Bauhaus-Universität Weimar
Fakultät Medien
Studiengang Mediensysteme

Algorithmen zur Analyse und Identifizierung thematisch verwandter Webseiten im Web

Bachelorarbeit

Christiane Glimm
Geboren am 19.07.1984 in Jena

Matrikelnummer 51697

1. Gutachter: Prof. Dr. Benno Stein
2. Gutachter: Prof. Dr. Volker Rodehorst
Betreuer: Dr. Matthias Hagen

Datum der Abgabe: 1. Oktober 2013

Erklärung

Hiermit versichere ich, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Weimar, den 1. Oktober 2013

.....
Christiane Glimm

Zusammenfassung

In dieser Arbeit werden Verfahren zur Suche thematisch verwandter Dokumente entwickelt, untersucht und miteinander verglichen. Bei diesen Verfahren handelt es sich um die Link-Crawling Technik, die Key-Query Technik und die Google-Related Technik. Das erste Verfahren, die Link-Crawling Technik, arbeitet mit den Hyperlinks eines Dokumentes und benutzt diese als Verweise auf Webseiten mit potenziell thematisch verwandtem Inhalt. Die Key-Query Technik basiert auf sogenannten Keyqueries. Keyqueries sind Kombinationen aus Worten, die aus dem Anfragedokument extrahiert werden und eine wichtige Rolle im Text des Dokuments spielen. Die dritte Technik, die Google-Related Technik, verwendet die von Google bereitgestellte Option zur Suche nach verwandten Webseiten zu einem Anfragedokument bzw. Adresse. Zur Verwendung dieser Techniken wurde ein gemeinsames Interface entwickelt. Mit Hilfe von Hypothesen zu den einzelnen Verfahren wird eine Evaluierung der Techniken konzipiert und durchgeführt. Diese umfasst die Erstellung eines Korpus mit unterschiedlichen Arten von Anfragedokumenten und die manuelle Klassifikation der Suchergebnisse. Die Auswertung und Diskussion der Ergebnisse zeigt, dass die Auswahl einer geeigneten Suchtechnik direkt von der Art des Anfragedokumentes abhängt. Die Evaluierung zeigt die Anwendbarkeit der drei Verfahren, jedoch auch ihre Limitierungen und Möglichkeiten der Verbesserung.

Inhaltsverzeichnis

Kapitel 1

Einleitung

In seiner erst relativ kurzen Existenz hat sich das Internet von einer exklusiven Informationsplattform für Forschungseinrichtungen und Universitäten zu einem frei veröffentlichten Korpus von mindestens 4,42 Milliarden Webseiten[1], die inzwischen bei Google und Bing indiziert sind, entwickelt. Heute kann jede Privatperson Informationen im Web publizieren. Bekannte Suchmaschinen, wie z.B. Google, ermöglichen es gezielt bestimmte Informationen im Web zu finden. Der Nutzer formuliert dazu eine Suchanfrage, zu der die Suchmaschine eine Liste an Ergebnissen liefert. Ungleich schwerer ist die Suche, wenn der Nutzer ähnliche Inhalte zu einem Dokument sucht, die Formulierung einer entsprechenden Suchanfrage jedoch nicht eindeutig oder nicht möglich ist. Beispiele dafür sind das Finden eines Buches mit ähnlichem Inhalt oder eine automatisierte Recherche zu einem bestimmten Thema.

Ziel dieser Arbeit ist es, verschiedene Verfahren zur Suche thematisch verwandter Dokumente zu entwickeln, zu untersuchen und miteinander zu vergleichen. Ein gemeinsames Interface soll die Suche mit diesen Verfahren ermöglichen und deren Evaluierung unterstützen. Folgende zwei Schwerpunkte stehen dabei im Vordergrund. Der erste Schwerpunkt ist die Entwicklung von Algorithmen zur Suche von Dokumenten, die potenziell thematisch verwandten Inhalt zu einem Anfragedokument besitzen. Der zweite Schwerpunkt der Arbeit bildet die Evaluierung und der Vergleich dieser Suchalgorithmen.

Im Rahmen dieser Arbeit sollen drei Suchtechniken zum Einsatz kommen. Zwei der Techniken basieren auf der Analyse des Anfragedokuments. Während die erste Methode die Hyperlinks einer Webseite auswertet (*Link-Crawling* Technik), untersucht die zweite Technik (*Key-Query* Technik) den Text des Anfragedokuments. Dazu werden aus dem Text eine Menge an Schlüsselwörtern und Phrasen extrahiert, welche im Anschluss zur Bildung einer geeigneten Suchanfrage verwendet werden. Als dritte Technik wird die von Google be-

[1] `http://www.worldwidewebsize.com/` (Stand: 30.September 2013)

reitgestellte Technik der Suche nach verwandten Webseiten genutzt (*Google-Related* Technik).

Da diese drei Techniken jeweils einen unterschiedlichen Ansatz verwenden, wird die Evaluierung an einem Korpus aus verschiedenen Anfragedokumenten durchgeführt. Dieser Korpus setzt sich aus Gruppen von Dokumenten zusammen, welche jeweils exemplarisch für eine bestimmte Art von Webdokumenten stehen. Hypothesen über die Eignung der Suchtechniken bei unterschiedlichen Arten von Anfragedokumenten helfen die Evaluierung zu konzipieren. Die Ergebnisse der durchgeführten Evaluierung werden im Anschluss analysiert und diskutiert.

Die Suche nach thematisch verwandten Dokumenten im Web ist Gegenstand aktueller Forschung. Kapitel 2 gibt einen Überblick über die aktuellen Arbeiten auf diesem Gebiet. Der darauf folgende Hauptteil dieser Arbeit ist in drei Kapitel aufgeteilt. Kapitel 3 beschäftigt sich mit der Entwicklung und Implementierung der verschiedenen Suchverfahren. Danach werden in Kapitel 4 die Konzeption und Durchführung der Evaluierung beschrieben sowie die ermittelten Ergebnisse vorgestellt. Im Anschluss werden die Ergebnisse in Kapitel 5 diskutiert und ein Ausblick auf mögliche Weiterentwicklungen gegeben. Im Kapitel 6 wird die Arbeit und deren Ergebnisse zusammengefasst.

Kapitel 2

Verwandte Arbeiten

Dieses Kapitel soll einen Überblick über den Stand der Forschung über die in dieser Arbeit verwendeten Techniken geben.

Hagen und Stein [HS11] beschäftigten sich mit dem Auffinden von Dokumenten, welche potenziell identische bzw. paraphrasierte Textpassagen eines Anfragedokumentes enthalten. Dabei werden aus einem Dokument Schlüsselwörter ermittelt, die dessen Inhalt repräsentieren. Aus diesen Wörtern wird eine Suchanfrage formuliert. Die Ergebnisse der Anfrage werden einem System zur Erkennung von Textwiederbenutzung zur detaillierten Analyse übergeben. In dem Artikel steht die Suche nach einer guten Anfrageformulierung aus den Schlüsselwörtern im Mittelpunkt. Die Gruppe präsentierte eine neue Anfrageformulierungsstrategie, welche 70% der Anfragen gegenüber der vorher publizierten Strategie einspart. Die in dieser Arbeit untersuchte Key-Query Technik basiert ebenfalls auf dem Ansatz, repräsentative Schlüsselwörter für ein Dokument zu finden.

Das Verfahren von Dasdan et al. [DDKD09] befasst sich mit dem Auffinden von Dubletten (fast identische Dokumente) in einem Korpus. Um dieses Problem zu lösen, schlagen sie einen Ansatz vor, der sich aus drei Schritten zusammensetzt. Im ersten Schritt wird eine Anfragesignatur erzeugt, welche charakteristisch für das Dokument ist. Im zweiten Schritt wird diese Anfragesignatur in einem Korpus gesucht. Im letzten Schritt werden die zurückgegebenen Ergebnisse zusammengefasst und auf ihre Ähnlichkeit zum Startdokument validiert. Ihren Ansatz untersuchen sie in umfangreichen Experimenten und zeigen, dass ihr Verfahren gut mit verschiedenen Korpussen und Sprachen funktioniert. Auch dieser Ansatz ist der Key-Query Technik ähnlich.

Der Artikel von Bendersky und Croft [BC09] beschäftigt sich mit dem Auffinden von wiederverwendeten Textpassagen. Im Gegensatz zu früher publizierten Techniken zur Erkennung von Textwiederverwendungen, die auf relativ kleinen und homogenen Korpussen getestet wurden, wenden Bendersky

und Croft diese Techniken auf Websuche an. Hauptbeitrag des Artikels ist eine neue Technik, die originale Quellen von Textpassagen im Web erkennen kann. Kernelemente ihres Ansatzes sind dabei das Einsortieren der Dokumente in eine Zeitlinie und die Evolution des Textes. Zusätzlich untersuchen sie eine Technik, welche mit Hilfe einer Linkanalyse erkennt, ob es sich um eine seriöse und relevante Quelle handelt. Ihre experimentellen Resultate zeigen, dass die vorgeschlagene Technik in der Größe des Webs operieren kann, bedeutend genauere Ergebnisse liefert als eine gewöhnliche Websuche und eine Repräsentation zum Verfolgen des Informationsflusses erstellen kann.

Lee et al. [LJSL10] präsentieren einen Ansatz zum Identifizieren von sogenannten „News stories“. Ihr Ansatz basiert auf der Tatsache, dass die Anfragen in einer Blog-Suchmaschine mit einer signifikanten Häufigkeit News-Themen sind. Zentraler Beitrag dieses Artikels ist die Einbeziehung des Nutzers bzw. von Nutzeranfragen zur Klassifikation einer bestimmten Art von Webinhalten. Die in dieser Arbeit untersuchte Google-Related Technik scheint auf einem solchen Ansatz zu basieren.

Der Artikel von Fuhr et al. [FLSG11] stellt die theoretischen Grundlagen für das optimale Clustering (Gruppieren) von Dokumenten vor. Dokumente werden zu einer Gruppe zusammengefasst, wenn sie als ähnlich klassifiziert werden. Dies ist der Fall, wenn sie das Ergebnis der selben Suchanfrage sind. Ein optimales System zum Clustering von Dokumenten besteht dabei aus drei Komponenten. Die erste Komponente ist ein Set von Anfragen, die zweite Komponente ist eine wahrscheinlichkeitsbasierte Suchmethode und die dritte eine Dokumentenähnlichkeitsmetrik. Die Arbeit stellt ein Framework zum Clustern der Dokumente auf Basis der genannten Komponenten vor. Die Idee die Ergebnisse einer Suchanfrage als ähnlich zu klassifizieren ist Basis des Ansatzes der Keyqueries.

Pickens et al. [PCG10] schlagen die Technik des *reverted Indexing* vor. Bei traditionellen Suchmaschinen werden invertierte Listen verwendet. Dabei wird jedem potentiellen Suchbegriff eine Liste mit Dokumenten zugeordnet und dessen Häufigkeit und Ort im Dokument gespeichert. Bei einer Suchanfrage werden die Ergebnislisten kombiniert und die Dokumente mit der höchsten Häufigkeit zurückgegeben. In ihrem Artikel drehen Pickens et al. diese Technik um, das heißt, einem Set von Ergebnisdokumenten wird eine Anfrage zugeordnet. Damit lassen sich zu einem Dokument mehrere mögliche Suchanfragen zuordnen. Der Benutzer bekommt somit eine Möglichkeit geboten, sich Suchanfragen zu einem Anfragedokument generieren zu lassen und diese gegebenenfalls zu verfeinern oder zu verbessern. Da mit Hilfe dieses Ansatzes Suchanfragen passend zu einem Dokument generiert werden können, sollten die entsprechenden Suchanfragen bei einer Suche zu thematisch ähnlichen Dokumenten führen. Dieser Ansatz ist jedoch im Rahmen dieser Arbeit nicht realisierbar, da dies

eine Indizierung von Suchanfragen für alle Dokumente im Korpus erfordert. Da diese Arbeit sich mit der Suche im Web beschäftigt, wäre Zugriff auf den Korpus einer Suchmaschine notwendig.

Yang et al. [YBD+09] extrahieren Schlüsselphrasen aus einem Dokument. Diese Schlüsselphrasen werden mit dem Ziel, thematisch verwandte Inhalte zu einem Anfragedokument zu finden, als Anfrage an eine Suchmaschine gestellt. Yang et al. stellen zwei Verfahren zum Extrahieren dieser Phrasen vor und zeigen weitere Techniken, um diese Phrasen zu ergänzen. Hierzu werden externe Quellen wie beispielsweise Wikipedia herangezogen. In einer umfangreichen Studie zeigen sie die Effizienz und Effektivität ihrer Techniken zum automatischen Sammeln von verwandten Dokumenten.

Das Tool von Golshan et al. [GLT12], „SOFIA SEARCH“, stellt eine völlig automatische Suche und Sammlung von thematisch verwandter Literatur vor. Das System startet mit einer initialen Menge an wissenschaftlichen Artikeln, die ein Benutzer der Suchmaschine übergibt. „SOFIA SEARCH“ durchsucht das Web nach Artikeln, die verwandten Inhalt besitzen könnten, und evaluiert deren Relevanz. Das Ziel ist es, Wissenschaftler und Anwender bei der Suche nach relevanter Literatur in ihrem Themengebiet zu unterstützen. Das Tool stellt ebenfalls Module für die Evaluierung und das Ranking von Autoren und Artikeln bereit. Die Arbeit von Golshan et al. fokussiert ausschließlich auf wissenschaftliche Artikel und ist somit nicht für generelle Suche nach thematisch verwandten Dokumenten geeignet.

Einige der an dieser Stelle vorgestellten Arbeiten befassen sich mit der Erzeugung und Optimierung von Suchanfragen. Die in dieser Arbeit verwendete Technik der *Key-Query* [GHMS13] ist eine Weiterentwicklung dieser Vorgehensweise und wird in Kapitel 3.2 näher beschrieben.

Kapitel 3

Suchalgorithmen

Im Rahmen dieser Arbeit wurden drei Techniken zur Suche nach thematisch verwandten Webseiten implementiert und evaluiert. Dieses Kapitel erläutert diese Techniken im Detail. Die Technik des *Link-Crawling* wird in Kapitel 3.1 beschrieben. Im darauf folgenden Kapitel wird auf die *Key-Query* Technik näher eingegangen und in Kapitel 3.3 die Umsetzung der *Google-Related* Technik beschrieben. Aufgrund der Eigenschaften dieser verschiedenen Techniken können bereits vor der Evaluierung Annahmen über die Suchergebnisse getroffen werden. Die entsprechenden Hypothesen dazu werden in Kapitel 3.4 aufgestellt und deren Korrektheit im später folgenden Kapitel 4.3 diskutiert.

Um einem Nutzer die Verwendung der drei Suchtechniken (Link-Crawling, Key-Query und Google-Related) zu ermöglichen, wurden diese in ein Java-Programm integriert und auf einem Server zur Verfügung gestellt. In Abbildung 3.1 ist die Startseite dieses Programms und die Auswahl der Suchtechniken zu sehen. Der Nutzer gibt die Adresse (URL) des Anfragedokumentes über ein HTML-Formular in das Programm ein. Des Weiteren muss die Art der Suche eingestellt werden.

3.1 Link-Crawling Technik

Die Technik des *Link-Crawling* verwendet den gleichen Ansatz wie die ersten Suchmaschinen mit Volltextindex [ACGM+01] für das World Wide Web. Diese Technik nutzt die Hyperlinks in Dokumenten, um neue Dokumente im Web zu finden.

3.1.1 Funktionsweise

Diese Suchtechnik basiert auf der Annahme, dass sich aus einem Anfragedokument externe Hyperlinks extrahieren lassen und diese auf Webseiten oder

Abbildung 3.1: Startseite der Suche

Dokumente mit thematisch verwandtem Inhalt verweisen. Um die externen Hyperlinks zu extrahieren, ist es wichtig, das Anfragedokument zu analysieren. Die Dokumentanalyse besteht aus folgenden elementaren Arbeitsschritten: *Extraktion* und *Klassifikation*.

Extraktion Im ersten Schritt der Analyse wird der HTML-Quellcode des Anfragedokumentes auf den Server kopiert. Dieser Quellcode wird dann bezüglich des Anker-Elements (<a>) untersucht. Wird ein Anker-Element gefunden, wird dieses auf die Existenz des Hypertext-Referenz-Attributes (<a href=““>) hin überprüft. Bei erfolgreicher Suche, wird der Inhalt des href-Attributes extrahiert und an die Klassifikation übergeben.

Klassifikation Nach der Ermittlung der Hyperlinks, müssen diese noch in so genannte interne bzw. externe Hyperlinks unterteilt werden. Diese Aufteilung ist nötig, da nicht nur das Anfragedokument, sondern auch dessen Unterseiten, untersucht werden. Die Adresse des Anfragedokumentes wird in eine Domain und einen Pfad zerlegt. Dokumente in diesem Pfad werden als *interne* Hyperlinks bezeichnet. Für diese wird angenommen, dass sie auf Dokumente verweisen, die zur Ermittlung thematisch verwandter Webseiten relevant sind. *Externe* Links sind Hyperlinks, die nicht im Pfad sind oder auf eine andere Domain verweisen. Abbildung 3.2 illustriert dieses Vorgehen am Beispiel der Webseite `http://de.selfhtml.org/`.

Die gefundenen externen Hyperlinks werden als thematisch verwandte Webseiten zum Anfragedokument angenommen und als Ergebnisse ausgegeben. Die internen Hyperlinks werden rekursiv als Eingangsparameter der Routine (Extraktion und Klassifikation) übergeben. Diese werden ebenfalls auf interne bzw. externe Hyperlinks hin untersucht (siehe Abbildung 3.3). Endlosschlei-

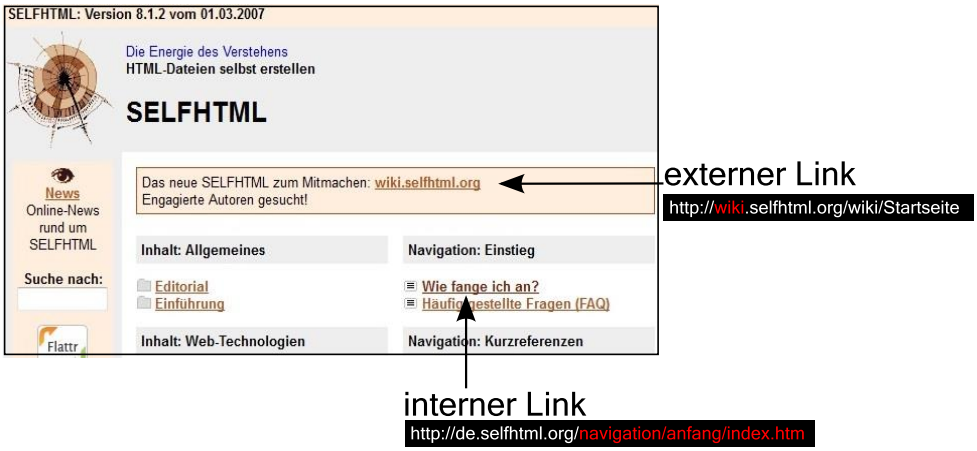

Abbildung 3.2: Diese Abbildung zeigt die Startseite von „SELFHTML“(„`http://de.selfhtml.org`“). Bei dem Hyperlink „Wie fange ich an?“ handelt es sich um einen internen Link, da ausgehend vom Wurzelverzeichnis in den Unterpfad „/navigation/anfang/index.htm“ verlinkt wird. Der Hyperlink „wiki.selfhtml.org“ ist ein externer Link, da auf eine andere Domain verwiesen wird.

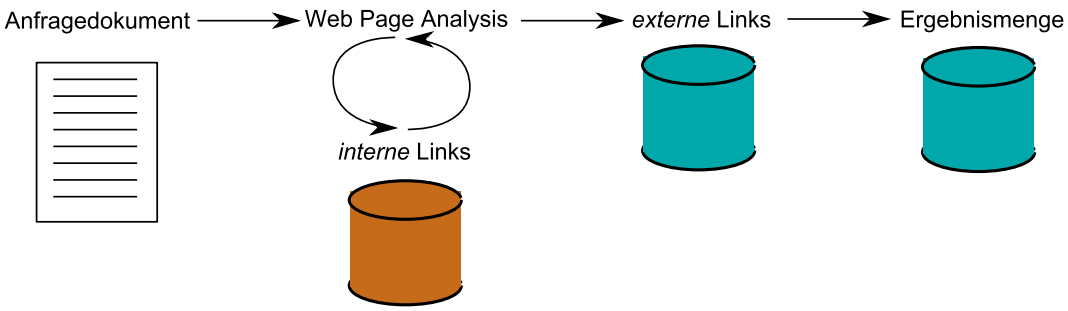

Abbildung 3.3: Algorithmus der Link-Crawling Technik. Der Arbeitsschritt *Web Page Analysis* beinhaltet die Extraktion und Klassifikation der Hyperlinks.

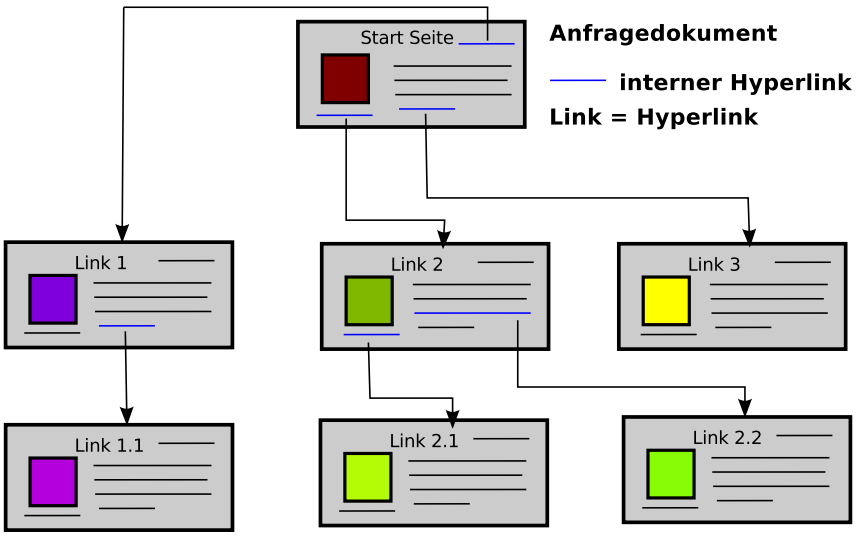

Abbildung 3.4: Suchgraph der Link-Crawling Technik

fen werden vermieden, indem interne Hyperlinks gesammelt und nur einmal untersucht werden. Der Algorithmus endet, wenn alle internen Hyperlinks untersucht wurden oder die Anzahl externer Hyperlinks ein gewünschtes Maximum erreicht hat. Diese Obergrenze wurde im Rahmen dieser Arbeit auf 200 Hyperlinks festgesetzt, ist jedoch beliebig anpassbar.

Aufgrund der rekursiven Vorgehensweise ergibt sich automatisch eine Reihenfolge der Ergebnisse, die im Allgemeinen nicht dem Grad der Ähnlichkeit entspricht. Dennoch ergibt sich eine Anordnung, welche der Tiefe der Verlinkung der Dokumente entspricht. Da alle gefundenen internen Hyperlinks im FIFO-Prinzip (*first in, first out*) abgearbeitet werden, entspricht dies einer *Breitensuche* im Suchgraph, wie in Abbildung 3.4 illustriert. Unter der Annahme, dass sich mit der Tiefe der Verlinkung auch der Inhalt ändert bzw. spezialisiert, kann jedoch davon ausgegangen werden, dass die Relevanz der Ergebnisse mit der Tiefe abnimmt.

3.1.2 Implementierungsdetails und Limitierungen

Nach der erläuterten Verfahrensweise wurde die Link-Crawling Technik implementiert. Dies erfordert die Behandlung von vielen Spezialfällen, die im Rahmen dieser Arbeit nicht alle auf Grund des enormen Zeitaufwandes, umgesetzt wurden. Diese Spezialfälle betreffen virtuelles Hosting, die unterschiedlichen Vorkommensformen von Hyperlinks und spezielle Webseitenprogrammierung. Auf einige Spezialfälle und Limitierungen wird in folgenden Abschnitten genauer eingegangen.

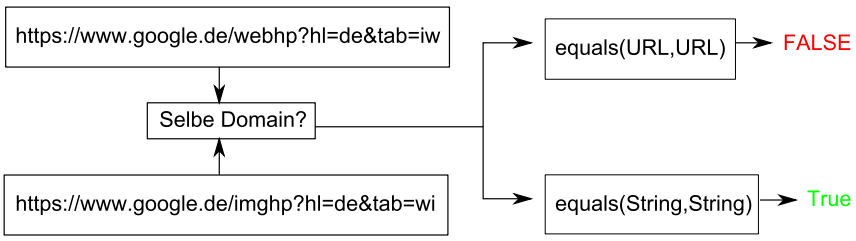

Abbildung 3.5: Probleme bei der Bestimmung der Domain. Durch die Benutzung von virtuellen Hosts ist es nicht möglich, zwei URL Adressen als identisch mit der equals()-Funktion der URL-Klasse von Java zu identifizieren, da diese Funktion die IP-Adressen der Hosts vergleicht auf denen die Dokumente abgerufen wurden. Durch virtuelles Hosting kann es aber passieren das zwei identische Webseiten von unterschiedlichen virtuellen Hosts abgerufen wurden.

Durch virtuelles Hosting ist es möglich, dass zwei Adressen mit der gleichen Domain in unterschiedlichen IP-Adressen aufgelöst werden können. Da die *equals()*-Funktion der verwendeten Klasse java.net.URL die Adressen der zu vergleichenden Domain vor dem Vergleich auflöst, können scheinbar gleiche Hyperlinks als ungleich bewertet werden [Ora12]. Auch die alternative *sameFile()*-Funktion löst dieses Problem nicht, da sie auf erstere Methode zurückgreift. Die Ähnlichkeit von Adressen wird daher ausschließlich über einen String-Vergleich gelöst.

Für die Klassifikation von internen Hyperlinks sind einige Spezialfälle zu beachten, da das Suchen der Adresse des Anfragedokuments in einem gefundenen Hyperlink nicht ausreichend ist. Interne Hyperlinks können auch relativ zum Wurzelverzeichnis oder einer beliebigen Basis angegeben werden. Das HTML-Dokument wird daher auf die Angabe einer Basis hin untersucht und diese an den relativen Hyperlink angefügt. Das Resultat wird der Klassifikation nach internen oder externen Hyperlinks übergeben.

Adressen können beliebige Zusatzinformationen, zum Beispiel Weiterleitungshinweise oder Quelladressen, enthalten. Da diese Adressen jedoch zum gleichen Dokument verweisen können, wurde auch dafür eine Behandlung implementiert. In Abbildung 3.1 ist die Abfrage zur Unterscheidung in interne und externe Hyperlinks dargestellt. Im letzten Block der Abbildung ist die if-Abfrage „befindet sich ein '= http'oder ein '- https' in dem Link“ zu sehen. Dies behandelt den Fall, dass Hyperlinks mit Weiterleitungshinweisen versehen sind und die Zieladresse verwendet werden soll.

Die Verarbeitung von alternativen Formen des Dokumentaufbaues, wie z. B. HTML-Frames, Flash und JavaScript, wurde auf Grund der Komplexität einer vollständigen Analyse nicht implementiert. Eine Untersuchung, an erster Stelle von JavaScript-Code, ist wünschenswert und könnte in einer zukünftigen

Arbeit umgesetzt werden. Die Verwendung von HTML-Frames sowie Flash als Gestaltungsformen von Webseiten ist relativ selten und veraltet. In den durchgeführten Tests bestätigte sich diese Annahme und auch wenn dies theoretisch zu fehlenden Ergebnissen führen kann, wurde auf eine Analyse von HTML-Frames und Flash verzichtet.

```
if(Adresse des Anfragedokumentes im Hyperlink){
  if("=http" oder ein "-https" im Hyperlink)
  {Hyperlink ist intern}
  else{Hyperlink ist extern}}
else{
  if("http" oder "https" im Hyperlink)
  {Hyperlink ist extern}
  else{
    if("/" an erster Stelle im Hyperlink){
      if(Adresse des Anfragedokumentes in
         zusammengesetzten Hyperlink)
      {Hyperlink ist intern}
      else{Hyperlink ist extern}}
    else{Hyperlink ist intern}}}
```

Algorithmus 3.1: Pseudocode zur Klassifizierung von internen und externen Hyperlinks

Ebenfalls ignoriert werden Adressen, die einen Client-Fehler verursachen, wie z. B. inaktive Hyperlinks und unautorisierter Zugriff. Da diese Webseiten nicht existieren und nicht analysiert werden können, stellt dies keine Limitierung der Suchtechnik dar.

Unter Beachtung der beschriebenen Lösungen und Limitierungen hat sich die Technik des Link-Crawling als ein störunanfälliger Algorithmus bewährt. Die Methode findet robust vorhandene Hyperlinks und verfolgt unter Beachtung genannter Spezialfälle alle internen Hyperlinks zur weiteren Analyse.

3.2 Key-Query Technik

Die zweite Technik basiert auf der Hypothese, dass zwei Dokumente thematisch verwandt sind, wenn sie beide Ergebnis der selben Suchanfrage (*query*) sind [FLSG11]. Unter dieser Annahme sind alle Ergebnisse einer Suchanfrage an eine Web-Suchmaschine, z.B. Google oder Bing, thematisch miteinander verwandt. Eine Suchanfrage besteht im Allgemeinen aus einem oder mehreren Schlüsselwörtern (*keywords*). Die Grundidee der sogenannten Key-Query Technik ist es, aus einem Anfragedokument die Schlüsselwörter zu extrahieren, mit

denen sich eine Suchanfrage formulieren lässt, die als Ergebnis das Dokument selbst enthält [GHMS13], den sogenannten *Keyqueries*. Alle weiteren Suchergebnisse der Suchanfrage mit den extrahierten Schlüsselworten sind, unter der im ersten Satz genannten Hypothese, thematisch mit dem Anfragedokument verwandt. Im weiteren Verlauf dieser Arbeit bezeichnet der Begriff „Keyquery“ bzw. „Keyqueries“ die Schlüsselwörter einer Suchanfrage, während der Begriff „Key-Query Technik“ das gesamte Suchverfahren, welches Keyqueries verwendet, benennt.

3.2.1 Funktionsweise

Ausgehend von einem Anfragedokument kann die Key-Query Technik in zwei Arbeitsschritte unterteilt werden. Der erste Schritt erläutert das Extrahieren der Keyqueries aus dem Anfragedokument, wie im nachfolgenden Abschnitt beschrieben. Im zweiten Schritt werden diese Keyqueries verwendet, um thematisch verwandte Dokumente zu suchen, wie im Abschnitt „Auffinden der thematisch verwandten Webseiten“ näher erläutert wird.

Bildung der Keyqueries

Keyqueries sind Worte oder Phrasen, die bei einer Anfrage an eine Suchmaschine das Anfragedokument unter den ersten r Ergebnissen zurück gibt. Im Allgemeinem ist die Variable r frei wählbar. In der aktuellen Implementierung ist r auf 10 festgelegt, d.h. das Anfragedokument muss sich unter den ersten 10 Ergebnissen befinden, am Beispiel der Suchmaschine Google folglich auf der ersten Ergebnisseite. Um die Keyqueries eines Dokumentes zu bilden, wird in einem ersten Schritt der Text des Ausgangsdokumentes extrahiert. Bilder, Textformatierung und Metainformationen werden verworfen. Der extrahierte Text wird im Anschluss mit einem Part-of-Speech Tagger auf eine beschränkte Anzahl von Wortarten reduziert.

Ein Part-of-Speech Tagger [Bri92] ist ein Werkzeug, welches Texte in Worte zerlegt, diese bezüglich ihrer Wortart klassifiziert und filtert. Bei der Technik der *Keyqueries* werden folgende 3 Wortarten beibehalten: Substantive (inklusive zusammengesetzte Substantive), Adjektive und Konjunktionen. Alle anderen Wörter werden aus dem Text entfernt. Außerdem werden alle Adjektive gelöscht, die nicht in direkter Verbindung mit einem Substantiv stehen. In Abbildung 3.6 sind die ersten beiden Schritte bei der Ermittlung der Keyqueries dargestellt.

Mit den verbleibenden Wörtern aus dem Ausgangstext wird dann mit Hilfe des TextRank-Algorithmus von Mihalcea und Tarau [MT04] ein Graph modelliert. Die Wörter bilden die Knoten w_i des Graphs. Den Kanten zwischen den

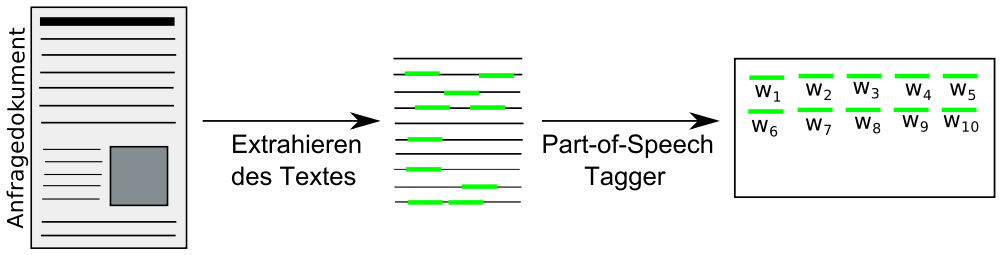

Abbildung 3.6: Erste Schritte bei der Ermittlung der Keyqueries. Im ersten Schritt wird der Text eines Dokumentes extrahiert, um danach mit Hilfe eines Part-of-Speech Taggers den Text auf wichtige Bestandteile zu reduzieren.

Knoten w_i und w_j wird eine Gewichtung $k_{i,j}$ zugewiesen, die sich wie folgt berechnet:

$$k_{i,j} = 1/(1+e) \tag{3.1}$$

Die Distanz e ist ein Maß für die Entfernung zwischen den Wörtern w_i und w_j, d.h. für die Anzahl der Wörter, die zwischen den zwei Wörtern stehen. Für die Ermittlung dieser Distanz wird das Ausgangsdokument verwendet. Die Initialisierung der Knoten erfolgt mit einer Gewichtung $g_i = 1$. Somit entsteht ein gerichteter und gewichteter Graph, wie in Abbildung 3.7 gezeigt. Zur Ermittlung der finalen Gewichte der Knoten wird der iterative PageRank-Algorithmus [PBMW99] auf den Graph angewendet. In einem Iterationsschritt dieses Algorithmus berechnet sich das Gewicht für jeden Knoten in Abhängigkeit zu seinen Nachbarknoten neu. Formel 3.2 zeigt die Berechnungsvorschrift für einen Knoten pro Iteration. Der Parameter d ist ein Dämpfungsfaktor und wird auf 0.85 festgelegt.

$$g_i = \frac{1}{A(w_i)} \sum_{j}^{A(w_i)} k_{i,j} \cdot d \cdot g_j \tag{3.2}$$

Das Gewicht des Knotens g_i berechnet sich wie folgt: über alle Nachbarknoten wird das Gewicht der Kante $k_{i,j}$ mit dem Dämpfungsfaktor d und dem Gewicht des eingehenden Knotens g_j multipliziert und aufsummiert. Das Ergebnis dividiert durch die Anzahl der Nachbarknoten ergibt eine Normalisierung der Knotengewichte. Im Allgemeinen konvergieren die Knotengewichte nach einer bestimmten Anzahl von Iterationsschritten. Das bedeutet, dass sich die Gewichte der Knoten pro Iteration nur noch minimal ändern. Wenn eines der zwei folgenden Kriterien erfüllt ist, wird der Algorithmus zur Berechnung der Knotengewichte abgebrochen. Das erste Kriterium ist erfüllt, wenn sich das Ergebnis der Knotengewichte nicht mehr als $\epsilon = 10^{-5}$ ändert, d.h. alle Kno-

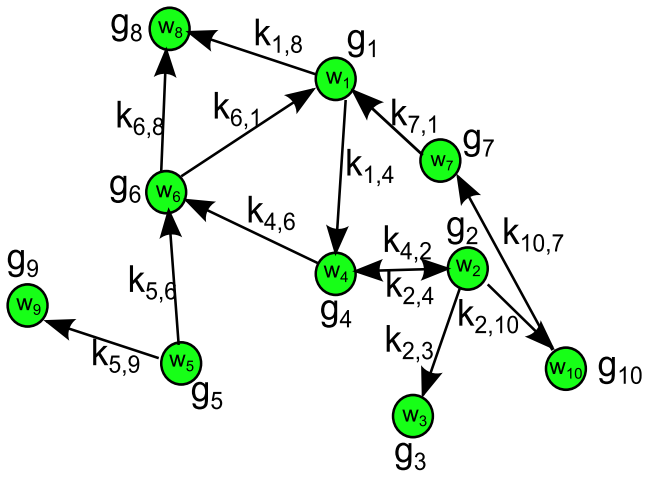

Abbildung 3.7: Gewichteter Ausgangsgraph

tengewichte konvergiert sind. Das zweite Kriterium bricht das Verfahren ab, falls der Algorithmus nach 100 Iterationen noch nicht konvergiert ist.

Im resultierenden Graph ist jedem Knoten ein Gewicht zugeordnet, welches ein Maß für die Wichtigkeit des entsprechenden Wortes für das Ausgangsdokument ist. Da bestimmte Wörter semantisch in einer bestimmten Reihenfolge zusammengehören, jedoch bei der Generierung des Graphen voneinander getrennt wurden, bietet es sich an, diese wieder zusammenzufassen. Abbildung 3.8 zeigt ein Beispiel für die unterschiedlich semantische Bedeutung gleicher Worte in Abhängigkeit von ihrer Reihenfolge.

Das Identifizieren dieser zusammengehörenden Wortgruppen, im Folgenden als *Keyphrasen* bezeichnet, erfolgt über das Gewicht der Kanten. Wörter, die im Ausgangsdokument direkt nebeneinander stehen, sind im Graph auf zwei Knoten w_i,w_j abgebildet, welche über eine Kante mit dem Gewicht $k_{i,j} = 1$ verbunden sind. Zur Bildung von Keyphrasen werden alle Kanten mit dem Gewicht 1 gesucht und durch das Zusammenfassen der dazugehörigen Knoten eliminiert (Abbildung 3.9). Dem resultierenden Knoten w'_i werden die Wörter der Ausgangsknoten w_i,w_j entsprechend der Richtung der eliminierten Kante zugewiesen. Das Gewicht g'_i von Knoten w'_i ergibt sich aus der der Summe der Gewichte der Ausgangsknoten w_i,w_j.

Durch das Eliminieren von Kanten können Knoten entstehen, die aus mehreren Wörtern bestehenden Keyphrasen beinhalten. Im Rahmen dieser Arbeit wurden Keyphrasen auf eine maximale Wortanzahl von 3 beschränkt. Nicht zusammengefasste Knoten beinhalten ein Wort, welches im Folgenden auch als *Keyword* bezeichnet wird. Die Knoten des resultierenden Graphen werden anhand ihrer Gewichtung sortiert. Die sieben Keyphrasen bzw. Keywords mit dem höchsten Gewicht werden zur Generierung von Keyqueries genutzt.

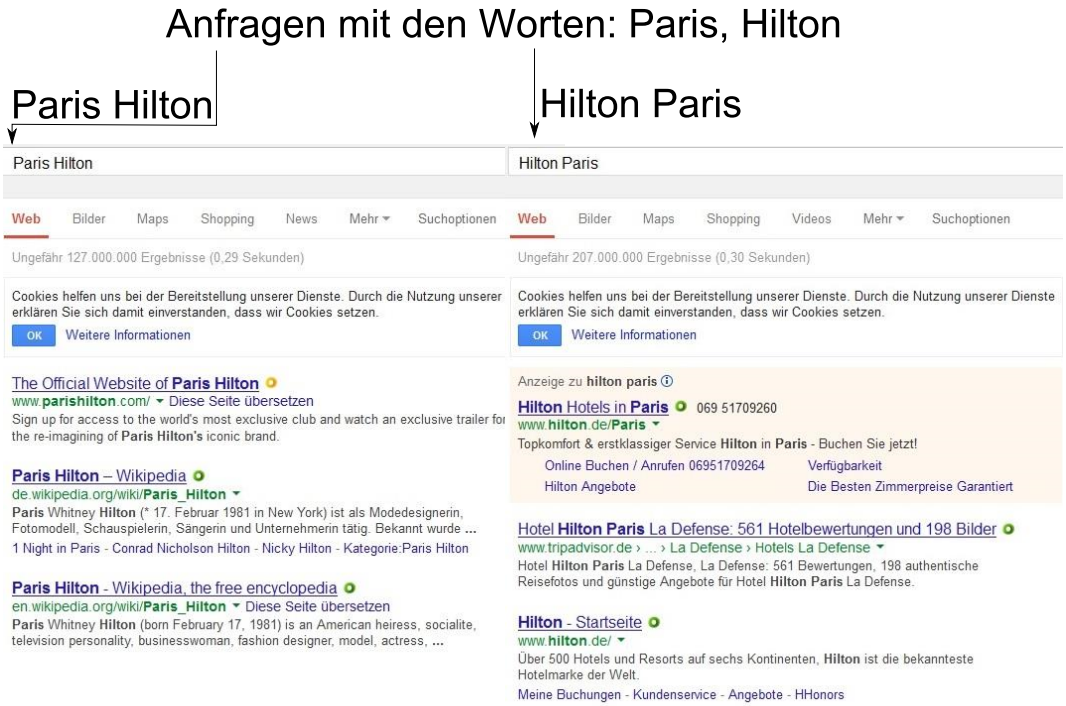

Abbildung 3.8: Die Abbildung zeigt ein Beispiel für die unterschiedliche Bedeutung von Wortpaaren, wenn deren Reihenfolge geändert wurde.

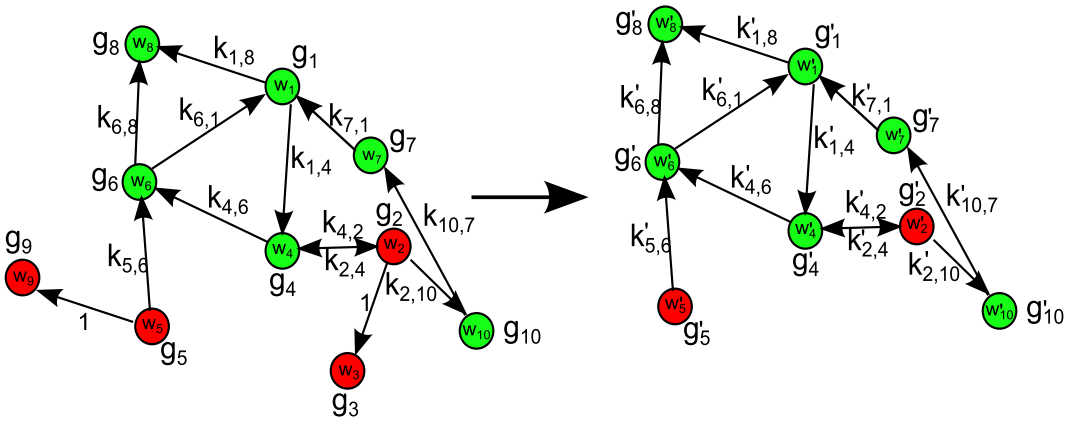

Abbildung 3.9: Generierung von Keyphrasen. Die Knoten w_5 und w_9 bzw. w_2 und w_3 sind jeweils über Kanten mit dem Gewicht 1 verbunden und werden zusammengefasst.

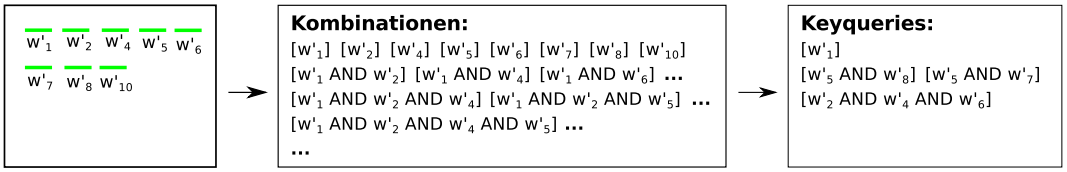

Abbildung 3.10: Erstellung der Keyqueries aus den 7 Knoten mit dem höchsten Gewicht des Graphen.

Keyqueries sind alle Kombinationen aus diesen 7 Keywords bzw. Keyphrasen, die bei einer Suchanfrage das Anfragedokument unter den ersten 10 Ergebnissen zurückgeben. Die verwendeten Kombinationen enthalten keine Wiederholungen und unterschiedliche Permutationen werden nicht berücksichtigt. Zu Beginn werden die Kombinationen getestet, die nur aus einem Keyword bzw. Keyphrase bestehen. In weiteren Schritten werden Kombinationen aus einer steigenden Anzahl (2,3,4 usw.) von Keywords bzw. Keyphrasen gebildet. Kombinationen, die bereits klassifizierte Keyqueries enthalten, werden ausgeschlossen. Ebenfalls ausgeschlossen werden Kombinationen, die keine Suchergebnisse zurückliefern. Das Verfahren stellt Suchanfragen, bis keine weiteren Kombinationen unter genannten Einschränkungen gebildet werden können. Das Ergebnis ist eine Liste von Keyqueries für das Anfragedokument.

Auffinden der thematisch verwandten Webseiten

Mit Hilfe der ermittelten Keyqueries des Anfragedokumentes werden in diesem Schritt thematisch verwandte Webseiten ermittelt. Die einzelnen Keyqueries werden als Suchanfrage an eine Websuchmaschine gestellt. Die Suchergebnisse der einzelnen Suchanfragen werden zu einer Ergebnisliste zusammengefasst. Das Anfragedokument selbst wird aus dieser Ergebnisliste herausgefiltert. Laut Anfangshypothese dieses Ansatzes sind die Dokumente der Ergebnisliste thematisch verwandt mit dem Anfragedokument.

3.2.2 Implementierungsdetails und Limitierungen

Die Implementierung der Suchanfrage wurde mit Python realisiert. Als Websuchmaschine wurde Google verwendet. Die Zielmenge für die Anzahl der Dokumente in der Ergebnisliste wurde auf 200 Dokumenten festgelegt. Da die Suchergebnisse aller Keyqueries gleichberechtigt berücksichtigt werden sollen, wurde von jeder Keyquery die gleiche Anzahl an Ergebnissen verwendet. Diese Anzahl ergibt sich durch die Division der gewünschten Zielmenge von 200 Dokumenten mit der Anzahl der gefundenen Keyqueries.

Bei dieser Suchtechnik ergeben sich Limitierungen hauptsächlich durch die

verwendete Suchmaschine. Google versucht zu verhindern, dass Suchanfragen automatisch gestellt werden. Erkennt Google, dass die Suchanfrage automatisch durch eine Software gestellt wurde, wird die IP des Rechners von dem die Anfrage gestellt wurde, gesperrt. Weitere Anfragen von diesem Rechner werden blockiert und liefern keine Ergebnisse. Diese Sperrung dauert bis zu ca. 6 Stunden. Dieser Wert ist ein Erfahrungswert, da Google keine Informationen zur Dauer der Sperrung gibt. Eine weitere Beobachtung ist, dass die selbe Anfrage, oft hintereinander gestellt, ebenfalls zu einer Sperrung führen kann. Motivation dieser Vorgehensweise ist es vermutlich, dass sich Google davor schützen will, dass die Suchmaschine missbräuchlich genutzt wird, d. h. Informationen automatisiert ausgelesen werden.

Die Wahrscheinlichkeit einer Sperrung konnte reduziert werden, in dem die Anfragen mit einem festen Zeitabstand gestellt wurden. Ziel dieser verzögerten Anfragen ist es, einen menschlichen Nutzer vorzutäuschen und Google die Erkennung eines automatischen Vorganges zu erschweren. Die Wartezeit zur Verzögerung der Anfragen wurde auf 6 Sekunden festgesetzt. Nachteil dieser Vorgehensweise ist die gesteigerte Gesamtdauer aller Anfragen. Da diese an mehreren Stellen dieser Suchtechnik verwendet werden, u.a. bei der Ermittlung der Keyqueries, erhöht sich die Laufzeit stark. Die theoretische maximale Dauer zur Ermittlung der Keyqueries hängt von der Anzahl möglicher Kombinationen $K = 127$ (ohne Wiederholung) der 7 Keywords bzw. Keyphrases ab. Diese ergibt sich wie folgt:

$$K = \binom{7}{1} + \binom{7}{2} + \binom{7}{3} + \binom{7}{4} + \binom{7}{5} + \binom{7}{6} + \binom{7}{7} \tag{3.3}$$

Da zwischen jeder Anfrage einer Kombination eine Wartezeit von 6 Sekunden liegt, ergibt sich eine *worst-case*-Abschätzung für die Laufzeit der Keyquery-Ermittlung von $126 \cdot 6s = 756s$, also etwa 13 Minuten.

Die detaillierten Laufzeiten aus der Evaluierung des Algorithmus sind in Kapitel 4.3.2 beschrieben. Eine alternative Lösung wäre die Benutzung einer anderen Suchmaschine, die nicht über diese Limitierung verfügt. Dies würde die Laufzeit der Key-Query Technik enorm beschleunigen.

3.3 Google-Related Technik

Die dritte Technik für die Suche nach thematisch verwandten Webseiten ist die von Google bereitgestellte Funktion [1] zur Suche nach verwandte Webseiten. Über die Arbeitsweise dieser Technik gibt es leider keine Veröffentlichungen

[1] `http://www.googleguide.com/advanced_operators_reference.html` (Stand: 11.September 2013)

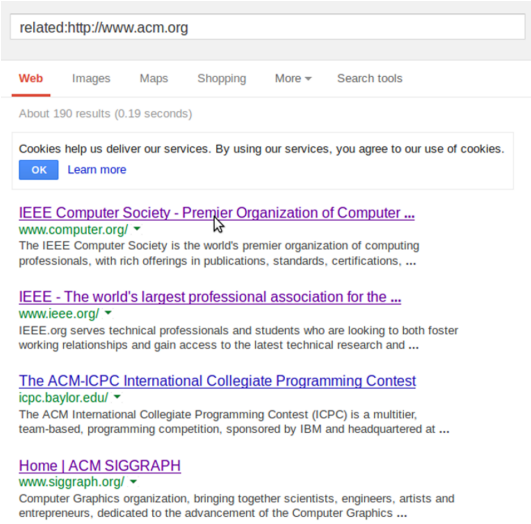

Abbildung 3.11: Beispiel einer Google-Suche mit dem Schlüsselwort „related:“

oder sonstige Informationen. Daher lassen sich keine konkreten Aussagen dazu machen, auf welcher Grundlage Google entscheidet, welche Dokumente als ähnlich eingestuft werden.

3.3.1 Funktionsweise

Zur Nutzung von Google's Suche nach ähnlichen Webseiten, muss vor die Adresse des Anfragedokumentes das Schlüsselwort „related:“ ohne Leerzeichen angefügt werden. Bei einer so gestellten Suchanfrage gibt Google eine Ergebnisliste mit ähnlichen Webseiten zurück. In Abbildung 3.11 ist beispielhaft eine solche Suchanfrage mit der Adresse `http://www.acm.org` gezeigt.

3.3.2 Implementierungsdetails und Limitierungen

Die Integration dieser Suchtechnik in das gemeinsame Interface, welches in Abbildung 3.1 dargestellt ist, wurde mit der Script-Sprache Python realisiert. In der aktuellen Implementierung werden pro Suchanfrage maximal die ersten 200 Dokumente der Ergebnisliste verwendet.

Die Haupteinschränkung dieser Technik ist, wie auch bei der Key-Query Technik, die Sperrung durch Google. Die Details dieser Limitierung wurden bereits in Kapitel 3.2.2 ausführlich diskutiert. Der Umfang der Suchanfragen ist jedoch wesentlich kleiner als bei der Key-Query Technik, da keine Suchanfragen zur Verifizierung der Keyqueries gestellt werden müssen.

3.4 Hypothesen

In den Kapiteln 3.1 bis 3.3 wurde ausführlich auf die Methoden eingegangen, die zum Auffinden von thematisch verwandten Dokumenten implementiert bzw. verwendet wurden. Zur Konzeption der Evaluierung dieser Algorithmen wurden Hypothesen über die Eignung dieser Verfahren für verschiedene Anwendungsfälle aufgestellt. Diese Hypothesen basieren auf der grundlegenden Funktionsweise der Ansätze.

3.4.1 Hypothesen zur Link-Crawling Technik

Die Technik des Link-Crawling basiert auf der Analyse von Hyperlinks im Anfragedokument. Aus dieser starken Abhängigkeit ergeben sich folgende Hypothesen für diesen Ansatz:

- Auf Grund vieler gegenseitiger Verweise untereinander, eignet sich die Link-Crawling Technik für stark vernetzte Gemeinschaften, wie z.B. politische Gruppen, Vereine und andere Interessensgruppen.
- Webauftritte von kommerziellen Anbietern, wie z. B. Shops oder Nachrichtenportale, eignen sich wegen ihrer Konkurrenzsituation nicht.
- Die in dieser Arbeit beschriebene Link-Crawling Technik liefert keine Ergebnisse für wissenschaftliche Artikel, da diese statt Hyperlinks eine andere Form der Referenzierung verwenden, welche im Rahmen dieser Arbeit nicht implementiert wurde.

3.4.2 Hypothesen zur Key-Query Technik

Dieser Ansatz basiert auf der Analyse von Fließtext. Der Inhalt des Anfragedokumentes sollte sich dazu eignen, die für Keyqueries notwendigen Schlüsselworte zu extrahieren. Daraus werden folgende Hypothesen abgeleitet.

- Die Key-Query Technik eignet sich für wissenschaftliche Artikel, da diese im Allgemeinen ausreichend Fließtext mit einem starken Fokus auf ein spezielles Thema beinhalten.
- Die Generierung geeigneter Keyqueries von Nachrichtenportalen ist auf Grund ihrer großen thematischen Vielfalt nur eingeschränkt möglich.
- Für monothematische Webblogs sind gute Resultate zu erwarten. Blogs, deren Inhalt sich auf subjektive Erlebnisse beschränkt, sind ungeeignet.

3.4.3 Hypothesen zur Google-Related Technik

Auf Grund der fehlenden Kenntnis über die Funktionsweise von Google's Suche nach ähnlichen Webseiten, wurden die Hypothesen zu dieser Technik auf Grundlage von Pilottests aufgestellt. Google verwendet vermutlich zusätzlich zum Suchkorpus gesammelte Informationen zum Nutzerverhalten und kann so Rückschlüsse auf ähnliche Inhalte ziehen.

- Häufig besuchte Webseiten, wie Nachrichtenportale oder beliebte Webblogs, eignen sich für die Google-Related Technik.
- Dokumente, die Spezialthemen für einen stark eingeschränkten Nutzerkreis beinhalten, sollten keine zufriedenstellenden Ergebnisse liefern. Beispiele hierfür sind wissenschaftliche Artikel.

Kapitel 4

Evaluierung

Auf Grundlage der Hypothesen, welche im vorangegangenen Kapitel 3.4 aufgestellt wurden, wurde die Evaluierung der Techniken konzipiert und umgesetzt. Dies umfasst u.a. die Erstellung eines geeigneten Korpus von Anfragedokumenten. Dieser Schritt wird in Kapitel 4.1 näher beschrieben. Anhand des erstellten Korpus wurde ein Testablauf geplant und durchgeführt. Dieser ist in Kapitel 4.2 dargestellt. Die Ergebnisse der Tests werden in Kapitel 4.3 vorgestellt und diskutiert.

4.1 Erstellung des Korpus

Auf der Grundlage der Hypothese wurden verschiedene Arten von Webseiten zur Evaluierung der Techniken ausgewählt. Zu jeder dieser Arten wurden mögliche Anfragedokumente gesammelt und zu einem Korpus zusammengefasst. Folgende Arten sind in dem Korpus enthalten: wissenschaftliche Artikel, Nachrichten-Seiten im Web und Webseiten mit rechtsextremistischem Hintergrund.

Im Korpus stehen wissenschaftliche Artikel exemplarisch für alle Dokumente mit spezieller Thematik, welche viel Fließtext enthalten und für einen kleinen Nutzerkreis interessant sind. Als weitere Art wurde eine Menge von Nachrichten-Seiten dem Korpus hinzugefügt. Diese repräsentieren alle Dokumente mit vielfältigem Inhalt und kurzen Texten. Stellvertretend für stark vernetzte Interessengruppen wurden Webseiten mit rechtsextremistischem Hintergrund gesammelt und in den Korpus mit aufgenommen.

Der Begriff Rechtsextremismus beschreibt kein homogenes ideologisches Konzept. Es gibt dazu keine einheitliche Definition [Nan08]. In dieser Arbeit wird der Begriff als eine Gesinnung verwendet, die einen übersteigerter Nationalismus, Fremdenfeindlichkeit, Antisemitismus, ein autoritär-konservatives, hierarchisches Familien- und Gesellschaftsbild und die Ablehnung der De-

mokratie in ihren Grundzügen unterstützt. Der Begriff löste den Begriff des Rechtsradikalismus ab und ist seit den 70 Jahren ein Begriff der politischen Alltagssprache geworden [Jas06]. Die Auswahl dieser Interessengruppe ist zusätzlich motiviert durch den Bedarf, neue Inhalte innerhalb dieser Gruppe automatisiert zu finden. Dies ist notwendig, da die Inhalte möglicherweise verfassungsfeindlich sind oder zum Zwecke des Jugendschutzes beobachtet werden sollten. Diese ideologische Gruppe drängt verstärkt in das Web, wie *jugendschutz.net*[1] in ihrem Bericht von 2013 darstellt[2]. Da Rechtsextremismus in vielen Erscheinungsformen im Web auftritt, wurden exemplarisch zwei Arten von Webrepräsentationen ausgewählt: Webblogs (Hauptvertreter sind freie Gruppen und Kameradschaften) und öffentliche, politische Webseiten am Beispiel der NPD.

Der Inhalt des entstandenen Korpus kann wie folgt zusammengefasst werden:

- Wissenschaftliche Artikel
- Nachrichten-Seiten
- Webblogs (mit rechtsextremistischem Hintergrund)
- NPD-Seiten

Aus jeder dieser vier Kategorien wurden 25 Dokumente exemplarisch als Anfragedokumente verwendet. Somit ergibt sich ein Testkorpus von 100 Anfragedokumenten, auf dem die Evaluierung der drei Suchtechniken durchgeführt wurde.

4.2 Testablauf

Die Evaluierung der Suchtechniken (Link-Crawling, Key-Query und Google-Related) erfolgte mit Hilfe des Testkorpus, welcher in Kapitel 4.1 beschrieben wurde. Für jedes der 100 Anfragedokumente aus dem Korpus wurde jede Suchtechnik angewendet. Dadurch ergaben sich eine Anzahl von insgesamt 300 Anfragen. Jede Anfrage liefert maximal 200 Ergebnisse, wovon die ersten 20 manuell ausgewertet wurden. Dies bedeutet eine maximale Anzahl von 6000 Ergebnisseiten. Die manuelle Auswertung der Ergebnisseiten erfolgt über eine Klassifizierung in eine der drei folgende Kategorien:

[1] Jugendschutz.net ist ein Unternehmen, welches von den Jugendministerien der Bundesländer gegründet wurde.

[2] `http://www.hass-im-netz.info/fileadmin/dateien/pk2013/bericht2012.pdf`, Stand: 17.September 2013

1. *Thematisch verwandte Webseite*: Das Ergebnisdokument enthält thematisch verwandten Inhalt mit dem Anfragedokument

2. *Thematisch nicht verwandte Webseite*: Das Ergebnisdokument enthält keinen thematisch verwandten Inhalt mit dem Anfragedokument

3. *Nicht klassifizierbar*: Der Inhalt des Ergebnisdokumentes ist nicht eindeutig mit dem Inhalt des Anfragedokumentes thematisch verwandt. Ebenfalls in diese Kategorie fallen Ergebnisse, die bei der Evaluierung nicht abrufbar waren.

Im Allgemeinen ist eine manuelle Auswertung von Ergebnissen subjektiv. Zur Einschränkung unterschiedlicher Bewertungen von Ergebnisdokumenten auf Grund subjektiver Wahrnehmung, wurden vor der Evaluierung folgende Kriterien als Richtlinien zur Klassifizierung definiert.

Wissenschaftliche Artikel: Die Bewertung inhaltlicher Überschneidungen bei wissenschaftlichen Artikeln erfordert oft tiefere Fachkenntnisse des jeweiligen Forschungsgebietes. Daher wird ein Dokument als thematisch verwandt klassifiziert, falls es sich ebenfalls um einen wissenschaftlichen Artikel handelt und die Zusammenfassung des Artikels darauf hindeutet, dass es sich mit den gleichen oder ähnlichen Forschungsfragen beschäftigt.

Nachrichten-Seiten: Webseiten mit journalistischem Hintergrund sind thematisch verwandt.

Webblogs mit rechtsextremistischem Hintergrund: Thematisch verwandt sind alle Dokumente und Webseiten mit offensichtlich rechtsextremistischen Hintergrund. Dies umfasst auch Blogs, soziale Netzwerke, Partei-Seiten und Shops.

NPD-Seiten: Thematisch verwandt sind alle Dokumente und Webseiten mit offensichtlich rechtsextremistischen Hintergrund. Dies umfasst auch Blogs, soziale Netzwerke, Partei-Seiten und Shops.

4.3 Ergebnisse

Die Evaluierung der Suchtechniken teilt sich auf in die Untersuchung der Laufzeit des Algorithmus, der Quantität und der Qualität der Ergebnisse. Eine Analyse dieser dient der Bestätigung oder Widerlegung der Hypothesen aus Kapitel 3.4.

Tabelle 4.1

Testdatensatz	**Klassifizierung**				**Laufzeit**
	verwandt	nicht verwandt	nicht klass.bar	Gesamt	Mittelwert in Sek.
Wiss. Artikel	0	0	0	0	0,92
Nachrichten-Seiten	18	443	39	500	6,79
Webblogs	173	70	248	491	5,57
NPD-Seiten	289	24	14	327	10,32

4.3.1 Link-Crawling Technik

Tabelle 4.1 zeigt die klassifizierten Ergebnisse und die Laufzeit der Link-Crawling Technik für die vier untersuchten Dokumenttypen. Im Folgenden wird die Eignung dieses Verfahrens für jeden Dokumenttyp einzeln diskutiert.

Wissenschaftliche Artikel: Wie bereits in Kapitel 3.1 beschrieben, beinhalten wissenschaftliche Artikel sehr selten Hyperlinks. Dass das Verfahren keine Ergebnisse liefert, liegt daran, dass sich die Extraktion von Hyperlinks auf die Auszeichnungssprache HTML beschränkt. Eine Erweiterung auf andere Verlinkungsformen, wie zum Beispiel wissenschaftliche Referenzen oder Hyperlinks in PDF-Dokumente, ist wünschenswert.

Nachrichten-Seiten: Die Quantität der Ergebnisse der Link-Crawling Technik ist sehr zufriedenstellend. Die Anzahl gefundener Dokumente erreichte die für die Evaluierung festgesetzte Obergrenze von 500 Dokumenten. Die Qualität der Ergebnisse ist jedoch schlecht, da nur 3,6% der Dokumente thematisch verwandten Inhalt zu dem Anfragedokument aufweisen, während 88,6% als eindeutig nicht thematisch verwandt klassifiziert wurden. Dies bestätigt die Hypothese, dass sich Nachrichten-Seiten wegen ihrer Konkurrenzsituation nicht verlinken und sich die Link-Crawling Technik daher nicht für diese eignet.

Webblogs: Die Quantität der Ergebnisse für Webblogs erreichte mit 491 von 500 Dokumenten fast die vorher definierte Obergrenze. Die Qualität variiert jedoch stark. Mit 50,5% der Ergebnisse sind mehr als die Hälfte der gefundenen Dokumente nicht klassifizierbar. Dies liegt vor allem daran, dass viele Hyperlinks nicht aktiv waren. Thematisch verwandt waren 35,2% der Ergebnisse. Die Hypothese, dass die Link-Crawling Technik für stark vernetzte Interessengruppen gut funktioniert, konnte somit größtenteils bestätigt werden. Abbildung 4.1 zeigt ein Beispiel für die Vernetzung solcher Webblogs.

Abbildung 4.1: Startseite von Kameradschaftsdienst vom 21.09.2013. Die rot umrandeten Hyperlinks verweisen auf ebenfalls rechtsextremistische Inhalte.

Tabelle 4.2

Testdatensatz	**Klassifizierung**				**Laufzeit**
	verwandt	nicht verwandt	nicht klass.bar	Gesamt	Mittelwert in Sek.
Wiss. Artikel	82	107	71	260	116,57
Nachrichten-Seiten	0	0	0	0	184,35
Webblogs	52	219	37	308	25,44
NPD-Seiten	0	0	0	0	129,63

NPD-Seiten: Die Anzahl der insgesamt extrahierten Hyperlinks ist mit 391 von 500 Ergebnissen als gut zu bewerten. Auch die Qualität der gefundenen Dokumente ist sehr gut. Mit 88,4% wies ein Großteil der Ergebnisse ebenfalls rechtsextremistische Inhalte auf. Dies bestätigt ebenfalls die Hypothese, dass sich die Link-Crawling Technik besonders für die Suche thematisch verwandter Inhalte bei stark vernetzten Interessengruppen eignet.

Zusammenfassung: Die Hypothesen für die Link-Crawling Technik konnten bestätigt werden. Sie funktioniert am besten bei stark vernetzten Interessengruppen. Für wissenschaftliche Dokumente eignet sich die Suchtechnik in ihrer aktuellen Implementierung jedoch nicht. Die durchschnittliche Laufzeit dieses Ansatzes ist im Vergleich zu den anderen Suchtechniken relativ gering. Pro Ergebnisdokument wurde eine Laufzeit von etwa 5 bis 10 Sekunden benötigt.

4.3.2 Key-Query Technik

Tabelle 4.2 zeigt die klassifizierten Ergebnisse der Key-Query Technik für die vier untersuchten Dokumenttypen. Die durchschnittlichen Laufzeiten pro Ergebnis sind ebenfalls in der Tabelle aufgeführt. Wie erwartet ist die durchschnittliche Zeit pro gefundenem Ergebnis mit 2-3 Minuten relativ hoch. Ein Großteil dieser Zeit ist für die Ermittlung der Keyqueries notwendig. Die qualitativen und quantitativen Resultate der Key-Query Technik werden für jeden Dokumenttyp einzeln analysiert.

Wissenschaftliche Artikel: Für diese Art von Dokumenten liegt die Qualität der Ergebnisse leicht unter den Erwartungen. Der Anteil thematisch verwandter Dokumente in der Ergebnismenge beträgt nur 31,5%. Eine Ursache dafür sind die fehlende Zugangsberechtigungen für einige der wissenschaftlichen Plattformen. Auf einige Ergebnisdokumente bestand kein Zugriff und

Abbildung 4.2: Startseite der FAZ am 21.09.2013. Die Markierungen zeigen die unterschiedlichen Themen, mit denen sich die Webseite befasst.

wurden somit als nicht klassifizierbar eingeordnet. Auch quantitativ erzielte die Suchtechnik nur bedingt die erwarteten Ergebnisse. Für etwa 54% der Anfragedokumente konnten Keyqueries generiert werden.

Nachrichten-Seiten: Aus Nachrichten-Seiten ließen sich in den Tests keine Keyqueries extrahieren. Dafür kann es unterschiedliche Gründe geben. Ein Grund ist wahrscheinlich die Diversität der Seiten. Diese sind sehr stark in viele kleine Themenfelder segmentiert, wie in Abbildung 4.2 zu sehen ist. Durch die Vermischung verschiedener Inhalte können keine sinnvoll zusammenhängenden Kombinationen von Keywords gebildet werden. Des weiteren ist es möglich, dass häufige Aktualisierungen der Inhalte der Nachrichten-Seiten dazu führen, dass die bei Google indizierte Version nicht konsistent mit der aktuellen Version der Webseite ist. Die Suchanfrage zur Prüfung einer potenziellen Keyquery liefert somit kein korrektes Ergebnis.

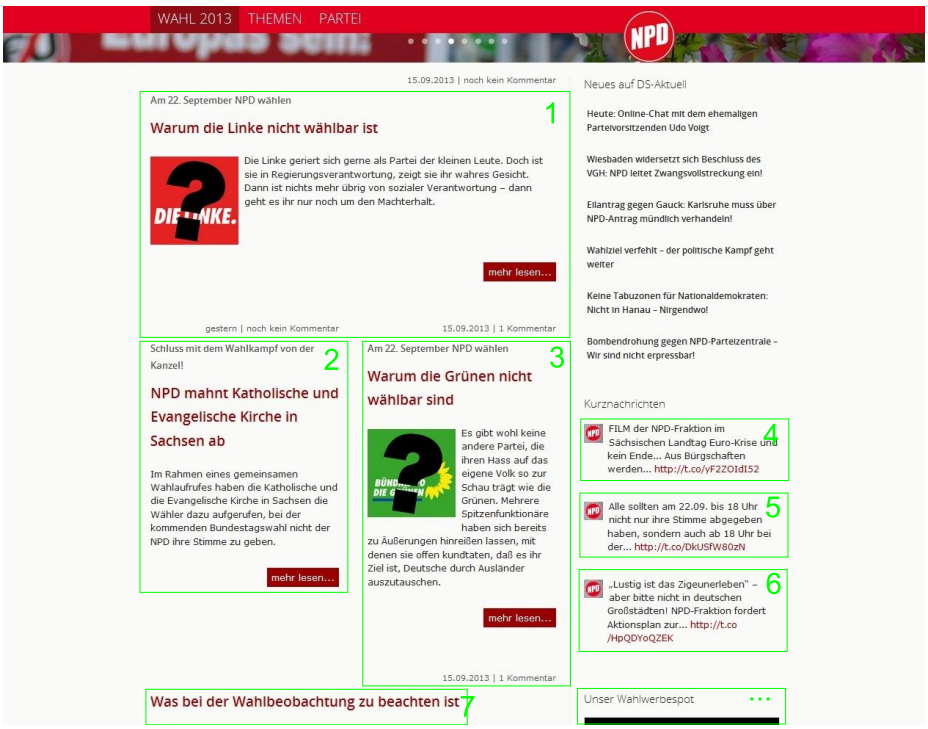

Abbildung 4.3: Startseite des NPD Bundesverbandes am 20.09.2013. Die Markierungen zeigen die unterschiedlichen Themen, mit denen sich die Webseite befasst.

Webblogs: Die Anzahl der gefundenen Ergebnisse übertrifft mit 308 von 500 möglichen Dokumenten die Erwartungen. Dieses Resultat ist erstaunlicherweise besser als das Resultat bei wissenschaftlichen Artikeln. Dies bedeutet, dass für Webblogs mehr Keyqueries generiert werden konnten. Jedoch ist nur ein geringer Prozentsatz von 16,9% mit dem Anfragedokument thematisch verwandt. Dies bedeutet, dass die extrahierten Keywords den Inhalt nicht ausreichend gut beschreiben. Für die erstaunlich geringe Laufzeit ist keine eindeutige Erklärung auf Grundlage der Ergebnisse möglich.

NPD-Seiten: Die Key-Query Technik konnte für NPD-Seiten keine Ergebnisse finden, da keine Keyqueries generiert werden konnten. Ein möglicher Grund dafür ist, wie bei Nachrichten-Seiten, die hohe Diversität der veröffentlichten Inhalte, wie in Abbildung 4.3 zu sehen ist.

Zusammenfassung: Die Key-Query Technik eignet sich für Dokumente mit monothematischen Inhalten. Im Gegensatz dazu funktionierte der Ansatz für Dokumente mit vielfältigen Inhalten nicht. Die Hypothesen bestätigten sich weitestgehend. Die Wartezeiten zwischen den Suchanfragen zur Vermeidung

Tabelle 4.3

Testdatensatz	**Klassifizierung**				**Laufzeit**
	verwandt	nicht verwandt	nicht klass.bar	Gesamt	Mittelwert in Sek.
Wiss. Artikel	0	0	0	0	7,38
Nachrichten-Seiten	406	91	3	500	0,83
Webblogs	0	0	0	0	24,23
NPD-Seiten	237	136	13	386	2,93

einer Sperrung durch Google, wie in Kapitel 3.2.2 beschrieben, führen zu einer hohen Laufzeit.

4.3.3 Google-Related Technik

Tabelle 4.3 zeigt die klassifizierten Ergebnisse und die Laufzeit der Google-Related Suche für die vier untersuchten Dokumenttypen. Im Folgenden wird die Eignung dieses Verfahrens für jeden Dokumenttyp einzeln diskutiert.

Wissenschaftliche Artikel: Es konnten keine Ergebnisse gefunden werden.

Nachrichten-Seiten: Die Ergebnisse stützen die Hypothese, dass die Google-Related Technik bei häufig gesuchten Anfragedokumenten gut funktioniert. Für alle Anfragedokumente konnte die Zielmenge an Ergebnissen gefunden werden. Davon sind 81,2% der gefundenen Webseiten thematisch verwandt mit dem Anfragedokument.

Webblogs: Es konnten keine Ergebnisse gefunden werden.

NPD-Seiten: Die NPD ist eine bundesweit agierende Partei und somit stark in der Öffentlichkeit vertreten. Von der hohen Anzahl der Ergebnisse (77,2% der Zielmenge) enthielten 61,4% diese Dokumente mit verwandten Inhalt.

Zusammenfassung: Die Suche mit der Google-Related Technik scheint für Webseiten mit populären und oft gesuchten Inhalten gut zu funktionieren. Die durchschnittliche Laufzeit ist geringer als bei der Key-Query und der Link-Crawling Technik.

Tabelle 4.4

	Link-Crawling	**Key-Query**	**Google-Related**
Gefundene Dokumente in % (500=100%)			
wiss. Artikel	0	52	0
Nachrichten-S.	100	0	100
Webblogs	98,2	61,6	0
NPD-S.	65,4	0	77,2
Thematisch verwandte Dokumente in %			
wiss. Artikel	0	31,5	0
Nachrichten-S.	3,6	0	81,2
Webblogs	35,2	16,9	0
NPD-S.	88,4	0	61,4
Gesamtwahrscheinlichkeit in %			
wiss. Artikel	0	16,4	0
Nachrichten-S.	3,6	0	81,2
Webblogs	34,6	10,4	0
NPD-S.	57,8	0	47,4

4.3.4 Zusammenfassung

Nach der Vorstellung der Ergebnisse der einzelnen Suchtechniken im letzten Kapitel, erfolgt in diesem Abschnitt ein direkter Vergleich der Suchtechniken untereinander in Abhängigkeit zu den Dokumenttypen. Tabelle 4.4 zeigt die Zusammenfassung der Ergebnisse.

Die Zusammenfassung zeigt, dass die Auswahl einer geeigneten Suchtechnik direkt vom Anfragedokument abhängt. Auf der Suche nach thematisch verwandten Dokumenten ist es hilfreich, bestimmte Eigenschaften des Anfragedokuments einschätzen zu können, um die Suchtechnik auszuwählen. Tabelle 4.4 zeigt die Gesamtwahrscheinlichkeiten für den Erfolg einer Suchtechnik in Abhängigkeit vom Dokumenttyp. Die Gesamtwahrscheinlichkeit setzt sich zusammen aus dem Produkt der Wahrscheinlichkeit für das Finden eines Dokuments und der Wahrscheinlichkeit, dass dieses thematisch verwandt ist.

Für Textdokumente, die sich mit einer einzelnen Thematik beschäftigen, wie z.B. wissenschaftliche Artikel, konnte ausschließlich die Key-Query Technik erfolgreiche Ergebnisse liefern. Für Webseiten mit hoher Diversität der Inhalte und hohem Bekanntheitsgrad, wie z.B. Nachrichten-Seiten, erzielte die Google-Related Technik die besten Ergebnisse. Auch die Link-Crawling Technik konnte

für diese Webseiten thematisch verwandte Ergebnisse finden, jedoch nur wenige. Für Webblogs (mit rechtsextremistischem Hintergrund) lieferten sowohl die Link-Crawling Technik als auch die Key-Query Technik gute Ergebnisse, wobei die Link-Crawling Technik vorzuziehen ist. Die Google-Related Technik funktionierte für diesen Dokumenttyp nicht. Webseiten stark vernetzter Interessengruppen, wie z.B. NPD-Seiten, eigneten sich für die Link-Crawling und Google-Related Suche in etwa gleich gut.

Kapitel 5

Diskussion und Ausblick

Kapitel 4.3 zeigt die Ergebnisse der Evaluierung. Die Hypothesen zu den Suchtechniken wurden darin im Wesentlichen bestätigt. Während die Ergebnisse die Stärken und Limitierungen der drei Verfahren zeigen, gibt es jedoch noch Aspekte, die einer Diskussion bedürfen.

Die Evaluierung der Link-Crawling Technik hat gezeigt, dass dieses Verfahren für bestimmte Inhalte unerwartet gut funktioniert. Dennoch bietet die beschriebene Implementierung Potenzial zur Erweiterung. Die Fokussierung auf Hyperlinks in HTML-Dokumenten ist eine nicht notwendige Einschränkung. Eine Unterstützung weiterer Formate (PDF, JavaScript, etc.) sowie alternative Formen der Referenzierung würden zu einer Verbesserung der Ergebnisse führen. Die Form der Referenzierung sollte sich nicht auf Hyperlinks beschränken und auch alternative Formen, wie z.B. wissenschaftliche Referenzen und Quellenangaben im Allgemeinen, einbeziehen.

Trotz der guten Ergebnisse der Link-Crawling Technik ist dieser Ansatz nicht generell für alle Dokumente anwendbar. Die Annahme des Verfahrens, dass ein Dokument auf andere Dokumente referenziert, ist nicht immer gegeben. Es besteht folglich eine hohe Abhängigkeit zwischen dem Erfolg des Verfahrens und der Existenz von Referenzen im Anfragedokument. Da sonstige Inhalte des Anfragedokumentes nicht betrachtet werden, funktioniert die Technik nur, wenn Referenzen vorhanden sind.

Eine Möglichkeit zur Abschwächung dieser Abhängigkeit ist die Kombination der Link-Crawling Technik mit anderen Ansätzen. Dabei könnte, neben den in dieser Arbeit vorgestellten Suchtechniken, auch eine Funktion von Google zum Einsatz kommen. Der Zusatz „link:“ bei einer Suche bei Google führt zu Dokumenten, welche auf das Anfragedokument verweisen. Eine Ausweitung der Suche auf diese Dokumente könnte zu zusätzlichen Ergebnissen führen.

Die Key-Query Technik ist ein vielversprechender Ansatz und die Ergebnisse zeigen, dass er für bestimmte Dokumente den anderen Ansätzen klar

überlegen ist. Für wissenschaftliche Artikel konnte die Key-Query Technik thematisch verwandte Dokumente finden, während die anderen beschriebenen Suchtechniken keine Ergebnisse lieferten.

Dennoch konnte die Key-Query Technik für zwei Dokumentarten aus dem Korpus zur Evaluierung keine Ergebnisse liefern. Dies ist darauf zurückzuführen, dass keine Keyqueries generiert werden konnten. Die Gründe dafür sind nicht eindeutig. Eine mögliche Ursache ist, dass für mannigfaltige Inhalte, wie z.B. Nachrichten-Seiten, die beschränkte Anzahl an Keywords bzw. Keyphrasen zur Erzeugung einer geeigneten Keyquery nicht ausreichend sind. Ob eine größere Anzahl verwendeter Keyphrasen zu einer Verbesserung der Ergebnisse führt, ist jedoch nicht klar und sollte Gegenstand weiterer Untersuchungen sein. Zu beachten wäre, dass eine solche Erweiterung zu einer Verlängerung der Laufzeit auf Grund einer höheren Anzahl an Suchabfragen führt.

Ein weiterer Grund für das Fehlschlagen der Key-Query Technik bei bestimmten Inhalten ist die mögliche Diskrepanz zwischen dem Anfragedokument und der indizierten Version der Suchmaschine. Darauf deuten einige der Ergebnisse der Evaluierung. Insbesondere für die im Testkorpus enthaltenen Nachrichten-Seiten ist es überraschend, dass keine Keyqueries generiert werden konnten. Die häufigen Aktualisierungen von Nachrichten-Seiten könnten dazu führen, dass nicht die aktuelle Version, sondern eine veraltete Version bei der verwendeten Suchmaschine indiziert ist. Da von der aktuellen Version die Keywords extrahiert werden, ist es möglich, dass sie sich stark von den Keywords der indizierten Version des Anfragedokuments unterscheiden.

Die hohen Laufzeiten der Key-Query Technik in der beschriebenen Implementierung sind, wie bereits besprochen, auf Limitierungen der automatischen Suche über die Suchmaschine Google zurückzuführen. Für zukünftige Arbeiten ist die Verwendung alternativer Suchmaschinen eine mögliche Lösung für dieses Problem.

Die Google-Related Technik erzielte in der Evaluierung gute Resultate für häufig aufgerufene Webseiten. Dies deutet darauf hin, dass Google zusätzliche Informationen über das Nutzerverhalten verwendet, um ähnliche Dokumente zu finden. Daraus ergeben sich für zukünftige Arbeiten zwei mögliche Erweiterungen der Link-Crawling Technik. Einerseits könnten die Ergebnisse mit Hilfe von Nutzerinformationen untereinander gewichtet werden. Andererseits könnten sich zusätzliche Hyperlinks potenziell ähnlicher Dokumente ermitteln lassen. Dabei könnten Hyperlinks untersucht werden, die von anderen Nutzern aufgerufen wurden, welche ebenfalls das Anfragedokument verwendeten.

Neben den diskutierten möglichen Verbesserungen der einzelnen Suchtechniken, bieten folgende Erweiterungen viel Potenzial für weitere Arbeiten. Die Ergebnisse der Evaluierung zeigen, dass jede Suchtechnik für unterschiedliche Dokumenttypen geeignet ist. Eine Analyse und Klassifikation des Anfragedo-

kumentes zur Auswahl einer geeigneten Suchtechnik ist sinnvoll.

In der Evaluierung der Suchtechniken wurden die Ergebnisse manuell bezüglich ihrer thematischen Verwandtschaft mit dem Anfragedokument klassifiziert. Dieser Schritt sollte in zukünftigen Implementierungen automatisiert werden. Nicht relevante Ergebnisse können mit Hilfe von Ähnlichkeitsmetriken oder anderen Ansätzen herausgefiltert werden. Qi et al. geben einen Überblick über verschiedenen Ansätze, dies zu realisieren [QD09].

Die automatische Analyse zur Auswahl einer geeigneten Suchtechnik und das Filtern der Ergebnisse können zu einer vollautomatischen Suche von thematisch verwandten Dokumenten zu einem Anfragedokument integriert werden. Ebenfalls ist das zusätzliche Aufbereiten der Ergebnisse durch Clustering-Techniken denkbar.

Kapitel 6

Zusammenfassung

In dieser Arbeit wurden Techniken zur Suche thematisch verwandter Dokumente entwickelt und untersucht. Mit Hilfe der Techniken Link-Crawling, Google-Related und Key-Query ist es möglich solche Dokumente im Web zu finden. Zur Verwendung dieser Techniken wurde ein gemeinsames Nutzerinterface entwickelt, welches die Eingabe eines Anfragedokumentes und die Auswahl einer gewünschten Suchtechnik erlaubt. Die Evaluierung der Verfahren wurde mit Hilfe dieses Interfaces realisiert.

Das erste Verfahren, die Link-Crawling Technik, arbeitet mit den Hyperlinks eines Dokumentes und benutzt diese als potenzielle Webseiten mit thematisch verwandten Inhalt (siehe Kapitel 3.1). Die Google-Related Technik verwendet die von Google bereitgestellte Option zur Suche nach verwandten Seiten zu einem Anfragedokument bzw. deren Adresse (siehe Kapitel 3.3). Das dritte Verfahren, die Key-Query Technik, arbeitet mit sogenannten Keyqueries, welche an der Bauhaus-Universität entwickelt wurden. Keyqueries sind Kombinationen aus Worten, die aus dem Text des Anfragedokumentes extrahiert werden und eine wichtige Rolle im Text des Dokuments spielen. Die Konzeption der Evaluierung erfolgte mit Hilfe der Aufstellung von Hypothesen über die Eignung der Techniken auf Basis ihrer Funktionsweisen.

Da diese drei Techniken jeweils einen unterschiedlichen Ansatz verwenden, wurde die Evaluierung an einem Korpus aus verschiedenen Anfragedokumenten durchgeführt. Dieser Korpus setzt sich zusammen aus 4 Gruppen von Dokumenten, welche jeweils exemplarisch für einen bestimmten Typ von Webdokumenten ausgewählt wurden. Für jedes der insgesamt 100 Anfragedokumente wurden alle Suchtechniken angewendet. Die jeweils ersten 20 Ergebnisse jeder Suche wurden bezüglich ihrer thematischen Verwandtschaft untersucht. Die bis zu 6000 Ergebnisdokumente wurden manuell nach vorher definierten Richtlinien klassifiziert. Mit Hilfe der Klassifikation der Ergebnisse konnte eine quantitative und qualitative Bewertung der Suchtechniken erfolgen.

Die Evaluierung zeigt, dass die Auswahl einer geeigneten Suchtechnik direkt vom Typ des Anfragedokuments abhängt. Dabei haben sich die Hypothesen zu den Suchtechniken weitestgehend bestätigt. Die Key-Query Technik kann bei Webblogs und bei Dokumenten, die sich mit einer einzelnen Thematik beschäftigen, wie z.B. wissenschaftliche Artikel, erfolgreiche Ergebnisse liefern. Die Link-Crawling Technik ist für verschiedene Dokumenttypen anwendbar, besonders für Webseiten von stark vernetzten Interessengruppen, wie z.B. Webblogs und Parteiwebseiten. Die Google-Related Technik eignet sich besonders gut für Nachrichten-Seiten, aber auch für Parteiwebseiten.

Die Umsetzung und Ergebnisse der drei Verfahren zeigen ihre Anwendbarkeit, jedoch auch ihre Limitierungen und Möglichkeiten der Verbesserung. Eine mögliche Kombination der Suchtechniken oder eine intelligente Auswahl eines geeigneten Suchverfahrens auf Grundlage des Anfragedokumentes würden eine einheitliche Suche für alle Dokumente ermöglichen. Das entwickelte Interface zur Suche thematisch verwandter Dokumente ist ein erster Schritt in diese Richtung.

Literaturverzeichnis

[ACGM+01] Arvind Arasu, Junghoo Cho, Hector Garcia-Molina, Andreas Paepcke und Sriram Raghavan. Searching the web. *ACM Trans. Internet Technol.*, 1(1):2–43, August 2001.

[BC09] Michael Bendersky und W. Bruce Croft. Finding text reuse on the web. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, Seiten 262–271, New York, NY, USA, 2009. ACM.

[Bri92] Eric Brill. A simple rule-based part of speech tagger. In *Proceedings of the workshop on Speech and Natural Language*, HLT '91, Seiten 112–116, Stroudsburg, PA, USA, 1992. Association for Computational Linguistics.

[DDKD09] Ali Dasdan, Paolo D'Alberto, Santanu Kolay und Chris Drome. Automatic retrieval of similar content using search engine query interface. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, Seiten 701–710, New York, NY, USA, 2009. ACM.

[FLSG11] Norbert Fuhr, Marc Lechtenfeld, Benno Stein und Tim Gollub. The Optimum Clustering Framework: Implementing the Cluster Hypothesis. *Information Retrieval*, 15(2):93–115, Juli 2011.

[GHMS13] Tim Gollub, Matthias Hagen, Maximilian Michel und Benno Stein. From Keywords to Keyqueries: Content Descriptors for the Web. In *36th International ACM Conference on Research and Development in Information Retrieval (SIGIR 13)*. ACM, Juli 2013.

[GLT12] Behzad Golshan, Theodoros Lappas und Evimaria Terzi. Sofia search: a tool for automating related-work search. In *Proceedings*

of the 2012 ACM SIGMOD International Conference on Management of Data, SIGMOD '12, Seiten 621–624, New York, NY, USA, 2012. ACM.

[HS11] Matthias Hagen und Benno Stein. Candidate Document Retrieval for Web-Scale Text Reuse Detection. In *18th International Symposium on String Processing and Information Retrieval (SPIRE 11)*, Band 7024 aus *Lecture Notes in Computer Science*, Seiten 356–367, Berlin Heidelberg New York, 2011. Springer.

[Jas06] Prof. Dr. Hans-Gerd Jaschke. Rechtsextremismus. `http://www.bpb.de/politik/extremismus/rechtsextremismus/41889/rechtsextremismus`, September 2006. Accessed: 29/01/2013.

[LJSL10] Yeha Lee, Hun-young Jung, Woosang Song und Jong-Hyeok Lee. Mining the blogosphere for top news stories identification. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, Seiten 395–402, New York, NY, USA, 2010. ACM.

[MT04] Rada Mihalcea und Paul Tarau. Textrank: Bringing order into texts. In Dekang Lin und Dekai Wu (Hrsg.), *Proceedings of EMNLP 2004*, Seiten 404–411, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[Nan08] Gabriele Nandlinger. Was ist rechtsextremismus? `http://www.bpb.de/politik/extremismus/rechtsextremismus/41312/was-ist-rechtsextrem`, July 2008. Accessed: 29/01/2013.

[Ora12] Oracle. Java™ platform standard ed. 7 - class url. `http://docs.oracle.com/javase/7/docs/api/java/net/URL.html#equals%28java.lang.Object%29`, September 2012. Accessed: 15/05/2013.

[PBMW99] Lawrence Page, Sergey Brin, Rajeev Motwani und Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.

[PCG10] Jeremy Pickens, Matthew Cooper und Gene Golovchinsky. Reverted indexing for feedback and expansion. In *Proceedings of the*

19th ACM international conference on Information and knowledge management, CIKM '10, Seiten 1049–1058, New York, NY, USA, 2010. ACM.

[QD09] Xiaoguang Qi und Brian D. Davison. Web page classification: Features and algorithms. *ACM Comput. Surv.*, 41(2):12:1–12:31, Februar 2009.

[YBD+09] Yin Yang, Nilesh Bansal, Wisam Dakka, Panagiotis Ipeirotis, Nick Koudas und Dimitris Papadias. Query by document. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, Seiten 34–43, New York, NY, USA, 2009. ACM.

Anhang A

Anhang

A.1 Datentabellen und Diagramme des Teilkorpus wissenschaftliche Artikel

Tabelle A.1: Qualitative Auswertung der wissenschaftlichen Artikel. Im ersten Abschnitt stehen die absolute Werte. Der zweite Abschnitt zeigt die prozentualen, klassifizierten Ergebnisse der pro Suchtechnik. Der dritte Abschnitt zeigt die prozentualen Werte der Suchtechniken pro Klassifizierung. Der letzte Abschnitt zeigt die prozentualen Werte der Klassifizierungen in Abhängigkeit zu der Suchtechnik.

Art der Suche	**Klassifizierung**			
	verwandt	nicht verwandt	nicht klass.bar	Gesamt
Absolute Werte				
Google-Related Technik	0	0	0	0
Key-Query Technik	82	107	71	260
Link-Crawling Technik	0	0	0	0
Gesamt	82	107	71	260
Prozentuale Werte der Klassifizierung pro Suchtechnik				
Google-Related Technik	0%	0%	0%	0%
Key-Query Technik	31,5%	41,2%	27.3%	100,0%
Link-Crawling Technik	0%	0%	0%	0%
Gesamt	31,5%	41,2%	27.3%	100,0%

Art der Suche	**Klassifizierung**			
	verwandt	nicht verwandt	nicht klass.bar	Gesamt
Prozentuale Werte der Suchtechniken pro Klassifizierung				
Google-Related Technik	0%	0%	0%	0%
Key-Query Technik	100,0%	100,0%	100,0%	100,0%
Link-Crawling Technik	0%	0%	0%	0%
Gesamt	100,0%	100,0%	100,0%	100,0%
Prozentuale Werte der Suchtechniken in Abhängigkeit zu der Klassifizierung				
Google-Related Technik	0%	0%	0%	0%
Key-Query Technik	31,5%	41,2%	27.3%	100,0%
Link-Crawling Technik	0%	0%	0%	0%
Gesamt	31,5%	41,2%	27.3%	100,0%

Tabelle A.2: Quantitative Auswertung der wissenschaftlichen Artikel; http://www.uni-weimar.de/medien/webis/publications/papers/

Start URL	**Art der Suchtechnik**		
	Google-Related	Key-Query	Link-Crawling
Gefundene Ergebnisse pro Anfrage (max. 20)			
.../stein_1999b.pdf	0	20	0
.../stein_1999d.pdf	0	0	0
.../stein_2001e.pdf	0	0	0
.../stein_2003a.pdf	0	0	0
.../stein_2003b.pdf	0	20	0
.../stein_2003d.pdf	0	0	0
.../stein_2007g.pdf	0	20	0
.../stein_2008a.pdf	0	20	0
.../stein_2008b.pdf	0	20	0
.../stein_2008i.pdf	0	20	0
.../stein_2010k.pdf	0	0	0
.../stein_2011a.pdf	0	20	0
.../stein_2011e.pdf	0	0	0
.../stein_2011n.pdf	0	0	0
.../stein_2011s.pdf	0	20	0
.../stein_2011u.pdf	0	0	0

Start URL	**Art der Suchtechnik**		
	Google-Related	Key-Query	Link-Crawling
.../stein_2011w.pdf	0	20	0
.../stein_2012g.pdf	0	0	0
.../stein_2012i.pdf	0	0	0
.../stein_2012q.pdf	0	0	0
.../stein_2012v.pdf	0	20	0
.../stein_2013e.pdf	0	0	0
.../stein_2013f.pdf	0	20	0
.../stein_2013g.pdf	0	20	0
.../stein_2013h.pdf	0	20	0
Gesamt	0	260	0
Statistische Auswertungen			
Mittelwert	0,0	10,4	0,0
Mittelwert in Prozent(20 = 100%)	0.0%	52%	0,0%
Median	0	20	0
Standartabweichung	0,0	10,20	0,0
Varianz	0,0	104	0,0

Tabelle A.3: Laufzeitauswertung der wissenschaftlichen Artikel

Start URL	**Art der Suchtechnik**		
	Google-Related	Key-Query	Link-Crawling
Durchschnittliche Zeit in s pro gefundenem Dokument			
.../stein_1999b.pdf	7,45	1,53	0,15
.../stein_1999d.pdf	8,09	93,82	0,34
.../stein_2001e.pdf	8,96	419,08	4,43
.../stein_2003a.pdf	7,29	207,29	1,13
.../stein_2003b.pdf	7,38	1,39	0,03
.../stein_2003d.pdf	6,72	304,14	0,43
.../stein_2007g.pdf	7,25	2,31	0,14
.../stein_2008a.pdf	7,36	2,16	0,48
.../stein_2008b.pdf	7,54	2,63	0,48

Start URL	**Art der Suchtechnik**		
	Google-Related	Key-Query	Link-Crawling
.../stein_2008i.pdf	7,77	1,42	1,72
.../stein_2010k.pdf	8,49	195,94	0,71
.../stein_2011a.pdf	6,93	1,19	0,28
.../stein_2011e.pdf	7,25	113,37	0,20
.../stein_2011n.pdf	7,42	202,41	0,98
.../stein_2011s.pdf	7,29	4,12	1,46
.../stein_2011u.pdf	7,87	409,54	1,27
.../stein_2011w.pdf	0,78	4,70	0,03
.../stein_2012g.pdf	8,52	211,11	1,11
.../stein_2012i.pdf	7,95	197,87	0,17
.../stein_2012q.pdf	6,95	131,59	0,53
.../stein_2012v.pdf	7,13	1,99	0,42
.../stein_2013e.pdf	7,40	399,24	0,56
.../stein_2013f.pdf	9,29	1,55	4,35
.../stein_2013g.pdf	7,36	1,80	0,29
.../stein_2013h.pdf	8,11	2,06	1,39
Statistische Auswertungen (in s)			
Mittelwert	7,38	116,57	0,92
Median	7,40	4,70	0,48
Standartabweichung	1,51	144,43	1,15
Varianz	2,29	20859,36	1,32

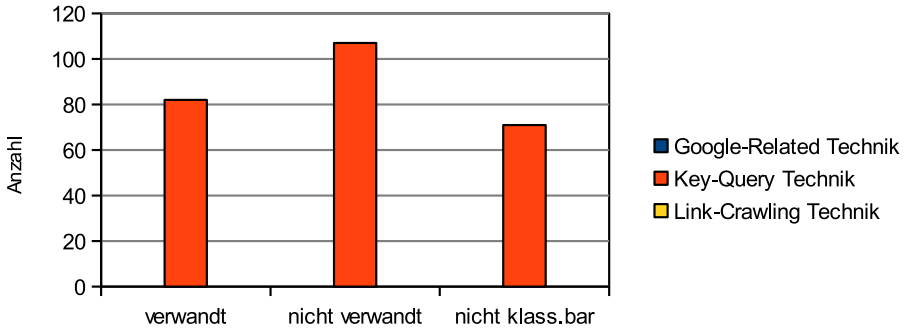

Abbildung A.1: Qualitative Auswertung der wissenschaftlichen Artikel

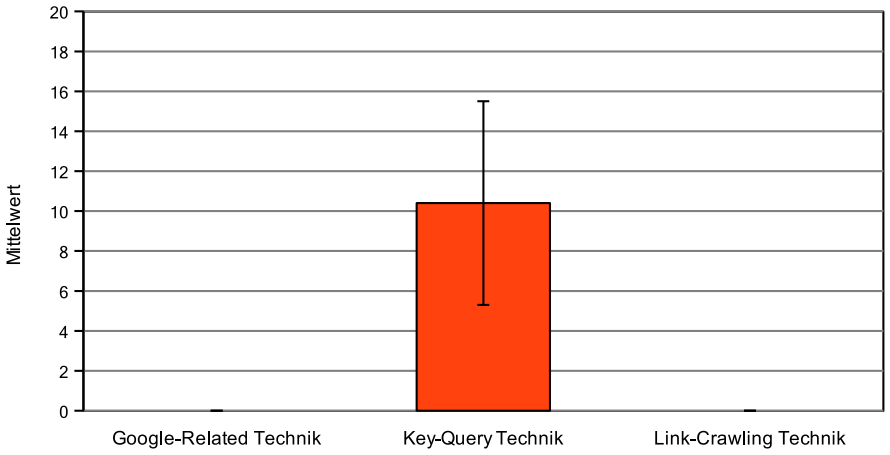

Abbildung A.2: Quantitative Auswertung der wissenschaftlichen Artikel

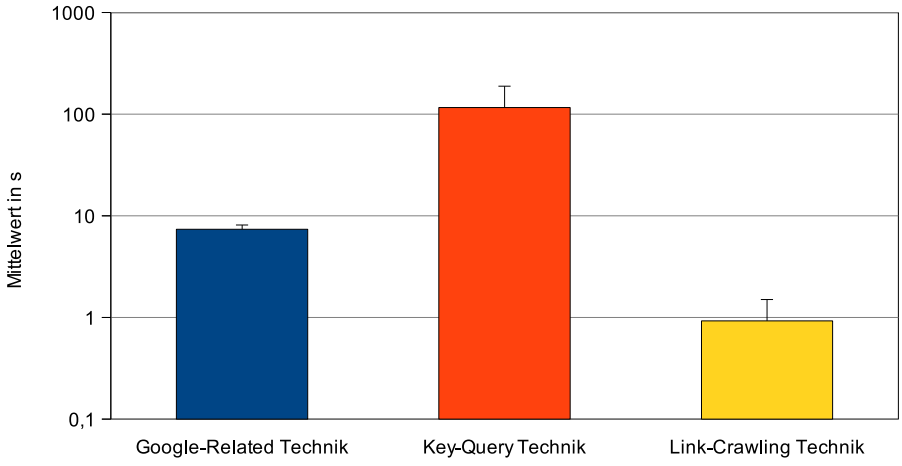

Abbildung A.3: Laufzeitauswertung der wissenschaftlichen Artikel

A.2 Datentabellen und Diagramme des Teilkorpus News-Seiten

Tabelle A.4: Qualitative Auswertung der News-Seiten. Im ersten Abschnitt stehen die absolute Werte. Der zweite Abschnitt zeigt die prozentualen, klassifizierten Ergebnisse der pro Suchtechnik. Der dritte Abschnitt zeigt die prozentualen Werte der Suchtechniken pro Klassifizierung. Der letzte Abschnitt zeigt die prozentualen Werte der Klassifizierungen in Abhängigkeit zu der Suchtechnik.

Art der Suche	**Klassifizierung**			
	verwandt	nicht verwandt	nicht klass.bar	Gesamt
Absolute Werte				
Google-Related Technik	406	91	3	500
Key-Query Technik	0	0	0	0
Link-Crawling Technik	18	443	39	500
Gesamt	424	534	42	1000
Prozentuale Werte der Klassifizierung pro Suchtechnik				
Google-Related Technik	81,2%	18,2%	0,6%	100,0%
Key-Query Technik	0%	0%	0%	0%
Link-Crawling Technik	3,6%	88,6%	7,8%	100,0%
Gesamt	42,4%	53,4%	4,2%	100,0%
Prozentuale Werte der Suchtechniken pro Klassifizierung				
Google-Related Technik	95,8%	17,0%	7,1%	50,0%
Key-Query Technik	0%	0%	0%	0%
Link-Crawling Technik	4,2%	83,0%	92,9%	50,0%
Gesamt	100,0%	100,0%	100,0%	100,0%
Prozentuale Werte der Suchtechniken in Abhängigkeit zu der Klassifizierung				
Google-Related Technik	40,6%	9,1%	0,3%	50,0%
Key-Query Technik	0%	0%	0%	0%
Link-Crawling Technik	1,8%	44,3%	3,9%	50,0%
Gesamt	42,4%	53,4%	4,2%	100,0%

Tabelle A.5: Quantitative Auswertung der News-Seiten

Start URL	**Art der Suchtechnik**		
	Google-Related	Key-Query	Link-Crawling
Gefundene Ergebnisse pro Anfrage (max. 20)			
http://www.allgemeiner-anzeiger.de/	20	0	20
http://www.bild.de/	20	0	20
http://www.faz.net/	20	0	20
http://www.focus.de/	20	0	20
http://www.handelsblatt.com/	20	0	20
http://www.heute.de/	20	0	20
http://www.hna.de/	20	0	20
http://www.lr-online.de/	20	0	20
http://www.mittelbayerische.de/	20	0	20
http://www.n-tv.de/	20	0	20
http://www.n24.de/	20	0	20
http://www.otz.de/	20	0	20
http://www.ovz-online.de/	20	0	20
http://www.rundschau-online.de/	20	0	20
http://www.spiegel.de/	20	0	20
http://www.stern.de/	20	0	20
http://www.stuttgarter-zeitung.de/	20	0	20
http://www.sueddeutsche.de/	20	0	20
http://www.svz.de/	20	0	20
http://www.sz-online.de/	20	0	20
http://www.tagesspiegel.de/	20	0	20
http://www.thueringer-allgemeine.de/	20	0	20
http://www.tlz.de/	20	0	20
http://www.welt.de/	20	0	20
http://www.zeit.de/	20	0	20
Gesamt	500	0	500
Statistische Auswertungen			
Mittelwert	20	0	20
Mittelwert in Prozent(20 = 100%)	100%	0%	100%
Median	20	0	20
Standartabweichung	0	0	0
Varianz	0	0	0

Tabelle A.6: Laufzeitauswertung der News-Seiten

Start URL	**Art der Suchtechnik**		
	Google-Related	Key-Query	Link-Crawling
Durchschnittliche Zeit in s pro gefundenem Dokument			
http://www.allgemeiner-anzeiger.de/	1,04	59,25	16,05
http://www.bild.de/	0,82	477,07	0,17
http://www.faz.net/	0,80	69,94	0,39
http://www.focus.de/	0,78	110,20	0,01
http://www.handelsblatt.com/	0,81	248,97	0,62
http://www.heute.de/	0,80	82,06	¿20h
http://www.hna.de/	0,82	84,90	0,08
http://www.lr-online.de/	0,79	94,95	0,48
http://www.mittelbayerische.de/	0,83	95,97	25,97
http://www.n-tv.de/	0,78	207,38	0,017
http://www.n24.de/	0,86	256,76	¿20h
http://www.otz.de/	0,85	521,64	83,04
http://www.ovz-online.de/	0,97	178,07	0,48
http://www.rundschau-online.de/	0,84	92,72	0,09
http://www.spiegel.de/	0,78	485,32	0,05
http://www.stern.de/	0,80	77,35	1,01
http://www.stuttgarter-zeitung.de/	0,79	76,74	0,60
http://www.sueddeutsche.de/	0,81	100,74	0,01
http://www.svz.de/	0,83	72,53	0,46
http://www.sz-online.de/	0,79	107,18	0,64
http://www.tagesspiegel.de/	0,80	100,50	10,83
http://www.thueringer-allgemeine.de/	0,83	186,69	4,81
http://www.tlz.de/	0,79	226,72	10,22
http://www.welt.de/	0,79	114,17	0,04
http://www.zeit.de/	0,86	480,91	0,03
Statistische Auswertungen (in s)			
Mittelwert	0,83	184,35	6,79
Median	0,81	107,18	0,48
Standartabweichung	0,06	148,46	17,86
Varianz	0,00	22041,47	318,95

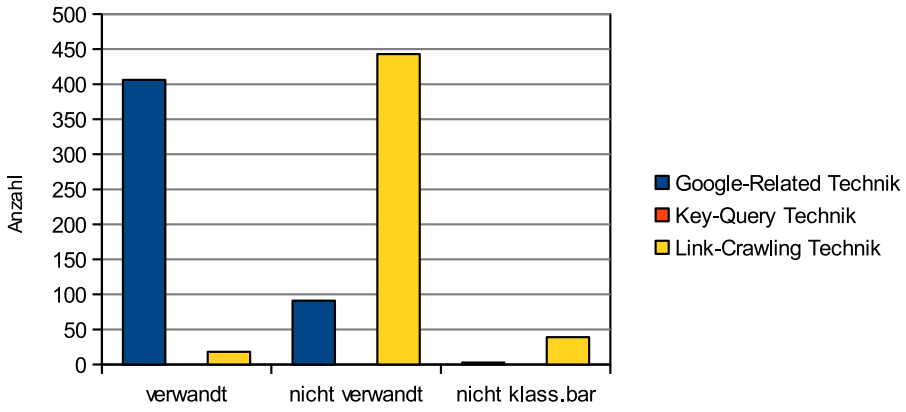

Abbildung A.4: Qualitative Auswertung der News-Seiten

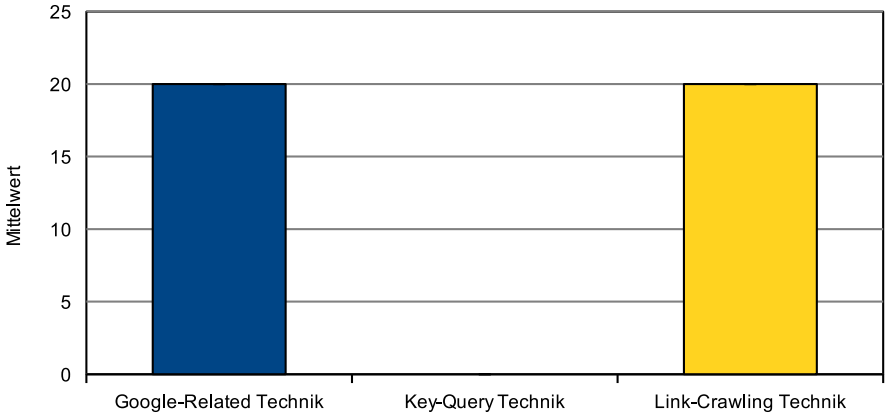

Abbildung A.5: Quantitative Auswertung der News-Seiten

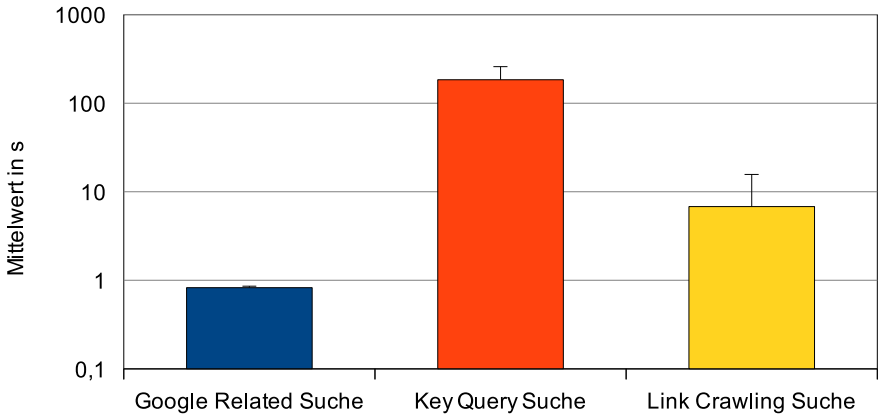

Abbildung A.6: Laufzeitauswertung der News-Seiten

A.3 Datentabellen und Diagramme des Teilkorpus Webblogs

Tabelle A.7: Qualitative Auswertung der Webblogs. Im ersten Abschnitt stehen die absolute Werte. Der zweite Abschnitt zeigt die prozentualen, klassifizierten Ergebnisse der pro Suchtechnik. Der dritte Abschnitt zeigt die prozentualen Werte der Suchtechniken pro Klassifizierung. Der letzte Abschnitt zeigt die prozentualen Werte der Klassifizierungen in Abhängigkeit zu der Suchtechnik.

Art der Suche	**Klassifizierung**			
	verwandt	nicht verwandt	nicht klass.bar	Gesamt
Absolute Werte				
Google-Related Technik	0	0	0	0
Key-Query Technik	52	219	37	308
Link-Crawling Technik	173	70	248	491
Gesamt	225	289	285	799
Prozentuale Werte der Klassifizierung pro Suchtechnik				
Google-Related Technik	0%	0%	0%	0%
Key-Query Technik	16,9%	71,1%	12,0%	100,0%
Link-Crawling Technik	35,2%	14,3%	50,5%	100,0%
Gesamt	28,2%	36,2%	35,6%	100,0%
Prozentuale Werte der Suchtechniken pro Klassifizierung				
Google-Related Technik	0%	0%	0%	0%
Key-Query Technik	23,1%	75,8%	13,0%	38,5%
Link-Crawling Technik	76,9	24,2%	87,0%	61,5%
Gesamt	100%	100%	100%	100%
Prozentuale Werte der Suchtechniken in Abhängigkeit zu der Klassifizierung				
Google-Related Technik	0%	0%	0%	0%
Key-Query Technik	6,5%	27,4%	4,6%	38,5%
Link-Crawling Technik	21,7	8,8%	31,0%	61,5%
Gesamt	28,2%	36,2%	35,6%	100%

Tabelle A.8: Quantitative Auswertung der Webblogs

Start URL	**Art der Suchtechnik**		
	Google-Related	Key-Query	Link-Crawling
Gefundene Ergebnisse pro Anfrage (max. 20)			
http://logr.org/ageutin/	0	7	20
http://logr.org/agkiel/	0	20	20
http://logr.org/agkredu/	0	8	20
http://logr.org/agmerseburg/	0	7	20
http://logr.org/ahlen/	0	20	20
http://logr.org/ansr/	0	4	20
http://logr.org/anstormarn/	0	18	20
http://logr.org/anwetzlar/	0	19	20
http://logr.org/derstaatsstreich/	0	4	20
http://logr.org/fkse/	0	15	20
http://logr.org/fnaluenen/	0	9	20
http://logr.org/fnkoeln/	0	17	20
http://logr.org/fnwug/	0	8	20
http://logr.org/freiesks/	0	4	20
http://logr.org/infoleipzig/	0	20	20
http://logr.org/infozwickau/	0	6	20
http://logr.org/kameradschaftsdienst/	0	4	20
http://logr.org/leerostfriesland/	0	20	20
http://logr.org/nasopremnitz/	0	14	20
http://logr.org/nsfkn/	0	20	20
http://logr.org/nsolfen/	0	4	20
http://logr.org/nsrastatt/	0	9	20
http://logr.org/sachsentage/	0	11	11
http://logr.org/strassenkunst/	0	20	20
http://logr.org/tddz/	0	20	20
Gesamt	0	308	491
Statistische Auswertungen			
Mittelwert	0	12,32	19,64
Mittelwert in %	0%	61,6%	98,2%
Median	0	11	20
Standardabweichung	0	6,54	1,80
Varianz	0	42,73	3,24

Tabelle A.9: Laufzeitauswertung der Webblogs

Start URL	**Art der Suchtechnik**		
	Google-Related	Key-Query	Link-Crawling
Durchschnittliche Zeit in s pro gefundenem Dokument			
http://logr.org/ageutin/	16,77	32,44	4,00
http://logr.org/agkiel/	32,97	6,23	1,98
http://logr.org/agkredu/	14,48	14,54	0,36
http://logr.org/agmerseburg/	17,29	11,47	4,38
http://logr.org/ahlen/	37,93	4,34	1,21
http://logr.org/ansr/	16,74	30,65	1,95
http://logr.org/anstormarn/	21,98	236,99	33,04
http://logr.org/anwetzlar/	24,37	11,20	1,96
http://logr.org/derstaatsstreich/	15,38	33,46	3,38
http://logr.org/fkse/	22,97	15,33	1,36
http://logr.org/fnaluenen/	15,45	20,25	0,92
http://logr.org/fnkoeln/	23,09	13,29	2,25
http://logr.org/fnwug/	16,24	17,66	3,58
http://logr.org/freiesks/	29,91	29,83	4,53
http://logr.org/infoleipzig/	57,92	4,62	3,47
http://logr.org/infozwickau/	17,46	29,62	5,93
http://logr.org/kameradschaftsdienst/	17,39	32,60	4,96
http://logr.org/leerostfriesland/	43,10	2,70	2,03
http://logr.org/nasopremnitz/	22,82	15,34	1,43
http://logr.org/nsfkn/	24,14	12,99	0,32
http://logr.org/nsolfen/	17,59	26,00	5,76
http://logr.org/nsrastatt/	15,99	9,78	29,24
http://logr.org/sachsentage/	23,03	9,44	14,76
http://logr.org/strassenkunst/	29,78	7,71	1,17
http://logr.org/tddz/	31,04	7,59	5,21
Statistische Auswertungen in Sekunden			
Mittelwert	24,23	25,44	5,57
Median	22,82	14,54	3,38
Standardabweichung	10,32	45,18	8,24
Varianz	106,43	2041,33	67,92

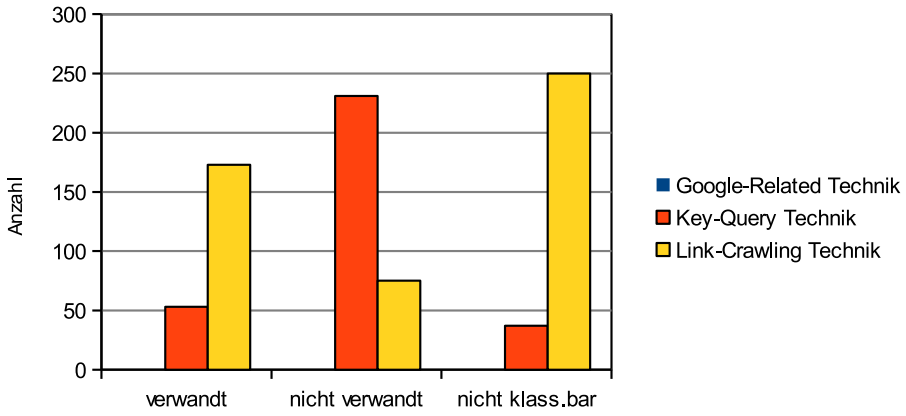

Abbildung A.7: Qualitative Auswertung der Webblogs

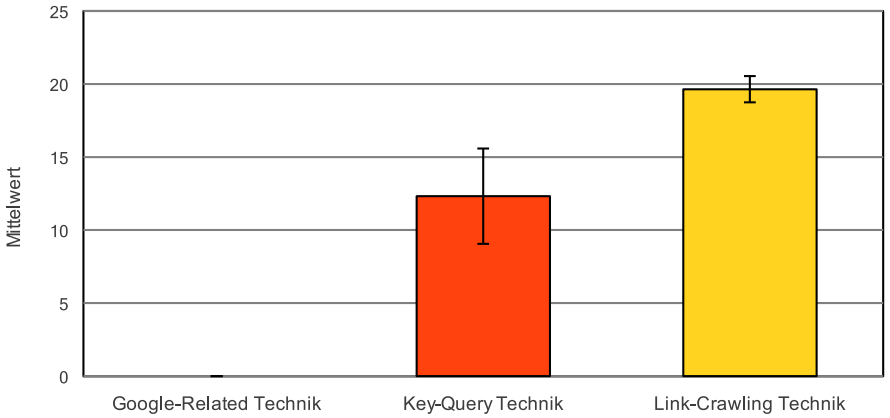

Abbildung A.8: Quantitative Auswertung der Webblogs

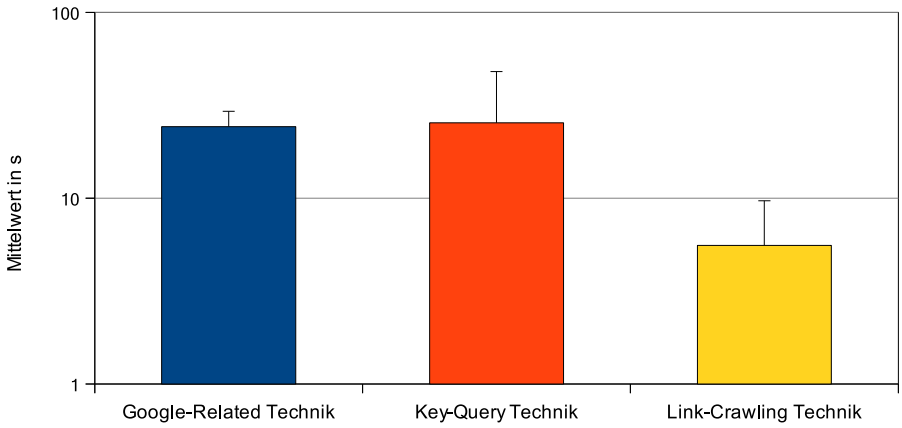

Abbildung A.9: Laufzeitauswertung der Webblogs

A.4 Datentabellen und Diagramme des Teilkorpus NPD-Seiten

Tabelle A.10: Qualitative Auswertung der NPD-Seiten. Im ersten Abschnitt stehen die absolute Werte. Der zweite Abschnitt zeigt die prozentualen, klassifizierten Ergebnisse der pro Suchtechnik. Der dritte Abschnitt zeigt die prozentualen Werte der Suchtechniken pro Klassifizierung. Der letzte Abschnitt zeigt die prozentualen Werte der Klassifizierungen in Abhängigkeit zu der Suchtechnik.

Art der Suche	**Klassifizierung**			
	verwandt	nicht verwandt	nicht klass.bar	Gesamt
Absolute Werte				
Google-Related Technik	237	136	13	386
Key-Query Technik	0	0	0	0
Link-Crawling Technik	289	24	14	327
Gesamt	526	160	27	713
Prozentuale Werte der Klassifizierung pro Suchtechnik				
Google-Related Technik	61,4%	35,2%	3,4%	100,0%
Key-Query Technik	0%	0%	0%	0%
Link-Crawling Technik	88,4%	7,3%	4,3%	100,0%
Gesamt	73,8%	22,4%	3,8%	100,0%
Prozentuale Werte der Suchtechniken pro Klassifizierung				
Google-Related Technik	45,1%	85,0%	48,1%	54,1%
Key-Query Technik	0%	0%	0%	0%
Link-Crawling Technik	54,9	15,0%	51,9%	45,9%
Gesamt	100%	100%	100%	100%
Prozentuale Werte der Suchtechniken in Abhängigkeit zu der Klassifizierung				
Google-Related Technik	33,2%	19,1%	1,8%	54,1%
Key-Query Technik	0%	0%	0%	0%
Link-Crawling Technik	40,5	3,4%	2,0%	45,9%
Gesamt	73,7%	22,5%	3,8%	100%

Tabelle A.11: Quantitative Auswertung der NPD-Seiten

Start URL	**Art der Suchtechnik**		
	Google-Related	Key-Query	Link-Crawling
Gefundene Ergebnisse pro Anfrage (max. 20)			
http://npd-fraktion-sachsen.de/	20	0	20
http://npd-saar.de/	2	0	1
http://www.npd-bayern.de/	20	0	20
http://www.npd-berlin.de/	2	0	20
http://www.npd-brandenburg.de/	1	0	20
http://www.npd-bremen.de/	20	0	12
http://www.npd-burgenlandkreis.de/	18	0	20
http://www.npd-bw.de/	20	0	0
http://www.npd-dresden.de/	20	0	0
http://www.npd-fraktion-mv.de/	20	0	5
http://www.npd-hamburg.de/	20	0	0
http://www.npd-hessen.de/	20	0	20
http://www.npd-in-rlp.de/	20	0	20
http://www.npd-kiel.de/	17	0	20
http://www.npd-koeln.de/	3	0	0
http://www.npd-mittelfranken.de/	19	0	20
http://www.npd-muenchen.de/	0	0	20
http://www.npd-mv.de/	20	0	8
http://www.npd-niedersachsen.de/	20	0	20
http://www.npd-nrw.de/	20	0	0
http://www.npd-rhein-sieg.de/	4	0	0
http://www.npd-sachsen-anhalt.de/	20	0	20
http://www.npd-sachsen.de/	20	0	20
http://www.npd.de/	20	0	20
http://www.npdfrankfurt.de/	20	0	20
Gesamt	386	0	326
Statistische Auswertungen			
Mittelwert	15,44	0	13,04
Mittelwert in Prozent(20 = 100%)	77,2%	0%	65,2%
Median	20	0	20
Standartabweichung	7,77	0	9,09
Varianz	60,34	0	82,62

Tabelle A.12: Laufzeitauswertung der NPD-Seiten

Start URL	**Art der Suchtechnik**		
	Google-Related	Key-Query	Link-Crawling
Durchschnittliche Zeit in s pro gefundenem Dokument			
http://npd-fraktion-sachsen.de/	0,93	88,39	0,58
http://npd-saar.de/	7,40	206,02	19,37
http://www.npd-bayern.de/	0,97	77,07	35,80
http://www.npd-berlin.de/	7,31	175,48	6,99
http://www.npd-brandenburg.de/	15,17	231,44	9,16
http://www.npd-bremen.de/	1,45	75,30	2,36
http://www.npd-burgenlandkreis.de/	1,23	81,75	3,30
http://www.npd-bw.de/	0,88	537,14	1,37
http://www.npd-dresden.de/	1,14	65,53	0,27
http://www.npd-fraktion-mv.de/	0,89	67,32	9,02
http://www.npd-hamburg.de/	0,87	83,94	0,73
http://www.npd-hessen.de/	1,11	230,12	26,75
http://www.npd-in-rlp.de/	1,13	77,58	36,77
http://www.npd-kiel.de/	1,32	31,156	6,93
http://www.npd-koeln.de/	4,93	66,26	0,63
http://www.npd-mittelfranken.de/	1,22	70,082	14,03
http://www.npd-muenchen.de/	14,98	7,23	21,04
http://www.npd-mv.de/	0,88	64,39	3,86
http://www.npd-niedersachsen.de/	1,03	90,20	19,07
http://www.npd-nrw.de/	0,86	152,99	0,49
http://www.npd-rhein-sieg.de/	3,78	69,44	0,28
http://www.npd-sachsen-anhalt.de/	0,87	94,49	8,32
http://www.npd-sachsen.de/	0,95	190,35	0,58
http://www.npd.de/	0,79	203,52	3,63
http://www.npdfrankfurt.de/	1,12	203,47	26,73
Statistische Auswertungen (in s)			
Mittelwert	2,93	129,63	10,32
Median	1,12	83,94	6,93
Standartabweichung	4,12	107,05	11,47
Varianz	17,00	11458,93	131,59

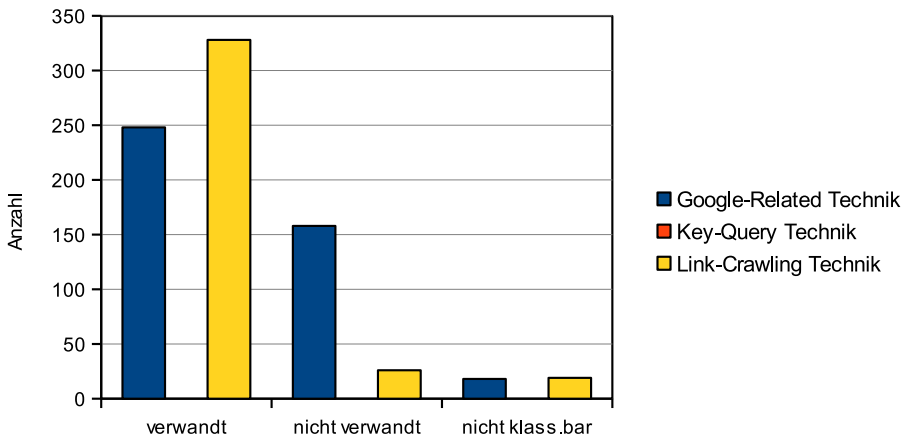

Abbildung A.10: Qualitative Auswertung der NPD-Seiten

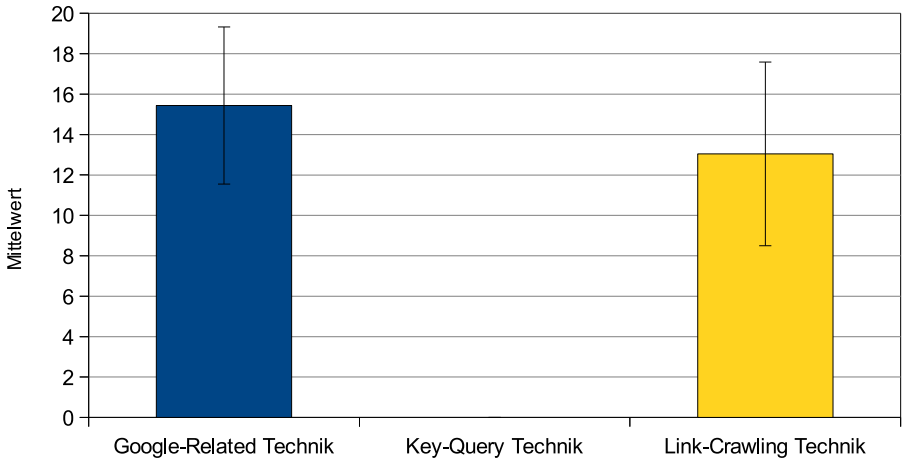

Abbildung A.11: Quantitative Auswertung der NPD-Seiten

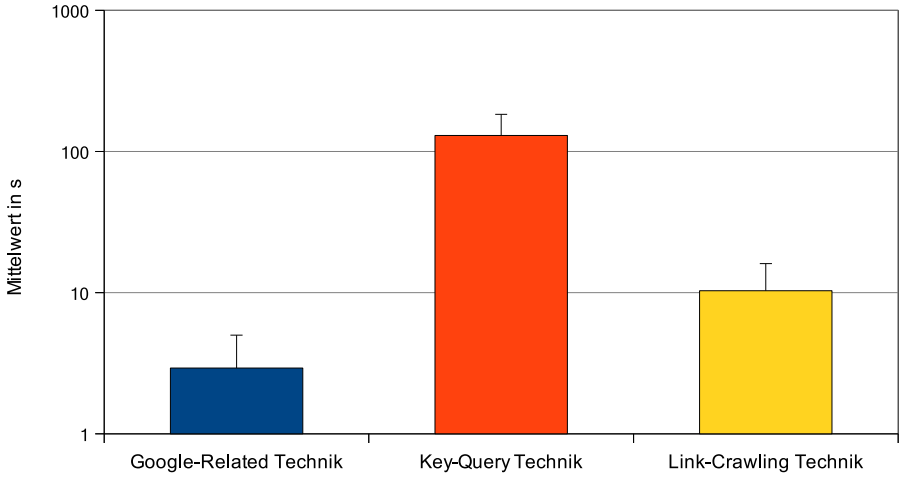

Abbildung A.12: Laufzeitauswertung der NPD-Seiten