TECHNISCHE UNIVERSITÄT
KAISERSLAUTERN

BACHELORARBEIT

Untersuchung von Querverweisen und semantischer Verwandschaft in Wikipedia

Matthias Streuber

BETREUER:
Prof. Dr. Prof. h.c. Andreas Dengel
Dipl. Inf. Darko Obradovic

9. November 2011

Inhaltsverzeichnis

Abbildungsverzeichnis

Zusammenfassung

Das Ermitteln und Bewerten von semantischen Beziehungen zwischen verschiedenen Konzepten ist eine Aufgabe, die zu lösen möglich, aber nicht einfach ist. Technische Probleme wie Laufzeit und Datenmengen, oder linguistische Probleme wie Polysemie und Synonymie erschweren eine solche semantische Interpretation. Es gibt jedoch viele Ansätze, wie man diese Probleme umgehen und eine annehmbare semantische Beziehung herleiten kann.

Wikipedia Link-based Measure und die *Google Similarity Distance* sind beides Verfahren, die im Rahmen dieser Bachelorarbeit hinsichtlich ihrer Leistung bei der Ähnlichkeitsbestimmung untersucht werden.

Abstract

The Identifcation and evaluation of semantic relationships between different concepts is a task which to solve is possible but not easy. Technical problems such as runtime and data sets, or linguistic problems such as polysemy and synonymy way of impede such a semantic interpretation. There are, however, many approaches how to handle these problems and derive an acceptable semantic relationship.

Within the scope of this bachelor thesis, *Wikipedia Link-based Measure* and the *Google Similarity Distance* are both methods, which are examines their performance in terms of similarity determination.

Kapitel 1

Einleitung

1.1 Motivation

Die Bestimmung semantischer Beziehungen zwischen zwei Objekten fällt uns Menschen weitestgehend leicht. Auf die Frage, in welcher Beziehung Katze und Maus stehen, oder was ein Buch und Papier verbindet, findet sich entsprechend schnell eine Antwort. Die Art und Aussagekraft der Antwort hängen allerdings vom jeweiligen Wissensumfang der befragten Person ab. Wenn diese Person beispielsweise noch nie ein Schachspiel gesehen hat, führt die Frage nach einer Beziehung zwischen „König“ und „Turm“ sicher zu einer anderen Antwort, als wenn ein professioneller Schachspieler diese Frage gestellt bekäme. Die subjektive Bewertung solcher Zusammenhänge ist also abhängig vom jeweils individuellen Wissen. Und eben jenes Wissen ist auch für Computer erforderlich, um vergleichbare Leistung bei der Bewertung von Beziehungen zu erzielen. Nur wenn dem System die Begrifflichkeiten eingepflegt und eine Art Beziehung vordefiniert wurden, ist der Computer in der Lage, die angefragte Beziehung direkt, oder als Folgerung verschiedener Fakten, zu liefern.

Wissensdatenbanken, die Informationen mit Relationen verwalten, sind bei weitem keine Seltenheit und aus dem Alltag vieler gar nicht mehr wegzudenken, sei es bei der Kaufberatung eines Produkts und den damit verbundenen Verweisen auf ähnliche empfehlenswerte Produkte, bei der Suche nach einem technischen Problem, bei der Suche nach wissenschaftlichen Konzepten, die einander referenzieren, oder schlicht und ergreifend bei einer Suchmaschine wie Google. Der Vorteil dieser Suchmaschine liegt darin, dass die einzelnen Seiten, und die uns angezeigte Ähnlichkeit verschiedener Seiten hinsichtlich einer Suchanfrage, nicht manuell, sondern

weitestgehend automatisiert eine Beziehung zueinander haben. Diese Art von automatisiertem, qualitativ hochwertigem Bewerten von Paarbeziehungen ist es, worauf die Forschung semantischer Ähnlichkeit hinarbeitet.

Doch ein erstes Problem liegt bereits in der dafür benötigten Datenmenge. Der Aufwand, das scheinbar unendliche Sammelsurium von Begriffen von Hand in ein System einzupflegen, ist enorm. Und das darauf folgende Bewerten von Beziehungen - nicht minder aufwändig - wäre wiederum abhängig von der subjektiven Wahrnehmung der Menschen, die diese Bewertung manuell einfügen.
Einfacher wäre es, bereits existierende und entsprechend große Datenmengen zu verwenden. Die Internet-Enzyklopädie Wikipedia stellt das unbestreitbar größte Vorkommen solchen Wissens zur Verfügung. Bereits mehr als eine Million Artikel sind in deutscher Sprache verfügbar, in der englischen Version sogar weit über drei Millionen. Nun stellt sich die Frage, ob das in Wikipedia mehr oder minder strukturierte Wissen tatsächlich in eine entsprechend auswertbare Form gebracht werden kann, und es somit einem Programm ermöglichen würde, eine semantische Beziehung zwischen zwei Objekten zu bewerten.

1.2 Historie und ähnliche Arbeiten

Die Aufgabe, semantische Beziehungen zu bestimmen, war schon im Jahre 1952 von Relevanz [5]. In theoretischen Ansätzen wurden verschiedene Verfahren entwickelt, um den semantischen Abstand zweier Objekte zu ermitteln, siehe [12, 9, 13, 2, 1]. Insofern ist es nicht verwunderlich, wie viele Forschungsgruppe sich mit dieser Thematik auseinandersetzen, zumal ab 2001, dem Gründungsjahr der Online-Enzyklopädie Wikipedia, ein neues Wissensmedium verfügbar war.

Doch bis dahin (ab 1995) wurden vorerst größere, von Hand angelegte Datenmengen verwendet. Ein Beispiel hierfür wäre das an der Princeton University entwickelte *WordNet* [8], ein Wortschatz englischer Sprache, beziehungsweise eine Datenbank, die rund 207000 Begriffe (Nomen, Verben, Adjektive) und ihre jeweiligen Synonyma miteinander verbindet. Durch diese manuell eingepflegten Beziehungen war es effektiv möglich, erste automatisierte semantische Analysen laufen zu lassen.

Kurze Zeit später (2000) wurden zu Analysezwecken weitere Thesauri erprobt, von Experten manuell angelegte Datenbanken, in denen eindeutige Begriffe und deren Beziehung untereinander durch Kategorien wie *Oberbegriff*, *engerer Sinn* und *verwand mit* beschrieben werden [4]. Aber auch

hier war man auf die händische Pflege der Datensätze angewiesen. Die Verfahren selbst jedoch waren vielversprechend.

Da sich nach fünf Jahren Entwicklungszeit Wikipedia aus den in Abschnitt 2.3 genannten Eigenschaften zu einer nützlichen Informationsplattform etabliert hat, wurden etliche neue Verfahren auf Basis dieser Internet-Enzyklopädie entwickelt. Zwei davon, GSD und WLM, werden im Verlauf dieser Arbeit näher betrachtet.

Die Entwicklung und Wartung von Thesaurus-System ist aufgrund des sich stets im Wandel befindlichen Wissens äußerst aufwendig. Daher versuchte eine Forschungsgruppe der University of Waikato auf Basis der Wikipedia einen automatisiert Thesaurus zu entwickeln, der die entsprechenden Thesaurus-Strukturen extrahiert [10].
Gleichzeitig arbeitete eine Arbeitsgruppe der Universität Karlsruhe (TH) an einer Erweiterung der Wikipedia-Link-Syntax namens *Semantic Wikipedia*, um es den Autoren relativ einfach zu ermöglichen, sich selbst am Prozess der semantischen Gliederung zu beteiligen. So wurde die recht einfache Wikipedia-Syntax für interne Links der Art erweitert, das objektbezogene Beziehungen angegeben werden konnten, zum Beispiel: „*London is the capital city of [[capital of::England]]*“ [15]

Das Projekt WikiRelate! [14] implementiert ein Verfahren zur Bestimmung semantischer Ähnlichkeit auf Basis von Wikipedia und vergleicht dieses mit eigenen Tests auf der Datenmenge von WordNet. Dabei wurde festgestellt, dass sich Wikipedia weitaus besser für solche Analysen eignet, als WordNet.

Ab 2007 schließlich schien der Trend von auf Thesaurus basierten Systemen nachgelassen zu haben. Viel mehr wurde sich jetzt mit dem Problem der Disambiguierung (siehe Abschnitt 2.2) und der Analyse semantischer Beziehungen auf Basis von Wikipedia auseinandergesetzt.

So entwickelten Razvan Bunsecu und Silviu Cucerzan unabhängig voneinander erste Disambiguierungsverfahren auf Wikipedia, die beide recht vielversprechende Ergebnisse lieferten [**? ?**]

Nahezu zeitgleich wurde ein Verfahren auf Basis expliziter Links entwickelt, welches sich *Explicit Semantic Analysis* (ESA) nennt [6]. Zusätzlich bedient sich ESA dem *Open Directory Project* (ODP), dem derzeit größten Online-Verzeichnis. Darin werden Seiten und Konzepte von freiwilligen Redakteuren einer Kategorie-Hierarchie zugewiesen. ESA verwendet und evaluiert sein Vorgehen auf Basis von sowohl Wikipedia-Datensätzen, als auch ODP-Kategorien.

Mit Wikify [7] wurde einer der ersten Online-Dienste zur Verfügung

gestellt, der es Benutzern ermöglichen soll, ihre in ein Textfeld kopierbaren Texte automatisch mit dazu passenden Wikipedia-Links bestücken zu lassen[1]. Relevante Begriffe sind all jene, die nicht auf einer sogenannten Blacklist stehen, also einer Liste von Worten, denen keine große Relevanz zugewiesen wird[2]. Die verbliebenen Worte werden in Wikipedia gesucht und bei einem oder mehreren Treffern auf Ähnlichkeit mit dem vorliegenden Text verglichen und entsprechend eingebunden.

Ohne Bezug zu Wikipedia stellte Anfang 2007 Rudi L. Cilibrasi die *Google Similarity Distance* [3] vor, ein semantisches Maß, welches die Suchmaschine Google intern für die Verarbeitung von Suchanfragen verwendet. Dies ist eines von zwei Verfahren, welches im weiteren Verlauf detaillierter vorgestellt werden wird.

Das andere Verfahren (2008) nennt sich *Wikipedia Link-based Measure* und arbeitet auf einer semantischen Interpretation von Verweisen zwischen einzelnen Wikipedia-Artikeln.

[1] `http://wikify.appointment.at`

[2] `http://www.world-english.org/english500.htm`

Kapitel 2

Theoretische Grundlagen

In diesem Abschnitt werden einige Begrifflichkeiten für den weiteren Gebrauch erklärt und die Funktionsweise von Wikipedia und der Nutzen der WordSimilarity-353 näher erläutert.

2.1 Terminologie

Konzept Der Begriff des Konzepts bezieht sich hier auf den erläuternden Repräsentanten eines Begriffs in Textform. Im Bezug auf die Enzyklopädie Wikipedia ist damit der Wikipedia-Artikel zu einem entsprechenden Begriff gemeint. Dieser Artikel beinhaltet weitere Informationen, die den Begriff näher erläutern und somit Bezug zu anderen Begriffen und deren Konzepten herstellen, sei es nun direkt oder indirekt.

Link Würde man sich die Menge aller Begriffe (und deren Konzepte) als Knoten eines gerichteten Graphen vorstellt, entsprächen die Kanten zwischen den Knoten den Verweisen zwischen den Konzepten. Übertragen auf Wikipedia werden dort Beziehungen zwischen Konzepten mittels Hyperlinks hergestellt. Ein Link ist also eine gerichtete Kante zwischen genau zwei Konzepten.

Ähnlichkeit Dieser Begriff ist aus zwei Fachbereichen zu betrachten. Mathematisch motiviert kann ein Konzept als Vektor dargestellt werden. Betrachtet man die Ähnlichkeit zweier Vektoren, ermittelt man ihren jeweiligen Abstand zueinander. Über die Kosinus-Ähnlichkeit wird der Winkel zwischen diesen beiden Vektoren analysiert: stehen

sie senkrecht zueinander, sind sie - um wieder auf Konzepte zu kommen - maximal unähnlich, je kleiner der Winkel, desto ähnlicher die Konzepte.
Philosophisch motiviert wird die Ähnlichkeit zwischen zwei Konzepten anhand unterschiedlicher und gemeinsamer Eigenschaften ausgemacht. In den später vorgestellten Verfahren entspräche eine solche Eigenschaft beispielsweise der Menge ausgehender Links zweier Konzepte.
Die in dieser Arbeit vorgestellten Verfahren bedienen sich in erster Linie der mathematischen Aspekte, aber die Entwicklung und Gestaltung der Verfahren ist philosophisch motiviert.

2.2 Disambiguierung

Ein wesentliches Problem bei der Bestimmung von Ähnlichkeiten zwischen Begriffen ist das der Ambiguität, der Mehrdeutigkeit. Schon seit 1950 wird auf dem Gebiet der automatisierten Disambiguierung geforscht [13]. Es ist, wie das in der Motivation genannte Beispiel des Schachspiels, sowohl vom Wissen des verarbeitenden Systems, als auch von einer kontextbezogenen Interpretation abhängig. Die Aussage „Der Kohl ist gute zwei Meter groß“ macht, sofern der Interpret vom entsprechenden Gemüse ausgeht, wenig Sinn. Fügt man aber der interpretierenden Wissensmenge Informationen über den Altbundeskanzler Helmut Kohl und dessen Körpergröße hinzu, kann der Interpret diese Aussage bestätigen.

Unter der Disambiguierung [18] versteht man das Differenzieren und Auflösen von Mehrdeutigkeiten auf sprachlicher Ebene. Zwei wesentliche Aspekte, die gerade bei der Bestimmung semantischer Ähnlichkeiten relevant sind:

Polysemie

Im Allgemeinen beschreibt Polysemie die einfache Mehrdeutigkeit, wie das „Kohl-Beispiel“ im oberen Abschnitt. Einem Begriff werden also mehrere Bedeutungen zugewiesen.
Das Wort „Läufer“ referenziert in der deutschen Wikipedia am 01.11.2011 29 Konzepte [1]. Um in diesem Fall die Mehrdeutigkeit zu beseitigen, wird entweder eine kontextbezogenen Interpretation durchgeführt, oder mindestens ein weiterer Begriff zur Eingrenzung der Bedeutung benötigt.

[1] `http://de.wikipedia.org/wiki/Läufer`

Synonymie

Unterschiedliche Begriffe mit gleicher Bedeutung werden Synonyme genannt, beispielsweise „Pferd“ und „Gaul“ . Hinsichtlich eines verarbeitenden Systems zur Bestimmung semantischer Ähnlichkeit ist es erforderlich, Synonyma zu erkennen, um die Sinnverwandtschaft bei der Ähnlichkeitsberechnung mit einzubeziehen.

2.3 Wikipedia

Die Online-Enzyklopädie Wikipedia ermöglicht es seit 2001 jedem, Wissen in Textform einzupflegen, zu korrigieren und zu aktualisieren. Die englische Version umfasst bereits drei Millionen Artikel. Das Konzept „Jeder darf mitmachen“führt nicht nur zu einem rapiden Wachstum der Wikipedia-Inhalte[2], sondern auch zu einer sich stets verbessernden Qualität selbiger. Doch nicht nur die Menge an Daten macht Wikipedia zu einer geeigneten Plattform Wissen verarbeitender Form, sondern auch der strukturierte und meist regelkonforme Aufbau der Artikel selbst. Einige wesentliche Vorteile hinsichtlich semantischer Analyse sind:

Linksyntax

Autoren sind dazu angehalten, in ihren Texten auf andere Artikel an passender Stelle zu verweisen, sollten diese relevant sein. Links [3], die nur auf andere Konzepte innerhalb Wikipedias verweisen, haben eine einfache Syntax: *[[Zielartikel | alternativer Text]]*. Sie lassen sich für die Zwecke dieser Arbeit leicht verarbeiten, und der alternative Text kann zusätzlich für semantische Interpretationen genutzt werden, beinhaltet er doch unter Umständen ein vom Autor gesetztes Synonym.

Disambiguierung

Das unter Abschnitt 2.2 vorgestellte Problem der Polysemie ist durch die Disambiguierungsseiten einfacher zu fassen. Sucht man beispielsweise nach „Läufer“, werden auf dieser Disambiguierungsseite alle Konzepte aufgelistet, die sich mit dem Begriff *Läufer* befassen. Um die Einzigartigkeit der Titel beizubehalten, werden diese mittels Kategorien erweitert (*Läufer (Schach)*)

[2] `http://de.wikipedia.org/wiki/Wikipedia:Statistik`
[3] `http://de.wikipedia.org/wiki/Hilfe:Links`

oder umbenannt (*Teppich*). Letztendlich findet hier aber keine automatisierte Disambiguierung statt, sondern lediglich eine Hilfestellung für den Benutzer geboten, um den Schritt der Disambiguierung selbst durchzuführen.

Weiterleitung

Auch dem Problem der Synonymie wird mittels Weiterleitungen ansatzweise entgegengewirkt. So leitet beispielsweise das Konzept *Redirect* auf *Weiterleitung* weiter, ohne selbst einen wesentlichen Inhalt zu besitzen.

Datenbank

Verfahren wie GSD (3.2) arbeiten mit eingehenden Links zu Konzepten. Theoretisch müssten nun alle Konzepte durchsucht und Links zu einem bestimmten Konzept extrahiert werden. Die Datenbankstruktur Wikipedias erleichtert dieses Vorgehen, da innerhalb des Systems eine Tabelle existiert, in der sämtliche Links von Konzept *a* zu Konzept *b* gespeichert werden. So kann auf zeitaufwendige Textsuchen verzichtet werden. Die praktische Anwendung wird im Abschnitt 4.3 näher erläutert.

Es folgt eine reduzierte Beschreibung der im Rahmen dieser Arbeit verwendeten Wikipedia-Tabellen [16], Spaltenbezeichner in runden Klammern

page *(page_id, page_title, page_is_redirect)* Der Titel, die dazu gehörige ID und eine Markierung, ob es sich um eine Weiterleitung handelt.

text *(old_id, old_text)* Eine Text-ID, die in einer anderen Tabelle mit der *page_id* verbunden wird, und der dazugehörige Text.

pagelinks *(pl_from, pl_title, pl_id)* Ein Link von der Seite mit der ID *pl_from* zu der Seite mit dem Titel *pl_title* und der ID *pl_id*, letzteres wurde eigens für diese Arbeit angelegt, um die Suche zu beschleunigen.

Artikelaufbau

Die einzelnen Artikel aus Wikipedia haben, bis auf kleinere Abweichungen, alle den selben Aufbau [17]: Zu Beginn fasst ein kurzer Text mit wenigen Zeilen die wesentlichen Inhalte des gesamten Artikels zusammen, ähnlich dem *Abstract* wissenschaftlicher Arbeiten.

Direkt darunter befindet sich bei kategoriebezogenen Themen eine Infobox. So finden sich dort zum Beispiel bei einem Artikel über ein Land Informationen zur Sprache, Hauptstadt, Einwohnerzahl, Fläche, Bruttoinlandsprodukt oder Regierungsform. Bei Tieren hingegen sind Informationen zu Gattung, Überfamilie, Unterfamilie oder wissenschaftlichen Namen angegeben. All diese Informationen sind wiederum Links zu anderen Konzepten. Allein diese Infobox beinhaltet schon ausreichend Informationen zur semantischen Kategorisierung des Artikelthemas.
Das Inhaltsverzeichnis eines Artikels bietet dem Leser einen Überblick über die folgenden Abschnitte. Und auch diese Information ist bei einer semantischen Analyse von Interesse, liefert es doch wesentliche Bestandteile des zu untersuchenden Konzepts.
Aus Datenverarbeitungssicht liegt ein weiterer Vorteil in der Struktur des Quelltextes. So sind eben vorgestellte Abschnitte durch Schlüsselworte klar voneinander getrennt. Die Infobox ist beispielsweise durch das Schlüsselwort $\{\{Infobox\}\}$ vom restlichen Seiteninhalt abgegrenzt. Im Bezug auf die Tierwelt existieren sogar besondere Infoboxen (Taxoboxen[4]), die ein festes Schema an Informationen zur systematischen Einordnung vorschreiben.

Kategorien

Wikipedia unterstützt im Ansatz auch schon eine Art Beziehungsdefinition. So ist es möglich, einen Artikel einer oder mehreren Kategorien zuzuweisen. Die Art der Beziehung entspricht dann einer *ist-ein*-Beziehung. Ein Subtyp der Kategorien[5] sind Listen. Wie der Name schon sagt, werden darin alle in einen Listentyp passenden Artikel eingeordnet[6].

Alle vorgestellten Eigenschaften sind natürlich abhängig von den jeweiligen Autoren. Doch je allgemeiner das Thema des Konzepts ist, desto größer ist die Zahl der Besucher und desto größer ist die Wahrscheinlichkeit, das grobe Fehler erkannt und beseitigt werden.

[4] `http://de.wikipedia.org/wiki/Wikipedia:Taxoboxen`
[5] `http://de.wikipedia.org/wiki/Spezial:Kategorienbaum/!Hauptkategorie`
[6] `http://de.wikipedia.org/wiki/Wikipedia:Listen`

2.4 WordSimilarity-353

WordSimilarity-353 [12] ist eine offene Sammlung englischer Wortpaare. Jedes dieser Wortpaare ist mit einem durch eine Menge von Menschen festgelegten Ähnlichkeitsmaß bewertet worden. Dieser Datensatz besteht aus zwei konkatenierten Listen von Wortpaaren:

Die erste Liste beinhaltet 153 Wortpaare und wurde von 13 Personen bewertet. 30 dieser Wortpaare entstammen der Sammlung G.A. Miller and W.G. Charles [9], allerdings wurden diese Paare von den genannten 13 Personen erneut bewertet.

Die zweite Liste besteht aus 200 Worten und wurde von 16 Personen bewertet. Bei der Zusammenführung beider Listen wurden die jeweiligen Bewertungen normiert. Ein Ähnlichkeitsmaß von 10 entspricht identischen Begriffen, ein Wert von 0 dagegen entspricht maximaler Unähnlichkeit.
Die Begriffe selbst sind weitestgehend englisch-nativer Art, wie zum Beispiel: Tiger, Professor, Train, Football, Physics, Food oder Coast. Ein Beispieldatensatz wäre dann ein Tripel wie (Tiger;Cat;7,35)

Dieser Datensatz wird recht häufig bei Testverfahren zur Bestimmung von Ähnlichkeiten verwendet, so zum Beispiel in [6, 11, 14, 3]

Kapitel 3

Implementierung

Nachdem die Begrifflichkeiten geklärt und die Beweggründe, nach denen Wikipedia als Hilfsmittel geeignet scheint, dargestellt wurden, geht es in diesem Abschnitt um die Vorstellung der im Rahmen dieser Arbeit getesteten Methoden zur Analyse semantischer Beziehungen.

3.1 WLM

Im Gegensatz zu den meisten der bereits genannten Verfahren, arbeitet *Wikipedia Link-based Measure* (WLM) [11] auf den explizit im Text angegebenen Links zu anderen Konzepten und vernachlässigt (zumindest, nachdem passende Konzepte gefunden wurden) den Kontext, indem diese Links stehen.
Der Vorteil dieser Herangehensweise liegt in der Möglichkeit, auf eine große Anzahl textverarbeitender Prozesse zu verzichten. Die Implementierung nutzt die bereits dargestellte einfache Syntax der Wikipedia-Konzeption zur Darstellung von Links. Das Verfahren arbeitet in zwei Schritten:

1. Finde in Konzept a zu jedem ausgehenden Link das entsprechende Konzept b_i und berechne das Gewicht von b_i zu a
   (siehe Algorithmus 1)

2. Berechne auf Basis der Menge an ausgehenden Links ein Ähnlichkeitsmaß anhand der Kosinus-Ähnlichkeit

Für jeden aus einem Konzept C ausgehenden Link wird also dessen Relevanz für das jeweilige Konzept C ermittelt. Durch den Quotienten $\frac{|W|}{|T|}$

Algorithmus 1 WLM: Gewicht

Require: *concept*
  $Links = getOutgoingLinks(concept)$
  $k = 0$
  $W = 2.670.068$
  for all Links **do**
    $secondConcept = getConceptByName(Link.name)$
    $T = countIncomingLinks(secondConcept)$
    $Link.weight = log(W/T)$
  end for

aus Algorithmus 1, wobei W der Menge aller Wikipedia-Artikel und T der Menge eingehender Links des jeweiligen Konzepts entspricht, erhalten alle Konzepte mit weniger eingehenden Links eine höhere Gewichtung. Eine Motivation hierfür ist die Annahme, die Ähnlichkeit zweier Konzepte werde durch die Seltenheit ihrer gemeinsamen Konzepte definiert.
So ist es beispielsweise nicht weiter beeindruckend, wenn beide betrachteten Konzepte auf das Konzept „Wissenschaft“ verweisen, wohl aber, wenn beide auf das „Hamiltonsches Prinzip“ zeigen, also eine Verfeinerung der Gemeinsamkeit. Allerdings kann diese Annahme auch zu Fehlschlüssen führen, wie spätere Auswertungen zeigen werden.

Der eben vorgestellte Algorithmus wird nun auf die beiden zu untersuchenden Konzepte angewandt. Die jeweils resultierenden Listen gewichteter Links werden als Vektoren betrachtet und mittels der Kosinus-Ähnlichkeit verglichen:

$$similarity = \frac{A \cdot B}{||A||\,||B||} \tag{3.1}$$

Das Ergebnis ist ein Wert zwischen 1 (die Konzepte sind identisch) und 0 (orthogonal, maximal unähnlich).

Anstatt von jedem Konzept den entsprechenden Text zu durchsuchen und Links zu extrahieren, Durch die in Abschnitt 2.3 erwähnte Tabelle innerhalb der Wikipedia-Datenbank kann der Suchaufwand erheblich reduziert werden. Das geänderte Vorgehen, mit vereinfachten SQL-Anfragen, ist in Algorithmus 2 zu sehen.

Algorithmus 2 WLM: vereinfachtes Gewicht

Require: *word*
  $concept = query(page \rightarrow page\_title \equiv word)$
  $Links = query(pagelinks \rightarrow pl\_from \equiv concept.page\_id)$;
  $k = 0$
  $W = 2.670.068$
  for all Links **do**
    $T = count(query(pagelinks \rightarrow pl\_id \equiv Link.pl\_id))$
    $Link.weight = log(W/T)$
  end for

3.2 Google Similarity Distance

Die generelle Idee, GSD für diese Zwecke zu verwenden, entspringt [11]. In dieser Annäherung wird das bewährte Ähnlichkeitsmaß der Google-Suchmaschine verwendet, beziehungsweise die darin verwendete *Google Similarity Distance* (GSD) [3], entwickelt nach einem theoretischen Ansatz der Kolmogorow-Komplexität [1]. Im Wesentlichen lässt sich das Problem der Linkanalyse von Wikipediakonzepten auf die Suche im Internet übertragen: Eine Website stellt ein Konzept dar, die eingehenden und ausgehenden Hyperlinks entsprechen den Verbindungen, also den Links der Konzepte. Der Vorteil des Google-Abstands ist dessen Schlichtheit. Websites, beziehungsweise Konzepte, welche gemeinsame eingehende Referenzen haben, sind sich ähnlich. Es werden keine Bewertungen der verweisenden Seiten vorgenommen, was gegenüber dem WLM-Verfahren geringeren Aufwand bedeutet.

Gesucht sei also der Abstand, oder auch NGD (Normalized Google Distance), zwischen den beiden Konzepten a und b, welcher sich wie folgt berechnen lässt:

$$NGD(a,b) = \frac{log\,(max\,(|A|,|B|)) - log\,(|A \cap B|)}{log\,(|W|) - log\,(min\,(|A|,|B|))} \tag{3.2}$$

wobei A und B die Mengen der jeweils auf a und b verweisenden Links darstellen und W - wie in WLM - die Menge aller Wikipedia-Konzepte beinhaltet. Je kleiner der Wert, umso geringer der Abstand und umso größer die Ähnlichkeit.

3.3 Implizite Beziehungen

Die bisher vorgestellten Verfahren bauen beide auf manuell festgelegten Links auf. Das ermittelte Ähnlichkeitsmaß ist also abhängig von der menschlichen Motivation, sich an die durch Wikipedia angegebenen Richtlinien zu halten, und relevante Konzepte miteinander zu verknüpfen.
Die Frage ist nun, ob auch nicht explizit verlinkte Begriffe eine sinnvolle Auswirkung auf die Bestimmung semantischer Ähnlichkeit haben. So ist beispielsweise anzunehmen, dass nicht jeder Autor jeden relevanten Substantiv mit einem Link zum entsprechenden Artikel versieht. Diesem „Problem"soll der Gebrauch impliziter Links entgegenwirken.

In diesem Ansatz wurde versucht, die Menge der expliziten Links eines Konzepts durch implizite Verweise zu erweitern. Zu diesem Zweck wurde auf der Datenbank der Konzepttexte ein Volltext-Index angelegt, um ein möglichst schnelles Ergebnis bei der Suche nach Worten zu erzielen. Diese implizite Linksuche soll die beiden vorgestellten Implementierungen erweitern, um besagte Abhängigkeit von manuell gesetzten Verweisen zu umgehen. Dabei ist natürlich anzunehmen, das auch semantisch irrelevante Konzepte in die Suche miteinbezogen werden. Gleichzeitig liefert die Volltextsuche einen von der Datenbank ermittelten Relevanzwert, der sich laut mySQL [1] für das gesuchte Wort W wie folgt errechnet:

$$w = \frac{log\,(dtf) + 1}{sumdtf} \cdot \frac{U}{1 + 0.0115 * U} \cdot log\left(\frac{N - nf}{nf}\right) \tag{3.3}$$

dtf: Anzahl Vorkommnisse von W innerhalb des Dokuments
$sumdtf$: Summe der $log\,(dtf) + 1$ für jedes Wort des Dokuments
U: Anzahl einzigartiger Worte innerhalb des Dokuments
N: Anzahl aller Dokumente
nf: Anzahl aller Dokumente, die W enthalten.

Diese Formel ist durch die des TFxIDF[2] motiviert und wurde lediglich durch Normalizierungsmechanismen erweitert, um so beispielsweise die Länge des Dokuments verglichen mit der Durchschnittslänge miteinzubeziehen.

[1] `http://forge.mysql.com/wiki/MySQL_Internals_Algorithms#Full-text_Search`
[2] `http://en.wikipedia.org/wiki/Tf-idf`

Kapitel 4

Evaluation

Es gilt nun die eben vorgestellten Verfahren zu testen. Nach einer kurzen Vorstellung der Testumgebung und den Testbedingungen folgt die Auswertung der einzelnen Verfahren. Im weiteren Verlauf werden versuchsweise Modifikationen vorgenommen und deren Auswirkung bewertet.
Dabei ist jeder Testlauf zweigeteilt. In Testlauf 1 werden Wortpaare der WordSimilarity-353 ausgewertet. Testlauf 2 hingegen prüft Wortpaare, die nach eigenem Ermessen eine äußerst geringe bis nicht existente Ähnlichkeit miteinander aufweisen. Dadurch soll geprüft werden, ob das jeweilige Verfahren auch in der Lage ist, Wortpaare semantisch schwacher Bindung als solche zu erkennen.

4.1 Testumgebung

Sämtliche Verfahren wurden in der Skriptsprache PHP implementiert. Die Skripte, ein Abbild der Wikipedia-Datensätze und die Testläufe wurden auf einem Rechner des DFKI (Deutsches Forschungszentrum für Künstliche Intelligenz) abgelegt, bearbeitet und durchgeführt.

Die nachfolgenden Ergebnisse wurden auf Basis der WordSimilarity-353 aus Abschnitt 2.4 durchgeführt.

Datenmenge

Wie zu Anfang begründet, verwendet dieser Evaluationsabschnitt Wikipedia als Datenmenge. Da Wikipedia seit knapp zwei Jahren keine Schnittstelle für das Verarbeiten von Anfragen mehr anbietet, wurde zu Testzwecken ein Abbild der Wikipedia-Datensätze vom 26. Mai 2011 heruntergе-

laden und in eine mySQL-Datenbank importiert. Diese Version beinhaltet 2.670.068 Artikel. Auf Grund von Importfehlern konnten rund 300.000 Artikel nicht vollständig eingeladen werden, was in Anbetracht der Menge an Links und Testläufen nur geringe Auswirkungen hat. Die in Abschnitt 2.3 genannte Tabelle, welche alle Links innerhalb des Wikipedia-Systems beinhaltet, umfasst 134.621.968 Einträge, das entspricht ungefähr 50 Links pro Konzept.

Testmenge

Bei der Auswertung ohne den Gebrauch impliziter Links wurden von den 353 Wortpaaren einige entfernt, da es in dieser Wikipedia-Version kein Konzept gibt, das sich beispielsweise mit „Trinken“ befasst. Die genaue Menge der getesteten Datensätze ist am Ende der Evaluation tabellarisch aufgeführt.

4.2 WLM

Das erste Testverfahren wurde auf allen 353 Wortpaaren ausgeführt, allerdings unter Angabe expliziter Konzepte, da die Disambiguierung aus 2.2 kein praktischer Bestandteil dieser Arbeit ist.

Testlauf 1

Nach dem 60. Wortpaar wurde die Versuchsreihe abgebrochen, da die bisherigen Ergebnisse des WLM-Verfahrens weit unter den erwarteten Werten lagen. Das Problem hierbei wird am Wortpaar *Tiger* und *Cat* erläutert:

Konzept	Anz. Links	∅ Gewicht/Link	Norm
Tiger	426	80	7729
Cat	463	35	12021

Tabelle 4.1: Exemplarische Auswertung des Wortpaares Tiger - Cat

Nach Formel 3.1 für die Kosinusähnlichkeit ergibt sich bei einem Skalarprodukt von 1539 gerade mal ein Ähnlichkeitswert von $0,16$. Angepasst an den Wertbereich der WordSimilarity-353 entspräche das einem Wert von $1,6$, verglichen mit der Expertenmeinung von $7,35$.

Ein Problem ist die relativ hohe Anzahl ausgehender Links, welche anhand ihrer Syntax aus dem Konzepttext extrahiert wurden, und die ver-

gleichsweise geringe Anzahl gemeinsamer ausgehender Links (27 in diesem Beispiel). In Abbildung 4.1 ist deutlich die Abweichung von der Expertenmeinung der WordSimilarity-353 zu erkennen. Niedrige Varianzen sind durch entsprechend geringe Werte in der Expertenmeinung zu erklären. Beispielsweise liefert das Wortpaar $(professor, cucumber)$ einen Expertenwert von $0,31$, das WLM-Verfahren hingegen $0,14$ (angepasst).

Das beste Ergebnis wurde für das Wortpaar $(King\_(Chess), Rook\_(Chess))$ erzielt, allerdings nur durch die Angabe expliziter Konzepte. Für das eigentliche Wortpaar $(King, Rook)$ wäre das WLM-Ergebnis auf Grund größerer Mengen ausgehender Links schlechter gewesen.

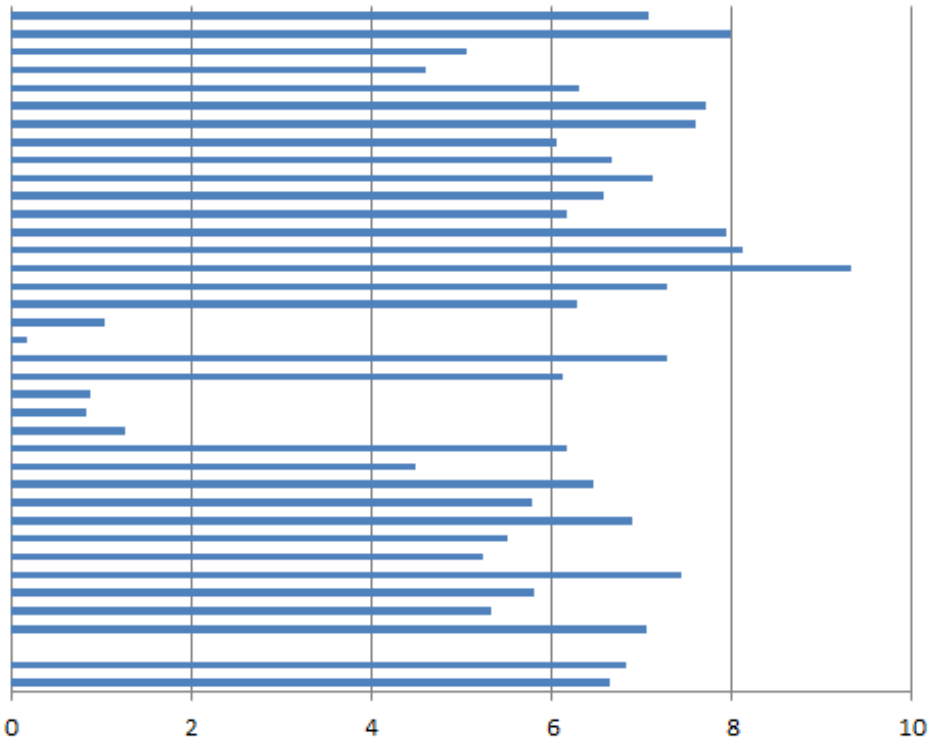

Abbildung 4.1: Varianz des WLM-Verfahrens

Testlauf 2

Abbildung 4.2 zeigt für sich genommen gute Ähnlichkeitswerte. Doch vergleicht man sie mit den Ergebnissen aus Testlauf 1, wird klar, dass auch diese Werte keine verwendbare Aussagekraft besitzen. Das nach eigenen Erwartungen gering zu bewertende Wortpaar $(Wine, Citizenship)$ hat eine semantisch engere Beziehung als das Wortpaar $(Footbal, Tennis)$ (0,05).

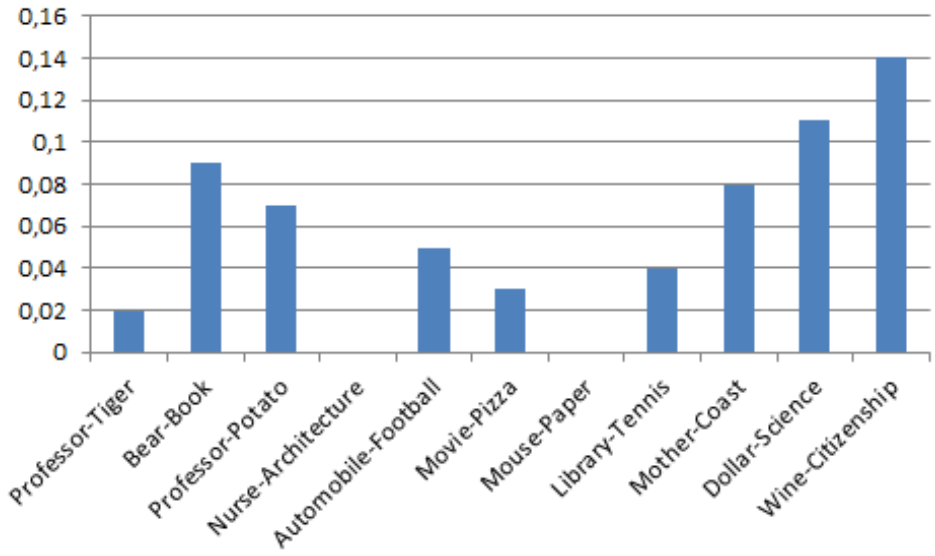

Abbildung 4.2: Varianz des WLM-Verfahrens auf Negativ-Fällen

4.3 GSD

Im Gegensatz zum WLM-Verfahren wurden in dieser Testreihe alle 353 Wortpaare abgearbeitet. Wie in Abschnitt 4.1 erwähnt, gibt es einige Wortpaare, denen kein eindeutiges Wikipedia-Konzept zugewiesen werden konnte, aber das Gesamtergebnis deckt sich trotzdem mit der Expertenmeinung.

Testlauf 1

In Abbildung 4.3 ist die absolute Abweichung der im GSD-Verfahren ermittelten Ähnlichkeitswerte zu denen der WordSimilarity-353 angegebenen Expertenmeinung dargestellt. Wie zu sehen ist, unterscheiden sich nur

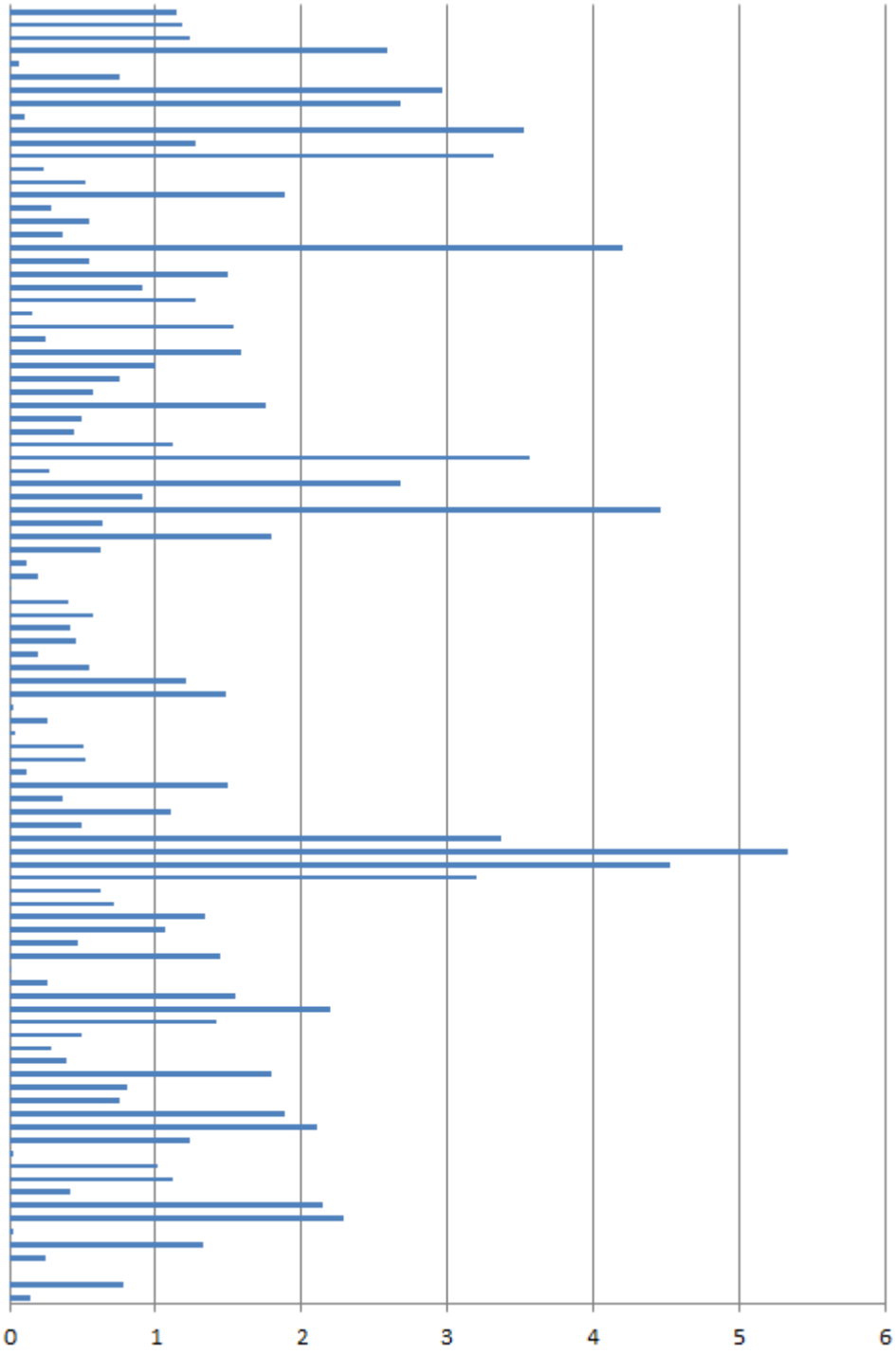

Abbildung 4.3: Varianz des GSD-Verfahrens

wenige der Wortpaare um einen Wert größer drei. Beispiele hierfür sind die Paare *(Monk, Slavery)*, *(Mars, Water)* und *(Shore,Woodland)*.

Bei dem Wortpaar *(Shore, Woodland)* haben die jeweiligen Konzepte nur 6, beziehungsweise 29 eingehende Links, und nur eine gemeinsame Verbindung. Für so geringe Zahlenwerte und eine entsprechend große Anzahl an Wikipedia-Einträgen (rund drei Millionen) ist der Quotient der Formel 3.2 entsprechend gering und somit die Ähnlichkeit sehr groß.

Die Abweichung bei $(Water, Mars)$ ist hingegen unerwartet. Diese Konzepte haben insgesamt 18 gemeinsame Links, fast ein Zehntel ihrer jeweils ausgehenden Links. Eine gewisse Relevanz ist da nicht von der Hand zu weisen. GSD ermittelt hier einen Wert von $7,4$, die Expertenmeinung hingegen schlägt eine Ähnlichkeit von $2,94$ vor. Ein gutes Beispiel dafür, wie kontextabhängig diese semantischen Bewertungen sind und wie stark sie sich von der persönlichen Meinung entfernen kann.

Testlauf 2

Wie in Abbildung 4.4 zu sehen ist, werden einige der Beispiele „missinterpretiert“ .

Dieser Vorwurf ist natürlich wieder kontextabhängig. Die Beziehungen existieren nachweislich, es ist nur die Frage, ob diese gewünscht ist. Der Vergleich von Bär und Tiger könnte einen durchschnittlichen Wert von 5 erzielen, wenn ich als Anwender aber insgeheim von der englischen Fantasie-Figur „Winnie the Pooh“ [1] ausgehe und dessen Freund „Tigger“meine, wäre der Vergleich ebenfalls missinterpretiert.

Zurück zur Auswertung. Betrachte man beispielsweise das Paar $(Library, Tennis)$, sind auf Anhieb keine Gemeinsamkeit zu erkennen. Doch haben diese Konzepte zwei gemeinsame Links: *Eindhoven* (mit 5933 eingehenden Links) und *Eindhoven University of Technology* (6130 eingehende Links). Tatsächlich war Eindhoven Gastgeber der Tennisweltmeisterschaft 1999, und das Fabrikgebäude „Witte Dame“ beherbergt zur Zeit eine Bibliothek. Ebenfalls ein Beispiel dafür, wie sehr sich die eigene Erwartung von den Messergebnissen unterscheidet. Es stellt sich natürlich die Frage, ob man auf die eben Beschriebene Beziehung überhaupt zurückgreifen möchte. Da die Konzepte explizit angegeben wurden, hätte auch ein Disambiguierungsschritt nichts an dem Ausgang dieses Testlaufs geändert.

[1] `http://en.wikipedia.org/wiki/Winnie-the-Pooh`

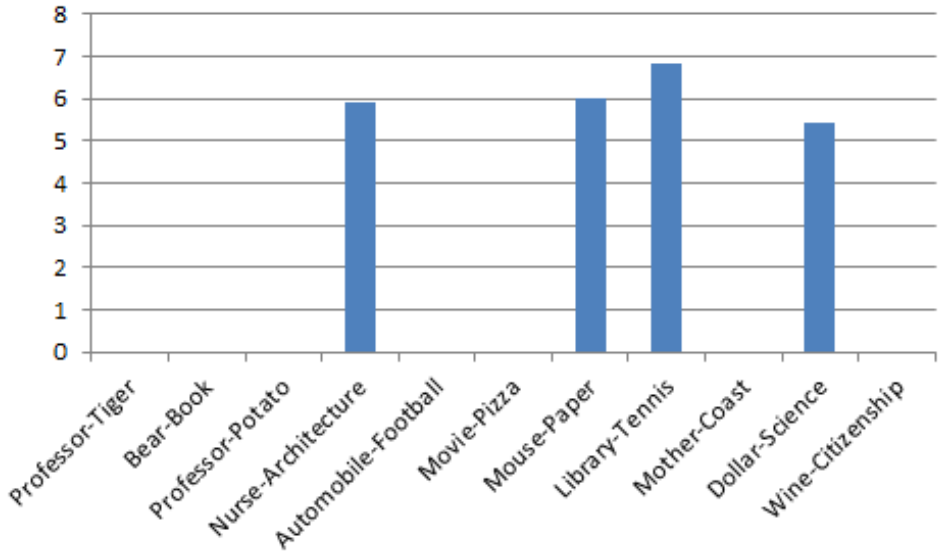

Abbildung 4.4: Varianz des GSD-Verfahrens auf Negativ-Fällen

4.4 Erweiterung durch implizite Links

Wie bereits erwähnt, arbeiten WLM und GSD auf expliziten Links. Nun stellt sich die Frage, ob die Verwendung impliziter Links in Kombination mit diesen Verfahren eine Verbesserung der Ergebnisse bewirkt. Zu diesem Zweck wurde ein Volltext-Index für die Textdatensätze angelegt, was eine schnellere Suche durch die Texte der Wikipedia-Datenbank ermöglicht.

Der große Vorteil dieser Erweiterung liegt darin, dass die jeweiligen Verfahren nicht mehr auf explizite Konzepte angewiesen sind. Es werden einfach sämtliche Konzepte nach den entsprechenden Begriffen durchsucht. Dadurch lässt sich auch die zu Beginn von Abschnitt 4.1 genannte Problematik des Datenimports und der daraus resultierenden lückenhaften Datenmenge beheben.
Gleichzeitig ist das auch die größte Schwäche dieser Erweiterung: die Notwendigkeit der Disambiguierung wird hier vollständig außer Acht gelassen. Wo in den bisherigen Implementierungen durch die Angabe expliziter Konzepte der Disambiguierungsprozess manuell durchgeführt wurde, wird durch das Einbinden impliziter Links auch eine größere Menge unpassender Artikel gefunden. Der Vergleich des Paares $(King, Rook)$ findet also Konzepte zum Titel König, zur Schachfigur, zur Musikgruppe und vie-

len mehr[2]. Inwiefern das Auswirkungen auf die Qualität der einzelnen Verfahren und deren Bewertungen hat, werden die folgenden Testläufe zeigen.

Testlauf 1

Nach Stichprobenartigen Versuchen wurde auf eine durchgehende Testreihe des WLM-Verfahrens mit impliziten Links verzichtet. Da die Ergebnisse der Volltextsuche wie in Abschnitt 3.3 genannt ein Gewicht beinhalten, sollte versucht werden, die Gewichte des WLM-Verfahrens mit diesen Volltext-Gewichten zu ersetzen. Tatsächlich gab es in 10 von 36 Fällen eine leichte Verbesserung der Ergebnisse, aber diese lagen immernoch weit unter den Expertenbewertungen der WordSimilarity-353.

Daraufhin wurde die implizite Linksuche in das GSD-Verfahren integriert. Wie in 3.2 erklärt, werden ausschließlich eingehende Links betrachtet, de ÜBernahme des Gewichts aus Formel 3.3 ist daher überflüssig. Die Menge eingehender Links zum Konzept *King* entspricht also dem Suchergebnis auf allen Konzepttexten nach dem entsprechenden Wort.

Abbildung 4.5 zeigt, dass sich implizite Links ohne den Disambiguierungsschritt nur bedingt eignen. Durch den genannten Vorteil der Unabhängigkeit expliziter Konzepte, findet die Suche nach impliziten Links zu jedem Wortpaar einen Ähnlichkeitswert, allerdings stark verfälscht. Die größte Diskrepanz ist bei dem Wortpaar $(Book, Library)$ zu erkennen. *Book* hat 11150 eingehende „Links“ *Library* 7280, doch nur 20 Konzepte haben beide Begriff gemein. Es ist insofern nicht verwunderlich, da ein Wort wie *Book* ohne Zusammenhang in vielen Konzepten zu finden ist[3], welche nicht wirklich mit einer Bibliothek in Verbindung gebracht werden können. Dennoch wäre es möglich, die über die Textsuche erlangten impliziten Links einem separaten Disambiguierungsvorgang zu unterziehen. um unpassende Verweise zu entfernen, aber gleichzeitig die Anzahl passender Treffer zu erhöhen, da vermutlich nicht alle Konzepte an den passenden Stellen von Autoren gesetzte Verweise enthalten.

Testlauf 2

Die Unterschiede zwischen den Ergebnissen der Negativ-Testfälle für GSD und das durch implizite Links erweiterte GSD sind gering, siehe Abbildung 4.6. Interessant sind an dieser Stelle die Abweichungen der Ähnlichkeitswerte. Beispielsweise hatte das Wortpaar $(Dollar - Science)$ im nor-

[2] `http://en.wikipedia.org/wiki/King_(disambiguation)`
[3] `http://en.wikipedia.org/wiki/Book_(disambiguation)`

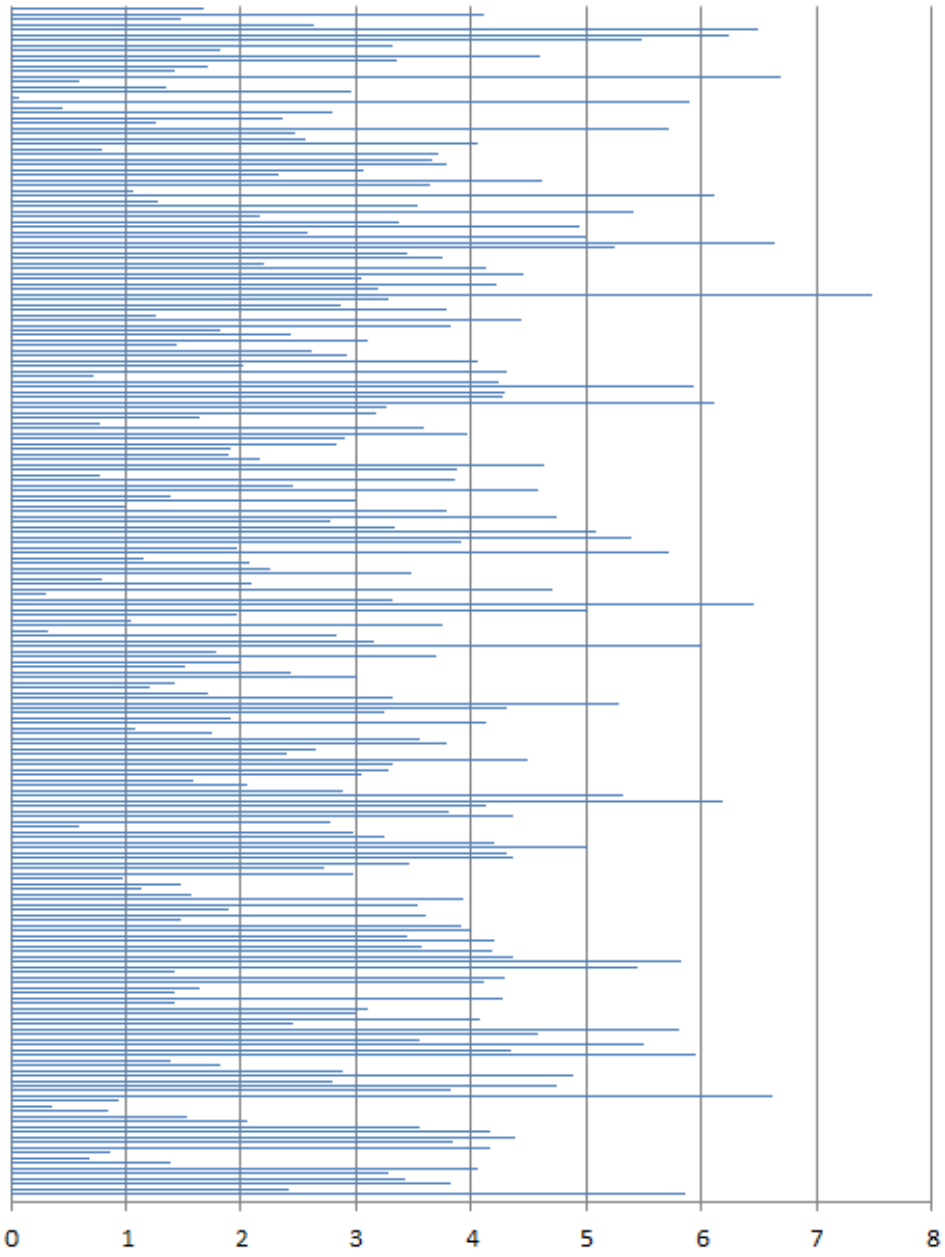

Abbildung 4.5: Varianz des GSD-Verfahrens erweitert mit impliziten Links

malen GSD-Verfahren eine Bewertung, nach Erweiterung durch implizite Links hingegen nicht mehr, bei dem Paar $(Professor, Tiger)$ entsprechend umgekehrt. Dafür gibt es zweierlei Erklärungen:

1. In der Wikipedia-Datenbanktabelle *pagelinks* fehlte einfach eine entsprechende gemeinsame Referenz zwischen den Wortpaaren, sodass keine Bewertung zustande kam. Eine Textsuche hingegen spürt diese Gemeinsamkeit auf.

2. Die unter *pagelinks* eingetragenen Links geben keine Information über ihren Ankertext, also den Text, der mit dem entsprechenden Verweis verknüpft ist. Gleichzeitig könnte dieser Link auf eine weiterleitende Seite zeigen, wohingegen die Textsuche diese Weiterleitung nicht nachverfolgen kann.

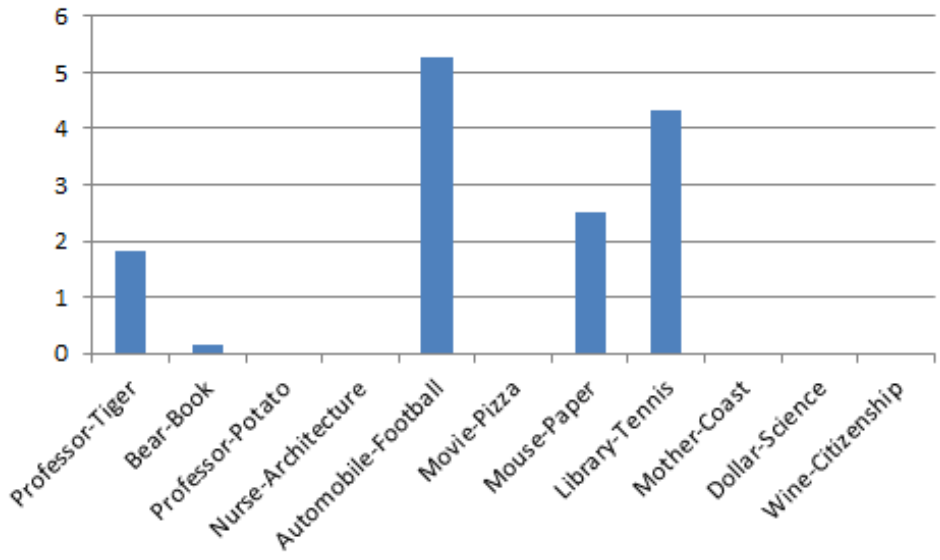

Abbildung 4.6: Varianz des GSD-Verfahrens erweitert mit impliziten Links auf Negativ-Fällen

4.5 Vergleich

Der Tabelle 4.2 ist klar zu entnehmen, dass sich das GSD-Verfahren sehr gut eignet, um eine Ähnlichkeitsbeziehung zwischen zwei Wörtern zu be-

stimmen. Die mit * markierten Spalten beinhalten kritische Datensätze. Damit sind jene Problemfälle gemeint, die auf mangelnde Konzepte zurückzuführen sind, bedingt durch sehr spezifische Begriffe aus der WordSimilarity-353, oder durch den fehlerhaften Import der Datensätze.

Es ist ganz klar zu erkennen, das GSD zwar bessere Ergebnisse liefert, aber dafür nur auf einem sehr kleinen Teil der Testmenge gearbeitet hat. Durch die Angabe expliziter Konzepte und expliziter Links wird der Testbereich ziemlich eingeschränkt.
Eine Erweiterung des GSD-Verfahrens durch implizierte Links erhöht zwar die Varianz, ermöglicht aber das Verarbeiten einer wesentlich größeren Testmenge, da eben keine Abhängigkeit zu expliziten Konzepten besteht.

Verfahren	Datensätze*	Anz. Werte ≤ 0	Varianz*	Datensätze	Varianz
WLM	61	23	5,985	38	5,665
GSD	353	254	4,465	99	**1,161**
GSD+Implicit	353	122	3,809	**231**	3,028

Tabelle 4.2: Vergleich der vorgestellten Verfahren

Kapitel 5

Fazit

5.1 Zusammenfassung

Nachdem im ersten Kapitel ein Überblick der bekannteren Verfahren hinsichtlich semantischer Interpretationen gegeben wurde, führte das zweite Kapitel einige theoretische Grundlagen und Modell vor. Der Schwerpunkt lag dabei auf potentiellen Problemen bei einer solchen semantischen Interpretation, der Wissensdatenbank Wikipedia, die für viele der bereits vorgestellten Verfahren eine essentielle Grundlage bildet, und zuletzt der WordSimilarity-353, die in diesem Bereich der Forschung häufig als Test- und Trainingsdatenmenge verwendet wird.

Im dritten Kapitel wurden die ausgewählten Verfahren und ihre Funktionsweise genauer vorgestellt, während sie im darauf folgenden Kapitel getestet und ausgewertet wurden.

Es kann festgehalten werden, dass der Gebrauch der *Google Similarity Distance* durchaus zu annehmbaren Ergebnisse bei der Berechnung semantischer Ähnlichkeit führt. Die Ergebnisse werden nochmals deutlich verbessert, wenn bei der Suche nach repräsentativen Konzepten ein Disambiguierungsschritt eingebaut wird.
Zusätzlich führt die Erweiterung durch implizite Links letztendlich dazu, dass GSD eine wesentlich größere Datenmenge abarbeiten kann und so flexibler in der Bewertung ist. Eine Kombination dieser Erweiterung mit der Disambiguierung würde sicherlich noch bessere Ergebnisse liefern.

5.2 Ausblick

Wenn Menschen sich in ihrer natürlichen Sprache schon missverstehen, wie soll dann ein Computer in der Lage sein, dieses Problem zu umgehen? Diese Arbeit hat gezeigt, dass es bis zu einem gewissen Grad möglich ist, Beziehungen einfacher Wortpaare zu bestimmen. Ein wesentliches Hindernis, das es zu überwinden gilt, ist die Disambiguierung. Wie in Abschnitt 1.2 referenziert, gibt es einige Ansätze, die sich damit auseinandersetzen, aber eine hundertprozentig fehlerfreie Lösung ist derzeit nicht gefunden. Den letzten Schritt in Anwendungen, die auf solche Verfahren aufbauen, muss also der Benutzer ausführen.

Gleichzeitig geben einem Datenplattformen wie Wikipedia sehr viele Informationen zu verschiedenen Konzepten. In dieser Arbeit außer Acht gelassen wurde zum Beispiel die Kategorie-Hierarchie, nach welcher einzelne Konzepte gruppiert werden könnten. Oder die Bilder- und Diskussionsseiten, die ebenfalls verwertbare Informationen beinhalten. Ansätze zur Kombination verschiedener Datenplattformen wie Wikipedia, WordNet und ODP scheinen hier die Lösung zu sein, denn die Stärke der einen ist gleichzeitig die Schwäche der anderen Plattformen.

Die stets sich verfeinernden Suchergebnisse wie beispielsweise von Google zeigen, wie sehr sich die semantische Analyse in den letzten Jahren fortentwickelt hat. Plattformen wie facebook[1] oder twitter[2] beinhalten ebenfalls strukturiertes Wissen, nicht nur auf sozialer, sondern auch auf organisatorischer Ebene.

Klar ist, das Wikipedia in künftigen Arbeiten dieser Art weiterhin als Informationsmedium verwendet werden wird, und sich die Querverweise innerhalb der Artikel durchaus für eine semantische Interpretation eignen.

[1] `http://www.facebook.com`

[2] `http://www.twitter.com`

Literaturverzeichnis

[1] C. H. Bennett, P. Gacs, Ming Li, M. B. Vitanyi, and W. H. Zurek. Information distance. *IEEE Transactions on Information Theory*, 44(4):1407–1423, July 1998.

[2] Alexander Budanitsky. Lexical semantic relatedness and its application in natural language processing. Technical report, University of Toronto, 1999.

[3] Rudi L. Cilibrasi and Paul M. B. Vitanyi. The Google Similarity Distance. *IEEE Trans. on Knowl. and Data Eng.*, 19(3):370–383, March 2007.

[4] Peter Clark, John Thompson, Heather Holmback, and Lisbeth Duncan. Exploiting a thesaurus-based semantic net for knowledge-based search. In *In Procs of 12th Conf. on Innovative Applications of AI (AAAI/IAAI'00*, pages 988–995, 2000.

[5] Charles E. and Osgood. The nature and measurement of meaning. *Psychological Bulletin*, 49(3):197 – 237, 1952.

[6] Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *In Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1606–1611, 2007.

[7] Rada Mihalcea and Andras Csomai. Wikify!: linking documents to encyclopedic knowledge. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242, New York, NY, USA, 2007. ACM.

[8] George A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–41, 1995.

[9] George A. Miller and Walter G. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6:1–28, 1991.

[10] David Milne, Olena Medelyan, and Ian H. Witten. Mining Domain-Specific Thesauri from Wikipedia: A Case Study. In *WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 442–448, Washington, DC, USA, 2006. IEEE Computer Society.

[11] David Milne and Ian H. Witten. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *In Proceedings of AAAI 2008*, 2008.

[12] Herbert Rubenstein and John B. Goodenough. Contextual correlates of synonymy. *Commun. ACM*, 8(10):627–633, October 1965.

[13] Mark Sanderson. Word sense disambiguation and information retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, pages 142–151, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[14] Michael Strube and Simone P. Ponzetto. WikiRelate! Computing Semantic Relatedness Using Wikipedia. July 2006.

[15] Max Völkel, Markus Krötzsch, Denny Vrandecic, Heiko Haller, and Rudi Studer. Semantic wikipedia. In *Proceedings of the 15th international conference on World Wide Web*, WWW '06, pages 585–594, New York, NY, USA, 2006. ACM.

[16] wikimedia. Wikimedia:database layout. `http://www.mediawiki.org/wiki/Manual:Database_layout` – visited last 4^{th} November 2011, November 2011.

[17] wikipedia. Wikipedia:aufbau eines artikels. `http://de.wikipedia.org/wiki/Wikipedia:Wie_schreibe_ich_gute_Artikel#Aufbau_eines_Artikels` – visited last 18^{th} October 2011, October 2011.

[18] wikipedia. Word-sense disambiguation. `http://en.wikipedia.org/wiki/Word-sense_disambiguation` – visited last 18^{th} October 2011, October 2011.

Ich versichere, dass ich diese Bachelorarbeit selbständig verfasst und nur die angegebenen Quellen und Hilfsmittel verwendet habe.

Kaiserslautern, den

. . . . . .
(Unterschrift des Kandidaten)