Technische Universität Graz
Dekanat für Informatik
Institut für Wissensmanagement

Datenaufbereitung und Entwicklung einer Abbruchsimulation zur Modellierung von Benutzernavigation

Masterarbeit

von

Markus HORWATH, BSc

Vorgelegt zur Erlangung des
akademischen Grades eines Master
der Studienrichtung Informatik

Graz, im März 2013

Betreuer der Masterarbeit:

Assoc.Prof. Dipl.-Ing. Dr.techn. Denis HELIC

.....

Deutsche Fassung:

Beschluss der Curricula-Kommission für Bachelor-, Master- und Diplomstudien vom 10.11.2008

Genehmigung des Senates am 1.12.2008

EIDESSTATTLICHE ERKLÄRUNG

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Graz, am 14. März 2013 ……

(Unterschrift)

Englische Fassung:

STATUTORY DECLARATION

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

14th March 2013 ……

(signature)

Danksagung

Zu Beginn möchte ich mich bei Herrn Dipl.-Ing. Dr.techn. Denis Helic bedanken, der als Betreuer während des Verfassens dieser Arbeit mit Rat und Tat zur Seite stand und mir diese Masterarbeit ermöglichte.

Ein Dank gilt auch Herrn Dipl.-Ing. Dr.techn. Markus Strohmaier, der mir durch die Einladung zu den wöchentlichen Treffen am Institut die Gelegenheit für interessante Einblicke in die Forschungsgebiete und das wissenschaftliche Arbeiten bot. Dadurch konnte ich zahlreiche Studierende und deren wissenschaftliche Arbeitsbereiche kennenlernen.

Da diese Arbeit auch den Abschluss meines Masterstudiums darstellt, gilt ein besonderer Dank auch meinen Studienkollegen Florian Geigl und Michael Eder, mit denen ich von Beginn an zahlreiche Lehrveranstaltungen meines Studiums erfolgreich absolvieren konnte.

Abschließend gilt ein außerordentlicher Dank meiner Familie, im Besonderen meinen Eltern Helga und Fritz Horwath, die mir das Studium ermöglicht haben.

Bei allen Bezeichnungen, die auf Personen bezogen sind, meint die Formulierung beide Geschlechter, unabhängig von der in der Formulierung verwendeten konkreten geschlechtsspezifischen Bezeichnung.

Kurzfassung

Der Trend der letzten Jahre im World Wide Web zeigte, dass Inhalte immer dynamischer und kurzlebiger werden. Jeder Anwender kann im sogenannten Web 2.0 Daten und Informationen beisteuern. Die rasante Verbreitung und die häufigen Veränderungen wirken sich auch maßgeblich auf die Struktur und die Entwicklung des Netzwerkes aus. Im Rahmen dieser Masterarbeit dienen die freie Enzyklopädie Wikipedia sowie ein Spiel namens Wikigame, das auf dieser Wissensdatenbank basiert, als Grundlagen für Forschungstätigkeiten. Es ist von Interesse, das Navigationsverhalten von Benutzern in solchen dynamischen Netzwerken zu untersuchen.

Um Auswertungen des Benutzerverhaltens durchführen zu können, müssen alle verfügbaren Daten geordnet, aufbereitet und in einem geeigneten Format zur Weiterverarbeitung bereitgestellt werden. Dies betrifft im Allgemeinen das Wikipedia-Netzwerk, Vergleichsdaten aus dem Wikigame und die Modellierung von Hintergrundwissen. Letzteres wird zur Ausführung von Simulationen benötigt und durch hierarchische Strukturen beschrieben. Ein Teil dieser Arbeit beschreibt die Verfahren, die zur Anwendung gekommen sind, um Daten als Basis bereitstellen zu können. Der im Rahmen des gesamten Projekts entwickelte Simulator für Benutzernavigation bietet mehrere Schnittstellen, um das Verhalten von Benutzern in einem Netzwerk zu parametrisieren. Dazu zählt u.a. eine Beschreibung der Abbruchrate. Bisherige Publikationen zielten häufig darauf ab, dass Simulationen ein Navigationsziel besonders effizient und rasch erreichen. Das Ziel dieser Arbeit ist es, eine automatische Navigation möglichst gut an ein vorhandenes empirisches Datenset anzupassen, um sich dem Benutzerverhalten anzunähern.

Die Evaluierung zeigt, dass durch die simulierte Abbruchrate das Verhalten von Benutzern sehr gut angenähert werden kann und folglich durch eine automatische Simulation Daten ohne Benutzerinteraktion generiert werden können. Diese Navigationssimulationen bieten zahlreiche Auswertungsmöglichkeiten, mit denen in weiterer Folge auch die Durchführung von Bewertungen – beispielsweise für Netzwerke oder Hintergrundwissen – ermöglicht wird.

Abstract

The trend over the last few years in the World Wide Web showed that content becomes more dynamic and short-lived. Each user can contribute data and information in a so-called Web 2.0. Rapid spread and frequent changes essentially affect the structure and development of the network. In the present thesis the free encyclopedia Wikipedia and a game called Wikigame that is based on this knowledge database are used as a basis for research activities. It is of interest to study the navigation behavior of users in these dynamic networks.

In order to perform analyses of user behavior all available data has to be ordered, prepared and placed in a suitable format for further processing. This generally relates to the Wikipedia network, comparing data from the Wikigame and modeling of background knowledge. The latter one is needed to run simulations and is specified by hierarchical structures. One part of this thesis describes the procedures that were used in order to provide data as a basis. A simulator for user navigation that was developed as a part of this project provides several interfaces to parameterize the behavior of users in a network. This includes among others a description of an attrition rate. Previous publications often targeted that simulations reach a navigation target efficiently and quickly. The goal of this study is to bring an automatic navigation as close as possible to an existing empirical data set in order to approach the behavior of users.

The evaluation shows that the simulated attrition rate is very useful to approximate user behavior and therefore data can be generated by an automatic simulation without any user interaction. The navigation simulations offer numerous analysis options, that subsequently also allow the execution of assessments, for example for networks or background knowledge.

Inhaltsverzeichnis

1. Einführung

Dieses Kapitel soll einen kurzen Überblick über das Themengebiet dieser Arbeit geben. Nach der Motivation wird auf die Aufgabenstellung und den Aufbau der Arbeit näher eingegangen.

1.1 Motivation

Als das World Wide Web im Jahre 1989 an der Europäischen Organisation für Kernforschung, kurz CERN, als Projekt entstand, ahnte noch niemand, welche Ausmaße es annehmen wird. Die Verwendung war ursprünglich nur dafür gedacht, Forschungsergebnisse auszutauschen. Das Projekt zielte speziell darauf ab, Inhalte miteinander zu vernetzen und mit sogenannten Hyperlinks zu verknüpfen. [1]

Während es Versuche vor 1989 nicht schafften, ein derartiges Projekt zu verbreiten, erfreute sich das World Wide Web, kurz WWW oder Web, immer größerer Beliebtheit. Wurden einst nur statische Inhalte, hauptsächlich Texte, im WWW angeboten, so kamen im Laufe der Zeit auch Grafiken und Bilder hinzu. Durch die kontinuierlich wachsende Verbreitung und die Erhöhung der Bandbreiten für Internetzugänge, veränderte sich auch der Inhalt. Neben Audiodateien wurden u.a. auch Videodaten angeboten. Bedingt durch das Wachstum, nahm auch die Vernetzung stark zu. Das Netzwerk wurde immer dichter.

Immer mehr Benutzern aus verschiedensten Arbeitsumfeldern war es möglich, Zugang zum World Wide Web zu erhalten. Durch die breitflächige Verwendung musste die Bedienung des Mediums stets intuitiv bleiben und u.a. den zeitlichen Veränderungen angepasst werden.

Da sich das Medium rasant weiterentwickelte, kam recht bald der Wunsch auf, dass auch Benutzer aktiv in die Inhalte des WWW eingreifen können. Das Web 2.0 war geboren. Die Inhalte wurden immer dynamischer und kurzlebiger, weil jeder Benutzer beliebigen Inhalt auf einfache Art und Weise veröffentlichen konnte. Die angeboten Web-2.0-Dienste sind heutzutage kaum überschaubar,

sie reichen von Video-Plattformen wie YouTube[1], über Blogs- und Wiki-Systemen bis hin zu sozialen Netzwerken wie Facebook[2]. [2]

Wiki-Systeme, in denen der Inhalt von einer Vielzahl an Benutzern frei bearbeitet werden kann, unterliegen häufigen Veränderungen. Nachdem jeder Anwender Informationen beisteuern kann, wirken sich Änderungen auch maßgeblich auf die Struktur und die Entwicklung des Netzwerkes aus. An genau dieser Stelle ist es nun von Interesse, das Verhalten von Benutzern zu untersuchen.

1.2 Themengebiet

Die Fragestellungen, die im Bereich Benutzerverhalten aufgeworfen werden, sind breit gefächert. Sie reichen von der generellen Untersuchung der Netzwerk-Struktur bis zur Nachmodellierung eines Benutzers.

Um Auswertungen über das Anwenderverhalten durchführen zu können, muss eine Analyse über aufgezeichnetes Navigationsverhalten vorgenommen werden. Aus den Erkenntnissen kann dann durch Parametrisierung versucht werden, sich dem Benutzerverhalten anzunähern. Die zugrundeliegenden Daten liefern die Basis für alle darauf aufsetzenden Analyseverfahren und Auswertungen.

In weiterer Folge wäre es sogar möglich, basierend auf den Auswertungen, die Netzwerkstruktur dahingehend anzupassen, um effizienteres Navigieren zu ermöglichen und die Benutzerfreundlichkeit zu erhöhen.

1.3 Aufgabenstellung

Ein Ziel dieser Masterarbeit ist es, für ein bestehendes Navigationsframework alle benötigten Eingabedaten zu liefern. Dies bezieht sich sowohl auf die Daten, die das Netzwerk an sich betreffen und beschreiben, als auch auf jene, die das aufgezeichnete Benutzerverhalten beinhalten. Neben der Bestandsanalyse von

[1] http://www.youtube.com
[2] http://www.facebook.com

existierenden Daten lagen hier vor allem die Augenmerke bei der Selektion, der Strukturierung und der Aufbereitung.

Des Weiteren wird eine Schnittstelle des Frameworks für Abbruchraten verwendet. Durch die Einführung von verschiedenen Typen, kann durch den Abbruchsimulator das menschliche Abbruchverhalten in der Navigation modelliert werden.

Parallel zu diesen beiden Bereichen fand auch das Entwickeln von diversen Tools statt, die u.a. durch das iterative Vorgehen entstanden sind. Diese wurden benötigt, um diverse Eingabedaten zu generieren, die vom Navigationsframework benötigt werden.

1.3.1 Projektübersicht

Aufgrund des Umfangs dieser Forschungsarbeit sind daraus neben dieser Masterarbeit auch noch zwei weitere Masterarbeiten, [3] und [4], entstanden. Abbildung 1 zeigt eine Übersicht des Gesamtprojekts. Der Ablauf und die Gliederung in die einzelnen Bereiche sind aus dem Flussdiagramm ersichtlich. Rot markiert sind darin jene Bereiche, die diese Masterarbeit betreffen. Blau hinterlegt sind Bereiche, in denen zusammengearbeitet wurde um gemeinsame Lösungen zu finden.

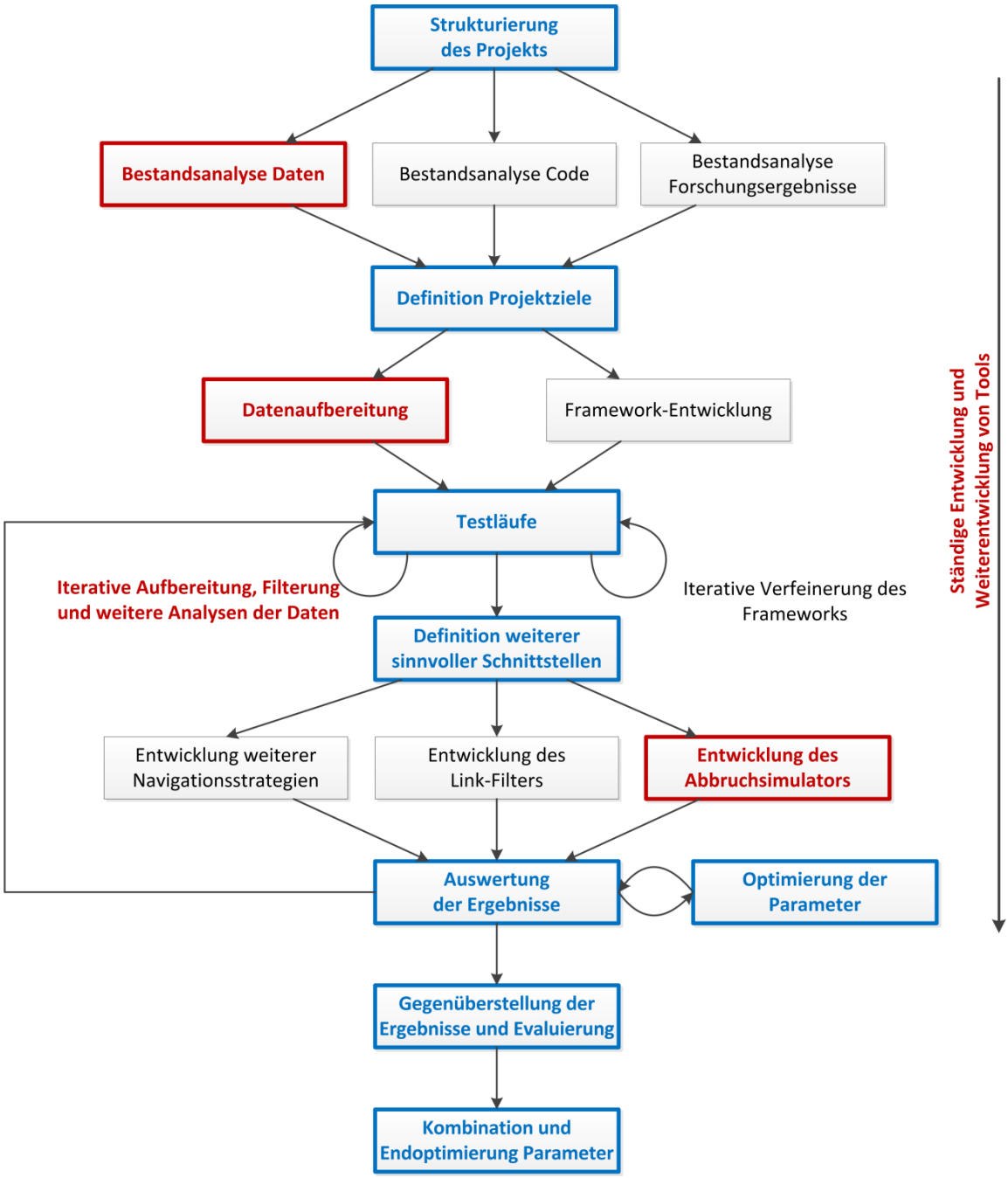

Abbildung 1: Projektübersicht

1.4 Aufbau der Arbeit

Im Allgemeinen besteht diese Arbeit aus neun Kapiteln. Nach diesem Einführungskapitel werden in Kapitel 2 die Grundlagen besprochen, um einen tieferen Einblick in das Themengebiet zu geben. Dazu zählen u.a. die Graphentheorie, die Navigation und nähere Ausführungen zur Enzyklopädie Wikipedia und dem Spiel Wikigame.

Im Kapitel 3 wird über bereits veröffentlichte Arbeiten geblickt. Dieser Abschnitt soll Einblicke in Forschungsergebnisse geben, die in einem Zusammenhang mit dieser Arbeit stehen.

Einen Überblick über die benötigten Daten und deren Repräsentation gibt Kapitel 4. Das darauffolgende Kapitel 5 gibt Einsicht in die verwendete Arbeitsumgebung, wie dem Framework und der benutzten Bibliothek.

Kapitel 6 beschäftigt sich mit der Datenaufbereitung. Beginnend vom Ursprung der Daten, über Analyse, Extrahierung und Erzeugung bis zur Verwendung für das Projekt wird darin auf alle benötigten Schritte und Tools Bezug genommen.

Die Schnittstellenerweiterung zur Entwicklung des Abbruchsimulators wird in Kapitel 7 beschrieben. Darin wird auf die verschiedenen Ansätze und die Evaluierung der Ergebnisse eingegangen.

In Kapitel 8 erfolgt noch eine Conclusio der Inhalte dieser Arbeit, in der auch das Fazit gezogen wird.

Letztendlich widmet sich Kapitel 9 Ansätzen, die in weiterführende Forschungsarbeiten übergehen.

2. Grundlagen

In diesem Kapitel werden grundlegende Begriffe erläutert, auf die in den folgenden Abschnitten dieser Arbeit Bezug genommen wird. Sie dienen dem besseren Verständnis, da die weiteren Inhalte darauf aufbauen.

2.1 Graphentheorie

Die Graphentheorie als Teilgebiet der Mathematik beschäftigt sich mit dem Aufbau und den Eigenschaften von Graphen, in denen Beziehungen zwischen diesen eine wichtige Rolle spielen. Viele mathematische Ansätze lassen sich mittels Graphen repräsentieren und bieten somit zahlreiche Vorteile. Zu diesen zählen beispielsweise die einfache Beschreibung oder die übersichtliche Visualisierbarkeit der Daten.

2.1.1 Allgemeines

Die der Graphentheorie zugrunde liegende Struktur nennt man *Graph*. Darunter versteht man eine abstrakte Darstellung einer Menge von Objekten und deren Verbindungen zueinander. Ein Graph setzt sich grundlegend aus *Knoten* und *Kanten* zusammen. Mit Knoten werden Objekte modelliert, die man auch als Punkte oder Ecken in einer Visualisierung bezeichnet. Jeweils zwei Knoten können durch Kanten miteinander verbunden werden, die häufig durch Linien oder Pfeile repräsentiert werden. Zwei Beispiele der graphischen Darstellung finden sich in Abbildung 2. Darin sind Knoten blau markiert und mit fortlaufender Nummerierung versehen, Kanten werden durch grüne Verbindungen repräsentiert. [5]

Bei den Kanten kann zwischen zwei Typen unterschieden werden. Ein *gerichteter Graph* besitzt Kanten, die von einem Knoten s auf einen Knoten t zeigen. Die Repräsentation erfolgt als Pfeil. Es kann auch eine weitere Kante von Knoten t wieder auf Knoten s zeigen, d.h. in die entgegen gesetzte Richtung. In *ungerichteten Graphen* ist die Richtung einer Kante nicht vorgegeben. Existiert eine Kante von Knoten s zu Knoten t, kann ein Übergang sowohl von s zu t als

auch von t zu s erfolgen. Eine Gegenüberstellung beider Graph-Typen ist in Abbildung 2 ersichtlich. [5]

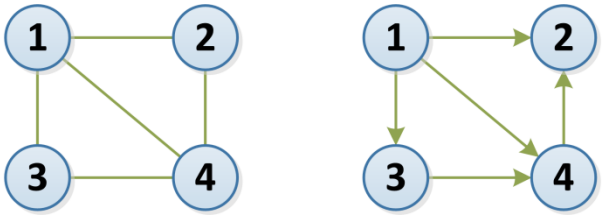

Abbildung 2: Ungerichteter Graph (links) und gerichteter Graph (rechts)

Aus den gegebenen Definitionen kann eine Kante nicht einen Knoten mit sich selbst verbinden, d.h. es ist keine Kante von Knoten s zu s möglich. Mehrfachkanten zwischen zwei Knoten sind ebenfalls nicht zulässig. Sollten die beiden letzten erwähnten Eigenschaften, in Abbildung 3 rot dargestellt, auf den Graphen zutreffen, so spricht man von einem *Multigraph*. In dieser Arbeit wird jedoch immer von *einfachen Graphen* ohne Mehrfachkanten sowie ohne rückbezügliche Kanten ausgegangen. [6]

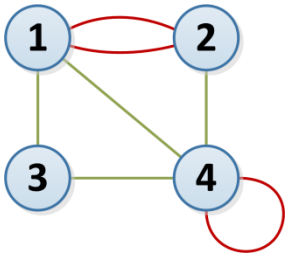

Abbildung 3: Multigraph

Ein weiteres Unterscheidungsmerkmal ist die *Endlichkeit* von Graphen. Besitzt der Graph eine endliche Anzahl an Knoten und Kanten, spricht man infolgedessen auch von einem endlichen Graphen. Unendliche Graphen werden häufig als Beschreibung für komplexere Problemstellungen benutzt und dienen eher als theoretisches Konstrukt. Wenn der Begriff Graph ohne weiteren Zusatz verwendet wird, ist im Allgemeinen immer von einer endlichen Struktur auszugehen. Dies trifft auch auf die Formulierungen in dieser Masterarbeit zu. [7]

Als *Weg* wird eine Folge von Knoten bezeichnet, die nacheinander besucht werden. Hierbei spricht man von einem Weg von Knoten s zu Knoten t, wenn s der Startknoten dieses Weges ist und t der Zielknoten desselben Weges. Be-

ginnend von s können alle ausgehenden Kanten gewählt werden, um zu einem nächsten Knoten im Graphen zu gelangen, der in den Weg aufgenommen wird. Nach einer bestimmten Anzahl an Schritten und besuchten Knoten im Graphen, endet dieser Weg schließlich in t. Es kann in einem Graphen nur ein Weg von s nach t existieren, wenn eine entsprechende Kantenfolge vorhanden ist. Daraus lässt sich auch das Erreichbarkeitsproblem in Graphen formulieren. Knoten t ist ausgehend von Knoten s nur dann erreichbar, wenn ein Weg existiert, andernfalls ist t nicht erreichbar. In den Graphen in Abbildung 4 ist von Knoten 11 ausgehend beispielsweise kein anderer Knoten erreichbar. [7] [8]

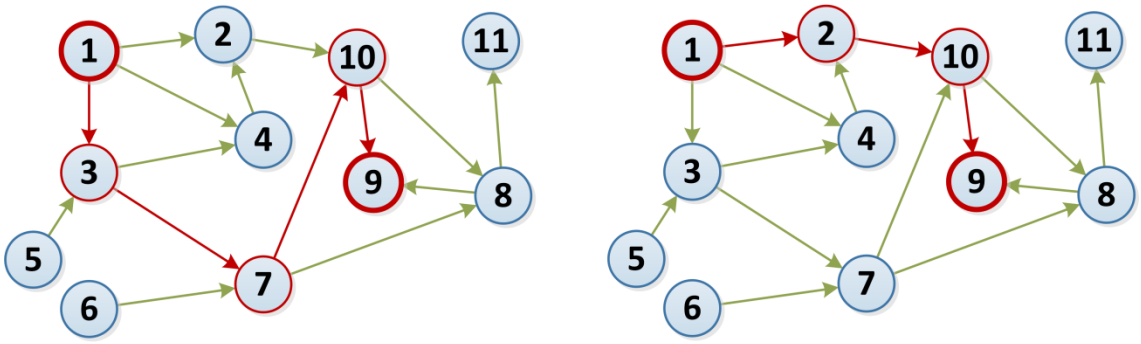

Abbildung 4: Zwei unterschiedliche Pfade mit selben Start- und Zielknoten

Ein *Pfad* bezeichnet einen Weg in einem Graphen, bei dem alle Knoten in einer Folge verschieden sind. In diesem Fall darf ein bereits besuchter Knoten kein weiteres Mal in die Folge aufgenommen werden. Abbildung 4 zeigt zwei mögliche Pfade zwischen dem Startknoten 1 und dem Zielknoten 9. Sowohl der Pfad, der die Knoten 1, 3, 7, 10 und 9 beinhaltet, als auch jener mit den Knoten 1, 2, 10 und 9 ist gültig. [7]

Da in den weiteren Kapiteln dieser Arbeit eine Unterscheidung zwischen Wegen und Pfaden nicht notwendig ist, wird der Begriff Pfad als Überbegriff angesehen, bei dem auch ein Wiederbesuch bereits hinzugefügter Knoten zum Weg erlaubt ist. Auch in Literaturquellen herrscht keine Eindeutigkeit der Definitionen.

Wenn es aus den Anforderungen nötig ist, können Kanten in einem Graphen auch mit *Gewichten* oder *Farben* versehen werden. Von Gewichten spricht man, wenn eine Kante mit einer rationalen oder reellen Zahl versehen ist. Dafür

lassen sich auch sogenannte Gewichtsfunktionen definieren. Bei einer Färbung von Kanten bzw. Knoten wird diesen jeweils eine Farbe zugeordnet. Je nach Anwendungsfall kann das Gewicht einer Kante z.B. auch in den Weg mit aufgenommen bzw. berücksichtigt werden. [8]

Als *Distanz* bzw. *Länge* eines Pfades wird die Anzahl der Kanten, die dieser Pfad beinhaltet, bezeichnet. Um nochmals auf die Abbildung 4 zurückzukommen: Der linke Pfad besitzt somit eine Distanz von vier, während der rechte eine Distanz von drei besitzt. Des Weiteren lässt sich eine *kürzeste Distanz* zwischen zwei Knoten s und t definieren, die auf dem kürzesten Pfad beider Knoten basiert. In solch einem Pfad ist die Distanz, d.h. die Anzahl der Kanten um von s zu t zu gelangen, minimal. Kürzeste Distanzen, gegeben durch den kürzesten Pfad zwischen zwei Knoten, können zum Beispiel mit einer Breitensuche, auf die auch der Algorithmus von Dijkstra aufbaut, berechnet werden. Ausgehend von einem gegebenen Startknoten s berechnet dieser die kürzesten Pfade und somit auch die Distanzen zu allen anderen Knoten, die von s aus erreichbar sind. Er zählt zu der Kategorie der Greedy-Algorithmen, die schrittweise arbeiten und in jedem Schritt denjenigen nächsten Schritt wählen, der zum jeweils aktuellen Zeitpunkt die größte Verbesserung oder das beste Ergebnis erzielt. [8] [5] [6]

Häufige Anwendungsbereiche für Graphen liegen in der Modellierung und Visualisierung von Netzwerken oder Abläufen. Von simplen Darstellungen z.B. von Beziehungen zwischen Menschen oder Verbindungen in sozialen Netzwerken bis hin zu komplexen Schaltvorgängen oder Visualisierungen in der Netzwerktechnik bieten Graphen die Möglichkeit, die Daten in eine abstraktere Struktur zu bringen, um damit besser arbeiten zu können.

2.1.2 Begriffsdefinitionen

Ein *Subgraph* oder *Teilgraph* T eines Graphen G beinhaltet eine Teilmenge der im ursprünglichen Graphen G enthaltenen Knoten und Kanten. Diese Definition resultiert aus der Gegebenheit, dass es häufig notwendig ist, nicht den Graphen als Ganzes zu betrachten, sondern nur den zur Beschreibung der Vorgänge relevanten Teilmenge. Ob und wann nur eine Komponente betrachtet wird,

hängt von der jeweiligen Verwendung ab. In Abbildung 5 kann man die blau hervorgehobenen Knoten sowie die grünen gerichteten Kanten als Teilgraph des gesamten Graphen betrachten. [7]

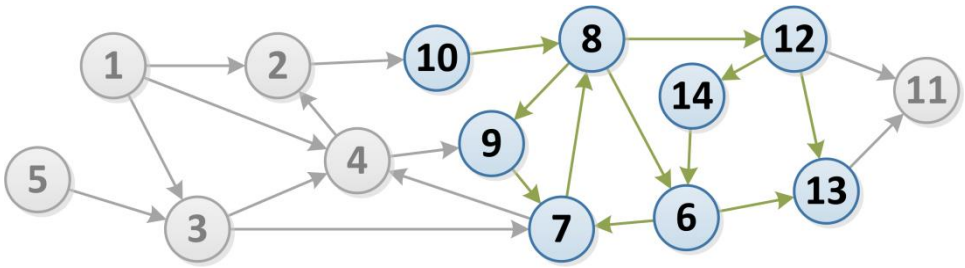

Abbildung 5: Teilgraph

Ein Weg, bei dem Start- und Endknoten ident sind, wird als *Zyklus* bezeichnet. Als *Kreis* wird ein Weg bezeichnet, bei dem der Startknoten s und der Endknoten t wie bei einem Zyklus ident sind und zusätzlich der Weg zwischen s und t einen Pfad bildet. Im Graph in Abbildung 6 (links) bildet der Pfad über die Knoten 1, 2, 4, 7, 3 und 1 einen Kreis. Existiert, wie in Abbildung 6 (rechts), kein Kreis im Graphen, besitzt dieser die Eigenschaft *kreisfrei*. [5]

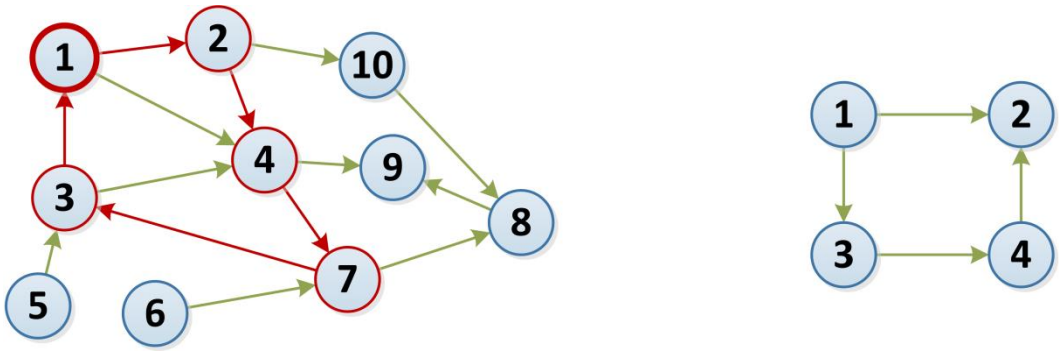

Abbildung 6: Kreis bzw. Zyklus (links) und kreisfreier Graph (rechts)

Die *Nachbarschaft* eines Knotens s bezeichnet die Menge all jener Knoten v_1 bis v_n, die von s aus über ausgehende Kanten in einem Schritt erreichbar sind. In dem Beispiel in Abbildung 7 (links) wären vom Knoten 4 die Nachbarschaftsknoten 2 und 9. Eine weitere Eigenschaft lässt sich aus der Nachbarschaft ableiten. Als *Vorgängerknoten* von einem Knoten s wird ein Knoten t bezeichnet, wenn eine Kante von s nach t existiert. *Nachfolgerknoten* ist ein Knoten t, wenn eine Kante vom betrachteten Knoten s zu t vorhanden ist.

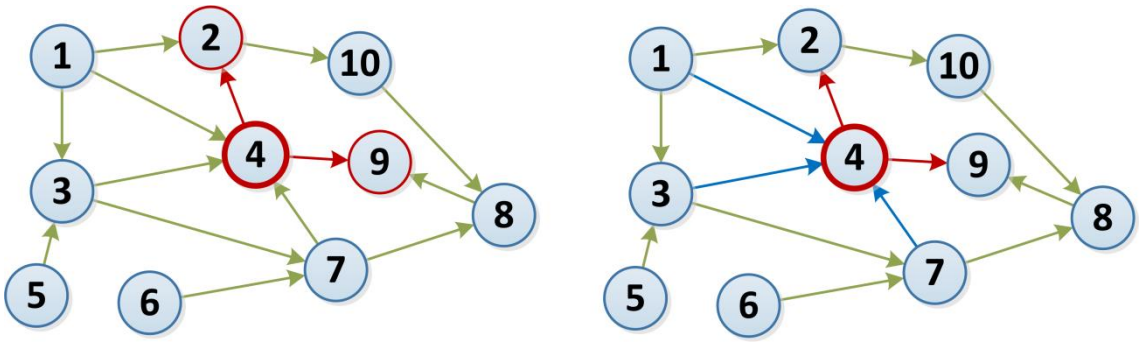

Abbildung 7: Nachbarschaft eines Knotens (links) und Knoten-Grade (rechts)

Der *Grad* eines Knotens s ist eine Eigenschaft, die die Anzahl der Kanten, die s beinhalten, beschreibt. In einem ungerichteten Graphen gibt der Grad die Anzahl an Kanten an, die einen betrachteten Knoten mit anderen Knoten verbinden. In einem gerichteten Graphen wird zusätzlich zwischen den eingehenden Kanten und den ausgehenden Kanten unterschieden. Die Anzahl der Kanten, die von anderen Knoten auf den betrachteten Knoten zeigen, fallen unter die Bezeichnung Eingangsgrad. Betrachtet man in Abbildung 7 (rechts) den Knoten 4, so sind die blau gefärbten Kanten von den Knoten 1, 3 und 7 die eingehenden Kanten und der Eingangsgrad beträgt drei. Die Anzahl der Kanten, die vom betrachteten Knoten zu anderen Knoten zeigen, wird unter dem Ausgangsgrad zusammengefasst. Jener beträgt in Abbildung 7 (rechts) zwei, da die ausgehenden Kanten von Knoten 4 mit den Knoten 2 und 9 verbunden sind. [7]

Ein *Wurzelknoten* ist jener Knoten in einem gerichteten Graphen, der von keinem anderen Knoten im Graphen erreichbar ist. Von der Wurzel sind jedoch alle anderen Knoten im Graphen erreichbar. Er besitzt somit einen Eingangsgrad von 0, mit anderen Worten keinen Vorgänger. Als *Blatt* wird ein Knoten vom Grad 1 bezeichnet, d.h. in einem gerichteten Graphen besitzt jener Knoten einen Eingangsgrad von 1 und einen Ausgangsgrad von 0. In Abbildung 8 (links) ist die Wurzel rot dargestellt und alle Blätter grün. [8]

Als *Baum* wird eine spezielle Form von zusammenhängenden Graphen bezeichnet, bei der die Knoten und Kanten hierarchisch angeordnet sind. Eine weitere Unterscheidung kann bezüglich der Kantentypen getroffen werden. Ein gerichteter Baum, wie in Abbildung 8 (links) ersichtlich, besitzt gerichtete Kan-

ten, einen Wurzelknoten mit nur ausgehenden Kanten, Blätter und ist kreisfrei. Ein ungerichteter Baum besteht aus ungerichteten Kanten, besitzt innere Knoten und Blätter und ist kreisfrei. Aus einem Baum kann keine Kante entfernt werden, ohne dass der Graph in zwei Komponenten zerfällt. [8]

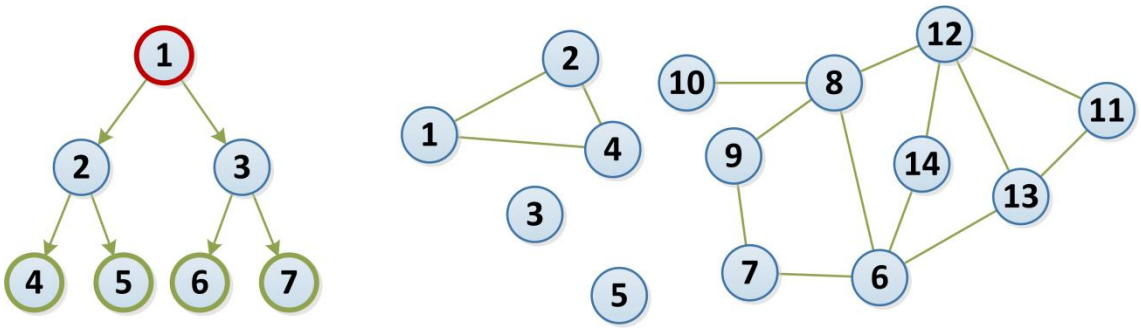

Abbildung 8: Baum-Struktur (links) und Graph mit vier Komponenten (rechts)

Die *Konnektivität* von Graphen beschreibt den Zusammenhang der Knoten in einem Graphen G. Sind je zwei beliebig gewählte Knoten in G durch einen Weg erreichbar, so heißt dieser Graph zusammenhängend. In ungerichteten Graphen trifft diese Eigenschaft zu, wenn für alle Knotenpaare ein ungerichteter Weg existiert. In Abbildung 8 (rechts) besteht der Graph aus vier Komponenten. Ein Weg von Knoten 1 zu Knoten 4 ist möglich, von Knoten 1 zu Knoten 6 hingegen nicht, da die Knoten in zwei verschiedenen Komponenten liegen. Bei gerichteten Graphen kann man eine weitere Einschränkung treffen. Von einem starken Zusammenhang spricht man, wenn ein gerichteter Weg zwischen je zwei gewählten Knoten existiert. Die stark zusammenhängende Komponente ist in Abbildung 9 (rechts) blau hervorgehoben. Repräsentiert man alle gerichteten Kanten in einem Graphen G als ungerichtete Kanten und ist dieser zugehörige ungerichtete Graph G_U zusammenhängend, spricht man von einem schwachen Zusammenhang von G. Abbildung 9 (links) zeigt zwei Komponenten, die jeweils schwach zusammenhängend sind. [5]

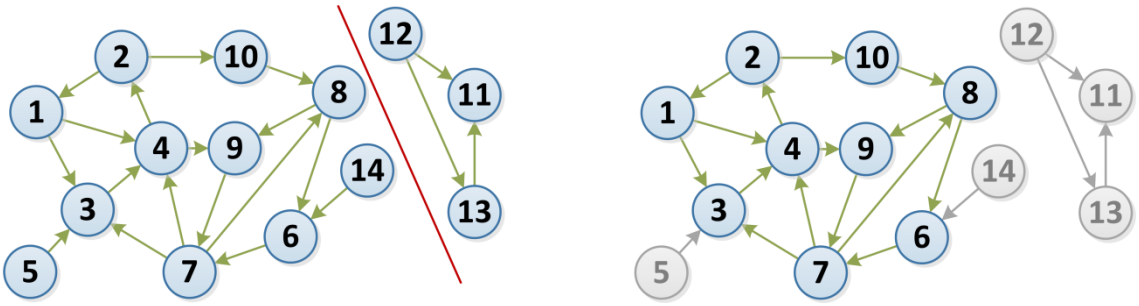

Abbildung 9: Schwacher (links) und starker Zusammenhang (rechts)

Die in diesem Abschnitt erwähnten Grundlagen zur Graphentheorie sind nicht vollständig und beinhalten nur jene Definitionen, die für diese Arbeit relevant sind. Für Graphen lassen sich zahlreiche weitere Eigenschaften definieren und mathematische Berechnungen ausführen. Dazu zählt beispielsweise auch der hierarchische Wert eines Knotens, worauf in Abschnitt 3.4.2 im Kapitel „Related Work“ näher eingegangen wird.

2.2 Navigation

Von Navigation spricht man, wenn man sich an einem Punkt A befindet und sich zu Punkt B bewegen möchte. Der Ablauf nach dem Start bei A bis zum Erreichen von B unterliegt verschiedenen Einflüssen, wie z.B. generelle Einschränkungen oder Störeinflüsse, die sich auf das Bewegungsverhalten auswirken können.

Zur besseren Veranschaulichung kann man das Netz des öffentlichen Verkehrs in einer größeren Stadt betrachten. In diesem Netz befinden sich zahlreiche Haltestellen, die als Einstiegspunkte, Knotenpunkte oder Zielpunkte fungieren. Vorgegebene Linien verbinden diese einzelnen Punkte und geben somit vor, welche Schritte möglich und erlaubt sind. Um nun von einer bestimmten Haltestelle A zur Zielhaltestelle B zu gelangen, gibt es häufig nicht nur eine Möglichkeit. Je komplexer das Netz aus Punkten und Verbindungen ist, desto umfangreicher sind auch die Navigationsmöglichkeiten. Kommt man in einem Verkehrsnetz z.B. zu einem Knotenpunkt, so ist es die Entscheidung des Fahrgastes, welche Linie dieser benutzten möchte. Es gibt effiziente Routen, die schnell und direkt ans Ziel führen, ineffiziente Routen, die Umwege benutzen

und auch ans Ziel führen oder Routen, die nicht erfolgreich sind und das Ziel daher nicht erreichen.

Netzpläne, wie sie in allen Städten mit öffentlichem Verkehr verfügbar sind, bringen einen guten Einblick in das Netzwerk. Diese Darstellungsform entspricht im Allgemeinen der eines Graphen. Haltestellen werden als Knoten betrachtet, Verbindungen zwischen Haltestellen als Kanten. Das Verhalten eines Fahrgastes kann als Pfad, genauer als Navigationspfad, betrachtet werden. Wie schon im Kapitel über Graphentheorie beschrieben, besitzt ein Pfad einen Startpunkt und einen Endpunkt, zwischen diesen beiden Knoten können beliebige andere Knoten durchlaufen werden.

2.2.1 Informationsnetzwerke

Nachdem sich Netzwerkstrukturen anhand von Graphen besonders gut darstellen lassen, kann auch die Navigation eines Benutzers durch ein Netzwerk dadurch beschrieben werden. Das World Wide Web lässt sich durch einen gerichteten Graphen repräsentieren, indem man alle Webseiten als Knoten und alle Links zwischen Webseiten als gerichtete Kanten darstellt. Eine Navigation von einer Startseite A über Seite B und C zu einer Zielseite D könnte durch einen Pfad, der die Kanten A → B, B → C und C → D beinhaltet, beschrieben werden. Dieser Pfad beinhaltet somit 4 Knoten und 3 Kanten. Die zugehörige Distanz beträgt 3, da der Benutzer 3 Links verfolgt hat, um von A zu D zu gelangen. Somit entspricht die Navigation von Seite A zu Seite D einem Navigationspfad mit dem Startknoten A und dem Zielknoten D.

Generell lässt sich über die Navigation von Benutzern in Netzwerken aussagen, dass diese verschiedene Ansätze verfolgen, um zum gewünschten Ergebnis zu gelangen. Abhängig von ihrem Hintergrundwissen zu Themenbereichen bevorzugen z.B. bestimmte Benutzergruppen eine Navigation über geographische, andere hingegen über themenspezifische Aspekte.

2.2.2 Hintergrundwissen

Damit ein Benutzer nicht ziellos durch ein Netzwerk irrt, benötigt dieser bestimmtes Wissen über die Struktur sowie Hintergrundinformationen. Ein Fahrgast kann eine zufällige Linie benutzen und hoffen, dass sie ihn an das gewünschte Ziel bringt, um an dieser Stelle das Beispiel des öffentlichen Verkehrs in Städten nochmals aufzugreifen. Diese Variante ist weder effizient noch entspricht sie der Realität. Viel mehr Sinn ergibt dieses Beispiel schon, wenn sich der Fahrgast Informationen einholt, auf welchem Wege sei dahingestellt. Es findet somit eine Kombination aus bereits vorhandenem und neuem Wissen statt. Durch diese Informationen wählt der Mensch seine „Route“ durch ein Netzwerk, die als menschlicher Navigationspfad dargestellt werden kann. Zahlreiche Einflüsse kennzeichnen somit diesen Pfad, der nicht willkürlich, sondern auf Basis von erlerntem Wissen entstanden ist.

Ein Mensch, der in einem Netzwerk navigiert, hat immer das Ziel vor Augen. In jedem Zeitpunkt möchte er – basierend auf dem Hintergrundwissen – jenen nächsten Schritt wählen, der den größtmöglichen Erfolg verspricht. Betrachtet man alle menschlichen Abläufe, die zum Ziel führen, sind diese äußerst komplex. Aus diesem Grund ist es auch sehr schwierig, das menschliche Wissen zu bestimmten Themenbereichen abzubilden oder zu modellieren.

Ein Ansatz, der häufig verfolgt wird, ist das Abbilden durch hierarchische Strukturen, u.a. in [9]. Solche Strukturen besitzen einen speziellen Aufbau, der alle Informationen in Relation stellt. Die Intuition des Menschen, Informationen zu ordnen und eine gewisse Anzahl an Objekten immer mit einem Überbegriff zu betiteln, kommt jener einer Hierarchie nahe.

Wie schon in den Grundlagen zur Graphentheorie erwähnt, entspricht diese hierarchische Gliederung einer speziellen Form eines Graphen, dem Baum. Durch diese Struktur lassen sich beliebig große Netzwerke in eine Baum-Struktur, beginnend von einer Wurzel bis hin zu den einzelnen Blättern, bringen. Weiteres zum Aufbau von Hierarchien als Hintergrundwissen befindet sich in Abschnitt 3.4.1 sowie die praktische Umsetzung in Abschnitt 6.6.

2.3 Wikipedia

Wikipedia[3], die freie Enzyklopädie, ist das größte und bekannteste webbasierende Nachschlagewerk der Welt. Das Konzept hinter Wikipedia ist einfach und zugleich einzigartig: Informationen und Artikel können von allen Besuchern betrachtet und auch bearbeitet werden. Durch die kollaborative Arbeit wächst die Menge an Informationen, die das Werk beinhaltet, stetig an und ist laufenden Änderungen unterworfen. [10]

Das Projekt wurde Anfang 2001 von Jimmy Wales und Larry Sanger gegründet und hat über die Jahre enorm an Popularität gewonnen. [11] Mit Stand vom November 2012 ist die größte Wikipedia die englisch-sprachige mit ca. 4,2 Millionen Artikeln, gefolgt von der deutschen Version mit rund 1,5 Millionen und der französischen Wikipedia mit ca. 1,3 Millionen Artikeln. Die Entwicklung der letzten Jahre zeigt, dass die Wachstumsrate monatlich etwa 1% beträgt. [12] Auf den verschiedensprachigen Wikipedia-Enzyklopädien erfolgt nicht die Übersetzung eines Inhalts von einer Sprache in die andere, sondern jede einzelne Wikipedia entwickelt sich durch die Autorenschaft individuell.

Eine frei zugängliche Wissensdatenbank bringt nicht nur Vorteile mit sich. Die Qualität der Inhalte ist ein häufiger Kritikpunkt. Viele Artikel, vor allem aktuelle Themen oder umstrittene Inhalte, sind von Vandalismus betroffen. Um gegen diese Probleme vorzugehen, können bestimmte Artikel nicht anonym bearbeitet werden oder müssen vor der Freigabe noch von Moderatoren bestätigt werden.

Wikipedia-Artikel können nicht als wissenschaftlich geprüft und daher auch nicht als vertrauenswürdig eingestuft werden. Zur Angabe als Quellen in wissenschaftlichen Arbeiten sind diese folglich nicht geeignet. Die Gründe liegen einerseits wiederum bei fehlenden Quellenangaben bei Inhalten und andererseits ist eine ständige Überprüfung aller Artikel, die durch die große Autorenschaft ständiger Veränderung unterzogen sind, nicht realisierbar.

[3] http://en.wikipedia.org

2.3.1 Technologie

Das System basiert auf einem sogenannten Wiki, einer Technologie, die es erlaubt, jede einzelne Seite durch einen einfachen Klick in den Bearbeitungsmodus umzuschalten, Änderungen vorzunehmen und anschließend abzuspeichern. Für die Wikipedia wurde das freie Softwarepaket MediaWiki[4] entwickelt, das – basierend auf PHP[5] und unter einer GPL[6] angeboten – auch für beliebige andere Zwecke verwendet werden kann. [10]

Der Zugriff auf die Website kann von jedem Gerät mit Verbindung zum Internet erfolgen. Dadurch ist gewährleistet, dass auch immer die aktuelle Version geladen wird, die z.B. auch tagesaktuelle Informationen beinhaltet. Da alle Inhalte frei verfügbar sind, besteht auch die Möglichkeit den gesamten Inhalt der Enzyklopädie herunterzuladen und offline zu verwenden. Für Entwickler und Forscher stehen weiters zahlreiche Daten wie z.B. zum Aufbau oder zur Struktur bereit. Jene Daten wurden auch für diese Arbeit benötigt.

Um zu einem bestimmten Artikel zu gelangen, kann entweder die Suche benutzt werden oder eine Navigation über die auf den einzelnen Seiten verfügbaren Links erfolgen. Da Artikel zu einem bestimmten Themengebiet zugeordnet sind, kann durch eine hierarchische Navigation, beginnend von der Hauptseite, auch zum gewünschten Artikel gefunden werden.

2.3.2 Aufbau und Struktur

Die Gliederung der Wikipedia basiert auf Portalen und Kategorien. Durch diese Einteilungen soll es dem Benutzer einfacher gemacht werden, sich in der komplexen Wissensdatenbank zurechtzufinden. Beginnend von der Startseite hat man entweder die Möglichkeit, die Suche zu benutzen oder sich durch die Hauptportale, die auf der Startseite gelistet sind, zu klicken. In der englischen Wikipedia sind diese Portale „Arts“, „Biography“, „Geography“, „History“, „Mathematics“, „Science“, „Society“ und „Technology“. Außerdem gibt es noch

[4] http://www.mediawiki.org

[5] PHP: Hypertext Preprocessor – Skriptsprache zur Erstellung dynamischer Webanwendungen

[6] GNU General Public License – Lizenz für freie Software

zahlreiche Subportale, über die navigiert werden kann. All diese Portale sollen so aufgebaut sein, dass es für den Anwender möglich ist, von der Startseite über mehrere Schritte zu jedem gesuchten Artikel zu navigieren. Durch die hierarchische Gliederung besteht für den Benutzer die Möglichkeit, den Suchbereich immer weiter einzuschränken um schließlich zum gewünschten Ergebnis zu gelangen. Zum Artikel „Graz“ kommt man beispielsweise durch eine Navigation über das Portal „Geography“, mit einer anschließenden Navigation über „Europe“, zu „Austria“ und „Graz“. [10]

Neben den Portalen stehen auch noch Kategorien zur Verfügung. Jeder Artikel kann einer oder mehreren Kategorien zugeordnet werden, die am Ende des Artikelinhalts angezeigt werden. Jede Kategorie beinhaltet eine Liste aller Artikel, die ihr zugeordnet sind. Auch darüber lässt sich eine schnelle Navigation realisieren.

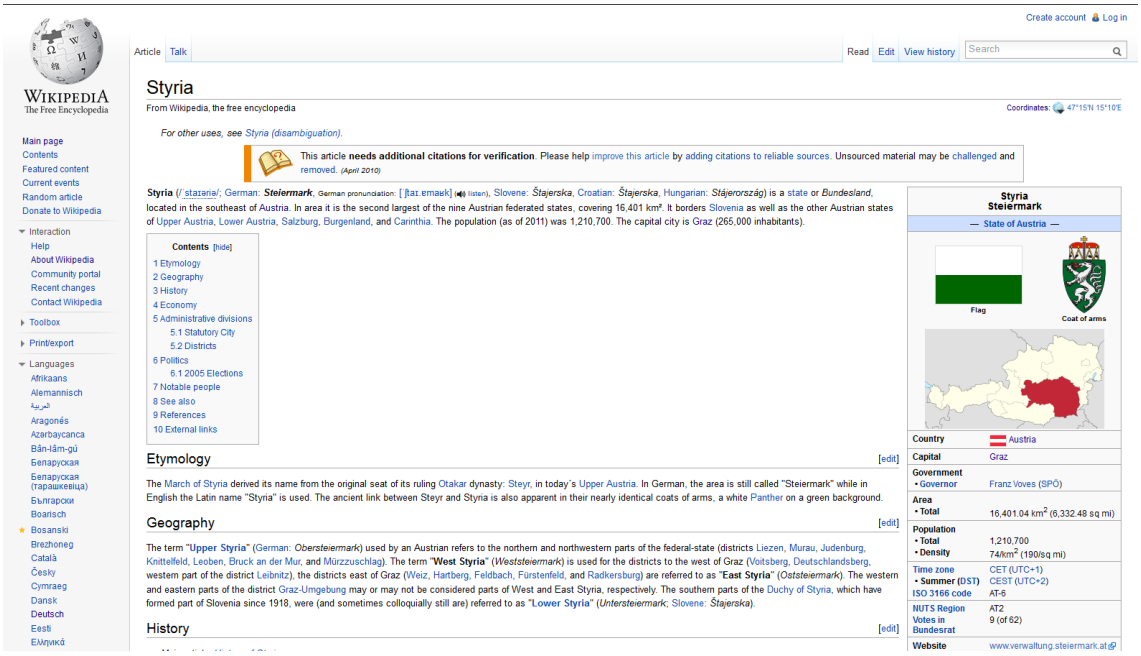

Abbildung 10: Wikipedia - Englischer Artikel zu "Styria"

Alle Wikipedia-Artikel sind in ähnlicher Form aufgebaut. Nach einer kurzen Beschreibung bzw. einer Themenbereich-Einordnung des Artikels folgt bei längeren Artikeln ein Inhaltsverzeichnis, das die anschließende Gliederung beinhaltet. Bei zahlreichen Artikeln, z.B. bei Städten, befinden sich auf der rechten Seite Infoboxen, die aufgrund der verwendeten Vorlagen immer diesel-

be Struktur aufweisen. Darin enthalten sind oft Daten und Fakten zum Artikel. Am Artikelende befinden sich häufig Quellenverweise oder externe Links und es werden – wie schon im Absatz zuvor beschrieben – die Kategorien gelistet. Abbildung 10 zeigt ein Beispiel für eine Artikelseite.

Zur Struktur in der Wikipedia kann man noch einige interessante Fakten erwähnen. Alle Seiten, die sich in dem Enzyklopädie-Netzwerk befinden, sind einem bestimmten Namensraum[7] („namespace“) zugeordnet. Mit der Ausnahme des Namensraumes „Main“ wird vor Seiten ein zugehöriger Prefix gesetzt, um die Seite eindeutig zu identifizieren, z.B. „Template:Flag“ für die Vorlage einer Flagge. [13] [10]

Nachfolgend eine Auflistung der verwendeten Namensräume:

- „Main“ beinhaltet alle Artikel, Begriffsklärungen und auch Weiterleitungen zu anderen Artikeln.
- „User“ stellt Seiten für Autoren, die frei gestaltet werden können, zur Verfügung.
- „Wikipedia“ enthält alle Informationen zum Projekt an sich. Dazu zählen Einführungen, Richtlinien, Abläufe etc.
- „File“ beinhaltet alle Beschreibungsseiten für Media-Dateien wie Bilder, Audio- oder Video-Dateien.
- „MediaWiki“ bietet Spezialseiten zur Software MediaWiki an.
- „Template“ beinhaltet Vorlagen, die auf diversen Seiten häufig verwendet werden. Dies können Informationsboxen, Navigationsboxen oder andere Standard-Texte sein.
- „Help“ bietet Hilfeseiten zur Verwendung der Wikipedia sowohl für Anwender bzw. Besucher als auch für Autoren.

[7] http://en.wikipedia.org/wiki/Wikipedia:Namespace

- „Category“ enthält Listen mit allen Seiten, die einer bestimmten Kategorie zugeordnet sind.

- „Portal“ bietet Übersichtsseiten zu einem bestimmten Themengebiet.

Es sind noch einige weitere Namensräume vorhanden, die jedoch nicht sehr häufig Anwendung finden und aus diesem Grund nicht in die Übersichtsliste aufgenommen wurden. Des Weiteren sollte auch noch erwähnt werden, dass es zu jeder Seite, egal in welchem Namensraum sie sich befindet, eine zugehörige Diskussionsseite vorhanden ist. Auf diesen Seiten ist meist Platz für den Meinungsaustausch unter den Autoren, wie z.B. Verbesserungsvorschläge, Kritiken etc.

Eine Vielzahl an Regeln gibt Wikipedia-Autoren vor, wie Artikel aufzubauen sind und worauf bei der Bearbeitung zu achten ist. Möchte man beispielsweise in einem Artikel einen Link zu „Austria“ setzen, so erfolgt dies durch die Verwendung folgender Syntax: [[Austria]]. Durch doppelte Apostrophen kann ein Text kursiv dargestellt werden: ''kursiver Text''. [10] Da vorgegebene Regeln, speziell zum Aufbau von Artikeln, nicht von allen Autoren berücksichtigt werden, ist an vielen Stellen Vorsicht geboten. Als Beispiele können die geringfügige unterschiedliche Verwendung der Wiki-Syntax oder die nicht konforme Setzung von Links zu anderen Artikeln bzw. zu externen Seiten angeführt werden.

Die Struktur von Wikipedia mit Artikeln und Links zwischen Artikeln ist hervorragend geeignet, dieses Wissen mittels eines Graphen zu repräsentieren. Darauf wird später in dieser Arbeit in Abschnitt 6 näher Bezug genommen.

2.4 Wikigame

Aufgrund der starken Verbreitung der Wikipedia, wurden über die Jahre hinweg auch Spiele, die auf der Online-Wissensdatenbank basieren, entwickelt. Die folgenden Projekte zählen zu den bekanntesten Spielen:

- WikiGame – Das Wiki-Spiel[8]
- The Wiki Game – Wikipedia Game[9]
- Wikispeedia[10]

„WikiGame – Das Wiki-Spiel“ ist ein Spiel, dessen Ziel es ist, Multiple-Choice-Fragen zu beantworten. Fragen können von Spielern selbst hinzugefügt werden und basieren auf dem Wissen der Wikipedia.

Bei „The Wiki Game – Wikipedia Game“ ist es die Aufgabe des Spielers, sich anhand von Links durch die Wikipedia-Seiten zu navigieren, wobei Start- und Zielseite vorgegeben sind. Als Mission wird die grundlegende Aufgabe bezeichnet, eine Navigation von einer Seite (Startseite) zu der anderen Seite (Zielseite) durchzuführen. Da dieses Spiel und das im Absatz zuvor erwähnte nahezu dieselbe Bezeichnung besitzen, sei darauf hingewiesen, dass Wikigame im weiteren Verlauf dieser Arbeit immer das Navigationsspiel „The Wiki Game“ bezeichnet.

Das Projekt Wikispeedia ist aus einer Forschungsarbeit an der McGill University in Montreal (Canada) entstanden, um generierte Benutzerdaten automatisch anzulernen. Die Funktionsweise dieses Spiels, eine Navigation von Seite A zu Seite B auszuführen, ist jener des Wikigames nahezu ident.

Wie auch Wikipedia, kann das Wikigame mit jedem internet-fähigen Gerät per Webbrowser aufgerufen werden. Die eingesetzten Technologien sind Python[11], Django[12], Redis[13], XMPP[14] und Strophe[15].

[8] http://www.wikigame.org/
[9] http://www.thewikigame.com/
[10] http://www.wikispeedia.net/
[11] eine universelle höhere Programmiersprache
[12] ein quelloffenes Web-Framework basierend auf Python
[13] Remote Dictionary Server, eine In-Memory-Datenbank
[14] Extensible Messaging and Presence Protocol
[15] Bibliothek für XMPP

Gespielt werden kann entweder als Gastbenutzer oder als registrierter Benutzer. Gastbenutzer bekommen eine eindeutige Identifikation als Benutzernamen zugeteilt. Bei einem erfolgreichen Spiel werden für registrierte Spieler automatisch Punkte auf das Konto gutgeschrieben, sodass auch eine Eintragung in verschiedene Ranglisten erfolgt. Die Anzahl der Punkte, die zwischen 50 und 1000 liegt, richtet sich nach der Platzierung bei einem Spiel.

2.4.1 Allgemeiner Spielablauf

Nach dem Start erhält der Spieler seine Mission, die den Start-Artikel und den gesuchten Ziel-Artikel beinhaltet. Beide wurden aus dem Wikipedia-Netzwerk zufällig gewählt und sind für eine bestimmte Zeit, ca. 2 Minuten, aktiv. Während dieser Zeit ist es die Aufgabe aller Spieler, die sich gerade im Spiel befinden, dieselbe Mission erfolgreich auszuführen. Mit einem Klick auf den Start-Artikel startet das Spiel mit dem Öffnen der zugehörigen Wikipedia-Artikelseite. Von nun an kann auf alle auf dieser Seite verfügbaren Links geklickt werden, um zur nächsten Seite – und somit immer näher an den Zielknoten – zu gelangen. Abbildung 11 zeigt ein aktives Spiel, dessen Ziel es ist, zum Artikel „Mick Jagger“ zu finden.

Neben dem Hauptbereich des Spieles, siehe Abbildung 11 links, befindet sich eine Liste, die jeweils die Anzahl der getätigten Klicks von aktiven Mitstreitern zeigt. Auch erfolgreiche Missionen, in denen Spieler den gesuchten Ziel-Artikel schon gefunden haben, werden angezeigt. Das Spiel endet entweder mit Erfolg, wenn die Zielseite innerhalb des Zeitfensters geöffnet wird, oder erfolglos, wenn entweder die Zeit abgelaufen ist oder der Spieler die Seite verlässt.

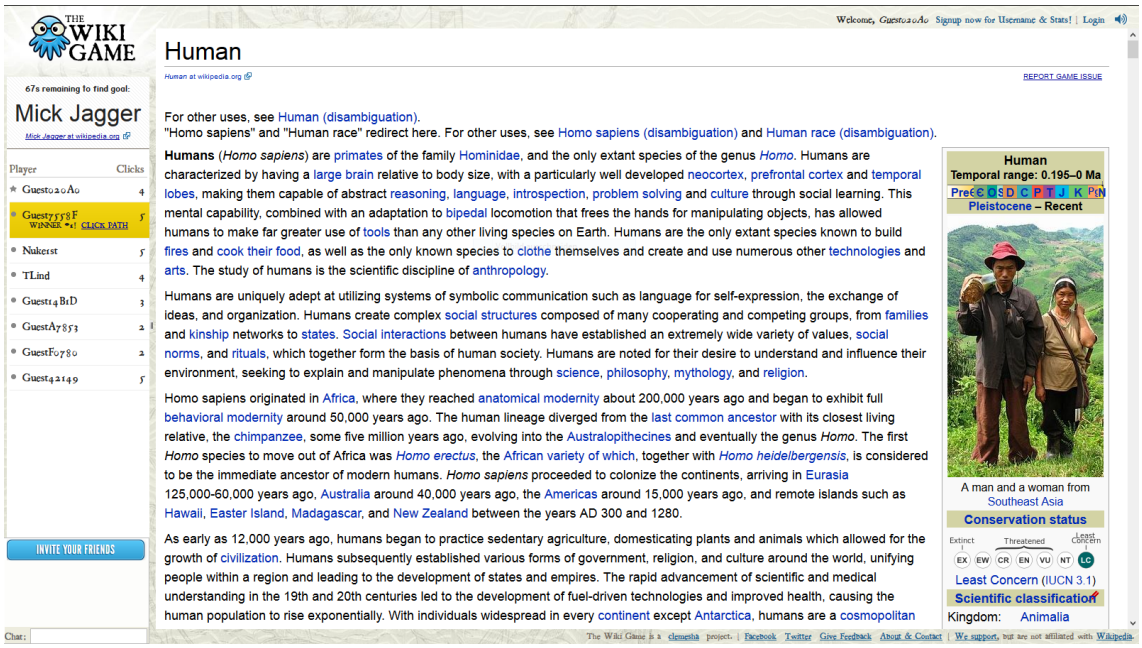

Abbildung 11: Wikigame - Spiel

Abbildung 12: Wikigame - Startseite mit Auswahl der Spieltypen

2.4.2 Spieltypen

Wikigame bietet fünf verschiedene Spieltypen an, wobei das Grundprinzip der Navigation immer dasselbe bleibt und sich die Spiele in kleinen Abänderungen voneinander unterscheiden. Die Auswahl ist in Abbildung 12 ersichtlich.

Der erste Typ wird „Speed Race“ genannt. In diesem Spieltyp muss der Spieler versuchen, schnellstmöglich die genannte Zielseite zu erreichen. Die Anzahl der Klicks zum Spielerfolg spielt keine Rolle.

„Least Clicks“ wird der zweite Typ genannt, bei dem – wie schon der Name verrät – mit möglichst wenigen Klicks die Zielseite erreicht werden soll. Auch in diesem Spiel ist ein Zeitfenster, innerhalb der das Ziel gefunden werden soll, vorgegeben. Sieger ist dennoch derjenige Spieler, der vom Start- zum Ziel-Artikel die wenigsten Klicks benötigt hat.

Ein weiterer Typ nennt sich „Six Degrees Of Wikipedia“. Ziel des Spieles ist es, vom Start- zum Ziel-Artikel in genau sechs Klicks zu gelangen. Dieser Modus stellt den Benutzer vor eine Herausforderung, da der Benutzer sein Klickverhalten dementsprechend anpassen muss. Warum die Länge sechs eines Pfades interessant ist, wird im Kapitel 3.3, in dem das Kleine-Welt-Phänomen betrachtet wird, erklärt.

„Five Clicks To Jesus“ ist der Name des vierten Spieltyps. Dabei wird nur die Startseite per Zufall gewählt, die Zielseite ist in jeder Mission der Wikipedia-Artikel „Jesus“. Diese Seite muss wie im zuvor beschriebenen Typ in einer bestimmten Anzahl an Klicks, nämlich genau in fünf, erreicht werden.

Die letzte Version des Spieles heißt „No United States“ und basiert darauf, dass im Spiel nicht über den Artikel „United States“ navigiert werden darf. Dieser Wikipedia-Artikel bietet eine Besonderheit, denn er ist in sehr vielen Artikeln verlinkt, d.h. er hat eine große Anzahl an eingehenden Links. Darüber hinaus besitzt der Artikel auch zahlreiche ausgehende Links. Da viele Spieler vor allem über die geographischen Informationen navigieren, stellt diese Einschränkung eine noch größere Bedeutung dar. Sollte in einer Mission „United States“ ge-

klickt werden, kann zwar noch weiter navigiert werden, aber das Spiel ist ab diesem Zeitpunkt schon verloren.

2.4.3 Einschränkungen und Abänderungen

Wikigame basiert nicht auf den vollständigen und aktuellen Inhalten der Wikipedia. Informationen dazu, wie oft ein Datenabgleich mit Wikipedia erfolgt, sind nicht bekannt. Bei vielen Artikeln fehlen bestimmte Links oder Infoboxen, die oft relevante Links zu weiteren Seiten beinhalten würden. Durch Veränderungen bei der Darstellung der Artikelseiten, entspricht die Formatierung bei vielen Artikeln nicht jenen der Wikipedia. Ob dies eine bewusste Veränderung ist oder ob es sich um Fehler bei der automatisierten Darstellung der originalen Wikipedia-Artikel handelt, konnte nicht festgestellt werden. Um das Spiel nicht zu umgehen, wird des Weiteren keine Suche angeboten.

Durch detailliertere Recherchen wurde festgestellt, dass es auch möglich ist, das Spiel teilweise recht einfach zu manipulieren und sich somit über „Links“ weiter zu bewegen, die im Artikel nicht existieren. Diese erwähnten Umstände können die Klickpfade der Benutzer verfälschen. Nähere Erklärungen dazu finden sich in Kapitel 6.

3. Related Work

In diesem Kapitel wird auf bereits veröffentlichte Arbeiten und Forschungsergebnisse näher eingegangen, die in einem Verhältnis zum Themengebiet dieser Arbeit stehen.

3.1 Verhaltensanalyse von Benutzern im Internet

Musste vor Jahrzehnten in zahlreichen Büchern oder Dokumenten nachgeschlagen werden, um an bestimmte Informationen oder an Wissen zu gelangen, so sind diese heutzutage in der stark vernetzten Welt oft nur wenige Klicks entfernt, außerdem jederzeit und im Internet von überall aus abrufbar. Zahlreiche Informationssysteme und Wissensdatenbanken, wie z.B. Wikipedia, tragen zum schnellen Auffinden von Wissen bei. Auf der Suche nach Informationen verfolgen Menschen verschiedene Strategien, um auch an das Gesuchte zu gelangen. Suchmaschinen oder Links auf diversen Websites zählen dazu, ebenso wie die Benutzung von Navigationsschaltflächen im Webbrowser. Das Verhalten von Benutzern, das auf verschiedenen Strategien basiert, wird u.a. in [14], [15] und [16] analysiert.

3.1.1 Allgemeines

Im Allgemeinen kann, wie schon zuvor erwähnt, zwischen einer Suche und einer Navigation über Hyperlinks unterschieden werden. Ein typischer Suchvorgang eines Benutzers auf der Suche nach Informationen startet bei einer Suchmaschine im World Wide Web (WWW). Je nach Art, können hier Suchmaschinen, die den Großteil an zugänglichen Informationen des WWW durchsuchen, wie z.B. Google[16] oder spezielle Suchmaschinen, z.B. für Suchen nach Übersetzungen oder nach Personen, zum Einsatz kommen. Um Suchfunktionen nutzen zu können, muss eine Anfrage formuliert werden. Diese enthält meistens die Stichworte des Suchzieles. Der Erfolg eines Suchvorgangs hängt somit auch entscheidend von der Formulierung der Anfrage ab. Führt eine Suche nicht zum erwünschten Ergebnis, so kann eine Umformulierung, Ergän-

[16] http://www.google.com

zung oder Verfeinerung der Stichworte erfolgversprechender sein. Auch im Bereich Usability entwickeln sich Suchmaschinen weiter, um Anfragen benutzerfreundlicher zu gestalten und trotzdem automatisiert interpretierbar zu bleiben.

In [16] wurde eine Studie mit zweitausend freiwilligen Teilnehmern durchgeführt, um das Suchverhalten im Web zu analysieren. Aufgrund des längeren Zeitraumes von fünf Monaten konnten die Ergebnisse an Aussagekraft gewinnen. Die Menge an aufgezeichneten Interaktionen von Benutzern wurde auf Gemeinsamkeiten untersucht, um so das Verhalten kategorisieren zu können. Darin wurde aufgezeigt, dass Benutzer eher selten mit ihrer ersten Suchanfrage erfolgreich sind und dadurch die weiteren Links durchsuchen, zurücknavigieren oder bei Misserfolg ihre Anfrage neu formulieren.

Die Ergebnisse in [17] haben gezeigt, dass eine gute Suchmaschine für Benutzer nicht ausreichend ist, um an ihre Informationen zu gelangen. Bei der Studie haben 39% der Teilnehmer bei Suchen eine Suchanfrage mit Schlüsselwörtern formuliert. Auch wenn das genaue Ziel der Suche bekannt ist, werden häufig kurze, lokale Links in Kombination mit dem Benutzer-Hintergrundwissen benutzt, um schließlich an die Informationen zu gelangen. Dieses Verhalten wird als sogenannte Orientierung bezeichnet, in der der Benutzer zuerst z.B. durch einen Suchanfrage auf eine Seite gelangt, und von dort aus die weitere Navigation erfolgt. In [18] wird dargelegt, dass Benutzer in 65% der Fälle eine Navigation über Links bevorzugen.

Somit spielt auch der zweite Bereich, die Navigation mittels Links, neben der Suche eine wichtige Rolle bei der Analyse von Benutzerverhalten. Nahezu alle Webseiten beinhalten eine Vielzahl an Links, wobei es die Entscheidung des Benutzers ist, welcher am relevantesten ist. Bezug drauf wird in [14] genommen: Menschen sind sehr gut darin, bei Entscheidungen die beste oder eine möglichst gute zu wählen. Das menschliche Hintergrundwissen, das sehr ausgeprägt, breit gefächert und in bestimmten Bereichen vertieft ist, ist hierbei sehr hilfreich. In [14] wird nach einem Algorithmus gesucht, dessen Navigationsverhalten Menschen übertrifft. In diesem Projekt hingegen möchte man dem

menschlichen Verhalten möglichst nahekommen und den Algorithmus dahingehend verfeinern und adaptieren.

3.1.2 Wikipedia

Voß hat in [19] die grundlegenden Strukturen der Wikipedia untersucht. Neben den Artikeln, den Autoren und den Bearbeitungen der Seiten wurde darin vor allem auch die Linkstruktur analysiert. Schon zur Veröffentlichung des Papers im Jahr 2005 zeichnete sich die rasante Entwicklung der freien Enzyklopädie ab. Die darin enthaltenen Analysen zeigen die starke Vernetzung der Artikelseiten.

In [20] möchte Helic mehr Einblick in das Navigationsverhalten von Benutzern zu einer bestimmten Zielseite auf Wikipedia erhalten. Das Verhalten ist von vielen Faktoren abhängig, wie dem menschlichen Hintergrundwissen, dem aktuellen Informationsbedürfnis und auch den Informations-, Navigations- und Netzwerkstrukturen. Dabei wurde ein Wikipedia-Netzwerk untersucht, das ca. 10 Millionen Knoten und rund 250 Millionen Links beinhaltet. Der effektive Durchmesser, d.h. die kürzeste Pfadlänge unter der 90% aller Knoten des Netzwerks erreicht werden, beträgt weniger als sechs. Der größere Teil des Netzwerks, etwa 55%, ist stark zusammenhängend. Dies bedeutet, dass innerhalb dieser Komponente kurze Pfade zwischen zwei zufällig gewählten Knoten existieren. Die schwach zusammenhängende Komponente beinhaltet ca. 99% des gesamten Netzwerks.

3.2 Analyse von Klickpfaden

Wie in [21] aufgezeigt und in [20] näher betrachtet, basiert das Navigationsverhalten von Menschen in komplexen Netzwerken aus zwei Phasen. Benutzer starten bei einer Navigation häufig an einem Knoten, der einen geringen Grad besitzt. Von diesem bewegen sie sich in Phase eins über Links zu übergeordneten Knoten, d.h. welche die einen höheren Grad besitzen. Diese werden auch Hubs genannt, von wo aus die zweite Phase beginnt, nämlich die Navigation zur Zielseite. Die erste wird als Zoom-Out-Phase bezeichnet, die zweite als Zoom-In-Phase. In der Analyse eines Klickpfades ist das Zoom-Out durch

immer größer werdende Grade der bereits besuchten Knoten zu erkennen, das Zoom-In durch immer kleiner werdende Grade bis zum Erreichen des Zielknotens. An einem Hub erfolgt der Wechsel der beiden Phasen. Aus diesem Grund wird darin auch die Wichtigkeit von Hubs im Navigationsverhalten hervorgehoben. Des Weiteren wurde in [20] gezeigt, dass die durchschnittliche Klickpfad-Länge bei erfolgreichen Pfaden aus dem Wikigame (6,27) nur knapp über dem effektiven Durchmesser des Netzwerkes (5,7) liegt.

Eine weitere Untersuchung, die sich den Unterschieden zwischen automatischer und menschlicher Navigation widmet, wurde von West und Leskovec in [14] vorgenommen. Die Forscher möchten analysieren, ob ein breites Hintergrundwissen für eine möglichst gute Navigation benötigt wird. Aus diesem Grund wurden mehrere Ansätze, die auf einfachen Netzwerk-Eigenschaften basieren, getestet. Neben heuristischen Ansätzen wurde auch auf maschinelles Lernen gesetzt. Erstere Ansätze wurden durch die Navigation aufgrund des besten Ausgangsgrades einer Seite, durch TF-IDF-Maße[17] und einer Kombination dieser beiden modelliert. Es wurde gezeigt, dass eine automatische Navigation basierend auf diesen Methoden zum Teil besser abläuft. Im Gegensatz zu den 30.000 Wikipedia-Navigationspfaden der Benutzer, die im Durchschnitt die doppelte Länge des kürzesten Pfades besaßen, betrug die Länge der automatisch generierten Pfade nur das Eineinhalbfache.

3.3 Kleine-Welt-Phänomen

Der Begriff Kleine-Welt-Phänomen wurde durch die Forschungsarbeit des US-amerikanischen Psychologen Stanley Milgram in [22] geprägt. Er führte gemeinsam mit Kollegen an der Harvard University in den 1960er Jahren die ersten Experimente über die Vernetzung unserer Gesellschaft durch. Ziel des ursprünglichen Experiments war es, zwischen zwei zufällig ausgewählten Menschen, die sich nicht kennen und sich an verschiedenen Orten in den Vereinigen Staaten von Amerika befinden, eine möglichst kurze Kette an Bekanntschaften zu finden, die diese beiden Personen miteinander verbindet.

[17] term frequency - inverse document frequency – ein Maß, das eine Aussage zur Relevanz eines Begriffes zu Dokumenten in einer Kollektion aus Dokumenten trifft

Milgram suchte per Zufall Personen aus, die als Startpersonen fungierten. Jene Personen wurden instruiert, einen Brief an eine Zielperson in Sharon, Massachusetts, USA weiterzuleiten. Die einzigen Informationen, die sie zur Zielperson erhielten, waren der Name, die Adresse, das Beschäftigungsumfeld bzw. Beruf und ein paar weitere persönliche Informationen. Die Aufgabe der Startpersonen war es nun, diesen Brief an eine bestimmte frei wählbare Person aus dem Bekanntenkreis weiterzuleiten, keinesfalls die Zielperson – außer sie ist eine engere Bekanntschaft. Die Auswahl sollte immer auf jene Person fallen, die sie als beste Wahl empfanden, um den Brief möglichst nahe ans Ziel weiterzuleiten. Jene Personen, die den Brief erhielten, hatten dieselbe Aufgabe – eine Weiterleitung des Briefes an eine Bekanntschaft – durchzuführen. Diese Kette wurde fortgesetzt, bis der Brief im erfolgreichen Fall die Zielperson erreicht. Ein Beispiel für den Verlauf einer Briefkette ist in Abbildung 13 dargestellt. Rund ein Drittel der abgesendeten Briefe kamen beim Empfänger an. [23] [5]

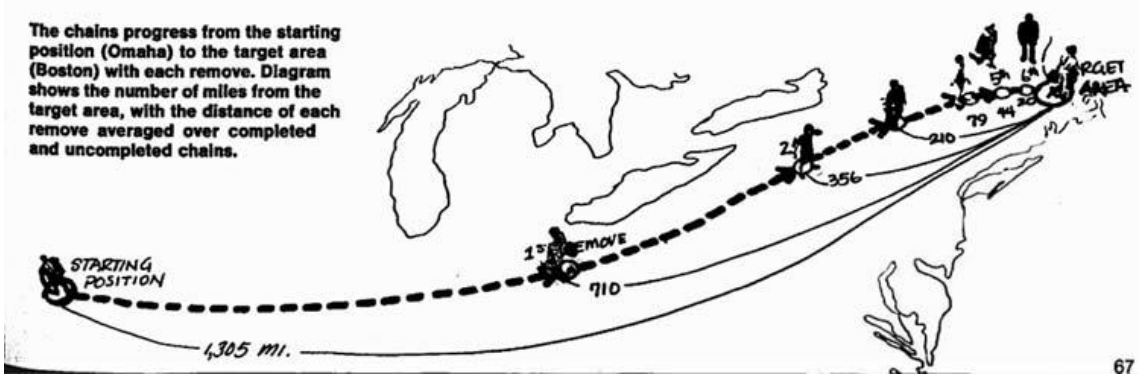

Abbildung 13: Original-Abbildung aus Milgrams Veröffentlichung [22]

3.3.1 Ergebnisse von Milgrams Experiment

Das Ergebnis dieses Experiments erstaunte die Forscher so sehr, dass Milgram von einem Phänomen sprach. Die durchschnittliche Schrittlänge bzw. Länge eines Pfades von der Start- zur Zielperson aller weitergeleiteten Briefe betrug sechs. Im Englischen wird auch oft von „six degrees of separation“, d.h. dass über sechs Verbindungen in einem Netzwerk zwei beliebige Personen miteinander verbunden sind, gesprochen. Dies ließ die Forscher zum Entschluss kommen, dass kurze Pfade zwischen allen oder nahezu allen Personen im globalen Freundschaftsnetzwerk existieren. [5]

Milgrams Experiment konnte zwei wichtige Fakten in Bezug auf große soziale Netzwerke aufzeigen. Einerseits existiert die Tatsache, dass es zwischen zwei beliebigen Personen kurze Pfade gibt. Andererseits kommt auch der Fähigkeit der Menschen, diese kürzesten Pfade effektiv zu finden, ohne dass sie ein Hintergrundwissen über das vollständige Netzwerk besitzen, eine große Bedeutung zu. Ein ausgedehntes soziales Netzwerk beinhaltet genügend Verknüpfungen geographischer und sozialer Art, sodass diese Pfade ermöglicht werden. [5]

3.3.2 Erweiterungen des Experiments

Dodds, Muhamad und Watts haben in ihrer Publikation [24] einige Jahrzehnte nach Milgram ein weiteres Experiment durchgeführt. Dabei wurden die grundlegenden Versuchsbedingungen allerdings abgewandelt, um ein repräsentatives Ergebnis zu erhalten. Statt einer einzigen Zielperson, wurde allen Teilnehmern eine von achtzehn Zielpersonen aus dreizehn verschiedenen Ländern per Zufall zugeteilt. Die Aufgabenstellung war dieselbe: Jeder, der einen Brief erhielt, soll diesen an eine Bekanntschaft weiterleiten. 25 Prozent der Startpersonen, die sich auf einer Webseite dazu registrieren konnten, nahmen auch am Experiment teil und sendeten den Brief ab. Um mehr Hintergründe über das Verhalten der teilnehmenden Personen zu erhalten, musste jeder die Art, den Ursprung und die Stärke der Bekanntschaft sowie den Grund für die Weiterleitung angeben. Die Forscher gelangten zu der Erkenntnis, dass an erster Stelle die geographische Nähe der Bekanntschaft für die Weiterleitung ausschlaggebend war, gefolgt von den naheliegenden Tätigkeitsfeldern der Zielpersonen. Während in den ersten Schritten die geographische Nähe dominierend war, so kam danach immer mehr das Tätigkeitsfeld bei der Auswahl ins Spiel.

Die Länge der erfolgreichen Ketten lag zwischen eins und zehn, mit einem Durchschnitt von rund vier. Da hier nur die erfolgreichen Pfade betrachtet wurden und kürzere Ketten häufiger vorkamen, konnte in der Forschungsarbeit die effektive Länge mit sieben Schritten berechnet werden. Eine weitere Erkenntnis aus [24] ist, dass sogenannte Hubs, d.h. Personen, die sehr viele Bekanntschaften bzw. eine große Anzahl an Verbindungen zu Personen haben,

eine untergeordnete Rolle spielen. Dies spiegelt sich in der Auswahl der Personen wider, indem Teilnehmer Briefe im Allgemeinen gesehen auch nicht an Hubs senden.

Wie in [25] erwähnt, spricht man von einem Kleine-Welt-Netzwerk, wenn nahezu jedes zufällig gewählte Paar über eine kleine Anzahl an Knoten bzw. Kanten miteinander verbunden ist. Man kann ein solches Netzwerk auch definieren, indem man das Wachstum der durchschnittlichen Distanz L zwischen zwei Knoten proportional zum Logarithmus der Anzahl der Knoten N beschreibt: $L \sim log(N)$.

Bezüglich der Abbruchrate, in [24] „attrition rate“ genannt, von Ketten wurde festgestellt, dass diese bis zu einer gewissen Länge, im untersuchten Fall etwa sieben, nahezu konstant ist und danach stark abfällt. Begründet wird diese Rate in der Analyse durch drei Punkte. Teilnehmer brachen zufällig ab, durch einen bestimmten Grad an Teilnahmslosigkeit. Bei langen Ketten kam es häufig vor, dass Briefe verloren gingen oder es möglicherweise gar nicht möglich war, das Ziel zu erreichen. Kurze Ketten wurden abgebrochen, da das Ziel sehr weit entfernt schien und daher eine Weiterleitung von den Personen nicht in Betracht gezogen wurde. Somit wurde auch gezeigt, dass die Netzwerkstruktur allein nicht ausschlaggebend für eine erfolgreiche Navigation ist, sondern von individuellen Anreizen der Teilnehmer abhängig ist.

In ihrer Publikation [26] haben Watts und Strogatz 1998 ein Modell veröffentlicht, das die zwei wesentlichen Aspekte von sozialen Netzwerken modelliert. Der erste Bereich beschreibt das Prinzip, dass Menschen zu anderen Menschen Verbindungen haben, die wie sie selbst sind und sich im näheren Umfeld befinden. Der zweite Typ bildet jene Verbindungen nach, die Menschen zu Bekanntschaften haben, die weiter von ihnen entfernt sind. Während erstere Verbindungen als Dreiecke in einem Graphen darstellbar sind, sind schwächere Beziehungen, sogenannte „weak ties“, als weitreichendere Kanten dargestellt. Wie in Abbildung 14 ersichtlich, zeigt das Raster die grundlegende Struktur und die zufälligen längeren Kanten stellen die langen Distanzen dar.

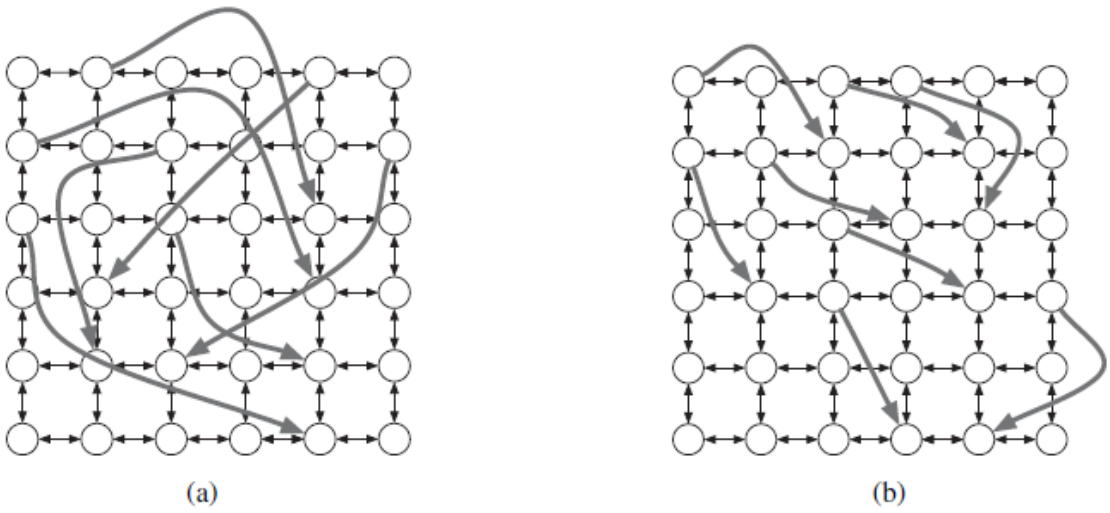

Abbildung 14: Watts-Strogatz-Modell - lange und kurze Zufallskanten [5]

3.4 Dezentralisierte Suche („Decentralized Search“)

Benutzer, die wie im Experiment von Milgram aus Abschnitt 3.3 handeln, wenden dezentralisierte Suchmethoden („Decentralized Search“) an. Dabei basiert die jeweilige Entscheidung nur auf lokalem Wissen, da es unmöglich ist, detaillierte Informationen über das gesamte Netzwerk zu erhalten. Die Ergebnisse haben gezeigt, dass Benutzer gut darin sind, nahezu kürzeste Pfade nur durch dieses lokale Hintergrundwissen zu finden. [5]

Um wirklich den kürzesten Pfad zwischen zwei Personen in einem sozialen Netzwerk zu finden, müsste in Milgrams Experiment der Brief von jeder Person, die einen Brief weitersendet, nicht nur eine Person gewählt werden, sondern eine Weiterleitung an alle Bekanntschaften erfolgen. Dabei spricht man von einer Breitensuche. Das gesamte Netzwerk wird dadurch „geflutet“, sodass von einer Startperson ausgehend alle Bekanntschaften und von all diesen wiederum alle Bekannte erreicht werden usw. Dadurch wäre der versandte Brief am schnellsten bei der Zielperson angekommen. [5]

In [25] wird die Grundidee der dezentralisierten Suche nochmals genauer beschrieben. In jedem Suchschritt, in dem man sich befindet, hat man nur Wissen über die zugrundliegende Rasterstruktur des Netzwerkes, den Ort des

Zielknotens und die eigenen weitreichenden Kanten. Über alle anderen Kanten in diesem Netzwerk ist im betrachteten Suchschritt keine Information verfügbar.

3.4.1 Hierarchien als Hintergrundwissen

Eine Variante, Hintergrundwissen darzustellen, ist neben dem zuvor erwähnten Rastermodell auch die Verwendung von Hierarchien. In dieser n-ären[18] Baumstruktur befinden sich die Knoten an den Blättern. Die Distanz zwischen einem Knotenpaar ist auf den kleinsten gemeinsamen Vorfahren zurückzuführen. Während ein Rastermodell besonders für eine Abstraktion des geographischen Bereichs geeignet ist, so nähert eine hierarchische Struktur das Umfeld bzw. den Tätigkeitsbereich an. Hierarchien eignen sich gut, um bestimmte Mengen an Information zu einem Überbegriff zusammenzufassen. [25]

Die Idee, Hierarchien als Hintergrundwissen einzusetzen, stammt von Kleinberg [9]. Er stellte alle Knoten in einem Informationsnetzwerk in dieser Baumstruktur dar, wobei ähnliche Knoten sich auch in der Hierarchie näher beieinander befinden.

3.4.2 Hierarchische Werte für Knoten

Um Knoten zu bewerten bzw. ein Maß für die Bedeutung eines Knotens zu finden, wurden verschiedene Verteilungen analysiert. In [20] wurde das Verhältnis von Eingangsgrad d_{in} und Ausgangsgrad d_{out} wie folgt definiert: $\frac{d_{in}}{d_{out}}$. Es wird auch darauf verwiesen, dass diese einfache Berechnung bessere Ergebnisse verspricht, als beispielsweise PageRank oder HITS. Die beiden letzteren sind ebenfalls Bewertungsverfahren für verlinkte Dokumente, setzen allerdings auf komplexere Berechnungen.

In [27] wird ein Algorithmus vorgestellt, der Hierarchien aus dem Netzwerk extrahiert. Hierbei wird das Netzwerk durchlaufen und bei Links anhand von Kriterien entschieden, ob derjenige Link für die hierarchische Struktur relevant ist. Um dies entscheiden zu können, wird ein hierarchischer Wert für einen

[18] In einem n-ären Baum darf kein Knoten mehr als n Nachfolgerknoten besitzen.

Knoten berechnet. Über- oder unterschreitet dieser rechnerische Wert die Schwellwerte, so wird er in die Hierarchie nicht eingeordnet. Die verwendete Berechnung ist eine Erweiterung der bereits in [20] vorgestellten Verhältniswertes: $\frac{d_{in}}{d_{out}}\sqrt{d_{in}}$. Der Wurzelterm berücksichtigt, dass nicht nur das reine Verhältnis zu tragen kommt, denn ohne diesen würde eine Seite z.B. mit 30 Eingangs- und 10 Ausgangskanten denselben Wert erhalten wie eine Seite mit 3000 Eingangs- und 1000 Ausgangskanten.

4. Datenrepräsentation

In diesem Kapitel wird definiert, welches Format die benötigten und verwendeten Daten besitzen. Die Datenaufbereitung, die eine Teilaufgabe dieser Arbeit war und in Kapitel 6 beschrieben wird, baut auf diesen Datenformaten auf.

4.1 Allgemeines

Das von Eder in [3] entwickelte Framework MUN, welches in Abschnitt 5.2 noch näher beschrieben wird, erwartet ein definiertes Format an Eingabedaten. Die Vorgabe basiert auf der Bibliothek SNAP, auf die ebenfalls noch in Abschnitt 5.1 Bezug genommen wird. Aufgrund von bereits implementierten Methoden dieser Bibliothek, wurde die Struktur dieser Daten zum Teil daraus übernommen, um eine gute Kompatibilität zu bieten.

4.2 Wikipedia-Netzwerk

Ein Teil der Datenrepräsentation bezieht sich auf das Wikipedia-Netzwerk. Wie bereits erwähnt, erfolgt die Modellierung durch Knoten, die Artikel repräsentieren, und durch Kanten, die den Links zwischen Artikeln entsprechen. Jeder Knoten besitzt eine eindeutige Identifikationsnummer, die einer natürlichen Zahl entspricht und nachfolgend zur einfacheren Beschreibung mit ID abgekürzt wird. Es existiert eine eindeutige Zuordnung zwischen Artikel-Titel und Artikel-ID (Knoten-ID).

4.2.1 Graph

Das Format eines Graphen entspricht einer Datei, in der jede Zeile eine gerichtete Kante folgendermaßen beschreibt: Der ID des Knotens, von dem die Kante ausgeht, gefolgt durch einen Leerraum und abschließend die ID des Knotens, zu dem die Kante zeigt. Der Leerraum kann mit Leerzeichen oder Tabulatoren gefüllt werden. Listing 1 zeigt ein Beispiel dieser Repräsentation. Jene Datei enthält zwölf Zeilen, die zwölf Kanten im Graphen entsprechen. Die erste Zeile stellt z.B. eine Verbindungskante von Knoten 1 zu Knoten 2 dar.

Eine getrennte Angabe aller benötigten Knoten-IDs ist nicht nötig, da diese durch die Angabe von entsprechenden Kanten beim Einlesen implizit angelegt werden. Eine Kantenliste, gespeichert in einer Datei, modelliert somit einen Graphen.

```
1     2
1     3
1     4
1     5
3     6
3     7
4     5
4     7
4     8
5     9
5     10
11    9
```

Listing 1: Datei zur Beschreibung eines Graphen

Zur Veranschaulichung enthält Abbildung 15 die Visualisierung jenes Graphen aus Listing 1. Elf Knoten werden durch die Kantenliste beschrieben. Wie ersichtlich, werden alle Kanten gerichtet dargestellt. Eine Kante von Knoten 2 zu Knoten 1, d.h. eine Verbindung in die entgegengesetzte Richtung, wie sie in Zeile 1 in Listing 1 beschrieben wird, müsste durch eine weitere Zeile erfolgen.

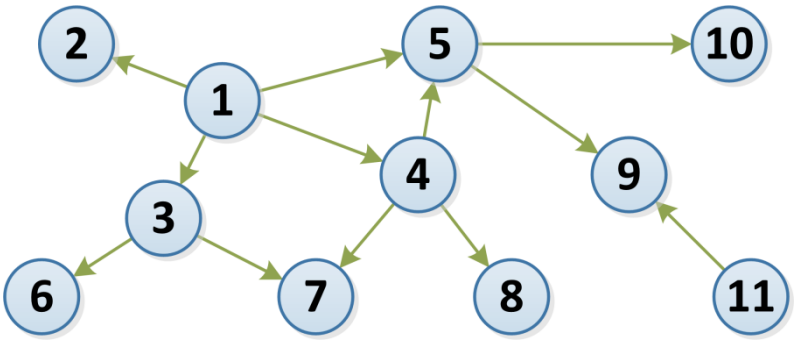

Abbildung 15: Visualisierung einer Kantenliste

Der in diesem Projekt verwendete Graph basiert auf einem Export der Wikipedia-Daten vom 15. November 2011. Da die Artikel in Wikipedia durch Links sehr stark vernetzt sind, wächst die Dateigröße der Listen enorm an. Der Wikipedia-Graph besitzt rund 3,8 Millionen Knoten und ca. 232 Millionen Kanten. Jene Datei, die diesen Graphen repräsentiert, besteht somit aus einer Kantenliste mit

232 Millionen Zeilen und besitzt eine Größe von ca. 3,6 GB. Aus diesem Grund war es notwendig, statt einer Textdatei, die Kantenlisten auch in einem binären Format abzuspeichern. In [4] nimmt Geigl Bezug auf den Konverter zwischen den beiden Formaten. Die Laufzeit beim Einlesen der Daten lässt sich dadurch deutlich verkürzen sowie der Verbrauch des Arbeitsspeichers reduziert sich. Auch die Dateigröße des Graphen verringert sich im Binärformat auf 1,9 GB, das fast der Hälfte entspricht.

Weitere Informationen zur Größe und Struktur des in diesem Projekt verwendeten Wikipedia-Graphen finden sich in Kapitel 6.

4.2.2 Hierarchie

Hierarchien sind Graphen, die eine Baumstruktur besitzen. Aus diesem Grund können sie wie einfache Graphen in Abschnitt 4.2.1 beschrieben werden. Sie besitzen einen Wurzelknoten, von dem ausgehend alle Knoten über einen Pfad erreichbar sind. Aufgrund der Definition kann ein Baum nie mehr Kanten als Knoten besitzen, wodurch die Länge der Kantenliste durch die Anzahl der Knoten des Graphen begrenzt ist. In Listing 2 ist eine Hierarchie-Datei zu sehen, die anhand der Breitensuche (siehe Abschnitt 6.6.1) beginnend von Knoten 1 (Wurzel) aus dem Graphen in Listing 1 erstellt wurde.

```
1    2
1    3
1    4
1    5
3    6
4    7
4    8
5    9
5    10
```

Listing 2: Datei zur Beschreibung einer Hierarchie

Abbildung 16 zeigt die zugehörige Visualisierung von Listing 2. Dabei fällt auf, dass die Hierarchie den Knoten 11 nicht beinhaltet, da dieser im Graphen von Knoten 1 aus nicht erreichbar ist.

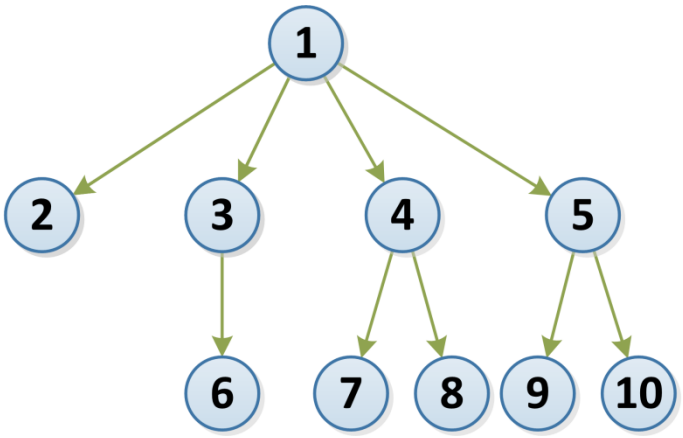

Abbildung 16: Visualisierung einer Hierarchie

4.3 Wikigame-Daten

Weitere Daten, deren Repräsentation in diesem Projekt Verwendung finden, sind jene des Wikigames. Die Artikel bzw. Knoten sind wie auch in Abschnitt 4.2 durch IDs und derselben Zuordnung repräsentiert.

4.3.1 Navigationspfade

In einer Datei, die Klickpfade beinhaltet, entspricht jede Zeile einer Navigation. Diese hat folgenden Aufbau: Zu Beginn wird die ID des Startknotens, gefolgt von der ID des Zielknotens angegeben. Danach ist der komplette Pfad, über den der Benutzer navigiert hat, enthalten. In Listing 3 sind sieben Navigationspfade dargestellt. Beim ersten Pfad, der von Knoten 1 zu Knoten 5 führt, ist der Weg durch einen einzigen direkten Schritt erfolgt. Der zweite Pfad beinhaltet dieselben Start- und Zielknoten, allerdings ist hier eine Navigation über Knoten 4 erfolgt, d.h. die Kanten sind 1 → 4 und 4 → 5.

```
1    5    1    5
1    5    1    4    5
1    10   1    4    5    10
1    10   1    5    9
4    10   4    5
4    10   4    5    10
1    7    1
```

Listing 3: Datei zur Beschreibung der Klickpfade von mehreren Benutzern

Durch die Angabe des Zielknotens an zweiter Stelle einer Zeile kann zwischen jenen Pfaden, die das Ziel erreicht haben und jenen, die das Ziel nicht erreicht haben, unterschieden werden. Diese Unterscheidung ist notwendig, da die Analysen auf der Aufteilung zwischen erfolgreichen und nicht erfolgreichen Navigationspfaden basieren.

Des Weiteren sind bei den Navigationspfaden zwei Besonderheiten zu berücksichtigen. Die erste bezieht sich darauf, dass Benutzer beim Navigieren im Netzwerk die Zurück- und Vorwärts-Schaltflächen des Browsers benutzen können, um zu bereits besuchten Seiten zurück zu gelangen. Jene Klicks werden in den Pfaden allerdings nicht mitgespeichert. Weitere Ausnahmen sind Weiterleitungen. Sollte ein Benutzer von einer Seite A auf eine Seite B gelangen, die automatisch auf Seite C weiterleitet, so dürfen im Navigationspfad nicht beide Schritte enthalten sein. Dies würde nicht dem Benutzerverhalten entsprechen, da nur ein Klick erfolgt ist. In jenem Fall beinhaltet die Datei nur die Seiten-ID, auf die weitergeleitet wurde.

4.3.2 Paare

Eine weitere Repräsentationsform die benötigt wird, ist eine Datei, die alle Start- und Zielknoten von Navigationspfaden beinhaltet. Wie auch in den bereits beschriebenen Formaten, entspricht auch hier eine Zeile einem Datensatz. Nach den beiden IDs der Start- und Zielknoten wird zusätzlich noch der Eingangsgrad des Startknotens und der Eingangsgrad des Zielknotens angegeben. Listing 4 zeigt sechs Paare. Das erste enthaltene Paar besitzt z.B. Startknoten-ID 1 und Zielknoten-ID 5, der Eingangsgrad von Knoten 1 beträgt 0, jener von Knoten 5 entspricht 2. Dieselben Paare können mehrfach enthalten sein, wenn mehrere Navigationspfade dieselben Knoten beinhalten.

```
1    5    0    2
1    5    0    2
1    10   0    1
4    10   1    1
4    10   1    1
1    3    0    1
```

Listing 4: Datei zur Beschreibung der Paare von Navigationspfaden

4.4 Kürzeste Distanzen

Ein weiteres Format dient der Darstellung von kürzesten Distanzen zwischen zwei Knoten. Diese Datei ist grundlegend in Abschnitte unterteilt. Ein neuer Abschnitt beginnt jeweils mit der Sequenz „-1“. In derselben Zeile ist, getrennt durch einen Leerraum, eine Startknoten-ID angegeben. Die darauffolgenden Zeilen beinhalten den Zielknoten mit der zugehörigen kürzesten Distanz. Eine Erklärung kann anhand des Beispiels in Listing 5 erfolgen. Der erste Abschnitt beginnt mit „-1“ und legt den Startknoten mit der ID 1 fest. Die zweite Zeile sagt aus, dass die Distanz von Knoten 1 zu Knoten 2 einer Länge von 1 entspricht. Die Distanz zwischen Knoten 1 und Knoten 6 beträgt 2, wie aus Zeile 3 hervorgeht. Danach erfolgt noch eine weitere Distanz-Angabe, bevor ein neuer Abschnitt beginnt, der ID 4 als Startknoten besitzt.

```
-1   1
2    1
6    2
10   2
-1   4
7    1
8    1
9    2
10   2
```

Listing 5: Datei zur Beschreibung von kürzesten Distanzen im Graphen

5. Arbeitsumgebung

Dieses Kapitel beschreibt die Arbeitsumgebung, in der die Entwicklung dieses Projekts stattfand. Das von Eder in [3] entwickelte Framework MUN zur Navigationssimulation setzt auf der Bibliothek SNAP auf.

5.1 Bibliothek SNAP

Die Stanford Network Analysis Platform[19], kurz SNAP genannt, ist eine in C++ geschriebene Bibliothek, die an der Stanford University in Kalifornien (USA) entwickelt wurde. Sie wird zur Netzwerk-Analyse und zum Aufbau bzw. zur Repräsentation von Graphen verwendet.

Die Bibliothek, die seit 2004 entwickelt wird, ist besonders effizient und bietet Verwendung für extrem große Netzwerke mit über 100 Millionen Knoten und Milliarden Kanten. Mit Stand vom Jänner 2013 ist SNAP in der Version 1.11 verfügbar und kann von der Projekt-Website frei heruntergeladen und verwendet werden. Sie steht unter der BSD-Lizenz.

SNAP bietet zahlreiche Funktionen zum Einlesen, Modifizieren und Speichern von gerichteten und ungerichteten Graphen. Zur Visualisierung von Daten lässt sich eine Verwendung mit Gnuplot[20], Graphviz[21] oder NodeXL[22] einfach realisieren.

Der Einsatz dieser Bibliothek ist weitreichend. Es sind zahlreiche Beispieldatensätze auf der Website[23] frei verfügbar, die beispielsweise Daten von Facebook, Google+ oder Twitter modellieren – sowohl durch gerichtete als auch durch ungerichtete Graphen.

---

[19] http://snap.stanford.edu

[20] ein Paket, das Daten grafisch darstellt und ausgibt - Anwendung v.a. für die grafische Darstellung von strukturellen Eigenschaften in Netzwerken

[21] ein Programmpaket zur Visualisierung von Objekten und Beziehungen – Anwendung v.a. zur Darstellung von kleinen Graphen

[22] eine grafische Oberfläche, um SNAP mit Microsoft Office bzw. Excel einfach bedienen zu können

[23] http://snap.stanford.edu/data/index.html

Auch zu Wikipedia sind mehrere Datensätze verfügbar. Hierbei werden z.B. Umfragen oder Diskussionsseiten angeboten. Auch die gesamte Geschichte aller Wikipedia-Bearbeitungen von Autoren wurde modelliert. Darin sind 2,3 Millionen Benutzer, 3,5 Millionen Seiten und 250 Millionen Kanten enthalten. Dies zeigt, dass mit dieser Bibliothek eine enorme Datenmenge verarbeitet werden kann. Das Größenverhältnis der zuletzt erwähnten Datenmodellierung entspricht in etwa jener, die in diesem Projekt Anwendung findet.

5.2 Framework MUN

Im Rahmen dieses Projektes wurde ein eigenes Framework entwickelt. Es trägt den Namen Modeling User Navigation, abgekürzt MUN, und wurde von Eder in [3] erstellt und zur Verfügung gestellt. Geschrieben in C++ bietet es durch die Unterstützung der Bibliothek SNAP zahlreiche Funktionen zur Modellierung von Benutzernavigation.

Eine Dokumentation des gesamten Codes wurde mittels Doxygen[24] realisiert. Sowohl alle Klassen als auch deren Methoden inklusive Parameter und Rückgabewerte sind ausführlich dokumentiert.

Der Ablauf nach dem Start des Frameworks gliedert sich in drei größere Schritte. Anfangs werden die Eingabedaten eingelesen, danach erfolgt die Navigationssimulation an sich. Nach dem Ablauf dieser beiden Schritte, bietet das Framework Evaluierungsmöglichkeiten der Ergebnisse. Die Konfiguration des Programmablaufs erfolgt durch eine Vielzahl an Parametern. Diese können entweder in einer Konfigurationsdatei angegeben werden, oder direkt beim Aufruf auf der Kommandozeile mitgegeben werden.

Um das Framework lauffähig zu machen, werden grundsätzlich zwei Eingabedateien benötigt. Eine beinhaltet den Graphen, z.B. das Wikipedia-Netzwerk, und eine andere die Hierarchie, die auf dem Graphen basiert.

Neben diesen Dateien besteht noch die Möglichkeit, weitere Dateien anzugeben. Diese sind jedoch optional, da im Falle keiner Angabe jene Daten durch

[24] ein freies Software-Dokumentationswerkzeug - http://www.stack.nl/~dimitri/doxygen/

das Framework berechnet werden. Zum einen betrifft dies die Datei, die alle Navigationspaare beinhaltet, d.h. jeweils Start- und Zielknoten. Dadurch wird dem Navigator mitgeteilt, welche Navigationen erfolgen sollen. Zum anderen kann jene Datei, die die kürzesten Distanzen zwischen Knotenpaaren beinhaltet, ebenfalls beim Durchlauf berechnet werden. Da vor allem letztere Datei aufwändige Berechnungen mit sich zieht, steigt dadurch die Laufzeit stark an.

Auf alle erwähnten Dateien und deren Erstellung wird in dieser Arbeit in Kapitel 6 Bezug genommen. Bei der Datenaufbereitung werden alle benötigten Dateien detailliert beschrieben.

Nach dem Einlesen bzw. Generieren der benötigten Eingabedaten erfolgt die Navigationssimulation, die auf der definierten Konfiguration und den Knotenpaaren basiert. Sogenannte Node-Selektoren, die auf einer Schnittstelle des Frameworks basieren, wurden von Geigl in [4] implementiert. Anhand dieser entscheidet das Framework, basierend auf Wahrscheinlichkeiten und Konfigurationsmöglichkeiten, wie die einzelnen Navigationspfade erstellt werden.

Um sich den Benutzern bezüglich Abbruchverhalten anzunähern, bietet das Framework dafür eine weitere Schnittstelle. In dieser Arbeit wurden verschiedene Abbruchraten implementiert und getestet. Eine ausführliche Beschreibung befindet sich in Kapitel 7.

Auch auf die anschließende Evaluierung wird in [4] eingegangen sowie ebenfalls in dieser Arbeit in Kapitel 7. Hierbei wurden mehrere Plots zur Visualisierung der Ergebnisse erstellt. Um einen Vergleich zwischen generierten Pfaden und Benutzerpfaden – in diesem Projekt aus dem Wikigame – ziehen zu können, stellen Diagramme auch die Analyseergebnisse jener beiden Kollektionen gegenüber.

5.3 Skripte

Da im Zuge der Datenaufbereitung (siehe Kapitel 6) zahlreiche Iterationsschritte notwendig waren um an die endgültigen Daten zu gelangen, wurden mehrere

Skripte entwickelt. Zur Anwendung kamen Bash[25]-Skripte als auch PHP-Skripte.

Auf der Bash wurden zum effizienten Umgang mit Dateien zahlreiche Unix-Werkzeuge wie awk, sed oder grep eingesetzt. Es wurde vor allem beim Parsen der Grunddaten auf reguläre Ausdrücke gesetzt.

Skripte, geschrieben in PHP, boten sich ebenfalls durch den gut konfigurierbaren Zugriff auf die Datenbankschnittstelle an.

5.4 Datenbank - MySQL

Bei MySQL handelt es sich um ein relationales Datenbankverwaltungssystem. Dieses wurde eingesetzt um mehrere vorliegende Grunddaten, sowohl von Wikipedia als auch vom Wikigame, einzulesen um danach effizienten Zugriff zu erlauben. Datenbank-Abfragen wurden dazu formuliert und in weiterer Folge verfeinert. Auch hierzu finden sich Details in Kapitel 6.

[25] Bourne-again shell – Benutzerschnittstelle in Unix-/Linux-Systemen

6. Datenaufbereitung

Ein Hauptthema dieser Arbeit behandelt dieses Kapitel über Datenaufbereitung. Es wird darin beschrieben, auf welche Art und Weise Daten generiert, verarbeitet und aufbereitet wurden, damit sie dem Eingabeformat des Frameworks entsprechen.

6.1 Übersicht des gesamten Ablaufes

Der Ablauf der Datenaufbereitung gliedert sich in mehrere Abschnitte, die nacheinander durchlaufen werden. Die Unterabschnitte in diesem Kapitel entsprechen den einzelnen Bereichen.

Abbildung 17 zeigt eine Übersicht, die Einblick in die einzelnen Schritte gibt. Nachfolgend eine kurze Erläuterung, bevor eine detailliertere Beschreibung folgt.

Begonnen wurde mit einer Bestandsanalyse aller Daten, sowohl das Wikipedia-Netzwerk als auch das Wikigame betreffend. Nach der Auswahl von geeigneten Daten erfolgten der Download und die anschließende Analyse. Unter Berücksichtigung von Einschränkungen konnte der Wikipedia-Graph extrahiert werden. Da es durch mehrere Evaluierungen zu Unstimmigkeiten kam, wurde eine Bereinigung durchgeführt.

Nachdem die Grunddaten des Wikipedia-Netzwerks vorhanden waren, konnte mit der Erzeugung von Hierarchien begonnen werden. Die Erstellung basiert auf einer Breitensuche bzw. auf hierarchischen Werten.

Navigationspfade wurden anschließend aus den zur Verfügung gestellten Wikigame-Daten gewonnen und danach ebenfalls bereinigt.

Im Anschluss wurden noch weitere benötigte Eingabedaten, wie z.B. Paare und kürzeste Distanzen, erzeugt.

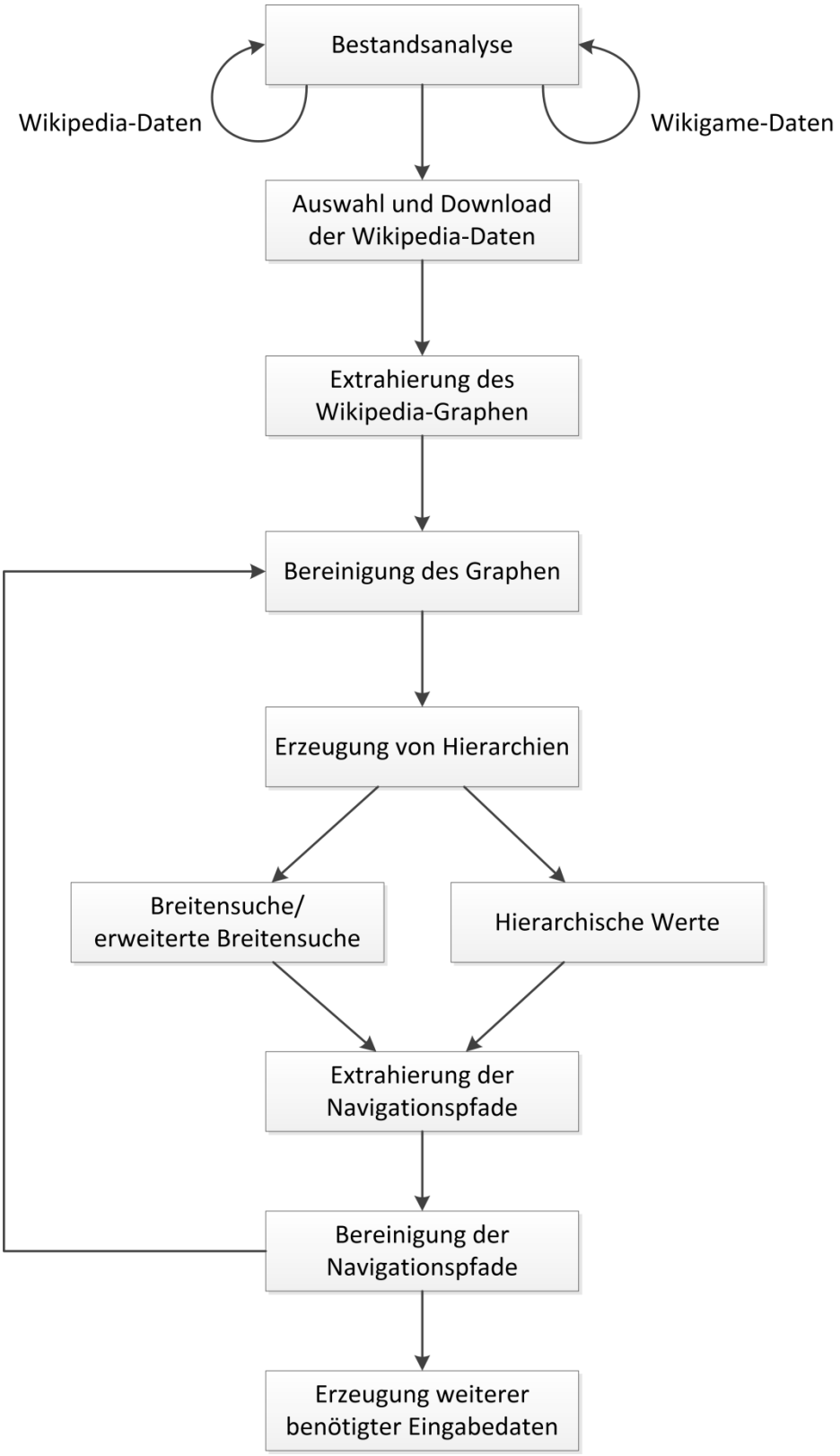

Abbildung 17: Flussdiagramm - Ablauf der Datenaufbereitung

6.1.1 Tools

In diesem Kapitel werden auch zahlreiche Tools beschrieben, die durch die Datenaufbereitung benötigt wurden und im Rahmen dieser Arbeit implementiert wurden. Da es sich hierbei um Programme handelt, die häufig nur bei der Erzeugung von Grunddaten benötigt werden, wurden diese nicht in das Framework inkludiert. Als Beispiele können jene Tools zur Auflösung von Weiterleitungen oder zur Erzeugung von Hierarchien genannt werden.

Alle beschriebenen Tools können eigenständig gestartet und ausgeführt werden. Sie sind allerdings stark an das Framework MUN sowie die Bibliothek SNAP gekoppelt, da sie Klassen und auch Methoden daraus verwenden.

6.2 Bestandsanalyse

Um grundsätzlich an die Bestandsanalyse von Daten heranzugehen, musste neben der Verfügbarkeit auch auf die Beschaffenheit geachtet werden. Da man hier von einzelnen Files spricht, die von mehreren hundert MB bis zu 10 GB an Größe besitzen, musste die Verarbeitung dieser auch dementsprechend effizient ablaufen. Auch bei der Selektion der Daten flossen u.a. diese Merkmale ein.

6.2.1 Wikipedia

Bei der Repräsentation von Wikipedia wurde nicht auf bereits vorhandene bearbeitete Daten gesetzt, sondern eine Neugenerierung gewählt. Ein dafürsprechender Grund ist, dass dadurch die Herkunft genauestens bekannt ist und von einer Original-Quelle stammt. Dadurch sind unbekannte oder nicht erwähnte Modifikationen, obgleich sie bewusst oder unbewusst erfolgten, ausgeschlossen. Des Weiteren konnte durch die Erstellung eine Analyse der einzelnen Datensätze erfolgen, um die Vor- und Nachteile jeweils abzuwägen.

Die Wikimedia Foundation, die als Non-Profit-Organisation u.a. die Wissensdatenbank Wikipedia betreibt, bietet auf ihren Webseiten[26] zahlreiche sogenannte Datenbank-Dumps zum Download an. Bei diesen Dumps handelt es sich um Auszüge aus der Datenbank, die in bestimmten periodischen Abständen aus dem aktuellen Datenbestand erfolgen. Diese sind frei zugänglich und können für alle Sprachen abgerufen werden. Mit Stand vom Dezember 2012 stehen die letzten 14 Dumps[27] der englischen Wikipedia zur Verfügung, die vom 15. November 2011 bis zum 1. Dezember 2012 reichen. Somit wird monatlich ein Dump bereitgestellt, der anschließend für ca. ein Jahr verfügbar ist.

Datenbankauszüge stehen für alle Sprachen, in denen eine Wikipedia angeboten wird, zur Verfügung. In diesem Projekt wird die englische Wikipedia als zugrunde liegende Quelle gewählt. Wie bereits in den Grundlagen in Abschnitt 2.3 erwähnt, ist jene von der Anzahl ihrer Artikel die größte derzeit verfügbare Enzyklopädie.

Für eine erste Analyse empfiehlt es sich daher, nicht die englischsprachigen Dumps zu wählen, sondern jene, die eine weitaus geringere Menge an Daten beinhalten. Um sich eine Übersicht zu verschaffen, wurde daher die Wikipedia der Sprache Twi[28] gewählt. Tabelle 1 zeigt einen Vergleich der Wikipedia-Dumps in den Sprachen Englisch, Deutsch und Twi. Die Anzahl der Artikel entspricht reinen Artikel-Seiten und die Größe der Artikel-Dumps ist im komprimierten Format angegeben.

Tabelle 1: Vergleich Wikipedia-Dumps vom Jänner 2013

Sprache	Anzahl Artikel	Größe Artikel-Dump
Englisch	4.149.223	9,0 GB
Deutsch	1.540.474	2,9 GB
Twi	151	0,2 MB

[26] http://dumps.wikimedia.org/

[27] http://dumps.wikimedia.org/enwiki/

[28] http://tw.wikipedia.org/

Nicht nur die Dumps der Artikelseiten stehen zum Download bereit, sondern über 25 verschiedene Exporte mit unterschiedlichen Inhalten werden angeboten. All jene Datenbankauszüge richten sich in erster Linie an Entwickler und Forscher, um Einblicke in die Struktur hinter Wikipedia zu erhalten und die zugehörigen Datensätze, wie in dieser Arbeit, weiterverarbeiten zu können.

Dumps sind komprimiert verfügbar und besitzen entweder das SQL[29]- oder das XML[30]-Format. SQL-Dumps bieten den Vorteil, dass sie direkt nach dem Entpacken in eine MySQL-Datenbank eingelesen werden können, ohne dass dazu zusätzliche Software benötigt wird.

Listing 6 beinhaltet ein Beispiel für einen SQL-Datenbank-Dump. Es wird die zu erstellende Tabelle „pagelinks“ mit den zugehörigen Feldern beschrieben, anschließend die zugehörigen Inhalte.

```
...
CREATE TABLE `pagelinks` (
  `pl_from` int(8) unsigned NOT NULL DEFAULT '0',
  `pl_namespace` int(11) NOT NULL DEFAULT '0',
  `pl_title` varbinary(255) NOT NULL DEFAULT '',
  ...
) ENGINE=InnoDB DEFAULT CHARSET=binary;
...
INSERT INTO `pagelinks` VALUES
(3,10,'Pending_deletion'),
(10,0,'Computer_accessibility'),
(10,4,'Subpages'),
(12,0,'1917_October_Revolution'),
...
```

Listing 6: Auszug aus einem SQL-Dump

Die zweite Form von Dumps sind jene, die eine XML-Struktur besitzen. Durch das weit verbreitete und häufig verwendete Dateiformat, bietet es eine hohe Kompatibilität zu anderer Software, im speziellen XML-Parsern. Ein direkter Import in eine Datenbank kann jedoch im Gegensatz zum SQL-Format nicht erfolgen. Wikimedia stellt Tools in PHP und Java zur Verfügung, mit denen ein Import in eine MySQL-Datenbank möglich ist.

[29] Structured Query Language

[30] Extensible Markup Language

Ein Auszug aus dem XML-Dump „pages-articles“ wird in Listing 7 dargestellt. Darin sind der Inhalt und einige Metadaten enthalten.

```
...
<page>
 <title>AccessibleComputing</title>
 <id>10</id>
 <redirect />
 <revision>
  <id>381202555</id>
  <timestamp>2010-08-26T22:38:36Z</timestamp>
  <contributor>
   <username>OlEnglish</username>
   <id>7181920</id>
  </contributor>
  <minor />
  <comment>...</comment>
  <text xml:space="preserve">
   #REDIRECT [[Computer accessibility]] ...
  </text>
 </revision>
</page>
...
```

Listing 7: Auszug aus einem XML-Dump

In Tabelle 2 sind angebotene Dumps gelistet, aus denen relevante Informationen extrahiert werden können. Die übrigen Exporte wurden für dieses Projekt nicht benötigt. Die angegeben Größen sind ebenfalls wieder in komprimierter Form angegeben. Es kommen die Komprimierungsverfahren bzip2 und gzip zur Anwendung.

Tabelle 2: Auswahl an verfügbaren Wikipedia-Dumps

Dump-Name	Beschreibung	Größe	Format
pages-articles	vollständige Inhalte von Artikel, Vorlagen, Datei-Beschreibungen etc.	9.0 GB	XML
page	Seiten-Daten (ID, Titel, etc.)	909,9 MB	SQL
redirect	Weiterleitungen	78,8 MB	SQL
pagelinks	Links zwischen Seiten	4,7 GB	SQL

Der Dump „pages-articles“ beinhaltet die vollständigen Inhalte aller Artikel, die sich in der Wikipedia befinden. In Listing 7 ist bereits ein Abschnitt aus diesem

Auszug beschrieben. Neben dem Titel, der ID und dem vollständigen Inhalt der aktuellen Revision, sind auch noch Metadaten, die sich z.B. auf den Autor, das Änderungsdatum oder Weiterleitungseigenschaften beziehen, enthalten.

Detailinformationen zu den einzelnen Seiten befinden sich in „page“. In dieser Tabelle sind vor allem die ID, der Titel und der Namensraum jeder Seite von Bedeutung. Wie später noch erwähnt, dient diese Tabelle als Basis für die Knoten-ID-Zuordnung im Graphen.

Weiterleitungen sind im SQL-Dump „redirect“ abgebildet. Darin werden sowohl die Weiterleitungsseite als auch die Zielseite und der zugehörige Namensraum gelistet.

Ferner steht noch der Dump „pagelinks“ zur Verfügung, der alle Links zwischen Artikeln beinhaltet. Ein darin enthaltener Datensatz ist ähnlich wie jener in den Weiterleitungsdaten aufgebaut und beinhaltet die Seite, die den Link enthält, samt Namensraum und die Zielseite, zu der der entsprechende Link zeigt.

6.2.2 Wikigame

Die Daten für das Wikigame sind nicht frei verfügbar bzw. nicht per Download zu beziehen, sondern wurden beim Autor der Spiel-Website angefordert. Alex Clemesha[31] stellte einen Datenbank-Dump für Forschungsarbeiten zur Verfügung. Der Datenbank-Export besitzt eine Größe von 1,9 GB und beinhaltet 19 Tabellen. Es sind darin rund 780.000 eindeutige Missionen des Spieles enthalten und insgesamt ca. 2,3 Millionen Navigations- bzw. Klickpfade. Die Daten beziehen sich auf einen Zeitraum von Februar 2009 bis September 2011.

Im Rahmen dieser Arbeit wurden zwei Tabellen daraus benötigt: Zum einen jene, die alle Missionen mit Start- und Zielartikel beinhaltet, und zum anderen jene, die alle Benutzerklicks mit zeitlichem Ablauf zur Verfügung stellt.

[31] http://www.clemesha.org/

6.3 Auswahl und Download der Wikipedia-Daten

Wie bereits genannt, stehen – mit Stand vom Dezember 2012 – 14 Dumps zur Auswahl. Es wurden der älteste verfügbare Dump vom 15. November 2011 und der aktuellste Dump vom 1. Dezember 2012 herangezogen. Der Vergleich hat einen enormen Unterschied gezeigt, wodurch die Auswahl schließlich auf den ältesten fiel. Begründet ist diese Wahl u.a. dadurch, dass dieser Datenbank-Export mit jenem Zeitraum, in dem die Wikigame-Daten erstellt wurden, nahezu übereinstimmt. Der Wikipedia-Graph und alle darauf aufbauenden Daten basieren somit auf dem Dump vom 15. November 2011.

Der Download erfolgte von der Wikimedia-Website[32] direkt und beinhaltet die nachfolgend angeführten Dumps:

- enwiki-20111115-pages-articles.xml
- enwiki-20111115-page.sql
- enwiki-20111115-redirect.sql
- enwiki-20111115-pagelinks.sql

6.4 Extrahierung des Wikipedia-Graphen

Die Extrahierung des Wikipedia-Graphen kann in zwei Bereiche aufgeteilt werden. Es soll einerseits eine Kantenliste aller Links in Wikipedia erstellt werden und andererseits eine Kantenliste, die nur Weiterleitungslinks enthält. Diesen Listen liegt eine eindeutige Auflösung von Knoten-IDs zugrunde.

6.4.1 Zuordnung von IDs

Von besonderer Wichtigkeit in diesem Projekt ist die eindeutige Zuordnung von Knoten-IDs. Aus diesem Grund wurde der Dump „page“ herangezogen und in eine MySQL-Datenbank importiert. Dadurch können vor allem Abfragen, die von einem Artikel-Titel die zugehörige ID ausgeben, als auch die umgekehrte

[32] http://dumps.wikimedia.org/enwiki/20111115/ - Aufruf am 15. Dezember 2012

Auflösung von ID auf dessen Titel erfolgen. Des Weiteren sind Überprüfungen auf Weiterleitungen oder auf Namensraum-Zugehörigkeit möglich.

Um einen effizienten Import und schnelle Abfragen zu gewährleisten, mussten am Datenbank-Server einige Modifikationen vorgenommen werden, die vor allem Überprüfungen und Einschränkungen betrafen. Limits für Abfragen und temporäre Tabellen wurden angehoben sowie Prüfungen auf eindeutige Primärschlüssel aufgehoben.

6.4.2 Kantenliste mit Links

Zur Erstellung einer Kantenliste mit allen enthaltenen Links werden die Dumps „pagelinks“ und „page“ herangezogen. Auf einen Import von „pagelinks“ in eine MySQL-Datenbank wurde verzichtet, da auf diese Daten in weiterer Folge nicht wieder zugegriffen werden muss. Aus diesem Grund wurde ein Skript geschrieben, das aus dem Dump jene Inhalte mit einem Parser und regulären Ausdrücken herausfiltert und diese in eine Datei ausgibt. Eine Beispielausgabe von fünf Links ist in Listing 8 ersichtlich. Betrachtet man die erste Zeile, so stellt dieser Datensatz eine Link und somit eine Kante von Knoten-ID 10 zu jenem Knoten mit dem Titel „Computer_accessibility“ dar.

```
...
10   Computer_accessibility
12   1917_October_Revolution
12   1919_United_States_anarchist_bombings
12   19th_century_philosophy
12   6_February_1934_crisis
...
```

Listing 8: Auszug der Kantenliste der Wikipedia-Daten ohne ID-Auflösung

Anhand des Dumps „page“ wurden anschließend alle Titel der Linkziele durch ihre eindeutig definiere ID ersetzt. Dieses nachträgliche Ändern ist notwendig, weil der Wikipedia-Dump nicht die ID des Linkzieles beinhaltet, sondern nur den Titel. Nach der Umsetzung besitzt die Liste jenes Format, das in Listing 9 ersichtlich ist. Der Titel „Computer_accessibility“ wurde z.B. auf die ID 411964 aufgelöst.

```
...
10   411964
12   3721152
12   11284618
12   28357259
12   5013592
...
```

Listing 9: Auszug der Kantenliste der Wikipedia-Daten

Würde der Graph durch diese Kantenliste repräsentiert werden, so sind darin etwa 8,9 Millionen Knoten enthalten. Da laut Wikipedia-Statistik zu diesem Zeitpunkt nur rund 3,8 Millionen Artikel vorhanden waren, entspricht dies nicht dem gewünschten Ergebnis.

6.4.3 Kantenliste mit Weiterleitungen

In weiterer Folge wurde festgestellt, dass in der erstellten Kantenliste Weiterleitungen enthalten sind. Summiert man alle Weiterleitungslinks in Wikipedia auf, so beträgt diese Anzahl 5,1 Millionen. Würde man diese Weiterleitungslinks im Wikipedia-Graphen auflösen, entspricht dieses Ergebnis der Anzahl der Artikel, ca. 3,8 Millionen. Um nun anschließend eine Bereinigung durchführen zu können, wurde eine weitere Kantenliste, die nur aus Weiterleitungslinks besteht, erstellt. Listing 10 zeigt wieder ein Beispiel: ID 10 mit dem Titel „AccessibleComputing“ ist beispielsweise eine Weiterleitung auf ID 411964 mit dem Titel „Computer_accessibility“.

```
...
10   411964
13   13813
14   12681
15   66468
...
```

Listing 10: Auszug der Kantenliste mit Weiterleitungen

6.5 Bereinigung des Graphen

Nachdem die beiden Kantenlisten aus Abschnitt 6.4 erzeugt wurden, wurde die Bereinigung durch zwei geschriebene Tools durchgeführt. Um die Laufzeit bei den Modifikationen zu verkürzen, wurde auch hier das Binärformat eingesetzt.

6.5.1 Auflösung von Weiterleitungen

Bei der Auflösung von Weiterleitungen kommt ein implementiertes Tool zum Einsatz, das einen Graphen und eine geeignete Weiterleitungsliste, wie im Abschnitt 6.4 beschrieben, entgegen nimmt. Der Vorgang lässt sich am besten anhand von Abbildung 18 (links) erklären. In diesem Beispiel existiert von den Seiten 1, 2 und 3 jeweils ein Link zu Seite 5, die automatisch auf Seite 6 weiterleitet. Der Weiterleitungslink ist rot dargestellt. Seite 4 besitzt hingegen einen Link, der direkt auf Seite 6 zeigt. Nun ist es das Ziel, die Weiterleitungsseite 5 zu entfernen und die Links direkt auf Seite 6 zeigen zu lassen. Das Ergebnis nach dem Durchlauf dieses Tools ist in Abbildung 18 (rechts) zu sehen. Knoten 5 existiert nicht mehr, stattdessen zeigen alle Links von den Seiten 1, 2, 3 und 4 direkt auf Seite 6.

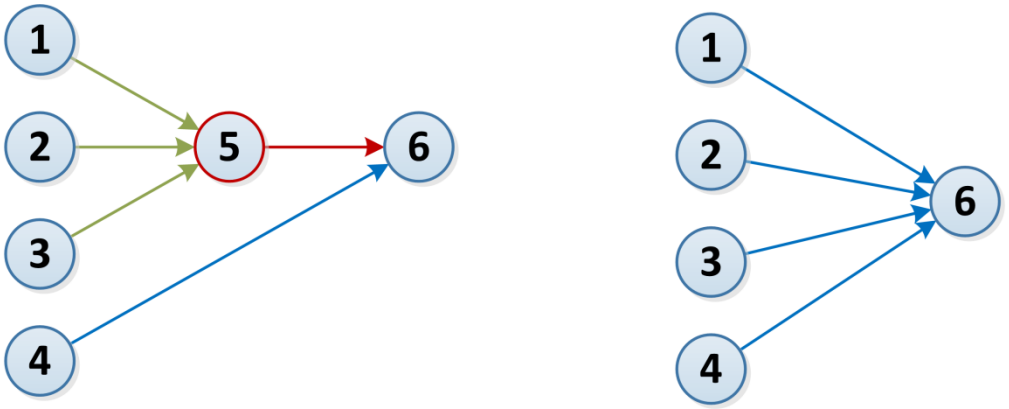

Abbildung 18: Auflösung von Weiterleitungen (links davor, rechts danach)

Jene geänderte Variante entspricht genau dem gewünschten Ergebnis bei Navigationspfaden. Um von einer Seite über eine Weiterleitung auf eine andere Seite zu gelangen, soll nicht durch zwei Schritte bzw. Kanten abgebildet werden, sondern lediglich durch eine Kante.

Listing 11 zeigt die Ausgabe des Tools. Es wurden wie bereits angeführt über 5 Millionen Knoten entfernt, das entspricht einer Reduktion von ca. 57%.

```
...
Removed 5126475 redirecting nodes...
Graph without redirects has:
     3826901 Nodes / 232499481 Edges
Nodes -57.26%
Edges -4.19%
...
```

Listing 11: Ausgabe des Tools zur Entfernung der Weiterleitungen

6.5.2 Größten Teilgraphen extrahieren

Da eine Navigation zwischen einem Knotenpaar nur Sinn ergibt, wenn diese auch im Graphen über einen Pfad eine Verbindung besitzen, wurde ein weiteres Tool entwickelt. Damit ist es möglich, aus dem Graphen die größte zusammenhängende Komponente zu finden und diese getrennt in eine neue Kantenliste abzuspeichern. Die Eigenschaft, die die Verbindung zwischen beliebigen gewählten Knotenpaaren beschreibt, nennt man, wie schon in den Grundlagen erwähnt, schwachen bzw. starken Zusammenhang.

Jene größte schwach zusammenhängende Komponente des Graphen wurde durch dieses Tool zur Weiterverarbeitung extrahiert. Das Ergebnis war sehr eindeutig, da mit 99,99% nahezu der gesamte Graph darin enthalten war. Tabelle 3 zeigt die drei größten schwach zusammenhängenden Komponenten und die drei größten stark zusammenhängenden Komponenten im Vergleich. Auch die größte Komponente beim starken Zusammenhang mit einem Anteil von ca. 94,1% deutet auf den hohen Vernetzungsgrad des Graphen hin.

Tabelle 4 zeigt den Zusammenhang des Graphen vor und nach der Auflösung von Weiterleitungen aus Abschnitt 6.5.1. Während bei dem schwachen Zusammenhang keine Veränderung erkenntlich ist, ist der Unterschied bei der größten stark zusammenhängenden Komponente deutlich ersichtlich. Der Grund liegt bei den Weiterleitungsknoten, die meistens mehrere eingehende Kanten besitzen, jedoch nur eine ausgehende. Der Knoten, auf den weitergeleitet wird, besitzt im Regelfall keinen Link in entgegengesetzter Richtung, d.h.

zurück zum Weiterleitungsknoten. Diese Erkenntnis spiegelt sich somit im starken Zusammenhang des Wikipedia-Graphen wider.

Tabelle 3: Schwacher und starker Zusammenhang des Wikipedia-Graphen

Zusammenhang	Komponente	Knotenanzahl	Prozent
schwacher Zusammenhang	1 von 202	~ 3 826 000	~ 99,99 %
schwacher Zusammenhang	2 von 202	3	< 0,001 %
schwacher Zusammenhang	3 von 202	2	< 0,001 %
starker Zusammenhang	1 von 220 016	~ 3 600 000	~ 94,09 %
starker Zusammenhang	2 von 220 016	61	~ 0,002 %
starker Zusammenhang	3 von 220 016	53	~ 0,001 %

Tabelle 4: Größte Komponente vor bzw. nach der Weiterleitungsauflösung

Zusammenhang	Größe vor Auflösung	Größe nach Auflösung
schwach	~ 99,99 %	~ 99,99 %
stark	~ 62,72 %	~ 94,09 %

6.6 Erzeugung der Hierarchien

Einen weiteren Teil der Datenaufbereitung stellen die Hierarchien dar. Sie modellieren das Hintergrundwissen und können auf unterschiedliche Art und Weise aufgebaut werden, entsprechen jedoch immer den Eigenschaften eines Baumes. Durch die Struktur der Hierarchie wird vorgegeben, wie und an welcher Stelle die Navigation beeinflusst wird. Die Entscheidung, welche Knoten in einer Navigationssimulation, wie von Geigl in [4] beschrieben, gewählt werden, hängt somit von den berechneten Distanzen in der Hierarchie, d.h. den Pfadlängen zwischen den betrachteten Knoten, ab.

Nachdem ursprünglich davon ausgegangen wurde, dass Hierarchien von anderen Arbeitsgruppen zur Verfügung gestellt werden, musste die Erzeugung schließlich doch im Rahmen dieser Datenaufbereitung stattfinden. Grundsätzlich wird daher in dieser Arbeit zwischen zwei Hierarchietypen unterschieden: Erzeugung durch die Breitensuche und Erstellung basierend auf hierarchischen Werten. Diese können jedoch nicht die optimalen Ergebnisse bieten, da hierbei

weitere Analysen und Forschungsergebnisse betrachtet werden müssten. Die in den nächsten drei Abschnitten beschriebenen Hierarchien bieten die Basis, damit der Start des Frameworks MUN erfolgen kann und Evaluierungen durchgeführt werden können.

6.6.1 Hierarchieerstellung mittels Breitensuche

Das Tool zur Erstellung einer Hierarchie anhand der Breitensuche benötigt einen Graphen und einen Wurzelknoten als Eingabe. Ausgehend von der angegeben Wurzel wird anschließend eine Breitensuche durchgeführt. Dadurch werden ausgehend von diesen Knoten alle weiteren Links aufgerufen und danach in die nächste Ebene der Hierarchie eingefügt. Dasselbe Vorgehen erfolgt iterativ auf allen tieferen Ebenen, bis alle Knoten, die von der Wurzel aus erreichbar sind, enthalten sind.

In [4] analysiert Geigl den Wikipedia-Graphen auf jene Knoten mit den größten Graden, den größten Eingangs- bzw. Ausgangsgraden sowie den größten hierarchischen Werten. Die Ergebnisse sind in Tabelle 5 zusammengefasst. Alle darin vorkommenden Seiten-IDs wurden herangezogen, um eine Hierarchie mit diesem Tool zu erstellen. Es ist ersichtlich, dass diverse IDs in den vier verschiedenen Kategorien mehrmals aufscheinen. Vor allem der Artikel „United_States“, der bereits bei der Beschreibung eines Spieltyps des Wikigames in Abschnitt 2.4.2 erwähnt wurde, befindet sich an zweithöchster Stelle bei den Eingangsgraden.

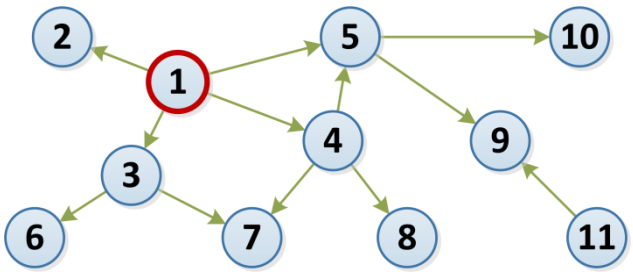

Abbildung 19: Beispiel eines Graphen

Ein einfaches Beispiel mit einer Hierarchie basierend auf der Breitensuche wurde bereits in Abschnitt 4.2.2 gezeigt. Abbildung 16 zeigt die Visualisierung der Baumstruktur des Graphen aus Abbildung 19 mit Knoten 1 als Wurzel.

Tabelle 5: Analyse des Wikipedia-Graphen auf Knoten-Eigenschaften

Typ	Rang	Grad/Wert	Seiten-ID	Seiten-Titel
Grad	1	671 260	48361	Geographic_coordinate_system
	2	520 327	3434750	United_States
	3	288 764	14919	International_Standard_Book_Number
Eingangsgrad	1	671 160	48361	Geographic_coordinate_system
	2	518 668	3434750	United_States
	3	288 384	14919	International_Standard_Book_Number
Ausgangsgrad	1	8 103	24669262	List_of_Italian_communes_(2009)
	2	5 515	21824714	List_of_municipalities_of_Brazil
	3	5 251	274621	Index_of_India-related_articles
hierarchischer Wert	1	5 498 435	48361	Geographic_coordinate_system
	2	2 350 949	16130497	Population_without_double_counting
	3	1 900 303,5	351656	Geographic_Names_Information_System

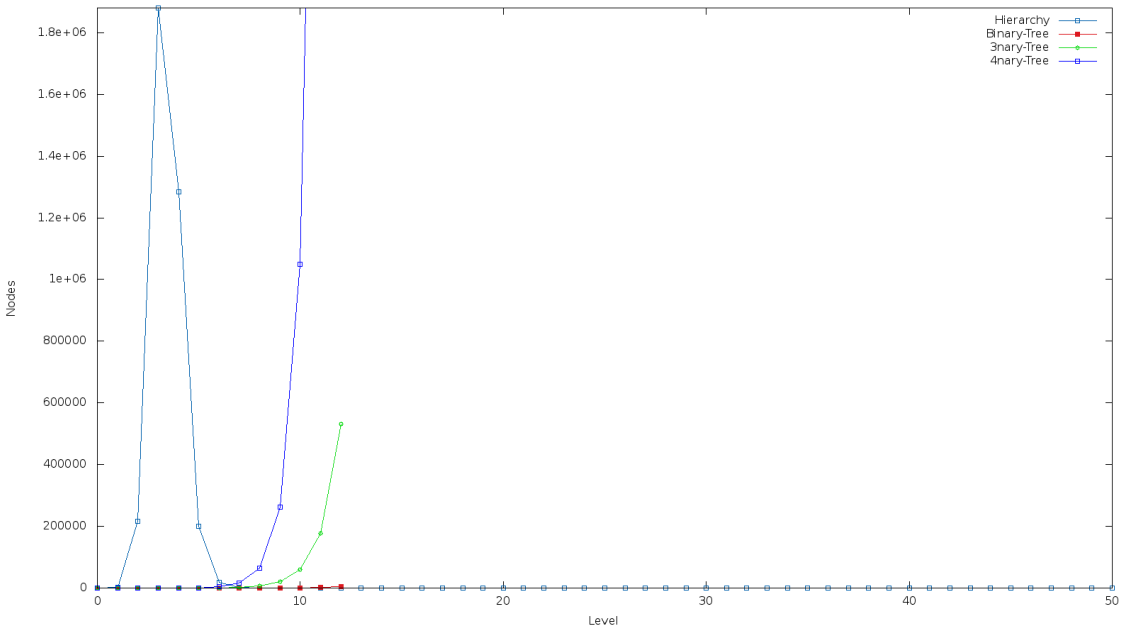

Abbildung 20: Anzahl der Knoten auf den Ebenen (Wurzel: "United_States")

Eine Analyse der Hierarchie für den Artikel „United_States“ als Wurzelknoten ist in Abbildung 20 ersichtlich. Entlang der X-Achse des Plots sind die einzelnen Ebenen des Baumes dargestellt und auf der Y-Achse die jeweilige Anzahl der Knoten. Es geht hervor, dass auf Ebene 1 nach dem Wurzelknoten bereits 1845

Knoten existieren, 215533 auf Ebene 2 und bereits über 1,8 Millionen auf Ebene 3, bevor die Anzahl danach wieder stark abnimmt. Ab Ebene 10 beträgt die Knotenanzahl pro Ebene nur noch weniger als 100. Im Vergleich sind in der Visualisierung neben dem Binärbaum auch 3-näre und 4-näre eingezeichnet.

Diese Methode zur Erstellung einer Hierarchie wurde auf alle in Tabelle 5 enthaltenen Artikelseiten, d.h. aus den vier Typen jeweils die besten drei, angewandt und ausgeführt. Daraus sind acht Hierarchien mit unterschiedlichen Wurzelknoten entstanden.

6.6.2 Hierarchieerstellung mittels hierarchischen Werten

Eine weitere Möglichkeit zur Erstellung einer Hierarchie mit hierarchischen Werten bietet ein anderes Tool. Diesem werden als Eingabedaten ein Graph und die Anzahl der jeweils möglichen Nachfolger pro Knoten übergeben.

Es wird anfangs von allen Knoten im Graphen ein sogenannter hierarchischer Wert, auch als „hierarchical score“ oder HS bezeichnet, berechnet, der auf dem Eingangsgrad d_{in} und dem Ausgangsgrad d_{out} basiert. Die Berechnung erfolgt durch $\frac{d_{in}}{d_{out}}\sqrt{d_{in}}$. Anschließend werden die hierarchischen Werte sortiert und jener Knoten mit dem größten Wert als Wurzel für die Hierarchie gewählt. Der n-äre Baum wird anschließend Ebene für Ebene mit Knoten aus der geordneten Liste aufgefüllt, wobei jeder Knoten nur mit n Nachfolgern besetzt wird.

```
5    3
5    4
3    1
3    2
4    6
4    7
1    8
1    9
2    10
2    11
```

Listing 12: Kantenliste einer HS-Hierarchie mit zwei Nachfolgern je Knoten

In Listing 12 wird ein Beispiel angeführt, bei dem eine Hierarchie durch einen Binärbaum, d.h. jeweils zwei Nachfolgerknoten, aufgebaut wird. Listing 13 zeigt

hingegen eine Hierarchie, die einen 3-nären Baum abbildet. Die graphische Darstellung und Gegenüberstellung der beiden Hierarchievarianten finden sich in Abbildung 21.

```
5     1
5     3
5     4
3     2
3     6
3     7
4     8
4     9
4     10
1     11
```

Listing 13: Kantenliste einer HS-Hierarchie mit drei Nachfolgern je Knoten

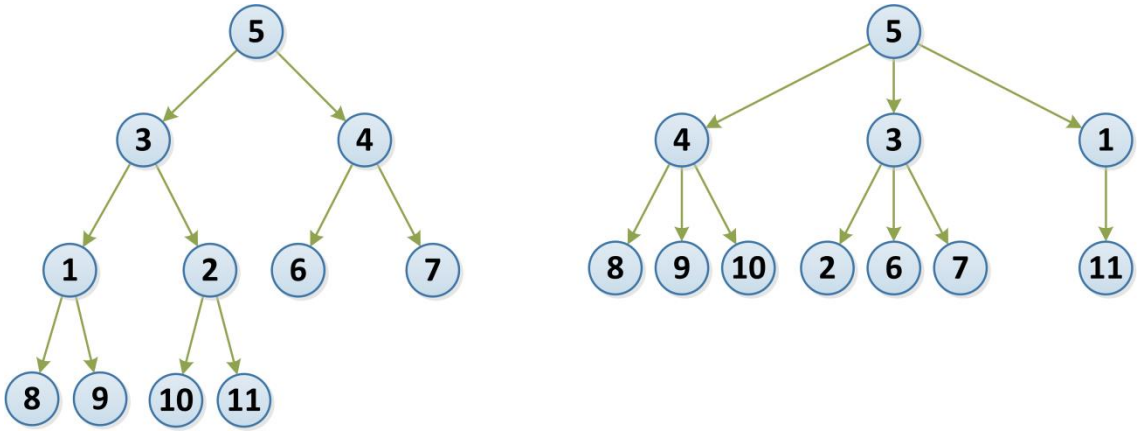

Abbildung 21: Hierarchien mit zwei bzw. drei Nachfolgerknoten

In Tabelle 6 ist ersichtlich, wie schnell die Ebenen von n-ären Bäumen aufgefüllt werden. Während Binärbäume eher langsam anwachsen, steigt bei höheren Werten für n die Knotenanzahl recht schnell an.

Anhand dieser Methode wurden insgesamt sechs Hierarchien aus dem Wikipedia-Graphen erstellt. Diese beinhalten eine Struktur basierend auf n-ären Bäumen mit Werten von 2, 3, 4, 5, 6 und 10 für n.

Tabelle 6: Max. Knoten-Anzahl auf den jeweiligen Ebenen n-ärer Bäume

Ebene	n = 2	n = 3	n = 4	n = 5	n = 10
0	1	1	1	1	1
1	2	3	4	5	10
2	4	9	16	25	100
3	8	27	64	125	1 000
4	16	81	256	625	10 000
5	32	243	1 024	3 125	100 000
6	64	729	4 096	15 625	1 000 000
7	128	2 187	16 384	78 125	10 000 000

6.6.3 Hierarchieerstellung mittels erweiterter Breitensuche

Da die beiden bereits beschriebenen Hierarchie-Ansätze nicht zum gewünschten Erfolg bei der Evaluierung der Ergebnisse von Geigl in [4] geführt haben, wurde gemeinsam an einer weiteren Methode zur Erstellung einer Hierarchie gearbeitet.

Bei der Verwendung von Hierarchien, die durch einfache Breitensuche erstellt wurden, zeigte sich bei den Versuchen, dass es besonders schwierig ist, Entscheidungen zu treffen, wenn auf einer bestimmten Ebene sehr viele Knoten existieren (vgl. Abbildung 20). Dem Tool, das Hierarchien durch die abgeänderte erweiterte Breitensuche erstellt, kann als Parameter die Anzahl der Nachfolgerknoten sowie der Wurzelknoten mitgegeben werden. Es werden dadurch pro Knoten nur n Kanten zu jenen weiteren Knoten eingefügt, die die n größten Eingangsgrade besitzen, sodass ein n-ärer Baum entsteht.

Abbildung 22 zeigt ein Beispiel, das auf dem Graphen aus Abbildung 19 basiert. Die Hierarchie wurde ausgehend von Knoten 1 erstellt. Es wurden z.B. nur die Knoten 3 und 5 als Nachfolgerknoten eingefügt, da sie im Graphen die größten Eingangsgrade besitzen. Die beiden Knoten 2 und 4 wurden vernachlässigt und nicht in die Hierarchie aufgenommen. Da die Knoten 2, 3 und 4 in diesem Fall denselben Eingangsgrad besitzen, wurde aus diesen ein zufälliger Knoten, in diesem Beispiel Knoten 3, eingefügt. Es sind in jenen Hierarchien dadurch

weniger Knoten vorhanden, jedoch existieren alle enthaltenen Kanten in der Hierarchie ebenfalls im zugehörigen Graphen.

In weiterer Folge wird eine Hierarchie, die durch die beschriebene erweiterte Breitensuche mit N Nachfolgern aufgebaut ist, als TopN-BFS[33]-Hierarchie bezeichnet. Eine einfache Hierarchie basierend auf Breitensuche, wie in Kapitel 6.6.1 erwähnt, wird BFS-Hierarchie genannt.

Zur Auswahl der Parameter für diese Methode der Hierarchie-Erzeugung wurde auf Ergebnisse der Erstellung mit einfacher Breitensuche zurückgegriffen. Daher wurden als Wurzelknoten die Artikel „Geographic_coordinate_system“ und „United_States“ gewählt und jeweils die Berechnungen mit fünf (Top5-BFS) bzw. zehn (Top10-BFS) Nachfolgerknoten durchgeführt. Daraus sind insgesamt vier Hierarchien entstanden. Geigl zeigt in [4], dass mit Hierarchien, die auf fünf oder zehn Nachfolgerknoten begrenzt sind, gute Ergebnisse erzielt werden.

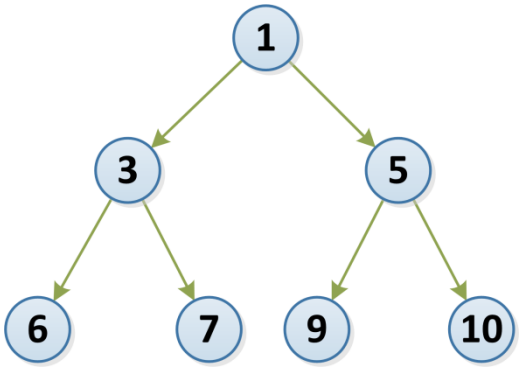

Abbildung 22: Erstellte Hierarchie durch erweiterte Breitensuche

6.7 Extrahierung der Navigationspfade

Die durch Spiele des Wikigames erzeugten Navigations- bzw. Klickpfade mussten anhand des bereitgestellten SQL-Dumps extrahiert werden. Dazu wurde ein Skript in PHP entwickelt, das die als einzelne Klicks in der Datenbank gespeicherten Datensätze sortiert und ausgibt.

Durch nähere Analysen wurde festgestellt, dass es zu zwei Inkonsistenzen in der Datenspeicherung gekommen ist. Da sehr viele gespeicherte Missionen,

[33] Breadth-first search, Breitensuche

d.h. Spiele von einzelnen Benutzern, nicht mit der korrekten Startseite als ersten Klick begannen, wurde der Klickverlauf der einzelnen Benutzer über mehrere Spiele hinweg näher betrachtet. Dabei fiel auf, dass nach dem Ablauf des Zeitlimits eines Spieles häufig der letzte Klick bereits als erster Klick im neuen darauffolgenden Spiel gewertet wird. Diese in keinem Zusammenhang stehenden Klicks wurden erkannt und aus den Klickpfaden entfernt.

Des Weiteren wurde vielfach über Seiten, die sich nicht im Namensraum für Artikel in der Wikipedia befinden, navigiert. Diese Klickpfade enden häufig z.B. auf einer Diskussionsseite. Alle Navigationspfade, die Seiten außerhalb des Artikel-Namensraums beinhalten, wurden ebenfalls gefiltert und sind in der finalen Klickpfad-Datei nicht enthalten.

Listing 14 zeigt ein Beispiel für Klicks außerhalb des Artikel-Namensraums. Dies sind jene Klicks, die – statt einer gültigen ID – „-1“ beinhalten. Das Format dieser Datei wurde bereits in Abschnitt 4.3.1 erklärt.

```
...
176724 6598 176724 9228 12305127  -1  -1  -1
176724 6598 176724 9228 -1  -1  -1 49564
176724 6598 176724
176724 6598 176724 9228 592645 7819 82804 18838 6598
...
```

Listing 14: Beispiel einer ungefilterten Klickpfad-Datei

Von den ursprünglich 2306024 extrahierten Navigationspfaden ohne Filterung konnten nach der Namensraumbegrenzung noch 2207571 gültige Pfade zur Weiterverarbeitung behalten werden.

6.8 Bereinigung der Navigationspfade

Neben den grundsätzlichen Einschränkungen beim Export der Klickpfade mussten weitere Maßnahmen getroffen werden, um aussagekräftige Navigationspfade zu erhalten. Um dieses Vorhaben zu ermöglichen, wurde ein weiteres Tool erstellt, das eine Bereinigung durchführt.

Das entwickelte Programm nimmt den Wikipedia-Graphen, eine Klickpfad-Datei, eine Weiterleitungsliste und weitere Parameter entgegen, um die nachfolgenden Schritte zu durchlaufen.

Am Beginn mussten die Weiterleitungen in den Navigationspfaden korrekt aufgelöst werden. Folgendes Navigationsverhalten wurde analysiert: Wenn ein Spieler über einen Link von Seite A auf Seite B folgt, die aber automatisch auf Seite C weiterleitet, so sind im Klickpfad nur die Knoten A und B enthalten, nicht jedoch C. Wie schon in den Abschnitten zuvor beschrieben, existieren die Weiterleitungsknoten im Wikipedia-Graphen nicht mehr. Daher wurden in diesem Projekt auch alle Weiterleitungen in den Klickpfaden angepasst. Im zuvor erwähnten Beispiel wäre dann Seite B nicht mehr im Pfad enthalten, sondern nur Knoten A und C.

Da die Klickpfad-Daten über einen längeren Zeitraum hinweg betrachtet werden, änderte sich in dieser Zeit auch die dem Wikigame zugrunde liegende Wikipedia-Datenbank. Daher wurden jene Klickpfade entfernt, die Knoten enthalten, die in dem verwendeten Wikipedia-Graphen vom 15. November 2011 nicht enthalten sind.

In weiterer Folge wurden die Klickpfade auch dahingehend untersucht, welche Strategien von den Spielern benutzt wurden, um von einer Seite zu einer anderen zu navigieren. Sollte es keine Möglichkeit geben, über Links von Seite A bzw. von zuvor besuchten Seiten zu Seite B zu gelangen, so spricht man von einer Teleportation von A zu B. Da weder Suche noch sonstige Möglichkeiten im Wikigame bestehen, die solche ermöglichen, wurde erneut eine Analyse durchgeführt.

Die Analyse hat gezeigt, dass diese Navigationen auf zwei Gründe zurückzuführen sind:

- Der Link hat vor einiger Zeit existiert, wurde aber gelöscht.
- Das Spiel wurde manipuliert.

Es wurde eine Liste mit allen Teleportationen erstellt und jeweils die Häufigkeit des Vorkommens in allen Pfaden errechnet. Nach einer stichprobenartigen

Untersuchung wurde festgelegt, dass ein Schwellwert definiert werden soll. Befindet sich eine Teleportation in mehr als zehn Klickpfaden, so kann davon ausgegangen werden, dass dieser Link zum Spielzeitpunkt existierte und bis zum Erstellen des Wikipedia-Graphen gelöscht wurde. All jene Kanten, die diese fehlenden Links repräsentieren, wurden in den Graphen, d.h. in die Kantenliste, nachträglich eingefügt. Alle beschriebenen Schritte von Abschnitt 6.5 bis 6.8 mussten daher erneut ausgeführt werden, da der Graph dadurch erweitert wurde.

Nun sollte noch kurz Bezug auf jene Teleportationen genommen werden, die unter dem Schwellwert liegen. Häufig sind darin „Links“ enthalten, die in der gesamten Menge an Klickpfaden nur ein einziges Mal vorkommen. Aus diesem Grund wurde das Online-Spiel Wikigame auf Manipulierbarkeit untersucht. Es wurde festgestellt, dass durch einen kleinen Hack ungültige Navigationspfade entstehen können.

Eine kurze Erklärung zum Cheaten im Spiel: Die Seite, in der der Benutzer seine vorgegebene Mission spielen kann, wird in einem Inline-Frame geladen, d.h. eine eigene unabhängige Webseite, die als Unterdokument auf der Hauptseite geladen wird. Diese Website besitzt eine eigene URL[34] und kann im Browser über das Kontextmenü in einem neuen Fenster geöffnet werden. Ein Beispiel dieser URL ist in Listing 15 ersichtlich. Die aktuelle Seite ist rot markiert.

```
http://www.thewikigame.com/wiki/American_Hockey_
League?game_type=speed-race&game_uuid=
cdcdd69ab37b434aa05af2917d1b684f&ts=1359108722133
```

Listing 15: Direkte URL einer Seite im Wikigame

Durch eine einfache Änderung in der Browser-Adressleiste auf die Zielseite wird ein ungültiger Link von einer beliebigen Seite zu einer anderen, in diesem Fall die Zielseite, vorgetäuscht. Nach dem Aufruf muss nur noch ein Klick zu einer beliebigen Seite im Spiel erfolgen, und schon erreicht man jede Seite in nur

[34] Uniform Resource Locator – identifiziert und beschreibt eine Ressource im Internet, z.B. eine Webseite durch eine Webadresse

einem Klick. Abbildung 23 zeigt einen Screenshot der Website nach einem Sieg durch einen ungültigen Klickpfad, der durch Manipulation entstanden ist. Ein Link vom Artikel „Pseudoephedrine“ zu „Internal Revenue Code“ existiert nicht.

Abbildung 23: Ungültiger Klickpfad im Wikigame

Im Zuge der Analyse von Wikigame im Rahmen dieses Projekts wurde öfter auch eine Mission im Spiel korrekt gespielt. Nach jedem Spiel sind die Klickpfade für alle anderen Teilnehmer ersichtlich. Auch hier waren bei Klickpfaden anderer Spieler teilweise bewusste Manipulationen erkenntlich. Somit wurde angenommen, dass jene Pfade mit Teleportationen nicht der Benutzernavigation entsprechen und wurden vollständig aus der Pfaddatei gelöscht.

Tabelle 7 zeigt die Navigationspfade nach der Abfolge an Modifikationen. Von insgesamt rund 1,9 Millionen verbleibenden Klickpfaden sind 41,4% erfolgreiche und 58,6% erfolglose Pfade.

Tabelle 7: Bereinigte Klickpfade aus dem Wikigame

	Anzahl Pfade	Relative Angabe
erfolgreiche Pfade	786 845	~ 41,4 %
nicht erfolgreiche Pfade	1 111 556	~ 58,6 %
Pfade gesamt	1 898 401	100,0 %

6.9 Aktualisierung des Graphen

Durch die Erkenntnisse aus dem vorhergehenden Abschnitt wurden zum bestehenden Graphen über 14.000 Kanten hinzugefügt. Der endgültige Wikipedia-Graph, der für alle Berechnungen herangezogen wird, beinhaltet somit 3.826.689 Knoten und 232.513.920 Kanten.

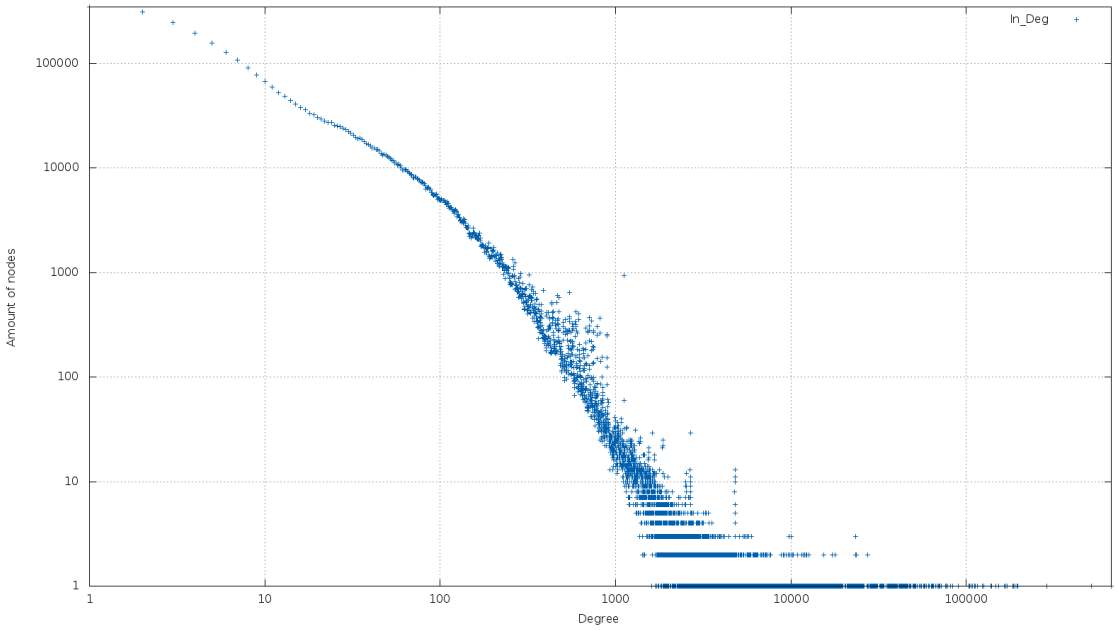

Abbildung 24: Verteilung der Knoten-Eingangsgrade im Wikipedia-Graphen

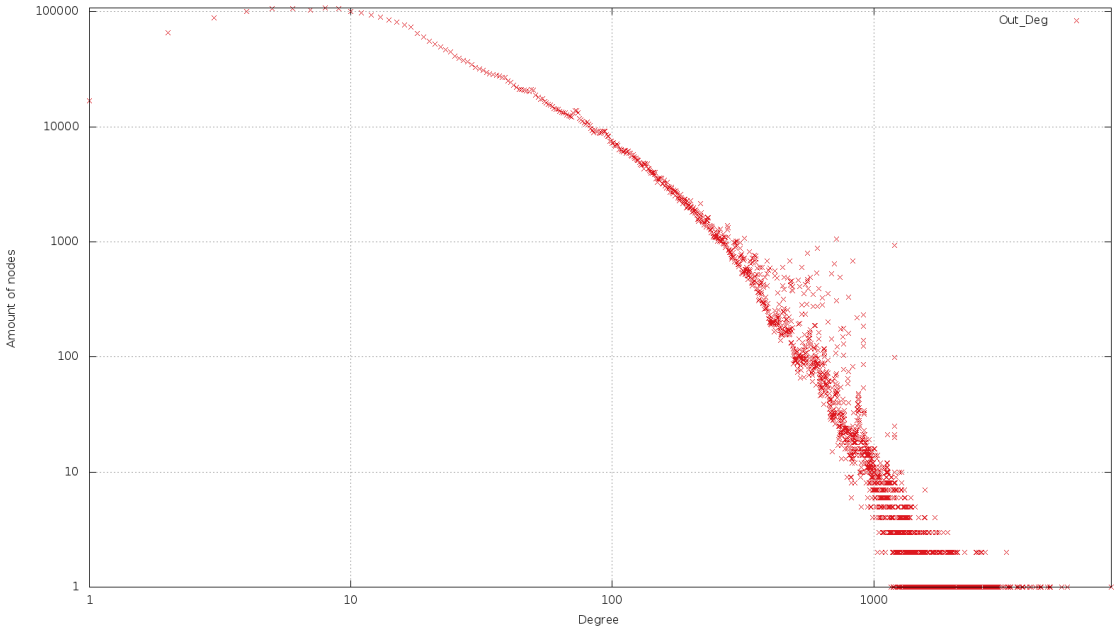

Abbildung 25: Verteilung der Knoten-Ausgangsgrade im Wikipedia-Graphen

Eine Analyse der Knoten-Eingangsgrade des Graphen in Abbildung 24 zeigt, dass es viele Knoten gibt, die wenige eingehende Links besitzen. Im Gegensatz dazu gibt es einzelne Seiten, die über eine Million Eingangslinks besitzen. Abbildung 25 stellt die Anzahl der Ausgangslinks pro Knoten dar. Die meisten Artikel enthalten knapp unter 10 ausgehende Links, jedoch gibt es auch mehrere mit über 1000.

6.10 Erzeugung weiterer benötigter Eingabedaten

Neben allen bisher beschriebenen Eingabedaten werden noch zwei weitere benötigt, die anhand von Tools erzeugt werden können.

6.10.1 Generierung von Paaren

Da der Navigator Knotenpaare, d.h. Kombinationen aus Start- und Zielknoten, für die Navigation benötigt, stellt ein Tool diese Funktionalität zur Verfügung. Es nimmt einen Graphen und eine Klickpfad-Datei entgegen und erstellt eine Ausgabedatei, die – wie bereits in Abschnitt 4.3.2 erwähnt – Paare enthält.

Dadurch kann aus den Klickpfaden des Wikigames jene Datei mit Paaren erzeugt werden, damit der Navigator genau dieselbe Ausgangssituation vorfindet.

6.10.2 Berechnung von kürzesten Distanzen

Die Erstellung der kürzesten Distanzen nimmt aufgrund der komplexen Berechnung viel Zeit in Anspruch. Aus diesem Grund wurde auch hier auf ein Tool gesetzt, das die kürzesten Distanzen in einem Graphen berechnet und ausgibt.

Um die gewünschten Distanzen zu erhalten, wird eine Breitensuche vom jeweiligen Startknoten durchgeführt, bis alle benötigten Zielknoten und somit auch die jeweiligen kürzesten Distanzen gefunden wurden.

Das Tool bietet mehrere Funktionen an: Es kann beispielsweise eine Klickpfaddatei oder eine Datei mit Knotenpaaren als Paramater übergeben werden, für

die kürzeste Pfade berechnet werden sollen. Die Breitensuche wird anschließend für alle enthaltenen Paare ausgeführt.

Um beim Hinzufügen von weiteren Paaren nicht alle Berechnungen erneut ausführen zu müssen, kann auch eine bereits bestehende Datei mit kürzeste Distanzen, wie in Abschnitt 4.4 dargestellt, als zusätzlicher Parameter definiert werden. Nach der Analyse werden nur die Änderungen neu berechnet und zur Datei mit den kürzesten Distanzen hinzugefügt.

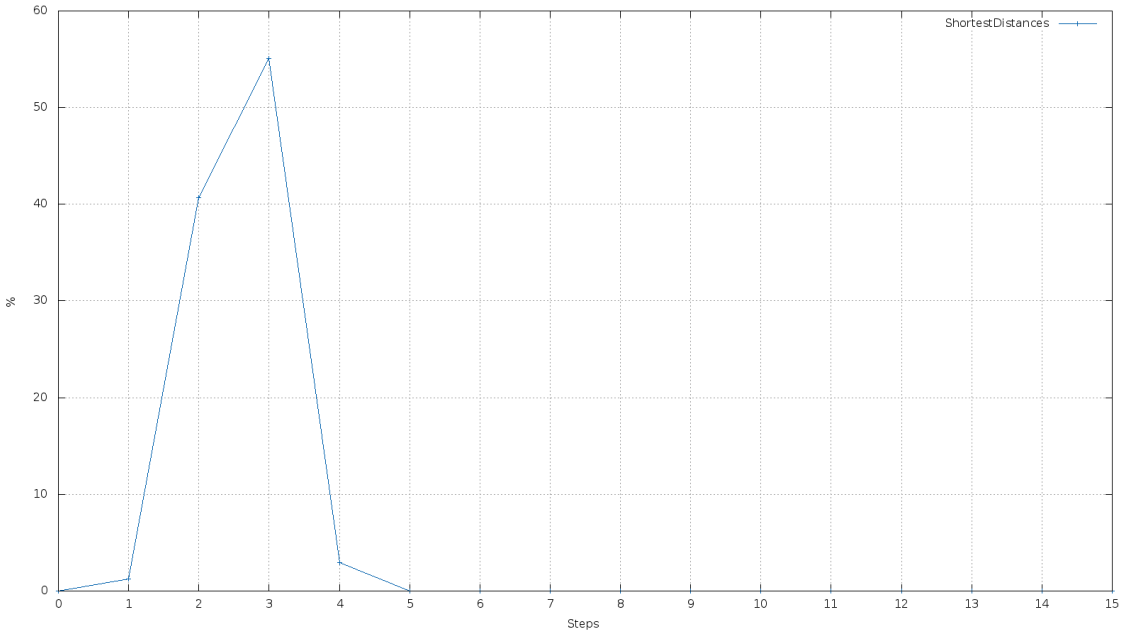

Abbildung 26: Kürzeste Distanzen im Graphen

Tabelle 8: Kürzeste Distanzen im Graphen

Länge	Anzahl
1	5089
2	164420
3	222869
4	11903
5	92

Abbildung 26 visualisiert die Verteilung der kürzesten Distanzen der Knotenpaare aus den Wikigame-Klickpfaden. Die exakten Werte sind in Tabelle 8 ersichtlich. Es ist daraus ablesbar, dass aus 404373 eindeutigen Spielen kein einziges

einen kürzesten Pfad besitzt, der länger als fünf ist. Nach [25] ist die durchschnittliche Länge $L \sim log(N)$ bei N Knoten in einem Kleine-Welt-Netzwerk. Mit einem tatsächlichen Durchschnitt der berechneten Paare von rund 2,6 liegt das vorhandene Wikipedia-Netzwerk deutlich unter den berechenbaren 6,6.

6.11 Konfiguration

Abschließend wird in diesem Abschnitt noch eine Übersicht über die Konfigurationsmöglichkeiten des Frameworks MUN in Bezug auf die erzeugten Daten gegeben.

Parameter zur Beschreibung des Graphen:

- Datei, die den Graphen enthält
  graph.file:/path/to/graph
- Graph ist gerichtet („d“ - directed) oder ungerichtet („u“ - undirected)
  graph.type:d

Parameter zur Beschreibung der Hierarchie:

- Datei, die die Hierarchie enthält
  hier.file:/path/to/hierarchy
- Hierarchie ist gerichtet („d“ - directed) oder ungerichtet („u“ - undirected)
  hier.type:d

Parameter zur Beschreibung der Knotenpaare:

- Datei, die die Knotenpaare enthält (optional)
  pairs.inputfile:/path/to/pairs

Parameter zur Beschreibung der Vergleichspfade bzw. Klickpfade:

- Datei, die die Klickpfade enthält (optional)
  paths.comparewith:/path/to/clickpaths

Parameter zur Beschreibung der kürzesten Distanzen:

- Datei, die die kürzesten Distanzen enthält (optional)
  shortestd.inputfile:/path/to/shortestdistances

Die angeführten Konfigurationsparameter beschreiben die Eingangsdaten, eine ausführliche Dokumentation aller weiteren Parameter beschreibt Eder in [3].

7. Abbruchsimulator / Attrition Rate

Dieses Kapitel beschäftigt sich mit einer Erweiterung des Frameworks MUN, einem Abbruchsimulator. Es wird auf die Implementierung, die Umsetzung und die Evaluierung der Ergebnisse eingegangen.

7.1 Motivation

Die Grundidee des Frameworks MUN ist es, anhand von zahlreichen Parametern eine Navigationssimulation durchführen zu können. Diese soll Möglichkeiten bieten, sich an das Verhalten von Benutzern anzupassen, d.h. die Navigation von Benutzern zu modellieren. Um die Eigenschaften nachzubilden, wurde von Geigl in [4] u.a. ein Link-Restrictor, der basierend auf diversen Merkmalen eine Einschränkung einer Linkauswahl trifft, entwickelt. Neben dieser Erweiterung erschien es ebenso sinnvoll, das Abbruchverhalten von Benutzern abzubilden.

Um dieses Vorhaben zu verdeutlichen, zeigt Abbildung 27 das Benutzerverhalten im Wikigame auf. In der Grafik sind die Pfadlängen der einzelnen Klickpfade dargestellt. Blau angezeigt werden dabei die erfolgreichen Pfade. Dabei ist zu erkennen, dass rund 1,1% dieser Pfade mit einem Klick zum Erfolg führen, die meisten Klickpfade (ca. 20,6%) jedoch vier Klicks benötigen. Im weiteren Verlauf nehmen die erfolgreichen Pfade stark ab, sodass sie ab einer Länge von 12 bereits unter ein Prozent abfallen. Die rot dargestellten erfolglosen Navigationspfade besitzen ihr Maximum, jeweils rund 15,2%, bei einer Länge von 3 bzw. 4. Davor ist ein leichter Anstieg erkennbar, ein Gefälle – ähnlich den erfolgreichen – ist danach auch hier ersichtlich. Grün dargestellt ist schließlich noch die Summe beider Kategorien. Des Weiteren ist daraus die Abbruchrate der Benutzer ersichtlich, die annährend einer quadratischen Funktion entspricht. Jene Rate gibt an, wie viele der jeweils verbleibenden Navigationspfade im betrachteten Schritt abgebrochen werden.

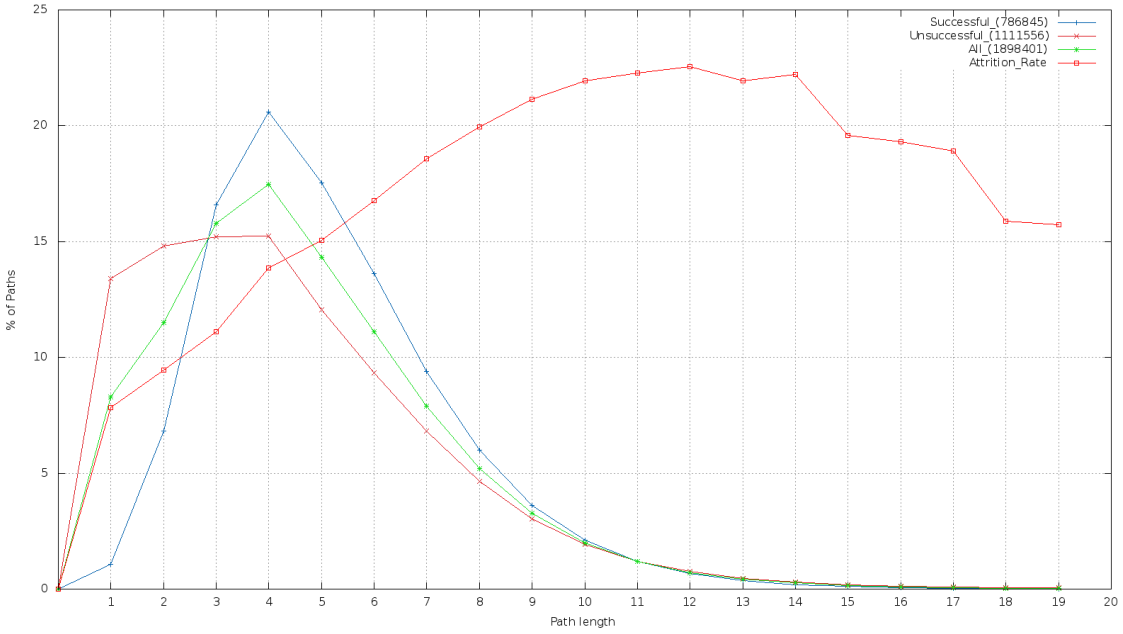

Abbildung 27: Länge der Navigationspfade im Wikigame

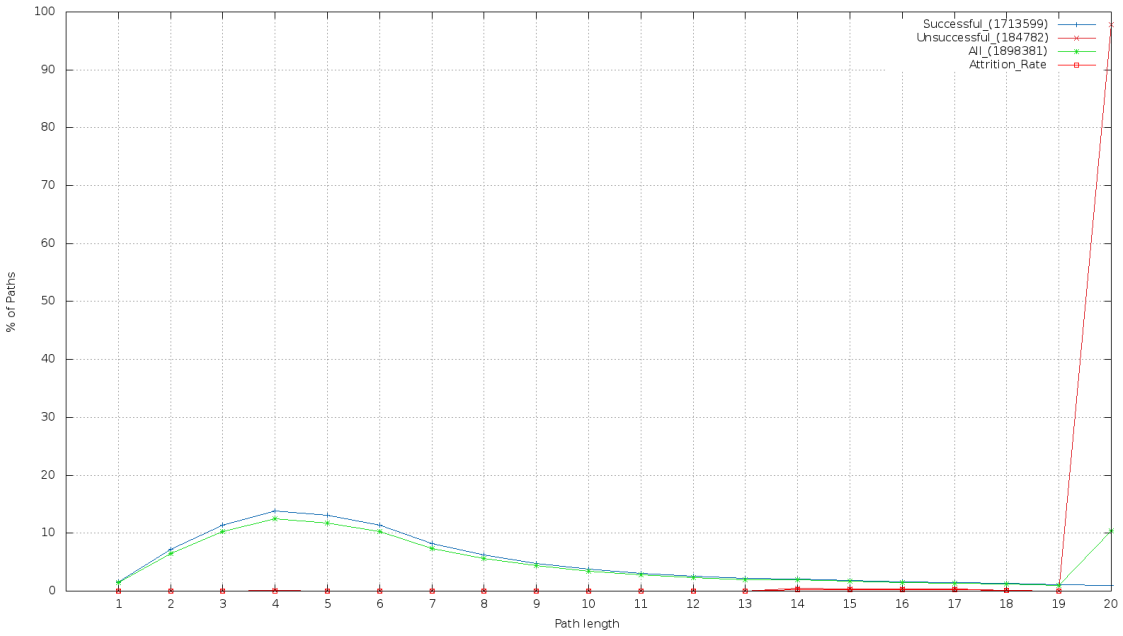

Abbildung 28: Länge von generierten Navigationspfaden

Abbildung 28 stellt Pfadlängen dar, die vom Framework MUN durch eine Navigation des Greedy-Algorithmus mit der Top5-BFS-Hierarchie „United_States“ erstellt wurden. Dieser sogenannte Greedy-Node-Selector wurde von Geigl in [4] implementiert und wählt zu jedem Zeitpunkt der Navigationssimulation jenen Knoten aus, der den größten Erfolg verspricht, d.h. die geringste Distanz zum Zielknoten besitzt.

In jener Abbildung ist auch zu erkennen, dass nahezu alle erfolglosen Navigationspfade, ca. 98 %, bei 20 Schritten enden. Dies liegt daran, dass in der Simulation dieser Wert als Grenze gesetzt wurde, bei der jede Navigation abbricht.

Motiviert durch diese Gegebenheit ist das Ziel, eine Abbruchsimulation zu modellieren, die den menschlichen Klickpfaden aus dem Wikigame nahekommt.

7.2 Implementierung

Der Abbruchsimulator implementiert eine Schnittstelle des Frameworks, sodass eine Erweiterung durch zusätzliche Typen jederzeit möglich ist. Die Auswahl des Typs bzw. der Abbruchrate ist durch die Konfigurationsparameter, die in Abschnitt 7.3 erklärt werden, gegeben.

Ist der Abbruchsimulator gesetzt, so wird am Beginn jedes Navigationsschrittes überprüft, ob der aktuelle Pfad abgebrochen werden soll. Die Entscheidung basiert auf einer Methode, die vom jeweiligen Typ implementiert und aufgerufen wird. Der Typ gibt den grundlegenden Verlauf der Abbruchrate vor, die Entscheidung basiert jedoch auf einem Zufallszahlenwert, der als Seed[35] die aktuelle Uhrzeit benutzt. Diese Implementierung gewährleistet dadurch auch geringfügige Abweichungen bei den Ergebnissen. Der interne Rückgabewert zur Bestimmung des Abbruchs ist immer ein eindeutiger Boolscher Wert. Des Weiteren sollte noch erwähnt werden, dass im nullten Schritt kein Abbruch erfolgt.

Da das Framework nicht nur auf die Verwendung für Navigationssimulationen in Wikipedia-Netzwerken ausgelegt ist, wurden vier unterschiedliche Verlaufe des Abbruchs implementiert, um auch eine allgemeine Anwendung zu ermöglichen.

7.2.1 Konstanter Verlauf

Der konstante Verlauf ist der einfachste Typ zur Bestimmung des Abbruchs. Ist diese Version aktiviert, so wird im jedem Schritt ein bestimmter gleichbleibender Anteil aller Pfade abgebrochen. Abbildung 29 zeigt diesen Verlauf mit einem

[35] Initialwert für einen Zufallszahlengenerator

Resistenzwert von 0,9. Das bedeutet in diesem Fall, dass die Rate konstant 90% beträgt und ein Abbruch in jedem Schritt gleich wahrscheinlich ist.

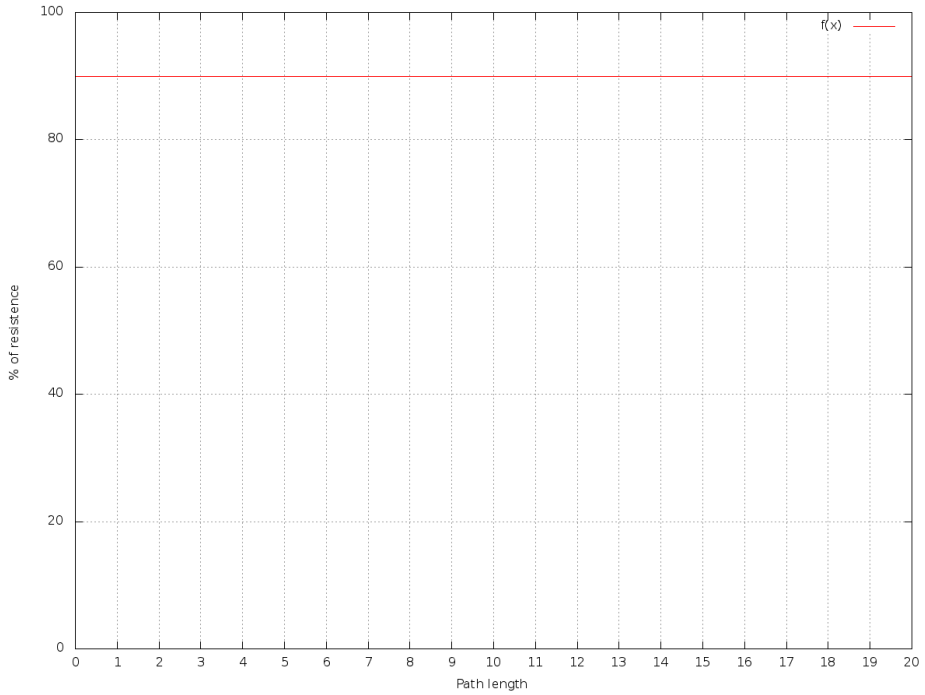

Abbildung 29: Konstante Abbruchrate

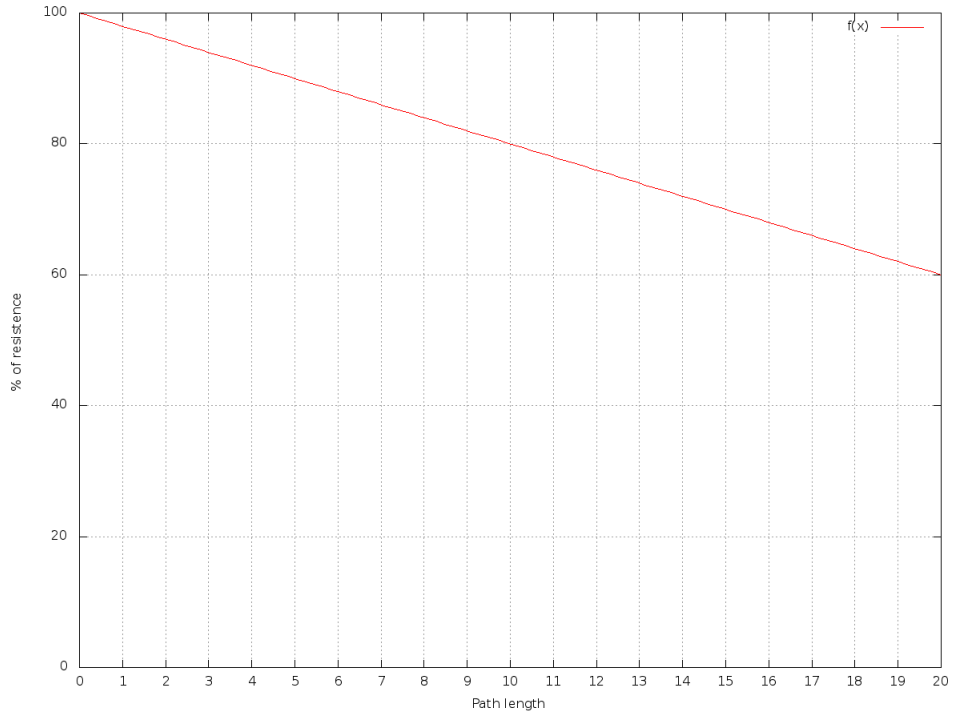

Abbildung 30: Lineare Abbruchrate

7.2.2 Linearer Verlauf

Eine leichte Veränderung der Rate bringt ein linearer Verlauf. Durch die Angabe eines Resistenzwertes kann über den Verlauf der Funktion bestimmt werden. Abbildung 30 zeigt ein Beispiel, bei dem die Rate über die Pfadlänge hinweg linear abnimmt.

7.2.3 Quadratischer Verlauf

Ein weiteres Modell stellt der quadratische Verlauf dar. Die Basis dafür ist eine Funktion zweiten Grades, die die Abbruchrate bestimmt und u.a. über einen Resistenzparameter definiert wird. Aufgrund des Verlaufs besteht neben der quadratischen Funktion an sich noch eine weitere Konfigurationsmöglichkeit. Da die beschreibende Parabel, wie in Abbildung 31 rot dargestellt, wieder auf einen Wert von 100% ansteigt, kann ein Limit gesetzt werden, um auch die Rate danach konstant zu halten. D.h. während z.B. im Intervall 1 bis 10 die Beschreibung durch die Parabel (rot) erfolgt, wird die Rate im Intervall 10 bis unendlich durch die grün dargestellte Funktion konstant modelliert.

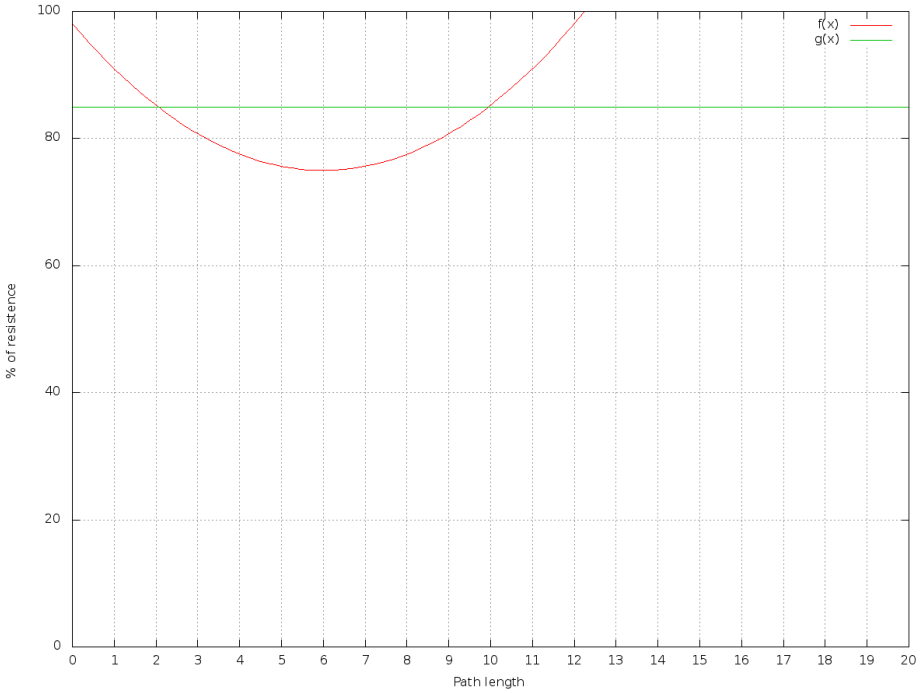

Abbildung 31: Quadratische Abbruchrate

7.2.4 Exponentieller Verlauf

Als vierter Verlauf wurde noch die exponentielle Form implementiert. Durch den Resistenzparameter definiert, stellt sie eine exponentiell abnehmende Funktion dar. Diese Rate kann vor allem verwendet werden, wenn ein schnelles Abbruchverhalten modelliert werden soll. Abbildung 32 stellt ein Beispiel dieser Abbruchrate dar.

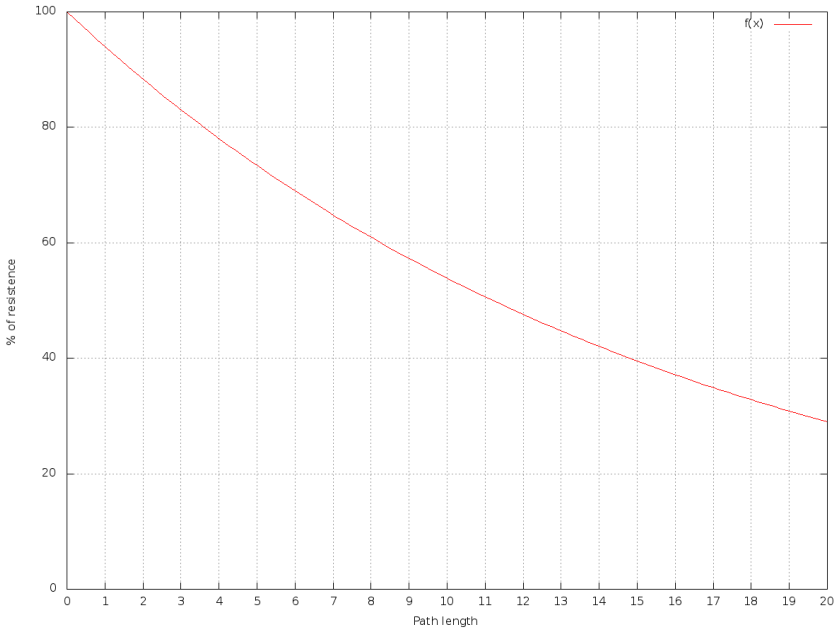

Abbildung 32: Exponentielle Abbruchrate

7.3 Konfiguration

Der Abbruchsimulator kann über mehrere nachfolgend beschriebene Parameter, die entweder in der Konfigurationsdatei des Frameworks MUN oder über die Kommandozeile beim Start definiert werden können, konfiguriert werden.

Um den Simulator zu aktivieren, muss der Typ gesetzt werden. Dies geschieht über den Parameter *attritionsim.type*. Es kann die Auswahl zwischen *constant*, *linear*, *quadratic* oder *exponential* getroffen werden.

Durch die Angabe einer Resistenz im Intervall 0 und 1 über den Parameter *attritionsim.resistance* kann definiert werden, wie wahrscheinlich ein Abbruch erfolgen soll. Ein hoher Wert verspricht folglich hohe Resistenz gegen Abbrü-

che, bei einem niedrigen Wert wird die Wahrscheinlichkeit für einen Abbruch größer. Zur Beschreibung und Veränderung vor allem für den quadratischen Verlauf sind noch drei weitere Parameter verfügbar. Mit *attritionsim.limit* wird ein lineares Limit, wie bereits in Abschnitt 7.2.3 beschrieben, gesetzt. Die quadratische Funktion kann anhand von *attritionsim.shiftx* und *attritionsim.shifty* entlang der beiden Achsen verschoben werden.

Um bei einem Start des Frameworks die Aktivierung des Abbruchsimulators überprüfen zu können, werden vor der Navigationssimulation alle aktiven Parameter und deren Belegungen ausgegeben. Eine Beispiel-Ausgabe ist in Listing 16 ersichtlich.

```
:: Attrition simulation activated
   Type: quadratic
   Resistance: 0.92
   Shift X: 7
   Shift Y: 0.75
   Limit: 0.8
```

Listing 16: Ausgabe bei aktiviertem Abbruchsimulator

7.4 Ergebnisse/Evaluierung

Um nun eine Konfiguration zu finden, die jenen Klickpfaden des Wikigames nahe kommt, wurden vor allem konstante und quadratische Verläufe näher betrachtet. Nachdem anfangs die Simulation mit einer konstanten Abbruchrate durchgeführt wurde, konnte durch nähere Betrachtungen und Analysen festgestellt werden, dass eine quadratische Rate mit einem zusätzlichen konstanten Limit zu den besten Ergebnissen führt.

Durch zahlreiche Versuche und Parameteranpassungen wurde eine Konfiguration gefunden, die eine sehr gute Annäherung an die Benutzerdaten aus dem Wikigame erlaubt. In jener Konfiguration wurde ausschließlich die Navigationsstrategie „Greedy“ eingesetzt, als Hintergrundwissen diente die Top5-BFS-Hierarchie mit dem Wurzelknoten „United_States“. Als Abbruchsimulation wurde die quadratische Rate mit einem Resistenzfaktor von 0,96 und einem linearen Limit von 0,8 gewählt. Die Verschiebung der quadratischen Funktion

entlang der X-Achse wurde auf 12 gesetzt und die Y-Achsenverschiebung betrug 0,75.

Die graphische Darstellung der simulierten Navigationspfadlängen ist in Abbildung 33 ersichtlich. Darin sind sowohl die erfolgreichen als auch die nicht erfolgreichen Pfadlängen eingetragen, deren Verhältnis zueinander jenem aus den Vergleichsdaten sehr nahe kommt. Auch die Abbruchrate lässt sich deutlich als quadratische Form erkennen, die nach dem Maximum bei einer Länge von 12 nicht mehr stark abnimmt, sondern eher konstant bleibt.

Der Erfolg ist vor allem bei der Gegenüberstellung mit den Wikigame-Vergleichspfaden in Abbildung 34 deutlich ersichtlich. Jene Datenpunkte, die durch ein Quadrat visualisiert werden, entsprechen den durch den Navigator generierten Pfaden. Im Diagramm werden diese auch als erste Kollektion („first collection“) bezeichnet. Die Vergleichspfade das Wikigames hingegen werden als zweite Kollektion („second collection“) und durch kreisförmige Datenpunkte dargestellt. In nahezu allen Bereichen zeigen die beiden Kollektionen einen sehr ähnlichen Verlauf.

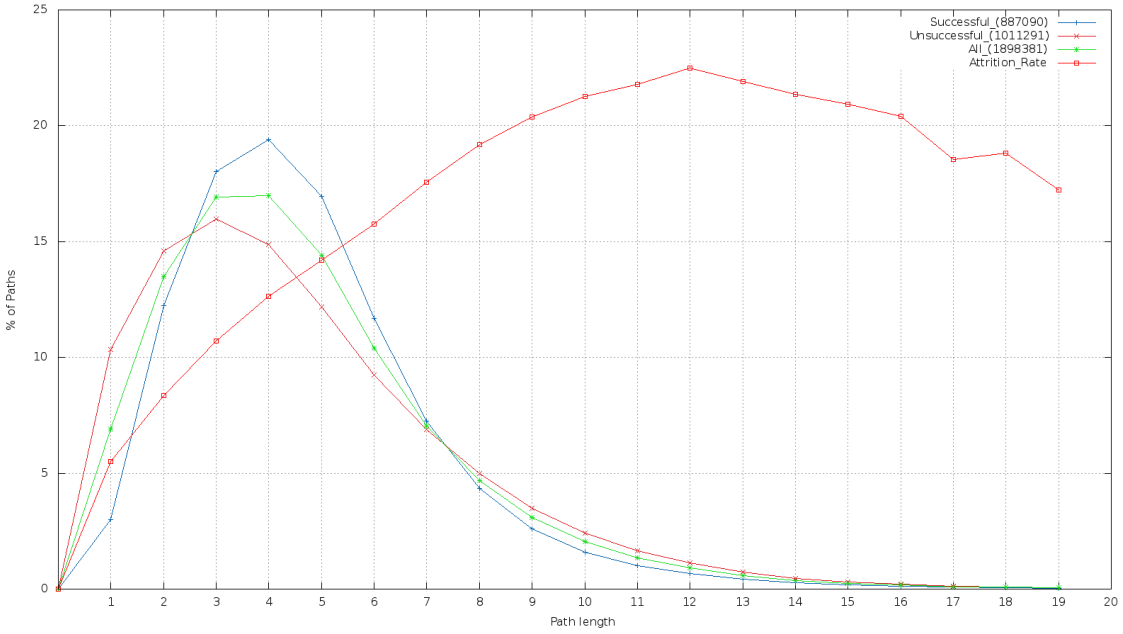

Abbildung 33: Annäherung durch quadratische Abbruchrate

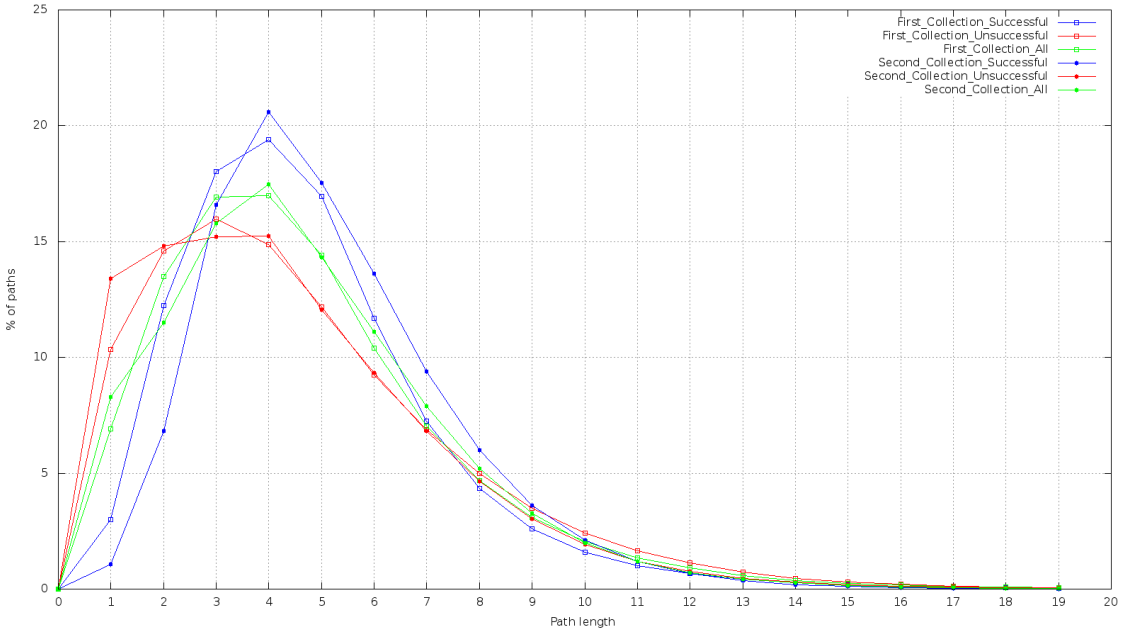

Abbildung 34: Vergleich mit den Wikigame-Klickpfaden

Im direkten Vergleich zwischen Abbildung 28 ohne Abbruchsimulator und Abbildung 33 mit aktiver Abbruchrate ist eine enorme Verbesserung erkennbar. Dies zeigt, dass durch die Verwendung dieses Simulators bei der automatischen Generierung von Navigationspfaden das Benutzerverhalten sehr gut angenähert werden kann. Auch die graphisch dargestellte Abbruchrate der Simulation in Abbildung 33 und der Wikigame-Daten in Abbildung 27 zeigen sich annäherungsweise ident.

Zum rechnerischen Vergleich zwischen den generierten Navigationspfaden und den Klickpfaden aus dem Wikigame wurden beide Verteilungen betrachtet. Anschließend wurde die Kullback-Leibler-Divergenz – auch KL-Divergenz genannt – berechnet, die Aufschluss über die Unterschiedlichkeit zweier Wahrscheinlichkeitsverteilungen gibt. Die Berechnung erfolgte anhand der Formel:

$$KL(P,Q) = \sum_{x \in X} P(x) \cdot \log \frac{P(x)}{Q(x)}$$

Um einen Vergleich zu ermöglichen, wurde zusätzlich sogenanntes Smoothing der Verteilungen angewandt. Dadurch ist gewährleistet, dass alle Werte aus beiden Sets berücksichtigt werden können, auch wenn der entsprechende Wert in einem Set den Wert null besitzt.

Tabelle 9 zeigt eine Übersicht der vier betrachteten Navigator-Konfigurationen. Es sind jeweils der Navigationstyp, die verwendete Hierarchie sowie die Parameter für die Abbruchsimulation angegeben. Die berechneten Werte für die KL-Divergenz zwischen generierten Pfaden und Wikigame-Pfaden sind angeführt.

Tabelle 9: Vergleich der Kullback-Leibler-Divergenz (KL)

Navigationstyp	*Random*	*Random*	*Greedy*	*Greedy*
Hierarchie	*Top5-BFS „United_States“*	*Top5-BFS „United_States“*	*Top5-BFS „United_States“*	*Top5-BFS „United_States“*
Abbruchsimulation	*(keine)*	*quadratisch*	*(keine)*	*quadratisch*
attritionsim.resistance	*-*	*0,96*	*-*	*0,96*
attritionsim.shiftx	*-*	*12*	*-*	*12*
attritionsim.shifty	*-*	*0,75*	*-*	*0,75*
attritionsim.limit	*-*	*0,8*	*-*	*0,8*
KL Länge alle Pfade	6,6053	0,1509	0,3230	0,0108
KL Länge erfolgreiche	0,8401	0,4560	0,1701	0,0349
KL Länge nicht erfolgr.	6,9029	0,1900	6,4484	0,0161
KL ∅ Grad alle Pfade	0,6250	0,5549	0,3075	0,3104
KL ∅ Grad Schritt 1	0,8714	0,6542	0,5755	0,5729
KL ∅ Grad Schritt 2	0,9344	0,6756	0,7461	0,7444
KL ∅ Grad Schritt 3	0,9238	0,6462	0,7068	0,7095
KL ∅ Grad Schritt 4	0,8404	0,5486	0,5640	0,5622
KL ∅ Grad Schritt 5	0,7374	0,4355	0,4452	0,4460
KL ∅ Grad Schritt 6	0,6862	0,3610	0,3751	0,3734

Die KL-Werte zeigen eindeutig die Verbesserung durch eine aktivierte Abbruchsimulation. Jene KL-Divergenzen der Pfadlängen bei der Greedy-Navigation mit quadratischer Abbruchrate weisen Werte kleiner als 0,04 auf, das auf eine deutliche Ähnlichkeit mit der Vergleichsverteilung aus den Wikigame-Pfaden hinweist. Auch aus dem Vergleich der Durchschnittseingangsgrade in den einzelnen Schritten geht eine Verbesserung hervor.

7.5 Kombination des Abbruchsimulators mit dem Link-Restrictor

Die alleinige Anwendung eines Abbruchsimulators durch einen Greedy-Navigator kann das Benutzerverhalten nur teilweise annähern. Aus diesem Grund wurde eine Kombination mit der von Geigl in [4] entwickelten Framework-Erweiterung, einem Link-Restrictor, in Betracht gezogen.

Durch die Verwendung des Link-Restrictors, der die Anzahl der auswählbaren Links in jedem Schritt der Navigationssimulation einschränkt, konnte eine geeignete Konfiguration gefunden werden. Die bereits beschriebenen Parameter mit der besten Performance aus Abschnitt 7.4 wurden um den Link-Restrictor, der sich in zwei Phasen aufteilt, erweitert. In der ersten Phase wird in jedem Schritt auf jene Knoten mit den niedrigsten Eingangsgrad (Parameter „amount“: 99.3 %) eingeschränkt, in der zweiten Phase auf jene mit dem höchsten Eingangsgrad (Parameter „amount“: 40.0 %). Ein Wechsel der Phase erfolgt nach dem ersten Navigationsschritt, die Einschränkung wird bis zum siebenten Schritt durchgeführt. Eine detaillierte Beschreibung befindet sich in [4].

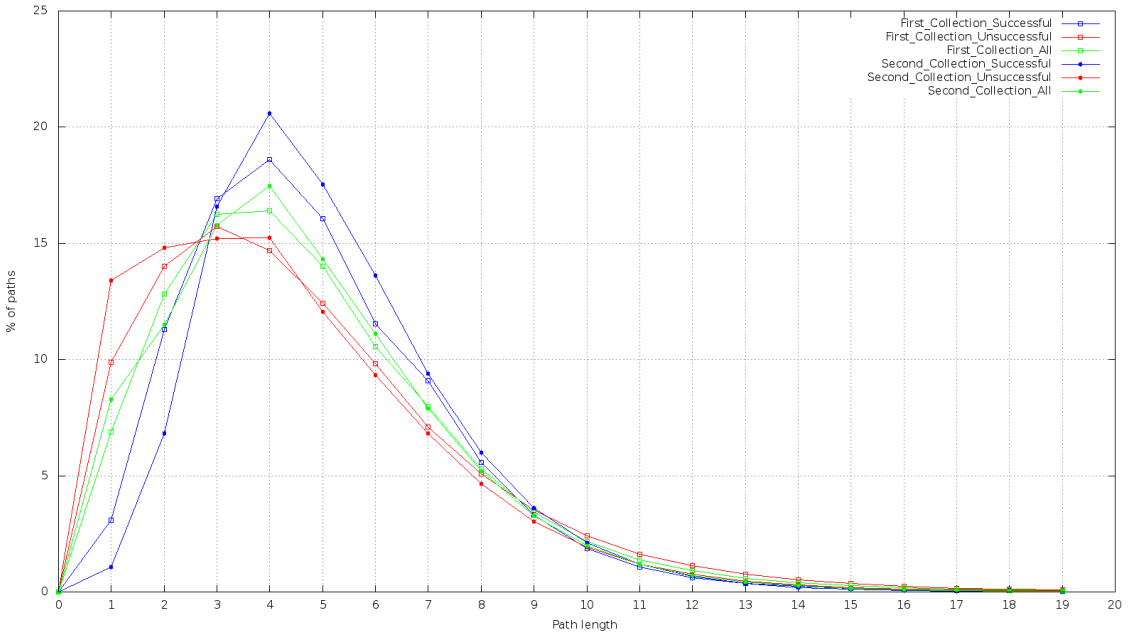

Abbildung 35: Abbruchsimulator in Kombination mit Link-Restrictor

Abbildung 35 zeigt den Längenvergleich zwischen den generierten Pfaden und den Klickpfaden der Benutzer aus dem Wikigame. Darin ist ersichtlich, dass die blau dargestellten erfolgreichen Pfade sehr gut an die Vergleichspfade angenä-

hert werden konnten. Auch die nicht erfolgreichen rot gefärbten Pfade liefern eine gute Annäherung, wobei der Verlauf auch nur geringfügig abweicht. Eine ausführliche Auswertung beschreibt Geigl ebenfalls in [4].

Die in diesem Abschnitt gezeigten Ergebnisse weisen bereits eine große Ähnlichkeit zwischen den durch den Simulator erstellten Pfaden und den Vergleichspfaden des Wikipedia-Spieles auf.

8. Conclusio

Dieses Projekt hat gezeigt, dass sich die Materie der Benutzermodellierung sehr komplex darstellt. Mit dieser Arbeit und dem Gesamtprojekt konnten zahlreiche interessante Forschungsergebnisse erstellt werden, die u.a. auch als Basis für weitere Tätigkeiten auf diesem Gebiet dienen.

Bei der Datenaufbereitung ist es von besonderer Wichtigkeit, dass die Herkunft und die Qualität der Daten genau hinterfragt werden muss. Von diesen sind alle weiteren Schritte abhängig, da sie die Basis dafür zur Verfügung stellen. Aus diesem Grund wurde besonders ausführlich auf die Datenaufbereitung Bezug genommen. Die Daten des Wikigames haben gezeigt, dass man sich nicht darauf verlassen soll, dass alle Daten korrekt verarbeitet werden, sondern weitere Analysen Aufschluss über die Beschaffenheit geben können. Durch die umfangreichen Datensätze, die zur Simulation verwendet wurden, konnten repräsentative Ergebnisse erzielt werden.

Die iterative Entwicklung verschiedener Hierarchien, die das Hintergrundwissen modellieren, hat sich ebenfalls als Erfolg herausgestellt. Durch die Zusammenarbeit konnten Strukturen gefunden werden, die sich sehr gut zur Navigationssimulation eignen. Das Verhältnis von Tiefe und Breite spielt bei Hierarchien eine entscheidende Rolle.

Der Abbruchsimulator zeigt auf, dass eine große Auswahl an möglichen Konfigurationen existiert. Auch hier erweist sich durch zahlreiche Versuche die Annäherung an den Benutzer als sehr komplex. Durch diverse Analysen gelang es, eine geeignete Annäherung zu finden, die im Rahmen des Frameworks zur Navigationsmodellierung verwendet werden kann. Ebenfalls wurde die Kombinierbarkeit mit anderen Erweiterungen, im Speziellen dem Link-Restrictor, demonstriert.

9. Future Work

Diese Arbeit bietet mehrere Ansätze für weitere Forschungstätigkeiten in diesen Themenbereichen. Diese betreffen sowohl die Datenaufbereitung als auch Erweiterungen des Abbruchsimulators.

Bezüglich der Datenaufbereitung zur Simulation bzw. zu Vergleichszwecken könnte eine Differenzierung der Daten aus dem Wikigame getroffen werden. Das Wikigame bietet sechs verschiedene Spieltypen an, die sich von den Zielen mehr oder weniger unterscheiden. Hierbei könnten die einzelnen Navigationspfade getrennt extrahiert und analysiert werden, um eventuell weitere Einblicke in das Benutzerverhalten zu erlangen.

Auch bei der Erstellung von hierarchisch aufgebautem Hintergrundwissen besteht noch Forschungspotential. Wie bereits in der Datenaufbereitung erwähnt, erzielte eine Hierarchie, die anhand der Breitensuche und einer Einschränkung der Knotenanzahl, bereits deutlich bessere Erfolgsraten. Es könnten diverse andere Ansätze zur Erzeugung noch erforscht werden, die dann anhand des Frameworks und den implementierten Erweiterungen getestet werden können.

Ebenfalls bietet der Abbruchsimulator noch weitere Möglichkeiten zur Erweiterung an. So könnten weitere Daten in die Simulation einfließen, die die Entscheidungen beeinflussen. Das Framework erlaubt es durch eine Analyse der Navigationspfade mit einer gewissen Wahrscheinlichkeit den Klick eines Benutzers in Kategorien einzuteilen. Basierend auf diesen erhobenen Daten könnte das Abbruchverhalten bzw. die Rate verändert werden, wenn der Benutzer beispielsweise öfter zu bereits besuchten Artikeln zurücknavigiert. Ebenso könnte die Zeit, die ein Benutzer auf einzelnen Seiten verbringt, Hinweise auf das Verhalten liefern.

Die hier erwähnten Forschungsansätze zeigen nur einen Teil des Potentials, das dieses Gesamtprojekt noch an Möglichkeiten in sich birgt.

Literaturverzeichnis

[1] T. Berners-Lee und M. Fischetti, Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor, Paw Prints, 2008.

[2] T. Alby, Web 2.0: Konzepte, Anwendungen, Technologien, München: Hanser Verlag, 2008.

[3] M. Eder, *Entwicklung eines Frameworks zur Navigationssimulation,* Graz: Technische Universität Graz, noch nicht veröffentlicht.

[4] F. Geigl, *Analyse und Modellierung des menschlichen Navigationsverhaltens in einem Wikipedia-Netzwerk,* Graz: Technische Universität Graz, noch nicht veröffentlicht.

[5] D. Easley und J. Kleinberg, Networks, Crowds, and Markets: Reasoning About a Highly Connected World, Cambridge University Press, 2010.

[6] M. Newman, Networks: An Introduction, Oxford University Press, 2009.

[7] Ø. Ore, Theory of Graphs, American Mathematical Society, 1962.

[8] R. Diestel, Graph Theory, Berlin: Springer-Verlag, 2010.

[9] J. Kleinberg, „Small-World Phenomena and the Dynamics of Information“, in *Proceedings of the 2001 Neural Information Processing Systems Conference*, Vancouver, 2001.

[10] P. Ayers, C. Matthews und B. Yates, How Wikipedia Works: And How You Can Be a Part of It, San Francisco: No Starch Press, 2008.

[11] M. Völkel, M. Krötzsch, D. Vrandecic, H. Haller und R. Studer, „Semantic Wikipedia“, in *Proceedings of the 2006 international symposium on Wikis*, Odense, 2006.

[12] „Wikipedia-Statistik - Tables - Anzahl Artikel (offiziell)“, [Online]. Available: http://stats.wikimedia.org/DE/TablesArticlesTotal.htm. [Zugriff am 18. Januar 2013].

[13] „Wikipedia:Namespace - Wikipedia, the free encyclopedia“, [Online]. Available: http://en.wikipedia.org/wiki/Wikipedia:Namespace. [Zugriff am 18. Januar 2013].

[14] R. West und J. Leskovec, „Automatic versus Human Navigation in Information Networks“, in *Proceedings of the 23rd ACM conference on Hypertext and social media*, Milwaukee, 2012.

[15] R. West und J. Leskovec, „Human Wayfinding in Information Networks“, in *Proceedings of the 21st international conference on World Wide Web*, Lyon, 2012.

[16] R. W. White und S. M. Drucker, „Investigating Behavioral Variability in Web Search“, in *Proceedings of the 16th international conference on World Wide Web*, Banff, 2007.

[17] J. Teevan, C. Alvarado, M. S. Ackerman und D. R. Karger, „The Perfect Search Engine Is Not Enough: A Study of Orienteering Behavior in Directed Search“, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Wien, 2004.

[18] D. F. Gleich, P. G. Constantine, A. D. Flaxman und A. Gunawardana, „Tracking the Random Surfer: Empirically Measured Teleportation Parameters in PageRank“, in *Proceedings of the 19th international conference on World wide web*, Raleigh, 2010.

[19] J. Voß, „Measuring Wikipedia“, in *Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics*, Stockholm, 2005.

[20] D. Helic, „Analyzing User Click Paths in a Wikipedia Navigation Game“, in *Proceedings of the 35th International Convention of Information Communication Technology, Electronics and Microelectronics*, Opatija, 2012.

[21] M. Boguñá, D. Krioukov und K. C. Claffy, „Navigability of complex networks“, *Nature Physics,* pp. 74-80, 2009.

[22] S. Milgram, „The small-world problem“, *Psychology Today,* p. 60–67, 1967.

[23] J. Kleinberg, „The small-world phenomenon: An algorithmic perspective“, in *Proceedings of the 32nd ACM Symposium on Theory of Computing*, Portland, 2000.

[24] P. S. Dodds, R. Muhamad und D. J. Watts, „An Experimental Study of Search in Global Social Networks“, *Science (Vol. 301),* pp. 827-829, 2003.

[25] J. Kleinberg, „Complex Networks and Decentralized Search Algorithms“, in *Proceedings of the International Congress of Mathematicians*, Madrid, 2006.

[26] D. J. Watts und S. H. Strogatz, „Collective dynamics of “small-world” networks“, *Nature,* pp. 440-442, 1998.

[27] C. Trattner, P. Singer, D. Helic und M. Strohmaier, „Exploring the Differences and Similarities between Hierarchical Decentralized Search and Human Navigation in Information Networks“, in *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies*, Graz, 2012.

Abbildungsverzeichnis

Tabellenverzeichnis

Listingverzeichnis

Abkürzungsverzeichnis

BFS	Breadth-First Search
BSD	Berkeley Software Distribution
bzw.	beziehungsweise
ca.	circa
d.h.	das heißt
GB	Gigabyte
HITS	Hyperlink-Induced Topic Search
ID	Identification (Identifikationsnummer)
KL	Kullback-Leibler
MB	Megabyte
MUN	Modeling User Navigation
MySQL	My Structured Query Language
PHP	PHP: Hypertext Preprocessor
SNAP	Stanford Network Analysis Platform
SQL	Structured Query Language
u.a.	unter anderem
URL	Uniform Resource Locator
usw.	und so weiter
v.a.	vor allem
vgl.	vergleiche
WWW	World Wide Web
XML	Extensible Markup Language
z.B.	zum Beispiel