arbeit von Interesse ist. Diese Vorgehensweise führt zu einem besseren Verständis der folgenden Kapitel, auch für nicht mit dem Thema vertraute Personen. Eine Abgrenzung zu meiner eigenen Forschungsarbeit wird dadurch ebenfalls gewährleistet.

In **Kapitel 3** werden Begriffe und Strategien eingeführt und erklärt, die in der weiteren Arbeit zur Anwendung kommen sollen. Dabei handelt es sich um lokale wie globale Metriken für eindimensionale Netzwerke, die bereits in anderen Arbeiten beschrieben wurden.

In **Kapitel 4** werden Metriken neu entwickelt, um mit diesen themenspezifische Ergebnisse zu erhalten. Des Weiteren werden die Probleme und deren Lösungen beschrieben, die sich bei der praktischen Umsetzung des gesamten Projekts ergeben haben.

In **Kapitel 5** sollen die Experimente und deren Ergebnisse dargestellt und diskutiert werden.

In **Kapitel 6** werden die Ergebnisse dieser Arbeit zusammengefasst.

Auf der beigelegten **CD** ist der JAVA-Quellcode zum Einlesen des Datensatzes, zur Simulation der Kommunikation und der Implementierung der Metriken für die Netzwerkauswertung vorhanden. Des Weiteren befinden sich die Quelldaten des sozialen Netzwerks

und Dateien mit Anfragen an selbiges auf der CD.

Kapitel 2

Stand der Wissenschaft

In diesem Kapitel wird der aktuelle Stand der Forschung im Themenbereich eindimensionale soziale Netzwerke dargestellt, soweit er für diese Arbeit von Interesse ist. Diese Vorgehensweise führt zu einem besseren Verständis der folgenden Kapitel, auch für nicht mit dem Thema vertraute Personen. Eine Abgrenzung zu meiner eigenen Forschungsarbeit wird dadurch ebenfalls gewährleistet.

Zu Beginn dieses Kapitels werden Kleine-Welt-Netzwerke näher beleuchtet, damit diese später hin zur Mehrdimensionalität erweitert werden können. Besonders die Eigenschaften, die soziale Netzwerke zu Kleine-Welt-Netzwerken werden lassen, werden herausgearbeitet, denn Kleine-Welt-Netzwerke sind der Netzwerktyp, nach deren Aufbauprinzip real existierende Personen miteinander verbunden sind.

Auch verschiedene Modelle zur automatischen Erzeugung von Kleine-Welt-Netzwerken werden in diesem Kapitel auf ihre Tauglichkeit hin überprüft. Diese Tauglichkeit stellt sich durch eine möglichst große Nähe zur Realität dar. Denn nur mit guten Modellen kann man gute Ergebnisse erhalten, welche man benötigt, um aussagekräftige Thesen aufstellen oder widerlegen zu können. Die theoretischen Zusammenhänge werden mit Hilfe von Beispielen aus dem ganz normalen Alltag leichter verständlich gemacht.

Wenn das soziale Netzwerk einmal vorhanden ist, muss eine Strategie entwickelt werden, mit der man Personen finden kann, die Fragen zu verschiedenen Themengebieten möglichst umfassend und schnell beantworten können, denn das ist der eigentliche Sinn

von sozialen Netzwerken. Für diesen Zweck werden Strategien aufgezeigt, die gewählt werden können, um solche Personen zu finden [BCK+07]. Der zu diesem Zweck bereits entwickelte REMINDIN-Algorithmus [Tem06] wird vorgestellt und es wird hinterfragt ob dieser Algorithmus, die Realität abbildet und er somit für die vorliegende Arbeit einsetzbar ist.

Im nächsten Schritt werden bereits entwickelte Maßzahlen eingeführt, mit denen man Eigenschaften eindimensionaler sozialer Netzwerke genau messen und so auch verschiedene Netzwerke bewerten und vergleichen kann. Diese Maßzahlen sollen später auch auf mehrdimensionale soziale Netzwerke angewandt werden, um feststellen zu können ob, und wenn ja - wo, es signifikante Unterschiede zwischen ein- und mehrdimensionalen sozialen Netzwerken gibt.

2.1 Kleine-Welt-Netzwerke

Um ein reales soziales Netzwerk zu simulieren, muss man wissen, wie ein solches Netzwerk aufgebaut ist.

Eine Studie aus der Soziologie von Stanley Milgram [Mil67] hat gezeigt, dass solche Netzwerke immer nach dem gleichen Prinzip, dem Kleine-Welt-Phänomen (Small World Phenomenon) funktionieren. Dieses Phänomen beschreibt die Tatsache, dass jede Person einen ähnlich aufgebauten Bekanntenkreis hat. Der Aufbau besteht darin, dass die Person sehr viele Bekannte aus ihrem nahen persönlichen Umfeld hat, z.B. Personen aus dem selben Ort, von der gleichen Schule oder der selben Arbeitsstelle. Dagegen pflegen die meisten Menschen nur wenige Kontakte zu Freunden bzw. Bekannten, die weiter von ihnen entfernt leben. In diesen Personenkreis würden beispielsweise Urlaubsbekanntschaften oder Personen, die aus dem nahen Umfeld weggezogen sind, fallen. Diese Kontakte werden als '*long-range-Kontakte*' bezeichnet.

Das diese Art der Verteilung der Realität entspricht, kann man beispielsweise auf der Internetplattform wer-kennt-wen[1] nachvollziehen auf der reale Personen ihre Bekanntschaften pflegen können. Dort wird auch eine Karte mit den Wohnorten aller Bekanntschaften

[1] www.wer-kennt-wen.de

erzeugt, welche die Studie empirisch bestätigt. Es wird auch eine weitere von Milgram beschriebene Eigenschaft eines Kleine-Welt-Netzwerkes sichtbar, nämlich die Clustering-Eigenschaft sozialer Netzwerke. Diese Eigenschaft besagt, dass es eine große Deckungsgleichheit der Bekanntenkreise zweier Personen gibt, die sich selbst kennen. Mit einem Beispiel aus der Realität lässt sich auch diese Eigenschaft schnell verdeutlichen. Wenn sich zwei Personen aus ihrer ehemaligen Schulklasse kennen, dann kennt auch jeder von ihnen seine Mitschüler. Da es sich bei diesen Mitschülern aber um die gleichen Personen handelt, überschneidet sich der Bekanntenkreis der beiden Personen sehr stark.

Über dieses Beziehungsgeflecht, besonders über die '*long-range-Kontakte*', kann man nun Pfade von einer beliebigen Startperson zu einer beliebigen Zielperson herausfinden. Die Länge des Pfades ist in einem Kleine-Welt-Netzwerk immer polylogarithmisch. Das bedeutet, dass bei Anzahl von *n* Personen bzw. Knoten jeder andere Knoten mit log n Schritten erreicht werden kann. [Kle06]

Dabei ist es die herausragende Eigenschaft von Small World-Netzwerken, dass die kürzest möglichen Pfade in den seltensten Fällen länger sind als sechs Personen. Daher ist das Small Word Phänomen auch unter dem Begriff „Six Degrees of Separation "bekannt. [Mil67]

2.1.1 Modelle

Das Modell von Watts und Strogatz

Das Modell, dass von Watts und Strogatz zur Simulation eines sozialen Netzwerkes entwickelt wurde ist folgendermaßen aufgebaut. Alle Personen haben einen Kontakt zu ihrem jeweiligen Nachbarn. Des Weiteren haben sie auch einen Kontakt zum Nachbarn ihres Nachbarn. Da es sich um ein eindimensionales Modell handelt, ergeben sich also für jede Person vier Kontakte. Dies gilt auch für die erste und die letzte Person in der Reihe, da alle Personen in einem Ring angeordnet sind. Nun wird die Zufallszahl *p* eingeführt diese besitzt einen Wert zwischen Null und eins. Die Zufallszahl *p* bestimmt die Wahrscheinlichkeit mit der ein Kontakt zwischen einer Start- und einer Zielperson aufgelöst wird, um dann einen neuen Kontakt zwischen der selben Startperson und einer anderen Zielperson

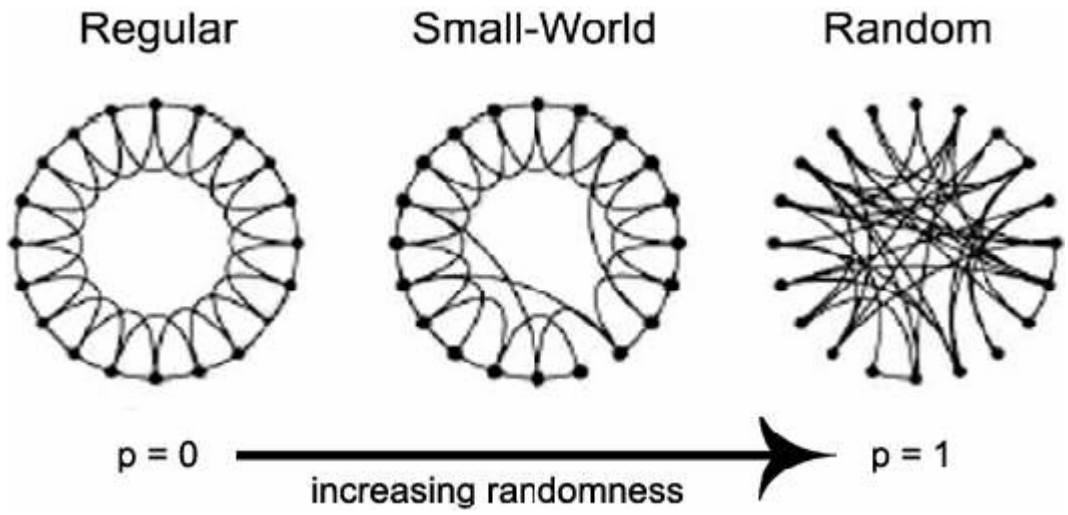

Bild 2.1: Links: Regulärer Graph (ohne long-range-Kontakte); Mitte: Kleine-Welt-Netzwerk (mit einigen long-range Kontakten); Rechts: Zufallsgraph (mit sehr vielen long-range Kontakten) [WS98]

herzustellen. Siehe Bild 2.1

Daher kann schon ein relativ kleiner Wert für *p* ein Kleine-Welt-Netzwerk erzeugen, wie es in der Realität vorkommt. Dabei stellen die Kanten, die am Anfang vorhanden waren, die räumlich nahen Kontakte dar und die neu erstellten Kanten bilden die '*long-range-Kontakte*'.

Doch Watts und Strogatz haben einen Fakt aus der realen Welt in ihrem Modell nicht berücksichtigt. Es werden nicht nur kurze und long-range-Kontakte zwischen einzelnen Personen geknüpft, sondern die long-range-Kontakte unterscheiden sich auch untereinander. Das bedeutet, dass es in der Realität viel wahrscheinlicher ist, dass eine Person Kontakte zu einer anderen hat, die nur 100 Kilometer entfernt lebt, als zu einer Person die 500 Kilometer oder noch weiter entfernt lebt. Daher ist der Aufwand den kürzesten Pfad zu finden auch nicht $\log(n)$, wie es verlangt wird, damit ein Kleine-Welt-Netzwerk entsteht, sondern $n^{2/3}$, also wesentlich höher. [Kle06]

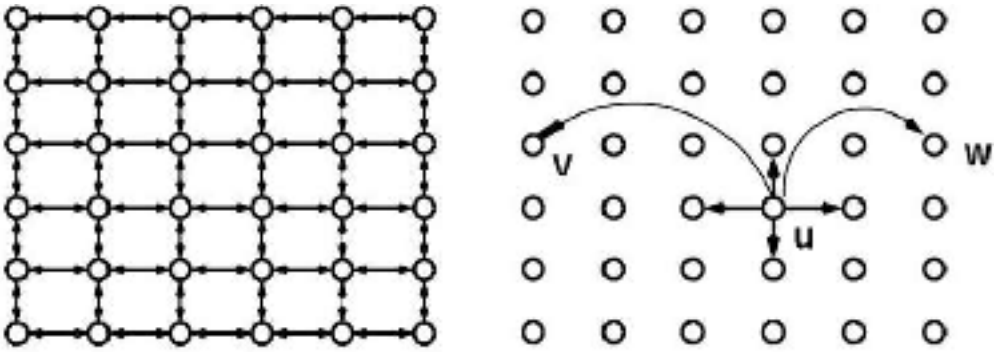

Bild 2.2: Gitterbasiertes Modell nach Kleinberg mit lokalen Kontakten (links) und '*long-range-Kontakten*' (rechts). [Gru06]

Gitterbasiertes Modell

Aus dem oben beschriebenen Grund hat Jon Kleinberg [Kle06] dem Modell einen zusätzlichen Parameter hinzugefügt. Dieser Parameter α sorgt dafür, dass long-range-Kontakte mit zunehmender Länge immer seltener werden.

Das Modell basiert, wie in Bild 2.2 gezeigt, auf einer Anordnung der Knoten in einer $n \times n$-Matrix. Jeder der Knoten erhält kurze Kontakte zu seinen vier Nachbarn. Des weitern erhält jeder Knoten keinen oder einen long-range-Kontakt. Die Entscheidung, ob ein Knoten *u* einen long-range-Kontakt zu einem Knoten *v* erhält, hängt dabei von deren Entfernung und vom bereits erwähnten Parameter α ab. Dabei wird die Entfernung über die Cityblockdistanz berechnet. Diese Entfernung wird mit $-\alpha$ potenziert.

$$(|u_1 - v_1| + |u_2 - v_2|)^{-\alpha}$$

Die Zahl, die sich dann ergibt, wird als Wahrscheinlichkeit betrachtet, mit welcher der '*long-range-Kontakt*' erstellt wird. Bei gegebenem α ergibt sich daraus, dass die Wahrscheinlichkeit, mit der ein '*long-range-Kontakt*' erstellt wird immer geringer wird, je länger dieser long-range-Kontakt sein soll. Daher kommt Kleinbergs Modell der Realität sehr nahe.

Kleinberg zeigt in seiner Arbeit auch, dass $\alpha = 2$ der ideale Wert ist, um ein reales Netzwerk zu simulieren.

2.2 Hubs und Authorities

Hubs sind in einem sozialen Netzwerk Personen, die sehr viele andere Menschen kennen. Im Vergleich dazu selbst allerdings von wenig anderen Personen gekannt werden. Authorities hingegen sind in einem sozialen Netzwerk Personen, die von sehr vielen anderen Personen gekannt werden, selbst aber relativ wenige andere Personen kennen. In der realen Welt trifft diese Beschreibung vor allem auf Prominente zu. Beispiele hierfür wären der *Papst* oder der *US-Präsident*. Ein Beispiel aus der realen Welt für einen Hub lässt sich hingegen nur konstruieren, da reale Bekanntschaften im Allgemeinen nur bidirektional sind. Es wäre aber möglich z.B. den Besitzer einer Modelagentur, der seine Kunden in einer Datenbank vorrätig hat und diese damit 'kennt' als Hub zu betrachten, denn die Models kennen den Besitzer nicht unbedingt.

2.3 Anfragerouting

In einem sozialen Netzwerk gibt es verschiedene Arten des Routings, also dem Finden von Wegen auf denen Informationen über das Netz möglichst schnell und praktisch gefunden bzw. verbreitet werden können. Bei der Problemstellung, die in dieser Diplomarbeit zu behandeln ist, muss der erste Schritt um eine Antwort auf eine Frage zu finden darin bestehen eine Person ausfindig zu machen, die mit möglichst hoher Wahrscheinlichkeit eine Antwort zur gestellten Frage geben kann.

Um nun eine solche Person zu finden, beschreiben Gerhard Weikum et al. [BCK+07] verschiedene Vorgehensweisen. Diese Verfahren zum Anfragerouting sind dabei folgendermaßen benannt:

- **Semantisches Anfragerouting:** Hier werden gezielt Personen befragt, bei denen man ausgehend von ihren bisherigen Interessen und Betätigungen davon ausgehen kann, dass sie auch zu der gestellten Anfrage eine brauchbare Antwort geben können.

  Es können auch Personen ausgewählt werden, die einen engen Kontakt zu einer oder mehreren Personen haben, die die obigen Ansprüche erfüllen.

- **Soziales Anfragerouting:** Bei diesem Verfahren werden gezielt die engen Kontakte der fragenden Person genutzt, um die Anfrage zu verbreiten.

- **Spirituelles Anfragerouting:** Mit dieser Art des Suchens soll nach dem Willen der Autoren ein Bruder im Geiste gefunden werden. Das bedeutet, dass eine bisher unbekannte Person ermittelt werden soll, die der anfragenden Person dem Verhalten nach ähnlich ist.

2.3.1 Semantisches Anfragerouting

Die technische Realisierung erfolgt über ein Maß an Ähnlichkeit zwischen dem Inhalt der Anfrage und der dem Zielknoten zugeordneten Daten. Die Anfrage besteht hier aus Schlüsselwörtern, die aus ihrem Inhalt hervorgehen. In der Fachsprache des Information Retrievals wird jedes dieser Schlüsselwörter auch als Term bezeichnet. Die Daten des Zielknotens werden sowohl in der Form von Termen bereitgestellt, aber auch über sogenannte Tags. Bei diesen handelt es sich um zusätzliche Informationen, die den eigentlichen Daten hinzugefügt wurden. Beispielsweise könnten die Tags zu einem Zeitungsartikel Erscheinungsdatum, Zeitung, Autor bzw. Qualität sein. Um die Trefferquote zu erhöhen, sollten auch noch Statistiken erhoben werden, in denen z.B. darauf eingegangen wird, welcher Zielknoten eine Anfrage zu einem konkreten Term oder Tag wie oft zufriedenstellend beantwortet hat. Die besten Zielknoten sollten hier vorrangig benutzt werden.

2.3.2 Soziales Anfragerouting

Das Finden von engen Kontakten oder Freunden kann über die Mitgliedschaft in den gleichen Gruppen geschehen oder über eine explizite Freundesliste, wie es sie in bereits bestehenden sozialen Netzwerken, wie z.B. StudiVZ gibt. Ob ein Kontakt eng ist oder nicht könnte man auch aus der Zahl der bereits von zwei Knoten untereinander getätigten Anfragen schließen.

2.3.3 Spirituelles Anfragerouting

Die technische Realisierung geschieht über das Messen von Ähnlichkeiten im Gebrauch von Tags, der Anzahl der Kommentare, der Bewertungen bzw. über die Ähnlichkeit der Dokumente, auf die der Benutzer ein Lesezeichen gesetzt hat.

2.3.4 Welches Verfahren ist geeignet?

Experimente von Matthias Bender, Tom Crecelius et al. haben gezeigt, dass das semantische auf Tags basierende Anfragerouting dem sozialen sowie dem spirituellen weit überlegen ist.

Da es das Ziel dieser Diplomarbeit ist ein soziales Netzwerk nachzubilden, dass einem realen existierenden sozialen Netzwerk nahe kommt, sollten für das zu simulierende Netzwerk trotzdem alle drei Varianten des Anfragerroutings berücksichtigt werden.

Dies ist aus mehreren Gründen sinnvoll. Da die Knoten des Netzwerks verschiedene reale Personen repräsentieren, ist davon auszugehen, dass sich diese verschiedenen Personen auch bevorzugt verschiedener Anfragstrategien bedienen.

Des Weiteren ist es in der Realität nicht immer sinnvoll mit Hilfe des semantischen Anfrageroutings den besten Spezialisten auf diesem Fach zu suchen, da beispielsweise die Fragestellung doch recht einfach ist oder die Spezialisten so überlastet sind, dass sie Anfragen nur schleppend oder gar nicht bearbeiten. Daher wäre es durchaus sinnvoll sich mit weniger guten Experten zufriedenzugeben, da einer von diesen aufgrund der größeren Anzahl schneller zu finden ist. Sollte man doch eine Person mit größerem Fachwissen benötigen, kann man auch weiter suchen.

Wählt man hingegen die Person unter seinen Freunden oder seinen Brüdern im Geiste aus, die auf dem Gebiet der Anfrage die besten Kenntnisse hat, so kann man evtl. schneller zu einem brauchbaren Ergebnis kommen, da die Kenntnisse des Freundes evtl. schon ausreichen, er nicht so überlastet ist und es auch in seinem eigenen Interesse ist die Anfrage zu beantworten, da er im Gegenzug auf Hilfe hoffen kann, wenn er selbst einmal eine Anfrage stellen sollte.

2.4 Die Algorithmen REMINDIN und INGA

Beim REMINDIN-Algorithmus handelt es ich um einen, in der Dissertation von Christoph Tempich entwickelten Algorithmus [Tem06]. Im Zuge dieser Dissertation wurde der Algorithmus auch implementiert. Der (**I**nterest-based **N**ode **G**rouping **A**lgorithms)-Algorithmus stellt eine Weiterentwicklung zu REMINDIN-Algortihmus dar [LTQ+05]. Beide dienen dazu das Anfrageverhalten realer Personen auf digitalisierten, sozialen Netzwerken zu simulieren [Tem06] [LTQ+05].

Dabei wird das soziale Netzwerk mit Hilfe eines Peer2Peer-Netzwerks dargestellt. Eine reale Person wird dabei durch genau einen Peer repräsentiert. Zu Begin besitzen in einem solchen Netzwerk alle Peers die gleiche Priorität, denn alle Peers besitzen auf allen Wissensgebieten noch keine Fähigkeiten. Durch eine Initialisierung wird eine Startkonfiguration erzeugt, die der Realität nicht unähnlich ist. Durch diese Konfiguration soll dargestellt werden, dass einige Peers mehr bzw. anderes Wissen besitzen als andere. Dabei wird davon ausgegangen, dass eine Person sich nur für eine überschaubare Anzahl von Themen interessiert [Tem06]. Mit dieser Einstellung wird verhindert, dass jede Person nach kurzer Zeit fast alles Wissen besitzt.

Durch die eigentliche Anwendung des REMINDIN- bzw. des INGA-Algorithmus ergeben sich allerdings einige Personen, die von Ihren Mitmenschen als Experten auf bestimmten Fachgebieten wahrgenommen werden und deshalb häufiger von anderen Personen kontaktiert werden.

In Teil III seiner Arbeit entwickelt Christoph Tempich die folgenden sechs Punkte (soziale Metaphern), welche die Überlegungen realer Personen repräsentieren, an welche andere Person in ihrem sozialen Umfeld sie ihre Anfrage am besten richten, um eine befriedigende Antwort zu erhalten (freie Übersetzung nach Tempich [Tem06] S. 162f)

1. Eine Anfrage wird an die Person gestellt, von der man annimmt, dass sie die Frage am besten beantworten kann. Wobei in der vorliegenden REMINDIN-Version die Person als am besten betrachtet wird, die das größte Fachwissen besitzt. Andere Eigenschaften, wie Zuverlässigkeit oder Kosten der Anfrage werden noch nicht berücksichtigt.

2. Eine Person wird als möglicher Wissender in einem bestimmten Fachbereich eingestuft, wenn sie bereits früher Anfragen aus dem gleichen Themengebiet beantwortet hat.
3. In einer generellen Annahme wird davon ausgegangen, dass sich Personen, die sich in einem bestimmten Themengebiet gut auskennen, sich auch in ähnlichen, z.B. einem generellerem Themengebieten auskennen.
4. Eine Person, die viele andere Personen kennt hat gute Chancen, dass sich darunter jemand befindet, der die Frage, die an sie gestellt wurde, beantworten kann.
5. Wenn eine Person (A) eine andere Person (B) fragt, wird sich B die Person A und ihre Frage merken, auch wenn er die Frage nicht beantworten konnte. Sollte später einmal der Fall eintreten, dass B die gleiche Frage beantwortet haben möchte, die A ihm vor längere Zeit bereits gestellt hat, kann B davon auszugehen, dass A diese Frage mittlerweile von einer andern Person korrekt beantwortet bekommen hat. Daher wäre es sinnvoll A um eine Antwort zu bitten.
6. In einigen Fällen fragt eine Person zufällig umher, weil niemand spezialisiertes aufzufinden ist oder weil die gefragte Person der fragenden Person nahe steht.

Um diese sozialen Metaphern in einem Peer2Peer-Netzwerk zu realisieren, wird zuerst eine gemeinsame Wissensbasis erstellt (siehe Tabelle 2.1). Diese Wissensbasis wird als Ontologie bezeichnet. Dort ist hinterlegt, welches Themengebiet zu welchem Oberthema gehört, so dass man auch ähnliche Themen finden kann.

Des Weiteren enthält jeder Peer Informationen darüber, über wie viele einzelne Dokumente er zu jedem einzelnen Thema verfügt. Diese Informationen sind in Tabelle 2.2 beschrieben. Aufgrund dieses Aufbaus kann dann ein geeigneter Peer ausgewählt werden, der eine Anfrage beantwortet bzw. weiterleitet (*vgl. soziale Metapher 1*).

Um einen geeigneten Peer zu finden wird das Netzwerk und die darin enthalten Peers in vier verschiedene Schichten eingeteilt, wobei sich jeder Peer die Schicht eines jeden anderen Peers merkt, mit dem er bereits kommuniziert hat.

Document Relations Document	***Relations*** *hasTopic => Topic*
Topic	
TourismActivity	
DestinationManagement	
TravelDistribution	
TourismTechnology	
BookingSystem	
GeographicalInformationSystem	

Tabelle 2.1: Beispiel für eine Ontologie. Die Verbindung der einzelnen Themengebiete funktioniert hier über die Relation: Thema besitzt Unterthema [Tem06] S. 165

Die vier Schichten erklären sich folgendermaßen (freie Übersetzung nach Tempich [Tem06] S. 163):

- Der beste Peer um eine Anfrage zu beantworten, ist eine solcher, der diese Anfrage oder eine ähnliche bereits einmal beantwortet hat. Diese Peers heißen *content provider* (*vgl. soziale Metapher 2*).
- Wenn keine '*content provider*' bekannt sind, dann werden Peers angefragt, die bereits eine ähnliche Anfrage in der Vergangenheit an den nun fragenden Peer gestellt haben, der diese aber damals nicht beantworten konnte. Die Annahme ist nun, dass dieser Peer in der Zwischenzeit einen anderen Peer gefunden hat, der ihm seine Anfrage beantworten konnte. Daher kann man von ihm nun einen content provider erfahren. Diese Peers heißen *recommender* (*vgl. soziale Metapher 5*).
- Falls keine der oben genannten Peers bekannt sein sollten, wird die Anfrage an einen Peer gestellt, der viele Beziehungen zu anderen Peers im sozialen Netzwerk hat, die wiederum ein möglichst breites Wissensgebiet abdecken sollten. Bei einem solchen Peer ist die Wahrscheinlichkeit sehr groß, dass sich unter ihnen ein content provider oder zumindest ein recommender befinden. Diese Peers erhält man aus einem sog. *boot-*

strapping network (*vgl. soziale Metapher 4*).

- Wenn keine der oben genannten Peers gefunden werden kann gibt es eine Standradstrategie, bei der alle Peers befragt werden, die in direkter Nachbarschaft zum fragenden Peer stehen. Um, falls eine große Anzahl von Nachbarn vorhanden ist, eine übermäßige Belastung des Netzwerkes zu vermeiden, werden einige Nachbarn zufällig ausgewählt. Die Strategie nennt sich *default network* (*vgl. soziale Metapher 6*).

Die Weiterentwicklung des INGA-Algorithmus besteht nun darin, daß INGA keinen zentralen Index benötigt um die Anfragen durch das Netzwerk zu routen. Im INGA-Algorithmus kennt jeder Peer die Namen und Interessen der Peers mit denen er kommuniziert. Somit wird ein zentraler Index überflüssig, da das Routing von Anfragen und Antworten auch dezentral, über die einzelnen Peers, gesteuert werden kann. Aus diesem Grund bildet der INGA-Algorithmus auch die realen Verhältnisse besser ab als REMINDIN, denn in der Wirklichkeit gibt es auch keine Person, die die Interessen von allen anderen Personen kennt. [LTQ+05]

2.4.1 Bewertung

Der REMINDIN- sowie der INGA-Algorithmus verwenden, wie oben beschrieben, nahezu ausschließlich das Prinzip des semantischen Anfrageroutings. Das soziale Anfragerouting kommt wenig zum Tragen und das spirituelle Anfragerouting kommt überhaupt nicht vor.

Nach meiner Ansicht wird dadurch ein reales soziales Netzwerk nicht komplett abgebildet, da in der Realität die Freunde und Bekannten doch wesentlich bevorzugter befragt werden, als dieser Algorithmus es abbildet.

Beispielsweise liegt folgendes Problem vor: Der Abfluss ist verstopft und man bekommt ihn selbst nicht wieder frei. In der Realität würde man sich wie im REMINDIN-Algorithmus zuerst an einen content provider wenden, also jemand, der bereits einmal den Abfluss gereinigt oder etwas Ähnliches gemacht hat und mit dem man zufrieden war.

Falls dieser Schritt nicht zum Erfolg führt, würde man sich einen recommender suchen,

also jemand, von dem man weiß, dass er einmal einen Installateur beschäftigt hat oder der sich im Allgemeinen gut mit Handwerkern auskennt, wie z.B. ein Architekt. Diese Personen sind aber aller Voraussichicht nach bereits Freunde oder Bekannte, da man von einer beliebigen Person, die man auf der Straße trifft nicht weiß, ob sie bereits Installateure beschäftigt hat oder Architekt ist. Insofern ist hier das soziale Anfragerouting teilweise verarbeitet, da recommender nur dadurch gefunden werden, dass sie bereits, wie in der sozialen Metapher 5 beschrieben, mindestens einmal erfolglos angefragt haben. Dabei bleibt allerdings zu berücksichtigen, dass die sozialen Kontakte völlg zufällig geknüpft wurden.

Da der REMINDIN - Algorithmus immer versucht die Personen zu finden, die Anfragen am besten beantworten können, haben zwangsläufig Personen, die die gleichen Fragen stellen auch ähnliche Bekanntenkreise. Insofern bildet der REMINDIN-Algorithmus die Clustereigenschaft des Kleine-Welt-Schemas nach. Eine Abbildung geographischer Nähe, wie sie in der realen Welt ganz natürlich ist und die der Kleine-Welt-Eigenschaft sogar ihren Namen gab, findet im REMINDIN-Algorithmus hingegen überhaupt nicht statt.

Abschließend müsste meiner Ansicht nach das soziale und auch das spirituelle Anfragerouting stärker als bisher berücksichtigt werden, sowie eine geographische Komponente erzeugt werden, um der Realität näher zu kommen.

2.5 SwapSim

Ebenfalls im Zuge der Evaluierung des Standes der Wissenschaft wurde das Simulationstool SwapSim[2] betracht, dass von Christoph Tempich und anderen an der Universität in Karlsruhe entwickelt worden ist. Diese Umgebung soll dazu dienen verschiedenste Formen von Netzwerken zu simulieren.

Leider kann die vorliegenden Version des Tools auch nach ausgiebigen Tests zwar zum Compilieren, aber nicht annähernd fehlerfrei zum Laufen gebracht werden. Das liegt unter anderem daran, dass es keine ausreichende Anwenderdokumentation zu diesem Tool gibt bzw. auch nach längerem Suchen keine gefunden werden konnte. Aus diesen Gründen scheidet auch die Möglichkeit eines *learning by doing* aus.

[2] http://ontoware.org/projects/swapsim

Des Weiteren sind alle für das Projekt verantwortlichen Personen nicht mehr an der Universität in Karlsruhe zugegen und somit für eine ausführliche mündliche Einweisung nicht mehr zu erreichen.

Somit wäre nur die eigene Einarbeitung in den Code übrig geblieben. Dies ist jedoch aufgrund der großen Unübersichtlichkeit der Simulationsumgebung im Rahmen dieser Diplomarbeit nicht möglich. Daher habe ich mich entschieden SwapSim **nicht** zu verwenden.

Peer	Peer Resource	No. of Documents
2	TourismActivity	0
	DestinationManagement	0
	TravelDistribution	10
	TourismTechnology	0
	BookingSystem	0
	GeographicalInformationSystem	0
3	TourismActivity	10
	DestinationManagement	10
	TravelDistribution	10
	TourismTechnology	10
	BookingSystem	10
	GeographicalInformationSystem	10
	GeographicalInformationSystem $\wedge$ DestinationManagement	5
5	TourismActivity	30
	DestinationManagement	50
	TravelDistribution	100
	DestinationManagement $\wedge$ TravelDistribution	10
	TourismTechnology	0
	BookingSystem	0
	GeographicalInformationSystem	0
8	TourismActivity	0
	DestinationManagement	0
	TravelDistribution	0
	TourismTechnology	40
	BookingSystem	20
	GeographicalInformationSystem	100
	GeographicalInformationSystem $\wedge$ BookingSystem	10

Tabelle 2.2: Beispiel für die Verteilung von Wissen auf verschiedene Peers [Tem06] S. 166

Kapitel 3

Metriken

3.1 Einleitung

Metriken benutzt man um quantifizierbare Eigenschaften z.B. eines sozialen Netzwerks zu messen. Diese Eigenschaft, bzw. der Grad ihrer Erreichung, bestimmt eine Qualitätseigenschaft. [Ebe07] Metriken gibt es neben der Netzwerkanalyse u.a. auch auf dem Gebiet der Softwareentwicklung. In der Betriebswirtschaftslehre kennt man Metriken unter dem Begriff *Kennzahlen*.

Mit Hilfe von Metriken soll die Leistungsfähigkeit von sozialen Netzwerken **beurteilt** werden. Des Weiteren sollen Metriken **Vorhersagen** über das zukünftige bzw. eine **Kontrolle** über das Verhalten einer Person in einem Netzwerk oder eines ganzen Netzwerks geben können. Sie sollen dem Benutzer außerdem ein **Feedback** geben, damit dieser Unstimmigkeiten im Aufbau eines solchen Netzwerks erkennen kann.

Metriken in der Analyse von sozialen Netzwerken müssen verschiedene Anforderungen erfüllen. Beispielsweise sollen die Netzwerk unabhängig vom guten Willen und vom Sachverstand des Betrachters bewertet werden, d.h. der Messende darf keinen Einfluß haben. (**Ojektivität**). So muss sichergestellt werden, dass Netzwerke nicht bewusst oder unbewusst besser oder schlechter als andere bewertet werden, um so die **Vergleichbarkeit** der sozialen Netzwerke zu gewährleisten. Dem Ziel der Vergleichbarkeit dient auch die **Nor-**

mierung, denn sie dient dazu die Qualität von großen und kleinen Netzwerken unabhängig von ihrer Größe zu bewerten. Des Weiteren sollten Metriken **Zuverlässig** sein, d.h. sie sollten, bei gleicher Eingabe, stets das gleiche Ergebnis liefern. Das Ergebnis einer Metrik muss auch in einer möglichst kurzen Zeit vorliegen (**Effizienz**), damit die Benutzbarkeit **Nützlichkeit** der Metrik erhalten bleibt. Zu dieser Nützlichkeit gehört es natürlich auch, dass die Metrik ein Ergebnis liefert, welches für den Benutzer interessant ist.

In sozialen Netzwerken gibt es Metriken, die Werte für einen einzelnen Knoten messen sowie Metriken die Werte für das gesamte Netzwerk oder zumindest für Teilnetzwerke ermitteln. Die Metriken, die dabei Werte für das gesamte Netzwerk messen, werden als *globale Metriken* und solche für den einzelnen Knoten werden als *lokale Metriken* bezeichnet.

3.2 Metriken in der Netzwerkanalyse

3.2.1 Lokale Metriken

Clusterkoeffizient

Der *lokale Clusterkoeffizient* (C_i) eines Knotens n_i in einem ungerichteten Graphen *G* bezeichnet in der Graphentheorie den Quotienten aus der Anzahl der Kanten die zwischen ihm und seinen Nachbarn tatsächlich verlaufen (e_i) und der Anzahl der Kanten, die zwischen ihm und seinen Nachbarn maximal verlaufen könnten ($e_{\max}$). Wenn ein Knoten *n* Nachbarn hat ist $e_{\max} = n * (n+1)/2$ [WS98]. Der lokale Clusterkoeffizient $C_i(v)$ berechnet also folgendermaßen:

$$C_i(v) = \frac{e_i(v)}{e_{\max}} = \frac{e_i(v)}{n * (n+1)/2} = \frac{2e_i(v)}{n * (n+1)}$$

Der Knoten *A* aus Bild 3.1 hat einen lokalen Clusterkoeffizienten von 1, da er zwei Nachbarn *(B, C)* besitzt. Diese beiden Nachbarn wiederum haben auch untereinander einen Kontakt, so dass sich für die Knoten *A, B, C* ein vollständiger Graph ergibt, in dem jeder

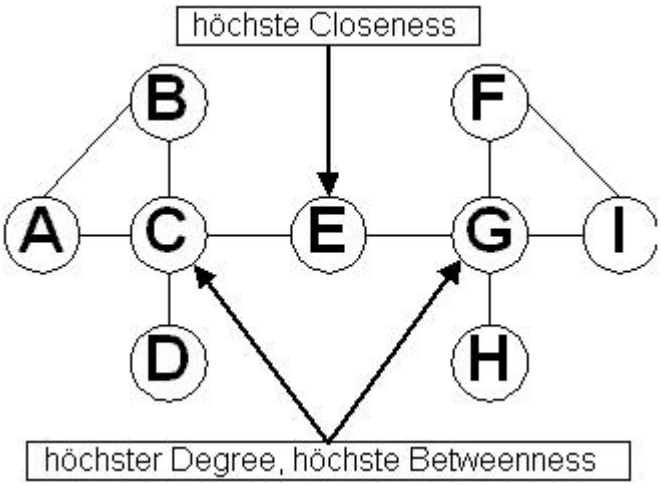

Bild 3.1: Beispiel zur Zentralität [Mut04]

mit jedem einen Kontakt hat. In die Formel eingesetzt ergibt sich daraus:

$$C_i(A) = \frac{2 * 3}{2 * (2 + 1)} = \frac{6}{6} = \mathbf{1}$$

Der Knoten *C* hat, genauso wie der Knoten *G* einen niedrigeren Clusterkoeffizienten, da zwischen seinen Nachbarn einige mögliche Kanten nicht existieren. Die Berechnung des Clusterkoeffizienten für die Knoten *C* bzw. *G* sieht folgendermaßen aus:

$$C_i(C) = \frac{2 * 5}{4 * (4 + 1)} = \frac{10}{20} = \frac{\mathbf{1}}{\mathbf{2}}$$

Zentralität

Es gibt drei verschiedene Arten der Zentralität. Alle verschiedenen Zentralitätsmaße kann man lokal, d.h. für jeden Knoten n_i einzeln bestimmen. [Mut04]

Grad Die *Degree Centrality* gibt an, wie viele direkte Kontakte ein Knoten besitzt. Je mehr Kontakte eine Person also in einem sozialen Netzwerk geknüpft hat, desto zentraler ist die Person nach dieser Metrik.

Der lokale Wert nennt sich *Actor Degree Centrality* (C_D). [WF07] Alternativ wird auch der Name *Degree* (d) verwendet.

$$C_D(n_i) = d(n_i)$$

Um das Maß zu standardisieren wird ein Quotient aus dem tatsächlichen und dem maximalen Degree-Wert eines Knotens gebildet. Der Maximalwert liegt bei $g - 1$, da in einem sozialen Netzwerk eine Person immer nur Kontakte zu anderen Personen aufbauen kann und nie zu sich selbst.

$$C'_D(n_i) = \frac{d(n_i)}{g-1}$$

Im Bild 3.1 hat beispielsweise der Knoten *I* eine standardisierte Actor Degree Centrality von 0,25.

$$C'_D(I) = \frac{2}{8} = \mathbf{0{,}25}$$

Nähe Die *closeness Centrality* gibt an, wie lang der kürzeste Weg von einem Knoten zu jeweils allen anderen ist. Aus den Längen der kürzesten Wege wird dann ein Durchschnittswert gebildet, der den Wert der Metrik darstellt. Je kleiner die Wert ist, desto zentraler liegt der betreffende Knoten.

Analog wird hier der lokale Wert als *Actor Closeness Centrality* (C_C) [WF07]

$$C_C(n_i) = \left[\sum_{j=1}^{g} d(n_i, n_j)\right]^{-1}$$

bezeichnet. Auch dieses Maß kann man wieder mit der Größe des Netzwerks $(g - 1)$ standardisieren. Auf diesem Weg erhält man:

$$C'_C(n_i) = \frac{g-1}{\left[\sum_{j=1}^{g} d(n_i, n_j)\right]}$$

In Bild 3.1 hat berechnet sich die Degree Centrality des Knoten *G* folgendermaßen:

$$C'_C(G) = \frac{8}{\left[\sum_{j=1}^{9} d(G, n_j)\right]} \approx \mathbf{0{,}533}$$

Der Knoten *F* besitzt hingegen nur einen Wert von $\approx$ **0,364**.

Zwischenraum Die *Betweenness Centrality* eines Knotens n_i gibt an auf wie vielen kürzesten Wegen aller anderen Personen diese Person liegt. Diese Zahl wird dann mit der Gesamtzahl der kürzesten Wege relativiert und bildet so die Metrik. Je größter die Anzahl der kürzesten Wege auf der eine Person liegt, desto zentraler bzw. prominenter ist eine Person in einem Netzwerk.

Auch hier wird der lokale Wert als *Actor Betweenness Centrality* [WF07] bezeichnet und folgendermaßen berechnet:

$$C_B(n_i) = \sum_{j<k} \frac{g_{jk}(n_i)}{g_{jk}}$$

Dabei wird g_{jk} als Anzahl der kürzesten Wege zwischen den Knoten (n_j) und (n_k) interpretiert und $g_{jk}(n_i)$ als Anzahl der kürzesten Wege, die über den Knoten n_i führen. Die Bedingung $j < k$ muss gelten, damit nur der kürzeste Weg von *A* zu *B* gefunden wird und nicht auch der von *B* zu *A*. Die standardisierte Form [WF07] sieht daher folgendermaßen aus:

$$C'_B(n_i) = \frac{C_B(n_i)}{(g-1)(g-2)/2}$$

Der Knoten *E* aus Bild 3.1 hat z.B. eine Actor Betweenness Centrality von

$$C'_B(E) = \frac{16}{28} \approx \mathbf{0{,}571}$$

3.2.2 Globale Metriken

Clusterkoeffizient

Der *globale* oder *Average Clusterkoeffizient* (C) lässt sich als Mittelwert der lokalen Clusterkoeffizienten aller Knoten berechnen. [WS98] Bei einer Gesamtzahl von *n* Knoten berechnet sich der globale Clusterkoeffizient nach der Formel:

$$C = \frac{1}{n} \sum_{i=1}^{n} C_i$$

Der Graph aus Bild 3.1 hätte somit einen globalen Clusterkoeffizienten von:

$$C = \frac{1}{9}(1 + 1 + \frac{1}{2} + 1 + \frac{2}{3} + 1 + \frac{1}{2} + 1 + 1) \approx \mathbf{0{,}852}$$

Kleine-Welt-Netzwerke haben einen sehr hohen durchschnittlichen Clusterkoeffizienten. In einem Zufallsgraphen ist der Clusterkoeffizient im Gegensatz zu natürlichen Netzwerken relativ gering.

Zentralität

Es gibt drei verschiedene Arten der Zentralität. Alle diese verschiedenen Zentralitätsmaße kann man global, also im Durchschnitt für die Gesamtzahl g der Knoten bestimmen. [Mut04]

Grad Die *Degree Centrality* gibt an, wie viele direkte Kontakte ein Knoten besitzt. Je mehr Kontakte also eine Person in einem sozialen Netzwerk geknüpft hat, desto zentraler ist die Person nach dieser Metrik.

Der globale Wert wird als *Group Degree Centrality* bezeichnet. Er besagt in wie weit sich das Netzwerk auf einen oder wenige zentrale Knoten stützt, zu denen alle anderen Knoten Kontakt haben. Die *Group Degree Centrality* wird mit folgender Formel berechnet [WF07]:

$$C_D = \frac{\sum_{i=1}^{g} [C_D(n^*) - C_D(n_i)]}{(g-1)(g-2)}$$

Dabei bezeichnet $C_D(n^*)$ die höchste Actor Degree Centrality, die im Graph vorzufinden ist. Dieses Maß erreicht seinen Maximalwert, falls ein sog. Sterngraph vorliegt. Das bedeutet, dass ein Knoten Kontakt zu allen anderen hat, sowie alle diese anderen Knoten wiederum ausschließlich Kontakt zu dem ersten Knoten haben. Daher hat der erste Knoten eine nicht zu überbietende Zentralität und ohne ihn würde das Netz zusammen brechen. Dieses Aufbauprinzip kam bei den ersten Versionen der Musiktauschbörse *Napster* zum Tragen. Daher war diese auch relativ leicht auszuschalten. In Bild 3.1 liegt die Group Degree Centrality bei

$$C_D = \frac{\sum_{i=1}^{9} [4 - C_D(n_i)]}{(8 * 7)} = \mathbf{\frac{1}{4}}$$

Die Group Degree Centrality ist nicht zu verwechseln mit dem *mittleren Grad (engl: mean*

degree) (k) eines Graphen.

$$k = \frac{\sum_{i=1}^{g} d_i}{g}$$

Der mean degree (k) ergibt sich aus dem Durchschnitt der degree-Werte aller Knoten im Netzwerk [AHK+07].

$$k = \frac{\sum_{i=1}^{g} d_i}{g}$$

Der mean degree für den in Bild 3.1 dargestellten Graphen berechnet sich daher folgendermaßen:

$$k = \frac{2+2+4+1+2+4+2+2+1}{9} \approx \mathbf{2{,}222}$$

Nähe Die *closeness Centrality* gibt an, wie lang der kürzeste Weg von einem Knoten zu jeweils allen anderen ist. Aus den Längen der kürzesten Weg wird dann ein Durchschnittswert gebildet, der den Wert der Metrik darstellt. Je kleiner die Wert ist, desto zentraler liegt der betreffende Knoten.

Diese Metrik soll, in einem sozialen Netzwerk, dafür genutzt werden zentrale Personen zu finden, die schnell mit möglichst vielen anderen Personen Kontakt aufnehmen können. Diese Personen werden als Hubs und Authorities bezeichnet.

Der globale Wert dieser Metrik wird als *Group Closeness Centrality* [WF07] bezeichnet.

$$C_C = \sum_{i=1}^{g} [C'_C(n^*) - C'_C(n_i)]$$

Dabei stellt $C'_C(n^*)$ die Person mit der höchsten Actor Closeness im Netzwerk dar. Auch hier gibt es wiederum eine Standardisierung des Maßes mit der höchstmöglichen Actor Closeness $[(g-2)(g-1)/(2g-3)]$ um die Vergleichbarkeit zu gewährleisten.

$$C_C = \frac{\sum_{i=1}^{g} [C'_C(n^*) - C'_C(n_i)]}{(g-2)(g-1)/(2g-3)}$$

Die Group Closeness Centrality des in Bild 3.1 abgebildeten Graph berechnet sich folgendermaßen:

$$C_C = \frac{\sum_{i=1}^{9} [0{,}533 - C'_C(n_i)]}{(7*8)/15} \approx \mathbf{3{,}82}$$

Zwischenraum Die *Betweenness Centrality* eines Knotens n_i gibt an, auf wie vielen kürzesten Wegen aller anderen Personen diese Person liegt. Diese Zahl wird dann mit der Gesamtzahl der kürzesten Wege relativiert und bildet so die Metrik. Je größter die Anzahl der kürzesten Wege auf der eine Person liegt, desto zentraler bzw. prominenter ist eine Person in einem Netzwerk.

Die globale Form wird als *Group Betweenness Centrality* [WF07] bezeichnet und sieht in ihrer standardisierten Form folgermaßen aus:

$$C_B = \frac{\sum_{i=1}^{g} [C_B(n^*) - C'_B(n_i)]}{(g-1)}$$

Dabei stellt $C_B(n^*)$ die Person mit der höchsten Actor Betweenness im Netzwerk dar. Die Betweenness Centrality des Graphen aus Bild 3.1 rechnet sich folgendermaßen:

$$C_B = \frac{\sum_{i=1}^{9} \left[\frac{17}{28} - C'_B(n_i)\right]}{(9-1)} = \frac{\frac{103}{28}}{8} \approx \mathbf{0{,}46}$$

Grad Korrelation

Die Grad Korrelation *(engl: Degree Correlation)* beschreibt die Summe der Wahrscheinlichkeiten, mit der sich ein Knoten des Grades d mit Knoten aller im Netzwerk vorhandenen unterschiedlichen Grade d' direkt verbindet. Kommt die Grad Korrelation k_{nn} dem betrachtetem Grad sehr nahe oder entspricht ihm sogar, dann deutet diese Eigenschaft auf einen geschlossenen Ring hin, d.h. Informationen dringen nur schlecht oder im Extremfall auch gar nicht aus dem Ring heraus. Andererseits gelangen sie aber auch nur schlecht in den Ring hinein. [AHK+07]

In einem sozialen Netzwerk kann ein solcher Wert auf mafiöse, korrupte oder korruptionsanfällige Strukturen hinweisen. [BK07] Die Formel für die Grad Korrelation lautet:

$$\langle k_{nn} \rangle = \sum_{d'=1}^{d_{max}} d' P\left(d' | d\right)$$

Die Grad Korrelation für den Grad 2 des in Bild 3.1 abgebildeten Graph berechnet sich folgendermaßen:

$$k_{nn}(2) = 1 * \frac{0}{10} + 2 * \frac{4}{10} + 3 * \frac{0}{10} + 4 * \frac{6}{10} = \frac{32}{10} = \mathbf{3{,}2}$$

Dabei sei zur Erklärung des vorstehenden Beispiels gesagt, dass keiner der 9 vorhanden Knoten den Grad 2 besitzt und sich gleichzeitig mit einem Knoten verbindet, der selbst nur den Grad 1 hat. Dafür gibt es allerdings vier Knoten, mit dem Grad 2, die sich wiederum mit einem anderen Knoten des selben Grades verbinden. Nach dem selben Prinzip wird für die Konten mit den Graden 3 und 4 vorgegangen. In der Addition aller Brüche muss der eins herauskommen, da es sich hier um Wahrscheinlichkeiten handelt. Im Nenner spiegelt sich die Gesamtzahl der Kanten wieder.

In Bild 3.2 wird beispielhaft die Grad Korrelation für zwei verschiedene soziale Netzwerke aufgezeigt. Dabei handelt es sich um zwei Unternetze aus dem real existierenden sozialen Netzwerk Cyworld[1]. Bei Cyworld gibt es zwei verschiedene Arten der Freundschaft:

1. Eine Freundschaft, die durch eine Einladung an eine andere Person deren Freund man werden möchte bzw. durch das Akzeptieren einer solchen Einladung entstanden ist. In Bild 3.2 als *Friends network* und in roter Farbe dargestellt.

2. Jedem seiner Freunde kann man durch hinzufügen einer besonderen Empfehlung *(Testimonial)*, als guten bzw. engen Freund ausweisen. Diese Art der Freundschaft war allerdings zu Anfang des Netzwerks beschränkt auf eine Anzahl von 101 testimonilas, die eine Person vergeben konnte. Diese Obergrenze wurde aber mit dem sich vergrößernden sozialen Netzwerk leicht angehoben. In Bild 3.2 als *Testimonial network* und in grün dargestellt.[AHK$^+$07]

Da bei der zweiten Form der Freundschaft nur eine beschränkte Anzahl zulässig ist, kann man davon ausgehen, dass ein Testimonial nur unter besonders guten Freunden vergeben wird. Daher ist auch damit zu rechen, dass eine Person, die ein solches Testimonial erhalten hat, der Person die es ihr gegeben hat wiederum ein Testimonial zurückgibt. Diese Eigenschaft verändert die Streuung der Grad Korrelation erheblich, denn nun ist das entstehen von Hubs und Authorities erheblich erschwert.

Das Fehlen von Hubs und Authorities lässt auch an der Grafik 3.2 ablesen. Da bei den Daten aus dem Testimonialnetzwerk der Grad k mit der Grad Korrelation $k_{nn}(k)$ in etwa übereinstimmen und sich so annähernd eine Winkelhalbierende ergibt.

[1] www.cyworld.com; größter und ältester Anbieter von Internetdienstleistungen im Bereich sozialer Netzwerke in Korea [AHK$^+$07]

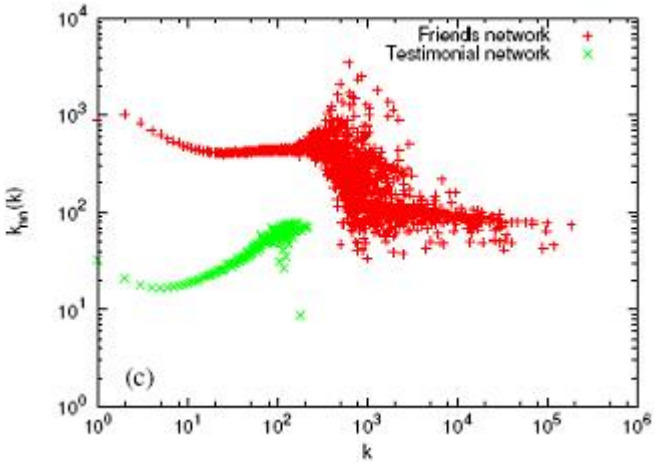

Bild 3.2: Grad Korrellation in Cyworld [AHK+07]

Im Friendsnetwork nimmt die Grad Korrelation sogar tendenziell ab. Was bedeutet, dass Personen die von vielen Leuten gekannt werden selbst wiederum nur wenige andere Personen kennen. Daher handelt es sich hier um Personen die in sozialen Netzwerken Freunde ähnlich wie Trophäen *'sammeln'*.

Degree Of Separation

Mit dem Degree of Separation lässt sich beschreiben wie gut ein Netzwerk vernetzt ist. Dies gibt, wie bereits oben beschrieben, die Anzahl der Knoten an, die zwischen zwei verschiedenen Knoten liegen. Je größer der Degree of Separation ist, desto schlechter dringen Informationen durch ein Netzwerk. In seiner globalen Ausprägung wird der Degree of Separation durch den Durchschnitt der Degrees of Separation aller möglichen Knotenpaare gebildet. Der Degree of Separation eines Knotenpaares wäre somit die lokale Ausprägung dieser Metrik. Diese ist aber für die wissenschaftliche Auswertung nicht interessant.

Da es in großen Netzwerken praktisch unmöglich ist alle lokalen Degrees of Separation festzustellen, um so den globalen Wert zu ermitteln, wird eine Näherungsformel vorgeschlagen, um den Degree of Separation zu berechnen. [NSW01]

$$\frac{\log(N/n_1)}{\log(n_2/n_1)} + 1$$

Dabei ist N die Gesamtzahl der Knoten im Netzwerk und n_1 bzw. n_2 beschreiben die durchschnittliche Anzahl der ersten bzw. zweiten Nachbarn. Der Degree of Separation für den in Bild 3.1 abgebildeten Graphen, berechnet sich nach dieser Näherungsformel folgendermaßen:

$$\frac{\log(9/\frac{20}{9})}{\log(\frac{42}{9}/\frac{20}{9})} + 1 = \frac{\log(\frac{81}{20})}{\log(\frac{21}{10})} + 1 \approx \mathbf{2{,}885}$$

Da der Degree of Separation für den in Bild 3.1 abgebildeten Graphen, wenn man ihn exakt ausrechnet bei $\approx$ **2,089** liegt, ist davon auszugehen, dass die Näherungsformel brauchbare Ergebnisse liefert. Zu diesem Ergebnis sind auch Ahn, Han et. al. gekommen [AHK$^+$07], da sie mit Hilfe dieser Formel die sozialen Netzwerke MySpace und orkut analysierten.

Prestige

Obwohl bei den verschiedenen Metriken zur Zentralität zwischen eingehenden und ausgehenden Kanten kein Unterschied gemacht wurde, kann es diesen geben. Für den Fall, dass es im betrachteten sozialen Netzwerk eine Unterscheidung zwischen den beiden Kantenarten eingehend und ausgehend existiert, kann dieser Unterschied auch dazu benutzt werden, um daraus diverse Metriken abzuleiten. Diese fimieren unter dem Oberbegriff *Prestige*-Metriken. [WF07]

Da aber diese Arbeit zeitlich begrenzt ist, können die Prestige-Metriken an dieser Stelle nicht mehr näher betrachtet werden. Daher werden nur die Ergebnisse, der bereits oben erwähnten Metriken, zum Clusterkoeffizient und zur Zentralität im Sinne dieser Arbeit weiter betrachtet.

Kapitel 4

Eigener Ansatz

In diesem Kapitel werden Metriken neu entwickelt, um mit diesen themenspezifische Ergebnisse zu erhalten. Des Weiteren werden die Probleme und deren Lösungen beschrieben, die sich bei der praktischen Umsetzung des gesamten Projekts ergeben haben.

4.1 Themenspezifische Metriken

Jede der bisher vorgestellten Metriken ist eine allgemeine Metrik, mit der man einzig und allein die Beziehungen zwischen den Knoten im Bezug auf die Gesamtheit der vorhandenen Interessensgebiete messen kann. Die eigentliche Fragestellung dieser Arbeit ist aber auf welche Art und Weise sich Personen mit den gleichen oder ähnlichen Interessen in einem sozialen Netzwerk finden. Daher ist eine Methode zu entwerfen, mit der man eine Antwort auf diese Fragestellung geben kann.

Ein geeigneter Weg dahin besteht darin, dass nur die Personen, welche Auskunft zu einem speziellen Themengebiet geben können, getrennt von allen anderen Personen, betrachtet werden. Dies wird erreicht, indem alle anderen im Netzwerk vorhandenen Personen ausgeblendet werden. Bei einem solchen Vorgehen müssen auch die Beziehungen, die von den ausgeblendeten Personen ausgehen, ausgeblendet werden, da sie sonst ins Leere laufen würden.

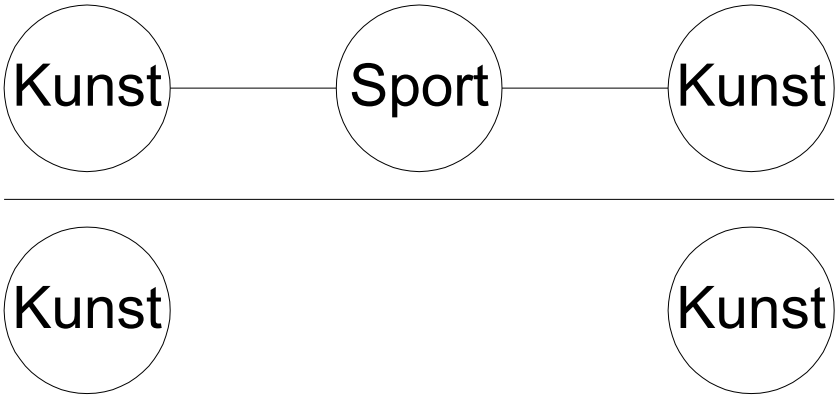

Bild 4.1: Ausblendvorgang bei Grad 1

Nun wird ein Interessensgebiet (z.B. *'Kunst'*) spezifiziert für das gemessen werden soll, wie gut sich die Personen gefunden haben, die sich für eben dieses Gebiet interessieren. Alle Knoten, die Personen repräsentieren, welche sich nicht für *'Kunst'* interessieren, sowie die Kanten, die von ihnen ausgehen, werden nun ausgeblendet (siehe Bild 4.1). Der sich aus den Ausblendungen ergebende Untergraph wird, mit den gleichen Metriken, die bereits in Abschnitt 3.2 eingeführt wurden, ausgewertet.

Nachfolgend werden an das alle Personen beinhaltende, ursprüngliche soziale Netzwerk Anfragen, wie in Kapitel 2 beschrieben, gestellt. Damit soll reales menschliches Verhalten simuliert werden. Während dieses Prozesses wird auch das Knüpfen neuer Bekanntschaften simuliert. Diese werden dann wiederum durch neue Kanten dargestellt. Im nächsten Schritt wird wieder der gleiche Ausblendvorgang durchgeführt, der bereits oben beschrieben wurde. Da neue Kanten zum Graph hinzugefügt wurden, hat sich dieser nun verändert und kann wiederum mit den Metriken aus Abschnitt 3.2 ausgewertet werden. Der Unterschied zwischen den beiden Messungen bildet ab, wie gut oder wie schlecht sich Personen in einem sozialen Netzwerk finden, die gleiche oder ähnliche Interessen haben.

In der Realität ist damit zu rechnen ist, dass Personen, die sich für das gleiche Thema interessieren und mindestens einen gemeinsamen Bekannten haben, sich im Laufe der Zeit, über diesen gemeinsamen Bekannten, ebenfalls kennenlernen. Um diesen Prozess zu simulieren, kann man zwischen Personen, auf die diese Kriterien zutreffen neue Beziehungen

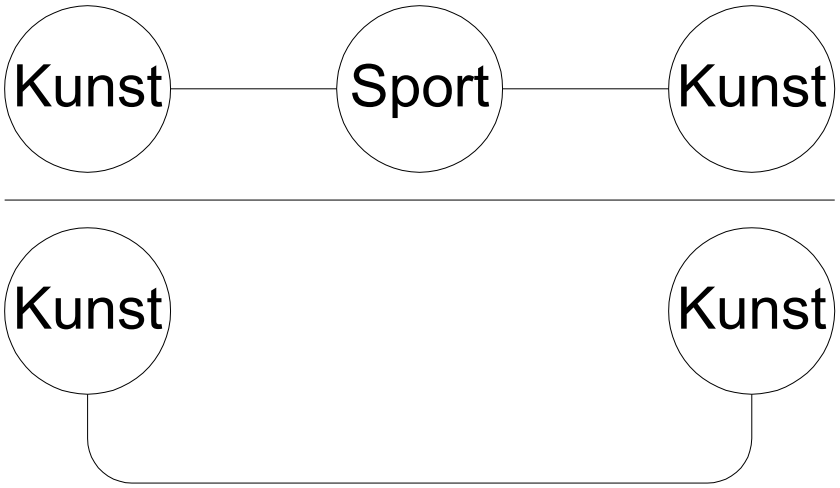

Bild 4.2: Ausblendvorgang bei Grad 2

in das soziale Netzwerk einfügen (siehe Bild 4.2).

Nun geht der Prozess wie oben beschrieben weiter. Alle Personen, die sich nicht für das spezifizierte Thema interessieren, werden ausgeblendet. Da keine neue interessierte Person hinzugefügt wurde, bleibt die Anzahl der Personen nach der Ausblendung gleich. Nur die Anzahl der Kanten ist gestiegen, denn einige wurden soeben hinzugefügt. Da jetzt beim Start der Simulation mehr Knoten vorhanden sind als vorher, ist die Chance gestiegen, dass sich Personen mit gleichen Interessen finden.

Da bei der zweiten Metrik auch Kontakte zu Nachbarn von Nachbarn, also zu Nachbarn zweiten Grades aufgenommen werden, nennt sich diese themenspezifische Metrik ***'Grad 2-themenspezifische Metrik'***. Dementsprechend nennt sich die weiter oben beschriebene Metrik ***'Grad 1-themenspezifische Metrik'***. Dieser Logik folgend könnten auch themenspezifische Metriken mit dem Grad 3, 4, 5 usw. angewendet werden. Bei einer solchen Vorgehensweise ist allerdings noch zu klären, ab welchem Grad die Realitätsnähe verloren geht.

4.2 Daten

Um die im Kapitel 2 beschriebenen Metriken für mehrdimensionale soziale Netzwerke erweitern und testen zu können, werden reale Daten benötigt. Dies ist notwendig um möglichst exakte Ergebnisse zu erhalten. Solche Daten werden für diese Arbeit aus dem **D**irectory**MOZ**illa-Datensatz[1] bezogen. Der DMOZ-Datensatz hat den Vorteil, dass er im Gegensatz zu anderen sozialen Netzwerken im Internet, wie beispielsweise YouTube[2] oder StudiVZ[3] seine Daten, für den Nutzer kostenfrei, offenlegt.

4.3 Was ist DMOZ?

Bei DMOZ handelt es sich um ein Open-Directory Projekt, welches eine Linksammlung darstellt. In dieser Linksammlung können Teilnehmer Webseiten manuell in Kategorien und Unterkategorien einsortieren. Die Kategorien entsprechen dabei dem Inhalt der Seiten. Neben einzelnen Webseiten können die Benutzer (in DMOZ *Editoren* genannt) auch selbst Unterkategorien erstellen, falls noch keine passende für eine spezielle Webseite existieren sollte. Zusätzlich können auch Querverweise zu ähnlichen Themen, die sich allerdings in anderen Kategorien befinden, von jedem Editor erzeugt werden.

Das Projekt DMOZ gibt es in nahezu allen, auf der Erde gesprochenen, Sprachen. Jede Kategorie gibt es also in mehreren Sprachen. Diese Kategorien sind über die oben beschriebenen Querverweise miteinander verlinkt. Dabei ist es aber nicht unbedingt notwendig, dass eine Kategorie, die z.B. in Deutsch existiert auch in allen anderen Sprachen vorhanden ist. Denn dazu müssten bereits beim Anlegen der Kategorie deren Name in allen Übersetztungen bekannt sein, was ein realer Editor niemals leisten kann, denn niemand kennt alle Übersetzungen des Kategorienennamens in alle verschiedenen Sprachen. Außerdem stoßen einige Kategorien auch in verschiedenen Sprachräumen auf wenig bis gar kein Interesse.

In dieser Arbeit führt der Sprachenmix allerdings dazu, dass nicht nur Personen nicht mit-

[1] www.dmoz.org
[2] www.youtube.com
[3] www.studivz.net

einander kommunizieren können, die unterschiedliche Interessensgebiete besitzen, sondern auch Personen mit gleichen oder ähnlichen Interessen können nicht miteinander kommunizieren, wenn sie keine gemeinsame Sprache sprechen. Daher wird hier der Einfachheit halber davon ausgegangen, dass sich jede Person mit jeder anderen Person verständigen kann. Diese Annahme ist nicht völlig abwegig, da es mit Englisch eine besonders im Internet weit verbreitete Mittlersprache gibt. Des Weiteren haben Personen mit den gleichen Interessen auch die jeweilige Fachsprache um sich verständigen zu können. Außerdem gibt es bereits heute im Internet brauchbare Tools um ganze Webseiten von einer in eine andere Sprache zu übersetzen.

Der DMOZ Datensatz steht als Datei im RDF-Format, einem XML-Dialekt zur Speicherung von Daten, zur Verfügung. Die, für diese Arbeit benötigten, Hauptklassen des DMOZ Datensatzes sind *Topic* und *Alias* ([Tem06] S. 186):

- **Topic:** Der Tag Topic repräsentiert die Themenhierarchie im DMOZ Datensatz. In seinen Attributen und Unterelementen werden der Themenname in unterschiedlichen Sprachen, sowie die Editoren des Themengebietes festgehalten. Die Eigenschaften *related* (ähnliches Themengebiet), *symbolic* (anderer Name für das gleiche Themengebiet) und *narrow* (Unterthema) beschreiben Beziehungen zu anderen Themengebieten und Aliasen.
- **Alias:** Über Alias werden Themengebiete hinterlegt, die dem behandelten Themengebiet gleichen. Es können ein Name und die Zieladresse des ähnlichen Themengebietes in der DMOZ-Datenstruktur angegeben werden.

Der Datensatz hat einige Eigenschaften, die ihn für eine Evaluieren interessant erscheinen lassen.

- Im Gegensatz zu vielen anderen Datensätzen gibt es sehr viele Beziehungen zwischen den einzelnen Themengebieten. Vor allen Dingen gibt es nicht nur ein baumartiges und damit stringentes Beziehungsgeflecht, sondern es gibt auch einige Querverweise, die Ähnlichkeiten und Schnittmengen repräsentieren und damit '*siehe auch*'-Beziehungen ermöglichen. Diese Vielfalt gestattet das Anfragerouting realitätsnäher zu gestalten.

- Die Themengebiete besitzen Editoren (viele auch mehrere). So ergibt sich eine sehr einfache Möglichkeit den DMOZ - Datensatz auf einen künstlich generiertes Kleine-Welt-Netzwerk zu übertragen.

- Zu den einzelnen Themengebieten sind sehr viele Links vorhanden, so dass eine repräsentative Auswertung entstehen kann.

4.4 Initiale Verteilung der DMOZ-Benutzer

Bei Beginn der Simulation stellt sich die Frage, nach welchem Schema man die Benutzer, die man aus einem realen Datensatz, wie beispielsweise DMOZ, gewonnen hat auf ein künstlich generiertes Kleine-Welt-Netzwerk verteilt. Der Ansatz von Kleinberg sieht eine quadratische Matrix vor, in der die einzelnen Personen angeordnet sind, so dass jede dieser Personen Kontakte zu ihren jeweiligen vier direkten Nachbarn hat. Außerdem besitzt jede Person noch einen long-range-Kontakt zu einer fünften Person, die weiter entfernt angeordnet ist.

Wenn man nun reale Personen aus dem DMOZ Datensatz, mit den ihnen zugeordneten Interessen, versucht auf ein solches künstliches Netzwerk zu verteilen, dann stellt sich die Frage welche Person an welche Stelle kommt. Eine Zuordnung die Personen mit den selben Interessen an eine ähnliche Stelle im Netzwerk setzt, wäre nicht angebracht, da auch in der Realität eine solches Clustering nicht stattfindet. Aber auch eine Verteilung bei der die Themenbereiche immer abwechselnd angeordnet werden, so dass beispielsweise jeder zehnte Knoten ein bestimmtes Thema zugeordnet bekommt, wenn 10 Prozent der realen Personen sich für dieses Thema interessieren, entspricht nicht der Realität. Daher wäre eine zufällige Anordnung der Personen auf dieser Matrix, im Hinblick auf die Realitätsnähe, optimal, denn auch in der Realität gibt es Cluster, in denen gewisse Interessen häufiger vorkommen als in anderen. Ein Beispiel hierfür wäre, dass sich wesentlich mehr Personen in Stadtvierteln mit gehobenem Einkommen für Luxusgüter interessieren, als Personen in Stadtvierteln mit sehr niedrigem Einkommen. Ein anderes Beispiel wäre, dass es in Gelsenkirchen mehr Anhänger des *FC Schalke 04* pro 1000 Einwohner gibt, als in einer anderen Stadt.

Außerdem gestaltet sich eine gleichmäßige Verteilung der Interessen über die Matrix sehr schwierig, da viele Personen zwei oder mehr Interessen gleichzeitig besitzen, wie in Grafik 4.3 zu erkennen ist. Diese Grafik veranschaulicht die Anzahl der verschiedenen Interessen aller im DMOZ-Datensatz registrierten Benutzer.

Auch wenn es gelingen sollte die Erstinteressen gleichmäßig oder geclustert zu verteilen, wären dann die Zweit-, Dritt-, usw. Interessen doch wieder willkürlich verteilt. Des Weiteren müsste man zuerst definieren was ein Erstinteresse ist.

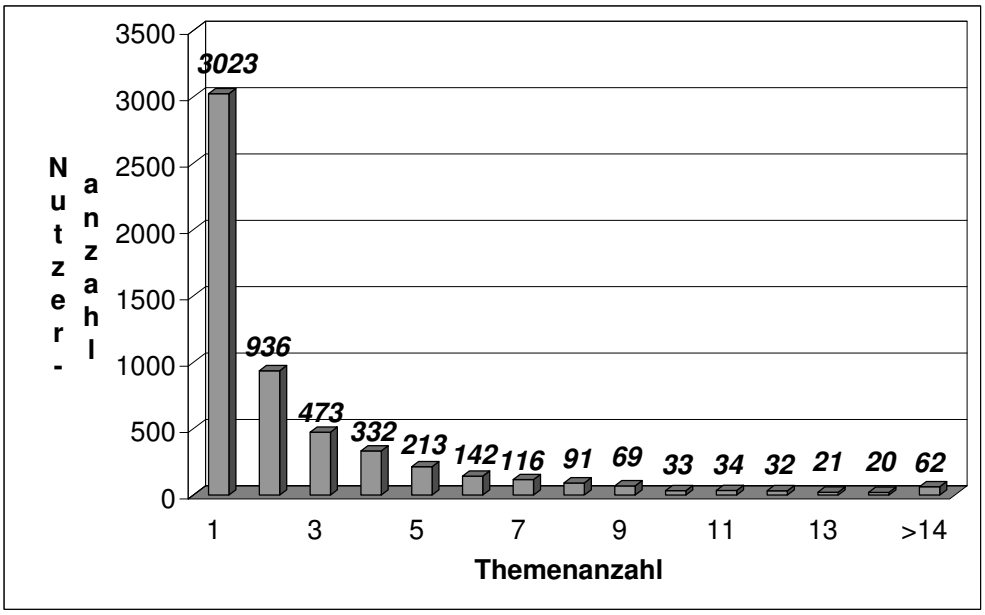

Bild 4.3: Wie viele Themengebiete werden von den einzelnen Editoren bearbeitet?

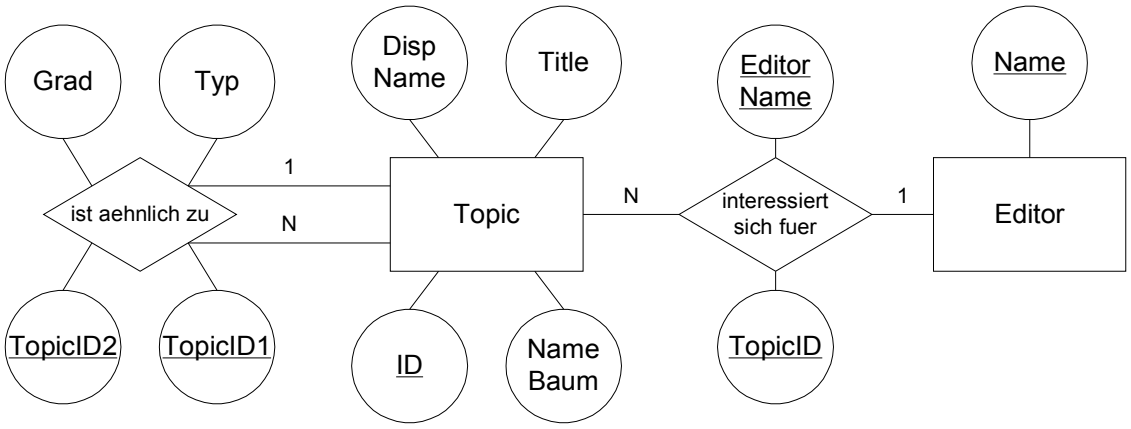

Bild 4.4: Entity-Relationship-Diagramm der Daten in der Datenbank

4.5 MySQL

Um die Daten aus dem DMOZ-Datensatz besser verwenden zu können, wurden diese in eine MySQL-Datenbank eingelesen. MySQL wurde als Datenbank für diese Arbeit ausgewählt, da es ich um eine frei verfügbare und leicht handhabbare Datenbank handelt.

In Bild 4.4 werden die Zusammenhänge im DMOZ-Datensatz in einem ER-Diagramm dargestellt. Die Entität Topic beinhaltet folgende Attribute zu den verschiedenen Themen.

- **ID:** Eine Zahl, die das Thema eindeutig identifiziert (z.B. *0000380507*).
- **NameBaum:** Der Name des Themas in der Baumstruktur (z.B. *Top/Regional/Europe/Germany/States/Rhineland-Palatinate/Localities/Mainz*).
- **Title:** Der eigentliche Name des Themas. Dieser entspricht meistens dem letzten Glied der Baumstruktur (z.B. *Mainz*).
- **DispName:** Manche Themen werden unter einem anderen Namen angezeigt als sie in der Datenbank abgespeichert werden (z.B. wird der Titel *Home_Improvement* in *Home Improvement* geändert. Dieser angezeigte Name wird unter DispName abgespeichert.

Die Entität Editor beinhaltet nur das Attribut Name. DMOZ stellt bereits sicher, dass jeder Name nur einmal vergeben wird, daher ist der Name in der Datenbank automatisch eindeutig.

Zwischen den verschiedenen Themen besteht eine Beziehung, die die Ähnlichkeit von zwei verschiedenen Themen beschreibt. Dabei kann ein Themengebiet ähnlich zu beliebig vielen anderen Themengebieten sein. Diese Beziehung hat vier Attribute:

- **TopicID1:** Erstes der beiden ähnlichen Themen.
- **TopicID2:** Zweites der beiden ähnlichen Themen.
- **Typ:** Art der Ähnlichkeit (*symbolic* oder *narrow*).
- **Grad:** Grad der Ähnlichkeit von 0 (große Ähnlichkeit) bis 2 (leichte Ähnlichkeit).

Zwischen den Editoren und den Themengebieten gibt es eine Beziehung, die das Interesse der Editoren für die unterschiedlichen Themengebiete festhält. Dabei kann sich ein Editor wiederum für beliebig viele Themengebiete interessieren. Diese Beziehung hat zwei Attribute:

- **EditorName:** Name des Editors.
- **TopicID:** Eindeutiger Identifizierer des Themengebietes für das sich der Editor interessiert.

Um die Daten einzulesen, wurde ein JAVA-Skript geschrieben, dass die Daten mit Hilfe des SAX-Parsers aus einer XML-Datei auslesen kann. Diese Daten wurden dann wiederum über die JDBC-Schnittstelle in die MySQL-Datenbank eingelesen. Die Übertragung von einer XML-Datei in eine Datenbank fand aus Gründen des einfacheren und schnelleren Zugriffs statt.

Es wurde bereits beim Einlesen der Daten darauf geachtet, dass Themen, die in mehreren Sprachen vorlagen als ein Thema behandelt werden. Dieses Vorgehen macht ein späteres Suchen nach Antworten in verschiedenen Sprachen unnötig.

Da sowohl der SAX-Parser als auch die JDBC-Schnittstelle für JAVA bereits kostenfrei zur Verfügung standen und das Einlesen der Quelldatei mit einer akzeptablen Geschwindigkeit ablief, lag es nahe zur Übertragung der Daten in die Datenbank eine Anwendung in JAVA zu schreiben.

Nachdem die Daten in die Datenbank eingelesen waren, ergab sich ein Bild, wie es in Tabelle 4.1 dargestellt ist.

Anzahl der...	***Werte***
...Themen in allen Sprachen	739.892
...Editoren	5.597
...Beziehungen zu verwandten Themen bzw. Unterthemen	869.312

Tabelle 4.1: Datenvolumen in DMOZ

Um in Java besser mit Graphen arbeiten zu können, wurde das *JUNG*-Package[4] verwendet, da es bereits eine Implementierung für den Small-World-Generator nach Kleinberg sowie für alle anderen Standardoperationen beinhaltet. Das *JUNG*-Package wurde in der Version 1.7.6 verwendet.

[4] http://jung.sourceforge.net/doc/api/index.html

Kapitel 5

Experimente und Ergebnisse

5.1 Grundaufbau

5.1.1 Versuchsaufbau und Versuchsdurchführung

Wie in Kapitel 2.1.1 beschrieben hat jeder Knoten 4 direkte Nachbarn und baut zusätzlich einen long-range-Kontakt zu einem weiter entfernten Knoten auf. Daher hat jeder Knoten mindestens 5 Nachbarn. Da jeder long-range-Kontakt, der von einem Knoten aufgebaut wird auch an einem anderen Knoten ankommen muss, ist es theoretisch möglich, dass sich alle Knoten mit ihrem long-range-Kontakt auf einen einzigen Knoten verbinden. Deshalb kann ein Konten maximal eine Kante weniger besitzen, als es Knoten im gesamten Netzwerk gibt. Dieser Fall ist aber sehr unwahrscheinlich.

Für die Auswertung der Versuche wurden die in Kapitel 3 beschriebenen Metriken in JAVA, unter Verwendung des Packages JUNG implementiert. Einzig auf die Betweenness Centrality wurde verzichtet, da sich deren Berechnung sehr aufwendig gestaltet. Jedem der im DMOZ-Datensatz vorhandenen Editoren wurde ein Knoten in einem Small-World-Netzwerk nach Kleinberg (siehe Kapitel 2.1.1) zugeordnet.

Da das Small-World-Modell von Kleinberg auf einer quadratischen Matrix basiert, konnten nicht alle 5597 im DMOZ-Datensatz vorhandenen Editoren für den Versuchsaufbau

verwendet werden. Das erklärt sich daraus, dass 5597 keine Quadratzahl ist. Daraus folgt, dass bei einer Verwendung einer quadratischen Matrix entweder Knoten übrig bleiben würden denen keine Editoren zugeteilt werden könnten oder es würden Editoren übrig bleiben, für die keine Knoten mehr frei sind. Es wurde dann mit 5476 die nächstkleinere Quadratzahl gewählt. Damit wurden die Benutzer auf einem zufällig erzeugten Small-World-Netzwerk mit einer Matrixkantenlänge von 74 verteilt. Die Verteilung erfolgte dabei alphabetisch. Die 121 überflüssigen Editoren wurden nicht berücksichtigt, da es sehr schwierig ist in einer Simulation mit *leeren* Knoten, also mit Knoten welche keinen Editor repräsentieren, umzugehen.

Der Informationsverlust durch das Weglassen von Editoren wurde dadurch minimiert, dass nur solche Editoren, die sich für nur ein einziges Thema interessieren weggelassen wurden. Auch hier wurden die 121 alphabetisch ersten Editoren auf die dieses Kriterium zutrifft weggelassen.

Für den von Kleinberg beschriebenen Parameter α, wurde der Wert 2 verwendet. Dieser Wert wurde gewählt, weil Kleinberg in seinem Artikel 2 als den Wert mit der größten Realitätsnähe beschreibt [Kle06].

5.1.2 Ergebnisse und Diskussion

Metriken	***Werte***
globaler Clusterkoeffizient	$\approx 29.09\%$
Group Degree Centrality	$\approx 0.091\%$
Mean Degree	6
Group Closeness Centrality	$\approx 3.431\%$
Degree Of Separation	≈ 7.78

Tabelle 5.1: Messergebnisse für das grundlegende Netzwerk.

Die Messungen zeigen ein zu erwartendes Ergebnis. Da pro Knoten zwei short- und eine long-range Kante generiert werden und diese auch jeweils bei einem anderen Knoten ankommen müssen, ist die Zahl der Kanten um das Sechsfache größer als die Zahl der Knoten. Daraus ergibt sich ein *mean Degree* von sechs.

Da die Zahl der long- gegenüber der Zahl der short-range Kanten noch relativ gering ist, ist das Netz, auf Grund der Anordnung der Knoten in Form einer Matrix mit verbundenen Nachbarn, noch sehr dezentral angelegt. Daher haben die Zentralitätsmaße *Group Degree Centrality* und *Group Closeness Centrality* noch sehr geringe Werte. Mit jeder neuen long-rang-Kante, die in den nächsten Experimenten hinzugefügt wird, wird sich später eine neue *'Abkürzung'* im Netzwerk ergeben. Diese führt dazu, dass die einzelnen Knoten besser erreichbar sind und somit die Zentralität zunimmt. Genau der gleiche Zusammenhang gilt auch für den *Degree Of Separation*, der einzige Unterschied ist hier, dass der *Degree Of Separation* mit jeder neuen Kanten abnimmt, denn der *Degree Of Separation* misst die durchschnittliche Weglänge zwischen zwei Knoten.

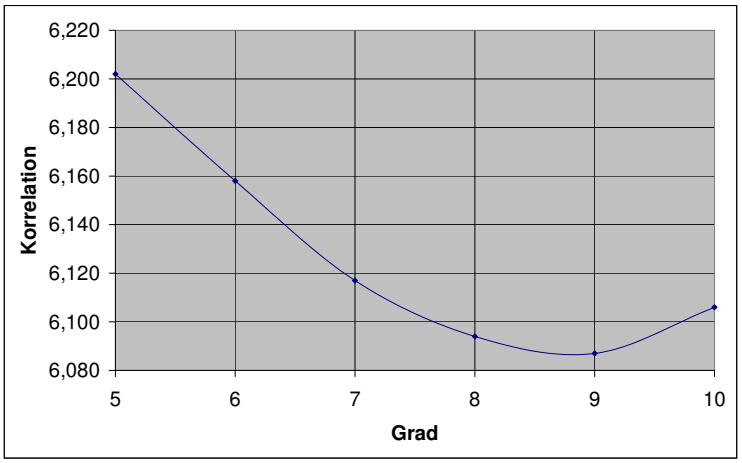

Bild 5.1: Grad Korrellation bei grundlegendem Aufbau

Auch die Grad Korrellation in Bild 5.1 bestätigt diesen Trend. Alle gemessenen Werte liegen bei ca. 6. Das bedeutet das die zufällige Verteilung der Kanten auf die Knoten funktioniert hat.

5.2 Remindin Klassik

5.2.1 Versuchsaufbau und Versuchsdurchführung

In einem ersten Experiment soll der Remindin-Algorithmus auf dem DMOZ-Datensatz angewendet werden, wie er in Kapitel 2 bzw. Kapitel 4 beschrieben worden ist. Alle anderen Parameter dieses Experiments sind dieselben, wie bereits in Kapitel 5.1 beschrieben.

Da die Anzahl der Knoten im Netzwerk gleich bleibt und nur neue Kanten durch Bootstprapping hinzukommen, wird der *Clusterkoeffizient*, die *Group Degree Centrality*, die *Group Closeness Centrality* sowie der *Mean Degree* steigen und der Degree Of Separation sinken.

ID	***Thema***
178279	Top/Arts/Television/Programs/Soap_Operas/Days_of_Our_Lives/Fan_Fiction
255279	Top/Business/Food_and_Related_Products/Frozen
380279	Top/Regional/Oceania/Australia/Victoria/Localities/C/Chewton
378279	Top/Regional/Oceania/Australia/Victoria/Localities/E/Elmhurst
382279	Top/Regional/Oceania/Australia/Victoria/Localities/M/Mernda
205279	Top/Science/Technology/Food_Science/Publications
119279	Top/Shopping/Antiques_and_Collectibles/Ephemera/Maps
327279	Top/Shopping/Home_and_Garden/Climate_Control/Fireplaces/Chimeneas
329279	Top/Shopping/Jewelry/Diamonds/Diamond_Jewelry
246279	Top/Society/Issues/Economic/Monopolies_and_Oligopolies/Microsoft
244279	Top/Society/Military/Veterans/Vietnam_War/Naval_Activities/Aviation
211279	Top/Society/Organizations/O

Tabelle 5.2: Beispiele der an das soziale Netzwerk gestellten Anfragen

Es wurden an jeden Knoten im Netzwerk genau 202 zufällig ausgewählte Anfragen aus verschiedenen Themengebieten gestellt. Einen Auszug aus den gestellten Anfragen ist in Tabelle 5.2 abgebildet. Das Prinzip der Anfragestellung ist dabei, dass der fragende Knoten ein Thema vorgibt, zu dem er einen Fachmann sucht. Dabei werden nur solche Editoren berücksichtigt, die Fachmann in exakt dem Themengebiet das vom Fragesteller vorgege-

ben wurde. Ein ähnliches Themengebiet wird nicht berücksichtigt. Beispielsweise können Anfragen zum Thema '*Fußball*' nicht von einem Editor mit dem Fachgebiet '*FC Bayern München*' beantwortet werden.

5.2.2 Ergebnisse und Diskussion

Metriken	***Werte***
globaler Clusterkoeffizient	$\approx 28,05\%$
Group Degree Centrality	$\approx 8,58\%$
Mean Degree	$\approx 6,61$
Group Closeness Centrality	$\approx 17,47\%$
Degree Of Separation	$\approx 6,98$
Neue Kanten	1686

Tabelle 5.3: Messergebnisse nach Anwendung des klassischen REMINDIN-Algorithmus

Alle im Versuchsaufbau erwarteten Ergebnisse sind eingetreten. Nur der *Clusterkoeffizient* ist nicht gestiegen sondern leicht gefallen.

Aus diesem Fakt sowie dem starken Anstieg der *Group Degree Centrality* kann geschlossen werden, dass sich verschiedene Experten für verschiedene Interessensgebiete herausgebildet haben. Diese Experten besitzen Kanten zu Personen, die von Ihnen eine Auskunft erhalten haben. Diese Personen sind aber untereinander nicht vernetzt. Daher ist der *lokale Clusterkoeffizient* für diese Experten auch besonders niedrig. Dieser Zusammenhang senkt dann auch den *globalen Clusterkoeffizienten*. Gleichzeitig haben solche Expertenkonstellationen eine hohe *Actor Degree Centrality*, da sie einen sterngraphartigen Aufbau besitzen. Damit bewirken Experten eine höhere *Group Degree Centrality*.

Auch die Grad Korellation in diesem Experiment (siehe Bild 5.2) bestätigt die Ergebnisse aus den anderen Metriken. Viele Knoten mit eher wenigen Kanten verbinden sich mit einigen wenigen Knoten die zentral liegen. Es bildet sich also die gleiche Struktur heraus wie im *Friends network* aus Bild 3.2.

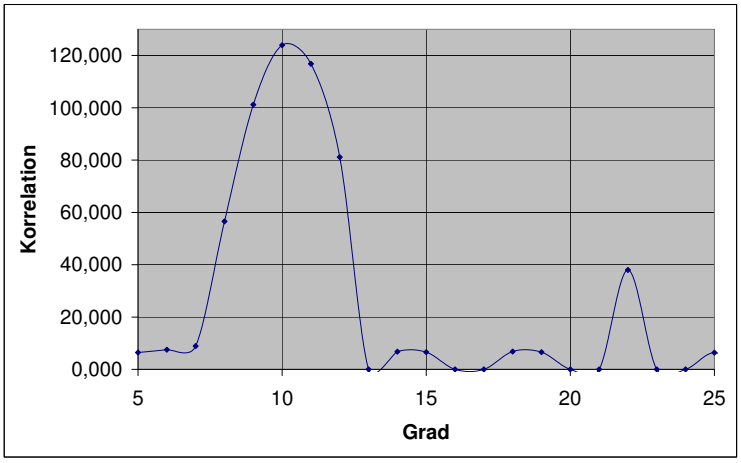

Bild 5.2: Grad Korrellation im Experiment 'Remindin Klassik'

5.3 Themenspezifische Fragen (*Grad: 1*)

5.3.1 Versuchsaufbau und Versuchsdurchführung

Um die Clusterbildung besser beobachten zu können, bietet es sich an nur Fragen zu einem Themenkomplex zu stellen. Bei diesem Experiment kommen die in Kapitel 4.1 vorgestellten themenspezifischen Metriken des ersten Grades zur Anwendung. *Kunst* ist das Themengebiet, auf das sich bei diesem Experiment konzentriert werden soll, denn an diesem Themengebiet ist die größte Gruppe von Editoren interessiert.

Der Ablauf des Experiments ist wie in Kapitel 4.1 beschrieben. Zuerst werden alle Editoren, die sich nicht für Kunst interessieren ausgeblendet. Dann werden die normalen Metriken zu Einsatz kommen. Als Nächstes wird der klassische REMINDIN-Algorithmus auf das Netzwerk mit allen Editoren angewendet. Allerdings werden hier keine 202 Fragen aus allen möglichen Themengebieten gestellt, sondern 202 Fragen alleine aus dem Themengebiet *Kunst*. Danach werden wieder alle Editoren ausgeblendet, die sich nicht für Kunst interessieren und die Metriken erneut angewendet. Durch einen Vergleich der Ergebnis-

se der beiden Messungen kann festgestellt werden, wie gut sich die kunstinteressierten Editoren gefunden haben.

5.3.2 Ergebnisse und Diskussion

Nach der Ausblendung aller Editoren, welche sich nicht für *Kunst* interessieren bleiben von insgesamt 5476 Editoren noch 884 Knoten übrig. Das entspricht $\approx 16,1\%$. Diese 884 Knoten sind mit insgesamt 408 Kanten verbunden. Die Metriken aus Kapitel 3 liefern das in Tabelle 5.4 beschriebene Ergebnis.

Metriken	***Werte***
globaler Clusterkoeffizient	$\approx 56,35\%$
Group Degree Centrality	$\approx 0,35\%$
mean degree	$0,92$
Group Closeness Centrality	nicht messbar
Degree Of Separation	nicht messbar

Tabelle 5.4: Messergebnisse nach Ausblendung aller nicht kunstinteressierten Editoren.

Aufgrund der zufälligen Verteilung der Editoren auf die verschiedenen Knoten des Netzwerks, gibt am Anfang der Simulation keine größere Zentralität einzelner Knoten. Dafür ist die Clusterbildung hoch, da Knoten die nur einen Nachbarn besitzen gleichzeitig den maximalen Clusterkoeffizienten besitzen, denn sie sind mit ihrem einzigen Nachbarn über genau eine Kante verbunden.

Da die Zahl der Kanten nur etwa halb so groß ist, wie die der Knoten, haben mindestens die Hälfte der Knoten keinen Kontakt zu einem anderen Knoten. Daraus folgt, dass auch kein Abstand zu diesen Knoten gemessen werden kann. Daher sind die *Group Closeness Centrality* sowie der *Degree Of Separation* nicht messbar, denn diese Maße basieren auf dem Abstand der Knoten zueinander.

Da in diesem Experiment nur wenige Kanten übrig geblieben sind, ist die Aussagekraft von Bild 5.3 gering. Auf Grund der Tatsache, dass die Werte für die Grade eng bei denen der Korrellation liegen handelt es sich hier um ein geschlossenes Netzwerk, ähnlich dem

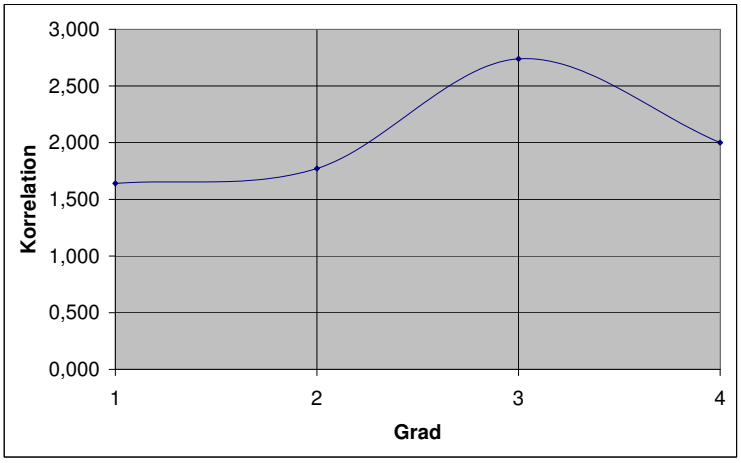

Bild 5.3: Grad-Korrellation bei themenspezifischen Fragen (Grad: 1)

Testimonial network aus Bild 3.2.

In Anbetracht der Tatsache, dass sich die meisten der Editoren für Kunst interessieren ist auch bei allen anderen Themen, für die sich eine relevante Zahl von Editoren interessieren, kein anderes Ergebnis für die verschiedenen Metriken zu erwarten.

Im nächsten Schritt wird der klassische REMINDIN-Algorithmus auf das gesamte Netzwerk, ohne Ausblendungen von einzelnen Editoren, angewendet. Dabei werden jedem Editor, wie bereits in Kapitel 5.2, genau 202 Fragen gestellt. Diese Fragen wurden per Zufallsgenerator vor der Anwendung des Algorithmus ausgewählt. Dabei entstehen insgesamt 34 neue Kanten. Nach der Anwendung des Algorithmus werden die Editoren, die sich nicht für Kunst interessieren wiederum ausgeblendet. Danach werden wieder die verschiedenen Metriken angewendet. Das Ergebnis dieser Messungen ist in Tabelle 5.5 nachzulesen.

Das neue soziale Netzwerk zeigt keine signifikanten Veränderungen zum sozialen Netzwerk vor der Anwendung des REMINDIN-Algorithmus. Die Tendenzen, die bei Anwendung des klassischen REMINDIN-Algorithmus in Kapitel 5.2 festgestellt wurden, treten

Metriken	***Werte***
globaler Clusterkoeffizient	$\approx 56,10\%$
Group Degree Centrality	$\approx 0,57\%$
mean degree	$0,93$
Group Closeness Centrality	nicht messbar
Degree Of Separation	nicht messbar
Neue Kanten	5

Tabelle 5.5: Messergebnisse nach Anwendung des REMINDIN-Algorithmus

hier nur in sehr stark abgeschwächter Form auf. Das insgesamt nur 34 neue Kanten erzeugt wurden von denen lediglich fünf zwischen zwei Kunst-Editoren verlaufen, erklärt die geringe Veränderung der Messergebnisse.

Da im Verlaufe dieses Experiments nur wenige Kanten hinzugekommen sind, zeigt auch die Grad Korrellation in Bild 5.4 keine signifikante Veränderung zur Grad Korrellation in Bild 5.3.

Wenn dieses Ergebnis in die reale Welt übertragen wird bedeutet es, dass Menschen mit einem speziellen Interessensgebiet nicht zueinander finden können, wenn die Kontakterstellung nur über dieses eine Thema funktioniert. Daher sind die Kontakte, die über andere Themen geknüpft wurden zwingend notwendig. Ein reales Beispiel für diesen Sachverhalt ist, dass sich zwei Personen, die sich für Handarbeiten interessieren sich nie gefunden hätten, wenn sie nicht auch die Kontakte von Personen aus ihrer Nähe (z.B. Ehepartner, die sich aus dem Sportverein kennen) nutzen könnten.

5.4 Themenspezifische Fragen (*Grad: 2*)

5.4.1 Versuchsaufbau und Versuchsdurchführung

Dieses Experiment läuft ab, wie das Experiment 5.3. Der einzige Unterschied ist das hier themenspezifische Metriken zweiten Grades zur Anwendung kommen sollen. Auch hier sind die gleichen Ergebnisse zu erwarten, wie bei den Experimenten des ersten Grades.

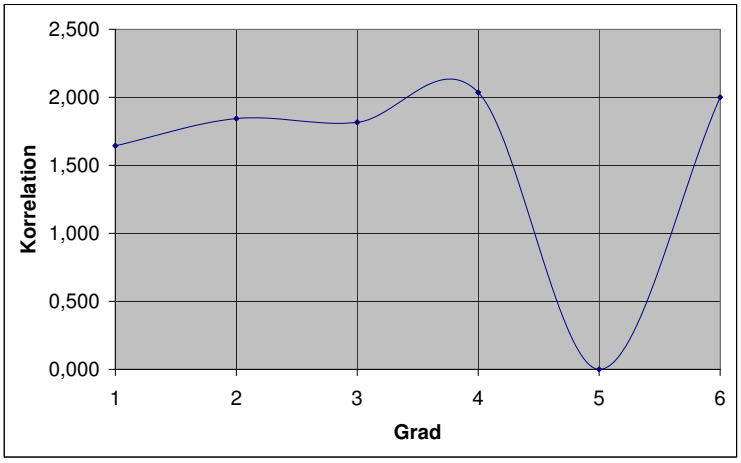

Bild 5.4: Grad Korrellation nach Anwendung des REMINDIN-Algorithmus

Wenn sich Personen mit gleichen Interessen schon kennen, brauchen sie untereinander keine neue Verbindungen mehr aufzubauen. Wenn sie sich noch nicht kennen, haben sie schlechte Chancen sich kennenzulernen. Das liegt daran, dass keine Kontakte zu anderen Personen, die sich nicht für Kunst interessieren aufgebaut werden. Diese Personen werden nicht erreicht, da nur Fragen zum Thema Kunst gestellt werden (siehe Experiment 5.3).

5.4.2 Ergebnisse und Diskussion

Nach der Ausblendung aller Editoren, welche sich nicht für *Kunst* interessieren bleiben von insgesamt 5476 Editoren auch in diesem Experiment noch 884 Knoten übrig. Das entspricht $\approx 16,1\%$. Diese 884 Knoten sind mit insgesamt 1842 Kanten verbunden. Die Zahl der Kanten ist in diesem Experiment größer, da auch Verbindungen zu kunstinteressierten Nachbarn zweiten Grades geknüpft werden. Die Metriken aus Kapitel 3 liefern das in Tabelle 5.6 beschriebene Ergebnis.

Aufgrund der zufälligen Verteilung der Editoren auf die verschiedenen Knoten des Netz-

Metriken	***Werte***
globaler Clusterkoeffizient	$\approx 45,19\%$
Group Degree Centrality	$\approx 1,12\%$
Mean Degree	$\approx 4,17$
Group Closeness Centrality	nicht messbar
Degree Of Separation	$\approx 6,76$

Tabelle 5.6: Messergebnisse nach Ausblendung aller nicht kunstinteressierten Editoren.

werks, gibt es zu Anfang der Simulation keine größere Zentralität einzelner Knoten. Dafür ist die Clusterbildung aber hoch, da Knoten, die nur wenige Nachbarn besitzen gleichzeitig einen hohen Clusterkoeffizienten besitzen.

Aufgrund der Tatsache, dass die Werte für die Grade eng bei denen der Korrellation liegen (siehe 5.5), handelt es sich auch hier um ein geschlossenes Netzwerk, ähnlich dem *Testimonial network* aus Bild 3.2. Das hier ein geschlossener Ring entstanden ist erklärt sich daraus, dass hier nur die Editoren betrachtet werden, die sich für ein und das selbe Thema interessieren.

Metriken	***Werte***
globaler Clusterkoeffizient	$\approx 44,62\%$
Group Degree Centrality	$\approx 1,11\%$
Mean Degree	$\approx 4,23$
Group Closeness Centrality	nicht messbar
Degree Of Separation	$\approx 6,76$
Neue Kanten	26

Tabelle 5.7: Messergebnisse nach Anwendung des klassischen REMINDIN-Algorithmus

Auch hier zeigt das soziale Netzwerk zeigt keine signifikanten Veränderungen zum sozialen Netzwerk vor der Anwendung des REMINDIN-Algorithmus. Die Tendenzen, die bei Anwendung des klassischen REMINDIN-Algorithmus in Kapitel 5.2 festgestellt wurden, treten hier nur in sehr stark abgeschwächter Form auf. Das insgesamt nur 210 neue Kanten erzeugt wurden, von denen lediglich 26 zwischen zwei Kunst-Editoren verlaufen, erklärt die geringe Veränderung der Messergebnisse.

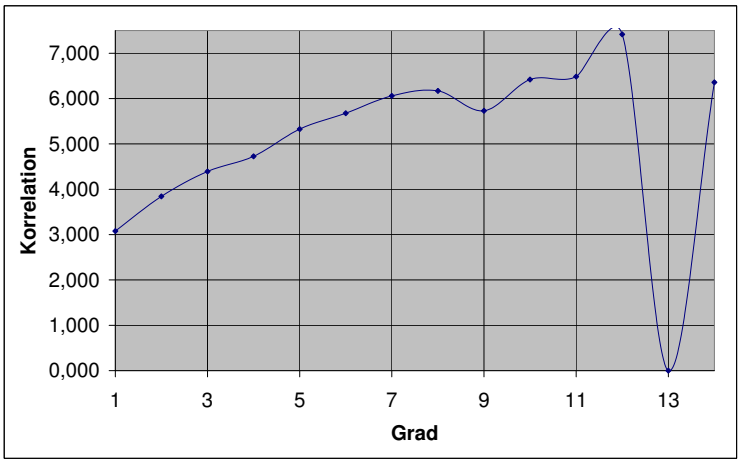

Bild 5.5: Grad Korrellation bei themenspezifischen Fragen (Grad: 2)

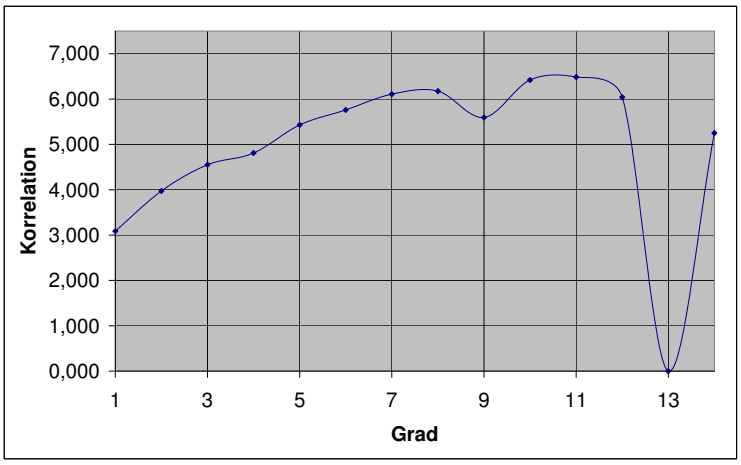

Bild 5.6: Grad Korrellation nach Anwendung des REMINDIN-Algorithmus

Da im Verlaufe dieses Experiments nur wenige Kanten hinzugekommen sind, zeigt auch die Grad Korrellation in Bild 5.6 keine signifikante Veränderung zur Grad Korrellation in Bild 5.5.

Damit ergibt sich kein signifikanter Unterschied zwischen den Werten der themenspezifischen Metriken ersten und zweiten Grades.

5.5 Lernsimulation

5.5.1 Versuchsaufbau und Versuchsdurchführung

Im Unterschied zu Experiment 5.2 soll es hier möglich sein, dass eine Person, wenn sie eine Anfrage zu einem Thema gestellt und eine passende Antwort bekommen hat, selbst zu einem *Fachmann* für dieses Themengebiet wird. Alle anderen Parameter sollen sich gegenüber 5.2 nicht ändern. Das Experiment simuliert den in der realen Welt vorhandenen Lerneffekt, denn eine Person merkt sich Antworten auf Fragen, die sie einmal erhalten hat und muss nicht immer erneut die selbe Frage stellen.

Bei diesem Experiment sind die selben Ergebnisse zu erwarten, wie bei Experiment 5.2. Die einzige Abweichung ist bei den Zentralitätsmaßen *Group Degree Centrality* und *Group Closeness Centrality* zu erwarten, da hier das Wissen einer Person weiter verbreitet wird, hat die nächste Person, die die gleiche Anfrage stellt schon zwei Personen zur Auswahl haben, von denen sie eine Antwort auf ihre Anfrage erhalten könnte. Daher knüpfen nicht mehr so viele Personen Beziehungen zu einer einzigen Person, denn die Informationen können auch über andere Personen erlangt werden. Daraus folgt, dass die Zentralität der einzelnen Knoten abnimmt.

5.5.2 Ergebnisse und Diskussion

Die Ergebnisse liegen im Bereich, der in 5.5.1 erwartet wurde. Weil im Verhältnis zu 5.2 nur wenige neue Kanten hinzugefügt wurden, haben sich *Mean Degree* und *Degree Of Separation* nur leicht gegenüber dem grundlegenden Aufbau (siehe 5.1) geändert.

Metriken	***Werte***
globaler Clusterkoeffizient	$\approx 28,99\%$
Group Degree Centrality	$\approx 0,66\%$
Mean Degree	$\approx 6,04$
Group Closeness Centrality	$\approx 5,79\%$
Degree Of Separation	$\approx 7,72$
Neue Kanten	119
Neue Fachleute	205

Tabelle 5.8: Messergebnisse für das Netzwerk mit zusätzlichen Fachleuten.

Das es jetzt 205 Fachleute zusätzlich gibt, hat dazu geführt, dass die Zentralität gegenüber dem grundlegenden Graphen (siehe 5.1) stärker ausgeprägt ist. Gegenüber dem klassischen REMINDIN-Algorithmus allerdings, ist die Zentralität nicht stark ausgeprägt. Daraus lässt sich folgender Zusammenhang ableiten: ***Je weiter das Wissen der einzelnen Personen steigt, desto mehr nehmen die Kontakte ab***.

In die reale Welt übertragen bedeuten die Ergebnisse dieses Experiments, dass Personen mit geringem Wissen verstärkt auf die Bildung von sozialen Kontakten angewiesen sind. Das gilt ebenfalls für Personen mit geringen handwerklichen Fähigkeiten oder mit beschränkten finanziellen Möglichkeiten. Dieses Phänomen zeigt sich z.B. bei der Bildung von Gewerkschaften und anderer Lobbygruppen. Es konnte auch in den Mangelwirtschaften des früheren Ostblocks beobachtet werden. In der ehemaligen DDR konnten viele Produkte nur mit Kontakten in die westliche Welt oder in die obere Schicht der Gesellschaft bezogen werden. Daher hatten die Besitzer dieser Kontakte viele 'Freunde'. Wer alle Produkte beziehen wollte, musste viele Personen kennen.

Daraus resultierte für die in der DDR lebenden Personen die zwingende Notwendigkeit ihre Gesellschaft mit einer hohen Zentralität, einem hohen *Mean Degree* und einem geringen *Degree Of Separation* zu organisieren. Damit bildete die von der damaligen Staatsführung verschuldete, katastrophale wirtschaftliche Lage gleichzeitig eine der Grundlagen für das System der staatlichen Überwachung und Bespitzelung. Denn eng vernetzte Gesellschaften lassen sich viel leichter kontrollieren, da fast alle Personen über wenige zentrale Personen (sog. Multiplikatoren) erreichbar sind. Als nach dem Ende der DDR die wirt-

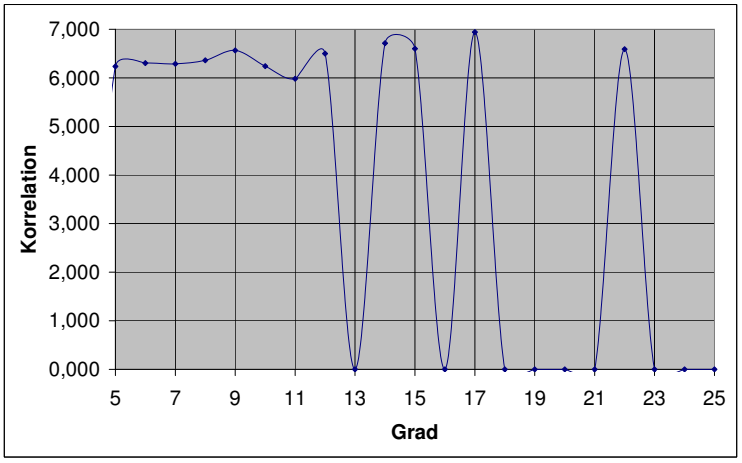

Bild 5.7: Grad Korrellation bei Fachmann-Experiment

schaftlichen Verhältnisse besser wurden, waren all diese Kontakte nicht mehr zwingend nötig. Daher hätten nach den Ergebnissen dieses Experimentes die Kontakte einschlafen müssen. Auch dieses Phänomen ist seit der Wiedervereinigung Deutschlands in der ehemaligen DDR zu beobachten.

Da in diesem Experiment das Wissen auf alle Editoren verteilt wird, haben sich keine signifikanten Schwankungen in der Grad Korrellation (siehe Bild 5.7) ergeben, denn wenn es viele Experten gibt, besteht nicht die Notwendigkeit eine spezielle Personen zu kontaktieren, um eine Anfrage beantwortet zu bekommen.

Kapitel 6

Zusammenfassung

6.1 Ergebnisse der Arbeit

Neben dem im Rahmen dieser Arbeit entwickelten, auf dem REMINDIN-Algorithmus basierenden Simulator und den in Kapitel 5 durchgeführten Experimenten sind vor allem die themenspezifischen Metriken, die in Abschnitt 4.1 entwickelt wurden, als Ergebnis dieser Arbeit zu betrachten.

Bei der Anwendung dieser Metriken hat sich herausgestellt, dass sehr wenige neue Kontakte entstehen, wenn Personen nur über ein einziges Themengebiet miteinander sprechen. Damit ist festzustellen, dass es keine Fachgebiete gibt, die von geringerem Nutzen sind als andere, weil sie beispielsweise weniger Gewinn erwirtschaften. Damit Menschen, die sich etwas zu sagen haben zueinander finden ist es notwendig, dass sie auch mit Menschen kommunizieren, die nicht zur Lösung ihres Problems beitragen können.

6.2 Ausblick

Um die themenspezifischen Metriken besser Nutzen zu können, ist es notwendig eine Strategie zum Umgang mit Personen zu finden, die keinen Kontakt zu anderen Personen besitzen. Hier muss herausgearbeitet werden, ob der Grad der themenspezifischen Metriken

beliebig erhöht werden kann. Das Entstehen von unerreichbare Personen würde damit verhindert. Andererseits geht, wenn Verbindungen zwischen weit entfernt liegenden Personen geknüpft werden, ein Stück Realitätsnähe verloren. Ein anderer Ansatz, zur Lösung dieses Problems, besteht darin Personen, die keine Kontakte besitzen ebenfalls auszublenden.

Des Weiteren können wären neben den in Kapitel 5 durchgeführten Experimenten noch zusätzliche Experimente denkbar. Wegen der zeitlichen Begrenztheit der Arbeit musste auf diese verzichtet werden.

Beispielsweise wurde das Faktum, dass Menschen nur eine begrenzte Anzahl an Kontakten pflegen können in dieser Arbeit nicht berücksichtigt. Um dieses Experiment durchzuführen, müssen in die Simulation Obergrenzen für Content Provider, Recommender und für Kontakte zu anderen Personen im Netzwerk eingefügt werden. Bei Überschreitung der Obergrenzen werden die unwichtigsten Kontakte überschrieben. Dafür muss eine Definition des Begriffs Wichtigkeit getroffen werden. Dieses Vorgehen simuliert das Einschlafen von Kontakten in der realen Welt.

Ein weiteres, noch offenes Experiment ergibt sich, wenn in die Simulation einfließt, dass Themen Unterthemen haben können. So kann z.B. eine Frage über *Fußball*, mit großer Wahrscheinlichkeit, auch von Personen mit dem Fachgebiet *FC Bayern München* beantwortet werden. Umgekehrt kann eine Frage zu einem beliebigen Verein nicht von jeder Person mit dem Fachgebiet *Fußball* beantwortet werden.

Dem Forschungsinteresse im Bezug auf weitere Experimente sind keine Grenzen gesetzt.

Da alle in dieser Arbeit durchgeführten Experimente Ergebnisse lieferten, die in der Realität wiedererkannt werden, ergibt sich, dass der REMINDIN-Algorithmus von Christoph Tempich [Tem06] und das Small-World-Modell von Jon Kleinberg [Kle06] realitätsnahe Ergebnisse liefern. In wie weit die Ergebnisse tatsächlich der Realität entsprechen, muss noch geklärt werden. Das Entwickeln von realitätsabbildenden Algorithmen und Modellen fällt in den Bereich der Soziologie.

Literaturverzeichnis

[AHK+07] Yong-Yeol Ahn, Seungyeop Han, Haewoon Kwak, Sue Moon, and Hawoong Jeong. Analysis of topological characteristics of huge online social networking services. In *Proceedings of WWW*, 2007.

[BCK+07] Matthias Bender, Tom Crecelius, Mouna Kacimi, Sebastian Michel, Josiane Xavier Parreira, and Gerhard Weikum. Peer-to-peer information search: Semantic, social, or spiritual? Technical report, Max-Planck Institute for Iformatics, Saarbruecken, Germany, 2007.

[BK07] Phillippe Blanchard and Tyll Krüger. Die ausbreitung von korruption als verallgemeinerter edidemischer prozess. *Forschungsmagazin*, 1:72–77, 2007.

[Ebe07] Prof. Dr. Jürgen Ebert. *Softwaretechnik II*. Universität Koblenz-Landau, 2006/2007.

[Gru06] Miriam Grunwald. Small world phenomenon. In *Seminar Analyse komplexer Informationssysteme - Wie Leute über Dinge reden*. Universität Koblenz-Landau, 2006.

[Kle06] Jon Kleinberg. Complex networks and decentralized search algorithms. In *Proceedings of the International Congress of Mathematicians (ICM)*, 2006.

[LTQ+05] Alexander Löser, Christoph Tempich, Bastian Quilitz, Wolf-Tilo Balke, Steffen Staab, and Wolfgang Nejdl. Searching dynamic communities with personal indexes. In *Proceedings of the 4th International Semantic Web Conference (ISWC 2005), Galway, Ireland*, 2005.

[Mil67] Stanley Milgram. The small world problem. *Psychology Today*, 2:60–67, 1967.

[Mut04] Peter Mutschke. Autorennetzwerke: Verfahren der netzwerkanalyse als mehrwertdienste für informationssysteme. *IZ-Arbeitsbericht*, Nr. 32:11, 2004.

[NSW01] M. E. J. Newman, Steven H. Strogatz, and Duncan J. Watts. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E*, 64:026118, 2001.

[Tem06] Christoph Tempich. *Ontology Engineering and Routing in Distributed Knowledge Management Applications*. PhD thesis, Universität Fridericiana zu Karlsruhe, 2006.

[WF07] Stanley Wasserman and Kathrine Faust. *Social Network Analysis*. Cambridge University Press, 2007.

[WS98] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393:440–442, 1998.