UNIVERSITÄT DORTMUND
FACHBEREICH INFORMATIK

Diplomarbeit

PageTracker
ein Agent zum Wiederauffinden von Webseiten mit intelligenten Suchanfragen

Nils Malzahn

Diplomarbeit am
Fachbereich Informatik
der Universität Dortmund

28. September 2003

Betreuer:
Prof. Dr. Katharina Morik
Dipl.-Inform. Stefan Haustein

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

1.1 Motivation

Das Internet beherbergt mittlerweile eine Vielzahl von Dokumenten, die von Nutzern des Internets gelesen und referenziert werden. Dies geschieht entweder durch Einträge in persönlichen Bookmark–Listen der Internetnutzer, durch Verweise in (Web–)Dokumenten oder durch Verweise in (selbst erstellten) Dokumentenmanagementsystemen.

Die Dynamik des Internets bedingt, dass sich unter der einmal eingetragenen URL gar kein Inhalt oder ein veränderter Inhalt befindet, der nicht mehr der ursprünglichen Motivation für den gesetzten Verweis entspricht. Ein Beispiel für dieses Problem sind etwa die Kurs-Unterlagen zu Lehrveranstaltungen. Nach Ablauf des Semesters pflegen sie nicht gelöscht, wohl aber verlagert zu werden. Sobald der Nutzer die Änderung bemerkt, wird er den fehlerhaften Verweis entweder löschen oder versuchen den verschwundenen Inhalt wiederzufinden. Für solche Suchen soll aber in Agenten–Systemen nicht ausgerechnet ein Mensch eingesetzt werden!

Wenn nun unter einer URL eine Information, die zu einem Ontologiepunkt passt, nicht mehr zu finden ist, wo ist sie dann hin gewandert? Wie kann ein Agent die neue URL finden? Welche Heuristiken lassen sich angeben? Diesen Fragestellungen soll im Rahmen dieser Arbeit nachgegangen werden.

1.2 Aufgabenstellung

Ziel der vorliegenden Diplomarbeit ist es, einen Agenten zu erstellen, der den neuen Fundort eines vorgegebenen Dokuments *wiederfindet*, falls das Dokument mit der gesuchten Information noch existiert.

Nach der Installation und der Eingabe der zu überwachenden URLs soll der Agent keinen weiteren Benutzereingriff bei der Suche nach verschwundenen Informationen benötigen. Da während der Suche kein Benutzereingriff notwendig ist, kann der Agent als server–seitige Applikation implementiert werden, so dass die tatsächliche Suchzeit von geringer Bedeutung ist. Andererseits kann der Agent nicht das gesamte erreichbare Internet absuchen. Das zu suchende Dokument sollte also nach möglichst wenigen Schritten gefunden werden. Ein Schritt entspricht einem verfolgten Verweis vom Startpunkt der Suche.

Um die Arbeitsersparnis durch die automatisierte Suche nicht durch einen hohen Installationsaufwand zu kompensieren, soll der Agent mit möglichst wenig Wissen über den Sachbereich der zu suchenden URLs auskommen können.

Verschiedene Benutzer können verschiedener Ansicht über das Vorhandensein einer Information unter einer URL sein. Das bedeutet, dass der Agent bzgl. eines zu definierenden Ähnlichkeitsmaßes, das das Informationsbedürfnis widerspiegelt, konfigurierbar sein muss. Das Ähnlichkeitsmaß wird dem Agenten vom Benutzer vorgegeben. Da verschiedene Benutzerinteressen unter Umständen verschiedene Suchstrategien notwendig machen, soll auch die Integration neuer Suchstrategien mit möglichst geringem Aufwand möglich sein.

Diese Arbeit beschränkt sich auf die Auswertung von HTML-Texten. Grafiken, JavaScript und ähnliche Web-Technologien werden nicht analysiert. Auch passwortgeschützte Seiten sind für den Agenten Sackgassen. Aus der Vielzahl der denkbaren Ähnlichkeitsbegriffe für zwei Dokumente wird der Agent sich in der vorliegenden Diplomarbeit auf das Wiederauffinden von inhaltsgleichen Dokumenten beschränken. Das Finden von z. B. struktur- oder themengleichen Dokumenten ist nicht Ziel der Diplomarbeit.

1.3 Aufbau der Diplomarbeit

Das Kapitel 2 zeigt verschiedene Verfahren zur Suche von Dokumenten im Internet auf. Dazu werden zunächst theorieorientierte Erkenntnisse erläutert (Abschnitt 2.1), die die Grundlage für die vorgestellten verweis- (s. 2.2) und suchmaschinenbasierten (s. 2.3) Suchverfahren bilden.

Nach einer Diskussion der in Kapitel 2 vorgestellten Ergebnisse und Verfahren, gibt Kapitel 3 einen Überblick über die Funktionsweise PageTrackers. Im Anschluss werden die, durch die Diskussion motivierten, verwendeten Suchverfahren beschrieben.

In Kapitel 4 wird dann der Versuchsaufbau und die Ergebnisse der durchgeführten Experimente erläutert. Sie bilden die Basis für die anschließende Evaluation des Agenten, der im Rahmen der Diplomarbeit entstanden ist.

Kapitel 5 fasst die wesentlichen Ergebnisse der Diplomarbeit aus Kapitel 3 und 4 zusammen. Es schließt mit einem Ausblick auf die möglichen Erweiterungs- und Vertiefungsmöglichkeiten der vorliegenden Arbeit.

2 Suchverfahren

Dieses Kapitel beschäftigt sich mit verwandten Arbeiten zur Suche im Internet. Da die in den Abschnitten 2.2 und 2.3 vorgestellten Arbeiten implizit oder explizit bestimmte Eigenschaften des World–Wide–Web voraussetzen, werden in Abschnitt 2.1 zunächst einige allgemeine Überlegungen zu den Strukturen und Eigenschaften des World–Wide–Webs vorgestellt.

Die Suche nach Dokumenten lässt sich grob in zwei Klassen einteilen:

- Die verweisbasierte Suche nach bestimmten Dokumenten.
  Die gesuchten Dokumente gehören meistens zu einem bestimmten Themengebiet und es sollen möglichst nur passende Dokumente besucht werden. Die Anfrage an die jeweils genutzte Suchmaschine steht schon zum Zeitpunkt der Entwicklung fest.
- Die Nutzung einer indexbasierten Suchmaschine.
  Indexbasierte Suchmaschinen sammeln Informationen über alle Dokumente, die sie erreichen können. Zum Zeitpunkt der Entwicklung der Suchmaschine steht nur die Index-Struktur fest, aber nicht die Anfrage.

Beide Verfahrensansätze könnten bei der Suche nach einem bestimmten Dokument hilfreich sein.

In Abschnitt 2.2 wird zunächst auf die verweisbasierte Suche eingegangen. In Abschnitt 2.3 folgt die Beschreibung der Arbeitsweise von indexbasierten Suchmaschinen.

Abgeschlossen wird das Kapitel durch eine Zusammenfassung der Vor– und Nachteile der beschriebenen Verfahren.

2.1 Strukturen im WWW

2.1.1 Annahmen

Die grundlegende Annahme, auf der die folgenden Überlegungen aufbauen, ist, dass das World–Wide–Web (WWW) als Graph modelliert werden kann.

Um weitere Aussagen über die Eigenschaften dieses WWW-Graphen machen zu können, benutzten Albert, Jeong und Barabási ([AJB99]) ein lokales Verbindungsmaß, um ein topologisches Modell zu entwickeln, welches es ermöglichen sollte, die Eigenschaften des großen Netzes zu erforschen und zu charakterisieren.

In den folgenden Abschnitten werden weitere Eigenschaften des benutzten Graphen-Modells vorgestellt. Alle Ansätze gehen davon aus, dass die Verteilung der Links in

den Dokumenten sog. Potenzgesetz-Verteilungen[1] unterliegen. Sie kommen jedoch zu unterschiedlich Ergebnissen bzgl. der Exponenten der Verteilungsfunktion (vgl. [AJB99, FFF99, BKM+00]).

2.1.2 Der Durchmesser des WWW

Ein Maß zur Charakterisierung eines Graphen ist sein Durchmesser. Der Durchmesser eines Graphen ist der längste Pfad unter allen kürzesten Pfaden, die im Graphen existieren. Da das World–Wide–Web (WWW) zwar endlich, aber aufgrund seiner Dynamik und Größe nicht vollständig erfassbar ist, wird in diesem Fall der durchschnittliche Abstand zweier beliebiger Knoten im Graph als Durchmesser bezeichnet.

Um die lokale Verbindungsstruktur des WWW zu erfassen, wurde in [AJB99] ein Roboter entsandt, der rekursiv die Links besuchter Dokumente aufsammelt. Es wurde dazu das gesamte Netz der `nd.edu`–Domain, das zum Zeitpunkt des Experiments 325279 Dokumente und 1469680 Links enthielt, erfasst.

Die gesammelten Daten wurden genutzt, um die Wahrscheinlichkeiten $P_{in}(k)$ und $P_{out}(k)$ zu ermitteln. $P_{in}(k)$ ist die Wahrscheinlichkeit, dass in k Dokumenten Verweise auf ein Dokument S existieren. $P_{out}(k)$ bedeutet, dass ein Dokument Verweise auf k andere Dokumente enthält.

Es stellte sich heraus, dass die Wahrscheinlichkeiten einer Potenzgesetz-Verteilung gehorchen (s. Abb. 2.1).

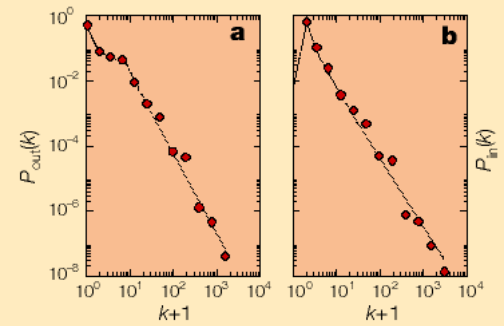


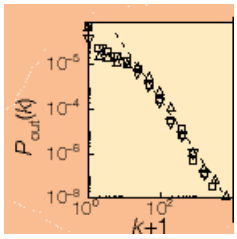
Abbildung 2.1: Link–Verteilung im World-Wide-Web (Quelle: [AJB99])

Das bedeutet einerseits, dass es signifikant viele Dokumente gibt, die viele Links enthalten. Andererseits gibt es auch eine große Anzahl von Dokumenten, auf die von vielen anderen Dokumenten verwiesen wird.

Um die Ergebnisse aus der `nd.edu`–Domain zu validieren, wurden Kontrollversuche auf den Domains von `whitehouse.gov` und `yahoo.com` gemacht (s. Abb. 2.2). Die Kontrollversuche bestätigten die ersten Ergebnisse.

Mit der gefundenen Verteilung wurde ein Graph modelliert, der dieser Verteilung gehorcht. In diesem Graph wurde die Länge der kürzesten Pfade bestimmt. Es stellte

[1] Potenzgesetzverteilungen sind Verteilungen der Form: $P(X = x) \sim x^{-a}$



Dreiecke: `yahoo.com`
Quadrate: `whitehouse.gov`

Abbildung 2.2: Kontrollversuche (Quelle: [AJB99])

sich heraus, dass die durchschnittliche Länge $< d >$ der kürzesten Pfade zwischen allen Knoten des Graphen sich durch $< d >= 0.35 + 2.06log(N)$ berechnen lässt. Für die Gesamtmenge der aktuell (in 2003) geschätzten, öffentlich erreichbaren 2 Billionen Webdokumente würde der kürzeste Pfad zwischen zwei beliebigen Dokumenten durchschnittlich 26 Schritte lang sein, sofern in jedem Knoten auf dem Pfad der richtige Link ausgewählt würde.

Der Durchmesser eines solchen Graphen ist also, gemessen an der Gesamtanzahl der Dokumente, klein.

2.1.3 Ist das gesamte WWW miteinander verbunden?

Wenn das Web also einen Durchmesser von 26 hat, kann dann von einer beliebigen Web-Seite jede beliebige andere Web-Seite erreicht werden? Diese Frage wird in [BKM$^+$00, KRR$^+$00] beantwortet.

Dort wird das in Abb. 2.3 gezeigte Modell des Web-Graphen zugrunde gelegt.

Das Modell geht davon aus, dass das Web im Wesentlichen in vier Teile zerteilt werden kann. In eine starke Zusammenhangskomponente SCC, deren Knoten alle so miteinander verbunden sind, dass von jedem Knoten in SCC jeder andere Knoten in SCC zu erreichen ist. Eine IN–Dokumenten–Menge, die alle Dokumente enthält, die in die Menge SCC hinein verweisen, jedoch nicht untereinander verbunden sind, sowie eine OUT–Dokumenten–Menge, die diejenigen Dokumente enthält, auf die von Knoten aus SCC verwiesen wird. Die Dokumente in OUT sind nicht untereinander verbunden. Schließlich gibt es eine Reihe von Links, die auf Dokumente verweisen, die nicht in den SCC hinein verweisen und auch nicht von diesem erreicht werden können. Sie werden „Tendrils“ genannt. Daneben gibt es noch eine, verhältnismäßig kleine Menge von nicht verbundenen, isolierten Dokumenten, die nicht weiter betrachtet werden. Zwischen den Mengen IN und OUT existieren sog. Tubes, die darstellen sollen, dass es Dokumente in IN gibt, die direkt auf Dokumente in OUT verweisen.

Pfade existieren im Modell von [KRR$^+$00] nur zwischen Knoten $u \in IN \cup SCC$ und Knoten $v \in SCC \cup OUT$. Bei den in Abb. 2.3 dargestellten Größenverhältnissen der Mengen von $\|OUT\| = \|IN\| = \|Tendrils\| = 44$Mio. und $\|SCC\| = 56$Mio. beträgt die Wahrscheinlichkeit, dass bei einem zufälligen, blinden Ziehen eines Dokuments aus der Gesamtdokumentenmenge, der Knoten $u \in IN \cup SCC$ ist, $P(u \in IN \cup SCC) \approx 0.5$. Entsprechend ist die Wahrscheinlichkeit, dass der Knoten $v \in SCC \cup OUT$ ist,

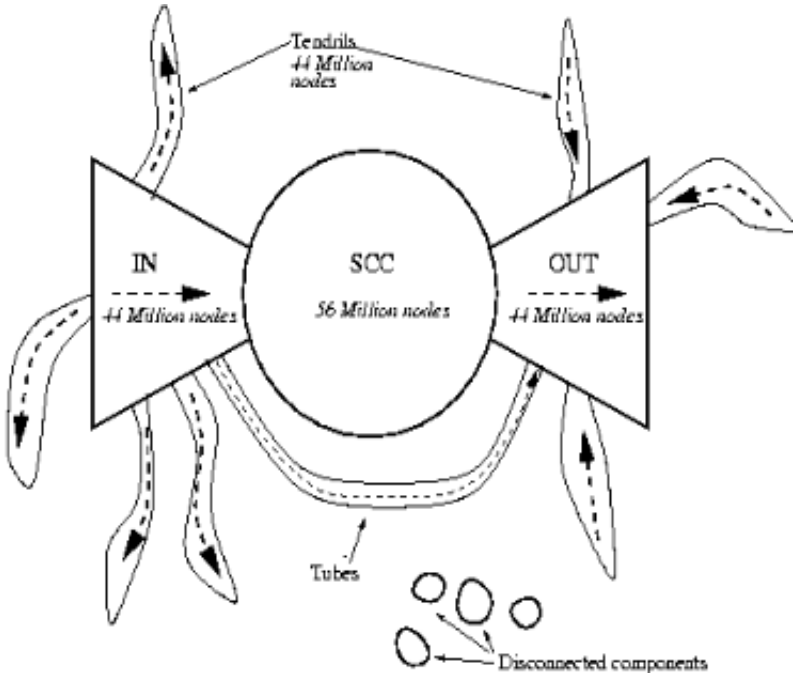


Abbildung 2.3: Das Web als „Fliege“ (Quelle: [BKM+00])

$P(v \in SCC \cup OUT) \approx 0.5$ (vgl. [KRR+00]). Bei dieser Art der Bestimmung der beiden Dokumente, ist die Wahl des einen Dokuments unabhängig von der Wahl des anderen. Da Pfade nur zwischen Knoten u und v mit den zuvor definierten Eigenschaften existieren können, beträgt die Wahrscheinlichkeit eines Pfads zwischen u und v

$$\begin{aligned} P\Big(\exists : \mathrm{Pfad}(u,v)\Big) &= P(u \in IN \cup SCC) \cdot P(v \in SCC \cup OUT) \\ &= 0.25 \end{aligned} \tag{2.1}$$

$$\begin{aligned} P\Big(\neg\exists : \mathrm{Pfad}(u,v)\Big) &= 1 - P\Big(\exists \mathrm{Pfad}(u,v)\Big) \\ &= 0.75 \end{aligned} \tag{2.2}$$

Dieses Ergebnis zeigt, dass ca. 75% des World Wide Web gar nicht miteinander verbunden sind. Die Ergebnisse aus 2.1.2, müssen daher relativiert werden, denn bei nicht verbundenen Dokumenten, gibt es keinen kürzesten Pfad oder die Länge wird als unendlich definiert.

Edge type	In-links (directed)	Out-links (directed)	Undirected
Average connected distance	16.12	16.18	6.83

Tabelle 2.1: Durchmesser des WWW nach [BKM+00]

Tabelle 2.1 zeigt die Durchmesser, die [BKM+00] bei den Experimenten ermittelten. Der Versuchsaufbau war dem aus [AJB99] ähnlich, jedoch wurde auf einer erheblich

größeren Dokumentenmenge gearbeitet. Die Spalte Undirected gibt den Durchmesser an, der existierte, wenn die Verweise ungerichtet wären, so dass eine Navigation zwischen den Dokumenten in beliebiger Richtung möglich wäre.

Dieses Ergebnis hat zwei Konsequenzen:

1. Es können offenbar nicht beliebige Seiten von beliebigen anderen Seiten erreicht werden.

2. Wenn zwei Dokumente überhaupt miteinander verbunden sind, dann ist der durchschnittliche Abstand zwischen diesen beiden Dokumenten kleiner als der in [AJB99] errechnete.

2.1.4 Welche Dokumente werden verbunden?

Web-Autoren neigen dazu, in ihren Dokumenten auf andere Dokumente zu verweisen, die ihrem thematisch nahe stehen. Das ist nicht überraschend, sondern gleicht der Art und Weise wie z. B. wissenschaftliche Arbeiten andere Arbeiten zitieren, die einerseits thematisch nahe an der eigenen sind und andererseits geeignet sind, die eigenen Aussagen zu belegen (vgl. [Hay00a, Hay00b, CHH98, NW01]). Das bedeutet, dass die Relevanz eines Dokuments bzgl. eines gegebenen Themas mit der Entfernung zu einer Seite, die als Referenz für ein Thema betrachtet wird, abnimmt. Oder umgekehrt formuliert: Je kürzer der Abstand zu einer solchen Referenzseite, desto wahrscheinlicher ist es, dass sich die Seite ebenfalls mit dem gegebenen Thema beschäftigt.

[BCHR01] beschreibt eine weitere Eigenschaft: Es werden mehr Verweise in die eigene (Top–Level–) Domäne gesetzt als in andere Domänen.

2.1.4.1 Aussagekraft von Verweisen

Nach den Erkenntnissen aus 2.1.4 stellt sich die Frage, ob die Verweise genutzt werden können, um Informationen über dahinter liegende Dokumente zu bekommen ohne diese betrachten zu müssen. Zu diesem Zweck erfolgreich genutzt werden z. B. der Ankertext eines Verweises zusammen mit den Worten um ihn herum (vgl. 2.2.1.3, 2.2.1.4, 2.3.2.2) oder die Verweisstruktur (vgl. 2.2.1.1, 2.2.1.2). Es wird z. B. untersucht, welche Dokumente auf das zu bewertende Dokument verweisen ([PBMW98, CDK$^+$99]. Alternativ kann berücksichtigt werden, auf welche Dokumente das aktuelle Dokument verweist ([Kle99]).

Bei dieser Art der Bewertung wird ein Dokument nicht ausgehend von seinem Inhalt, sondern aufgrund der Dokumente, die auf das Dokument verweisen bewertet. Die Bewertungsfunktion macht sich dabei die Verweise und die Verweisstruktur innerhalb derer das zu bewertende Dokument liegt zu nutze.

Gut gesetzte und beschriftete Verweise, können also zur Analyse von Dokumenten genutzt werden, ohne dass diese Dokumente selbst betrachtet werden müssen.

2.2 Verweisbasierte Suche

Dieser Abschnitt stellt verschiedene Suchverfahren vor, die versuchen zu einem gegebenen Dokument andere passende Dokumente zu finden. Alle hier beschriebenen Verfahren nutzen einen Webspider, um die Dokumente aus dem WWW zu holen. Die Verfahren unterscheiden sich im Wesentlichen in der Reihenfolge, in der die bisher gefundenen Verweise abgearbeitet werden. Zur Suche werden im Grunde die in der Künstlichen Intelligenz (KI) gebräuchlichen Verfahren, A^* (s. 2.2.1) und Breitensuche (s. 2.2.3) verwendet.

Die Abschnitte 2.2.1.3 bis 2.2.1.5 stellen Verfahren vor, die sich vor allem auf Klassifikatoren stützen.

Die meisten Verfahren gehen davon aus, dass ein guter Startpunkt gefunden wurde, von dem aus weitere relevante Dokumente gefunden werden können. Wie der geeignete Startpunkt gefunden wird, ist jedoch unterschiedlich. In einigen Ansätzen werden die Startpunkte manuell ausgewählt (vgl. 2.2.1.4). In anderen werden indexbasierte Suchmaschinen nach Schlüsselwörtern gefragt (vgl. 2.2.1.2).

2.2.1 Best First Search

Diese Art der Suche bewertet je nach Verfahren URLs oder Dokumente. Die URL oder das Dokument mit der besten Bewertung wird ausgewählt, um die Suche fortzusetzen.

Offensichtlich sind die Bewertungsverfahren die wichtigsten Komponenten in diesen Systemen, da es die Suchreihenfolge und den Erfolg des Verfahrens wesentlich beeinflusst. Im folgenden werden einige erfolgreiche Bewertungsverfahren erläutert.

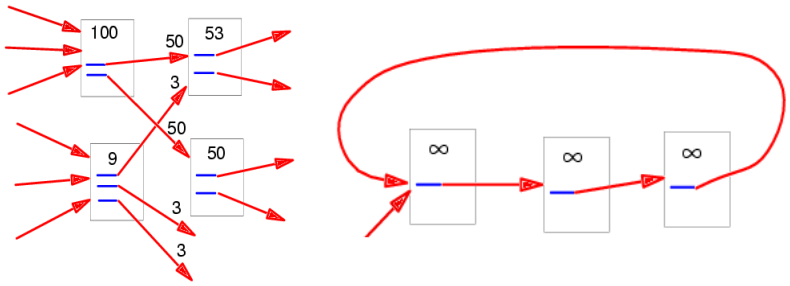
2.2.1.1 PageRank

PageRank ist ein Verfahren, welches Dokumente aufgrund ihrer Position in einer gegebenen Linkstruktur bewertet (vgl. [PBMW98]). Zur Ermittlung des PageRank $PR(D)$ eines Dokuments D wird folgende, zunächst vereinfachte, rekursive Formel benutzt (vgl. [Hua]):

$$\begin{aligned} PR(D) &= (1-d) + d \cdot \left(\frac{PR(T_1)}{C(T_1)} + \ldots + \frac{PR(T_n)}{C(T_n)} \right) && \text{[BP98]} \\ \text{oder} & \\ PR(D) &= d \cdot \sum_{t \in T} \frac{PR(t)}{C(t)} && \text{[PBMW98]} \end{aligned} \tag{2.3}$$

Dabei ist d ein Dämpfungsfaktor, $C(T_i)$ die Anzahl der Verweise in einem Dokument T_i und die Menge $T := \{T_1, \ \ldots, T_n\}$, die Menge der Dokumente, die auf D verweisen. Die Division des PageRanks durch die Anzahl der Verweise führt dazu, dass der PageRank von T_i gleichmäßig über alle Verweise des Dokuments T_i verteilt wird (s. Abb. 2.4(a)).

Die Berechnung des PageRanks kann auch anders formuliert werden (vgl. [Hua]): Sei A eine $n \times n$–Matrix und n die Anzahl der Dokumente. Dann ist $A_{u,v} = \frac{1}{C(u)}$, falls ein



(a) PageRankverteilung (b) PageRank-Schleifen

Abbildung 2.4: PageRank (Quelle: [PBMW98])

Verweis von u zu v existiert und $A_{u,v} = 0$ sonst. R seien die PageRanks aller Dokumente u als Spaltenvektor untereinander geschrieben. Dann lässt sich R so schreiben:

$$R = d \cdot A \cdot R \tag{2.4}$$

R ist also ein Eigenvektor der Verbindungsmatrix A mit Eigenwert d.

Ein Problem stellen die in Abb. 2.4(b) gezeigten Schleifen dar. Sie führen in den in Gleichung 2.3 gezeigten Formeln zu Endlosschleifen.

Um dieses Verhalten zu verhindern (vgl. [PBMW98]), wird in die PageRank-Berechnung ein Vektor $E(u)$ eingefügt (s. Gl. (2.5)).

$$PR^{'}(u) = d \cdot \sum_{v \in B_u} \frac{PR^{'}(v)}{C(v)} + d \cdot E(u) \tag{2.5}$$

B_u ist in Gleichung (2.5) die Menge der Dokumente v mit einem Verweis auf das Dokument u.

$E(u)$ stellt ein zufälliges Springen zu anderen Dokumenten dar. Zusätzlich wird verlangt, dass die Betragssummennorm (auch L_1-Norm genannt) des PageRanks $\|PR^{'}\|_1 = 1$ ist *und* der Dämpfungsfaktor d maximiert wird. Ein unendlich großer PageRank wird so verhindert, da d jeweils so angepasst werden muss, dass die Bedingung $\|PR^{'}\|_1 = 1$ erfüllt wird.

PageRank arbeitet gut auf großen Datenmengen [PBMW98], jedoch benötigt die Berechnung verhältnismäßig viel Zeit [NW01] und sollte daher offline erfolgen. Auf kleinen Datenmengen ist die Aussagekraft des PageRanks eingeschränkt, da in diesem Fall einzelne Dokumente ein zu hohes Gewicht erhalten. Es fehlen andere Dokumente, die die Verweiskraft dieser Dokumente relativieren (vgl. [CGMP98]).

PageRank basiert in der Grundversion nur auf der Linkstruktur und berücksichtigt die Anfrage dabei nicht. Dadurch ist PageRank zunächst nicht zur themenspezifischen Suche geeignet. [Hav02] versucht dieses Problem zu lösen, indem der Vektor E(u) in Gleichung (2.5) angepasst wird. Statt einer gleichmäßigen Verteilung der Werte der Einträge in E(u), werden Einträge, die zu Dokumenten, welche zu dem gesuchten Thema passen, auf $\frac{1}{N_R}$ (N_R ist die Gesamtanzahl der Referenzdokumente für ein Thema) und alle anderen Einträge auf Null gesetzt. Die Ermittlung einer initialen Menge von Seiten, die

zum gewünschten Thema passen, geschieht von Hand. Eine Benutzeranfrage wird einem Themenschwerpunkt zugeordnet. Die Zuordnung erfolgt z. B. durch einen Klassifikator. Durch den Themenschwerpunkt werden die zuvor errechneten themensensitiven PageRank-Werte für die Anfrage festgelegt. Die Bewertung für ein Dokument berechnet sich aus der zugehörigkeitsgewichteten Summe der Einzelwerte für jeden Themenschwerpunkt.

Da die PageRank–Berechnung auch auf aktuellen Rechnern lange dauert, nimmt folglich auch die themenbasierte Suche mit Hilfe von PageRank–Algorithmen viel Rechenzeit in Anspruch. Um diese Rechenzeit zu verkürzen, wurde in [KHMG03, KHG03] ein Verfahren vorgestellt, welches sich zu Nutze macht, dass innerhalb einer Domäne die meisten Dokumente in die eigene Domäne verweisen. Das führt laut [KHG03] dazu, dass der PageRank vieler Dokumente schnell konvergiert. Der sogenannte „BlockRank“–Algorithmus versucht diese Eigenschaft auszunutzen. Sobald der PageRank eines Dokuments konvergiert, wird in den nachfolgenden Iterationen dieser PageRank nicht erneut berechnet.

Ferner kann laut [KHMG03] der PageRank der Dokumente innerhalb eines Domänenblocks zunächst getrennt vom restlichen Netzwerk berechnet werden. In einem weiteren Schritt werden die gebildeten Blöcke als Dokumente im Sinne von PageRank aufgefasst. Für diese Block–Dokumente werden PageRanks (die „BlockRanks“) gebildet. Im dritten Schritt können die BlockRanks mit den PageRanks der Dokumente innerhalb der Blöcke verrechnet werden, um eine Approximation der globalen PageRanks zu erhalten.

Die Approximation dient als Startverteilung der globalen PageRank–Berechnung, die mit diesen Startwerten schneller konvergieren soll.

Durch die Bildung von Blöcken, kann das PageRank–Verfahren besser parallelisiert bzw. verteilt werden (s. [KHMG03]), so dass schon dadurch eine Geschwindigkeitssteigerung erreicht wird.

2.2.1.2 HITS

Einen ähnlichen Ansatz wie PageRank (s. 2.2.1.1) verfolgt Hyperlink–Induced–Topic–Search (HITS) [CDK+99]. Bei diesem Ansatz werden zwei verschiedene Arten von Knoten im Dokumentverweisgraphen unterschieden.

Authorities sind Dokumente, die sehr relevant für ein gegebenes Thema sind. Sie enthalten die gewünschten Informationen.

Hubs sind Dokumente, die Sammlungen von Verweisen auf Authorities enthalten. Sie bilden gute Startpunkte für eine Informationssuche.

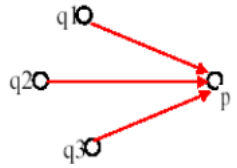
Um Hubs und Authorities zu finden, werden zunächst z. B. 200 Dokumente durch eine Suchmaschinenabfrage (z. B. Altavista[2]) mit der gegebenen Anfrage als Startmenge gesammelt. Die zurückgelieferten Dokumente müssen nicht alle passend zum Thema sein, aber es ist wahrscheinlich, dass einige dieser Dokumente Verweise auf sehr gute Authorities im Gebiet der Suchanfrage haben. Daher wird die Startmenge um alle Dokumente erweitert, auf die von den Startmengendokumenten verwiesen wird. Zusätzlich werden Dokumente hinzugenommen, die auf Dokumente in der Startmenge verweisen.

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

Um Verfälschungen des Ergebnisses zu vermeiden, werden alle Links zwischen Dokumenten, die aus derselben Domain stammen, aus der Startmenge entfernt. Solche Verweise kommen häufig vor und sagen i. A. nichts über die Güte des Dokuments als Hub oder Authority aus.

Zur Berechnung der Authorities und Hubs, wird jedem Dokument P ein nicht–negatives Authority–Gewicht x_p und ein nicht-negatives Hub–Gewicht y_p zugewiesen. Dabei sind die relativen Werte zueinander wichtig — nicht die absoluten Größen. Im in [CDK+99] beschriebenen Ansatz, werden die Gewichte nach oben begrenzt. Die Summe der Quadrate aller Gewichte soll 1 sein. Als Ergebnis der Berechnung wird ein Dokument mit einem relativ großen Authority–Gewicht als Authorität bzgl. eines Themas betrachtet und ein Dokument mit einem relativ großen Hub–Gewicht als guter Hub bezeichnet.

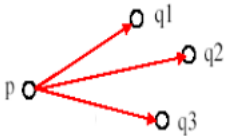
Zu Beginn der Berechnung wird allen Gewichten aller Dokumente derselbe konstante positive Wert zugewiesen. Im Zuge der Berechnung werden die Gewichte wie folgt aktualisiert:



$$x_p = \sum_{q,q \to p} y_q \tag{2.6}$$

q_i sind Hubs, p ist Authority

Abbildung 2.5: Authority–Score–Berechnung



$$y_p = \sum_{q,p \to q} x_q \tag{2.7}$$

p ist Hub, q_i sind Authorities

Abbildung 2.6: Hub–Score–Berechnung

Die rekursiven Gleichungen konvergieren, wenn x und y positiv sind, was durch die beschriebene Initialisierung erreicht wird. Wenn das Verweisgefüge als Adjazenzmatrix A aufgeschrieben wird und entsprechend die Hub– und Authority–Werte gemeinsam als Vektor aufgefasst werden, dann konvergiert $\vec{x}$ gegen den dominanten Eigenvektor von $A^T A$ und $\vec{y}$ gegen den entsprechenden Eigenvektor von AA^T.

HITS sucht, im Gegensatz zu PageRank (vgl. 2.2.1.1), aufgrund der Initialisierung durch Suchmaschinenergebnisse sofort themenspezifisch. Es kann jedoch aufgrund der benutzten Technik zu einer unerwünschten Generalisierung des gesuchten Themas kommen, weil bei sehr speziellen Anfragen nur sehr wenige Authorities zur Verfügung stehen, so dass das Verfahren zu einem übergeordneten oder im schlimmsten Fall zu einem beigeordneten Begriff generalisiert, je nachdem wie die zur Initialisierung befragte Suchmaschine mit der Anfrage umgeht. Es kann eine Themenverschiebung erfolgen, wenn die Hubs zu einem Thema auch mehrere andere Themen behandeln. Da die Hub–Bewertungen im beschriebenen Verfahren über alle Links gleich weiterpropagiert werden, kann eine Überbewertung nicht zum gesuchten Thema gehörender Verweise erfolgen. Das führt zu einer

Kettenreaktion, so dass auch die weiterverweisenden Dokumente dieser Fehl–Verweise höhere Werte erhalten als erwünscht. Das dritte Problem sind Verweise, die in fast jedem Dokument zu finden sind (z. B. Verweise auf den Acrobat Reader). Das führt dazu, dass die Adobe–Website für den Algorithmus stets als Authority erscheint, obwohl sie mit den meisten Suchthemen nichts gemeinsam hat.

Diesem Problem soll mit den folgenden Heuristiken begegnet werden:

- der Text um die Links herum wird zusätzlich in die Bewertung einbezogen. Wenn der Text z. B. Begriffe aus der Anfrage enthält, wird der Link höher gewichtet.
- Große Hub–Seiten werden künstlich in kleinere Segmente aufgeteilt, dadurch soll die Themenvermischung verhindert werden.

2.2.1.3 Webwatcher

Der in [AFJM95, JFM97] vorgestellte Agent Webwatcher soll Benutzern des Internets helfen für sie relevante Dokumente zu finden. Dazu versucht WebWatcher die folgende Zielfunktion zu lernen:

$$\text{LinkUtility} : \text{Page} \times \text{Goal} \times \text{User} \times \text{Link} \tag{2.8}$$

wobei *Page* das aktuelle Dokument ist, *Goal* die Information, die vom Benutzer gesucht wird, $User$ die Identität des Nutzers und *Link* einer der Links, die auf der Seite *Page* vorhanden sind. *LinkUtility* ist die Wahrscheinlichkeit, dass das Folgen des Links in *Page* das Ziel *Goal* für den Nutzer $User$ erfüllt. Das Lernen dieser Funktion ist relativ schwierig, da die Trainingsdaten schwer zu sammeln sind. Daher wird die Funktion (2.8) durch die Funktion

$$\text{UserChoice} : \text{Page} \times \text{Goal} \times \text{Link} \tag{2.9}$$

approximiert. $UserChoice$ ist die Wahrscheinlichkeit, dass ein beliebiger Benutzer den Verweis *Link* auf der Seite *Page* verfolgen wird. Dabei werden die Benutzer nicht mehr unterschieden und das Verfolgen von *Link* impliziert nicht mehr, dass es das Ziel erfüllen wird.

Um die Seite, das Ziel und den Link zu repräsentieren, benutzt [AFJM95] den Verweis-Text, Wörter aus dem Satz, der den Verweis umgibt, sowie Wörter in den Überschriften des Absatzes, der den Verweis umschließt und schließlich die Wörter der Suchanfrage. Diese Wörter werden in einem booleschen Vektor zusammengefasst. Dieser Vektor kann mit Verfahren wie Winnow, Wordstat und TF–IDF in Kombination mit dem Cosinusmaß weiterverarbeitet werden.

Als Trainingsdaten werden dem Agenten die vom Benutzer verfolgten und nicht verfolgten Links verschiedener Seiten zur Verfügung gestellt.

In [JFM97] wird zusätzlich ein Reinforcement Learning basierter Ansatz vorgestellt, der dem in 2.2.1.4 vorgestellten Ansatz ähnlich ist, jedoch ein gewichtetes 3–nearest–neighbour–Modell (s. [Mit97]) zur Klassifikation nutzt.

2.2.1.4 Reinforcement Learning

Sowohl der von [RM99] als auch der von [MB00] vorgeschlagene Ansatz nutzt Reinforcement Learning, genauer Q–Learning (s. [WD92]), zur Bestimmung der Suchstrategie.

Ziel des Reinforcement Learnings ist die Erstellung eines Handlungsplans, der einen optimalen Handlungsablauf beschreibt und daher den Erfolg maximiert.

Gegeben sind eine Menge von Zuständen S, eine Reihe von Aktionen A und eine Zustandsüberführungsfunktion $T : S \times A \rightarrow S$. Gesucht ist derjenige Handlungsplan $\pi : S \rightarrow A$, für den der Erfolg

$$V(s) = \sum_{t=0}^{\infty} \gamma^t \cdot r_t (0 \leq \gamma < 1) \tag{2.10}$$

maximal ist. r_t ist der Erfolg zum Zeitpunkt t und γ ist ein Dämpfungsfaktor, der bewirken soll, dass späte Erfolge weniger wert sind als frühe Erfolge. Da sich die Gleichung (2.10) in dieser Form nicht berechnen lässt, wird stattdessen die Gleichung

$$Q^*(s,a) = R(s,a) + \gamma \cdot V^*(T(s,a)) \tag{2.11}$$

berechnet. Q^* ist die Belohnung die erlangt wird, wenn im Zustand s die Aktion a und anschließend die optimale Handlungsreihenfolge ausgeführt wird. $R(s,a)$ gibt dabei die Belohnung an, die der Lerner für die Handlung a in s bekommt, an. Durch diese Neuformulierung des Problems muss nun das folgende Problem gelöst werden:

$$\pi^*(s) = \max\{Q^*(s,a)\} \tag{2.12}$$

Das bedeutet, dass in jedem Zustand die Aktion ausgewählt werden muss, die den besten Q^*–Wert erwarten lässt. Für einen Webspider, der sich durch Reinforcement Learning weiterentwickeln soll, entspricht die Menge S der Menge der zu findenden Dokumente (als Zielzustände) vereinigt mit der Menge der gefundenen Links. Eine Aktion entspricht dem Verfolgen eines Verweises. Für jeden bekannten Verweis existiert eine solche Aktion. Die Menge aller Aktionen ist die Menge A. Eine Belohnung r wird vergeben, wenn ein zu suchendes Dokument erreicht wird.

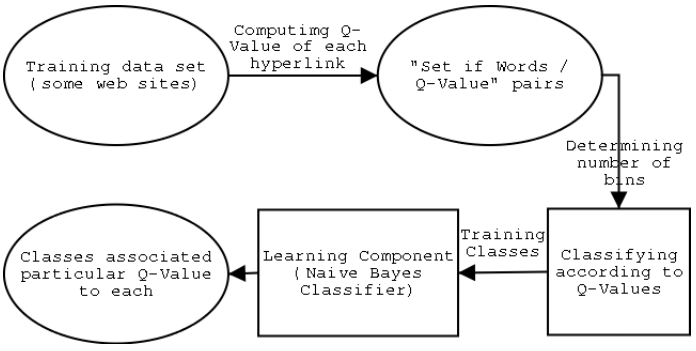


Abbildung 2.7: Reinforcement Learning nach [RM99] (Quelle: [BNAP])

Die Menge S ist zu groß, um sie handhabbar zu gestalten. Daher unterscheidet [RM99] die Zustände überhaupt nicht, so dass die Menge S nur einen Zustand enthält. Die zur Verfügung stehenden Aktionen werden ausschliesslich durch die Linktexte und Wörter

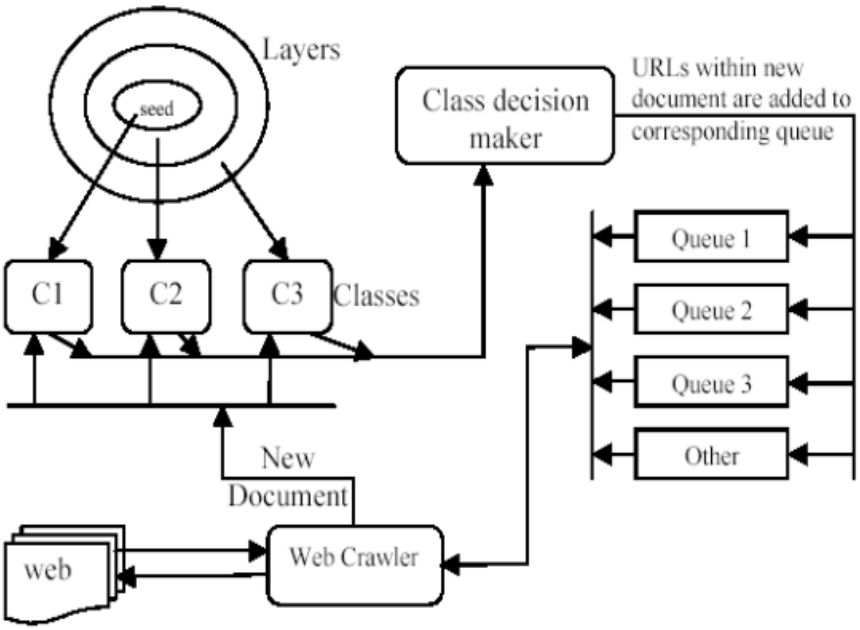


Abbildung 2.8: Context Focused Crawler (Quelle: [BNAP])

in der unmittelbaren Umgebung unterschieden (also z. B. nicht durch ihre Position im Dokument), so dass die Funktion $Q^*(s, a)$ zu einer Funktion degeneriert, die Wortmengen zu Q–Werten zuordnet. Diese Zuordnungsfunktion muss gelernt werden.

[RM99] lernt diese Funktion durch Naive Bayes Klassifizierer, die den Links Q–Werte zuweisen, wenn dieser zu ihrer Klasse gehört. Diese Klassifizierer müssen zunächst offline auf einer Menge von vorgegeben Trainingsdaten angelernt werden. Durch die Anzahl der Klassifizierer lässt sich einstellen wie sehr zukünftige Belohnungen berücksichtigt werden. Ein einfacher Spider, der nur sofortige Belohnungen berücksichtigt, unterscheidet nur zwei Klassen: Belohnung=1 und Belohnung≠1. Wenn mehr Klassen zugelassen werden, dann können auch gezielt Dokumente besucht werden, die erst später Erfolg versprechen.

[MB00] lernt diese Zuordnungsfunktion mit einem neuronalen feedforward Netzwerk. Die Eingabeknoten entsprechen in diesem Netz Schlüsselwörtern, die der Benutzer des Agenten vorgegeben hat. Die Eingabewerte für einen Verweis in die Knoten des Netzwerks werden aufgrund der Häufigkeit des Vorkommens und der Entfernung der vorgebenen Schlüsselwörter vom Verweis berechnet.

2.2.1.5 Context Focused Crawler

Beim Context Focused Crawler ([DCL+00]) werden als erstes eine Reihe von Zieldokumenten vom Benutzer vorgegeben. Anschliessend werden Suchmaschinen wie z. B. Google[3] genutzt, um Dokumente zu finden die Verweise (sog. Backlinks) auf die vorgegebenen Dokumente enthalten. Diese Backlinks werden dann aufgrund ihrer Entfernung zu den vorgegebenen Zieldokumenten in Klassen eingeteilt, die der Entfernung zum Ziel

[3] `http://www.google.com`

entsprechen. Es gibt also eine Klasse für alle Dokumente die einen Verweis von einem Zieldokument entfernt sind, eine für die, die zwei entfernt sind, usw. Für diese Klassen werden Naive Bayes Klassifizierer angelernt. Mit Hilfe dieser Klassifizierer werden während der Suche URLs in Warteschlangen eingeordnet. Jeder Warteschlange ist eine Klasse zugeordnet. Die Warteschlangen werden der Reihe nach abgearbeitet, d. h. es werden zunächst alle URLs in der ersten Warteschlange abgearbeitet und nur, wenn darin keine URL zu finden ist, wird zur nächsten übergegangen. Dadurch soll ein schneller Weg zum Ziel garantiert werden. Der gesamte Ablauf der Suche des Context Focused Crawler ist in Abbildung 2.8 dargestellt.

2.2.2 Suche in der Tiefe

Der in [CHH98] vorgestellte Ansatz hat es sich zum Ziel gesetzt, relevante Dokumente in einer möglichst großen Tiefe zu finden. [CHH98] legt dar, dass seiner Meinung nach, Dokumente, die durch Breitensuchen (s. 2.2.3) gefunden werden, so leicht zu finden sind, dass sich der programmiertechnische Aufwand für einen Agenten nicht lohnt. Chang fordert daher, dass auch weiter entfernte relevante Dokumente gefunden werden. Er geht davon aus, dass nach einer Reihe von irrelevanten Dokumenten wieder relevante Dokumente erreicht werden (vgl. Abb. 2.9). Daher stellt er einen Algorithmus vor, bei

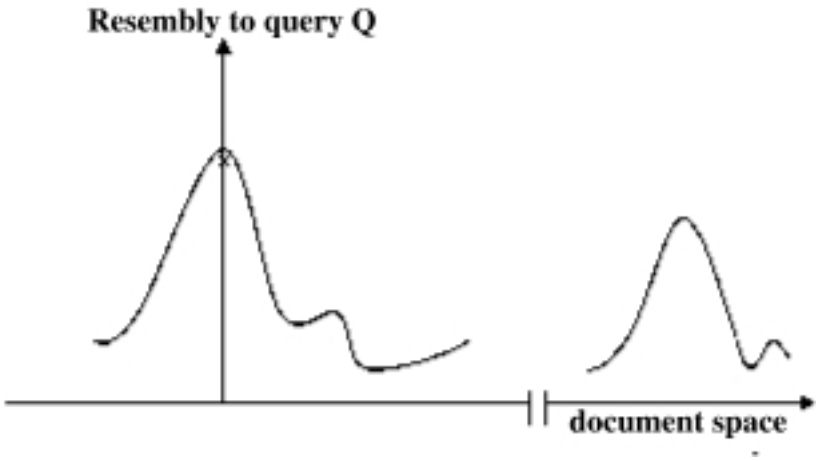


Abbildung 2.9: Relevanz von Dokumenten im Abstand zum Ursprungsdokument (Quelle: [CHH98])

dem versucht wird, möglichst viele Dokumente in grosser Tiefe zu finden, wobei maximal 200 Dokumente besucht werden sollen.

Es handelt sich bei diesem Ansatz nicht um eine Tiefensuche im eigentlichen Sinne!

Der Algorithmus weist vielmehr eine große Ähnlichkeit zu dem in [HJM+98] vorgestellten SharkSearch–Ansatz (s. auch 2.2.3) auf, wobei die Auswahl der zu besuchenden Links mit Hilfe eines probabilistischen Ansatzes verbessert wurden.

2.2.3 Breitensuche

Die Breitensuche ist ein Standardverfahren der Graphentheorie. Es werden die besuchten Links in der Reihenfolge ihres Auffindens abgearbeitet. Das bedeutet, dass jeder neu gefundene Link hinten an die Warteschlange der abzuarbeitenden Links angehängt wird.

Diese Art der Suche scheint zunächst wenig intelligent, doch andererseits wurde in 2.1.4 festgestellt, dass Seiten mit hoher Relevanz vor allem in geringer Entfernung zum Ursprungsdokument gefunden werden. Da die Breitensuche zunächst alle Dokumente der Tiefe 1 absucht und dann erst eine Ebene tiefer sucht, ist die Wahrscheinlichkeit, relevante Dokumente zu finden, hoch und rechtfertigt ihren Einsatz.

Durch die Breitensuche werden viele Dokumente besucht und geprüft. Das benötigt viel Zeit, doch [NW01] rechtfertigt den Einsatz der Breitensuche dadurch, dass fortgeschrittene Verfahren wie z. B. PageRank (vgl. 2.2.1.1) erhebliche Rechenzeit beanspruchen, um den PageRank einer Seite zu bestimmen. Diese Rechenzeit könne genauso gut für die Breitensuche verwendet werden.

Verfeinerungen des grundsätzlichen Breitensuchen–Algorithmus' — FishSearch und seine Verbesserung SharkSearch — werden in [DP94] und [HJM+98] vorgestellt. Dabei wird von der Analogie eines Fischschwarms ausgegangen. Wenn Fischschwärme Futter finden — im Kontext der webbasierten Suche entspricht das Futter den relevanten Seiten — dann vermehren sich die Fische. Sollte das Futter versiegen — die Relevanz der gefundenen Seiten sinkt —, so sterben die Fische. SharkSearch berücksichtigt zusätzlich die Relevanz der Elterdokumente, so dass Dokumente, die an sich nicht relevant sind, nicht zum sofortigen „aussterben“ des Fischs führen. Es handelt sich dabei eher um einen Multiagentenansatz, der in [HJM+98] auch expliziert wird.

Eine weitere Variante der Breitensuche unter Verwendung genetischer Algorithmen ist der in [Men97] vorgestellte Ansatz ARACHNID. Hier wird wie bei SharkSearch die Breitensuche auf mehrere Agenten verteilt, die aufgrund ihrer Energiepotentiale weiterleben. Im Unterschied zu SharkSearch erlaubt ARACHNID, dass die Einzelagenten verschiedene Suchstrategien verfolgen. Nur Agenten mit erfolgreichen Suchstrategien werden überhaupt geklont.

2.3 Suchmaschinenbasierte Suche

Suchmaschinen lassen sich grob in drei Klassen einteilen:

1. katalogbasierte Suchmaschinen
2. indexbasierte Suchmaschinen
3. Meta–Suchmaschinen

2.3.1 Katalogbasierte Suchmaschinen

Katalogbasierte Suchmaschinen werden auch als Web-Portal bezeichnet (vgl. [BNAP]). Web-Portale nehmen eine Einordnung der in ihnen enthaltenen Dokumente in eine Themenstruktur vor (vgl. [Loo00]). Dabei darf es vorkommen, dass ein Dokument in mehrere Gebiete eingeordnet wird. Die Einordnung kann von Menschen per Hand oder automatisch durch ein Programm (s. z. B. [Bor00, MNRS00, CDI98]) erfolgen. Die Einordnung durch Menschen kann noch einmal unterteilt werden: Die Einteilung kann durch wenige „Fachexperten“ oder „Administratoren“ oder von vielen Menschen, die sich selbst zum Experten auf dem von ihnen gewählten Gebiet ernennen (vgl. [Net99]) geschehen. Verzeichnisse der letzten Art werden „OpenDirectories“ genannt.

Der Vorteil der OpenDirectories soll die Arbeitserleichterung für den Einzelnen und das schnellere Wachstum des Verzeichnisses sein. Der Nachteil ist die mitunter schlechte Zuordnung von Dokumenten zu Themengebieten. Dieser Nachteil kann durch die ausschließliche Zuordnung durch nachgewiesene Fachexperten vermieden werden. Jedoch pflegen dann weniger Personen das Verzeichnis, so dass es neue Seiten und Veränderungen langsamer erfasst.

Durch Navigation durch die Verzeichnisstruktur soll in den Web-Portalen gesucht werden. Oft wird auch eine Schlüsselwortsuche in allen Einträgen des Verzeichnisses angeboten. Dabei werden jedoch nicht die Volltexte der erfassten Dokumente durchsucht, sondern die Textfragmente dieser Dokumente, die als Zusammenfassung angeboten werden (vgl. [Loo00]). Wie der Text zusammengefasst wird, ist von Katalog zu Katalog unterschiedlich. Es werden z. B. die ersten Sätze des Dokuments als Zusammenfassung angeboten oder ein Katalogverwalter fasst den Inhalt des verwiesenen Dokuments mit seinen eigenen Worten zusammen.

Beispiele für aktuelle katalogbasierte Suchmaschinen sind: Yahoo[4], das Open Directory Project[5] und LinuxLinks[6].

2.3.2 Indexbasierte Suchmaschinen

Indexbasierte Suchmaschinen versuchen einen sehr großen Teil der Dokumente des Internets zu erfassen, damit auf ihnen eine Volltextsuche durchgeführt werden kann.

Der Ablauf der Dokumentbeschaffung einer indexbasierten Suchmaschine (vgl. Abbildung 2.10) ist den in Kapitel 2.2 vorgestellten Ansätzen ähnlich. Bei den hier vorgestellten Suchmaschinen ist jedoch keine themenspezifische Suche die Motivation für die Datensammlung, sondern die Erstellung einer möglichst großen Dokumentensammlung und die Kenntnis der Dokumentenfundorte *auf Vorrat*. Mit diesem Wissensvorrat sollen später Anfragen, die zum Zeitpunkt des Sammelns noch nicht bekannt sind, beantwortet werden können.

Im Gegensatz zu den Ansätzen in Kapitel 2.2 sollen nicht nur Dokumente besucht werden, die zu einem Themengebiet passen, sondern möglichst viele unterschiedliche Dokumente in möglichst kurzer Zeit erfasst werden.

[4] `http://www.yahoo.com`
[5] `http://www.dmoz.org`
[6] `http://www.linuxlinks.com`

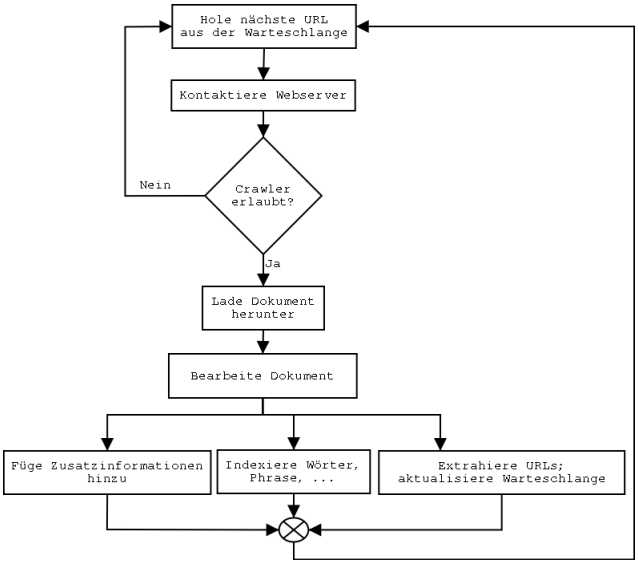


Abbildung 2.10: Ablauf eines „Webcrawls“ Quelle:[BNAP]

[Hua] zerlegt eine Suchmaschine in drei Komponenten: einem Indexierer, einem Dokumentsammler und einem Anfrage-Server. Der Indexierer bearbeitet die geladenen Dokumente und repräsentiert sie in einer effizienten Suchdatenstruktur. Der Dokumentensammler sammelt neue Dokumente und der Anfrage-Server nimmt die Benutzeranfragen entgegen.

Das Herzstück der indexbasierten Suchmaschinen sind die Web Crawler. Web Crawler sind diejenigen Programme, die die Datenbanken der Suchmaschinen aufbauen. Im folgenden werden die Web Crawler der Suchmaschinen Altavista (s. 2.3.2.1) und Google (s. 2.3.2.2) näher beschrieben, um einen Eindruck von der Arbeitsweise dieser beiden indexbasierten Suchmaschinen zu bekommen.

2.3.2.1 Mercator (Altavista)

Der Mercator Web Crawler ist ein skalierbarer, erweiterbarer, in Java geschriebener Web Crawler, der ursprünglich zu Forschungszwecken genutzt werden sollte, um Statistiken über das Internet zu erstellen (vgl. [HN99]. Mit skalierbar ist gemeint, dass der Crawler in der Lage sein soll, sehr viele Dokumente aus dem Internet herunter zu laden und zu verwalten. Die Erweiterbarkeit soll sich in einem modularen Aufbau des Crawlers widerspiegeln, so dass Dritte den Crawler an ihre eigenen Bedürfnisse anpassen können.

Die grundsätzliche Arbeitsweise des Mercator Web Crawlers gleicht der eines jeden Web Crawlers (vgl. Abb. 2.10):

- Entferne eine URL aus der URL-Liste.
- Bestimme die IP-Adresse des Host Namens.
- Lade das betreffende Dokument herunter.
- Extrahiere die im Dokument enthaltenen Verweise.

Für jeden Verweis:

- Bilde eine absolute URL (wandle relative Verweise um, wenn es notwendig ist)
- Füge die URL der URL-Liste hinzu, wenn sie nicht schon zuvor besucht wurde.
- Verarbeite die Seite weiter, wenn es gewünscht wird.

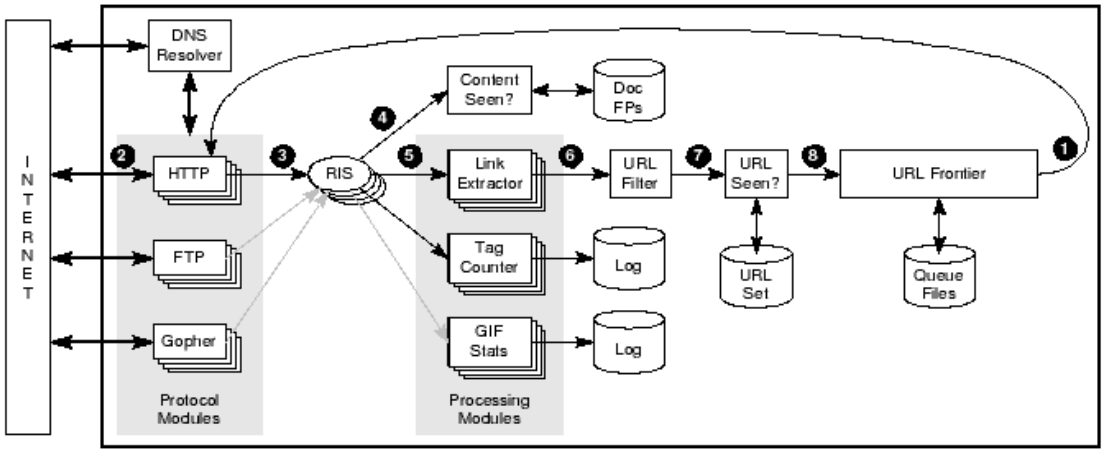


Abbildung 2.11: Architektur von Mercator (Quelle: [HN99])

Der Grund–Algorithmus basiert auf verschiedenen Komponenten (vgl. [HN99, NH01]). Die Nummern entsprechen denen in Abbildung 2.11.

1. Die sog. „URL Frontier“ stellt die Liste der noch abzuarbeitetenden URLs dar. Diese Liste wird von allen Arbeits-Threads gemeinsam genutzt. Aus ihr wird die aktuell zu bearbeitende URL entfernt.

   Der Aufbau der „URL Frontier“ bestimmt die Suchstrategie. Diese Komponente kann prinzipiell durch benutzereigene ersetzt werden. So ist es möglich, in Mercator verschiedene Strategien zu implementieren. Die Standard-Strategie, die auch von den meisten anderen Web Crawler verwendet wird, ist die Breitensuche. Um die Breitensuche zu realisieren, werden die neu gefundenen URLs in einer First–In–First–Out–Warteschlange (FIFO) verwaltet. Da der Mercator Crawler ein Multi-Threaded Crawler ist, muss dafür Sorge getragen werden, dass die besuchten Web-Server nicht durch eine Anfragenflut überlastet werden. Um das zu gewährleisten, wird jedem Mercator–Thread eine Unterwarteschlange zugeteilt, die sicherstellt, dass jeder Web-Server zu jedem Zeitpunkt von höchstens einem Arbeitsthread befragt wird. Es wird also immer nur eine Anfrage zur selben Zeit an einen Web-Server gestellt.

2. Im zweiten Arbeitsschritt wird versucht, das zur URL gehörige Übertragungsprotokoll zu identifizieren. Je nach Protokolltyp (z. B. HTTP oder FTP) wird ein entsprechendes Protokollmodul mit dem Beschaffen des Dokuments beauftragt. Die vorhandenen Module können vom Benutzer des Mercator Web Crawlers konfiguriert und durch eigene erweitert werden.

3. Die heruntergeladenen Daten werden in einen sog. „RewindInputStream (RIS)“ hinein gelesen, um die nachfolgenden Dokument-Analysen zu ermöglichen.

4. In diesem Schritt wird getestet, ob der heruntergeladene Inhalt schon einmal heruntergeladen wurde (z. B. von einer anderen URL). Falls das der Fall ist, wird das Dokument nicht weiterbearbeitet.

   Dieser Schritt verhindert, dass Dokumente, die aufgrund so genannter Mirror–Sites häufig im Netz vorkommen, die Geschwindigkeit des Web Crawlers senken. Das erneute Verarbeiten des Dokuments bringt dem Crawler–Betreiber keinen Gewinn, da für die darüber liegende Suchmaschine im Allgemeinen ein Exemplar des Dokuments als Fundstelle reicht.

   Mercator implementiert diese Doppelte-Inhalte-Überprüfung durch die Generierung einer 64-Bit–Checksumme der Dokumente. Bei dieser Checksumme ist es laut [HN99] beweisbar sehr unwahrscheinlich, dass für zwei unterschiedliche Zeichenketten dieselben Checksummen berechnet werden.

5. Im Schritt 5 werden die Dokument–Inhalte durch so genannte „Processing Modules“ analysiert. Je nach Multipurpose Internet Mail Extensions–Type (MIME-Type), wird ein entsprechendes Modul vom Web Crawler auf dem Dokumentinhalt ausgeführt. Ebenso wie die Protokollmodule lassen sich auch die Verarbeitungsmodule vom Benutzer erweitern und ergänzen.

6. Die im Schritt 5 begonnene Analyse der gefundenen Verweise wird fortgesetzt. Es wird geprüft, ob die URL einem benutzerdefinierten URL-Filter genügt. So kann der Benutzer den laufenden Suchvorgang z. B. auf Server einer bestimmten Domain einschränken.

7. Wenn eine URL die Filterkriterien des vorangegangenen Schritts erfüllt, wird geprüft, ob die neu gefundene URL zuvor schon einmal gefunden wurde. Ist das der Fall, wird die URL nicht erneut in die Warteschlange aufgenommen.

8. Wenn die neue URL bisher unbekannt war, wird sie in diesem Schritt zur „URL Frontier“ hinzugefügt und der Gesamtarbeitsablauf beginnt bei Schritt 1 erneut.

Laut [Lab01] ist der Mercator Web Crawler nun in die „AltaVista Search Engine 3“ integriert worden. Das ist insofern bedauerlich, weil dadurch der Zugang zu diesem Web Crawler stark eingeschränkt wurde. So steht auf der Informationsanfrageseite[7] von Altavista Software Solutions (Stand: 11.05.2003) geschrieben:

> **NOTE:** AltaVista Software's Products are available for corporate enterprise implementations exclusively. They are not intended for nor made available to individual users.

Damit steht der Mercator Web Crawler nicht mehr für die allgemeine Öffentlichkeit zur Weiterentwicklung zur Verfügung.

2.3.2.2 Google

Google ist eine weitere und im Moment die populärste[8] Suchmaschine. Die Entwickler dieser Suchmaschine beschreiben ihr Ziel in [BP98] ähnlich wie die Entwickler von Mercator (vgl. 2.3.2.1). Sie wollten eine Suchmaschine entwickeln, die in der Lage ist, sehr

[7] `http://solutions.altavista.com/en/company/requestinfo_en.shtml`
[8] bezogen auf die Nutzerzahlen lt. `http://www.google.de/intl/de/press/metrics.html`

große Datenmengen zu handhaben, um den späteren Nutzern gute Ergebnisse zu liefern. Ihre Absicht, viele Dokumente zu indexieren, brachte sie auch auf den Namen „Google“, der nach der Aussage von [BP98] eine im Englischen häufig vorkommende Schreibweise von „Googol“, also der Zahl 10^{100} sei.

Nach dem Willen seiner Entwickler soll Google sowohl möglichst viele Seiten in möglichst kurzer Zeit erfassen, als auch einen möglichst hohen Precision–Wert (die Zahl der gelieferten Ergebnisse, die zur Anfrage passen) liefern. Dafür nehmen sie laut [BP98] auch einen schwächeren Recall–Wert (Zahl der möglichen Dokumente die zur Anfrage passen) in Kauf.

Um eine hohe Precision zu erreichen nutzt Google sowohl den Pagerank–Algorithmus (s. 2.2.1.1) als auch die Verweisinformation aus dem Ankertext. Google verbindet den Ankertext nicht nur mit dem Dokument in dem der Verweis steht, sondern auch mit dem Dokument auf das der Verweis zeigt. Die Idee ist, dass der Text eines Verweises, der auf ein Dokument zeigt, dieses Dokument beschreibt (vgl. 2.1.4 und 2.1.4.1). Das erlaubt es, dass sogar Dokumente als Suchergebnis zurückgeliefert werden, die eigentlich noch gar nicht vollständig analysiert wurden ([BP98]), für die aber schon aufgrund der Ankertexte aus anderen Dokumenten Beschreibungen existieren.

Zusätzlich zu Pagerank und den Ankertexten nutzt Google weitere Eigenschaften der Webdokumente:

- die Fundortinformation, so dass das lokale Umfeld in die Dokumentbewertung miteinbezogen werden kann (siehe auch 2.1.4).
- die visuelle Präsentation von Wörtern. Wörter in Überschriften oder auch fett gesetzte Wörter zum Beispiel, werden von Google stärker gewichtet als andere.
- Der volle Rohtext der HTML–Seiten wird im Datenspeicher von Google gehalten.

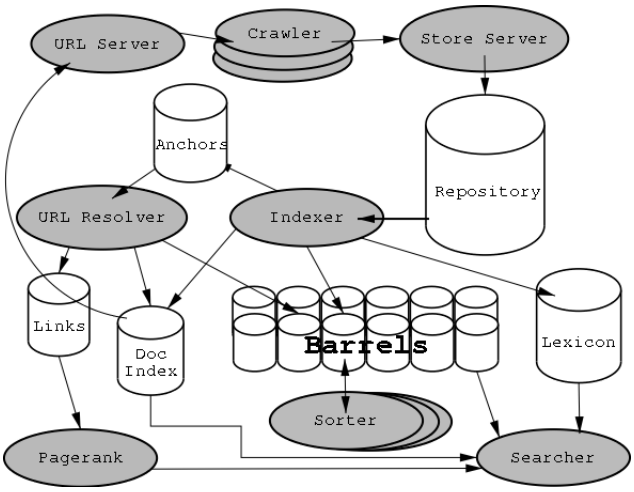


Abbildung 2.12: Architektur von Google (Quelle: [BP98])

Ähnlich wie Mercator (s. 2.3.2.1) wird das „Web Crawling“ von mehreren Web Crawlern gleichzeitig durchgeführt. Es gibt einen sog. „URL Server“ (s. Abbildung 2.12), der zu holende Dokument–URLs an die Crawler verteilt. Diese laden die Dokumente aus dem Internet herunter und speichern sie im zentralen Repository.

Die Daten im Repository werden dann vom „Indexer“ weiterverarbeitet. Der Indexer erstellt aus den Dokumenten Worttupelmengen. Die Tupel beinhalten das Wort selbst, die Position dieses Worts im Text, die Schriftgröße relativ zur Standardschriftgröße im Dokument und die Groß-/Kleinschreibung des Worts im vorliegenden Dokument. Die Worttupelmengen werden dann in die sog. „Barrels“ einsortiert. So wird ein Vorwärtsindex erzeugt. Zusätzlich extrahiert der Indexer die Verweise aus den Dokumenten und speichert Start- und Zielpunkt des Verweises zusammen mit dem Ankertext in einer „Anchors“–Datei zur späteren Auswertung.

Die Verweise im „Anchor“ werden ggf. in absolute URLs umgewandelt und als Start–Zielpunkt–Paare in die Verweisdatenbank eingetragen. Zusätzlich werden die Ankertexte der Verweise ebenfalls in Worttupel umgewandelt, die in Beziehung gesetzt werden, mit dem Dokument auf das der Verweis zeigt und dann in die „Barrels“ einsortiert.

Die Verweisdatenbank „Links“ wird genutzt um die Pageranks aller bekannten Dokumente zu berechnen.

Der Sortierer („Sorter“) erzeugt aus den „Barrels“, die nach Dokumenten sortiert sind, schließlich einen inversen Index, der nach Wörtern sortiert ist. Dieser inverse Index gibt an, in welchen Dokumenten ein bestimmtes Wort zu finden ist. Für die Suchmaschinenabfrage (“Searcher“) wird der inverse Index zusammen mit dem Pagerank genutzt, um Dokumente aufzufinden, die zur Anfrage passen.

Die Anfragesprache von Google ist vergleichsweise mächtig. Laut [Goo02] unterstützt Google neben einer AND–Verknüpfung (Standardeinstellung der Suchwortverknüpfung) und einer OR-Verknüpfung auch Möglichkeiten zum Ausschluss von Wörtern. Zusätzlich kann die Suche auf bestimmte Domains oder Zeiträume eingeschränkt werden. Google kann aufgefordert werden, bestimmte Wörter ausschließlich im Dokumenttitel, in Verweis–URLs und in Dokument–URLs zu suchen. Außerdem ist Google in der Lage nach so genannten Backlinks zu suchen. Das bedeutet, dass Dokumente, die auf eine bestimmte URL verweisen als Ergebnis geliefert werden. Der Benutzer kann die Kandidatenmenge auf Dokumente einer bestimmten Sprache oder auf bestimmte Filetypen, die anhand ihrer Endung erkannt werden, einschränken. Die Sprachbestimmung eines Dokuments erfolgt laut [Goo02] durch eine Kombination der sog. Top–Level–Domain–Endung mit der geographischen Lage, die Google aus der IP–Adresse des Dokumentenservers ableitet.

Der Mächtigkeit der Anfrage steht die Beschränkung der Anfragenlänge auf zehn Suchwörter gegenüber. Der Benutzer muss also die „richtigen“ zehn Kriterien für eine erfolgreiche Suche nach dem Wunschdokument auswählen — auch wenn mehr Kriterien zur Verfügung stehen.

Google ist mittlerweile eine kommerzielle Suchmaschine, so dass der Source–Code der Crawler nicht frei verfügbar ist. Jedoch bietet Google eine für nicht–kommerzielle Zwecke kostenlose SOAP[9]–Schnittstelle an, die es erlaubt Google–Suchanfragen aus eigenen Programmen heraus zu starten. Zusätzlich existiert eine entsprechende JAVA–Klassenbibliothek, die die entsprechenden SOAP–Schnittstelle kapselt(s. [Goo03]).

[9] Simple Object Access Protocol [BEK+]

2.3.3 Meta–Suchmaschinen

Meta–Suchmaschinen versuchen die Ergebnisse verschiedener anderer Suchmaschinen zusammenzufassen. Dazu wird die vom Benutzer gestellte Anfrage an die der Meta–Suchmaschine bekannten Suchmaschinen, weitergeleitet.

Im Zusammenhang mit Meta–Suchmaschinen kann eine Ebenenordnung von Suchmaschinen eingeführt werden. Es werden Suchmaschinen 1. - 4. Ordnung unterschieden.

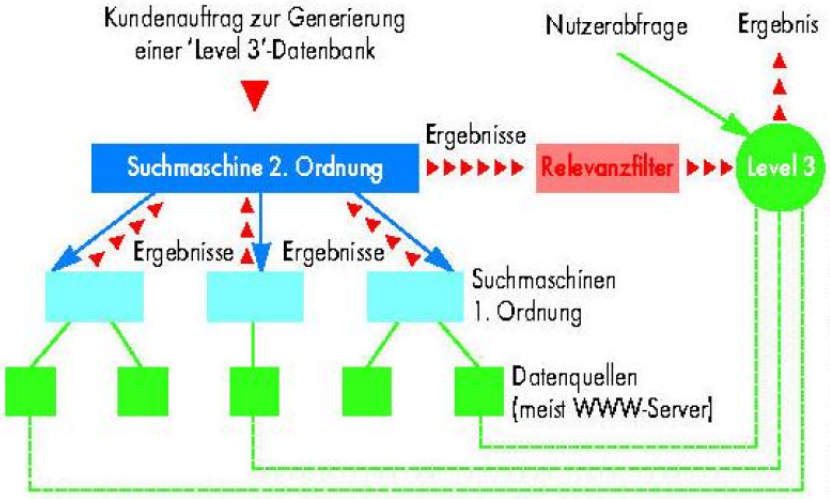


Abbildung 2.13: Suchmaschinen 1.– 3. Ordnung (Quelle: [SB98])

Suchmaschinen 1. Ordnung sind Suchmaschinen wie sie in 2.3.1 und 2.3.2 beschrieben werden.

Suchmaschinen 2. Ordnung fragen Suchmaschinen erster Ordnung ab und fassen deren Ergebnisse zusammen. Die gefundenen Dokumente werden jedoch nicht auf Anfragenrelevanz oder Verfügbarkeit geprüft.

Suchmaschinen 3. Ordnung nutzen die Ergebnisse von Suchmaschinen 2. Ordnung und laden die gefundenen Dokumente herunter. Diese Fundstellen werden mit dem Relevanzfilter der Suchmaschine untersucht. Der Relevanzfilter grenzt die Funddokumente auf ein bestimmtes Themengebiet ein, so dass davon gesprochen werden kann, dass eine neue Suchmaschine für ein bestimmtes Thema entstanden ist.

Suchmaschinen 4. Ordnung nehmen eine Suchanfrage entgegen und fragen aufgrund der Suchanfrage die (spezialisierten) Suchmaschinen ab, die voraussichtlich die besten Ergebnisse liefern. Suchmaschinen 4. Ordnung sind im Moment noch das Ziel aktiver Forschung (vgl. [MIN01]).

Suchmaschinen der 2., 3. und 4. Ordnung werden Meta–Suchmaschinen genannt.

In [SBS98] wird vorgeschlagen, dass als Meta–Suchmaschine nur solche Suchmaschinen betrachtet werden sollen, die die folgenden sieben Kriterien erfüllen:

Parallele Suche Die Meta–Suchmaschine soll ihre Anfragen parallel an die von ihr berücksichtigten Suchmaschinen stellen.

Ergebniszusammenfassung Die Ergebnisse sollen zusammengeführt und einheitlich dargestellt, statt einfach untereinander aufgelistet werden.

Doubletten–Eliminierung Fundstellen die von mehreren Suchmaschinen gefunden wurden, sollen erkannt und gekennzeichnet bzw. eliminiert werden.

Vollständige Operatoren Es sollen mindestens die Operatoren AND und OR von der Meta–Suchmaschine unterstützt werden.

Vermeidung von Informationsverlusten beim Transfer der Ergebnisse von den Suchmaschinen zum Benutzer. Die Informationen einer Suchmaschine zu einer Fundstelle (z. B. der Zusammenfassungstext) sollen auch in der Meta–Suchmaschine angezeigt werden, so dass die Meta–Suchmaschine nicht weniger Informationen liefert als die Einzel-Suchmaschinen.

Transparenz Der Benutzer der Meta–Suchmaschine sollte sich nicht mit den Eigenheiten der unterliegenden Suchmaschinen beschäftigen müssen.

Vollständige Suche Solange eine der unterliegenden Suchmaschinen noch Ergebnisse liefern kann, soll auch die Meta–Suchmaschine Ergebnisse liefern.

Dieser Forderungskatalog stellt hohe Anforderungen an Meta–Suchmaschinen. Er zeigt auch ihre Schwäche: Sie sollen die Suche vereinheitlichen. Das bedeutet aber auch, dass die Einflussmöglichkeiten des Nutzers auf die Suche beschränkt sind. Statt die üblicherweise mächtigen, jedoch nicht standardisierten Anfragesprachen der Suchmaschinen erster Ordnung nutzen zu können, muss der Nutzer die oft vergleichsweise beschränkten Möglichkeiten der Anfragesprache der Meta–Suchmaschine nutzen. Meta–Suchmaschinen bieten häufig nur die Schnittmenge der Anfragesprachen-Elemente der von ihr genutzten Suchmaschinen an. Alternativ gibt die Meta–Suchmaschine die vom Benutzer eingegebene Anfrage einfach an alle ihr bekannten Suchmaschinen weiter und überlässt es ihnen, ob sie etwas mit der Syntax anfangen können. Das ist insofern problematisch, als dass einige Suchmaschinen den AND-Operator als Default-Operator implementieren und andere die Anfrage als Phrase auffassen oder gar keine Ergebnisse liefern, weil sie die übergebene Syntax nicht verstehen.

3 Realisierung

Die Aufgabe, „verschollene“ Dokumente wiederzufinden, löst keins der in Kapitel 2 vorgestellten Suchverfahren. Trotzdem sind die dort angesprochenen Probleme mit der gestellten Aufgabe verwandt. Sie befassen sie sich mit einer themenspezifischen Suche, deren Ziel es ist *möglichst viele* Dokumente zu einem bestimmten Thema zu finden. Die Aufgabe *ein bestimmtes* Dokument wiederzufinden wird von ihnen nicht behandelt.

Es muss nun entschieden werden, welche der vorgestellten Verfahren und Erkenntnisse genutzt oder so angepasst werden können, um die Aufgabe zu lösen.

Aus den in 2.1.3 vorgestellten Ergebnissen zur Verbundenheit des Internets folgt, dass es nicht ausreicht nur einen Startpunkt für die Suche nach einem Dokument zu betrachten, da nicht jedes beliebige Dokument von jedem anderen Dokument erreichbar ist. Das bedeutet, dass für die Suche nach einem Dokument im Allgemeinen mehrere Einstiegspunkte für eine Suche benötigt werden. Diese Einstiegspunkte können durch Suchmaschinen (s. 2.3) gewonnen werden (vgl. 2.2.1.2 u. 3.4) oder durch Analyse der URL, um in derselben Domain Dokumente als Ausgangspunkt der Suche zu finden (s. 3.4.2).

Für eine erschöpfende Breitensuche ist das World–Wide–Web zu groß. Jedoch sollen durch einfache tiefenbeschränkte Breitensuchen gute Ergebnisse erzielt werden können (vgl. 2.2.3 u. 3.4.2.1). Zur Einschränkung der zu besuchenden Verweise, können offenbar Anker-Texte und Verweis–URLs einbezogen werden (s. 2.1.4.1). Zur weiteren Verfeinerung der Auswahl können die Textteile um die Verweise herum — also der Kontext — hilfreich sein.

Wenn es gelingt, Strukturen des WWW im System als Graph abzubilden, dann kann die Verweis-Struktur Hilfen zur Auswahl der zu verfolgenden Verweise geben (s. 2.1.4.1 und 2.1.4).

Eine weitere Unterstützung der Suche kann die Strukturinformation sein, wenn so genannte Backlinks ausgenutzt werden können, um Dokumente zu finden, die auf das gesuchte Dokument verweisen. Werden diese Information vor dem “Verschwinden“ des Dokuments gesammelt, können die verweisenden Dokumente genutzt werden, um den neuen Ort des verschwundenen Dokuments zu bestimmen, falls ein anderer Dokument-Autor sein Dokument schon dahingehend gepflegt hat, dass seine Verweise wieder auf die richtigen Stellen verweisen.

Die in Abschnitt 2.2 vorgestellten Verfahren zur themenspezifischen Suche sind erfolgreich getestet worden. Ob sie so angepasst werden können, dass sie direkt auf die Themenstellung der Diplomarbeit (s. 1.2) passen, ist fraglich. Es gelingt vermutlich am einfachsten, wenn die wiederzufindenden Dokumente tatsächlich zu *einem* Themengebiet gehören. In diesem Fall sind die Verfahren, die Klassifizierer anlernen müssen, anwendbar. Sollten die Dokumente keinen Zusammenhang haben, so werden die Methoden in

Ermangelung von Trainingsdaten versagen. Das bedeutet jedoch, dass während der Suche stets die Unsicherheit vorhanden ist, ob

1. die zu überwachenden URLs wirklich zu einem oder nur wenigen Themengebiet(-en) gehörten,

2. die Erkennung der Themengebiete bzw. die Klassifikation der einzelnen URLs in die Themengebiete korrekt war und

3. die Menge der Traininigsdokumente genügend groß und unkorreliert war.

Hinzu kommt, dass die Trainingsmengen für die Klassifizierer bei den vorgestellten Verfahren im Allgemeinen vom Benutzer zunächst vorgegeben werden müssen.

Interaktive Agenten wie z. B. 2.2.1.3 lösen das „Cold-Start-Problem“ zwar anders, jedoch wird in diesem Fall während des Agentenlaufs eine Interaktion mit dem Benutzer notwendig, die im Falle von PageTracker nicht gewünscht ist (s. 1.2).

Die in Abschnitt 2.3.1 vorgestellten katalogbasierten Suchmaschinen sind zur Wiederfindung der Seiten nicht geeignet, da nur die Zusammenfassungstexte nach den Anfragetermen durchsucht werden. Diese Zusammenfassungstexte sind dabei nicht einmal immer Texte, die aus den verwiesenen Dokumenten extrahiert werden, sondern z. B. bei Yahoo Texte, die das verwiesene Dokument beschreiben. Dort müssen also nicht die Wörter vorkommen, die auch im Dokument vorkommen. Hinzu kommt der üblicherweise lange Wartungszyklus der Verweise in den Verzeichnissen, wenn sie von Menschen gepflegt werden.

Meta–Suchmaschinen wie sie in Abschnitt 2.3.3 vorgestellt werden, sind wegen ihrer oft beschränkten Anfragesprachen eher nicht geeignet eine spezielle Seite wiederzufinden. Meta–Suchmaschinen erlauben es nicht, die speziellen Möglichkeiten einer indexbasierten Suchmaschine zu nutzen, um beispielsweise das Indexierungsdatum oder die Suche eines speziellen Teils einer Seite in die Suchanfrage mit einzubeziehen.

Es erscheint also sinnvoller, indexbasierte Suchmaschinen direkt zu nutzen. Wenn Suchmaschinen genutzt werden, ist der Ranking-Mechanismus der Suchmaschine für den Erfolg ausschlaggebend. Ist der Verweis auf das richtige Dokument erst sehr spät in den Suchergebnissen zu finden, müssen möglicherweise ebenso viele Verweise und Dokumente untersucht werden wie bei einer Breitensuche.

Moderne Suchmaschinen wie Google kombinieren für ihr Ranking verschiedene Verfahren, die auch bei den beschriebenen verweisbasierten Verfahren (s. 2.2) zum Einsatz kommen. Google (s. 2.3.2.2) kombiniert z. B. Pagerank (s. 2.2.1.1) und gewichtet die Dokumente zusätzlich unter Berücksichtigung von Ankertexten von Verweisen, die auf das zu gewichtete Dokument verweisen. Zudem werden Begriffe, die in Überschriften oder Dokumententiteln stehen, ähnlich wie z. B. in 2.2.1.3 zur Ermittlung des Rangs bezüglich einer gegebenen Suchanfrage herangezogen.

Google hat laut `http://www.Google.de` im April 2003 ca. 3,08 Milliarden Dokumente erfasst. Zusammen mit der zur Verfügung gestellten Anfrageschnittstelle, bietet sich die Verwendung von Google als Ansatzpunkt zur Suche nach den verlorenen Dokumenten an, denn die Methoden, die von verweisbasierten Verfahren zur Anordnung der zu

verfolgenden Verweise genutzt werden, werden von Google auf dieser großen Dokumentenkollektion als Ranking-Strategie angewendet. Da einige der Verfahren besser funktionieren, wenn die Dokumentenkollektion größer ist, ist die Google zur Verfügung stehende große Kollektion ein Vorteil. Diese Dokumentenkollektion wird ständig durch die Spider in Google aktualisiert. Durch die Verwendung von BlockRank (s. S. 10) erfolgt die Erfassung neuer Dokumente noch schneller, so dass die Aktualität und die Bewertung verbessert wird.

Durch die große Dokumentenkollektion und den Einsatz viel versprechender, in anderen Suchverfahren erfolgreich eingesetzter Bewertungsverfahren, sind Anfragen an Google am erfolgversprechendsten. Sie sollen daher im Rahmen der Diplomarbeit weiter verfolgt werden.

Mit der Nutzung von Google ergeben sich jedoch neue Problemstellungen:

- Ist es möglich, ein zu suchendes Dokument so auf zehn Wörter zu reduzieren, dass es unter den ersten zehn Suchergebnissen von Google auftaucht?
- Welche Wörter aus dem Dokument sind die richtigen?
- Wie kann eine Suchanfrage so geändert werden, dass die nächste Anfrage erfolgreicher ist?

Natürlich kann es sein, dass das gesuchte Dokument noch nicht oder noch nicht wieder von Google indexiert worden ist. In diesem Fall scheint eine tiefenbeschränkte Breitensuche (s. 2.2.3) ausgehend von verschiedenen Ansatzpunkten aussichtsreich und schnell zu sein.

Es folgt eine Beschreibung der wichtigsten Komponenten aus denen sich der Agent PageTracker zusammensetzt. Im Anschluss werden in Abschnitt 3.2 der prinzipielle Arbeitsablauf des Agenten und die verwendete Datenstruktur vorgestellt. Da PageTracker nach Dokumenten sucht, die einem Originaldokument sehr ähnlich sind, werden in Abschnitt 3.3 zwei Maße erläutert, die grundsätzlich als Ähnlichkeitsmaße in Betracht kommen. Den Abschluss des Kapitels bildet die Beschreibung der Suchstrategeme, die von PageTracker verwendet werden, um die gesuchten Dokumente zu finden.

3.1 Die Grundlage

Die Grundlage der Implementierung von PageTracker ist die Klassenbibliothek BotIShelly ([BFG+00]).

Die Bibliothek setzt sich im Wesentlichen aus sieben Komponenten zusammen:

1. einer zentralen Steuerung, auch Systemkontrolle genannt,
2. einem Planer und der zugehörigen Planausführung,
3. den Operatoren, die dem Planer zur Verfügung stehen,
4. einer Wissenskomponente, die durch eine A–/T–Box–Kombination dargestellt wird,
5. einer Komponente für die Datenbeschaffung aus dem Internet,
6. einer Komponente für die Datenanalyse und
7. einer Komponente zur Interaktion mit dem Benutzer des Agenten.

BotIShelly erlaubt grundsätzlich die parallele Ausführung mehrerer Planer und Planausführer gleichzeitig. Ich habe mich jedoch entschieden, immer nur genau einen Planer oder genau einen Planausführer zu jedem Zeitpunkt laufen zu lassen, um Probleme mit der Synchronisierung von vorne herein auszuschließen. Das schränkt die Funktionalität jedoch nicht ein, da im Falle einer Parallelisierung der Abgleich der verschiedenen A–Boxen erst nach dem Abschluss der Suche erfolgte. So ist sichergestellt, dass nicht zwei Planausführer gleichzeitig dasselbe Dokument suchen.

BotIShelly bietet seinen Nutzern zwei verschiedene Planer an. Eine Variante des Graph-Plan-Algorithmus von [BF95] und eine einfachere Variante, den sog. „PlannerTypeA“. Der „PlannerTypeA“ führt die zur Verfügung stehenden Operatoren gemäß ihrer Priorität aus, solange die Suche nicht beendet wird.

PageTracker setzt den PlannerTypeA ein, weil die zur Verfügung gestellten Operatoren fast alle gegeneinander austauschbar sind und ihre Signaturen es nicht erlauben, längere Planketten zu bilden.

Aus Sicht z. B. des Graphplaners erfüllt jeder einzelne Strategem–Operator für sich alleine schon das Suchziel, da diese Operatoren fertig bearbeitete URLs in der Nachbedingung stehen haben. Die Pläne wären kurz, aber der Rechenaufwand erheblich höher als beim „PlannerTypeA“. Zudem wäre die Kandidatenauswahl schwieriger zu modellieren, da ein Graph–Planer den Auswahloperator (s. 3.4.3.1) in seiner jetzigen Form nach jedem einzelnen Operator in die Pläne einbauen würde. Der Auswahloperator hätte aber immer nur einen Kandidaten zur Auswahl.

Dem „PlannerTypeA“ kann durch eine entsprechende Priorisierung der Operatoren mitgeteilt werden, in welcher Reihenfolge die Operatoren zur Ausführung gelangen sollen, so dass der „Auswahl-“ und der „FertigOperator“ erst am Schluss ausgeführt werden. So hat der Auswahl–Operator auch tatsächlich eine Menge, aus der er eine Auswahl treffen kann.

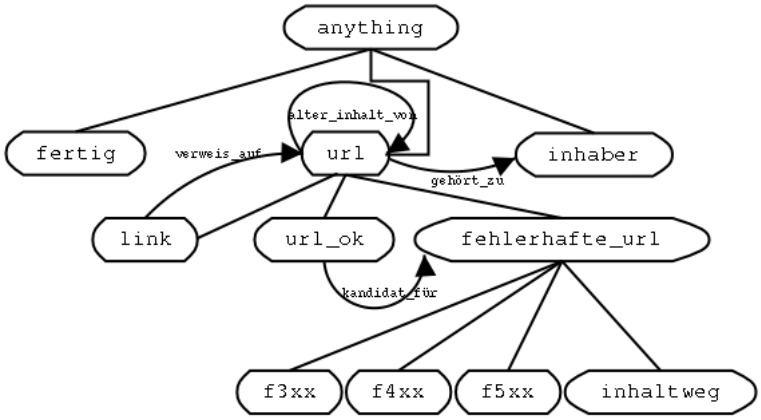
Bis auf den Auswahl–Operator und den „FertigOperator“ implementieren die Operatoren jeweils ein Suchstrategem. Diese werden in den Abschnitten 3.4.1 bis 3.4.2.1 näher beschrieben.

Die Wissensrepräsentationskomponente bestimmt die zur Verfügung stehenden Prädikate für die Operatoren. Diese Komponente wurde ohne Änderungen aus BotIShelly übernommen. Die für PageTracker benötigten Daten werden in einer KL–ONE-ähnlichen A–/T–Box–Struktur (s. [BS85]) abgelegt. Die Modellierung der T–Box stellt daher einen wesentlichen Bestandteil der Agentenerstellung dar. Die von PageTracker verwendete T–Box-Struktur ist in Abbildung 3.1 dargestellt.

Für PageTracker sind die folgenden Konzepte am wichtigsten:

url Dieses Konzept ist das Oberkonzept für die drei folgenden Konzepte. Die Instanzen zu diesem Konzept repräsentieren die Menge der zu überwachenden Dokumente. Das eigentlich Dokument „hängt“ an diesen Instanzen (s. [BFG+00] Kap. 5.3).

url_ok Dieses Konzept repräsentiert alle Dokumente, die keine Fehler aufweisen, aber potenziell fehleranfällig sind und daher von Zeit zu Zeit von einem URL–Checker überprüft werden müssen.



Linien mit Pfeil repräsentieren Rollen
Linien ohne Pfeil repräsentieren von unten nach oben gelesen `isA`-Beziehungen.

Abbildung 3.1: T–Box-Struktur von PageTracker

fehlerhafte_url Dieses Konzept repräsentiert alle Dokumente, die entweder nicht mehr den erwarteten Inhalt aufweisen oder die aus sonstigen Gründen nicht erreichbar sind. Die Instanzen dieses Konzepts enthalten den neuen Inhalt (bzw. die Fehlermeldung) der gegebenen URL. Der alte, erwartete Inhalt ist über die entsprechende Rolle `alter_inhalt_von` (s. 3.1) erreichbar.

link repräsentiert diejenigen URLs, die zwar nicht überprüft werden müssen, die aber in der Vergangenheit einen Verweis auf zu überprüfende Dokumente enthielten. Dieses Konzept ist im Zusammenhang mit der in 3.4.2.1 beschriebenen Suchstrategie wichtig.

Die Konzepte `fertig` und `inhaber` werden verwendet, um das Ende des Arbeitsablaufs zu markieren, bzw. Zugehörigkeiten von URLs zu modellieren, um die Testläufe besser interpretierbar zu machen. Für die eigentliche Arbeit von PageTracker sind sie von geringer Bedeutung.

Die vom Konzept `fehlerhafte_url` subsumierten Konzepte `f3xx, f4xx, f5xx` und `inhaltweg` können von einem URLChecker (z. B. [Mas03]) genutzt werden, um speziellere Hinweise zu der Art des Fehlers zu geben, die eventuell von einem Suchstrategem genutzt werden könnten.

Zusätzlich zu den beschriebenen Konzepten existieren eine Reihe von Rollen, die Beziehungen zwischen den Konzepten herstellen. Sie dienen PageTracker dazu, Informationen über die zu überwachenden URLs mit diesen URLs zu verbinden. PageTracker nutzt die folgenden Rollen:

alter_inhalt_von verbindet zwei `url`s miteinander. Die erste Komponente repräsentiert den Inhalt, der früher unter der URL, die durch die zweite Komponente repräsentiert wird, zu finden war. Wenn die zweite Komponente Instanz zum Konzept `fehlerhafte_url` ist, wird der Agent nach dem Inhalt des Dokuments, das der Instanz der ersten Komponenten zugeordnet ist, suchen.

Diese Beziehung ist die wichtigste für PageTracker. PageTracker weiß auf Grund dieser Beziehung wonach er suchen soll.

gehoert_zu ist eine Rolle die eine `url` einem `inhaber` zuordnet. Momentan wird diese Rolle nur zur besseren Identifizierung während der Testläufe genutzt.

kandidat_fuer verbindet eine fehlerfreie URL mit einer Instanz vom Typ `fehlerhafte_url`. Diese Rolle wird für den Kandidatenauswahlprozess am Ende eines Suchlaufs benötigt. Die Strategem–Operatoren (s. 3.4.1 ff.) erzeugen Rolleninstanzen, aus denen am Ende der beste Kandidat ausgewählt wird (s. 3.4.3.1).

verweis_auf verbindet `link`s mit `url`s. Ein `link`, der mit einer `url` in der `verweis_auf`–Beziehung steht, repräsentiert ein Dokument in dem ein Verweis auf die zu überwachende URL vorhanden ist. Diese Beziehung ist vor allem für das in Abschnitt 3.4.2.1 beschriebene Suchstrategem wichtig.

Die Komponente zur Datenbeschaffung hat eine wesentliche Veränderung erfahren: Die geladenen Daten wurden in der in BotIShelly implementierten Version als Objekt–Daten zusammen mit der A–Box serialisiert. Die Daten werden jetzt in einer Datenbank abgelegt. Unterstützt werden MySQL-Datenbankmanagementsysteme und Oracle 8i. Diese Auslagerung der heruntergeladenen Daten entlastetet den Hauptspeicher erheblich und lässt größere Datenmengen zu, da nun nicht mehr alle heruntergeladenen Dokumentinformationen gleichzeitig im Hauptspeicher gehalten werden müssen.

Aus der Datenanalysekomponente, die hauptsächlich zur Textklassifikation dient, nutzt PageTracker die zur Verfügung stehenden Wörterbuch–Klassen, die Klassen zur Vektorisierung von Zeichenketten, sowie die auf diesen Textvektoren zur Verfügung gestellten Berechnungen von Vektorabständen und Winkeln zwischen zwei Vektoren (s. [BFG+00]).

Die Komponente zur Interaktion mit dem Benutzer wurde ganz herausgenommen, da PageTracker als Server Side Application nicht mit Endbenutzern interagieren, sondern seine Arbeit still verrichten soll (s. 1.2).

3.2 PageTrackers Arbeitsablauf

Der schematische Arbeitsablauf ist in Abbildung 3.2 zu sehen. Der Agent wird mit einer URL–Menge initialisiert. Die URLs sind als fehlerhaft oder gültig gekennzeichnet. Die Kennzeichnung erfolgt in den Test-Szenarien manuell und soll später durch einen URL-Checker ([Mas03]) geschehen. PageTracker wendet aufgrund des verwendeten Planers (s. 3.1) die zur Verfügung stehenden Operatoren gemäß ihrer Priorisierung auf diese URL-Menge an. Dabei wird für jeden Operator geprüft, ob die Vorbedingung zur Durchführung des Operators erfüllt werden kann. Die Vorbedingung prüft i.d.R., ob eine fehlerhafte URL existiert und ob die ggf. benötigten zusätzlichen Informationen über das gesuchte Dokument vorhanden sind. Die zusätzlichen Informationen variieren je nach Suchstrategem. Falls welche vorhanden sind, wird der Operator mit jeder vorhandenen fehlerhaften URL, zu der auch die notwendige Zusatzinformation vorhanden ist, ausgeführt. Falls keine fehlerhafte URL vorhanden ist oder die Zusatzinformationen fehlen, wird der nächste Operator ausprobiert.

Es existieren zwei Operatoren, die von diesem Schema abweichen. Der erste ist der SelectCandidateOperator (s. 3.4.3.1). Dieser Operator wählt aus den Ergebnissen der vorangegangenen Operatoren das beste aus. Operatoren erzeugen immer dann Kandidaten, wenn sie kein Dokument finden, welches einen so geringen Abstand zum Original hat, dass fast ausgeschlossen ist, dass ein anderer Operator ein Dokument mit noch geringerem Abstand findet. Damit nachfolgende Operatoren die Gelegenheit erhalten, dieses besser passende Dokument zu finden, wird ein Kandidat für die fehlerhafte_url erzeugt, statt sie sofort durch das beste von diesem Operator gefundene Dokument zu ersetzen.

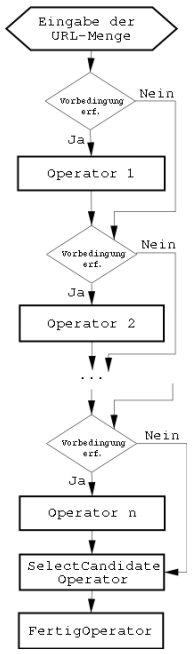


Abbildung 3.2: Arbeitsskizze von PageTracker

Ein Dokument, welches Kandidat für die Ersetzung der fehlerhaften_url ist, hat auf jeden Fall einen Abstand vom Originaldokument, der geringer ist als der maximal akzeptierte Abstand und größer ist als Null.

Der zweite ist der FertigOperator (s. 3.4.3.2). Dieser Operator setzt lediglich voraus, dass überhaupt eine URL existiert. Es ist irrelevant, ob sie fehlerhaft ist oder nicht. Der Operator überprüft intern, ob überhaupt noch fehlerhafte URLs existieren. Falls keine fehlerhaften URLs existieren, meldet PageTracker, dass er erfolgreich war. Sonst gibt er die URLs an, deren Dokumente er bisher nicht wiederfinden konnte.

Für den Fall, dass noch fehlerhafte URLs existieren, kann eine alternative Implementierung den sofortigen Neustart des Arbeitsablaufs vorsehen. Das ist jedoch nur sinnvoll, wenn entweder

(a) die Operatoren zusätzliche Informationen erzeugen, die ein anderer Operator nutzen kann, so dass dieser andere Operator mit der fehlerhaften URL verwendet werden kann, oder

(b) abzusehen ist, dass die Misserfolge des letzten Laufs durch temporäre Störungen bei verwendeten Suchmaschinen oder Web-Servern entstanden sind, die üblicherweise beim nächsten Aufruf behoben sind.

Da die implementierten Operatoren keine Informationen füreinander erzeugen, wird Fall (a) nicht eintreten. Da Fall (b) in den Testläufen nicht auftrat — auftretende Netzfehler waren von längerer Dauer — habe ich entschieden, dass die erneute Abarbeitung des Arbeitsablaufs erst mit einem Neustart des Agenten erfolgt. Dieser Neustart kann skriptgesteuert in nahezu beliebigen Zeitabständen erfolgen, so dass im Zweifel der Fall (b) auch abgedeckt wird.

3.3 Ähnlichkeitsmaße

Zur Bewertung der gefundenen Dokumente wird von PageTracker ein Ähnlichkeitsmaß genutzt.

Die Erstellung desjenigen Ähnlichkeitsmaßes, welches das Benutzerinteresse widerspiegelt, ist für sich genommen schon ein schwieriges Problem. Es wird in [Mas03] näher untersucht und behandelt.

Zur abschließenden Bewertung der Suchergebnisse wird daher in dieser Arbeit ein „perfektes Distanzmaß“ vorausgesetzt. Ein „perfektes Distanzmaß“ erstellt dasselbe Ranking zwischen den Ergebnissen, welches auch der Nutzer des Agenten erstellen würde. Dokumente, die vom Nutzer nicht akzeptiert werden, werden auch nicht von einem „perfekten Distanzmaß“ akzeptiert.

Da das beschriebene Maß nicht verfügbar ist, erfolgt die Abschlussbegutachtung im Rahmen dieser Diplomarbeit manuell.

Um dem Agenten trotzdem einen Anhaltspunkt zu geben, welche Dokumente als Ersatzdokumente für das gesuchte Dokument in Frage kommen, wurden zwei Maße getestet:

- die Levenstheindistanz (Editierdistanz)
- das Cosinusmaß

Diese Maße werden vom Agenten innerhalb der Operatoren zur Steuerung der Suche genutzt.

Da die Maße nicht „perfekt“ sind, können ihre Bewertungen Fehler generieren. So können unähnliche Dokumente als guter Ersatz bewertet werden oder sehr ähnliche Dokumente als Nicht-Kandidat aus der Kandidatenmenge herausfallen. Trotzdem ist die Verwendung eines solchen Hilfsmaßes sinnvoll, damit die notwendige Benutzerinteraktion gering bleibt und der Agent in der Lage ist zu erkennen, ob er die Suche nach einem bestimmten Dokument weiter intensivieren muss oder schon abbrechen kann, weil eins der gefundenen Dokumente dem Originaldokument (exakt) entspricht.

Die beiden getesteten Verfahren werden in den Abschnitten 3.3.1 und 3.3.2 genauer beschrieben. In den Testläufen hat sich das Cosinusmaß besser bewährt, da es die Dokumente, eine Suche nach einem möglichst inhaltsähnlichen Dokument vorausgesetzt, besser dem Nutzerwunsch gemäß bewertete.

3.3.1 Editierdistanz

Editierdistanzmaße basieren auf der Anzahl der Editieroperationen, die durchzuführen sind, um eine Symbolkette A in eine Symbolkette B umzuwandeln. Üblicherweise werden Wörter oder Buchstaben als Symbole betrachtet.

Die sog. Levensthein–Distanz [Lev65] gibt die minimalen Kosten an, die entstehen, wenn eine Symbolkette A in die Symbolkette B durch Hinzufüge-, Lösch- und Ersetzungsoperationen überführt wird. Wenn das Problem durch dynamische Programmierung gelöst wird, hat es eine Komplexität von $O(\|A\| \cdot \|B\|)$ (s. z. B. [Rah02]).

Sind die Kosten für jede der drei Änderungsoperationen gleich eins, gibt die Levensthein–Distanz die Anzahl der notwendigen Änderungsoperationen für die Überführung an. Die Kosten für die einzelnen Operationen dürfen jedoch auch verschieden sein, so dass andere Interpretationen der Distanz möglich sind.

3.3.2 Cosinusmaß

Eine Methode Dokumente zu repräsentieren ist, sie durch Wortvektoren darzustellen. Für jedes Wort, das in einem Dokument vorkommen kann, bzw. welches durch den Wortvektor erfasst werden soll, hat der Wortvektor eine Komponente. Als Einträge in die Komponente sind verschiedene Varianten denkbar. Die drei häufig benutzten Varianten sind:

- Boolsche Eintragungen, die nur Aussagen darüber treffen, ob ein Wort überhaupt in einem Dokument vorkam oder nicht,
- die Häufigkeit des Wortvorkommens, in dem Dokument zu dem der Wortvektor gehört, Term Frequency (TF) genannt, und
- die Häufigkeit des Wortvorkommens im Dokument bezogen auf die Häufigkeit des Wortvorkommens in allen Dokumenten.
  Die so genannte „Term Frequency Inverse Document Frequency“ (TF–IDF) wird folgendermaßen berechnet: Sei D die Gesamtmenge der Dokumente, $TF(t,d)$ die Häufigkeit des Vorkommens des Terms t im Dokument $d \in D$, und $DF(t)$ die Anzahl der Dokumente aus D die den Term t enthalten überhaupt enthalten, dann ist
  $$\text{TF–IDF(t)} := TF(t,d) \cdot \log_2 \left(\frac{\|D\|}{DF(t)} \right)$$

Das TF–IDF–Maß ist laut [Sal89] die beste Annäherung dafür, wie charakteristisch ein Wort für einen Text ist. Das TF–Maß ist eine schwächere Annäherung. Die boolsche Auswahl lässt eine solche Unterscheidung gar nicht zu.

Auf den so gestalteten Wortvektoren kann das Cosinusmaß ([SM83]) angewendet werden. Das Cosinusmaß berechnet einfach den Cosinus des Winkels zwischen zwei Wortvektoren $\vec{w_1}$ und $\vec{w_2}$.

$$cos(\phi) = \frac{\vec{w_1} \cdot \vec{w_2}}{\|\vec{w_1}\| \cdot \|\vec{w_2}\|}, \phi \in [0, \pi[$$

Größere Werte bedeuten größere Ähnlichkeit zwischen den verglichenen Dokumenten.

3.4 Suchstrategeme

Dieser Abschnitt befasst sich mit den einzelnen Komponenten der Suchstrategie, die PageTracker verwendet, um die „verlorenen“ Dokumentinhalte wiederzufinden. Die Strategie besteht darin, erfolgversprechenden Strategeme in der Reihenfolge ihrer Priorität aneinanderzureihen. Da jedes Strategem von einem Operator (s. [BFG+00] S. 165ff.) repräsentiert und implementiert wird, erstellt der Planer (s. [BFG+00] S. 180ff.) eine Strategie und lässt sie vom Planausführer (s. [BFG+00] S. 184ff.) realisieren.

Die Strategeme lassen sich grob in drei Klassen aufteilen:

1. Suchmaschinenbasierte Strategeme (s. 3.4.1),
2. Verweisbasierte Strategeme (s. 3.4.2) und
3. Strukturbasierte Strategeme (s. 3.4.2.1).

Innerhalb einiger Strategeme werden implizit Annahmen über die Art des Ähnlichkeitsmaßes getroffen. So werden die Strategeme 3.4.1.1, 3.4.1.3 und 3.4.1.4 gute Ergebnisse erzielen, wenn inhaltsähnliche Dokumente wiedergefunden werden sollen, während sie bei strukturähnlichen Suchzielen vermutlich schlechter abschneiden werden. Die verweisbasierten Suchstrategeme (s. 3.4.2) machen keine impliziten Annahmen über das verwendete Ähnlichkeitsmaß.

Abgeschlossen wird der Abschnitt durch eine Erläuterung des SelectCandidateOperators und des FertigOperators, die eine Sonderrolle unter den verwendeten Operatoren einnehmen.

3.4.1 Suchmaschinenbasierte Strategeme

Die verwendeten Suchstrategeme basieren letztendlich auf Textvergleich. Das bedeutet, dass Seiten, die im Wesentlichen aus Bildern bestehen, nur schwer wieder zu finden sind. Wenn ein Dokument zu großen Teilen aus Grafiken besteht, können keine aussagekräftigen Text–Anfragen generiert werden. Ein Beispiel für ein problematisches Dokument ist die Hauptseite des Deutschen Forschungsnetzes in der Version[1] vom November 2001. Das Dokument besteht im Wesentlichen aus Grafiken und aus nur fünf unterschiedlichen Wörtern. Die Wörter sind so allgemein, dass eine Suchanfrage zu viele Ergebnisse liefert.

Wenn eine Seite überwiegend aus Text besteht, dann muss das Problem der richtigen Auswahl der Wörter dieser Seite gelöst werden, da

1. die verwendete Suchmaschine, gemessen an der Anzahl Wörter in einem durchschnittlichen Text, nur eine geringe Anzahl von Suchworten zulässt und
2. das Dokument vermutlich eine Veränderung erfahren hat, so dass nicht mehr alle Wörter, die im Ursprungsdokument vorkamen auch in der neuen Version des Dokuments zu finden sind.

Eine weitere Problematik wirft die Filterfunktion von Google auf. Google ist in der Lage Verweise auf ähnliche Seiten zu unterdrücken. Das verringert die Suchergebnisanzahl und soll dazu führen, dass der Benutzer nicht mehrmals dieselbe Seite unter anderen URLs präsentiert bekommt. Jedoch führt das von Google verwendete Ähnlichkeitsmaß gelegentlich dazu, dass die Seite, die der gesuchten am ähnlichsten gewesen wäre, herausgefiltert wird. In Tabelle 3.1 ist ein solches Beispiel zu sehen. Bei eingeschalteter Filterfunktion wird nur die englische Version des zu suchenden Dokuments gefunden. Ist die Filterfunktion deaktiviert, wird auch das besser passende deutschsprachige Dokument gefunden. Das Abschalten der Filterfunktion verschärft jedoch auch die Anforderungen an die Suchbegriffe. Wenn die „doppelten“ Ergebnisse nicht mehr herausgefiltert werden,

[1] `http://web.archive.org/web/20011125211613/http://www.dfn.de/index.html`

Original	`http://web.archive.org/web/20011024125720/http://www-ai.cs.uni-dortmund.de/FORSCHUNG/`
Anfrage	Katharina Machine Knowledge Vol Klingspor Etal Acquisition Journal Kaiser Models
	Fund
	`http://www-ai.cs.uni-dortmund.de/FORSCHUNG/index.eng.html`

(a) mit Filter

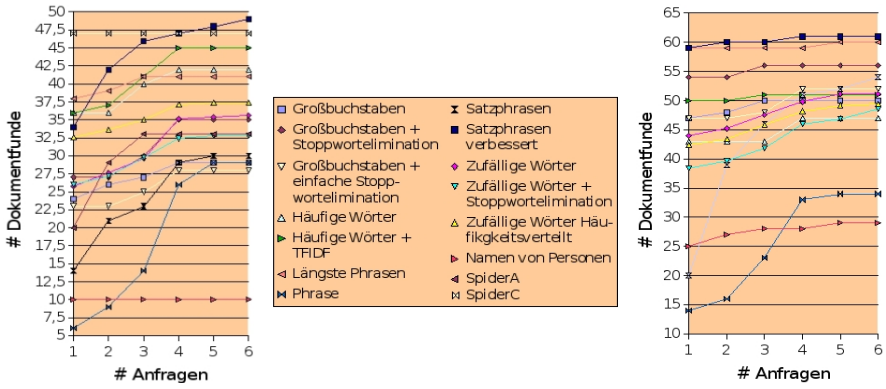
Original	`http://web.archive.org/web/20011024125720/http://www-ai.cs.uni-dortmund.de/FORSCHUNG/`
Anfrage	Katharina Machine Knowledge Vol Klingspor Etal Acquisition Journal Kaiser Models
	Fund
	`http://www-ai.cs.uni-dortmund.de/FORSCHUNG/index.eng.html`
	`http://www-ai.cs.uni-dortmund.de/FORSCHUNG/`
	`http://www-ai.cs.uni-dortmund.de/PERSONAL/morik.eng.html`
	`http://www-ai.cs.uni-dortmund.de/PERSONAL/morik.html`

(b) ohne Filter

Tabelle 3.1: Ergebnisse einer Suchanfrage *mit* und *ohne* Google–Filter

sind potenziell weniger andersartige Seiten unter den ersten zehn Ergebnissen, die von PageTracker überhaupt betrachtet werden.

Ein weiteres Problemfeld ist der Umgang mit Anfragen, die entweder viele oder wenige Ergebnisse liefern, ohne dass das richtige Dokument gefunden wird. Wenn zu viele Ergebnisse zurückgeliefert werden, war die Anfrage offenbar nicht speziell genug. Wenn zu wenige Ergebnisse zurückgegeben werden, waren vermutlich Suchbegriffe in der Anfrage, die im neuen Dokument nicht mehr vorkommen. Es kann natürlich auch sein, dass Google das gesuchte Dokument nicht kennt.



(a) Universität (b) WDR

Abbildung 3.3: Abhängigkeit der Dokumentenfunde von der Anfragenanzahl

Es gibt viele Möglichkeiten die Suchanfragen zu redefinieren. Die Operatoren unternehmen fünf Versuche eine Anfrage so zu redefinieren, dass sich das gesuchte Dokument unter den ersten zehn Ergebnissen befindet. Dazu können z. B. Suchbegriffe aus der Suchanfrage entfernt werden, wenn Google nicht genügend Ergebnisse liefert. Wenn zu viele Ergebnisse zurückgeliefert werden, können bis zu fünf Suchbegriffe ausgetauscht werden oder auch fünf andere Phrasen ausprobiert werden.

Der Operator steuert das Anfrageverhalten auf Basis der Suchergebnisse der letzten gestellten Anfrage. Es kann also vorkommen, dass die Anfrage zunächst verfeinert und dann wieder verallgemeinert wird (s. Tab. 3.2), um das gewünschte Ziel zu erreichen. Insgesamt werden jedoch nicht mehr als fünf Veränderungen an der ursprünglichen Anfrage vorgenommen. Es wäre auch möglich mehr als 5 Redefinitionen zuzulassen, jedoch hat sich in Vorversuchen herausgestellt, dass die verwendeten Strategeme (s. Kap. 3.4) nach 5 Redefinitionen keine nennenswerten Ergebnisverbesserungen erzielten. Abbildung 3.3 zeigt im Vorgriff auf Ergebnisse in Kapitel 4.4, dass die meisten Strategeme, keine zusätzlichen Dokumente mehr finden, wenn sie mehr als vier Anfragen ($\hat{=}$ 3 Redefinitionen) stellen dürfen.

Original	http://www-ai.cs.uni-dortmund.de/LEHRE/DIPLOM/diplom_fertig.html	
Anfrage	[2003] [2002] [2001] [2000] [1999] [1998] [1997] [1996] [1995] [1994]	
	Fund	**Bewertung**
	http://www.aegis.com/archive/papers.asp	89
	http://www.aegis.com/archive/other.asp	89
	http://www.walnet.org/csis/news/regions.html	89
	http://www.wavelengthmagazine.com/copyright.php	89
	http://www.census.gov/mrts/www/mrts.html	88
	http://www.eastman.com/News_Center/News_Archive/CorpNews_Archive.asp	89
	http://www.langmaker.com/db/mdl_index_year.htm	89
	http://www.catawbacountync.gov/depts/tax/taxfact.pdf	359
	http://www.irsemploymentreview.com/browse.asp	88
	http://www.ca5.uscourts.gov/oparchdt.cfm	89
Anfrage	[2003] [2002] [2001] [2000] [1999] [1998] [1997] [1996] [1995]	
	Fund	**Bewertung**
	http://w3.access.gpo.gov/eop/	89
	http://www.aegis.com/archive/papers.asp	89
	http://www.aegis.com/archive/other.asp	89
	http://www.aegis.com/archive/wire.asp	89
	http://www.census.gov/mrts/www/mrts.html	88
	http://www.wavelengthmagazine.com/copyright.php	89
	http://www.dcb.unibe.ch/groups/reymond/publications.html	88
	http://www.catawbacountync.gov/depts/tax/taxfact.pdf	359
	http://www.langmaker.com/db/mdl_index_year.htm	89
	http://www.eastman.com/News_Center/News_Archive/CorpNews_Archive.asp	89
Anfrage	[2003] [2002] [2001] [2000] [1999] [1998] [1997] [1996] [1995] inurl:diplom_fertig.html	
	Fund	**Bewertung**
	http://www-ai.cs.uni-dortmund.de/LEHRE/DIPLOM/diplom_fertig.html	71
	http://www-ai.informatik.uni-dortmund.de/LEHRE/DIPLOM/diplom_fertig.html	71

Tabelle 3.2: Beispiel für das wechselnde Anfrageverhalten

Das Standardverfahren zur Redefinition, generalisiert Suchanfragen, indem es Suchbegriffe aus der ursprünglichen Suchanfrage streicht. Gestrichen wird von hinten nach vorne, da die Suchbegriffe in der Suchanfrage gemäß ihrer Güte bzgl. des Auswahlkriteriums angeordnet sind. Diejenige Wörter, die dem Auswahlkriterium am besten genügen, stehen vorne in der Suchanfrage und sollen möglichst lange erhalten bleiben. Das Verfahren streicht nacheinander erst einen, dann zwei usw. bis zu fünf Suchbegriffen aus der vorherigen Anfrage. Dieses progressive Vorgehen ist dadurch motiviert, dass zuvor evtl. vorgenommene Spezialisierungszusätze, mitgestrichen werden sollen. Sollten zuvor keine Spezialisierungsschritte unternommen worden sein, so wird mit jedem Schritt aggressiver generalisiert. Die Gefahr einer Übergeneralisierung wird dabei in Kauf genommen.

Zur Spezialisierung wird der Suchanfrage der letzte Teil der Original-URL angehängt. Das ist i.d.R. der Dateiname des Dokuments, das gesucht wird. Wenn schon zehn Suchbegriffe in der Suchanfrage stehen, wird der letzte Suchbegriff durch die URL ersetzt.

3.4.1.1 Großbuchstaben

Abkürzungen und Eigennamen werden meistens groß geschrieben. Diese groß geschriebenen Wörter charakterisieren jedoch oft einen Text auch über mehrere Veränderungsschritte hinweg. Abkürzungen werden seltener falsch geschrieben und müssen daher nicht korrigiert werden und sie gehören meistens zu einer Fachsprache, die der Autor auch in einer neuen Version eines Textes nicht verlassen möchte.

Während Abkürzungen eher in eine Fachsprache gehören und so mehrere Kandidaten für ein Dokument generieren, werden weitere groß geschriebene Wörter die Kandidatenmenge einschränken.

Ein Problem stellen Satzenden dar, da neue Sätze prinzipiell mit einem groß geschriebenen Wort beginnen. Diese Wörter sind jedoch häufig keine Sinn tragenden Wörter, sondern Stoppwörter[2]. Um dem Problem zu begegnen, wird versucht die Stoppwörter zu eliminieren bevor die Suchbegriffe ausgewählt werden. Die Stoppwörter werden in zwei Varianten eliminiert. Eine arbeitet mit einer Stoppwortliste, die vom Benutzer vorgegeben wird. Die andere eliminiert alle Wörter, die weniger als vier Buchstaben hat. Die zweite Variante ist sicherlich fehleranfälliger, aber trotzdem als Annäherung an eine echte Stoppwortliste verwendbar, wenn keine echte zur Verfügung steht.

Der Operator bevorzugt Wörter mit vielen großen Buchstaben. Je mehr Großbuchstaben ein Wort enthält, desto eher wird es ausgewählt. Üblicherweise handelt es sich dabei um Abkürzungen oder Eigennamen.

Zur Redefinition der Suchanfrage wird das Standardverfahren genutzt.

3.4.1.2 Namen von Personen

Dieses Strategem spezialisiert das in 3.4.1.1 vorgestellte Verfahren noch weiter. Es wird versucht im Originaldokument Eigennamen von Menschen zu identifizieren. Aus den ersten fünf gefundenen Namen wird eine Suchanfrage gebildet. Die Beschränkung auf fünf ergibt sich aus der Beschränkung Googles auf zehn zugelassene Suchbegriffe. Vor- und Nachname müssen jeweils als ein Begriff gezählt werden. Sollte ein Name aus mehr als einem Vornamen bestehen, werden entsprechend weniger Namen ausgewählt.

Zur Eigennamenerkennung wird ein Wörterbuch bekannter, häufiger Vornamen genutzt. Wenn ein Vorname im Text gefunden wird, wird überprüft, ob das folgende Wort mit einem großen Buchstaben beginnt. Ist das der Fall, so geht die Heuristik davon aus, dass es sich bei diesem Wort um einen Nachnamen handelt. Folgt auf einen Vornamen ein oder mehrere weitere Vornamen und schließlich ein Nachname, so geht das Verfahren davon aus, dass es sich um einen längeren Namen handelt.

Das beschriebene Verfahren zur Namenserkennung ist einfach, ausreichend präzise und wird auch in komplexeren Verfahren zur „Named Entity Recognition“ eingesetzt (vgl. z. B. [Gri95, GWH+95]). Vor allem für Personalübersichtsdokumente und Dokumente mit vielen Literaturverweisen sollte dieses Strategem gute Ergebnisse liefern.

Zur Redefinition der Suchanfrage wird das Standardverfahren genutzt.

[2] Stoppwörter sind Wörter, die in fast jedem Text vorkommen wie z.B: der, die, das usw.

3.4.1.3 Häufige Wörter

Die Annahme, dass wichtige und charakterisierende Wörter häufiger in einem Dokument vorkommen als andere, ist eine Annahme, die in der Textklassifikation schon häufig erfolgreich gemacht wurde. Aus diesem Grund wird in diesem Strategem versucht die Suchmaschine mit den am häufigsten im gesuchten Dokument vorkommenden Wörtern, die nicht Teil einer Stoppwortliste sind, zu befragen. Da diese Wörter laut Annahme charakteristisch für das Dokument sein müssten, sollten sich die Wörter auch in neueren Versionen des Dokuments wiederfinden, vorausgesetzt, es werden inhaltsähnliche Dokumente gesucht. Für die Identifikation der Suchbegriffe wird eine TF–IDF–ähnliches Bewertungsmaß für die Wörter genutzt.

Statt das in Abschnitt 3.3.2 beschriebene „echte“ TF–IDF zu benutzen wird versucht die Suchbegriffe nach folgendem Verfahren auszuwählen:

> Wähle diejenigen Wörter, die am häufigsten in der bekannten Version des Dokuments vorkommen *und* keine Stoppwörter sind *und* die in weniger als 50% aller zu überwachenden Dokumente vorkommen. Stoppwörter werden durch eine Stoppwortliste vorgegeben.

Die Entscheidung, zunächst nicht „echtes“ TF–IDF zu verwenden, war dadurch begründet, dass die Dokumentfrequenzen der einzelnen Wörter nicht vorliegen. Andererseits ist klar, dass Wörter, die in mehr als der Hälfte der gesehenen Dokumente vorkommen, nicht zur Spezifikation des Dokuments beitragen.

Im zweiten Schritt — bei größeren Dokumentkollektionen — stellte sich heraus, dass die Berechnung der inversen Dokumentfrequenz aus den gesehenen Dokumenten offenbar doch zur besseren Differenzierung zwischen den Suchwörtern ausreicht und bessere Ergebnisse ermöglicht (s. 4.4.3).

Zur Redefinition der Suchanfrage der beiden Varianten des Strategems wird das Standardverfahren genutzt.

3.4.1.4 Phrasensuche

Phrasen sind für Google eine Aneinanderreihung von Wörtern, die durch Anführungszeichen eingeschlossen werden. Die Auswahl einer charakteristischen Phrase, um ein Dokument wiederzufinden, scheint erfolgversprechend.

Dieses Strategem implementiert ein Verfahren, das das in [Eul01] vorgeschlagene Verfahren zur Satzauswahl (vgl. [Eul01] Kap. 4.3) zugrunde legt. Die Grundidee ist, dass charakteristische Phrasen aus charakteristischen Wörtern bestehen. Wie charakteristisch ein Wort für ein Dokument ist, wird in diesem Operator durch TF-IDF (s. 3.3.2) bestimmt. Also ergibt sich die Spezifizität S einer Phrase P, welche aus einer Menge von Wörtern W_P besteht, für ein Dokument D aus der Gleichung:

$$S_D(P) = \sum_{w \in W_P} S_W(w) \tag{3.1}$$

mit

$$S_W(w) = \begin{cases} 0: & w \quad \text{ist Stoppwort lt. Stoppwortliste} \\ \text{TF–IDF}(w): & \text{sonst} \end{cases}$$

Da Google nicht mehr als zehn Suchwörter akzeptiert, dürfen die Phrasen nur aus zehn Wörtern bestehen. Für die Zerlegung des Originaldokuments in Zehn-Wort-Ketten werden zwei verschiedene Verfahren betrachtet:

1. Das Dokument wird zunächst in Sätze zerlegt. Ein Satzende wird durch die Zeichen `.!?:` markiert. Für jeden der gefundenen Sätze, wird nun die Wortkette aus höchstens zehn aufeinander folgenden Wörtern aus dem Satz gesucht, die den höchsten Wert $S_D(P)$ nach Gleichung (3.1) aufweist. Die Suchphrase ist diejenige Wortkette, die unter allen Wortketten aus allen Sätzen die beste Bewertung erreicht hat.

   Die Phrasen sind auf Sätze beschränkt.

2. Das Dokument wird in Wortketten aus zehn Wörter zerlegt. Es wird zunächst das erste Wort zusammen mit den neun folgenden Wörtern betrachtet, dann das zweite mit den folgenden neun Wörtern usw. Satzzeichen werden ignoriert, so dass die Wortketten sich auch über mehrere Sätze erstrecken können. Für jede der so gebildeten Wortketten wird wird nach Gleichung (3.1) die Phrasenbewertung errechnet. Diejenige Phrase mit der besten Bewertung wird für die Suchanfrage verwendet.

Zur Spezialisierung wird in beiden Fällen das Standardverfahren verwendet. Zur Generalisierung wird in 1. ein anderes Verfahren verwendet. Es wird der Satz ausgewählt, der das beste Bewertungsergebnis unter den noch nicht gewählten Sätzen hat.

Zusätzlich wird noch eine dritte Variante der Phrasensuche implementiert: Das Dokument wird zunächst in Sätze zerlegt. Von diesen Sätzen wird derjenige Satz mit den meisten Zeichen ausgewählt. Die ersten zehn Wörter dieses Satzes werden als Suchphrase benutzt. Zur Redefinition der Suchanfrage wird das Standardverfahren genutzt.

Diese Variante entspricht in etwa dem, was ein Mensch macht, wenn er einen Text durch Phrasensuche wiederfinden möchte. Er sucht sich einen langen Satz und kopiert ihn in das Abfrageformular. Dort wird Google dem Nutzer mitteilen, dass es nur die Wörter bis zum zehnten Begriff verwendet und den Rest ignoriert, aber die Ergebnisse sind oft brauchbar.

3.4.1.5 Zufällige Auswahl von Suchbegriffen

In den vorangegangenen Abschnitten wurden bestimmte Annahmen gemacht, warum die Auswahl bestimmter Suchbegriffe günstig sei. Nun sollen keine Annahmen gemacht werden. Es werden zufällig Wörter aus dem Text gezogen. Es werden drei Varianten unterschieden:

1. Alle Wörter des Originaldokuments sind unabhängig von der Häufigkeit ihres Auftretens gleich wahrscheinlich.

2. Alle Wörter des Originaldokuments sind unabhängig von der Häufigkeit ihres Auftretens gleich wahrscheinlich. Stoppwörter werden jedoch zuvor mit Hilfe einer Stoppwortliste eliminiert.

3. Die Wahrscheinlichkeit, dass ein Wort ausgewählt wird, ist abhängig von der Häufigkeit mit der es im Originaldokument vorkommt. Häufige Wörter werden mit einer größeren Wahrscheinlichkeit ausgewählt.

Die Zufallszahlen werden durch den Java-eigenen Zufallszahlengenerator erzeugt.

Ziel dieser Vorgehensweise ist es, zu überprüfen, ob die vorgeschlagenen Annahmen in den Abschnitten 3.4.1.1 bis 3.4.1.4 tatsächlich Verbesserungen bezüglich der Suchergebnisse liefern oder ob die Annahme ggf. sogar Hindernisse darstellen und wenn ja, in welchen Fällen.

3.4.1.6 Suche nach Verweisen

Dieses Strategem soll versuchen, ein Dokument wieder zu finden, indem es Google nach Dokumenten befragt, die dieselben Verweise enthalten. Es wird nach der URL, auf die verwiesen wird, gefragt, da nur diese Art der Verweisanfrage von einem Dokument auf andere unterstützt wird.

Damit eignet sich dieser Operator vermutlich prinzipiell auch, um strukturgleiche Seiten wiederzufinden.

Zur Redefinition wird das Standardverfahren leicht variiert: Statt Wörter werden Verweis-URLs aus der Suchanfrage herausgestrichen.

3.4.2 Verweisbasierte Strategeme

Strategeme, die beginnend bei ausgewählten Startpunkten auf der Verfolgung von Verweisen basieren, sollen in diesem Abschnitt beschrieben werden.

Bei den verweisbasierten Suchen ist es notwendig, die richtigen Verweise weiterzuverfolgen und Sackgassen möglichst früh zu erkennen (s. 2.2). Die fortgeschrittenen Verfahren verlangen, dass der Benutzer sie zunächst mit vielen Informationen über das gesuchte Themengebiet versorgt, damit Klassifikatoren angelernt werden können (s. 2.2.1.4, 2.2.1.5 und 2.2.2) oder das Verfahren braucht eine große Datenkollektion, um gute Ergebnisse zu erreichen (s. 2.2.1.1 und 2.2.1.2).

Andererseits bieten die Eigenschaften der Verweisstrukturen (s. 2.1.4) gute Voraussetzungen für gute Ergebnisse durch einfache Breitensuchen (s. 2.2.3).

3.4.2.1 Einfache Breitensuche

Ausgehend von einem oder mehreren Startpunkten, verfolgt die Breitensuche alle Verweise, die maximal drei Schritte von diesem Startpunkt entfernt sind. Verweise deren URL schon einmal besucht wurden, werden kein zweites Mal verfolgt.

Bei der Auswahl der Startpunkte sind drei Varianten denkbar:

(A) Von der URL aus, die als fehlerhaft klassifiziert wurde, schneidet der Operator nacheinander die Teile hinter einem „/“ weg. Dann wird versucht ein Dokument von der entstehenden URL zu laden. Gelingt dies, dann ist diese URL ein Startpunkt für die Breitensuche. Hier wird die Tatsache ausgenutzt, dass sich die Verweisstrukturen häufig auch in den Verzeichnisstrukturen auf dem betreffenden Web-Server widerspiegeln. Oft werden URLs, die keinen Dateinamen am Ende enthalten vom

Web-Server automatisch einer Default–Dokument zugeordnet. Bei diesem Default–Dokument handelt es sich häufig um eine Übersichtsseite mit Verweisen auf untergeordnete Themengebiete.

Beispiel:
Die URL `http://www-ai.cs.uni-dortmund.de/LEHRE/DIPLOM/OFFEN/agenten.html` sei fehlerhaft. Dann generiert der Operator nacheinander die Startpunkte:

- `http://www-ai.cs.uni-dortmund.de/LEHRE/DIPLOM/OFFEN/`
- `http://www-ai.cs.uni-dortmund.de/LEHRE/DIPLOM/`
- `http://www-ai.cs.uni-dortmund.de/LEHRE/`
- `http://www-ai.cs.uni-dortmund.de/`

(B) Von der URL aus, an der zuvor einmal der richtige Inhalt gelegen hat, wird dasselbe Verfahren wie in (A) angewendet.

Diese Variante macht nur dann Sinn, wenn z. B. im Laufe des Lebenszyklus des Dokuments eine Weiterleitung gesetzt wurde und diese nun fehlerhaft ist. In diesem Fall unterscheidet sich die URL, die als fehlerhaft klassifiziert wurde, von derjenigen, die dem so genannten alten Inhalt der URL zugeordnet ist. Dieser Fall sollte jedoch vom URL-Checker in [Mas03] behandelt werden und wird daher hier nicht weiter verfolgt.

(C) Bestimmt die fehlerhafte URL als Startpunkt. Dieser Operator ist dann erfolgreich, wenn sich an der fehlerhaften URL ein gültiges Dokument befindet und dieses Dokument einen Verweis auf das eigentlich unter der URL gesuchte Dokument enthält. Das ist z. B. dann der Fall, wenn statt einer HTTP-Weiterleitung (s. [FGM+99]) einfach ein Link auf den neuen Ort des Dokuments angegeben wird.

3.4.2.2 Rückverweise

Hinter diesem Strategem steckt die Idee, die Intelligenz und den Fleiß *anderer* Menschen oder Agenten auszunutzen. Es gilt die Dokumente, die auf ein überwachtes, noch vorhandenes Dokument verweisen, zu erfassen und den Ankertext des Verweises zu speichern. Verschwindet das zu überwachende Dokument, werden die zuvor erfassten Dokumente aufgesucht, um mit Hilfe des gespeicherten Ankertextes und des von ihm beschriebenen Verweises das verschwundene Dokument wiederzufinden. Die Hoffnung ist, dass nach dem verschwinden des gesuchten Dokuments sich schon ein Verwalter der anderen Dokumente die Mühe gemacht hat, das entschwundene Dokument zu suchen, es gefunden und seine Verweise angepasst hat.

Je einzigartiger der Ankertext auf den referenzierenden Dokumenten ist, desto weniger Verweise muss PageTracker bei diesem Operator verfolgen. Probleme treten vor allem dann auf, wenn der Ankertext leer ist oder die Verweise nicht durch Texte, sondern z. B. durch Grafiken gekennzeichnet sind.

3.4.3 Abschluss der Suche

Jede Suche wird durch die Ausführung von zwei Operatoren abgeschlossen. Zunächst wird unter den Kandidaten, die die Strategem-Operatoren erzeugt haben, der beste ausgesucht (s. 3.4.3.1). Zum endgültigen Abschluss der Suche wird überprüft, ob alle „verschwundenen“ Dokumente wiedergefunden wurden (s. 3.4.3.2).

3.4.3.1 Kandidatenauswahl

Derjenige Kandidat wird als Ersatz für das „verschwundene“ Dokument ausgewählt, der die höchste Konfidenz für die Richtigkeit aufweist.

Die Konfidenz errechnet sich durch:

$$\text{Konfidenz(C)} := \frac{\text{maximal zulässige Distanz} - \text{Distanz}(O, C)}{\text{maximal zulässige Distanz}} \tag{3.2}$$

Dabei ist C das Kandidaten–Dokument und O das Originaldokument. Die Strategem–Operatoren stellen sicher, dass keine Kandidaten erzeugt werden, die eine größere Distanz vom Originaldokument aufweisen, als zulässig ist.

3.4.3.2 Erfolg der Suche

Der Fertig–Operator wird am Ende jeder Suche ausgeführt. Aufgrund der fehlenden Möglichkeit zur Negation von Prädikaten in BotIShelly ([BFG+00]), erfolgt die Negation innerhalb des Operators, statt in der Vorbedingung.

Der Operator ist erfolgreich, wenn die benutzte A–Box (s. S. 28) *keine* Instanz des Konzepts fehlerhafte_url (s. S. 28) enthält. Wenn noch Instanzen dieses Konzepts existieren, scheitert er und gibt die betreffenden Instanzen in der Log-Datei aus.

Wird der Agent in einer Endlos–Schleife betrieben, so dass er immer wieder die zur Verfügung gestellten Operatoren nacheinander ausführt, bis alle fehlerhaften URLs eliminiert wurden, dann endet die Suche erst dann, wenn dieser Operator nicht fehlschlägt.

Im Standardablauf (s. Abb. 3.2) dient der Operator lediglich zur Ausgabe der im letzten Lauf nicht gefundenen Dokumente, da nach einem Durchlauf die Suche beendet wird.

4 Evaluation

In diesem Kapitel wird erörtert, wie erfolgreich der Agent PageTracker bei seiner Suche nach verschwundenen Dokumenten ist.

Dazu werden zunächst die verschiedenen Situationen (Fälle) beschrieben in denen nach einem Dokument gesucht wird. Aufbauend auf die Fallbeschreibungen wird in 4.2 erläutert wie sich die Testdaten zusammensetzen, die das Verhalten des Agenten testen sollen.

Abschnitt 4.3 beschreibt die Bewertungskriterien mit denen die Testläufe ausgewertet werden.

In Abschnitt 4.4 werden die Ergebnisse einzelner Strategeme getrennt voneinander vorgestellt und analysiert.

Den Abschluss des Kapitels bildet eine Diskussion verschiedener Strategien und die von ihnen genutzten Synergie–Effekte. Strategien werden durch die Aneinanderreihungen von Strategemen gebildet.

Die vollständigen Testdaten und Ergebnistabellen sind in den Anhängen A und B abgedruckt.

4.1 Testfälle

Es können im Wesentlichen folgende Fälle beim Lauf des Agenten auftreten:

1. Das Dokument befindet sich immer noch an derselben Stelle wie beim letzten Agentenlauf und wurde
   a) nicht geändert.
   b) mit Änderungen versehen.
   c) vollkommen neu gestaltet.
2. Das Dokument wurde innerhalb der Dokumenthierarchie verschoben. Es befindet sich also auf demselben Web-Server und wurde
   a) nicht geändert.
   b) mit Änderungen versehen.
   c) vollkommen neu gestaltet.

3. Das Dokument ist verschoben worden, befindet sich aber nicht mehr auf demselben Web-Server und wurde
   a) nicht geändert.
   b) mit Änderungen versehen.
   c) vollkommen neu gestaltet.
4. Das Dokument existiert überhaupt nicht mehr.

Die Fälle 1a und 1b testet, ob PageTracker Dokumente an ihren alten Plätzen wiederfindet, wenn sie angeblich verschwunden sind.

Fall 2a sollte mit den von PageTracker zur Verfügung gestellten Spidern (s. 3.4.2) zu lösen sein, weil alle zu testenden Güte-Kriterien eine exakt gleiche Seite identifizieren können sollten. Da das vorher bekannte Dokument inklusive seines alten Aufenthaltortes den Ausgangspunkt darstellt, sollte es auch zu finden sein. Der Fall 2b unterscheidet sich nur insofern von Fall 2a, als dass zusätzlich das Ähnlichkeitsmaß die „eigentlich“ richtige Seite erkennen muss.

Für den Fall 3 reichen die Spider im Allgemeinen nicht aus, da sonst auf das Dokument, an seinem neuen Ort, von einem anderen Dokument, in der vom Spider erreichbaren Dokumentstruktur, verwiesen werden muss. Im allgemeinen Fall müssen Suchmaschinen den Einstiegspunkt bieten, damit keine Benutzer–Interaktion notwendig wird, wie in 1.2 gefordert.

Für die vorgestellten suchmaschinenbasierten Operatoren ist die Unterscheidung in die Fälle 1,2 und 3 nicht wesentlich, da die suchmaschinenbasierten Strategeme die Ortsinformation fast nicht nutzen. Die einzige Information, die aus der URL entnommen wird, ist der Dokumentdateiname, wenn bei der Redefinition nach dem Standardverfahren spezialisiert (s. 3.4.1, S.36) wird. Welchen Einfluss diese Redefinitionsmöglichkeit auf die Ergebnisse hat, wird dadurch untersucht, dass beim Auswerten der Suchläufe, diejenigen Ergebnisse, die einen Spezialisierungsschritt enthalten besonders gekennzeichnet (s.Anhang B) werden. Wenn der Dateiname während des Verschiebens geändert wurde, werden die Dokumente, die gefunden wurden, weil mit Hilfe des Dateinamens spezialisiert wurde, nicht gefunden. Die Auswirkung auf die Ergebnisse der einzelnen Strategeme werden in den Abschnitt 4.4 diskutiert.

Durch diese Vorgehensweise kann außerdem der Erfolg bzw. die Wichtigkeit der Spezialisierung bewertet werden.

Bis auf den evtl. Einfluss der Dokumentnamensänderung stellt das Verschieben der Dokumente an andere Orte für die suchmaschinenbasierten Strategeme also keinen Unterschied dar. Für diese Strategeme ist vor allen Dingen die Unterscheidung (a), (b) und (c) wichtig. Die Fälle (a) sind die einfachsten, da hier alle aus dem Originaldokument herausgegriffenen Suchwörter noch im zu findenden Dokument vorkommen.

Die Fälle (b) sind schwieriger, weil das zu findende Dokument nicht mehr alle Wörter des Originaldokuments enthalten muss. Es können daher nicht beliebige Wörter des Originaldokuments zur Suche verwendet werden.

Die Fälle (c) sind die problematischsten, da bei einer vollkommenen Neugestaltung des Dokuments evtl. auch die Information verloren geht, aufgrund derer ursprünglich auf das

Dokument verwiesen wurde. Weil die Auswahl der Suchwörter für die Suchmaschinenanfrage sich auf die Begriffe beschränken muss, die am wahrscheinlichsten auch in der neuen Version des Dokuments vorkommen, scheint eine Auswahl der Wörter aufgrund ihrer Häufigkeit mit oder ohne Berücksichtigung der Dokumenthäufigkeit viel versprechend.

Der Fall 4 ist insofern interessant, dass der Agent auch keine anderen Seiten als die nun verlorene „wiederfinden“ sollte. Das Problem reduziert sich auf die korrekte Erkennung eines Dokuments. Dieses Problem soll jedoch in [Mas03] untersucht und gelöst werden. Es wird hier daher nicht weiter behandelt.

4.2 Beschreibung der Testmengen

Zu den im vorangegangen Abschnitt erläuterten Testfällen, sollen Testmengen erstellt werden. Die Erstellung der Testmengen geschieht wie folgt:

1. **Testfälle 1a, 2a, 3a** Für die Fälle 1a, 2a und 3a werden die Web-Seiten des Lehrstuhls für Künstliche Intelligenz, des Fachbereichs Informatik an der Universität Dortmund, sowie Seiten aus dem MLnet[1] verwendet. Dabei werden paarweise Vertauschungen vorgenommen, so dass dem Agenten z. B. vorgegaukelt wird, dass sich im Dokument „offene Diplomarbeiten“[2] nun der Inhalt der „abgeschlossenen Diplome“[3] verbirgt.

2. **Testfälle 1b, 1c** Für die Fälle 1b und 1c wird die Internet-Archiv-Maschine Wayback [AI01] herangezogen. Dabei wird dem Agenten vorgegaukelt, dass sich eine alte Version eines Dokuments, die im Archiv der Wayback-Maschine verfügbar ist, unter der aktuellen URL zu finden ist. Der Agent soll dann das aktuelle Dokument an der bekannten Stelle wiederfinden. Je nach Größe der vorgenommenen Änderungen an den Dokumenten fallen die so erzeugten Testfälle unter den Fall 1b oder 1c. Es wird zum Beispiel der Inhalt des Start–Dokuments des Lehrstuhls für Künstliche Intelligenz an der Universität Dortmund[4] in der Version vom 26.09.2001 aus dem Wayback-Archiv entnommen und dem Agenten als „verschwundener“ Inhalt des Dokuments unter der URL `http://www-ai.cs.uni-dortmund.de` vorgesetzt. Das Dokument wurde in der Zwischenzeit geändert. Die gesuchte Version befindet sich aktuell unter der URL `http://www-ai.cs.uni-dortmund.de`. Das weiß der Agent jedoch erst nachdem er es dort wiedergefunden hat.

3. **Testfälle 2b, 2c,3b, 3c** Hier kann das Vorgehen aus der ersten Testmengendefinition übernommen und mit der zweiten Testmenge kombiniert werden. Es wird dem Agenten gegenüber behauptet, dass der Inhalt sich verschoben hat, und zusätzlich wird ein, in der Wayback-Maschine gefundener Inhalt als Originaldokumentinhalt ausgegeben.

[1] `http://www.mlnet.org`
[2] `http://www-ai.cs.uni-dortmund.de/LEHRE/DIPLOM/OFFEN/`
[3] `http://www-ai.cs.uni-dortmund.de/LEHRE/DIPLOM/diplom_fertig.html`
[4] `http://www-ai.cs.uni-dortmund.de`

Für die vorgestellten suchmaschinenbasierten Strattegeme unterscheidet sich diese Menge nicht von Menge 2. Die Behandlung unterschiedlicher Dokumentdateinamen, die durch eine Verschibung zu Stande kommen können, geschieht wie in Abschnitt 4.1 beschrieben.

Für das SpiderA–Strategem (s. 3.4.2.1) ist die virtuelle Verschiebung des Dokuments wesentlich. Der Inhalt des Originaldokuments ist für den Suchvorgang jedoch insofern unwichtig, als dass aus diesem keine Informationen für die Suche genutzt wird, außer um ihn mit den neu gefundenen Dokumenten zu vergleichen. Für den SpiderA gleicht diese Testmenge also der Menge 1.

Für das SpiderC–Strategem (s. 3.4.2.1) ist die virtuelle Verschiebung ebenfalls wichtig, da zunächst an der alten Stelle noch ein Mal nachgeschaut wird, ob das Dokument wirklich nicht mehr an der zuletzt bekannten Stelle zu finden ist. Diese Überprüfung wird durch die virtuelle Verschiebung, ein funktionierendes Ähnlichkeitsmaß vorausgesetzt, immer fehlschlagen. Im Anschluss daran wird der alte Ort des Dokuments nicht mehr betrachtet und nur der Inhalt des Dokuments ist entscheidend. Das bedeutet, dass sich diese Testmenge für den SpiderC so darstellt wie die Menge 2 mit dem Unterschied, dass die Suche fehlschlägt, wenn das Dokument keinen Verweis auf sich selbst enthält.

Das Strategem, das Rückverweise nutzt (s. 3.4.2.1), wird auf dieser Testmenge keine Erfolge haben, da sich weder die Inhalte noch die Orte real verändert haben. Das bedeutet, dass andere Nutzer des Internets, auf die dieses Strategem angewiesen ist, keine Veranlassung sehen werden, ihre Verweise anzupassen.

Da abzusehen ist, dass diese Testmenge für keines der zu testenden Strategeme im Kern eine Situation darstellt, die nicht schon von den beiden anderen Testmengen abgedeckt ist, werden keine weiteren Testdaten für diese Testmenge gebraucht.

4. **Testfall 4** Wie zuvor beschrieben, wird dieser Fall nicht weiter untersucht und daher keine Testmengen für diesen Fall generiert.

Die Dokumente, deren Inhalte aus dem Archiv der Wayback-Maschine entnommen werden, werden von der Suchmaschine Google[5] nicht indexiert, so dass die Seiten nicht im Wayback–Archiv wiedergefunden werden. Das vereinfacht die Testmengenerstellung. Schließlich sollen die neuen Dokumente an den neuen Orten wiedergefunden werden.

Als Grunddaten für die Testmengen dienen einerseits eine Auswahl von Universitäts–Dokumenten und andererseits als Kontrollmenge eine Reihe von Dokumenten vom Web–Server des Westdeutschen Rundfunks[6].

Die in der Universitätsdaten–Testmenge enthaltenen Daten sind durch drei Vorgänge entstanden:

[5] `http://www.google.de`
[6] `http://www.wdr.de`

1. Mit einem Spider wurden ausgehend vom Startdokument `www-ai.cs.uni-dortmund.de` bis zu einer Tiefe von zwei, verschiedene URLs aufgesammelt. Anschließend wurden alte Versionen dieser Dokumente im Archiv der Wayback-Maschine aus dem Jahr 2001 gesucht. Diese alten Versionen der Dokumente werden als alte Inhalte genutzt.

2. Einige URLs zu Dokumenten von Personen, die am MLnet teilnehmen, wurden ausgewählt. Als alte Inhalte wurden Dokumente ausgewählt, die einen Bezug zur Lehre haben und in derselben Domain liegen wie das fehlerhafte Dokument.

3. Aktuelle Dokumente aus den Webdokumenten des Lehrstuhls für künstliche Intelligenz werden ausgewählt. Durch paarweises vertauschen von Inhalt und Ort zweier Dokumente, werden Dokumente virtuell verschoben.

Die Testdaten, die wie in 1. beschrieben erstellt wurden, testen den Umgang des Agenten mit veränderten Dokumenten. Die Testdaten unter 2 und 3 prüfen, wie sich veränderte Orte auf die Suchergebnisse auswirken. Sie sind vor allem als Testdaten für die Spider gedacht.

Die Kontrollmenge dient dazu, zu prüfen, ob die Ergebnisse aus der Universitäts–Testmenge übertragbar sind auf andere Sach- und Themenbereiche. Aus diesem Grund wurde eine Menge von Dokumenten gewählt, deren Inhalt sich mit verschiedenen Themen aus den WDR–Themenbereichen Politik, Wirtschaft und Freizeit beschäftigt. Aufgrund der Namenskonventionen des Westdeutschen Rundfunks zur Benennung seiner Dokumente, fallen die Dokumente mehrheitlich in die Testmengenkategorie 1. Es gibt fast keine zwei Versionen eines Dokuments gleichen Namens. „Fast“ deshalb, weil es einen dynamischen, täglich aktualisierten Teil im Layout jedes Dokuments gibt und Übersichtsdokumente gelegentlich ein neues Thema aufnehmen. Die Schwierigkeit für den Agenten liegt bei dieser Testmenge darin, dass die Themen aktuell sind und von vielen anderen Rundfunksendern auch behandelt werden. Die Suchanfragen müssen daher so gestaltet werden, dass nicht die Dokumente anderer Sender zum gleichen Thema wiedergefunden werden.

Außerdem kann mit dieser Testmenge untersucht werden, wie empfindlich die suchmaschinenbasierten Strategeme darauf reagieren, wenn die Suchmaschine nicht viel Zeit hatte, um die Dokumente zu erfassen. Dazu wird das Aufsammeln der Testdaten und die anschließende Suche nach den „verschwundenen“ Dokumenten zeitnah zueinander stattfinden.

Der Agent sucht jeweils 62 Dokumente aus der Universitätsdokumentenmenge und der WDR-Dokumenten-Menge.

4.3 Bewertungskriterien

Zur Bewertung der Testergebnisse müssen Bewertungskriterien aufgestellt werden. Da die Aufgabenstellung (s. 1.2) den Schwerpunkt auf das Finden der richtigen Dokumente legt und keine Benutzerinteraktion gewünscht ist, müssen die Bewertungskriterien diesen Schwerpunkt berücksichtigen.

Der Maßstab für den Erfolg des Agenten ist daher die Anzahl der wiedergefundenen Dokumente, gemessen an allen Dokumenten die der Agent finden soll.

Oft wird zusätzlich zu diesem Maß im Information Retrieval die Anzahl der „richtig“ gefundenen Dokumente (Precision) und der Prozentsatz der Dokumente von denen bekannt ist, dass das Suchsystem sie hätte finden können, wenn die gegebene Anfrage vorlag (Recall), angegeben. Diese beiden Maße sind im vorliegenden Fall aber nicht aussagekräftig. Das Maß Precision bewertet das Ähnlichkeitsmaß, welches entweder als perfekt (Precision ist 100%) vorausgesetzt wird oder nicht Gegenstand dieser Arbeit ist. Der Recall–Wert ist im vorliegenden Fall entweder 0 oder 1. Entweder wird das gesuchte Dokument gefunden oder eben nicht. Da es nicht Ziel des Agenten ist, möglichst viele Fundorte desselben Dokuments zu finden, sondern nur eines, ist dieses Maß nicht anwendbar.

Um herauszufinden wie sehr sich ein Dokument ändern darf, damit es noch gefunden werden kann, wird vor Beginn der Suche a priori eine einfache Bewertung der Schwierigkeit der Suche nach einem Dokument vorgenommen. Da im Test-Szenario bekannt ist, welches Dokument an welcher Stelle wiedergefunden werden soll[7], kann für jedes Dokumentpaar ⟨Originaldokument, neues Dokument⟩ die Anzahl der Wörter gezählt werden, die in der neuen Version des Dokuments nicht mehr vorkommen. Unterschiedliche Vorkommenshäufigkeiten werden nicht beachtet, solange das Wort im neuen Dokument überhaupt vorkommt.

Dieses Maß zur a–priori–Abschätzung zu nutzen ist sinnvoll, weil die verwendete Suchmaschine nach Dokumenten sucht, die dieselben Wörter wie die Anfrage enthält. Der Agent wählt aus dem alten Dokument verschiedene Wörter aus, um das gesuchte Dokument zu finden. Je weniger Worte aus dem alten Dokument im neuen Dokument überhaupt noch vorkommen, desto schwieriger ist es für ihn die richtige Auswahl zu treffen. Befragt der Agent die Suchmaschine mit einem Wort, das nicht mehr in der neuen Version des gesuchten Dokuments vorkommt, wird die Suche fehlschlagen.

Die Abschätzung wird wie folgt berechnet: Sei O die Menge der Nicht–Stopp–Wörter im Originaldokument und N die Menge der Nicht–Stopp–Wörter des neuen Dokuments, dann ist $S(O, N)$ die a–priori–Abschätzung:

$$S(O, N) = \frac{\|O \setminus N\|}{\|O\|} * 100 \tag{4.1}$$

Anhand der berechneten Schwierigkeitsabschätzung werden die zu suchenden Dokumente in drei Klassen eingeordnet:

einfach wiederzufinden. Weniger als 30% der Nicht–Stopp–Wörter des Originaldokuments sind nicht mehr im neuen Dokument enthalten.

schwierig wiederzufinden. Bis zu 60% der Nicht–Stopp–Wörter des Originaldokuments sind nicht mehr im neuen Dokument enthalten.

sehr schwierig wiederzufinden. Mehr als 60% der Nicht–Stopp–Wörter des Originaldokuments sind nicht mehr im neuen Dokument enthalten.

Die A–Priori–Abschätzungen für die genutzten Testdaten sind in Anhang C angegeben.

[7] bis auf Standard-Ersetzungen wie z. B. *.cs.* durch *.informatik.* oder `www-ai.cs.uni-dortmund.de/` und `www-ai.cs.uni-dortmund.de/index.html`

Um die Güte der Abschätzung zu beurteilen, wurde eine a posteriori Feststellung der Schwierigkeit vorgenommen. Dokumente die von

0-2 Operatoren wiedergefunden wurden, werden als *sehr schwierig* wiederzufinden beurteilt.

3-7 Operatoren wiedergefunden wurden, werden als *schwierig* wiederzufinden beurteilt.

8-13 Operatoren wiedergefunden wurden, werden als *einfach* wiederzufinden beurteilt.

Zur Berechnung von Precision- und Recall–Werten ([Fuh00], S. 28) werden die Universitäts–Testdatenmenge und die WDR-Testdatenmenge zusammengenommen betrachtet. Es ergeben sich für die drei Kategorien, die in den Tabellen 4.1(a) bis 4.1(c) abgebildeten Kontingenztabellen zusammen mit der zugehörigen Precision/Recall–Tabelle 4.1(d). Werden die in Tabelle 4.1 gezeigten Ergebnisse in Form einer Makro-Bewertung (s.

	richtige Kategorie	falsche Kategorie
zugeordnet	4	1
nicht zugeordnet	3	116

(a) 0-2 Funde

	richtige Kategorie	falsche Kategorie
zugeordnet	8	1
nicht zugeordnet	13	102

(b) 3-7 Funde

	richtige Kategorie	falsche Kategorie
zugeordnet	95	15
nicht zugeordnet	1	13

(c) 8-13 Funde

	Precision	Recall
0-2 Funde	80,00%	57,14%
3-7 Funde	88,88%	38,09%
8-13 Funde	86,36%	98,96%

(d) Precision / Recall

Tabelle 4.1: Kontingenztabellen für die Schwierigkeitsabschätzung

[Fuh00]) zusammengefasst, ergibt sich ein Precision–Wert von 85,08% und ein Recall-Wert von 65,73%.

Die A–Priori–Abschätzung ist also hilfreich, um zu erkennen, warum einige Dokumente wesentlich besser wiedergefunden werden als andere. Jedoch berücksichtigt die in Gleichung (4.1) angegebene Formel nicht, wie viele Wörter insgesamt im Dokument vorkommen.

Wenn ein Dokument nur aus wenigen Wörtern besteht, kann es passieren, dass die zur Verfügung stehenden Wörter nicht ausreichen, um ein Dokument eindeutig wiederzufinden, obwohl alle Wörter auch in der neuen Version des Dokuments vorkommen. Die Suchmaschine kennt viele Dokumente, in denen die benutzten Wörter vorkommen, so dass das richtige Dokument nicht in den ersten zehn Suchergebnisrängen vorkommt. Ein Beispiel für so ein problematisches Dokument war am 16.08.2001 unter der URL `http://www.dfn.de/home.html` zu finden. Es enthält nur fünf verschiedene Wörter. Der restliche Inhalt wird durch Grafiken ausgedrückt. Am 11.06.2003 war das Dokument nicht wieder auffindbar (s. Tab. 4.2(a)), da viele andere Dokumente die fünf zur Verfügung stehenden Begriffe auch verwendeten. Dann kam es offenbar zu einer Neuberechnung des Rangs, denn nun ist das Dokument trotz der geringen Wortanzahl auffindbar (s. Tab. 4.2(b)).

Original	http://web.archive.org/web/20011214030127/http://www.dfn.de/home.html
Anfrage	dfn newsletter deutsches forschungsnetz verein
	Fund
	http://www.uni-protokolle.de/nachrichten/id/8316/
	http://www.golem.de/0207/20753.html
	http://www.berufsbildung.de/h_links_organisationen/dfn.php
	http://www.berlinews.de/archiv-2002/1570.shtml
	http://www.berlinews.de/archiv/543.shtml
	http://www.ub.fu-berlin.de/internetquellen/suchdienste/speziell/internet.html
	http://news.zdnet.de/zdnetde/newsv2/story/0,,t101-s2124752,00.html
	http://www.itschau.de/0207/20753.html
	http://www.chemie.de/info/presse/clb.php3?language=d
	http://www.teltarif.de/a/dfn/

(a) Suchlauf am 11.06.2003

Original	http://web.archive.org/web/20011214030127/http://www.dfn.de/home.html
Anfrage	dfn newsletter deutsches forschungsnetz verein
	Fund
	http://www.dfn.de/
	http://www.uni-protokolle.de/nachrichten/id/8316/
	http://www.berlinews.de/archiv-2002/1570.shtml
	http://www.berlinews.de/archiv/543.shtml
	http://www.ub.fu-berlin.de/internetquellen/suchdienste/speziell/internet.html
	http://www.britischebotschaft.de/en/embassy/r&t/notes/rt-note00.1021_HighSpeed.html
	http://www.golem.de/0207/20753.html
	http://www.german-embassy.org.uk/new_media_and_telecommunicatio.html
	http://www.teltarif.de/a/dfn/
	http://www.chemie.de/info/presse/clb.php3?language=d

(b) Suchlauf am 20.07.2003

Tabelle 4.2: Das Wiederauffinden von Dokumenten ist schwierig, wenn es nur wenige Wörter im Dokument gibt

Offensichtlich sind für eine genauere Schätzung der Schwierigkeit noch andere Problemquellen zu berücksichtigen.

4.4 Strategem–Einzelergebnisse

In diesem Abschnitt werden die Ergebnisse der einzelnen Strategeme getrennt voneinander vorgestellt. Dazu wird zu jedem Strategem angegeben, wie viele Dokumente von diesem Strategem wiedergefunden wurden. Außerdem wird versucht zu erklären, warum ausgewählte Dokumente nicht wiedergefunden wurden.

Ein Dokument gilt als Wiedergefunden, wenn unter den ersten zehn Dokumenten, deren URLs die Suchmaschine als Ergebnis einer Anfrage zurück liefert, das Gesuchte enthalten ist. Wenn ein Strategem mehrere Anfragen generiert, um ein Dokument wieder zu finden, dann gilt das Dokument als gefunden, wenn das Dokument durch eine der Anfragen wiedergefunden wird.

Sei $Q(D)$ eine Anfrage, die das Dokument D wiederfinden soll, dann ist $R\,(Q(D))$ die Menge der ersten zehn Dokumente, deren URLs die Suchmaschine als Ergebnis der

Anfrage zurück liefert.

Die Funktion $E : D \to \{\text{gefunden}, \text{nicht gefunden}\}$ entscheidet, ob das Dokument gefunden wurde oder nicht:

$$E(D) = \begin{cases} \text{gefunden}, & \exists D' \in R\left(Q(D)\right), \text{ so dass } G(M, D, D') = 1 \\ \text{nicht gefunden}, & \text{sonst} \end{cases} \tag{4.2}$$

mit

$$G(M, D, D') = \begin{cases} 1, & 0 \leq M(D, D') \leq d \\ 0 & \text{sonst} \end{cases} \tag{4.3}$$

Die Funktion $M : D \times D' \to [0, \infty[$ stellt das verwendete Ähnlichkeitsmaß dar. Wenn das Maß perfekt wäre, dann könnte die maximal zulässige Distanz d auf Null gesetzt werden. Es würden dann nur noch Dokumente zugelassen, die laut Ähnlichkeitsmaß gleich sind. Da das Maß „perfekt“ wäre, kämen keine Fehlentscheidungen zustande. Beim verwendeten Cosinus–Maß müssen jedoch Intervalle zugelassen werden. Bei zu kleinem d kann es vorkommen, dass für das gesuchte Dokument eine größere Distanz errechnet wird, als zugelassen ist.

Im Rahmen der Testläufe zeigte sich, dass es gelegentlich vorkommt, dass der Operator die richtige URL und einige weitere URLs von der Suchmaschine zurückgeliefert bekommen hat und dann die falsche URL ausgewählt hat, weil das Maß für die falsche URL eine geringere Distanz ausgerechnet hatte, als für das gesuchte Dokument. Ein Beispiel ist

Original	`http://web.archive.org/web/20010424025243/http://www-ai.cs.uni-dortmund.de/FORSCHUNG/index.html`	
Anfrage	morik katharina machine learning knowledge vol klingspor etal acquisition kaiser	
	Fund	**Bewertung**
	`http://www-ai.cs.uni-dortmund.de/FORSCHUNG/index.eng.html`	53
	`http://www-ai.cs.uni-dortmund.de/PERSONAL/morik.eng.html`	73
	`http://www-ai.cs.uni-dortmund.de/PERSONAL/klingspor.eng.html`	76
	`http://www-ai.cs.uni-dortmund.de/FORSCHUNG/`	72
	`http://www-ai.cs.uni-dortmund.de/PERSONAL/morik.html`	74
	`http://www-ai.informatik.uni-dortmund.de/PERSONAL/morik.eng.html`	73

grün gekennzeichnet: richtiges Ergebnis
rot gekennzeichnet: gewähltes Ergebnis

Tabelle 4.3: Fehlbewertung der Ergebnisse

in Tabelle 4.3 zu sehen. Hier liegt das Problem darin, dass das Dokument hauptsächlich aus Literaturangaben besteht. Die englische Version unterscheidet sich von der deutschen hauptsächlich in einem einführenden Text, der in der jeweiligen Sprache abgefasst ist. Die deutsche Version des Dokuments wurde um einige Einträge erweitert, die englische hingegen nicht. Daher ist die aktuelle englische Version, der alten deutschen Version ähnlicher bzgl. des verwendeten Cosinusmaß (s. 3.3.2) als die neue deutsche Version.

Ein ähnliches Problem tritt auf, wenn das Distanzmaß nicht genau genug differenzieren kann. In Tabelle 4.4 wird zum Beispiel das Rasmus–Projekt–Dokument `http://lrb.cs.uni-dortmund.de/Projekte/rasmus/` genauso bewertet wie das eigentlich

Original	http://web.archive.org/web/20011128101650/http://ls1-www.cs.uni-dortmund.de/	
Anfrage	ls1 lehrstuhl rasmus ratgeber	
	Fund	**Bewertung**
	http://ls1-www.cs.uni-dortmund.de/	57
	http://ls1-www.informatik.uni-dortmund.de/	57
	http://lrb.cs.uni-dortmund.de/Projekte/rasmus/	57

Tabelle 4.4: Maßgenauigkeitsprobleme

gesuchte Dokument `http://ls1-www.cs.uni-dortmund.de/`. Durch höhere Rechengenauigkeiten kann versucht werden dieses Problem zu lösen, jedoch kann das tiefer liegende Problem nicht behandelt werden: Es kann vorkommen, dass Änderungen an einem Dokument, denselben gemessenen Abstand verursachen, den ein beliebiges anderes Dokument schon zuvor zum Originaldokument hatte.

Ein weiteres Problem stellen Framesets dar. Framesets sind Behälter für verschiedene Dokumente, die durch Verweise so zusammengefasst und dargestellt werden, so dass sie wie ein Dokument erscheinen. Die Suchmaschine erfasst die Einzeldokumente aus denen das Gesamtdokument (das Frameset) besteht jedoch einzeln. Wenn Bestandteile aus den verschiedenen Einzeldokumenten gleichzeitig gefragt werden, kann es vorkommen, dass kein Dokument als Ergebnis zurückgeliefert wird, weil kein Teildokument alle Begriffe gleichzeitig enthält. Werden nur Begriffe aus einem Teildokument als Suchanfrage genutzt, dann wird i.d.R. auch nur dieses Teildokument (s. Tab. 4.5) gefunden.

Original	http://sfbci.cs.uni-dortmund.de/home/German/frameset.html	
Anfrage	Finden sfB German Methoden über Contents 531 web service sonderforschungsbereich	
	Fund	**Bewertung**
	http://sfbci.uni-dortmund.de/home/German/top.html	35
	http://sfbci.cs.uni-dortmund.de/home/German/top.html	35

Tabelle 4.5: Probleme mit Framesets

4.4.1 Großbuchstaben

Abkürzungen und Eigennamen werden häufig groß geschrieben. Sie können als charakteristisch für ein Dokument angesehen werden. Die Ergebnisse des in 3.4.1.1 vorgestellten Strategems werden nun erörtert.

Das in Tabelle 4.6 gezeigte Ergebnis stammt von der Variante der Begriffsauswahl, die Stoppwörter nicht eliminiert. Die Zahlen in Klammern geben an, wie viele Dokumente gefunden werden, wenn das Strategem nicht die Möglichkeit hat eine Spezialisierungsoperation (s. 3.4.1, S. 36) auszuführen. Es werden weniger als die Hälfte aller gesuchten Dokumente der Testmenge mit den Universitätsseiten gefunden.

Im zweiten Schritt wurde versucht mit einer einfachen Stoppwortberücksichtigung bessere Ergebnisse zu erzielen. Das Ergebnis in Tabelle 4.7 zeigt jedoch, dass die Behauptung, dass alle Wörter mit weniger als vier Buchstaben Stoppwörter sind, das Ergebnis auf der Universitäts-Testmenge sogar verschlechtert.

gefunden	nicht gefunden
29 (28)	33 (34)

(a) Universität

gefunden	nicht gefunden
50 (48)	12 (14)

(b) WDR

Tabelle 4.6: Großbuchstaben ohne Stoppwortberücksichtigung

gefunden	nicht gefunden
28 (28)	34 (34)

(a) Universität

gefunden	nicht gefunden
52 (51)	10 (11)

(b) WDR

Tabelle 4.7: Großbuchstaben mit simpler Stoppwortberücksichtigung

Diese Verschlechterung ist in diesem Fall darauf zurückzuführen, dass der Operator, der keine Stoppwörter kennt, auch einen allein stehenden Gedankenstrich als Wort interpretiert. Er fragt Google daher nach einem Dokument, das folgende Wörter enthält:

```
| [2000] [1999] [1998] [1997] [1996] [1995] [1994]
 -=>Offene -
```

Der Operator mit der simplen Stoppworterkennung verwirft den Gedankenstrich und das führende Zeichen —, da diese beiden „Wörter“ weniger als vier Buchstaben haben. Stattdessen fügt der Operator zwei Wortketten hinzu, die dazuführen, dass das gesuchte Dokument nicht gefunden wird:

```
[2000] [1999] [1998] [1997] [1996] [1995] [1994]
 -=>Offene PS-GZ-Volltext PDF-GZ-Volltext
```

Die Ergebnisse der beiden Varianten unterscheiden sich jedoch nicht signifikant, da bis auf dieses Ergebnis keine weiteren bemerkenswerten Unterschiede in den Testläufen zu finden sind. Die simple Stoppworterkennung reicht also zur Verbesserung des Verfahrens nicht aus.

gefunden	nicht gefunden
35 (35)	27 (27)

(a) Universität

gefunden	nicht gefunden
56 (56)	6 (6)

(b) WDR

Tabelle 4.8: Großbuchstaben mit Stoppwortberücksichtigung

Wenn statt der simplen Stoppworterkennung Stoppwortlisten genutzt werden, verbessert sich das Resultat (s. Tabelle 4.8). Da das Eliminieren von Stoppwörtern in der Textklassifikation und im Information Retrieval schon lange üblich ist, um die Ergebnisse zu verbessern, überrascht das Ergebnis nicht.

Insgesamt sind auch die Einzelergebnisse des Operators mit Stoppwortberücksichtigung besser. Auf der Universitäts-Testmenge findet diese Variante bis auf das Dokument unter der URL `http://dekanat.cs.uni-dortmund.de/HaPra/index.html` eine Obermenge der gefundenen Dokumente der Variante ohne Stoppwortberücksichtigung. Bei diesem Dokument tritt das Problem auf, dass das gesuchte Dokument nicht unter den Suchergebnissen der gestellten Anfrage ist, aber die gefundenen Dokumente schon so gut zu

passen scheinen, dass keine weitere Verallgemeinerung der Suchanfrage vom Agenten vorgenommen wird. Bei genauerer Untersuchung, warum die Abfrage

```
HAPRA EPRA SS GB HaPra Digitalelektronischen Hardwarepraktikum
Teilnehmerlisten Vorbesprechungen
```

nicht das gewünschte Ergebnis enthält, fällt auf, dass die beiden Wörter „GB“ (Abkürzung für Geschossbau) und „SS“ (Abkürzung für Sommersemester) nicht mehr in der Neufassung des Dokuments vorkommen.

Der Operator, der keine Stoppwörter kennt, braucht vier Versuche um eine Abfrage zu generieren, die überhaupt Ergebnisse liefert:

```
. - HAPRA/EPRA HAPRA
```

Diese Anfrage enthält die beiden Wörter, die den vorherigen Operator zum Stolpern brachten nicht mehr, so dass das gesuchte Dokument gefunden wurde. Offenbar sind die Begriffe „HAPRA/EPRA“ speziell genug.

Die Variante mit der simplen Stoppworterkennung, profitiert auch dieses Mal von der Tatsache, dass die störenden Wörter weniger als vier Buchstaben haben, so dass folgende Anfrage erfolgreich ist:

```
HAPRA/EPRA HAPRA EPRA HaPra Hapra-Angelegenheiten
Online-Anmeldung Praktikumsplatz! Digitalelektronischen
Hardwarepraktikum Teilnehmerlisten
```

Auf der WDR-Testmenge findet die Variante mit Stoppwortelimination durch Stoppwortlisten ebenfalls bis auf zwei Ausnahmen dieselben Dokumente wie die Variante ohne Stoppwortelimination. Hier liegt die Ursache jedoch darin, dass das WDR-Dokumentlayout einen Bereich vorsieht, der die aktuellen Schlagzeilen zu einem Themengebiet enthält. Die Wahrscheinlichkeit, dass Wörter aus diesem Bereich nicht in der Version des Dokuments, die Google erfasst hat, vorkommen, ist hoch. Genau das ist bei der Suche nach den beiden Dokumenten, die nicht auch von der Strategemvariante mit der Stoppwortliste gefunden werden, passiert. Durch die Redefinitionen werden in diesen Fällen die Anfragen so übergeneralisiert, dass die anschließenden Spezialisierungen nicht mehr ausreichen.

Lässt man das Spezialisieren gar nicht zu, um zu überprüfen, wie abhängig der Erfolg dieses Strategems von der URL-Information ist, stellt sich heraus, dass sich das Ergebnis der Variante, die TF-IDF verwendet weder auf der Uni- noch auf der WDR-Testmenge ändert. Die Variante mit einfacher Stoppwortelimination braucht für ein Dokument der WDR-Testmenge die Möglichkeit zu spezialisieren. Nur bei der einfachsten Variante verschlechtert sich das Ergebnis sowohl auf der Universitäts-Testmenge (um 1) als auch auf der WDR-Testmenge (um 2). Insgesamt ist die Spezialisierungsoperation also nicht notwendig, um die Ergebnisse dieses Operators zu erzielen.

Trotz der nicht ganz vollständigen Überdeckung der Suchergebnisse der Variante ohne Stoppwortlisten durch die Variante mit Stoppwortliste, ist die zu letzt genannte Variante so erfolgreich, dass die beiden anderen nicht mehr verwendet werden müssen.

gefunden	nicht gefunden
10 (10)	52 (52)

(a) Universität

gefunden	nicht gefunden
29 (28)	33 (34)

(b) WDR

Tabelle 4.9: Eigennamensuche

4.4.2 Namen von Personen

Dieses Strategem (vgl. 3.4.1.2) scheint auf den ersten Blick (s. Tab. 4.9) nicht besonders erfolgreich zu sein. Es kommen nicht in allen untersuchten Dokumenten Namen von Personen vor. Falls sie doch vorkommen, sind sie nicht unbedingt für ein Dokument spezifisch. Namen wie z. B. „Heinrich Müller“ fördern jede Menge URLs zu Dokumenten zu Tage, die nichts mit der gesuchten Seite über einen Informatik-Vorkurs zu tun haben (s. Tab. 4.10).

Original	`http://web.archive.org/web/20011224114917/http://ls7-www.cs.uni-dortmund.de/VKInf/`
Anfrage	„Heinrich M?ller“
	Fund
	`http://ls7-www.cs.uni-dortmund.de/~mueller/`
	`http://ls7-www.cs.uni-dortmund.de/~mueller/Mueller_d.html`
	`http://www.eg.org/EG_Member_Home_2229`
	`http://www.fpp.co.uk/Himmler/Mueller/death151299.html`
	`http://www.whonamedit.com/synd.cfm/3038.html`
	`http://www.whonamedit.com/doctor.cfm/2564.html`
	`http://fano.ics.uci.edu/cites/Author/Heinrich-Mueller.html`
	`http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/m/M=uuml=ller:Heinrich.html`
	`http://www.diegeschichteberlins.de/personen/mueller2.shtml`
	`http://ls7-www.informatik.uni-dortmund.de/ls7/staff/staff`

Tabelle 4.10: „Heinrich Müller“ ist zu unspezifisch

Wenn mehrere Namen in einem Dokument vorkommen, wie auf Publikationsseiten, Personalseiten oder ähnlichen Dokumenten, kann dieses Strategem von großem Nutzen für die Wiederauffindung sein. So findet dieses Suchstrategem z. B. die Seite mit den „Bücher, an denen der LS8 beteiligt ist“ sehr zielsicher (s. Tab. 4.11). Da die WDR-

Original	`http://web.archive.org/web/20010430193146/http://www-ai.cs.uni-dortmund.de/FORSCHUNG/buecher.html`
Anfrage	“Katharina Morik“ “Stefan Wrobel“ “Werner Emde“ “Michael Kaiser“ “Volker Klingspor“
	Fund
	`http://www-ai.cs.uni-dortmund.de/FORSCHUNG/buecher.eng.html`
	`http://www-ai.cs.uni-dortmund.de/FORSCHUNG/buecher.html`

Tabelle 4.11: Spezialdokumente werden gefunden

Testmenge aus Dokumenten besteht, die sich mit Politik und Wirtschaft beschäftigen, kommen in diesen Dokumenten verstärkt Personennamen vor. Das erklärt, warum eine größere Teilmenge der Testmenge gefunden werden kann.

Wird die Auswirkung der fehlenden Spezialisierungsmöglichkeit betrachtet, dann ist festzustellen, dass sich das Ergebnis auf der Universitäts–Testmenge nicht verändert und auf der WDR–Testmenge ein Dokument weniger gefunden wird. Die Spezialisierungsoperation hat also keinen großen Anteil am Gesamterfolg des Strategems.

4.4.3 Häufige Wörter

Die Idee, dass häufig in einem Dokument vorkommende Wörter zur Wiederauffindung des Dokuments genutzt werden können (vgl. 3.4.1.3), ist richtig. Schon in der Variante, die die Worthäufigkeit ohne Beachtung der Dokumenthäufigkeit nutzt, werden mehr als zwei Drittel der gesuchten Dokumente gefunden (s. Tab. 4.12).

gefunden	nicht gefunden
42 (42)	20 (20)

(a) Universität

gefunden	nicht gefunden
47 (47)	15 (15)

(b) WDR

Tabelle 4.12: Häufige Wörter nur TF

Tabelle 4.13 zeigt, dass das Ergebnis noch besser wird, wenn die inverse Dokumenthäufigkeit berücksichtigt wird (s. Tab. 4.13). Ein Beispiel für die Verbesserung des Ergebnisses

gefunden	nicht gefunden
45 (44)	17 (18)

(a) Universität

gefunden	nicht gefunden
51 (51)	11 (11)

(b) WDR

Tabelle 4.13: Häufige Wörter mit TF–IDF

ist die englischsprachige Version der Leitseite des Lehrstuhls für Künstliche Intelligenz. Die folgende Anfrage liefert unter anderen einen Verweis auf das gesuchte Dokument:

```
unit mining intelligence data news research learning
```

Diese Anfragekombination erreicht der Operator ausgehend von der Anfrage:

```
unit mining intelligence data news research learning
participates cup center
```

Die Variante des Strategems, welches nur die Häufigkeiten der Wörter berücksichtigt startet hingegen mit folgender Anfrage:

```
2001 Unit Mining Intelligence Data News Research Learning
Participates CuP
```

und findet das gewünschte Dokument nicht.

Die Jahreszahl ist offensichtlich problematisch. Diese Jahreszahl wird von der TF–IDF–Variante des Strategems nicht gewählt, da sie auch in mehreren anderen Dokumenten, die der Agent kennt, vorkommt.

Da für das Ranking von Google die Reihenfolge der Suchbegriffe eine Rolle spielt, bewirkt das Wort „2001“, welches das gesuchte Dokument nicht sehr von anderen Dokumenten unterscheidet, eine Ergebnisverschlechterung, denn es werden andere Dokumente nach vorne gerankt. Hinzu kommt, dass gerade bei einem Übersichtsdokument die Jahreszahlen, die in der Dokumentversion aus dem Jahr 2001 vorkommen, in der aktuellen Version nicht mehr vorkommen. Dieses Phänomen wiederholt sich bei dem Übersichtsdokument des Lehrstuhl 12 (`http://ls12-www.cs.uni-dortmund.de`) und der deutschen Version

der Übersichtsseite des Lehrstuhls 8. Auch hier wird die Jahreszahl nach vorne sortiert, so dass das gesuchte Dokument in der neuen Version nicht mehr gefunden werden kann.

Angesichts der Tatsache, dass die Jahreszahl sich so störend auswirkt, scheint es sinnvoll Jahreszahlen gänzlich aus den Suchanfragen zu streichen, doch das ist nur dann sinnvoll, wenn immer die neuste Version eines Dokuments gefunden werden soll. Sonst trägt die Jahreszahl die notwendige Information, die es erlaubt das richtige Dokument aus einem Archiv auszuwählen. Vermutlich wäre es sinnvoll Zahlen nach hinten zu sortieren, damit sie während der Redefinitionsschritte möglichst schnell aussortiert werden können. Alternativ könnten die Redefinitionsverfahren Zahlen gesondert behandeln.

Werden die Ergebnisse dieses Strategems im Zusammenhang mit der a-priori–Abschätzung (s. Anhang C) betrachtet, fällt auf, dass auf der Universitäts–Testdatenmenge, die Variante mit TF–IDF–Berücksichtigung selbst dann noch Ergebnisse erzielt, wenn andere Strategeme nichts mehr finden. In den Klassen der *schwierig* und *sehr schwierig* wiederzufindenden Dokumente, taucht diese Strategemvariante überdurchschnittlich häufig auf (vgl. Anhang C.1).

Das Ergebnis ließ sich auf den WDR–Testdaten jedoch nicht verifizieren, doch das kann auch an der geringen Anzahl Dokumente in den genannten Schwierigkeits–Klassen liegen.

In diesem Zusammenhang kann festgestellt werden, dass der Erfolg dieses Strategems fast nicht von der Möglichkeit zu spezialisieren abhängt. Andererseits ist das einzige Dokument, welches nicht mehr wiedergefunden wird, wenn die Möglichkeit des Spezialisierens nicht mehr eingesetzt werden kann, das Dokument `http://www-ai.cs.uni-dortmund.de/index.eng.html`, welches sonst auch nur von diesem Strategem in der TF–IDF–Variante wiedergefunden wird. Hätte der Spezialisierungsschritt übersprungen werden können, wäre die Suche aber trotzdem erfolgreich gewesen, da nicht die Spezialisierung ermöglichte das Dokument zu finden, sondern der sich anschließende Spezialisierungsschritt. Ohne diesen Schritt kommt es aber nicht zur angesprochenen Spezialisierung.

Original	`http://web.archive.org/web/20011225001946/http://www-ai.cs.uni-dortmund.de/index.eng.html`
Anfrage	unit mining intelligence data news research learning participates cup center
	Fund
	[...]
Anfrage	unit mining intelligence data news research learning participates cup inurl:index.eng.html
	Fund
	[...]
Anfrage	unit mining intelligence data news research learning
	Fund
	`http://www.cs.bham.ac.uk/news/jobs/cercia-rf.03/`
	`http://www-ai.cs.uni-dortmund.de/index.eng.html`
	`http://www-ai.cs.uni-dortmund.de/CONFERENCE/dataMiningInPractice.htm`
	`http://www.kdnet.org/control/group?ORDER_BY=group_name`
	[...]

Tabelle 4.14: Der Spezialisierungsschritt ist nur mittelbar relevant

4.4.4 Phrasensuche

Dokumente wiederzufinden, indem nach Satzstücken aus diesen Dokumenten gesucht wird, scheint viel versprechend (vgl. 3.4.1.4) zu sein. Die Variante, die satzübergreifende Phrasen zur Suchanfragengenerierung nutzt, findet jedoch weniger als die Hälfte der gesuchten Dokumente wieder (s. Tab. 4.15). Das widerspricht dem Eindruck, der vielleicht

gefunden	nicht gefunden
29 (26)	33 (36)

(a) Universität

gefunden	nicht gefunden
34 (33)	28 (29)

(b) WDR

Tabelle 4.15: Phrasensuche mit TF–IDF

in eigenen Erfahrungen gewonnen wurde. Das Problem liegt in der Textrepräsentation. Die von PageTracker gewählte Repräsentation ist offensichtlich nicht dieselbe, die Google wählt, so dass Wörter, die für PageTracker eine Kette bilden, in Google offenbar keine bilden. Die Umwandlung von HTML-Dokumenten in Text kann also nicht nur durch einfaches Entfernen aller HTML-Tags geschehen. Werden Tabellen und Grafiken einfach weggelassen, stehen plötzlich Textteile, die zuvor durch Tabellenbegrenzungen oder Grafiken getrennt waren, zusammen oder erwecken den Anschein einen „Satz“ zu bilden. Diese Beobachtung wird durch das, an den anderen WDR-Ergebnissen gemessen, schlechte Ergebnis auf der WDR-Testmenge bestätigt. Unter 40 Dokumentfunden bleibt sonst nur das Eigennamen–Strategem.

Das Problem wird dadurch verschärft, dass die Wortkette satzübergreifend gebildet werden darf. Das wird auch daran deutlich, dass diese Variante zum Auffinden von Dokumenten sehr viele Redefinitionsschritte (s. Abb. 3.3) braucht, so dass die erfolgreiche Suchphrase häufig nur noch aus zwei statt zehn Wörtern besteht. Diese beiden Wörter stehen im Originaltext tatsächlich nebeneinander und so wird das Problem umgangen. Aber aus diesem regen Gebrauch der Redefinitionsschritten folgt auch die im Vergleich zu den anderen nicht-zufälligen Strategemen große Abhängigkeit von der Möglichkeit zu spezialisieren. Drei Dokumente werden auf der Universitäts–Testmenge nicht wiedergefunden, wenn der Dokument–Datei–Name nicht genutzt werden kann. Das kommt daher, dass die häufige Generalisierung schlussendlich durch einen Spezialisierungsschritt kompensiert werden muss, um erfolgreich zu sein. Die Dateinamen der betroffenen Dokumente („[...]diplom_fertig.html“ (2x) und „[...]mltxt.eng.html“) sind daher auch recht eindeutig im Gegensatz zu z. B. „index.html“.

gefunden	nicht gefunden
30 (30)	32 (32)

(a) Universität

gefunden	nicht gefunden
54 (54)	8 (8)

(b) WDR

Tabelle 4.16: Phrasensuche mit Sätzen

Die Variante, die sich auf Phrasen *eines* Satzes beschränkt, sollte demnach bessere Ergebnisse liefern. Das Ergebnis in Tabelle 4.16 scheint diese Vermutung jedoch nicht zu stützen. Wenn jedoch die Satzende-Erkennung ausgeweitet wird, so dass Tabellenzellen, Aufzählungszeichen, Bilder und Verweise als Satzenden interpretiert werden, dann verbessern sich die Ergebnisse erheblich (s. Tab. 4.17). Beide Varianten des Satzphrasen-

gefunden	nicht gefunden
49 (49)	13 (13)

(a) Universität

gefunden	nicht gefunden
61 (61)	1 (1)

(b) WDR

Tabelle 4.17: Phrasensuche mit Sätzen — Satzende–Erkennung ausgeweitet

Strategems nutzen die Spezialisierungsoperation gar nicht. Sie sind offenbar nicht von dem Dokument-Dateinamen abhängig.

gefunden	nicht gefunden
41 (41)	21 (21)

(a) Universität

gefunden	nicht gefunden
60 (59)	2 (3)

(b) WDR

Tabelle 4.18: Längste Phrasen

Die Ergebnisse der dritten und einfachste Variante der Phrasensuche siedeln sich zwischen der ersten und der zweiten verbesserten Variante an (s. Tab. 4.18). Unter dem Aspekt der problematischen Satzgrenzen betrachtet, ist das nicht überraschend. Da der längste Satz des Dokuments ausgewählt wird, beschränkt sich die Auswahl der Phrase auf die ersten zehn Wörter *eines* Satzes, so dass das Problem, dass Satzgrenzen nicht richtig erkannt werden, häufig umgangen wird. Es kann zwar vorkommen, dass ein Satz länger „gemacht“ wird, als er eigentlich ist, weil z. B. eine Tabellenzellengrenze nicht beachtet wird, doch dann kann es immer noch sein, dass die ersten zehn Wörter immer noch zu nur einem „echten“ Satz gehören.

4.4.5 Zufällige Auswahl von Suchbegriffen

Die bisher vorgestellten Ergebnisse waren aufgrund der deterministischen Verfahren reproduzierbar und eindeutig. Bei dem hier vorgestellten Strategem ist jedoch der Zufall entscheidend. Aus diesem Grund werden in den Tabellen 4.19 und 4.20 nicht die Ergebnisse eines Laufs, sondern der Erwartungswert der Stichprobe aus fünf Testläufen angegeben. Die Erwartungswerte werden nach der Formel für den erwartungstreuen Mittelwert einer Stichprobe (vgl. [Zei96]) geschätzt:

$$\mu = \frac{1}{n} \sum_{j=1}^{n} X_j \tag{4.4}$$

Dabei ist n die Anzahl der Testläufe insgesamt und X_j das Ergebnis für den Testlauf j.

	gefunden	nicht gefunden
$\boldsymbol{\mu}$	35,6 (35,2)	26,4 (26,8)
min	34 (34)	28 (28)
max	36 (36)	26 (26)

(a) Universität

	gefunden	nicht gefunden
$\boldsymbol{\mu}$	51,2 (48,6)	10,8 (13,4)
min	47 (43)	15 (20)
max	56 (53)	8 (11)

(b) WDR

Tabelle 4.19: Durchschnittlich gefundene Dokumente

	gefunden	nicht gefunden
μ	32,8 (32,0)	29,2 (30,0)
min	31 (30)	31 (32)
max	34 (34)	28 (28)

(a) Universität

	gefunden	nicht gefunden
μ	48,6 (45,6)	13,4 (16,4)
min	47 (43)	15 (19)
max	52 (49)	10 (13)

(b) WDR

Tabelle 4.20: Durchschnittlich gefundene Dokumente mit Stoppwortelimination

	gefunden	nicht gefunden
μ	37,4 (36,8)	24,6 (25,2)
min	34 (33)	28 (29)
max	39 (39)	23 (23)

(a) Universität

	gefunden	nicht gefunden
μ	49,4 (47,2)	12,6 (10,8)
min	47 (43)	15 (19)
max	52 (52)	10 (10)

(b) WDR

Tabelle 4.21: Durchschnittlich gefundene Dokumente bei einer Wahrscheinlichkeitsverteilung, die der Häufigkeitsverteilung entspricht

Die Ergebnisse aller drei Strategemvarianten (vgl. 3.4.1.5) liegen im Bereich der Ergebnisse, der Strategeme, die Suchbegriffe aufgrund der Häufigkeit (s. 3.4.1.3) ihres Vorkommens im Originaldokument auswählen. Sie sind auf der Universitätsmenge schlechter (s. 4.13 u. 4.21(a)) und auf der WDR-Testmenge im Erwartungswert sogar ein wenig besser als die Variante des „Häufige Wörter“–Strategems (s. Tab. 4.13 u. 4.19(b)), das TF-IDF einsetzt. Da die Auswahl der häufigsten Wörter eines Dokuments im Prinzip einen Spezialfall der zufälligen Auswahl darstellen, überraschen die ähnlichen Ergebnisse nicht. Es ist interessant, dass die Stoppwortelimination bei diesen Strategem offenbar die erzielten Ergebnisse verschlechtern. Das liegt daran, dass mit der Hinzunahme der Stoppwörter die Wahrscheinlichkeit steigt, das das ausgewählte Wort in dem neuen Dokument noch vorhanden ist. Das bedeutet in diesem Fall, dass die Hinzunahme eines Stoppworts zwar das Dokument nicht wesentlich besser kennzeichnet, aber auch nicht verhindert, dass die neue Version des Dokuments gefunden wird. Das gewählte Stoppwort wird höchstwahrscheinlich auch in der neuen Version vorkommen — so wie es in vielen anderen Dokumenten auch vorkommt.

Der Vergleich der drei Varianten des Strategems „Zufällige Wörter“ lässt keine eindeutige Reihung zu. Wenn die Ergebnisse anhand der jeweils gefundenen Dokumente angeordnet werden, dann dreht sich die Rangordnung beim Übergang von der Universitäts-Testmenge zur WDR-Testmenge um. Auf der WDR-Testmenge sind die Unterschiede zwischen den drei Varianten jedoch so gering, dass die Schwankungen zufällige Extreme sein können.

Wird das Verhalten der drei Varianten bei Wegfall der Spezialisierungsmöglichkeit betrachtet, fällt auf, dass bei diesem Strategem die größten Schwankungen unter allen suchmaschinenbasierten Strategemen auftreten können. Bis zu vier Dokumente pro Lauf werden nicht mehr gefunden, wenn das Strategem den Dokument–Dateinamen nicht mehr als Suchbegriff verwenden darf. Diese Schwankungsbreite überrascht nicht, da die zufällig ausgewählten Begriffe aus den Dokumenten das gesuchte Dokument nicht so gut spezifizieren kann, wie die guten vorher beschriebenen Strategeme. Aus diesem Grund werden häufiger Suchanfragen gestellt, die zu viele unerwünschte Ergebnisse liefern. Aus diesen

Ergebnissen kann häufig nur durch die Spezialisierung mit dem eher eindeutigen Dateinamen das richtige herausgefiltert werden. Wird die Betrachtung ausgeweitet und auch diejenigen Dokumente gezählt, bei denen eine Spezialisierung versucht und kein positives Ergebnis erreicht wurde, bestätigt sich der Eindruck, dass die zufälligen Strategeme im Vergleich zu den nicht-zufälligen Strategemen, häufiger versuchen zu spezialisieren. Auf der WDR-Testmenge sind sie dabei erfolgreicher, da die Dokument–Dateinamen eindeutiger sind. Daher hat der Wegfall dieser Spezialisierungsmöglichkeit auf dieser Testmenge größere Auswirkungen auf die Anzahl der gefundenen Dokumente (s. Tab. 4.20(b) bis 4.21(b)). Aus den einzelnen Testläufen der Strategeme ist ersichtlich, dass zwei Dokumente nur mit Hilfe eines Spezialisierungsschritts gefunden werden konnten (s. B.1.1 bis B.1.3 Nr. 26 u. Nr. 44 der Universitäts-Testdatenmenge). Eines der beiden Dokumente ist das Dokument `http://www-ai.cs.uni-dortmund.de/index.eng.html`. Es konnte schon vom Strategem „Häufige Wörter“ nicht mehr gefunden werden, wenn die Spezialisierung nicht anwendbar war. Für dieses Dokument scheint die Möglichkeit der Spezialisierung unerlässlich zu sein. Im Unterschied zu Abschnitt 4.4.3 ist in diesem Strategem die Kenntnis des gesuchten Dateinamens wesentlich.

Bei dem zweiten Dokument handelt es sich, um das Dokument `http://ls10-www.cs.uni-dortmund.de/LS10/Pages/sopra-specials/anmeldung/sopra.shtml`. Das wird vom einfachen Phrasen–Strategem jedoch auch ohne Spezialisierungsmaßnahmen gefunden.

Die Ergebnisse der einzelnen Testläufe sind in Anhang B.1.1 bis B.1.3 bzw. B.2.1 bis B.2.3 abgedruckt.

4.4.6 Suche nach Verweisen

Das Problem mit diesem Strategem (vgl. 3.4.1.6) ist, dass es mir auch in der Suchmaske, die Google im WWW zur Verfügung stellt, nicht gelungen ist, eine Anfrage zu formulieren, die Ergebnisse zurück liefert. Das Schlüsselwort „allinlinks:“ scheint nicht wie in [Goo02] beschrieben zu funktionieren. Aus diesem Grund kann dieses Strategem nicht evaluiert werden.

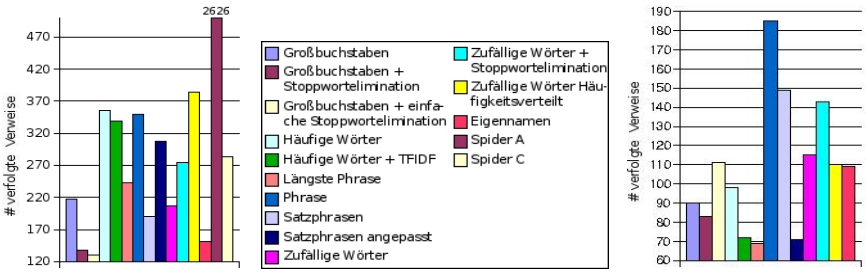
4.4.7 Einfache Breitensuche

Die einfache Breitensuche (vgl. 3.4.2.1) bis zu einer Tiefe von drei Verweisen vom Ursprungsdokument entfernt, findet ca. die Hälfte der Dokumente wieder (s. Tab. 4.22). Aufgrund der Testdatenmenge ist das auch zu erwarten gewesen, da sich die meisten Do-

gefunden	**nicht gefunden**
33	29

Tabelle 4.22: Breitensuche mit Hochschneiden – SpiderA

kumente auf demselben Server befinden. Die Strategie sich entlang der URL nach oben zu schneiden, hat sich ebenfalls bewährt, wenn auch Teilpfade der URL genutzt werden können. Wenn unter diesen Teilpfaden keine Dokumente existieren, die als Startpunkte genutzt werden können, dann ist die Tiefe drei oft nicht ausreichend, um das Dokument wiederzufinden.



(a) Universität (b) WDR

Abbildung 4.1: Von Strategemen verfolgte Verweise

Andererseits werden auch so überdurchschnittlich viele Verweise verfolgt (s. Abb. 4.1) und Dokumente betrachtet, so dass eine Ausweitung der Suche darüber hinaus nicht zu empfehlen ist. Hinzu kommt natürlich das Problem, dass nur Dokumente wiedergefunden werden können, auf die auch ein Verweis innerhalb der untersuchten Dokumente existiert.

gefunden	**nicht gefunden**
47	15

Tabelle 4.23: Breitensuche bis zur Tiefe eins – SpiderC

Die Ergebnisse der Spidervariante, die nur einen Link verfolgt, ausgehend vom fehlerhaften Dokument, sind natürlich besser. Da in den meisten Testdaten, die Dokumente nicht einmal virtuell verschoben wurden, wird das Dokument schon nach dem ersten Verweis wiedergefunden. Die gefundenen Dokumente sind i.d.R. bei diesem Testlauf nicht so interessant wie die nicht gefundenen.

Wurde das Dokument umbenannt und verschoben, dann findet dieser Spider nur das aktuelle Dokument wieder, nicht jedoch das besser passende Dokument, welches vielleicht umbenannt wurde. Zum Beispiel wird für die Lehre-Seite aus dem Jahr 2001 die aktuelle Lehre-Seite als Fundort ausgegeben, statt des besser passenden Dokuments `http://www-ai.cs.uni-dortmund.de/LEHRE/lehre_15_02_2002.html`.

Ein weiteres Problem beim Verschieben ist, dass unter der ursprünglichen URL gar kein Dokument mehr zu finden ist. In diesem Fall kann dieser Spider nichts tun, da ihm die Startpunkte ausgegangen sind (z. B. `www.dfn.de/home.html` ist keine gültige URL mehr). Da der Spider für diesen Fall auch nicht gedacht war, ist es auch nicht überraschend. Bei anderen nicht gefundenen Dokumenten handelt es sich schlicht um temporäre Netzprobleme. So fand der Spider das Dokument `http://www.uni-dortmund.de/UniDo/Personal/` während der regulären Testläufe (s. Anhang B.1) nicht wieder, weil die Zielseite zu diesem Zeitpunkt nicht verfügbar war. Etwas später war jedoch das in Abbildung 4.2 dargestellte Dokument unter der genannten URL verfügbar. Dieses Dokument enthält einen Verweis auf den neuen Ort des gesuchten Dokuments. Durch diesen Verweis ist der SpiderC in der Lage gewesen das Dokument zu finden. Damit erfüllt diese Spidervariante den Zweck für den sie entwickelt wurde und ist daher erfolgreich.

Der Einsatz des SpiderA–Strategem (in Anlehnung an die Aufzählung in 3.4.2.1) ist ins-

Abbildung 4.2: Ein Beispiel für den SpiderC

gesamt als Misserfolg zu werten. Es ist möglich, dass eine einfache Breitensuche genügend viele themenähnliche Dokumente findet, um den Einsatz in einem Agenten, der Dokumente zu einem bestimmten Thema finden soll, zu rechtfertigen (s. [NW01]), doch für die Suche nach bestimmten Dokumenten sind die vorgestellten Suchmaschinenstrategeme erfolgreicher und effizienter bezüglich der zu verfolgenden Verweise. Eine ähnliche Erfahrung wurde auch von C. Bordihn (s. [Bor00] Kap. 2.5.1) beim Erstellen seiner Testmengen gemacht. Dort sollte eine Verweis-Klassifikation vorgenommen werden. Zur Erstellung einer Trainingsmenge für den Klassifikator, sollte von einem gegebenen Startpunkt Verweise zu einem bestimmten Zielpunkt verfolgt werden. Die Suche musste auf Dokumente derselben Domain eingeschränkt werden, um in akzeptabler Zeit Ergebnisse zu erhalten. Die Suche nach dem ersten Dokumentenpaar wurde nach 40 Stunden ergebnislos abgebrochen.

Auf Grund der mangelnden Erfolgsaussichten, denen ein erheblicher Zeitaufwand gegenüber steht, wird diese Spidervariante nicht auf der WDR–Testmenge eingesetzt.

Die SpiderC–Variante ist beschränkter in ihren Möglichkeiten, kann aber trotzdem Erfolge aufweisen. Da das Strategem die „fehlerhafte“ URL als Startpunkt benutzt, findet sie Dokumente bei denen nach einem Verschieben ein Verweis auf den neuen Fundort unter der alten URL hinterlassen wurde. Dieses Strategem sollte also als letzte Alternative in einer Strategie verwendet werden.

Auf der WDR-Testdatenmenge bringt der Einsatz des SpiderC jedoch keine neuen Erkenntnisse, da keins der Dokumente in der beschriebenen Art und Weise verschoben wurde. Der SpiderC wird also alle Dokumente finden, da sie auch nicht virtuell verschoben wurden. Aus diesem Grund wird der SpiderC–Operator nicht auf der WDR-Testdatenmenge evaluiert.

4.4.8 Rückverweise

Dieses Strategem (3.4.2.2) konnte im Rahmen der Diplomarbeit nicht sinnvoll evaluiert werden, da keines der Testdatendokumente während der Betrachtungszeit real verschoben wurde, so dass die Verweise, die zuvor aufgesammelt wurden, alle noch gültig sind und auf dasselbe Dokument verweisen. Die Dokumente, die durch die Testmengendefinition „virtuell“ verschoben wurden, gelten nur für den Agenten als verschoben, so dass kein anderer Agent oder keine andere Person diese Verschiebung zur Kenntnis genommen und Verweise angepasst hat.

4.4.9 Einfluss der Redefinition auf die Strategeme

Generalisierung

Die Generalisierungsoperation ist in ihrer jeweiligen Ausprägung recht erfolgreich, wie Abbildung 3.3 im Zusammenhang mit den Tabellen im Anhang C zeigt. In Abbildung 3.3 ist zu sehen, dass die Anzahl der gefundenen Dokumente bis zu 4 Anfragen ($\hat{=}$ 3 Redefinitionen) stark ansteigt. Die zusammenfassenden Tabellen in Anhang C zeigen, dass der Anteil der Spezialisierungen an diesen Redefinitionen gering ist. Ohne die Möglichkeit einer Generalisierung würden die Ergebnisse also deutlich schlechter ausfallen.

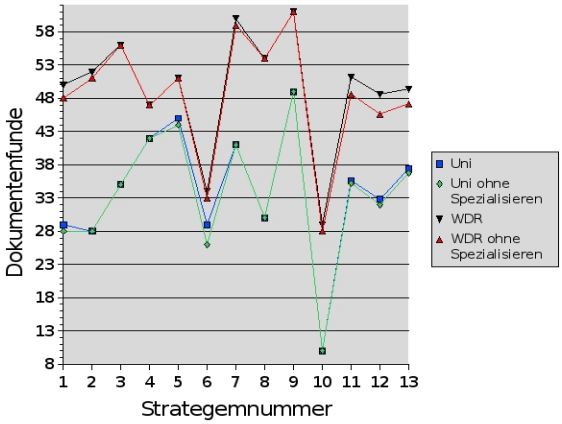
Sind die generierten Anfragen also zu speziell? Diese Frage ist nicht eindeutig zu beantworten, denn offensichtlich sind sie es, sonst würde eine Generalisierung keine Verbesserung erzielen. Andererseits sind Änderungen am Dokument erlaubt, die Wörter aus der bekannten Version des Dokuments nicht mehr erhalten. Da die Strategeme heuristische Annahmen machen, welche Wörter vermutlich nicht aus dem Dokument gestrichen wurden, kann es sein, dass einige der Wörter, die es ermöglichen würden, ein Dokument eindeutig zu identifizieren, gerade aus dem Dokument herausgestrichen wurden. Diese sollten dann auch aus der Anfrage herausgestrichen werden dürfen. Üblicherweise sind bei überspezialisierten Anfragen keine oder nur wenige Dokumente zu untersuchen bis der Agent feststellt, dass das gesuchte Dokument nicht unter den Kandidaten ist. Der Aufwand ist also gering im Vergleich zu übergeneralisierten Anfragen. In diesem Fall müssen mindestens zehn Dokumente gesichtet werden bevor PageTracker entscheidet, dass spezialisiert werden muss.

Spezialisierung

An dieser Stelle soll noch einmal der Einfluss des Spezialisierens auf die Strategeme im Zusammenhang erörtert werden, um herauszufinden, ob

(a) die Spezialisierungsoperation erfolgreich ist, und ob

(b) die Kenntnis des Dokument–Dateinamens wesentlich für den Erfolg der Strategeme ist.



Die Strategemnummern entsprechen den Nummern in Anhang B.

Abbildung 4.3: Dokumentfunde mit und ohne Spezialisieren

Abbildung 4.3 zeigt, dass sich die Ergebnisse mit und ohne Spezialisierungsmöglichkeit nicht wesentlich unterscheiden. Die Unterschiede treten am deutlichsten auf der WDR–Testdatenmenge auf. Dort vor allem bei den zufälligen Strategemen. Wenn die Tabellen der Testläufe in Anhang B.2.1 bis B.2.3 betrachtet werden, zeigt sich, dass zwar die zu erwartenden Funde je Lauf des Strategems geringer werden, aber kein Dokument in der WDR-Testdatenmenge zu finden ist, bei dem die Spezialisierung unbedingt notwendig war, um es überhaupt zu finden. Die Dokumente die mit Hilfe einer Spezialisierungsoperation gefunden wurden, wurden auch alle von einem anderen Lauf derselben Strategemvariante ohne Einsatz einer Spezialisierungsredefinition gefunden. Die Spezialisierung ist also nicht wesentlich für den Erfolg der Strategeme verantwortlich. Das bedeutet, dass die von den Strategemen erzeugten Anfragen i.d.R. so speziell sind, das ein schrittweises Generalisieren ausreicht, um das gesuchte Dokument zu finden. Um den Erfolg des Spezialisierens zu bewerten, werden der Anzahl der Dokumentensuchen mit dem Versuch einer Spezialisierung, die Anzahl der durch diese Spezialisierung gefundenen Dokumente gegenübergestellt. Von den 171 Suchen nach Dokumenten, die mindestens einen Spezialisierungsschritt enthalten, haben nur 60 ($\hat{=}$35%) anschließend das gesuchte Dokument gefunden. Aus den Betrachtungen zu den Strategemen in den vorangegangenen Abschnitten ist bekannt, dass unter diesen 60 erfolgreichen Funden nur zwei Ergebnisse sind, für die die Spezialisierung unbedingt erforderlich ist, weil sie nur mit Hilfe dieser Operation gefunden werden konnte. Insgesamt ist das vorgeschlagene Standardverfahren zur Spezialisierung (s. 3.4) also nicht erfolgreich, da die Ergebnisse i.d.R. auch anders erreicht werden können. In anderen Fällen wird auch durch die vorgenommene Spezialisierung keine Ergebnisverbesserung erzielt. Das bedeutet aber gleichzeitig, dass es möglich ist ohne Kenntnis des Ursprungsorts, nur durch die Kenntnis des Dokumenttextes, Anfragen zu formulieren, die fast alle Dokumente wiederfinden. Ein Starten von einer beliebigen URL aus, ist möglich. Das ist bei den verweisbasierten Strategemen nicht der Fall.

4.5 Strategie–Ergebnisse

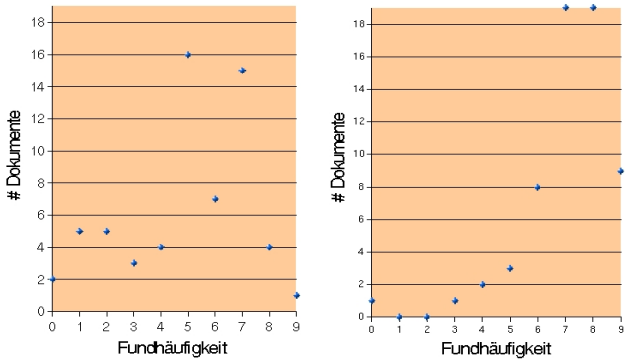
Eine Strategie ist eine Aneinanderreihung von Strategemen. In diesem Abschnitt sollen die verschiedene Strategien evaluiert werden. Dies geschieht, indem nachgeschaut wird, wie erfolgreich eine Zusammenlegung der Einzelergebnisse ist. Die Strategien werden immer für die Universitäts-Testdatenmenge entworfen und dann auf die WDR-Testdatenmenge übertragen. Das geschieht deshalb, weil überprüft werden soll, ob die Ergebnisse aus der Universitätsdatenmenge übertragbar sind.

4.5.1 Strategien ohne Zufallsstrategeme

Zunächst werden nur die nicht-zufälligen suchmaschinenbasierten Strategeme betrachtet.

In Abbildung 4.4 ist zu sehen wie viele Dokumente von wie vielen Strategem-Varianten gefunden werden. Es sind nur die nicht-zufälligen Strategeme und von den beiden Satzphrasen-Varianten nur die verbesserte berücksichtigt.

Wie in Abb. 4.4(a) zu sehen ist, wird jede Strategem-Kombination mindestens zwei Dokumente nicht finden, da keins der nicht-zufälligen suchmaschinenbasierten Strategeme



(a) Universität (b) WDR

Abbildung 4.4: Wie viele Dokumente werden wie oft gefunden?
Verwendetete Strategeme: Großbuchstaben, Häufige Wörter, Phrasensuche, Eigennamensuche

diese Dokumente wiederfindet. Außerdem wird nur ein Dokument von allen Strategemen gefunden. Es handelt sich um das Dokument unter `http://www-ai.cs.uni-dortmund.de/FORSCHUNG/REPORTS/reports.eng.html`.

Bei den WDR–Testdaten (vgl. Abb. 4.4(b)) wird ein Dokument gar nicht und neun Dokumente neunmal wiedergefunden. Das nicht wiedergefundene Dokument ist der WDR-Tagestipp[8]. Dieses Dokument ändert sich täglich und wird von Google nicht schnell genug erfasst. Erschwerend kommt hinzu, dass dieser Tagestipp nur sieben Tage lang als Dokument existiert und dann offenbar nicht archiviert wird.

Insgesamt sind die Dokumente aus der WDR-Testdatenmenge offenbar einfacher wiederzufinden, als die Dokumente aus der Universitätstestdatenmenge. Wie in Abbildung 4.4 zu sehen ist, werden die Dokumente dieser Menge durchschnittlich von mehr Operatoren gefunden. Dieses Ergebnis stimmt mit der Vorhersage der A-Priori-Bewertung (s. S. 48 bzw. Anhang C) überein.

gefunden	**nicht gefunden**
60 (59)	2 (3)

besuchte Verweise: 2231

(a) Universität

gefunden	**nicht gefunden**
61 (61)	1 (1)

besuchte Verweise: 2950

(b) WDR

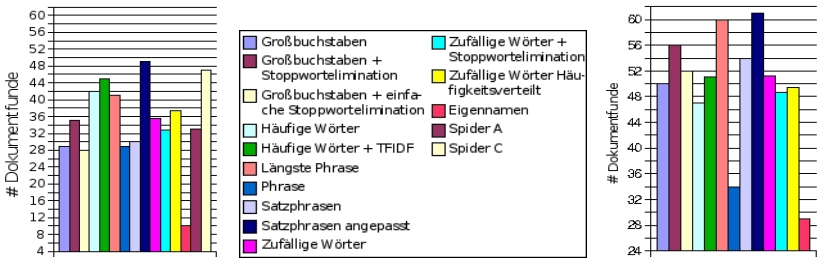
Tabelle 4.24: Ergebnis der Strategie alle nicht-zufälligen Operatoren zu nutzen

Aufgrund der in Abbildung 4.4 gezeigten Darstellung ist das Ergebnis der Strategie alle nicht-zufälligen Operatoren einzusetzen (s. Tab. 4.24) nicht überraschend.

Die Anzahl der untersuchten Verweise bleibt trotz der Vielzahl an Versuchen deutlich unter den 2626 untersuchten Verweisen der Spider–Variante A und findet mehr Dokumente wieder.

Da Abb. 4.4 zeigt, dass viele Dokumente von mehreren Strategemen und deren Varianten wiedergefunden werden, liegt es nahe, nicht alle Strategeme zu nutzen.

[8] `http://www.wdr.de/themen/computer/angeklickt/tagestipp/tagestipp.jhtml`



(a) Universität (b) WDR

Abbildung 4.5: Wie viele Dokumente werden durch welches Strategem gefunden?

Eine nahe liegende Strategie ist, aus jedem beschriebenen Strategem, die beste Variante auszuwählen. Die Strategie besteht aus den folgenden Strategem-Varianten (s. a. Abb. 4.5): `Häufige Wörter unter Berücksichtigung von TF--IDF`, `Großgeschriebene Wörter mit Stoppwortelimination`, `Namen von Personen` und `verbesserte Satzphrasen`.

gefunden	nicht gefunden
59 (58)	3 (2)

besuchte Verweise: 935

(a) Universität

gefunden	nicht gefunden
61 (61)	1 (1)

besuchte Verweise: 1299

(b) WDR

Tabelle 4.25: Ergebnis der Strategie die besten deterministischen Operatoren zu nutzen

Diese Strategie führt zu dem in Tabelle 4.25 gezeigten Ergebnis. Es wird ein Dokument der Universitäts–Testmenge weniger wiedergefunden, aber erheblich weniger Verweise untersucht. Je nach verwendetem Kostenmaß kann diese Strategie günstiger sein. Bei dem zusätzlich nicht gefundenen Dokument handelt es sich um die Leitseite des Fakultätentags Informatik[9]. Dieses Dokument wird nur von den beiden nicht gewählten Varianten des Phrasensuche-Strategems gefunden (s. B.1). Das Strategem, das die Eigennamensuche implementiert, könnte auch noch weggelassen werden; aber dann wird ein weiteres Dokument nicht gefunden. Dieses Strategem findet in dieser Strategie als einziges das Dokument über Bücher am Lehrstuhl 8.

Die WDR-Testdaten sind von der Änderung nicht betroffen, da das Satz-Phrasen-Strategem schon für sich genommen alle 61 überhaupt gefundenen Dokumente wiederfindet. Es werden jedoch sowohl auf der Universitätsdatenmenge als auch auf der WDR-Testdatenmenge, die Anzahl der untersuchten URLs auf weniger, als die Hälfte gegenüber der Strategie, alle nicht-zufälligen Strategeme zu verwenden, reduziert.

Schließlich kann noch die Strategie betrachtet werden, mit der die maximale Anzahl von Dokumenten mit minimalem Operatoreinsatz gefunden wird. Mit einer Strategie, die die folgenden Strategeme einsetzt, wird das erreicht: `Großbuchstaben mit Einsatz einer Stoppwortliste`, `Häufige Wörter mit TF-IDF`, `Längste Phrase` und `verbesserte Satzphrasen`. Das Ergebnis dieser Strategie ist in Tabelle 4.26 zu sehen. An dieser

[9] `http://www.ft-informatik.de`

gefunden	nicht gefunden
60 (59)	2 (3)

besuchte Verweise: 1027

(a) Universität

gefunden	nicht gefunden
61 (61)	1 (1)

besuchte Verweise: 1381

(b) WDR

Tabelle 4.26: So viele Dokumente wie möglich finden, so wenig Strategeme wie nötig nutzen

Strategie fällt auf, dass zwei Varianten des Phrasenstrategems gewählt werden. Der Anteil der „Längste Phrasen"-Strategem–Variante ist dabei recht gering. In dieser Strategie ist sie jedoch die einzige Chance die Dokumente `Sopra`[10] und `Bücher`[11] zu finden. Noch geringer ist der notwendige Anteil, der durch das Strategem „Häufige Wörter" geleistet wird. Dieses Strategem wird nur zur Wiederauffindung der englischen Startseite[12] des Lehrstuhls für Künstliche Intelligenz benötigt. Da die Strategie die Strategemvariante „Satzphrase" enthält, wird auch die maximal mögliche Anzahl der WDR–Dokumente wiedergefunden.

Bei allen bisher untersuchten Strategien spielt die Frage, ob ein Spezialisieren erlaubt ist und damit verbunden die Fragestellung, ob der Dokument–Dateiname sich geändert hat oder nicht, eine untergeordnete Rolle. Es unterscheiden sich nur die Ergebnisse auf der Universitätsdatenmenge. Diese unterscheiden sich nur in einem Dokument, welches mit Spezialisierungsoption nur vom Strategem Häufige Wörter wiedergefunden wird (s. 4.4.3).

4.5.2 Strategien mit Zufallsstrategemen

Wie in Abschnitt 4.4.5 schon festgestellt wurde, ist das Strategem der zufälligen Wortauswahl recht erfolgreich gewesen. Könnte es sein, dass eine Aneinanderreihung von mehreren Operatorläufen dieses Strategems zu ähnlich guten Ergebnissen führt wie die der nicht–zufälligen Strategie? Statt der Durchschnittswerte wie sie in 4.4.5 genutzt wurden, werden die fünf Testläufe nun als eine Strategie aufgefasst.

gefunden	nicht gefunden
52 (52)	10 (10)

besuchte Verweise: 1032

(a) Universität

gefunden	nicht gefunden
60 (60)	2 (2)

besuchte Verweise: 1850

(b) WDR

Tabelle 4.27: Zufällige Wörter als Strategie

Die Ergebnisse der verschiedenen Strategien, wenn die Strategeme zunächst voneinander isoliert betrachtet werden, d.h. die Strategien bestehen ausschließlich aus Aneinanderreihungen eines Zufallsstrategems, sind in den Tabellen 4.27, 4.28 und 4.29 zu sehen. Die Ergebnisse auf der WDR-Testdatenmenge sind ähnlich gut, wie die Ergebnisse der nichtzufälligen Strategien. Die Ergebnisse in 4.28(b) und 4.28(a) finden sogar genauso viele

[10] `http://ls10-www.cs.uni-dortmund.de/LS10/Pages/sopra.shtml`
[11] `http://www-ai.cs.uni-dortmund.de/FORSCHUNG/buecher.html`
[12] `http://www-ai.cs.uni-dortmund.de/index.eng.html`

gefunden	nicht gefunden
49 (48)	13 (14)

besuchte Verweise: 1373

(a) Universität

gefunden	nicht gefunden
61 (60)	1 (2)

besuchte Verweise: 2073

(b) WDR

Tabelle 4.28: Zufällige Wörter mit Stoppwortelimination als Strategie

Dokumente wie die erfolgreichsten Strategien in 4.5.1, aber die Anzahl der untersuchten URLs und der notwendigen Strategemläufe ist höher.

gefunden	nicht gefunden
55 (53)	7 (9)

besuchte Verweise: 1921

(a) Universität

gefunden	nicht gefunden
61 (61)	1 (1)

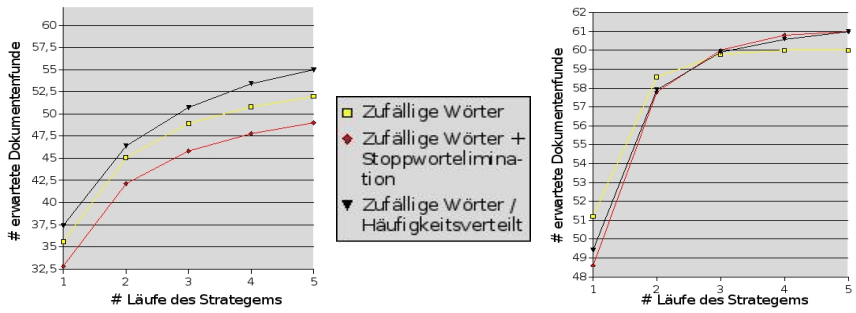
besuchte Verweise: 2522

(b) WDR

Tabelle 4.29: Zufällige Wörter mit Häufigkeitsverteilung als Strategie

Die Ergebnisse der zufälligen Strategien auf der Universitätsdatenmenge sind durchgehend schlechter als die der nicht–zufälligen Strategien. Sie finden weniger Dokumente und untersuchen mehr URLs als die nicht–zufälligen Strategie mit Ausnahme der Strategie alle nicht-zufälligen Strategeme einzusetzen.

Der Einfluss den die Redefinition durch Spezialisierung auf einzelne Testläufe hatte (s. 4.4.5), verliert sich offenbar in der Strategiebildung. Die einzige Ausnahme bildet die Suche nach dem Dokument „[...]popkomm2003.jhtml“ der WDR-Testdaten für die Strategie nur Strategeme vom Typ “Zufällige Wörter mit Stoppwortelimination“ einzusetzen. Dieses Dokument wird nicht gefunden, wenn Spezialisierungen nicht durchgeführt werden können. Alle anderen Dokumente werden auch ohne den Einsatz von Spezialisierungsschritten gefunden.



(a) Universität

(b) Universität

Abbildung 4.6: Erwartete Dokumentfunde bei Einsatz verschiedener Anzahlen von Operatoren

Abbildung 4.6 zeigt die Entwicklung der durchschnittlichen Strategem-Auswahlen. Da die Strategeme zufallsabhängig sind, werden nicht einfach die fünf Läufe aneinander gehängt, sondern es wird der Durchschnitt aus allen $\binom{5}{n}$ möglichen Kombinationen für Strategien mit n-Strategem-Wiederholungen berechnet. Wie zu sehen ist, werden für

die Universitätsdatenmenge alle Strategem-Läufe gebraucht, um das beste Ergebnis zu erreichen. Natürlich kann es in anderen Läufen auch passieren, dass schon der erste Strategemlauf alle Dokumente wiederfindet, aber das scheint nicht wahrscheinlich zu sein.

Der ausschließliche Einsatz von zufälligen Strategemen ist also nicht empfehlenswert, da sie nicht mehr Dokumente finden als die nicht-zufälligen und dabei mehr URLs untersuchen müssen.

5 Zusammenfassung und Ausblick

Im Rahmen der vorliegenden Arbeit wurde der Agent PageTracker geschaffen. Dieser Agent ist nach den Kriterien aus 2.3 eine Meta-Suchmaschine 3. Ordnung, da er eine andere Suchmaschine erster Ordnung befragt und deren Ergebnisse noch einmal mit einem eigenen Relevanzfilter untersucht.

Die vorliegende Arbeit weist nach, dass Suchmaschinenanfragen ein geeignetes Mittel sind, um nach Dokumenten, deren Ablageort sich verändert hat, zu suchen. Offenbar strukturieren Suchmaschinen wie Google das Internet durch die von ihnen eingesetzten Rankingverfahren so gut, dass Suchanfragen mit höchstens zehn Suchbegriffen ausreichen, um das gesuchte Dokument in den ersten zehn Rängen zu platzieren. Für die Lösung des Problems die zehn Wörter zu wählen, die es ermöglichen das Dokument wieder zu finden, konnten im Rahmen dieser Arbeit verschiedene erfolgreiche Verfahren angegeben und evaluiert werden. Das erfolgreichste Strategem ist eine spezialisierte Phrasensuche (s. 4.4.4). Zur Auswahl einer geeigneten Phrase wird ein Verfahren verwendet, das in [Eul01] zur Textzusammenfassung von Emails benutzt wurde. Es sollte in der dort beschriebenen Anwendung hochrelevante Sätze von anderen, nicht so relevanten Sätzen trennen. Hier wird dieses Verfahren genutzt, um eine Phrase auszuwählen, die für das vorliegende Dokument besonders relevant ist. Hinter dieser Idee steckt die Hoffnung, dass der gewählte Satz auch in einer Folgeversion unverändert vorkommen wird, da Kernaussagen in Dokumenten selten geändert werden.

Mit dem Erfolg der Phrasensuche eng verknüpft ist die Erkenntnis, dass die Repräsentation der Texte im Agenten derart sein sollte, dass zumindestens dieselben Satztrenner erkannt werden. Die verwendete Suchmaschine behauptet sonst, dass sie zur gestellten Suchanfrage nichts wisse. Das ist besonders problematisch, da z. B. das HTML-Tag <br> zur Erzeugung eines Zeilenumbruchs normalerweise nicht als Satztrenner betrachtet werden würde. Werden jedoch Phrasen über das Tag hinaus gebildet, findet Google kein Dokument, obwohl die Zeichenkette in einem Browser zusammenhängend erscheint.

Je nachdem welches Kostenmaß der Nutzer der Agenten verwendet, sollten jedoch auch die anderen Strategeme zum Einsatz kommen, denn die Satz-Phrasensuche findet nicht alle Dokumente. Wem die Anzahl der untersuchten Dokumente weniger wichtig ist, wenn die Chance auf eine höhere Dokumentenfund–Anzahl steigt, der sollte aus den anderen Strategemklassen auch jeweils eine Variante auswählen. Schließlich hat auch der SpiderC (s. 4.4.7) seine Einsatzfelder, wenn die Suchmaschinendatenbank den neuen Fundort eines Dokuments noch nicht erfasst hat, aber an der alten Stelle ein Verweis auf den neuen Ort existiert. Da der Agent Dokumente, die er sicher gefunden zu haben glaubt, nicht noch einmal mit anderen Strategemen sucht, sollten Strategeme, die erfahrungsgemäß viele Verweise abarbeiten müssen, um Erfolge zu erzielen, erst spät innerhalb der Strategie angeordnet werden.

Damit sind die in Kapitel 3 aufgeworfenen Fragen beantwortet:

Ist es möglich, ein zu suchendes Dokument so auf zehn Wörter zu reduzieren, dass es unter den ersten zehn Suchergebnissen von Google auftaucht?
Das Satzphrasen–Strategem löst diese Aufgabe für sich genommen schon recht gut. Zusammen mit anderen Strategemen sind die Ergebnisse sogar sehr gut.

Welche Wörter aus dem Dokument sind die richtigen?
Stoppwortelimination mit Hilfe von Stoppwortlisten und die Berücksichtigung der TF–IDF-Werte zur Auswahl der Suchbegriffe, verbessert die Ergebnisse sichtlich. Der für die TF–IDF–Berechnung wichtige IDF–Wert kann aus der Menge der zu überwachenden Dokumente gewonnen werden. Das kann automatisiert werden und stellt für den Nutzer des Agenten keinen zusätzlichen Aufwand dar. Für verschiedene Sprachen sind Stoppwortlisten im Internet verfügbar. In den Testläufen wurden diese Listen mit guten Ergebnissen einfach aneinandergehängt, um multilingual arbeiten zu können.

Im Zusammenhang mit dem in Kapitel 1.2 erwähnten Trade-off zwischen Sucherfolg und Installationsaufwand z. B. zur Erstellung oder Eingabe von Sachbereichswissen, ist bemerkenswert, dass der Agent nach Eingabe der zu überwachenden Dokumente, keine weiteren Informationen vom Benutzer über den Sachbereich benötigte und trotzdem gute Ergebnisse liefert.

Wie kann eine Suchanfrage so geändert werden, dass die nächste Anfrage erfolgreicher ist?
Diese Frage kann nicht pauschal beantwortet werden. Das Standardvorgehen, die Suchanfrage zu generalisieren, indem Begriffe aus der Anfrage herausgestrichen werden, wurde für die wortbasierten Strategeme erfolgreich eingesetzt. Das Generalisierungsverfahren blieb damit der Auswahlfunktion treu, da keine Wörter eingesetzt wurden, die in der Bewertung der Suchanfragenauswahl schlechter abgeschnitten hätten. Für die Suche nach Satz-Phrasen war jedoch das Austauschen der ganzen Phrase erfolgreicher. Damit blieb das Redefinitionsverfahren der grundlegenden Idee dieses Strategems treu, möglichst nach ganzen „Sätzen“ zu fragen.

Beide Änderungsansätze waren erfolgreich, denn beide Redefinitionsverfahren führten zu einer Verbesserung gegenüber dem Verhalten ohne Redefinition. Es ist zu vermuten, dass das Redefinitionsverfahren stets an das Suchverfahren angepasst werden sollte.

Das Vorgehen bei der Spezialisierung brachte zwar auch Verbesserungen der Suchergebnisse mit sich, doch in 65% der Fälle hat die Redefinition keine Änderung am Erfolg der Suche herbeigeführt. Andererseits wurde diese Redefinitionsvariante auch nur in 170 der 3100 ($\hat{=}$ 5,5%) untersuchten Dokument-Strategem-Kombinationen vom Agenten eingesetzt. Nur bei einem Dokument (<1% aller Dokumente) war die Spezialisierung erfolgsentscheidend.

In den Kapiteln 2.1.2 und 2.1.3 wurde ausgeführt, dass der Durchmesser des WWW sich zwischen 16 und 19 bewegt. Das entspricht der durchschnittlichen Anzahl von Verweisen, denen ein Spider folgen müsste, um von einem Dokument zu einem anderen zu kommen. In [BKM+00] (s. Tab. 2.1) wurde weiter ausgeführt, dass der Durchmesser

noch einmal deutlicher kleiner wäre (≈ 6), wenn die Verweise zwischen Dokumenten bi-direktional wären. Das ist im Allgemeinen jedoch nicht der Fall. Durch die Nutzung von Suchmaschinen kann aber ein ähnliches Ergebnis erreicht werden. Statt echte Verweise zu verfolgen, die direkt zwischen Start- und Ziel existieren, kann durch die Vorarbeit der Suchmaschinenbetreiber, der erste Verweis ein Verweis auf die Suchmaschine sein, welcher die richtige Anfrage enthält (s. Abb. 5.1). Um diesen Verweis zu generieren reicht

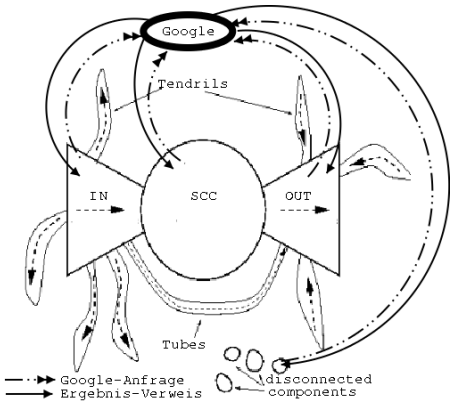


Abbildung 5.1: Die Web–Fliege in den Fängen einer Suchmaschine (in Anlehnung an [BKM+00])

die Kenntnis des Inhalts des Zieldokuments, wie die vorliegende Arbeit zeigt. Angenommen, die mit dem Verweis verbundene Suchanfrage liefert das Zieldokument, dann müssen nach diesem ersten Verweis höchstens 10 weitere Verweise verfolgt werden. Das entspricht einem Abstand von 11 Schritte zwischen Start und Ziel. Tatsächlich brauchen die getesteten Strategeme im Mittel nur 6,00 (6,39)[1] Verweise verfolgen. Diese durchschnittlich 6,00 (6,39)[1] Schritte enthalten alle Anfragen, die gestellt werden mussten, um das Dokument zu finden, falls es überhaupt gefunden wurde. Das entspricht in etwa den sechs Schritten, die in [BKM+00] als mittlere verbundene Distanz zwischen zwei beliebigen Dokumenten angegeben wurde, wenn alle existierenden Verweise bi-direktional wären.

Die in Kapitel 3 getroffene Entscheidung, Suchmaschinen zur Suche nach Dokumenten einzusetzen, war also richtig, da sie ein erhebliches Potenzial zur Verkürzung der mittleren Weglänge von einem Dokument zu einem anderen bieten, wenn aus dem Startdokument die Abkürzung zum Zieldokument errechnet werden kann. PageTracker kann diese Abkürzungen berechnen. Ein rein auf Spidern basierter Agent würde nach [BKM+00] im Mittel etwas mehr als 16 Verweise abarbeiten müssen, *wenn* in jedem Schritt der *richtige* Verweis gewählt wird. Die suchmaschinenbasierten Strategeme bleiben im Mittel alle deutlich unter diesem Wert. Die im SpiderA eingesetzte einfache Breitensuche hingegen benötigt für die gefundenen Dokumente im Schnitt 54 Schritte.

Eine interessante Frage ist, ob die Suchergebnisse mit steigender Anzahl von zu überwachenden Dokumenten weiter verbessert werden können, wie es die Verwendung von TF–IDF vielleicht erwarten lässt, weil die IDF–Werte besser abgeschätzt werden. Dazu sollte jedoch das verwendete Distanzmaß verbessert werden, damit die Testläufe vollautomatisch ablaufen können. Das eingesetzte Cosinusmaß ist zwar schon recht gut ge-

[1] Die in Klammern angegebenen Zahlen sind die Ergebnisse auf WDR-Testdatenmenge.

eignet, aber je mehr Änderungen am Dokument vorgenommen wurden, desto größer ist die Wahrscheinlichkeit, dass nach diesem Maß ein anderes Dokument als das erwartete einen kleineren Abstand bezüglich dieses Maßes hat.

Offen bleibt, ob die Ergebnisse dieser Arbeit auf andere Suchmaschinen, z. B. Altavista, übertragbar sind. Dazu müssten die Anfragen an die entsprechende Syntax angepasst werden. Die Abhängigkeit des Agenten von der Güte des Ranking–Algorithmus' der verwendeten Suchmaschine könnte dadurch überprüft werden. Das Beispiel der DFN-Homepage (s. 4.3) zeigt, dass es diese Abhängigkeit zumindestens bei kleinen Dokumenten gibt.

Literaturverzeichnis

[AFJM95] Armstrong, Robert, Dayne Freitag, Thorsten Joachims und Tom Mitchell: *WebWatcher: A Learning Apprentice for the World Wide Web.* In: *AAAI Spring Symposium on Information Gathering*, Seiten 6–12, 1995.

[AI01] Archive, Internet und Alexa Internet: *Wayback Machine.* `http://www.archive.org/`, Oktober 2001.

[AJB99] Albert, Réka, Hawoong Jeong und Albert-László Barabási: *Diameter oft the World-Wide Web.* Nature, 401, September 1999. www.nature.com.

[BCHR01] Bharat, Krishna, Bay-Wei Chang, Monika Henzinger und Matthias Ruhl: *Who Links to Whom: Mining Linkage between Web Sites.* In: *IEEE International Conference on Data Mining (ICDM '01)*, November 2001.

[BEK+] Box, Don, David Ehnebuske, Gopal Kakivaya, Andrew Layman, Noah Mendelsohn, Henrik Frystik Nielsen, Satish Thatte und Dave Winer: *Simple Object Access Protocol (SOAP) 1.1.* `http://www.w3.org/TR/SOAP/`.

[BF95] Blum, Avrim und Merrick Furst: *Fast Planning Through Planning Graph Analysis.* In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI 95)*, Seiten 1636–1642, 1995.

[BFG+00] Banken, Michael, Christian Fischbach, Oliver Geppert, Markus Hövener, Jens Jägersküpper, Volkher Kaschlun, Nils Malzahn, André Masloch, Ursula Mentel, Marina Podvoiskaia, Niels Schröter, Thorsten Joachims und Ralf Klinkenberg: *BotIshelly – Bibliothek zur Erstellung von Agenten für die Suche im Internet.* `http://www-ai.cs.uni-dortmund.de/LEHRE/PG/PG343`, Februar 2000.

[BKM+00] Broder, Andrei, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins und Janet Wiener: *Graph structure in the web.* In: *9th WWW Conference*, 2000.

[BNAP] Barfourosh, A. Abdollahzadeh, H.R. Motahary Nezhad, M. L. Anderson und D. Perlis: *Information Retrieval on the World Wide Web and Active Logic: A Survey and Problem Definition.*

[Bor00] Bordihn, Christian: *Ein lernender Agent zur kooperativen Erstellung und Verwaltung von themenorientierten Dokumentdatenbanken aus dem WWW.* Diplomarbeit, Fachbereich Informatik, Universität Dortmund, 2000.

[BP98] Brin, Sergey und Lawrence Page: *The anatomy of a large-scale hypertextual Web search engine*. Computer Networks and ISDN Systems, 30(1–7):107–117, 1998.

[BS85] Brachman, Ronald J. und James G. Schmolze: *An overview of the KL-One knowledge representation system*. Cognitive Science, 9(2):171–216, April–June 1985.

[CDI98] Chakrabarti, Soumen, Byron E. Dom und Piotr Indyk: *Enhanced hypertext categorization using hyperlinks*. In: Haas, Laura M. und Ashutosh Tiwary (Herausgeber): *Proceedings of SIGMOD-98, ACM International Conference on Management of Data*, Seiten 307–318, Seattle, US, 1998. ACM Press, New York, US.

[CDK+99] Chakrabarti, Soumen, Byron E. Dom, S. Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, Andrew Tomkins, David Gibson und Jon Kleinberg: *Mining the Web's Link Structure*. Computer, 32(8):60–67, 1999.

[CGMP98] Cho, Junghoo, Hector García-Molina und Lawrence Page: *Efficient crawling through URL ordering*. Computer Networks and ISDN Systems, 30(1–7):161–172, 1998.

[CHH98] Chang, Chia-Hui, Ching-Chi Hsu und Cheng-Lin Hou: *Exploiting Hyperlinks for Automatic Information Discovery on the WWW*. In: *In the Proceedings of the tenth IEEE International Conference on Tools with Artificial Intellligence*, Chien Tan Youth Activity Center, Taipei, Taiwan, November 1998.

[DCL+00] Diligenti, Michelangelo, Frans Coetzee, Steve Lawrence, C. Lee Giles und Marco Gori: *Focused Crawling using Context Graphs*. In: *26th International Conference on Very Large Databases, VLDB 2000*, Seiten 527–534, Cairo, Egypt, 10–14 September 2000.

[DP94] De Bra, P. M. E. und R. D. J. Post: *Information retrieval in the World-Wide Web: Making client-based searching feasible*. Computer Networks and ISDN Systems, 27(2):183–192, 1994.

[Eul01] Euler, Timm: *Informationsextraktion durch Zusammenfassung maschinell selektierter Textsegmente*. Diplomarbeit, Fachbereich Informatik, Universität Dortmund, 2001.

[FFF99] Faloutsos, Michalis, Petros Faloutsos und Christos Faloutsos: *On Power-law Relationships of the Internet Topology*. In: *SIGCOMM*, Seiten 251–262, 1999.

[FGM+99] Fielding, R., J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach und T. Berners-Lee: *RFC 2616: Hypertext Transfer Protocol – HTTP/1.1*, June 1999. `ftp://ftp.isi.edu/in-notes/rfc2616.txt`.

[Fuh00] Fuhr, Norbert: *Information Retrieval — Skriptum zur Vorlesung im WS 00/01*. `http://www.is.informatik.uni-duisburg.de/teaching/dortmund/lectures/ir_ws00-01/irskall.pdf`, Oktober 2000.

[Goo02] *Google Web APIs Reference*. `http://www.google.de/apis/reference.html`, 2002.

[Goo03] Google: *Google Web APIs*. `http://www.google.com/apis/`, 2003.

[Gri95] Grishman, Ralph: *The NYU System for MUC-6 or Where's the Syntax?* In: *Proceedings of the MUC-6 workshop*, Washington, November 1995.

[GWH+95] Gaizauskas, R., T. Wakao, K. Humphreys, H. Cunningham und Y. Wilks: *University of sheffield: Description of the lasie system as used for muc*, 1995.

[Hav02] Haveliwala, T.: *Topic-sensitive PageRank*. In: *Proceedings of the Eleventh International World Wide Web Conference*, Honolulu, Hawaii, May 2002.

[Hay00a] Hayes, Brian: *Graph Theory in Practice Part I.* American Scientist, 88(1):9–13, January-February 2000.

[Hay00b] Hayes, Brian: *Graph Theory in Practice Part II.* American Scientist, 88(2):104–109, March-April 2000.

[HJM+98] Hersovici, Michael, Michael Jacovi, Yoelle S. Maarek, Dan Pelleg, Menachem Shtalhaim und Sigalit Ur: *The Shark-Search Algorithm - An Application: Tailored Web Site Mapping*. In: *Proceedings of the 7th International World Wide Web Conferenc*, 1998.

[HN99] Heydon, Allan und Marc Najork: *Mercator: A Scalable, Extensible Web Crawler*. World Wide Web, 2(4):219–229, 1999.

[Hua] Huang, Lan: *A Survey On Web Information Retrieval Technologies*.

[JFM97] Joachims, Thorsten, Dayne Freitag und Tom M. Mitchell: *Web Watcher: A Tour Guide for the World Wide Web*. In: *IJCAI (1)*, Seiten 770–777, 1997.

[KHG03] Kamvar, Sepandar D., Taher H. Haveliwala und Gene H. Golub: *Adaptive Methods for the Computation of PageRank*. In: *Stanford University Technical Report*, 2003.

[KHMG03] Kamvar, Sepandar D., Taher H. Haveliwala, Christopher D. Manning und Gene H. Golub: *Exploiting the Block Structure of the Web for Computing PageRank*. In: *Stanford University Technical Report*, 2003.

[Kle99] Kleinberg, Jon M.: *Authoritative sources in a hyperlinked environment*. Journal of the ACM, 46(5):604–632, 1999.

[KRR+00] Kumar, Ravi, Prabhakar Raghavan, Sridhar Rajagopalan, D. Sivakumar, Andrew Tomkins und Eli Upfal: *The Web as a Graph*. In: *Proc. 19th ACM SIGACT-SIGMOD-AIGART Symp. Principles of Database Systems, PODS*, Seiten 1–10. ACM Press, 15–17 2000.

[Lab01] Lab, HP SRC Classic: *Home Page of the Mercator Web Crawler*. `http://research.compaq.com/SRC/mercator/`, 3 2001.

[Lev65] Levenstein, Vladimir I.: *Binary codes capable of correcting spurious insertions and deletions of ones.* Problems of Information Transmission, 1:8–17, 1965.

[Loo00] *LookOff Engine Ebook.* `http://www.lookoff.com/tactics/`, 10 2000.

[Mas03] Masloch, André: *Ein intelligenter URL-Checker.* Diplomarbeit, Fachbereich Informatik, Universität Dortmund, 2003. noch zu veröffentlichen.

[MB00] Menczer, Filippo und Richard K. Belew: *Adaptive retrieval agents: Internalizing local context and scaling up to the web.* Machine Learning, 39(2):203–242, May 2000.

[Men97] Menczer, Filippo: *ARACHNID: Adaptive Retrieval Agents Choosing Heuristic Neighborhoods for Information Discovery.* In: *Machine Learning: Proceedings of the Fourteenth International Conference*, Seiten 227–235, 1997.

[MIN01] *MIND - Resource Selection and Data Fusion for Multimedia International Digital Libraries.* http://www.mind-project.org/intro.html, 2001.

[Mit97] Mitchell, Tom M.: *Machine Learning*, Seiten 231–236. McGRAW-HILL, 1997.

[MNRS00] McCallum, Andrew K., Kamal Nigam, Jason Rennie und Kristie Seymore: *Automating the Construction of Internet Portals with Machine Learning.* Information Retrieval, 3(2):127–163, 2000.

[Net99] Netscape: *Über das Open Directory Project.* `http://dmoz.org/World/Deutsch/about.html`, 1999.

[NH01] Najork, Marc und Allan Heydon: *On High-Performance Web Crawling.* Technischer Bericht 173, Compaq Systems Research Center, 130 Lytton Avenue, Palo Alto California 94301, September 2001.

[NW01] Najork, Marc und Janet L. Wiener: *Breadth-First Crawling Yields High-Quality Pages.* In: *Proceedings of the 10th International World Wide Web Conference*, Seiten 114–118, Hong Kong, May 2001. Elsevier Science.

[PBMW98] Page, Lawrence, Sergey Brin, Rajeev Motwani und Terry Winograd: *The PageRank Citation Ranking: Bringing Order to the Web.* Technischer Bericht, Stanford Digital Library Technologies Project, 1998.

[Rah02] Rahm, Prof. Dr. E.: *Algorithmen und Datenstrukturen 2.* `http://dbs.uni-leipzig.de/en/skripte/ADS2/HTML/kap4-17.html`, Sommersemester 2002. Universität Leipzig, Institut für Informatik.

[RM99] Rennie, Jason und Andrew Kachites McCallum: *Using reinforcement learning to spider the Web efficiently.* In: Bratko, Ivan und Saso Dzeroski (Herausgeber): *Proceedings of ICML-99, 16th International Conference on Machine Learning*, Seiten 335–343, Bled, SL, 1999. Morgan Kaufmann Publishers, San Francisco, US.

[Sal89] Salton, Gerard: *Automatic text processing — the transformation, analysis, and retrieval of information by computer.* Addison-Wesley series in computer science. Addison-Wesley, 1989.

[SB98] Sander-Beuermann, Wolfgang: *Schatzsucher — Die Internet–Suchmaschinen der Zukunft.* c't Verlag Heinz Heise, (13):178–185, 1998.

[SBS98] Sander-Beuermann, Wolfgang und Mario Schomburg: *Internet Information Retrieval - The Further Development of Meta-Searchengine Technology.* In: Sadowsky, George und Mark Selby (Herausgeber): *Internet Summit*, Genf, 7 1998. Internet Society, `http://www.isoc.org/inet98/proceedings/`.

[SM83] Salton, Gerard und Michael J. McGill: *Introduction to modern information retrieval.* McGraw–Hill computer science series. McGraw–Hill, New York, 1983.

[WD92] Watkins, Christopher und Peter Dayan: *Q-Learning.* Machine Learning, 8:279–292, 1992.

[Zei96] Zeidler, E. (Herausgeber): *Teubner Taschenbuch der Mathematik*, Band 1, Kapitel 6.3.2, Seite 1062. B.G. Teubner Stuttgart, 1996.

Anhang A

Testmengen

Die Testmengen, die zur Evaluation genutzt wurden, sind hier abgedruckt. Die Daten sind in Form einer XML-Datei angegeben, wie sie auch genutzt wird, um die Daten dem Agenten zur Verfügung zu stellen. Jeder Eintrag für ein zu suchendes Dokument wird in `<urlentity> </urlentity>`-Paare geschachtelt. Innerhalb dieser Paare können beliebig viele `<url></url>` Paare auftauchen. Die Reihenfolge der `<url></url>`-Einträge ist entscheidend. Von oben nach unten interpretiert der Agent die URLs als zunehmend jünger. Das bedeutet, dass der letzte Eintrag auf jeden Fall als fehlerhafte URL behandelt wird und der vorhergegangene den alten Inhalt der fehlerhaften URL darstellt. Sind mehr als zwei `<url></url>`-Einträge vorhanden, wird der jeweilige Vorgänger als alter Inhalt des Nachfolgers betrachtet. Im folgenden Beispiel:

```
<urlentity name="53">
  <url>
  http://www.cs.helsinki.fi/kurssit/
 </url>
 <url>
  http://www.rni.helsinki.fi/~htt/
 </url>
</urlentity>
```

wird zum Beispiel die URL `http://www.rni.helsinki.fi/~htt/` als fehlerhaft deklariert und der Agent wird nach einem Dokument suchen, welches den Inhalt, des Dokuments, das unter `http://www.cs.helsinki.fi/kurssit/` zu finden ist, suchen.

Das Attribut `name` im Tag `<urltentity>` ist optional und dient nur einer vereinfachten Identifikation im Programmablauf. In diesem Dokument dienen sie zur Kennzeichnung der Testdatensätze durch Nummern.

Zusätzlich ist es möglich, dem Agenten vorzutäuschen, dass unter einer URL ein anderer Inhalt zu finden ist, als es tatsächlich der Fall ist. Bei Angabe der Daten

```
<urlentity name="03">
 <url content="http://web.archive.org/web/20011225002800/http://www-ai.cs.uni-dortmund.de/UNIVERSELL/">
  http://www-ai.cs.uni-dortmund.de/UNIVERSELL/
 </url>
 <url>
  http://www-ai.cs.uni-dortmund.de/UNIVERSELL/
 </url>
```

wird der Agent ein Dokument suchen, welches den Inhalt des Dokuments unter `http://web.archive.org/web/20011225002800/[...]/UNIVERSELL/` trägt, dessen alter Ort `http://www-ai.cs.uni-dortmund.de/UNIVERSELL/` ist und dessen fehlerhafte URL `http://www-ai.cs.uni-dortmund.de/UNIVERSELL/` ist. Diese Option wäre z. B. für

ein Spiderstrategem wichtig, welches die alte statt der fehlerhaften URL als Startpunkt nimmt. Das Attribut `content` ist optional und dient vor allem zur Generierung von Testfällen. In realen Kontexten wird die Option vermutlich nicht sinnvoll zum Einsatz kommen.

Für die suchmaschinenbasierten Operatoren verhält sich das Konstrukt genauso wie das folgende:

```
<urlentity name="03">
 <url>
  http://web.archive.org/web/20011225002800/http://www-ai.cs.uni-dortmund.de/UNIVERSELL/
 </url>
 <url>
  http://www-ai.cs.uni-dortmund.de/UNIVERSELL/
 </url>
```

A.1 Universitätsseiten

Die in dieser Testmenge enthaltenen Testdaten sind durch drei Vorgänge entstanden:

1. Mit einem Spider wurden ausgehend von `http://www-ai.cs.uni-dortmund.de` bis zu einer Tiefe von zwei, verschiedene URLs aufgesammelt. Anschließend wurden alte Versionen dieser Dokumente im Archiv der Wayback-Machine aus dem Jahr 2001 gesucht. Diese alten Versionen der Dokumente werden als alte Inhalte genutzt.
2. Einige URLs zu Dokumenten von Personen, die am ML-Net teilnehmen, wurden ausgewählt. Als alte Inhalte wurden Dokumente ausgewählt, die einen Bezug zur Lehre haben und in derselben Domain liegen wie das fehlerhafte Dokument.
3. Aktuelle Dokumente aus den Webdokumenten des Lehrstuhls für künstliche Intelligenz werden ausgewählt. Durch paarweises vertauschen von Inhalt und Ort zweier Dokumente, werden Dokumente virtuell verschoben.

Die Testdaten, die wie in 1. beschrieben erstellt wurden, testen den Umgang des Agenten mit veränderten Dokumenten. Die Testdaten unter 2 und 3 prüfen, wie sich veränderte Orte auf die Suchergebnisse auswirken. Sie sind vor allem als Testdaten für die Spider gedacht.

```
<urlentity name="01">
 <url content="http://web.archive.org/web/20011130192817/http://www-ai.cs.uni-dortmund.de/">
  http://www-ai.cs.uni-dortmund.de
 </url>
 <url>
  http://www-ai.cs.uni-dortmund.de
 </url>
</urlentity>
<urlentity name="02">
 <url content="http://web.archive.org/web/20011225145054/http://www-ai.cs.uni-dortmund.de/logo.html">
  http://www-ai.cs.uni-dortmund.de/logo.html
 </url>
 <url>
  http://www-ai.cs.uni-dortmund.de/logo.html
 </url>
</urlentity>
<urlentity name="03">
 <url content="http://web.archive.org/web/20011225002800/http://www-ai.cs.uni-dortmund.de/UNIVERSELL/">
  http://www-ai.cs.uni-dortmund.de/UNIVERSELL/
 </url>
 <url>
  http://www-ai.cs.uni-dortmund.de/UNIVERSELL/
 </url>
</urlentity>
<urlentity name="04">
 <url content="http://web.archive.org/web/20010717205202/http://www-ai.cs.uni-dortmund.de/UNIVERSELL/index.eng.html">
  http://www-ai.cs.uni-dortmund.de/UNIVERSELL/index.eng.html
 </url>
```

```
 <url>
  http://www-ai.cs.uni-dortmund.de/UNIVERSELL/index.eng.html
 </url>
</urlentity>
<urlentity name="05">
 <url content="http://web.archive.org/web/20011201025230/http://www-ai.cs.uni-dortmund.de/LEHRE/lehrveranstaltungen_alt.html">
  http://www-ai.cs.uni-dortmund.de/LEHRE/lehrveranstaltungen_alt.html
 </url>
 <url>
  http://www-ai.cs.uni-dortmund.de/LEHRE/lehrveranstaltungen_alt.html
 </url>
</urlentity>
<urlentity name="06">
 <url content="http://web.archive.org/web/20010424025243/http://www-ai.cs.uni-dortmund.de/FORSCHUNG/index.html">
  http://www-ai.cs.uni-dortmund.de/FORSCHUNG/index.html
 </url>
 <url>
  http://www-ai.cs.uni-dortmund.de/FORSCHUNG/index.html
 </url>
</urlentity>
<urlentity name="07">
 <url content="http://web.archive.org/web/20010420221332/http://www-ai.cs.uni-dortmund.de/FORSCHUNG/index.eng.html">
  http://www-ai.cs.uni-dortmund.de/FORSCHUNG/index.eng.html
 </url>
 <url>
  http://www-ai.cs.uni-dortmund.de/FORSCHUNG/index.eng.html
 </url>
</urlentity>
<urlentity name="08">
 <url content="http://web.archive.org/web/20010420163728/http://www-ai.cs.uni-dortmund.de/FORSCHUNG/MLTXT/mltxt.eng.html">
  http://www-ai.cs.uni-dortmund.de/FORSCHUNG/MLTXT/mltxt.eng.html
 </url>
 <url>
  http://www-ai.cs.uni-dortmund.de/FORSCHUNG/MLTXT/mltxt.eng.html
 </url>
</urlentity>
<urlentity name="09">
 <url content="http://web.archive.org/web/20011130142716/http://www.eunite.org/">
  http://www.eunite.org
 </url>
 <url>
  http://www.eunite.org
 </url>
</urlentity>
<urlentity name="10">
 <url content="http://web.archive.org/web/20011225124645/http://www-ai.cs.uni-dortmund.de/FORSCHUNG/PROJEKTE/SFB475A4/">
  http://www-ai.cs.uni-dortmund.de/FORSCHUNG/PROJEKTE/SFB475A4/
 </url>
 <url>
  http://www-ai.cs.uni-dortmund.de/FORSCHUNG/PROJEKTE/SFB475A4/
 </url>
</urlentity>
<urlentity name="11">
 <url content="http://web.archive.org/web/20010717215039/http://www-ai.cs.uni-dortmund.de/FORSCHUNG/PROJEKTE/SFB475C4/">
  http://www-ai.cs.uni-dortmund.de/FORSCHUNG/PROJEKTE/SFB475C4/
 </url>
 <url>
  http://www-ai.cs.uni-dortmund.de/FORSCHUNG/PROJEKTE/SFB475C4/
 </url>
</urlentity>
<urlentity name="12">
 <url content="http://web.archive.org/web/20010430193146/http://www-ai.cs.uni-dortmund.de/FORSCHUNG/buecher.html">
  http://www-ai.cs.uni-dortmund.de/FORSCHUNG/buecher.html
 </url>
 <url>
  http://www-ai.cs.uni-dortmund.de/FORSCHUNG/buecher.html
 </url>
</urlentity>
<urlentity name="13">
 <url content="http://web.archive.org/web/20010424030537/http://www-ai.cs.uni-dortmund.de/FORSCHUNG/REPORTS/reports.eng.html">
  http://www-ai.cs.uni-dortmund.de/FORSCHUNG/REPORTS/reports.eng.html
 </url>
 <url>
  http://www-ai.cs.uni-dortmund.de/FORSCHUNG/REPORTS/reports.eng.html
 </url>
</urlentity>
<urlentity name="14">
 <url content="http://web.archive.org/web/20011024193535/http://www.uni-dortmund.de/TOP/">
  http://www.uni-dortmund.de/TOP/
 </url>
 <url>
  http://www.uni-dortmund.de/TOP/
 </url>
</urlentity>
<urlentity name="15">
 <url content="http://web.archive.org/web/20011224114917/http://ls7-www.cs.uni-dortmund.de/VKInf/">
  http://ls7-www.cs.uni-dortmund.de/VKInf/
 </url>
 <url>
  http://ls7-www.cs.uni-dortmund.de/VKInf/
 </url>
</urlentity>
<urlentity name="16">
```

```
 <url content="http://web.archive.org/web/20011224101149/http://dekanat.cs.uni-dortmund.de/Studierende/Index.html">
  http://dekanat.cs.uni-dortmund.de/Studierende/Index.html
 </url>
 <url>
  http://dekanat.cs.uni-dortmund.de/Studierende/Index.html
 </url>
</urlentity>
<urlentity name="17">
 <url content="http://web.archive.org/web/20011219013640/http://fsinfo.cs.uni-dortmund.de/Studium/">
  http://fsinfo.cs.uni-dortmund.de/Studium/
 </url>
 <url>
  http://fsinfo.cs.uni-dortmund.de/Studium/
 </url>
</urlentity>
<urlentity name="18">
 <url content="http://web.archive.org/web/20011207195718/http://www.co.umist.ac.uk/dsd2002/">
  http://www.co.umist.ac.uk/dsd2002/
 </url>
 <url>
  http://www.co.umist.ac.uk/dsd2002/
 </url>
</urlentity>
<urlentity name="19">
 <url content="http://web.archive.org/web/20011128101650/http://ls1-www.cs.uni-dortmund.de/">
  http://ls1-www.cs.uni-dortmund.de/
 </url>
 <url>
  http://ls1-www.cs.uni-dortmund.de/
 </url>
</urlentity>
<urlentity name="20">
 <url content="http://web.archive.org/web/20010721090551/http://ls2-www.cs.uni-dortmund.de/">
  http://ls2-www.cs.uni-dortmund.de/
 </url>
 <url>
  http://ls2-www.cs.uni-dortmund.de/
 </url>
</urlentity>
<urlentity name="21">
 <url content="http://web.archive.org/web/20010924235605/http://ls4-www.cs.uni-dortmund.de/">
  http://ls4-www.cs.uni-dortmund.de/
 </url>
 <url>
  http://ls4-www.cs.uni-dortmund.de/
 </url>
</urlentity>
<urlentity name="22">
 <url content="http://web.archive.org/web/20010720035506/http://ls6-www.cs.uni-dortmund.de/">
  http://ls6-www.cs.uni-dortmund.de/
 </url>
 <url>
  http://ls6-www.cs.uni-dortmund.de/
 </url>
</urlentity>
<urlentity name="23">
 <url content="http://web.archive.org/web/20010516011851/http://ls12-www.cs.uni-dortmund.de/">
  http://ls12-www.cs.uni-dortmund.de/
 </url>
 <url>
  http://ls12-www.cs.uni-dortmund.de/
 </url>
</urlentity>
<urlentity name="24">
 <url content="http://web.archive.org/web/20010418145650/http://dekanat.cs.uni-dortmund.de/">
  http://dekanat.cs.uni-dortmund.de/
 </url>
 <url>
  http://dekanat.cs.uni-dortmund.de/
 </url>
</urlentity>
<urlentity name="25">
 <url content="http://web.archive.org/web/20011127041753/http://stl-www.cs.uni-dortmund.de/">
  http://stl-www.cs.uni-dortmund.de/
 </url>
 <url>
  http://stl-www.cs.uni-dortmund.de/
 </url>
</urlentity>
<urlentity name="26">
 <url content="http://web.archive.org/web/20011029121740/http://ls10-www.cs.uni-dortmund.de/LS10/Pages/sopra.shtml">
  http://ls10-www.cs.uni-dortmund.de/LS10/Pages/sopra.shtml
 </url>
 <url>
  http://ls10-www.cs.uni-dortmund.de/LS10/Pages/sopra.shtml
 </url>
</urlentity>
<urlentity name="27">
 <url content="http://web.archive.org/web/20011224094000/http://dekanat.cs.uni-dortmund.de/HaPra/index.html">
  http://dekanat.cs.uni-dortmund.de/HaPra/index.html
 </url>
 <url>
  http://dekanat.cs.uni-dortmund.de/HaPra/index.html
```

```
 </url>
</urlentity>
<urlentity name="28">
 <url content="http://web.archive.org/web/20011202194615/http://fsinfo.cs.uni-dortmund.de/">
  http://fsinfo.cs.uni-dortmund.de/
 </url>
 <url>
  http://fsinfo.cs.uni-dortmund.de/
 </url>
</urlentity>
<urlentity name="29">
 <url content="http://web.archive.org/web/20011125035908/http://dekanat.cs.uni-dortmund.de/ZPA/OeffnungszeitenZPA.shtml">
  http://dekanat.cs.uni-dortmund.de/ZPA/OeffnungszeitenZPA.shtml
 </url>
 <url>
  http://dekanat.cs.uni-dortmund.de/ZPA/OeffnungszeitenZPA.shtml
 </url>
</urlentity>
<urlentity name="30">
 <url content="http://web.archive.org/web/20011224094226/http://dekanat.cs.uni-dortmund.de/Ehemalige/Beitritt.html">
  http://dekanat.cs.uni-dortmund.de/Ehemalige/Beitritt.html
 </url>
 <url>
  http://dekanat.cs.uni-dortmund.de/Ehemalige/Beitritt.html
 </url>
</urlentity>
<urlentity name="31">
 <url content="http://web.archive.org/web/20011122165533/http://www.uni-dortmund.de/UniDo/Personal/">
  http://www.uni-dortmund.de/UniDo/Personal/
 </url>
 <url>
  http://www.uni-dortmund.de/UniDo/Personal/
 </url>
</urlentity>
<urlentity name="32">
 <url content="http://web.archive.org/web/20011205094316/http://www.ub.uni-dortmund.de/literatursuche/index.htm">
  http://www.ub.uni-dortmund.de/literatursuche/index.htm
 </url>
 <url>
  http://www.ub.uni-dortmund.de/literatursuche/index.htm
 </url>
</urlentity>
<urlentity name="33">
 <url content="http://web.archive.org/web/20011216193937/http://www.ft-informatik.de/index.html">
  http://www.ft-informatik.de/
 </url>
 <url>
  http://www.ft-informatik.de/
 </url>
</urlentity>
<urlentity name="34">
 <url content="http://web.archive.org/web/20011217182707/http://www.acm.org/">
  http://www.acm.org/
 </url>
 <url>
  http://www.acm.org/
 </url>
</urlentity>
<urlentity name="35">
 <url content="http://web.archive.org/web/20010823075028/http://www.isoc.org/">
  http://www.isoc.org/
 </url>
 <url>
  http://www.isoc.org/
 </url>
</urlentity>
<urlentity name="36">
 <url content="http://web.archive.org/web/20011214030127/http://www.dfn.de/home.html">
  http://www.dfn.de/home.html
 </url>
 <url>
  http://www.dfn.de/home.html
 </url>
</urlentity>
<urlentity name="37">
 <url content="http://web.archive.org/web/20011224220837/http://www-ai.cs.uni-dortmund.de/SOFTWARE/">
  http://www-ai.cs.uni-dortmund.de/SOFTWARE/
 </url>
 <url>
  http://www-ai.cs.uni-dortmund.de/SOFTWARE/
 </url>
</urlentity>
<urlentity name="38">
 <url content="http://web.archive.org/web/20011201211645/http://www-ai.cs.uni-dortmund.de/Harvest/brokers/www-ai/query.html">
  http://www-ai.cs.uni-dortmund.de/Harvest/brokers/www-ai/query.html
 </url>
 <url>
  http://www-ai.cs.uni-dortmund.de/Harvest/brokers/www-ai/query.html
 </url>
</urlentity>
<urlentity name="39">
<url content="http://web.archive.org/web/20010717224840/http://www-ai.cs.uni-dortmund.de/PERSONAL/morik.html">
  http://www-ai.cs.uni-dortmund.de/PERSONAL/morik.html
```

```
 </url>
 <url>
  http://www-ai.cs.uni-dortmund.de/PERSONAL/morik.html
 </url>
</urlentity>
<urlentity name="40">
 <url content="http://web.archive.org/web/20011224220416/http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/">
  http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/
 </url>
<url>
  http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/
 </url>
</urlentity>
<urlentity name="41">
 <url content="http://web.archive.org/web/20011201165136/http://www-ai.cs.uni-dortmund.de/FORSCHUNG/">
  http://www-ai.cs.uni-dortmund.de/FORSCHUNG/
 </url>
 <url>
  http://www-ai.cs.uni-dortmund.de/FORSCHUNG/
 </url>
</urlentity>
<urlentity name="42">
 <url content="http://web.archive.org/web/20011024144342/http://www-ai.cs.uni-dortmund.de/LEHRE/lehre.html">
  http://www-ai.cs.uni-dortmund.de/LEHRE/lehre.html
 </url>
 <url>
  http://www-ai.cs.uni-dortmund.de/LEHRE/lehre.html
 </url>
</urlentity>
<urlentity name="43">
<url content="http://web.archive.org/web/20011224215033/http://www-ai.cs.uni-dortmund.de/PERSONAL/personal.html">
  http://www-ai.cs.uni-dortmund.de/PERSONAL/personal.html
 </url>
 <url>
  http://www-ai.cs.uni-dortmund.de/PERSONAL/personal.html
 </url>
</urlentity>
<urlentity name="44">
 <url content="http://web.archive.org/web/20011225001946/http://www-ai.cs.uni-dortmund.de/index.eng.html">
  http://www-ai.cs.uni-dortmund.de/index.eng.html
 </url>
 <url>
  http://www-ai.cs.uni-dortmund.de/index.eng.html
 </url>
</urlentity>
<urlentity name="45">
 <url content="http://web.archive.org/web/20011225001030/http://www-ai.cs.uni-dortmund.de/SOFTWARE/SVM_LIGHT/">
  http://www-ai.cs.uni-dortmund.de/SOFTWARE/SVM_LIGHT/
 </url>
 <url>
  http://www-ai.cs.uni-dortmund.de/SOFTWARE/SVM_LIGHT/
 </url>
</urlentity>
<urlentity name="46">
 <url content="http://web.archive.org/web/20011224215935/http://www-ai.cs.uni-dortmund.de/SOFTWARE/IL/">
  http://www-ai.cs.uni-dortmund.de/SOFTWARE/IL/
 </url>
 <url>
  http://www-ai.cs.uni-dortmund.de/SOFTWARE/IL/
 </url>
</urlentity>
<urlentity name="47">
 <url content="http://web.archive.org/web/20011201215238/http://www-ai.cs.uni-dortmund.de/LEHRE/DIPLOM/diplom_fertig.html">
  http://www-ai.cs.uni-dortmund.de/LEHRE/DIPLOM/diplom_fertig.html
 </url>
 <url>
  http://www-ai.cs.uni-dortmund.de/LEHRE/DIPLOM/diplom_fertig.html
 </url>
</urlentity>
<urlentity name="48">
 <url>
  http://www.cs.kuleuven.ac.be/~hendrik/ML/
 </url>
 <url>
  http://www.cs.kuleuven.ac.be/~hendrik/
 </url>
</urlentity>
<urlentity name="49">
 <url>
  http://www.fi.muni.cz/usr/popelinsky/old.html.utf-8#courses
 </url>
 <url>
  http://www.fi.muni.cz/~popel/
 </url>
</urlentity>
<urlentity name="50">
 <url>
  http://www.aifb.uni-karlsruhe.de/Lehrangebot/Sommer2002/InfoC/script.html
 </url>
 <url>
  http://www.aifb.uni-karlsruhe.de/WBS/gst/
 </url>
</urlentity>
```

```
<urlentity name="51">
 <url>
  http://www.csd.abdn.ac.uk/~pedwards/teaching/CS5505/slides.html
 </url>
 <url>
  http://www.csd.abdn.ac.uk/~pedwards/
 </url>
</urlentity>
<urlentity name="52">
 <url>
  http://www.cs.bris.ac.uk/Teaching/Resources/COMS30106/
 </url>
 <url>
  http://www.cs.bris.ac.uk/~flach/
 </url>
</urlentity>
<urlentity name="53">
  <url>
  http://www.cs.helsinki.fi/kurssit/
 </url>
 <url>
  http://www.rni.helsinki.fi/~htt/
 </url>
</urlentity>
<urlentity name="54">
 <url>
  http://www-ai.cs.uni-dortmund.de/LEHRE/SEMINARE/INFORMATIONSEXTRAKTION/informationsExtraktion.html
 </url>
 <url>
  http://www-ai.cs.uni-dortmund.de/LEHRE/VORLESUNGEN/MLRN/ml.html
 </url>
</urlentity>
<urlentity name="55">
 <url>
  http://www-ai.cs.uni-dortmund.de/LEHRE/VORLESUNGEN/MLRN/ml.html
 </url>
 <url>
  http://www-ai.cs.uni-dortmund.de/LEHRE/SEMINARE/INFORMATIONSEXTRAKTION/informationsExtraktion.html
 </url>
</urlentity>
<urlentity name="56">
 <url>
  http://www-ai.cs.uni-dortmund.de/LEHRE/DIPLOM/diplom_fertig.html
 </url>
 <url>
  http://www-ai.cs.uni-dortmund.de/LEHRE/DIPLOM/OFFEN/
 </url>
</urlentity>
<urlentity name="57">
 <url>
  http://www-ai.cs.uni-dortmund.de/LEHRE/DIPLOM/OFFEN/
 </url>
 <url>
  http://www-ai.cs.uni-dortmund.de/LEHRE/DIPLOM/diplom_fertig.html
 </url>
</urlentity>
<urlentity name="58">
 <url>
  http://www-ai.cs.uni-dortmund.de/LEHRE/PG/
 </url>
 <url>
  http://www-ai.cs.uni-dortmund.de/LEHRE/PROMOTION/promotion_fertig.html
 </url>
</urlentity>
<urlentity name="59">
 <url>
  http://www-ai.cs.uni-dortmund.de/LEHRE/PROMOTION/promotion_fertig.html
 </url>
 <url>
  http://www-ai.cs.uni-dortmund.de/LEHRE/PG/
 </url>
</urlentity>
<urlentity name="60">
 <url>
  http://sfbci.cs.uni-dortmund.de/home/German/frameset.html
 </url>
 <url>
  http://www-ai.cs.uni-dortmund.de
 </url>
</urlentity>
<urlentity name="61">
 <url>
  http://www-ai.cs.uni-dortmund.de
 </url>
 <url>
  http://www-ai.cs.uni-dortmund.de/PERSONAL/HALL_OF_FAME/index.html
 </url>
</urlentity>
<urlentity name="62">
 <url>
  http://www-ai.cs.uni-dortmund.de
 </url>
 <url content="http://web.archive.org/web/20010926141703/http://www-ai.cs.uni-dortmund.de/">
```

```
  http://www-ai.cs.uni-dortmund.de
 </url>
</urlentity>
```

A.2 WDR

Diese Testdaten sollen kontrollieren, ob der Agent auch mit anderen als universitären Dokumenten zurechtkommt. Aus diesem Grund wurde eine Menge von Dokumenten gewählt, deren Inhalt sich mit verschiedenen Themen aus Politik, Wirtschaft und Freizeit beschäftigt. Aufgrund der Namenskonventionen des Westdeutschen Rundfunks zur Benennung seiner Dokumente, fallen die Dokumente mehrheitlich in die Testmengenkategorie 1 (s. 4.2). Es gibt fast keine zwei Versionen eines Dokuments gleichen Namens. „Fast“ deshalb, weil es einen dynamischen, täglich aktualisierten Teil im Layout jedes Dokuments gibt und Übersichtsdokumente gelegentlich ein neues Thema aufnehmen. Die Schwierigkeit für den Agenten liegt bei dieser Testmenge darin, dass die Themen aktuell sind und von vielen anderen Rundfunksendern auch behandelt werden. Die Suchanfragen müssen daher so gestaltet werden, dass nicht die Dokumente anderer Sender zum gleichen Thema wiedergefunden werden.

```
<urlentity name="01">
 <url>
  http://www.wdr.de/themen/politik/nrw/steinkohle/rag_ausgliederung.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/politik/nrw/steinkohle/rag_ausgliederung.jhtml
 </url>
</urlentity>
<urlentity name="02">
 <url>
  http://www.wdr.de/themen/politik/international/elfter_september/verfassungsschutz_cia/index.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/politik/international/elfter_september/verfassungsschutz_cia/index.jhtml
 </url>
</urlentity>
<urlentity name="03">
 <url>
  http://www.wdr.de/themen/politik/deutschland/stabilitaetspakt/index.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/politik/deutschland/stabilitaetspakt/index.jhtml
 </url>
</urlentity>
<urlentity name="04">
 <url>
  http://www.wdr.de/themen/politik/international/soldaten_afghanistan/landung_koeln.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/politik/international/soldaten_afghanistan/landung_koeln.jhtml
 </url>
</urlentity>
<urlentity name="05">
 <url>
  http://www.wdr.de/themen/politik/nrw/interview_2003/pinkwart/interview.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/politik/nrw/interview_2003/pinkwart/interview.jhtml
 </url>
</urlentity>
<urlentity name="06">
 <url>
  http://www.wdr.de/themen/politik/nrw/gemeindefinanzierung/reaktionen.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/politik/nrw/gemeindefinanzierung/reaktionen.jhtml
 </url>
</urlentity>
<urlentity name="07">
 <url>
  http://www.wdr.de/themen/politik/nrw/gemeindefinanzierung/ruettgers.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/politik/nrw/gemeindefinanzierung/ruettgers.jhtml
 </url>
</urlentity>
```

```
<urlentity name="08">
 <url>
  http://www.wdr.de/themen/politik/nrw/moellemann/inhalt.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/politik/nrw/moellemann/inhalt.jhtml
 </url>
</urlentity>
<urlentity name="09">
 <url>
  http://www.wdr.de/themen/politik/nrw/muellaffaere_spd/inhalt.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/politik/nrw/muellaffaere_spd/inhalt.jhtml
 </url>
</urlentity>
<urlentity name="10">
 <url>
  http://www.wdr.de/themen/homepages/irak.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/homepages/irak.jhtml
 </url>
</urlentity>
<urlentity name="11">
 <url>
  http://www.wdr.de/themen/homepages/d_usa.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/homepages/d_usa.jhtml
 </url>
</urlentity>
<urlentity name="12">
 <url>
  http://www.wdr.de/themen/wirtschaft/arbeit_und_tarifwesen/hartz_reformen/index.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/wirtschaft/arbeit_und_tarifwesen/hartz_reformen/index.jhtml
 </url>
</urlentity>
<urlentity name="13">
 <url>
  http://www.wdr.de/themen/wirtschaft/1/haushaltsgeraete/index.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/wirtschaft/1/haushaltsgeraete/index.jhtml
 </url>
</urlentity>
<urlentity name="14">
 <url>
  http://www.wdr.de/themen/wirtschaft/wirtschaftsbranche/ford/krise.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/wirtschaft/wirtschaftsbranche/ford/krise.jhtml
 </url>
</urlentity>
<urlentity name="15">
 <url>
  http://www.wdr.de/themen/wirtschaft/wirtschaftsbranche/rwe/usa_drohen.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/wirtschaft/wirtschaftsbranche/rwe/usa_drohen.jhtml
 </url>
</urlentity>
<urlentity name="16">
 <url>
  http://www.wdr.de/themen/wirtschaft/arbeit_und_tarifwesen/ich_ag/erfolg.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/wirtschaft/arbeit_und_tarifwesen/ich_ag/erfolg.jhtml
 </url>
</urlentity>
<urlentity name="17">
 <url>
  http://www.wdr.de/themen/wirtschaft/geld-_und_kreditwesen/westlb/index_030806.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/wirtschaft/geld-_und_kreditwesen/westlb/index_030806.jhtml
 </url>
</urlentity>
<urlentity name="18">
 <url>
  http://www.wdr.de/themen/wirtschaft/wirtschaftsbranche/luftfahrt/flughafen_weeze.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/wirtschaft/wirtschaftsbranche/luftfahrt/flughafen_weeze.jhtml
 </url>
</urlentity>
<urlentity name="19">
 <url>
  http://www.wdr.de/online/jobs/jobzeit/index.phtml?rubrikenstyle=wirtschaft
 </url>
 <url>
```

```
  http://www.wdr.de/online/jobs/jobzeit/index.phtml?rubrikenstyle=wirtschaft
 </url>
</urlentity>
<urlentity name="20">
 <url>
  http://www.wdr.de/themen/forschung/technik/solarflitzer/index.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/forschung/technik/solarflitzer/index.jhtml
 </url>
</urlentity>
<urlentity name="21">
 <url>
  http://www.wdr.de/themen/forschung/interdisziplinaer/virtuelle_bibliothek/index.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/forschung/interdisziplinaer/virtuelle_bibliothek/index.jhtml
 </url>
</urlentity>
<urlentity name="22">
 <url>
  http://www.wdr.de/themen/homepages/kleine_anfrage.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/homepages/kleine_anfrage.jhtml
 </url>
</urlentity>
<urlentity name="23">
 <url>
  http://www.wdr.de/themen/kultur/1/linkshaender_tag_2003/gaestebuch.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/kultur/1/linkshaender_tag_2003/gaestebuch.jhtml
 </url>
</urlentity>
<urlentity name="24">
 <url>
  http://www.wdr.de/themen/kultur/rundfunk/lilipuz_sommertour2003/index.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/kultur/rundfunk/lilipuz_sommertour2003/index.jhtml
 </url>
</urlentity>
<urlentity name="25">
 <url>
  http://www.wdr.de/themen/kultur/bildung_und_erziehung/bafoeg_missbrauch/index.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/kultur/bildung_und_erziehung/bafoeg_missbrauch/index.jhtml
 </url>
</urlentity>
<urlentity name="26">
 <url>
  http://www.wdr.de/themen/kultur/personen/ruge/index.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/kultur/personen/ruge/index.jhtml
 </url>
</urlentity>
<urlentity name="27">
 <url>
  http://www.wdr.de/themen/kultur/quiz/quiz_rechtschreibreform.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/kultur/quiz/quiz_rechtschreibreform.jhtml
 </url>
</urlentity>
<urlentity name="28">
 <url>
  http://www.wdr.de/themen/panorama/lifestyle/modemesse_reevolutions_2003/sommer.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/panorama/lifestyle/modemesse_reevolutions_2003/sommer.jhtml
 </url>
</urlentity>
<urlentity name="29">
 <url>
  http://www.wdr.de/themen/kultur/1/kinderschutzbund/interview.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/kultur/1/kinderschutzbund/interview.jhtml
 </url>
</urlentity>
<urlentity name="30">
 <url>
  http://www.wdr.de/themen/kultur/netzkultur/heimatinseln/inseln.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/kultur/netzkultur/heimatinseln/inseln.jhtml
 </url>
</urlentity>
<urlentity name="31">
 <url>
```

```
  http://www.wdr.de/themen/homepages/popkomm2003.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/homepages/popkomm2003.jhtml
 </url>
</urlentity>
<urlentity name="32">
 <url>
  http://www.wdr.de/themen/homepages/unicef.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/homepages/unicef.jhtml
 </url>
</urlentity>
<urlentity name="33">
 <url>
  http://www.wdr.de/themen/kultur/netzkultur/nrw_privat/wohnungen.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/kultur/netzkultur/nrw_privat/wohnungen.jhtml
 </url>
</urlentity>
<urlentity name="34">
 <url>
  http://www.wdr.de/themen/computer/internet/blaster/index.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/computer/internet/blaster/index.jhtml
 </url>
</urlentity>
<urlentity name="35">
 <url>
  http://www.wdr.de/themen/computer/software/virenticker/index.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/computer/software/virenticker/index.jhtml
 </url>
</urlentity>
<urlentity name="36">
 <url>
  http://www.wdr.de/themen/computer/schiebwoche/2003/index_33.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/computer/schiebwoche/2003/index_33.jhtml
 </url>
</urlentity>
<urlentity name="37">
 <url>
  http://www.wdr.de/themen/computer/internet/sicherheit/exploit.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/computer/internet/sicherheit/exploit.jhtml
 </url>
</urlentity>
<urlentity name="38">
 <url>
  http://www.wdr.de/themen/computer/internet/barrierefreies_internet/internetcafe_reportage.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/computer/internet/barrierefreies_internet/internetcafe_reportage.jhtml
 </url>
</urlentity>
<urlentity name="39">
 <url>
  http://www.wdr.de/themen/computer/angeklickt/tagestipp/tagestipp.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/computer/angeklickt/tagestipp/tagestipp.jhtml
 </url>
</urlentity>
<urlentity name="40">
 <url>
  http://www.wdr.de/themen/computer/angeklickt/webtv/index.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/computer/angeklickt/webtv/index.jhtml
 </url>
</urlentity>
<urlentity name="41">
 <url>
  http://www.wdr.de/themen/sport/1/holzfaeller/index.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/sport/1/holzfaeller/index.jhtml
 </url>
</urlentity>
<urlentity name="42">
 <url>
  http://www.wdr.de/themen/homepages/olympia_2012.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/homepages/olympia_2012.jhtml
 </url>
```

```
</urlentity>
<urlentity name="43">
 <url>
  http://www.wdr.de/themen/panorama/2/lottojackpot_nrw/index.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/panorama/2/lottojackpot_nrw/index.jhtml
 </url>
</urlentity>
<urlentity name="44">
 <url>
  http://www.wdr.de/themen/panorama/kriminalitaet02/gladbecker_geiseldrama_1988/index.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/panorama/kriminalitaet02/gladbecker_geiseldrama_1988/index.jhtml
 </url>
</urlentity>
<urlentity name="45">
 <url>
  http://www.wdr.de/themen/panorama/2/stromausfall_bielefeld/index.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/panorama/2/stromausfall_bielefeld/index.jhtml
 </url>
</urlentity>
<urlentity name="46">
 <url>
  http://www.wdr.de/themen/panorama/kriminalitaet02/missbrauch_moenchengladbach/index.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/panorama/kriminalitaet02/missbrauch_moenchengladbach/index.jhtml
 </url>
</urlentity>
<urlentity name="47">
 <url>
  http://www.wdr.de/themen/panorama/wetter/sommer_2003/abkuehlung_030814.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/panorama/wetter/sommer_2003/abkuehlung_030814.jhtml
 </url>
</urlentity>
<urlentity name="48">
 <url>
  http://www.wdr.de/themen/panorama/wetter/sommer_2003/waldbraende.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/panorama/wetter/sommer_2003/waldbraende.jhtml
 </url>
</urlentity>
<urlentity name="49">
 <url>
  http://www.wdr.de/themen/panorama/2/bombenentschaerfungen_nrw_2002/index.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/panorama/2/bombenentschaerfungen_nrw_2002/index.jhtml
 </url>
</urlentity>
<urlentity name="50">
 <url>
  http://www.wdr.de/themen/panorama/wetter/sommer_2003/hitze.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/panorama/wetter/sommer_2003/hitze.jhtml
 </url>
</urlentity>
<urlentity name="51">
 <url>
  http://www.wdr.de/themen/gesundheit/1/antipasti/index.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/gesundheit/1/antipasti/index.jhtml
 </url>
</urlentity>
<urlentity name="52">
 <url>
  http://www.wdr.de/themen/gesundheit/gesundheitswesen/gesetzliche_krankenkassen/sparzwang.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/gesundheit/gesundheitswesen/gesetzliche_krankenkassen/sparzwang.jhtml
 </url>
</urlentity>
<urlentity name="53">
 <url>
  http://www.wdr.de/themen/gesundheit/gesundheitswesen/gesetzliche_krankenkassen/bkk_beitragserhoehung.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/gesundheit/gesundheitswesen/gesetzliche_krankenkassen/bkk_beitragserhoehung.jhtml
 </url>
</urlentity>
<urlentity name="54">
 <url>
  http://www.wdr.de/themen/freizeit/freizeitgestaltung/insel_sommer/monkeys_island.jhtml
 </url>
```

```
 <url>
  http://www.wdr.de/themen/freizeit/freizeitgestaltung/insel_sommer/monkeys_island.jhtml
 </url>
</urlentity>
<urlentity name="55">
 <url>
  http://www.wdr.de/themen/freizeit/freizeitgestaltung/future_parade/index.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/freizeit/freizeitgestaltung/future_parade/index.jhtml
 </url>
</urlentity>
<urlentity name="56">
 <url>
  http://www.wdr.de/themen/freizeit/freizeitgestaltung/mundorgel_50jahre/interview_guildo_horn.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/freizeit/freizeitgestaltung/mundorgel_50jahre/interview_guildo_horn.jhtml
 </url>
</urlentity>
<urlentity name="57">
 <url>
  http://www.wdr.de/themen/verkehr/strasse/tunnel_b236/index.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/verkehr/strasse/tunnel_b236/index.jhtml
 </url>
</urlentity>
<urlentity name="58">
 <url>
  http://www.wdr.de/themen/verkehr/schiene/deutsche_bahn/maengel.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/verkehr/schiene/deutsche_bahn/maengel.jhtml
 </url>
</urlentity>
<urlentity name="59">
 <url>
  http://www.wdr.de/themen/verkehr/strasse/radarwarngeraete/autobahn_warnung.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/verkehr/strasse/radarwarngeraete/autobahn_warnung.jhtml
 </url>
</urlentity>
<urlentity name="60">
 <url>
  http://www.wdr.de/themen/verkehr/wasser/niedrigwasser_rhein_ruhr/niedrigwasser_0308.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/verkehr/wasser/niedrigwasser_rhein_ruhr/niedrigwasser_0308.jhtml
 </url>
</urlentity>
<urlentity name="61">
 <url>
  http://www.wdr.de/themen/panorama/2/verdaechtiger_koffer_koeln/index.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/panorama/2/verdaechtiger_koffer_koeln/index.jhtml
 </url>
</urlentity>
<urlentity name="62">
 <url>
  http://www.wdr.de/themen/verkehr/schiene/metrorapid/inhalt.jhtml
 </url>
 <url>
  http://www.wdr.de/themen/verkehr/schiene/metrorapid/inhalt.jhtml
 </url>
</urlentity>
```

Anhang B

Operatortestläufe

In diesem Kapitel werden die Ergebnisse der Testläufe der einzelnen Strategeme vorgestellt.

Zuordnung der Tabellenspalten zu den einzelnen Strategemen:

1. Großbuchstaben
2. Großbuchstaben + einfache Stoppwortelimination
3. Großbuchstaben + Stoppwortelimination
4. Häufige Wörter
5. Häufige Wörter + TF–IDF
6. Phrase
7. Längste Phrase
8. Satzphrase
9. verbesserte Satzphrase
10. Namen von Personen
11. Zufällige Wörter
12. Zufällige Wörter + Stoppwortelimination
13. Zufällige Wörter / Häufigkeitsverteilt
14. SpiderA
15. SpiderC

Die Einträge $\frac{a}{b}$ bedeuten für die Strategeme 1-10:

a sind die gestellten Anfragen an Google,

b sind die besuchten Dokumente.

Bei den SpiderOperatoren wird etwas anders gezählt:

a ist die Anzahl der genutzten Startpunkte.

b ist die Anzahl der besuchten Dokumente einschließlich des gesuchten bis das gesuchte Dokument gefunden wurde oder die Suche erfolglos war.

Wenn die Suche nicht erfolgreich war, wird das Ergebnis rot angegeben.

Die Feststellung, ob das durchschnittliche Ergebnis der Zufallsstrategeme (11-13) ein Gefunden oder ein Nichtgefunden ist, erfolgt durch einen Mehrheitsentscheid. Falls mindestens drei Testläufe eines Strategems zu einem Dokument das Dokument gefunden haben, wird es in der Zusammenfassung als gefunden gewertet, sonst nicht. Die Werte für die Eintragungen in diese Spalten, sind die Erwartungswerte, wie sie in den Abschnitten B.1.1 bis B.1.3 bzw. B.2.1 bis B.2.3 berechnet werden.

Die Erwartungswerte werden nach der Formel für den erwartungstreuen Mittelwert einer Stichprobe (vgl. [Zei96]) geschätzt:

$$\mu = \frac{1}{n} \sum_{j=1}^{n} X_j \tag{B.1}$$

Die Standardabweichung wird nach der Formel für das „Schätzen der Streuungsfunktion“ (vgl. [Zei96]):

$$\sigma = \sqrt{\frac{1}{n-1} \frac{1}{n} \sum_{j=1}^{n} X_j - \mu} \tag{B.2}$$

berechnet.

Sowohl Erwartungswert als auch Standardabweichung beziehen sich immer auf die Zeile in der sie angegeben sind. Das gilt insbesondere auch für die Erwartungswerte und Standardabweichungen der Gesamtergebnisse.

Es kann vorkommen, dass statt eines Ergebnisses ein „-“ in der Tabellenzelle steht. Das bedeutet, dass Google zum Zeitpunkt der Anfrage nicht erreichbar war oder dass nach den allen Anfrageversuchen kein ähnliches Dokument gefunden wurde. Ähnlich bedeutet in diesem Zusammenhang, dass kein Dokument einen Abstand von höchstens `85` zum Ursprungsdokument hatte. Wenn dieser Fall auftritt, wird das Dokument als nicht gefunden gewertet.

Um die Abhängigkeit der suchmaschinenbasierten Strategeme von der URL zu testen (s. 4.2 Menge 3), wurde bei diesen auch ausgewertet, welche Ergebnisse sich ergeben, wenn das Nutzen des Dokument–Dateinamens nicht möglich war. Das führt dazu, dass die Strategeme, die als Spezialisierungsschritt das Anfügen des Dateinamens implementieren, einige Dokumente nicht mehr finden. Ergebnisse die mit mindestens einem Spezialisierungsschritt zustande kamen, sind gelb hinterlegt. Wenn das Ergebnis als gefunden gekennzeichnet ist, dann bedeutet es, dass das Dokument ohne den Spezialisierungsschritt das Dokument wiedergefunden worden wäre. Ist das Ergebnis rot gekennzeichnet, dann hat auch die vorgenommen Spezialisierung nicht zur Wiederauffindung geführt. Die Eintragungen in den Spalten beziehen sich immer auf die Variante, in der das Spezialisieren erlaubt ist.

Die letzte Spalte gibt die Nr. des Testdatensatz aus Anhang A an zu der das Ergebnis gehört.

Am Ende der Tabellen wird jeweils die Anzahl der gefundenen Tabellen mit und ohne Spezialisierungsmöglichkeit der Operatoren angegeben.

B.1 Tabellarische Zusammenfassung – Universität

Die hier vorgestellten Ergebnisse stammen aus Testläufen vom 23. Juli 2003.

Ziel	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Nr
http://www-ai.cs.uni-dortmund.de	5 0	—	—	1 1	1 2	4 22	1 1	4 5	3 4	—	3,4 4,6	2,4 3	2,8 5,8	—	1 1	62
http://www-ai.cs.uni-dortmund.de	5 0	—	—	1 2	1 2	4 22	1 1	4 5	3 4	—	3,33 1,67	1 1,75	2 6,4	1 3	1 4	61
http://sfbci.cs.uni-dortmund.de/home/German/frameset.html	5 1	—	—	1 2	1 2	5 10	1 2	1 3	1 10	3 10	1,75 3,5	1,75 6,5	2,2 12,8	—	1 21	60
http://www-ai.cs.uni-dortmund.de/LEHRE/PROMOTION/promotion_fertig.html	1 2	1 1	1 1	—	—	4 0	1 10	6 0	1 6	1 1	1 1	1,6 2,6	1,6 4,4	2 18	1 16	59
http://www-ai.cs.uni-dortmund.de/LEHRE/PG/	1 2	1 2	1 2	1 2	1 10	1 10	1 10	1 10	1 3	—	1,2 2	2,4 2,4	1 3,6	3 96	1 30	58
http://www-ai.cs.uni-dortmund.de/LEHRE/DIPLOM/OFFEN/	1 1	1 1	1 1	1 1	1 1	1 1	1 2	1 2	1 2	—	1 2,4	1,6 4,4	1 5,2	3 17	1 13	57
http://www-ai.cs.uni-dortmund.de/LEHRE/DIPLOM/diplom_fertig.html	3 23	3 22	3 2	4 10	4 11	5 10	1 8	6 0	1 1	—	2,5 4,75	3 3,5	1 1	3 97	1 15	56
http://www-ai.cs.uni-dortmund.de/LEHRE/VORLESUNGEN/MLRN/ml.html	1 2	1 1	1 1	1 1	1 1	5 10	1 2	2 2	2 9	—	1,2 1,2	1,4 2	1 1,4	4 99	1 15	55
http://www-ai.cs.uni-dortmund.de/LEHRE/SEMINARE/INFORMATIONSEXTRAKTION/informationsExtraktion.html	5 0	—	1 1	3 1	3 1	1 1	1 1	1 1	1 3	1 1	1,75 1	1 1	1 2,6	4 101	—	54
http://www.rni.helsinki.fi/kurssit/	—	—	—	—	—	4 0	1 1	6 0	1 1	—	2 3,33	3,8 1,8	4 1,4	1 37	1 6	53
http://www.cs.bris.ac.uk/Teaching/Resources/COMS30106/	1 1	1 1	1 1	1 10	1 10	2 1	1 10	4 1	6 0	—	1 5,4	1 5,8	1 5,6	1 43	1 12	52
http://www.csd.abdn.ac.uk/~pedwards/teaching/CS5505/slides.html	1 1	1 1	1 1	—	—	5 10	1 3	6 0	2 1	1 3	4 5	3,8 3,4	3,2 1,2	—	1 20	51
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Nr

Ziel	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Nr
http://www.aifb.uni-karlsruhe.de/Lehrangebot/Sommer2002/InfoC/script.html	1 9	4 10	1 1	1 4	1 4	4 8	1 1	6 0	1 1	—	3,75 6,75	2,4 4,2	2,6 11,8	2 212	1 34	50
http://www.fi.muni.cz/~popel/old.html.utf-8#courses	3 7	4 8	—	1 9	1 9	1 8	—	1 8	1 7	1 7	1,67 4,33	1,6 5,6	1 7,8	1 45	1 13	49
http://www.cs.kuleuven.ac.be/~hendrik/ML	1 2	1 1	1 1	1 10	1 10	4 1	1 1	2 1	2 1	—	1 1,2	1 5,6	1,8 7,6	1 16	1 11	48
http://www-ai.cs.uni-dortmund.de/LEHRE/DIPLOM/diplom_fertig.html	1 10	1 9	4 2	3 6	3 6	2 10	3 3	6 0	1 1	—	1,75 2,5	4,75 1	1 1,2	3 97	1 1	47
http://www-ai.cs.uni-dortmund.de/SOFTWARE/IL/	1 7	1 9	4 7	1 9	1 10	—	1 10	2 10	1 10	1 3	1 3	1,33 6,33	1 2,2	1 98	1 1	46
http://www-ai.cs.uni-dortmund.de/SOFTWARE/SVM_LIGHT/	4 9	4 9	4 7	1 1	1 1	3 6	4 0	2 10	1 7	—	2,5 8	1,75 7,75	2,4 13,4	1 98	1 1	45
http://www-ai.cs.uni-dortmund.de/index.eng.html	5 0	—	—	6 49	4 29	4 10	4 0	6 0	6 13	1 8	5 28	5,75 10,75	4,6 20,6	1 84	1 1	44
http://www-ai.cs.uni-dortmund.de/PERSONAL/personal.html	5 0	1 9	1 1	1 2	1 2	3 2	1 10	6 0	1 1	—	3,4 5,6	4 3,2	4 7,2	2 22	1 1	43
http://www-ai.cs.uni-dortmund.de/LEHRE/lehre_15_02_2002.html	1 1	1 1	1 1	1 3	1 3	6 0	1 3	4 3	2 2	—	2,5 4	2,6 1,2	1,2 5,8	2 103	1 22	42
http://www-ai.cs.uni-dortmund.de/FORSCHUNG/	5 0	—	1 1	1 6	1 6	3 6	—	1 9	1 1	1 10	1 1,33	2,5 0,75	2,4 2,2	1 4	1 1	41
http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/	5 6	—	4 6	1 8	1 8	3 10	1 1	1 1	1 1	—	3,67 4,33	3 2	1 6,25	2 71	1 1	40
http://www-ai.cs.uni-dortmund.de/PERSONAL/morik.html	5 0	—	—	1 6	1 6	3 3	1 10	1 7	1 1	1 1	1 1,4	1 1	1 3,4	2 6	1 1	39
http://www-ai.cs.uni-dortmund.de/Harvest/brokers/www-ai/query.html	5 0	—	1 1	1 2	1 2	4 2	1 1	1 2	1 1	—	2,25 5,25	1,6 5,4	1,8 1,4	1 40	1 1	38
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**Nr**

Ziel	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Nr
http://www-ai.cs. uni-dortmund.de/ SOFTWARE/	5 0	—	—	3 2	3 2	4 4	3 2	4 2	1 2	1 5	1 3,5	2 2,25	2,2 5	1 10	1 1	37
http://www.dfn.de/	2 1	1 1	1 10	1 10	1 10	2 10	1 6	1 10	1 10	—	1 10	2,4 10,2	1 9,2	1 1	—	36
http://www.isoc.org/	—	—	3 5	1 8	1 8	5 0	1 9	1 9	1 3	3 10	3,25 4,5	1,8 6,4	2 6,2	—	1 1	35
http://www.acm.org/	—	—	1 3	1 8	1 8	4 9	1 10	3 8	1 10	—	2,67 4	3,33 9,33	1,6 10,6	—	1 1	34
http://www. ft-informatik.de/	—	—	—	5 20	5 19	4 8	4 0	3 4	1 4	—	4 8	3,5 9,5	3,2 11,4	—	1 1	33
http://www.ub. uni-dortmund.de/ literatursuche/index.htm	1 1	3 1	1 1	3 1	3 1	4 10	1 10	6 0	4 9	—	2,5 3	3,8 5	3,4 5,2	1 1	1 1	32
http://www.uni-dortmund. de/UniDo/Personal/	—	4 3	—	—	—	4 10	1 7	3 10	3 9	—	3,25 11,75	—	3,2 5,4	2 54	—	31
http://dekanat.cs. uni-dortmund.de/ Ehemalige/Beitritt.html	2 3	1 1	1 1	1 1	1 1	5 10	1 1	3 1	1 1	—	1,4 1	1,8 2,4	1,8 7,2	1 6	1 1	30
http://dekanat.cs. uni-dortmund.de/ZPA/ OeffnungszeitenZPA.shtml	1 2	4 1	4 1	3 2	3 2	5 10	1 1	5 1	2 1	—	2,8 4	3 6,6	3 3,4	1 61	1 1	29
http://fsinfo.cs. uni-dortmund.de/	5 0	—	—	1 10	1 10	5 0	1 10	2 0	1 3	—	1,6 4,2	2,6 5	1 4,2	—	1 1	28
http://dekanat.cs. uni-dortmund.de/HaPra/ index.html	4 10	1 1	2 2	4 9	4 9	4 0	1 1	4 1	3 1	1 10	3,4 3,6	3,8 13,8	3,4 8,2	1 1	1 1	27
http://ls10-www.cs. uni-dortmund.de/LS10/ Pages/sopra-specials/ anmeldung/sopra.shtml	5 16	—	4 10	1 10	1 10	3 0	1 1	3 10	3 10	1 8	3 2,67	4,4 13,4	3,4 8,6	1 35	1 2	26
http://stl-www.cs. uni-dortmund.de/	5 6	—	—	1 2	1 2	4 2	1 2	2 2	1 2	—	2 3,67	2 4,5	3,2 4,6	—	1 1	25
http://dekanat.cs. uni-dortmund.de/	5 0	—	3 11	—	—	4 10	—	6 0	6 0	—	—	5 2	1,6 5	—	1 1	24
http://ls12-www.cs. uni-dortmund.de/	5 0	—	—	1 4	1 7	4 4	1 5	3 3	2 6	—	3 7	4,5 9	3 11,4	—	1 1	23
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**Nr**

Ziel	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Nr
http://ls6-www.cs.uni-dortmund.de/	— —	— —	— —	3 1	3 1	5 15	2 1	1 0	1 2	1 10	1 1	3 3	3,6 2,8	— —	1 1	22
http://ls4-www.cs.uni-dortmund.de/	1 9	— —	— —	1 6	1 6	1 8	1 4	2 4	4 5	— —	1,67 4,33	4 9	3 6,6	— —	1 1	21
http://ls2-www.cs.uni-dortmund.de/	1 3	1 3	1 3	— —	— —	4 0	1 10	— —	3 11	— —	3 9,5	1,25 3,5	1 10	— —	1 1	20
http://ls1-www.cs.uni-dortmund.de/	5 0	— —	— —	4 3	4 3	5 0	— —	— —	5 10	— —	4 6,5	4,25 7,25	2 6,6	— —	1 1	19
http://www.co.umist.ac.uk/dsd2002/	1 9	1 7	1 3	1 10	1 10	4 0	1 10	1 10	1 10	1 9	1,2 7	1 8,8	3,6 8,8	1 79	1 1	18
http://fsinfo.cs.uni-dortmund.de/Studium/	1 1	1 1	1 1	1 3	1 3	4 0	1 2	4 0	1 2	1 10	1 1	1 2,8	1 3,6	1 28	1 1	17
http://dekanat.cs.uni-dortmund.de/Studierende/Index.html	4 35	— —	2 18	2 18	1 10	2 10	1 10	— —	1 10	— —	3 24	6 41,2	1 26,2	1 1	1 1	16
http://ls7-www.cs.uni-dortmund.de/VKInf/vorkurs_ws0102.shtml	— —	— —	1 1	1 5	2 5	5 2	1 8	2 10	1 10	1 10	2,5 6,75	1,5 5	3,2 5	1 31	1 7	15
http://www.uni-dortmund.de/TOP/	3 10	— —	4 13	— —	— —	1 10	— —	— —	3 20	— —	2,5 13,75	2 14,5	1,2 8,6	1 33	— —	14
http://www-ai.cs.uni-dortmund.de/FORSCHUNG/REPORTS/reports.eng.html	1 1	1 1	1 1	1 10	1 10	1 1	1 1	1 4	1 3	1 1	1 1	1 1	1 2,8	2 125	1 1	13
http://www-ai.cs.uni-dortmund.de/FORSCHUNG/buecher.html	5 0	— —	— —	4 9	4 9	4 6	1 1	6 0	4 10	1 2	3,5 7,75	3,2 5	1 5,2	1 121	1 1	12
http://www-ai.cs.uni-dortmund.de/FORSCHUNG/PROJEKTE/SFB475C4/	5 0	— —	— —	1 3	1 3	4 4	1 9	2 1	1 1	— —	1,75 2	1,8 1,2	2,4 2,8	2 117	1 1	11
http://www-ai.cs.uni-dortmund.de/FORSCHUNG/PROJEKTE/SFB475A4/	1 8	1 7	1 2	1 10	1 10	— —	1 10	6 0	1 10	1 2	1 1	3,5 3,5	1,4 3,4	2 116	1 1	10
http://www.eunite.org	1 10	1 10	1 10	1 10	1 10	1 10	1 10	— —	— —	— —	1 10	1 10	2,6 10	— —	1 1	09
http://www-ai.cs.uni-dortmund.de/FORSCHUNG/MLTXT/mltxt.eng.html	1 2	1 2	— —	1 10	1 10	5 10	1 6	6 0	1 7	3 10	2,5 3,5	2,2 3,2	1 4,4	2 99	1 1	08
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**Nr**

Ziel	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Nr
http://www-ai.cs.uni-dortmund.de/FORSCHUNG/index.eng.html	1 2	1 2	—	1 6	1 6	3 6	1 1	1 4	1 2	—	1 1,75	1 2,4	1,6 3,4	1 95	1 1	07
http://www-ai.cs.uni-dortmund.de/FORSCHUNG/index.html	5 0	—	—	1 6	1 6	3 6	1 1	1 9	1 5	1 10	1,6 1,2	3,5 5,25	1,2 1,4	1 1	1 1	06
http://www-ai.cs.uni-dortmund.de/LEHRE/lehrveranstaltungen_alt.html	1 1	4 1	1 1	1 3	1 3	4 10	1 1	6 1	1 10	—	1 1	2,33 0,67	1,4 4,4	2 98	1 1	05
http://www-ai.cs.uni-dortmund.de/UNIVERSELL/index.eng.html	1 1	1 1	1 1	1 1	1 1	4 1	1 1	3 1	2 8	—	2 6,25	1,8 4,6	2 2	1 100	1 1	04
http://www-ai.cs.uni-dortmund.de/UNIVERSELL/	1 1	1 1	1 1	1 1	1 1	4 1	—	3 1	1 1	—	1 1	1 1	1 6,2	1 4	1 1	03
http://www-ai.cs.uni-dortmund.de/logo.html	1 1	1 1	1 1	1 1	1 1	4 0	1 1	1 1	2 1	—	1 1,4	2 1	1,4 3,8	1 2	1 1	02
http://www-ai.cs.uni-dortmund.de	5 0	—	—	4 8	4 6	—	4 0	6 4	6 7	1 10	3 2	5 0	1 9,6	—	1 1	01
Gesamt gefunden	**29**	**28**	**35**	**42**	**45**	**29**	**41**	**30**	**49**	**10**	**35,6**	**32,8**	**37,4**	**33**	**47**	
ohne Spezialisieren	**28**	28	35	42	**44**	**26**	41	30	49	10	**35,2**	**32,0**	**36,8**	33	47	

B.1.1 Zufällige Wörter

Die hier vorgestellten Ergebnisse stammen aus Testläufen vom 28. Juli 2003.

Ziel	1	2	3	4	5	μ	σ	Nr
http://www-ai.cs.uni-dortmund.de	4 4	2 2	2 2	4 7	5 8	3,4 4,6	1,34 2,79	62
http://www-ai.cs.uni-dortmund.de	4 2	—	2 2	4 1	—	3,33 1,67	1,15 0,58	61
http://sfbci.cs.uni-dortmund.de/home/German/frameset.html	1 2	1 2	—	1 2	4 8	1,75 3,5	1,5 3	60
http://www-ai.cs.uni-dortmund.de/LEHRE/PROMOTION/promotion_fertig.html	1 1	1 1	1 1	1 1	1 1	1 1	0 0	59
http://www-ai.cs.uni-dortmund.de/LEHRE/PG/	1 2	1 2	1 2	1 2	2 2	1,2 2	0,45 0	58
http://www-ai.cs.uni-dortmund.de/LEHRE/DIPLOM/OFFEN/	1 1	1 1	1 8	1 1	1 1	1 2,4	0 3,13	57
http://www-ai.cs.uni-dortmund.de/LEHRE/DIPLOM/diplom_fertig.html	1 4	4 8	4 3	—	1 4	2,5 4,75	1,73 2,22	56
http://www-ai.cs.uni-dortmund.de/LEHRE/VORLESUNGEN/MLRN/ml.html	2 1	1 2	1 1	1 1	1 1	1,2 1,2	0,45 0,45	55
http://www-ai.cs.uni-dortmund.de/LEHRE/SEMINARE/INFORMATIONSEXTRAKTION/informationsExtraktion.html	1 1	4 1	1 1	1 1	—	1,75 1	1,5 0	54
http://www.rni.helsinki.fi/kurssit/	—	4 8	1 1	—	1 1	2 3,33	1,73 4,04	53
http://www.cs.bris.ac.uk/Teaching/Resources/COMS30106/	1 8	1 8	1 1	1 9	1 1	1 5,4	0 4,04	52
http://www.csd.abdn.ac.uk/~pedwards/teaching/CS5505/slides.html	—	—	—	—	4 5	4 5	– –	51
	1	2	3	4	5	μ	σ	Nr

Ziel	1	2	3	4	5	μ	σ	Nr
http://www.aifb.uni-karlsruhe.de/Lehrangebot/Sommer2002/InfoC/script.html	3 1	4 9	4 7	—	4 10	3,75 6,75	0,5 4,03	50
http://www.fi.muni.cz/~popel/old.html.utf-8#courses	3 6	1 1	1 6	—	—	1,67 4,33	1,15 2,89	49
http://www.cs.kuleuven.ac.be/~hendrik/ML	1 1	1 1	1 1	1 1	1 2	1 1,2	0 0,45	48
http://www-ai.cs.uni-dortmund.de/LEHRE/DIPLOM/diplom_fertig.html	—	4 4	1 2	1 2	1 2	1,75 2,5	1,5 1	47
http://www-ai.cs.uni-dortmund.de/SOFTWARE/IL/	—	1 3	1 3	1 3	1 3	1 3	0 0	46
http://www-ai.cs.uni-dortmund.de/SOFTWARE/SVM_LIGHT/	3 1	—	1 9	2 11	4 11	2,5 8	1,29 4,76	45
http://www-ai.cs.uni-dortmund.de/index.eng.html	—	6 33	—	2 19	6 32	5 28	2,65 7,81	44
http://www-ai.cs.uni-dortmund.de/PERSONAL/personal.html	6 17	1 2	2 1	4 4	4 4	3,4 5,6	1,95 6,5	43
http://www-ai.cs.uni-dortmund.de/LEHRE/lehre_15_02_2002.html	4 2	—	—	—	1 6	2,5 4	2,12 2,83	42
http://www-ai.cs.uni-dortmund.de/FORSCHUNG/	1 1	—	—	1 1	1 2	1 1,33	0 0,58	41
http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/	4 2	3 9	—	—	4 2	3,67 4,33	0,58 4,04	40
http://www-ai.cs.uni-dortmund.de/PERSONAL/morik.html	1 1	1 1	1 1	1 3	1 1	1 1,4	0 0,89	39
http://www-ai.cs.uni-dortmund.de/Harvest/brokers/www-ai/query.html	1 2	1 2	3 15	—	4 2	2,25 5,25	1,5 6,5	38
	1	**2**	**3**	**4**	**5**	μ	σ	**Nr**

Ziel	1	2	3	4	5	μ	σ	Nr
http://www-ai.cs. uni-dortmund.de/ SOFTWARE/	— —	1 8	1 2	1 2	1 2	1 3,5	0 3	37
http://www.dfn.de/home. html	— —	1 10	1 10	1 10	1 10	1 10	0 0	36
http://www.isoc.org/	3 1	4 8	2 1	4 8	— —	3,25 4,5	0,96 4,04	35
http://www.acm.org/	4 6	3 3	— —	— —	1 3	2,67 4	1,53 1,73	34
http://www. ft-informatik.de/	— —	4 8	— —	— —	— —	4 8	– –	33
http://www.ub. uni-dortmund.de/ literatursuche/index.htm	1 2	3 1	4 2	2 7	— —	2,5 3	1,29 2,71	32
http://www.uni-dortmund. de/UniDo/Personal/	4 9	3 9	2 13	— —	4 16	3,25 11,75	0,96 3,4	31
http://dekanat.cs. uni-dortmund.de/ Ehemalige/Beitritt.html	1 1	1 1	3 1	1 1	1 1	1,4 1	0,89 0	30
http://dekanat.cs. uni-dortmund.de/ZPA/ OeffnungszeitenZPA.shtml	4 2	1 5	1 2	4 5	4 6	2,8 4	1,64 1,87	29
http://fsinfo.cs. uni-dortmund.de/	2 3	1 9	1 4	1 4	3 1	1,6 4,2	0,89 2,95	28
http://dekanat.cs. uni-dortmund.de/HaPra/ index.html	4 10	4 3	4 2	1 1	4 2	3,4 3,6	1,34 3,65	27
http://ls10-www.cs. uni-dortmund.de/LS10/ Pages/sopra-specials/ anmeldung/sopra.shtml	— —	4 2	— —	4 2	1 4	3 2,67	1,73 1,15	26
http://stl-www.cs. uni-dortmund.de/	4 7	— —	— —	1 2	1 2	2 3,67	1,73 2,89	25
http://dekanat.cs. uni-dortmund.de/	— —	— —	— —	— —	— —	— —	– –	24
http://ls12-www.cs. uni-dortmund.de/	3 7	— —	— —	— —	— —	3 7	– –	23
http://ls6-www.cs. uni-dortmund.de/	— —	1 1	1 1	— —	— —	1 1	0 0	22
	1	2	3	4	5	μ	σ	Nr

Ziel	1	2	3	4	5	μ	σ	Nr
http://ls4-www.cs.uni-dortmund.de/	1 5	3 4	1 4	—	—	1,67 4,33	1,15 0,58	21
http://ls2-www.cs.uni-dortmund.de/	2 7	4 12	—	—	—	3 9,5	1,41 3,54	20
http://ls1-www.cs.uni-dortmund.de/	—	4 7	—	—	4 6	4 6,5	0 0,71	19
http://www.co.umist.ac.uk/dsd2002/	1 8	2 3	1 9	1 9	1 6	1,2 7	0,45 2,55	18
http://fsinfo.cs.uni-dortmund.de/Studium/	1 1	1 1	1 1	1 1	1 1	1 1	0 0	17
http://dekanat.cs.uni-dortmund.de/Studierende/Index.html	—	—	3 24	3 24	—	3 24	0 0	16
http://ls7-www.cs.uni-dortmund.de/VKInf/vorkurs_ws0102.shtml	1 5	1 2	4 10	—	4 10	2,5 6,75	1,73 3,95	15
http://www.uni-dortmund.de/TOP/	1 6	—	2 9	3 7	4 33	2,5 13,75	1,29 12,89	14
http://www-ai.cs.uni-dortmund.de/FORSCHUNG/REPORTS/reports.eng.html	1 1	1 1	1 1	1 1	1 1	1 1	0 0	13
http://www-ai.cs.uni-dortmund.de/FORSCHUNG/buecher.html	4 9	4 10	3 9	3 3	—	3,5 7,75	0,58 3,2	12
http://www-ai.cs.uni-dortmund.de/FORSCHUNG/PROJEKTE/SFB475C4/	1 1	1 1	4 5	—	1 1	1,75 2	1,5 2	11
http://www-ai.cs.uni-dortmund.de/FORSCHUNG/PROJEKTE/SFB475A4/	1 1	1 1	1 1	1 1	1 1	1 1	0 0	10
http://www.eunite.org	1 10	1 10	1 10	1 10	1 10	1 10	0 0	09
http://www-ai.cs.uni-dortmund.de/FORSCHUNG/MLTXT/mltxt.eng.html	1 2	4 4	4 6	1 2	—	2,5 3,5	1,73 1,91	08
http://www-ai.cs.uni-dortmund.de/FORSCHUNG/index.eng.html	1 2	—	1 3	1 1	1 1	1 1,75	0 0,96	07
	1	2	3	4	5	μ	σ	Nr

Ziel	1	2	3	4	5	μ	σ	Nr
http://www-ai.cs.uni-dortmund.de/FORSCHUNG/index.html	1 1	1 1	1 1	4 2	1 1	1,6 1,2	1,34 0,45	06
http://www-ai.cs.uni-dortmund.de/LEHRE/lehrveranstaltungen_alt.html	1 1	1 1	1 1	1 1	1 1	1 1	0 0	05
http://www-ai.cs.uni-dortmund.de/UNIVERSELL/index.eng.html	—	5 22	1 1	1 1	1 1	2 6,25	2 10,5	04
http://www-ai.cs.uni-dortmund.de/UNIVERSELL/	1 1	—	1 1	1 1	1 1	1 1	0 0	03
http://www-ai.cs.uni-dortmund.de/logo.html	1 1	1 1	1 1	1 1	1 3	1 1,4	0 0,89	02
http://www-ai.cs.uni-dortmund.de	—	—	—	—	3 2	3 2	– –	01
Gesamt gefunden	**36**	**36**	**36**	**34**	**36**	**35,6**	**0,89**	
ohne Spezialisieren	**35**	36	36	34	**35**	**35,2**	**0,84**	

B.1.2 Zufällige Wörter mit Stoppwortelimination

Die hier vorgestellten Ergebnisse stammen aus Testläufen vom 28. Juli 2003.

Ziel	1	2	3	4	5	μ	σ	Nr
http://www-ai.cs.uni-dortmund.de	4 9	1 2	5 0	1 2	1 2	2,4 3	1,95 3,46	62
http://www-ai.cs.uni-dortmund.de	1 1	1 2	1 2	—	1 2	1 1,75	0 0,5	61
http://sfbci.cs.uni-dortmund.de/home/German/frameset.html	1 3	1 9	4 9	1 5	—	1,75 6,5	1,5 3	60
http://www-ai.cs.uni-dortmund.de/LEHRE/PROMOTION/promotion_fertig.html	4 9	1 1	1 1	1 1	1 1	1,6 2,6	1,34 3,58	59
http://www-ai.cs.uni-dortmund.de/LEHRE/PG/	1 2	1 2	4 6	5 0	1 2	2,4 2,4	1,95 2,19	58
	1	**2**	**3**	**4**	**5**	μ	σ	**Nr**

Ziel	1	2	3	4	5	μ	σ	Nr
http://www-ai.cs.uni-dortmund.de/LEHRE/DIPLOM/OFFEN/	1 1	1 3	4 16	1 1	1 1	1,6 4,4	1,34 6,54	57
http://www-ai.cs.uni-dortmund.de/LEHRE/DIPLOM/diplom_fertig.html	6 8	2 2	3 2	1 2	—	3 3,5	2,16 3	56
http://www-ai.cs.uni-dortmund.de/LEHRE/VORLESUNGEN/MLRN/ml.html	3 4	1 1	1 1	1 1	1 3	1,4 2	0,89 1,41	55
http://www-ai.cs.uni-dortmund.de/LEHRE/SEMINARE/INFORMATIONSEXTRAKTION/informationsExtraktion.html	1 1	1 1	1 1	1 1	1 1	1 1	0 0	54
http://www.rni.helsinki.fi/kurssit/	4 7	5 0	2 1	3 1	5 0	3,8 1,8	1,3 2,95	53
http://www.cs.bris.ac.uk/Teaching/Resources/COMS30106/	1 6	1 3	1 8	1 5	1 7	1 5,8	0 1,92	52
http://www.csd.abdn.ac.uk/~pedwards/teaching/CS5505/slides.html	1 1	4 1	5 13	5 0	4 2	3,8 3,4	1,64 5,41	51
http://www.aifb.uni-karlsruhe.de/Lehrangebot/Sommer2002/InfoC/script.html	2 1	1 1	1 1	1 1	6 17	2,4 4,2	2,61 7,16	50
http://www.fi.muni.cz/~popel/old.html.utf-8#courses	1 6	4 9	1 6	1 1	1 6	1,6 5,6	1,34 2,88	49
http://www.cs.kuleuven.ac.be/~hendrik/ML	1 8	1 1	1 8	1 1	1 10	1 5,6	0 4,28	48
http://www-ai.cs.uni-dortmund.de/LEHRE/DIPLOM/diplom_fertig.html	5 0	4 4	—	5 0	5 0	4,75 1	0,5 2	47
http://www-ai.cs.uni-dortmund.de/SOFTWARE/IL/	1 3	1 8	2 8	—	—	1,33 6,33	0,58 2,89	46
http://www-ai.cs.uni-dortmund.de/SOFTWARE/SVM_LIGHT/	1 10	—	4 8	1 7	1 6	1,75 7,75	1,5 1,71	45
	1	**2**	**3**	**4**	**5**	μ	σ	**Nr**

Ziel	1	2	3	4	5	μ	σ	Nr
http://www-ai.cs. uni-dortmund.de/index. eng.html	5 12	6 11	—	6 9	6 11	5,75 10,75	0,5 1,26	44
http://www-ai.cs. uni-dortmund.de/ PERSONAL/personal.html	5 0	3 1	5 11	3 1	4 3	4 3,2	1 4,49	43
http://www-ai.cs. uni-dortmund.de/LEHRE/ lehre_15_02_2002.html	5 0	1 3	1 2	5 0	1 1	2,6 1,2	2,19 1,3	42
http://www-ai.cs. uni-dortmund.de/ FORSCHUNG/	5 0	3 1	1 1	—	1 1	2,5 0,75	1,91 0,5	41
http://www-ai.cs. uni-dortmund.de/ SOFTWARE/MYSVM/	—	5 0	5 0	1 2	1 6	3 2	2,31 2,83	40
http://www-ai.cs. uni-dortmund.de/ PERSONAL/morik.html	1 1	1 1	1 1	1 1	1 1	1 1	0 0	39
http://www-ai.cs. uni-dortmund.de/Harvest/ brokers/www-ai/query. html	1 2	3 21	1 1	1 1	2 2	1,6 5,4	0,89 8,73	38
http://www-ai.cs. uni-dortmund.de/ SOFTWARE/	—	1 2	1 2	1 2	5 3	2 2,25	2 0,5	37
http://www.dfn.de/home. html	3 6	1 7	3 9	4 20	1 9	2,4 10,2	1,34 5,63	36
http://www.isoc.org/	3 1	1 9	1 10	3 6	1 6	1,8 6,4	1,1 3,51	35
http://www.acm.org/	4 10	—	1 9	5 9	—	3,33 9,33	2,08 0,58	34
http://www. ft-informatik.de/	—	4 18	—	—	3 1	3,5 9,5	0,71 12,02	33
http://www.ub. uni-dortmund.de/ literatursuche/index.htm	4 10	5 0	1 5	5 0	4 10	3,8 5	1,64 5	32
http://www.uni-dortmund. de/UniDo/Personal/	—	—	—	—	—	—	—	31
http://dekanat.cs. uni-dortmund.de/ Ehemalige/Beitritt.html	1 1	5 8	1 1	1 1	1 1	1,8 2,4	1,79 3,13	30
http://dekanat.cs. uni-dortmund.de/ZPA/ OeffnungszeitenZPA.shtml	1 4	4 10	3 7	3 3	4 9	3 6,6	1,22 3,05	29
	1	**2**	**3**	**4**	**5**	μ	σ	**Nr**

Ziel	1	2	3	4	5	μ	σ	Nr
http://fsinfo.cs.uni-dortmund.de/	2 2	5 1	1 10	3 10	2 2	2,6 5	1,52 4,58	28
http://dekanat.cs.uni-dortmund.de/HaPra/index.html	1 5	4 20	6 25	4 11	4 8	3,8 13,8	1,79 8,41	27
http://ls10-www.cs.uni-dortmund.de/LS10/Pages/sopra-specials/anmeldung/sopra.shtml	3 9	6 22	4 11	4 15	4 10	4,4 13,4	1,52 5,32	26
http://stl-www.cs.uni-dortmund.de/	3 4	1 2	3 10	—	1 2	2 4,5	1,15 3,79	25
http://dekanat.cs.uni-dortmund.de/	—	—	—	—	5 2	5 2	– –	24
http://ls12-www.cs.uni-dortmund.de/	5 8	—	—	—	4 10	4,5 9	0,71 1,41	23
http://ls6-www.cs.uni-dortmund.de/	—	5 0	—	1 1	3 8	3 3	2 4,36	22
http://ls4-www.cs.uni-dortmund.de/	—	—	4 9	—	—	4 9	– –	21
http://ls2-www.cs.uni-dortmund.de/	—	2 7	1 2	1 2	1 3	1,25 3,5	0,5 2,38	20
http://ls1-www.cs.uni-dortmund.de/	5 0	—	4 10	4 9	4 10	4,25 7,25	0,5 4,86	19
http://www.co.umist.ac.uk/dsd2002/	1 10	1 10	1 9	1 9	1 6	1 8,8	0 1,64	18
http://fsinfo.cs.uni-dortmund.de/Studium/	1 1	1 1	1 2	1 1	1 9	1 2,8	0 3,49	17
http://dekanat.cs.uni-dortmund.de/Studierende/Index.html	6 32	6 46	5 36	6 42	6 50	6 41,2	0,71 7,29	16
http://ls7-www.cs.uni-dortmund.de/VKInf/vorkurs_ws0102.shtml	—	1 2	1 6	3 2	1 10	1,5 5	1 3,83	15
http://www.uni-dortmund.de/TOP/	—	3 19	—	—	1 10	2 14,5	1,41 6,36	14
http://www-ai.cs.uni-dortmund.de/FORSCHUNG/REPORTS/reports.eng.html	1 1	1 1	1 1	1 1	1 1	1 1	0 0	13
http://www-ai.cs.uni-dortmund.de/FORSCHUNG/buecher.html	3 8	3 10	1 5	4 2	5 0	3,2 5	1,48 4,12	12
	1	2	3	4	5	μ	σ	Nr

Ziel	1	2	3	4	5	μ	σ	Nr
http://www-ai.cs.uni-dortmund.de/FORSCHUNG/PROJEKTE/SFB475C4/	1 3	1 1	5 0	1 1	1 1	1,8 1,2	1,79 1,1	11
http://www-ai.cs.uni-dortmund.de/FORSCHUNG/PROJEKTE/SFB475A4/	1 1	1 1	1 1	—	11 1	3,5 1	5 0	10
http://www.eunite.org	1 10	1 10	1 10	1 10	1 10	1 10	0 0	09
http://www-ai.cs.uni-dortmund.de/FORSCHUNG/MLTXT/mltxt.eng.html	1 2	1 2	4 2	1 4	4 6	2,2 3,2	1,64 1,79	08
http://www-ai.cs.uni-dortmund.de/FORSCHUNG/index.eng.html	1 3	1 2	1 1	1 1	1 5	1 2,4	0 1,67	07
http://www-ai.cs.uni-dortmund.de/FORSCHUNG/index.html	6 7	1 1	1 1	—	6 12	3,5 5,25	2,89 5,32	06
http://www-ai.cs.uni-dortmund.de/LEHRE/lehrveranstaltungen_alt.html	1 1	—	5 0	—	1 1	2,33 0,67	2,31 0,58	05
http://www-ai.cs.uni-dortmund.de/UNIVERSELL/index.eng.html	1 5	1 1	5 15	1 1	1 1	1,8 4,6	1,79 6,07	04
http://www-ai.cs.uni-dortmund.de/UNIVERSELL/	—	1 1	1 1	1 1	1 1	1 1	0 0	03
http://www-ai.cs.uni-dortmund.de/logo.html	1 1	4 1	1 1	3 1	1 1	2 1	1,41 0	02
http://www-ai.cs.uni-dortmund.de	5 0	—	—	—	5 0	5 0	0 0	01
Gesamt gefunden	**31**	**32**	**33**	**34**	**34**	**32,8**	**1,3**	
ohne Spezialisieren	**30**	**31**	**32**	**33**	34	**32,0**	**1,58**	

B.1.3 Zufällige Wörter / Häufigkeitsverteilt

Die hier vorgestellten Ergebnisse stammen aus Testläufen vom 27. Juli 2003.

Ziel	1	2	3	4	5	μ	σ	Nr
http://www-ai.cs.uni-dortmund.de	1 2	6 19	1 1	5 0	1 7	2,8 5,8	2,49 7,85	62
http://www-ai.cs.uni-dortmund.de	1 2	1 1	1 2	6 25	1 2	2 6,4	2,24 10,41	61
http://sfbci.cs.uni-dortmund.de/home/German/frameset.html	4 19	1 9	1 7	4 28	1 1	2,2 12,8	1,64 10,69	60
http://www-ai.cs.uni-dortmund.de/LEHRE/PROMOTION/promotion_fertig.html	1 1	1 9	1 2	1 1	4 9	1,6 4,4	1,34 4,22	59
http://www-ai.cs.uni-dortmund.de/LEHRE/PG/	1 2	1 2	1 8	1 3	1 3	1 3,6	0 2,51	58
http://www-ai.cs.uni-dortmund.de/LEHRE/DIPLOM/OFFEN/	1 1	1 4	1 9	1 3	1 9	1 5,2	0 3,63	57
http://www-ai.cs.uni-dortmund.de/LEHRE/DIPLOM/diplom_fertig.html	1 1	1 1	1 1	1 1	1 1	1 1	0 0	56
http://www-ai.cs.uni-dortmund.de/LEHRE/VORLESUNGEN/MLRN/ml.html	1 1	1 3	1 1	1 1	1 1	1 1,4	0 0,89	55
http://www-ai.cs.uni-dortmund.de/LEHRE/SEMINARE/INFORMATIONSEXTRAKTION/informationsExtraktion.html	1 1	1 1	1 7	1 1	1 3	1 2,6	0 2,61	54
http://www.rni.helsinki.fi/kurssit/	5 0	5 0	1 1	4 6	5 0	4 1,4	1,73 2,61	53
http://www.cs.bris.ac.uk/Teaching/Resources/COMS30106/	1 1	1 6	1 1	1 10	1 10	1 5,6	0 4,51	52
http://www.csd.abdn.ac.uk/~pedwards/teaching/CS5505/slides.html	5 0	1 4	4 1	1 1	5 0	3,2 1,2	2,05 1,64	51
	1	2	3	4	5	μ	σ	Nr

Ziel	1	2	3	4	5	μ	σ	Nr
http://www.aifb.uni-karlsruhe.de/Lehrangebot/Sommer2002/InfoC/script.html	3 10	1 8	2 11	4 28	3 2	2,6 11,8	1,14 9,71	50
http://www.fi.muni.cz/~popel/old.html.utf-8#courses	1 7	1 7	1 7	1 9	1 9	1 7,8	0 1,1	49
http://www.cs.kuleuven.ac.be/~hendrik/ML	1 10	1 10	5 0	1 9	1 9	1,8 7,6	1,79 4,28	48
http://www-ai.cs.uni-dortmund.de/LEHRE/DIPLOM/diplom_fertig.html	1 1	1 1	1 2	1 1	1 1	1 1,2	0 0,45	47
http://www-ai.cs.uni-dortmund.de/SOFTWARE/IL/	1 2	1 2	1 2	1 2	1 3	1 2,2	0 0,45	46
http://www-ai.cs.uni-dortmund.de/SOFTWARE/SVM_LIGHT/	4 23	1 5	2 9	4 23	1 7	2,4 13,4	1,52 8,88	45
http://www-ai.cs.uni-dortmund.de/index.eng.html	5 22	5 10	5 17	2 17	6 37	4,6 20,6	1,52 10,11	44
http://www-ai.cs.uni-dortmund.de/PERSONAL/personal.html	6 13	4 10	5 2	4 9	1 2	4 7,2	1,87 4,97	43
http://www-ai.cs.uni-dortmund.de/LEHRE/lehre_15_02_2002.html	1 7	1 1	1 1	2 18	1 2	1,2 5,8	0,45 7,26	42
http://www-ai.cs.uni-dortmund.de/FORSCHUNG/	4 5	1 4	5 0	1 1	1 1	2,4 2,2	1,95 2,17	41
http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/	1 9	1 4	1 7	—	1 5	1 6,25	0 2,22	40
http://www-ai.cs.uni-dortmund.de/PERSONAL/morik.html	1 1	1 10	1 1	1 2	1 3	1 3,4	0 3,78	39
http://www-ai.cs.uni-dortmund.de/Harvest/brokers/www-ai/query.html	1 1	1 2	1 2	1 2	5 0	1,8 1,4	1,79 0,89	38
	1	**2**	**3**	**4**	**5**	μ	σ	**Nr**

Ziel	1	2	3	4	5	μ	σ	Nr
http://www-ai.cs.uni-dortmund.de/SOFTWARE/	1 2	1 2	1 2	5 15	3 4	2,2 5	1,79 5,66	37
http://www.dfn.de/home.html	1 9	1 10	1 9	1 10	1 8	1 9,2	0 0,84	36
http://www.isoc.org/	4 10	1 2	3 9	1 6	1 4	2 6,2	1,41 3,35	35
http://www.acm.org/	1 10	1 8	2 20	3 9	1 6	1,6 10,6	0,89 5,46	34
http://www.ft-informatik.de/	4 22	1 10	4 8	3 13	4 4	3,2 11,4	1,3 6,77	33
http://www.ub.uni-dortmund.de/literatursuche/index.htm	4 9	4 10	5 0	2 4	2 3	3,4 5,2	1,34 4,21	32
http://www.uni-dortmund.de/UniDo/Personal/	5 0	5 0	1 9	1 8	4 10	3,2 5,4	2,05 4,98	31
http://dekanat.cs.uni-dortmund.de/Ehemalige/Beitritt.html	1 1	1 6	4 6	1 7	2 16	1,8 7,2	1,3 5,45	30
http://dekanat.cs.uni-dortmund.de/ZPA/OeffnungszeitenZPA.shtml	1 1	6 1	3 3	4 11	1 1	3 3,4	2,12 4,34	29
http://fsinfo.cs.uni-dortmund.de/	1 10	1 1	1 5	1 4	1 1	1 4,2	0 3,7	28
http://dekanat.cs.uni-dortmund.de/HaPra/index.html	1 2	1 1	4 3	5 6	6 29	3,4 8,2	2,3 11,78	27
http://ls10-www.cs.uni-dortmund.de/LS10/Pages/sopra-specials/anmeldung/sopra.shtml	1 13	1 8	4 9	5 4	6 9	3,4 8,6	2,3 3,21	26
http://stl-www.cs.uni-dortmund.de/	4 2	6 5	4 8	1 2	1 6	3,2 4,6	2,17 2,61	25
http://dekanat.cs.uni-dortmund.de/	1 7	1 0	1 9	4 9	1 0	1,6 5	1,34 4,64	24
http://ls12-www.cs.uni-dortmund.de/	1 8	5 10	1 0	3 28	5 11	3 11,4	2 10,24	23
http://ls6-www.cs.uni-dortmund.de/	2 10	4 2	5 0	3 2	4 0	3,6 2,8	1,14 4,15	22
http://ls4-www.cs.uni-dortmund.de/	3 5	1 10	5 8	1 7	5 3	3 6,6	2 2,7	21
	1	**2**	**3**	**4**	**5**	μ	σ	**Nr**

Ziel	1	2	3	4	5	μ	σ	Nr
http://ls2-www.cs.uni-dortmund.de/	1 10	1 12	1 5	1 20	1 3	1 10	0 6,67	20
http://ls1-www.cs.uni-dortmund.de/	1 10	3 10	1 0	4 8	1 5	2 6,6	1,41 4,22	19
http://www.co.umist.ac.uk/dsd2002/	4 8	1 9	5 10	3 8	5 9	3,6 8,8	1,67 0,84	18
http://fsinfo.cs.uni-dortmund.de/Studium/	1 10	1 1	1 5	1 1	1 1	1 3,6	0 3,97	17
http://dekanat.cs.uni-dortmund.de/Studierende/Index.html	6 50	1 31	1 31	1 10	1 9	2 26,2	2,24 17,11	16
http://ls7-www.cs.uni-dortmund.de/VKInf/vorkurs_ws0102.shtml	6 3	4 10	4 4	1 4	1 4	3,2 5	2,17 2,83	15
http://www.uni-dortmund.de/TOP/	2 8	1 10	1 9	1 10	1 6	1,2 8,6	0,45 1,67	14
http://www-ai.cs.uni-dortmund.de/FORSCHUNG/REPORTS/reports.eng.html	1 1	1 1	1 1	1 9	1 2	1 2,8	0 3,49	13
http://www-ai.cs.uni-dortmund.de/FORSCHUNG/buecher.html	1 10	1 3	1 0	1 3	1 10	1 5,2	0 4,55	12
http://www-ai.cs.uni-dortmund.de/FORSCHUNG/PROJEKTE/SFB475C4/	1 1	1 1	5 10	1 1	4 1	2,4 2,8	1,95 4,02	11
http://www-ai.cs.uni-dortmund.de/FORSCHUNG/PROJEKTE/SFB475A4/	1 4	1 6	1 5	1 1	3 1	1,4 3,4	0,89 2,3	10
http://www.eunite.org	5 10	1 10	5 10	1 10	1 10	2,6 10	2,19 0	09
http://www-ai.cs.uni-dortmund.de/FORSCHUNG/MLTXT/mltxt.eng.html	1 4	1 10	1 2	1 2	1 4	1 4,4	0 3,29	08
http://www-ai.cs.uni-dortmund.de/FORSCHUNG/index.eng.html	1 5	4 1	1 9	1 1	1 1	1,6 3,4	1,34 3,58	07
http://www-ai.cs.uni-dortmund.de/FORSCHUNG/index.html	1 1	1 2	1 2	2 1	1 1	1,2 1,4	0,45 0,55	06
	1	**2**	**3**	**4**	**5**	μ	σ	**Nr**

Ziel	**1**	**2**	**3**	**4**	**5**	μ	σ	**Nr**
`http://www-ai.cs.uni-dortmund.de/LEHRE/lehrveranstaltungen_alt.html`	1 1	3 1	1 10	1 9	1 1	1,4 4,4	0,89 4,67	05
`http://www-ai.cs.uni-dortmund.de/UNIVERSELL/index.eng.html`	1 1	1 1	1 1	6 6	1 1	2 2	2,24 2,24	04
`http://www-ai.cs.uni-dortmund.de/UNIVERSELL/`	1 1	1 10	1 1	1 10	1 9	1 6,2	0 4,76	03
`http://www-ai.cs.uni-dortmund.de/logo.html`	1 1	1 5	1 1	3 10	1 2	1,4 3,8	0,89 3,83	02
`http://www-ai.cs.uni-dortmund.de`	1 4	1 10	1 10	1 16	1 8	1 9,6	0 4,34	01
Gesamt gefunden	**39**	**34**	**37**	**38**	**39**	**37,4**	**2,07**	
ohne Spezialisieren	**38**	**33**	37	**37**	39	**36,8**	**2,28**	

B.2 Tabellarische Zusammenfassung – WDR

1. Großbuchstaben
2. Großbuchstaben + einfache Stoppwortelimination
3. Großbuchstaben + Stoppwortelimination
4. Häufige Wörter
5. Häufige Wörter + TF–IDF
6. Phrase
7. Längste Phrase
8. Satzphrase
9. verbesserte Satzphrase
10. Namen von Personen
11. Zufällige Wörter
12. Zufällige Wörter + Stoppwortelimination
13. Zufällige Wörter / Häufigkeitsverteilt

Die hier vorgestellten Ergebnisse stammen aus Testläufen vom 16.-17. August 2003.

URL	1	2	3	4	5	6	7	8	9	10	11	12	13	Nr
http://www.wdr.de/themen/politik/nrw/steinkohle/rag_ausgliederung.jhtml	1 3	1 3	1 3	1 3	1 9	4 0	1 10	1 3	1 7	1 10	2 7,2	1,6 6,6	1 5,4	01
http://www.wdr.de/themen/politik/international/elfter_september/verfassungsschutz_cia/index.jhtml	1 2	1 1	1 7	1 2	1 2	4 10	1 2	1 10	1 4	1 1	1,6 2,6	1,2 1	2 13	02
http://www.wdr.de/themen/politik/deutschland/stabilitaetspakt/index.jhtml	1 1	1 3	1 2	1 2	1 7	4 10	1 2	2 2	1 5	1 3	1,8 8,4	2,6 7	1,4 9,2	03
http://www.wdr.de/themen/politik/international/soldaten_afghanistan/landung_koeln.jhtml	1 2	1 2	1 1	3 3	1 4	4 11	1 4	2 4	1 4	4 9	1,4 4	1,2 5,4	1,4 7,4	04
http://www.wdr.de/themen/politik/nrw/interview_2003/pinkwart/interview.jhtml	1 3	1 3	1 3	1 4	1 3	1 3	1 3	2 2	1 4	1 10	2 9	1,6 4,2	1 6,6	05
http://www.wdr.de/themen/politik/nrw/gemeindefinanzierung/reaktionen.jhtml	1 4	1 4	1 2	2 4	1 2	4 0	1 4	2 4	1 4	1 3	1 2,8	1,8 3,6	1 7,8	06
	1	2	3	4	5	6	7	8	9	10	11	12	13	Nr

URL	1	2	3	4	5	6	7	8	9	10	11	12	13	Nr
http://www.wdr.de/themen/politik/nrw/gemeindefinanzierung/ruettgers.jhtml	1 6	1 2	1 6	4 9	2 4	4 0	1 6	5 6	1 6	1 8	1,6 6,6	2 5,6	2,6 12,2	07
http://www.wdr.de/themen/politik/nrw/moellemann/inhalt.jhtml	1 5	1 5	1 9	1 10	1 10	4 10	1 9	6 0	1 8	3 10	2 3,6	2 4,2	1,2 7,8	08
http://www.wdr.de/themen/politik/nrw/muellaffaere_spd/inhalt.jhtml	1 7	1 7	1 9	1 10	1 10	4 9	5 40	6 0	1 10	4 0	1,8 5,6	2,2 5,8	2,2 7,6	09
http://www.wdr.de/themen/homepages/irak.jhtml	1 10	1 4	1 5	1 6	1 10	4 0	1 4	6 4	1 10	1 10	1 3	2,4 9,4	1,8 6	10
http://www.wdr.de/themen/homepages/d_usa.jhtml	1 29	1 10	1 10	1 10	1 10	4 10	1 5	1 5	1 5	1 8	1 2,8	2,2 4,8	1,8 7	11
http://www.wdr.de/themen/wirtschaft/arbeit_und_tarifwesen/hartz_reformen/index.jhtml	1 1	1 2	1 1	1 3	1 2	1 2	1 2	1 2	1 5	1 2	2,2 8	2,4 17	2,6 6,6	12
http://www.wdr.de/themen/wirtschaft/1/haushaltsgeraete/index.jhtml	1 1	1 3	1 1	1 4	1 7	4 0	1 4	4 4	1 6	1 0	3 7,4	2 8	3 8,6	13
http://www.wdr.de/themen/wirtschaft/wirtschaftsbranche/ford/krise.jhtml	1 2	1 2	1 2	4 10	1 7	1 7	1 7	1 7	1 8	1 7	2 7,4	3,2 5,8	2,4 16,8	14
http://www.wdr.de/themen/wirtschaft/wirtschaftsbranche/rwe/usa_drohen.jhtml	1 3	1 3	1 2	1 3	1 4	4 10	1 5	2 4	1 4	1 3	1,6 3,6	1,8 2	1,6 8,6	15
http://www.wdr.de/themen/wirtschaft/arbeit_und_tarifwesen/ich_ag/erfolg.jhtml	1 2	1 2	1 2	1 2	1 3	3 4	1 4	3 4	1 4	1 4	1,8 6,6	4 11	1,8 7,8	16
http://www.wdr.de/themen/wirtschaft/geld-_und_kreditwesen/westlb/index_030806.jhtml	1 2	1 2	1 2	1 8	1 7	5 12	1 3	1 7	1 8	3 0	3,6 10	2,4 10,4	3 14,6	17
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**Nr**

URL	1	2	3	4	5	6	7	8	9	10	11	12	13	Nr
http://www.wdr.de/themen/wirtschaft/wirtschaftsbranche/luftfahrt/flughafen_weeze.jhtml	1 1	1 1	1 1	1 10	1 5	1 3	1 4	1 3	1 3	3 4	2 6,4	3,2 6,8	1,6 7	18
http://www.wdr.de/online/jobs/jobzeit/index.phtml?rubrikenstyle=wirtschaft	1 3	1 3	1 3	1 3	1 3	1 3	1 3	1 3	1 2	3 0	1,6 6,4	2,4 11,8	1,2 6,2	19
http://www.wdr.de/themen/forschung/technik/solarflitzer/index.jhtml	1 2	1 3	1 3	1 3	1 3	1 1	1 3	1 1	1 5	2 3	2 3,8	1,4 2,8	1 4,2	20
http://www.wdr.de/themen/forschung/interdisziplinaer/virtuelle_bibliothek/index.jhtml	1 4	1 1	1 6	1 6	1 6	1 6	1 6	1 6	1 8	2 6	1 5,8	2,6 7,4	1 6,2	21
http://www.wdr.de/themen/homepages/kleine_anfrage.jhtml	6 40	6 24	1 2	1 10	1 10	4 7	1 9	6 0	1 9	2 9	2,6 6	2,6 6,6	1 3,4	22
http://www.wdr.de/themen/kultur/1/linkshaender_tag_2003/gaestebuch.jhtml	3 28	4 23	1 2	1 10	1 10	2 3	1 3	1 3	1 4	4 0	2 2,8	1 3,6	1 9,4	23
http://www.wdr.de/themen/kultur/rundfunk/lilipuz_sommertour2003/index.jhtml	1 2	1 2	1 2	2 2	1 3	4 8	1 3	4 3	1 2	1 10	3,4 7,6	4,2 15,8	1,4 7,2	24
http://www.wdr.de/themen/kultur/bildung_und_erziehung/bafoeg_missbrauch/index.jhtml	1 2	1 4	1 2	4 10	1 3	3 8	1 6	1 8	1 8	1 10	4,6 14,8	3 7,6	2,6 13,4	25
http://www.wdr.de/themen/kultur/personen/ruge/index.jhtml	1 2	1 2	1 2	3 8	1 3	4 3	1 3	3 3	1 3	1 3	1 1,8	1,8 6	1 5,2	26
http://www.wdr.de/themen/kultur/quiz/quiz_rechtschreibreform.jhtml	1 2	1 2	1 2	3 3	1 10	1 10	1 6	1 10	1 6	1 10	1 3,4	1,8 4	2 7,4	27
http://www.wdr.de/themen/panorama/lifestyle/modemesse_reevolutions_2003/sommer.jhtml	6 7	5 0	1 4	1 6	3 5	3 7	1 5	3 3	1 4	4 0	1,4 5,6	1,6 3,8	1 3,2	28
	1	2	3	4	5	6	7	8	9	10	11	12	13	Nr

URL	1	2	3	4	5	6	7	8	9	10	11	12	13	Nr
http://www.wdr.de/themen/kultur/1/kinderschutzbund/interview.jhtml	1 5	1 1	1 9	1 7	1 7	3 10	1 7	2 9	1 7	1 7	1 4,8	1,2 4,4	1,6 6,8	29
http://www.wdr.de/themen/kultur/netzkultur/heimatinseln/inseln.jhtml	1 3	1 2	1 3	1 3	1 3	2 5	1 3	3 4	1 2	2 7	1 4,4	2,2 3,6	1,6 4,2	30
http://www.wdr.de/themen/homepages/popkomm2003.jhtml	1 10	4 14	1 10	1 1	1 5	4 10	1 10	2 9	1 9	4 0	2,6 8,2	5,8 14,4	3,4 12	31
http://www.wdr.de/themen/homepages/unicef.jhtml	1 10	1 10	1 10	1 10	1 3	4 10	1 10	2 3	1 3	1 3	1,8 0,8	3,6 8,2	2,4 14,6	32
http://www.wdr.de/themen/kultur/netzkultur/nrw_privat/wohnungen.jhtml	1 8	1 8	1 9	1 8	1 8	6 0	1 10	2 10	1 10	1 10	1 8,4	1 9	1,2 10,4	33
http://www.wdr.de/themen/computer/internet/blaster/index.jhtml	1 1	1 2	1 2	1 4	1 10	1 2	1 2	1 2	1 5	1 10	2,2 7	1,4 4,8	1,6 8,2	34
http://www.wdr.de/themen/computer/software/virenticker/index.jhtml	1 10	3 10	1 1	1 10	1 10	1 5	1 5	2 5	1 6	4 0	1,8 2,6	2 3,4	1 7,4	35
http://www.wdr.de/themen/computer/schiebwoche/2003/index_33.jhtml	1 5	1 3	1 4	1 4	1 4	3 4	1 4	2 6	1 4	1 10	1 4	3 10,8	2,4 12,2	36
http://www.wdr.de/themen/computer/internet/sicherheit/exploit.jhtml	1 2	1 3	1 1	1 3	1 3	3 2	1 3	1 3	1 3	1 2	1 4	2,2 6,6	1 5,4	37
http://www.wdr.de/themen/computer/internet/barrierefreies_internet/internetcafe_reportage.jhtml	1 10	1 5	1 5	1 5	1 5	1 5	1 3	1 5	1 4	1 4	1 4,6	2,6 9,6	1,2 5	38
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**Nr**

URL	1	2	3	4	5	6	7	8	9	10	11	12	13	Nr
http://www.wdr.de/themen/computer/angeklickt/tagestipp/tagestipp.jhtml	3 14	5 0	4 10	4 8	4 9	4 0	4 3	6 0	6 0	1 0	2,6 1,8	4,4 13	4,8 17,2	39
http://www.wdr.de/themen/computer/angeklickt/webtv/index.jhtml	1 10	1 10	1 10	1 10	—	4 10	1 10	3 10	1 10	3 0	1 5	2,4 9,4	1 10	40
http://www.wdr.de/themen/sport/1/holzfaeller/index.jhtml	5 0	5 0	1 10	1 5	1 5	3 5	1 5	2 5	1 4	1 4	3 6,4	1 3,2	2,2 13	41
http://www.wdr.de/themen/homepages/olympia_2012.jhtml	1 4	1 4	1 4	1 10	1 10	4 0	1 8	1 10	1 10	1 8	1,8 1,8	3,6 7	1,6 7	42
http://www.wdr.de/themen/panorama/2/lottojackpot_nrw/index.jhtml	1 1	5 0	1 1	1 2	1 2	4 0	1 2	3 3	1 5	1 2	1 1,2	1 2	1,8 9,4	43
http://www.wdr.de/themen/panorama/kriminalitaet02/gladbecker_geiseldrama_1988/index.jhtml	1 2	4 2	1 3	1 5	1 10	4 0	1 3	2 3	1 4	4 0	1,4 2,6	2,6 7	1 4,6	44
http://www.wdr.de/themen/panorama/2/stromausfall_bielefeld/index.jhtml	1 3	4 1	6 30	1 4	1 4	3 3	1 9	2 4	1 6	1 6	2,4 15,4	3,6 12,4	1,8 7,2	45
http://www.wdr.de/themen/panorama/kriminalitaet02/missbrauch_moenchengladbach/index.jhtml	1 9	5 0	1 2	4 10	4 9	6 0	1 4	3 4	1 4	1 10	2,4 5,4	3,8 11	4 16,6	46
http://www.wdr.de/themen/panorama/wetter/sommer_2003/abkuehlung_030814.jhtml	1 9	1 1	1 10	3 2	1 3	1 2	1 4	2 2	1 4	3 0	2,8 12,8	1 5	1 4,6	47
http://www.wdr.de/themen/panorama/wetter/sommer_2003/waldbraende.jhtml	1 1	1 2	1 1	4 4	1 1	5 0	1 1	4 1	1 3	1 0	1,2 2,2	3,2 8	3 6	48
	1	2	3	4	5	6	7	8	9	10	11	12	13	Nr

URL	1	2	3	4	5	6	7	8	9	10	11	12	13	Nr
http://www.wdr.de/themen/panorama/2/bombenentschaerfungen_nrw_2002/index.jhtml	1 1	5 0	3 2	4 4	3 2	4 0	1 10	1 1	4 5	5 15	2,6 13,4	1,6 5,6	3,2 15,2	49
http://www.wdr.de/themen/panorama/wetter/sommer_2003/hitze.jhtml	1 10	1 1	1 9	6 14	1 3	1 1	1 4	1 1	1 2	2 10	1,8 2,8	3,2 4	2,2 2	50
http://www.wdr.de/themen/gesundheit/1/antipasti/index.jhtml	1 3	1 3	1 3	1 8	1 8	4 2	1 8	2 8	1 8	1 0	1,8 3,8	1 3,6	2 12,4	51
http://www.wdr.de/themen/gesundheit/gesundheitswesen/gesetzliche_krankenkassen/sparzwang.jhtml	1 2	1 2	1 3	1 4	1 6	4 7	1 10	2 4	1 5	1 1	1,4 4,8	1,6 2,8	1,8 4,4	52
http://www.wdr.de/themen/gesundheit/gesundheitswesen/gesetzliche_krankenkassen/bkk_beitragserhoehung.jhtml	1 9	1 9	1 9	1 8	1 10	1 10	1 5	2 5	1 6	1 1	1 2	1 2,8	1 5,4	53
http://www.wdr.de/themen/freizeit/freizeitgestaltung/insel_sommer/monkeys_island.jhtml	1 4	1 3	1 4	1 4	1 4	1 2	1 4	1 2	1 5	1 3	1,8 6	1,6 5,4	1,6 7,4	54
http://www.wdr.de/themen/freizeit/freizeitgestaltung/future_parade/index.jhtml	1 3	1 3	3 5	1 5	1 5	4 5	1 5	4 5	2 5	1 3	3 14,4	2 4,4	1 8,6	55
http://www.wdr.de/themen/freizeit/freizeitgestaltung/mundorgel_50jahre/interview_guildo_horn.jhtml	1 3	1 3	1 4	1 4	1 4	1 7	1 4	1 7	1 5	3 0	2,4 13,6	3,6 13,8	1,4 5	56
	1	2	3	4	5	6	7	8	9	10	11	12	13	Nr

URL	1	2	3	4	5	6	7	8	9	10	11	12	13	Nr
http://www.wdr.de/themen/verkehr/strasse/tunnel_b236/index.jhtml	3 10	1 2	4 10	1 1	1 3	1 4	1 10	1 3	1 5	1 9	2 6,6	1,4 1,6	2,4 7,8	57
http://www.wdr.de/themen/verkehr/schiene/deutsche_bahn/maengel.jhtml	1 2	1 2	1 3	1 5	1 9	1 5	1 3	1 3	1 4	1 0	2,4 5,6	3 6,4	1 7,6	58
http://www.wdr.de/themen/verkehr/strasse/radarwarngeraete/autobahn_warnung.jhtml	1 2	1 3	1 3	1 3	1 4	1 3	1 3	4 7	1 4	1 10	2,4 11,8	2 4,6	1,6 6,2	59
http://www.wdr.de/themen/verkehr/wasser/niedrigwasser_rhein_ruhr/niedrigwasser_0308.jhtml	2 12	1 2	2 10	1 3	1 3	4 0	1 10	6 3	1 3	1 3	2 7,6	1,6 3,6	1 2,6	60
http://www.wdr.de/themen/panorama/2/verdaechtiger_koffer_koeln/index.jhtml	3 5	4 5	6 31	1 6	1 7	4 9	1 6	2 6	1 7	1 0	1,8 5,6	4 9,8	1,75 7,25	61
http://www.wdr.de/themen/verkehr/schiene/metrorapid/inhalt.jhtml	6 29	4 10	1 10	1 8	1 10	4 10	1 10	6 0	1 7	4 0	1 5,6	1,6 5,4	2,2 6,4	62
Gesamt gefunden	50	52	56	47	51	34	60	54	61	29	51,2	48,6	49,4	
ohne Spezialisieren	48	51	56	47	51	33	59	54	61	28	48,6	45,6	47,2	
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**Nr**

B.2.1 Zufällige Wörter

Die hier vorgestellten Ergebnisse stammen aus Testläufen vom 19. August 2003.

Ziel	1	2	3	4	5	μ	σ	Nr
http://www.wdr.de/themen/politik/nrw/steinkohle/rag_ausgliederung.jhtml	1 4	1 4	1 5	6 20	1 3	2 7,2	2,24 7,19	01
http://www.wdr.de/themen/politik/international/elfter_september/verfassungsschutz_cia/index.jhtml	1 1	1 1	1 3	1 1	4 7	1,6 2,6	1,34 2,61	02
	1	**2**	**3**	**4**	**5**	μ	σ	**Nr**

Ziel	1	2	3	4	5	μ	σ	Nr
http://www.wdr.de/themen/politik/deutschland/stabilitaetspakt/index.jhtml	1 3	1 6	1 3	5 27	1 3	1,8 8,4	1,79 10,48	03
http://www.wdr.de/themen/politik/international/soldaten_afghanistan/landung_koeln.jhtml	1 5	1 5	1 1	1 1	3 8	1,4 4	0,89 3	04
http://www.wdr.de/themen/politik/nrw/interview_2003/pinkwart/interview.jhtml	1 5	1 8	1 10	4 9	3 13	2 9	1,41 2,92	05
http://www.wdr.de/themen/politik/nrw/gemeindefinanzierung/reaktionen.jhtml	1 3	1 1	1 3	1 3	1 4	1 2,8	0 1,1	06
http://www.wdr.de/themen/politik/nrw/gemeindefinanzierung/ruettgers.jhtml	1 1	3 25	2 5	1 1	1 1	1,6 6,6	0,89 10,43	07
http://www.wdr.de/themen/politik/nrw/moellemann/inhalt.jhtml	4 1	1 4	3 7	1 1	1 5	2 3,6	1,41 2,61	08
http://www.wdr.de/themen/politik/nrw/muellaffaere_spd/inhalt.jhtml	4 9	1 10	1 4	2 1	1 4	1,8 5,6	1,3 3,78	09
http://www.wdr.de/themen/homepages/irak.jhtml	1 1	1 5	1 1	1 1	1 7	1 3	0 2,83	10
http://www.wdr.de/themen/homepages/d_usa.jhtml	3 2	1 5	1 1	1 1	1 5	1 2,8	0 2,05	11
http://www.wdr.de/themen/wirtschaft/arbeit_und_tarifwesen/hartz_reformen/index.jhtml	1 1	1 8	3 12	1 10	5 9	2,2 8	1,79 4,18	12
	1	**2**	**3**	**4**	**5**	μ	σ	**Nr**

Ziel	1	2	3	4	5	μ	σ	Nr
http://www.wdr.de/themen/wirtschaft/1/haushaltsgeraete/index.jhtml	1 7	1 2	1 1	6 6	6 21	3 7,4	2,74 8,02	13
http://www.wdr.de/themen/wirtschaft/wirtschaftsbranche/ford/krise.jhtml	1 4	1 1	1 8	3 6	4 18	2 7,4	1,41 6,47	14
http://www.wdr.de/themen/wirtschaft/wirtschaftsbranche/rwe/usa_drohen.jhtml	4 10	1 3	1 1	1 3	1 1	1,6 3,6	1,34 3,71	15
http://www.wdr.de/themen/wirtschaft/arbeit_und_tarifwesen/ich_ag/erfolg.jhtml	1 6	1 2	1 3	1 1	5 21	1,8 6,6	1,79 8,26	16
http://www.wdr.de/themen/wirtschaft/geld-_und_kreditwesen/westlb/index_030806.jhtml	6 18	4 11	6 12	1 7	1 2	3,6 10	2,51 5,96	17
http://www.wdr.de/themen/wirtschaft/wirtschaftsbranche/luftfahrt/flughafen_weeze.jhtml	3 12	1 3	1 4	4 10	1 3	2 6,4	1,41 4,28	18
http://www.wdr.de/online/jobs/jobzeit/index.phtml?rubrikenstyle=wirtschaft	4 18	1 4	1 5	1 2	1 3	1,6 6,4	1,34 6,58	19
http://www.wdr.de/themen/forschung/technik/solarflitzer/index.jhtml	1 2	4 1	1 2	3 11	1 3	2 3,8	1,41 4,09	20
http://www.wdr.de/themen/forschung/interdisziplinaer/virtuelle_bibliothek/index.jhtml	1 7	1 5	1 6	1 7	1 4	1 5,8	0 1,3	21
http://www.wdr.de/themen/homepages/kleine_anfrage.jhtml	2 11	1 1	6 14	3 3	1 1	2,6 6	2,07 6,08	22
	1	**2**	**3**	**4**	**5**	μ	σ	**Nr**

Ziel	1	2	3	4	5	μ	σ	Nr
http://www.wdr.de/themen/kultur/1/linkshaender_tag_2003/gaestebuch.jhtml	1 2	1 1	6 6	1 3	1 2	2 2,8	2,24 1,92	23
http://www.wdr.de/themen/kultur/rundfunk/lilipuz_sommertour2003/index.jhtml	1 1	4 21	6 6	1 1	5 9	3,4 7,6	2,3 8,23	24
http://www.wdr.de/themen/kultur/bildung_und_erziehung/bafoeg_missbrauch/index.jhtml	6 4	4 11	6 28	6 21	1 10	4,6 14,8	2,19 9,58	25
http://www.wdr.de/themen/kultur/personen/ruge/index.jhtml	1 1	1 2	1 4	1 1	1 1	1 1,8	0 1,3	26
http://www.wdr.de/themen/kultur/quiz/quiz_rechtschreibreform.jhtml	1 6	1 1	1 8	1 1	1 1	1 3,4	0 3,36	27
http://www.wdr.de/themen/panorama/lifestyle/modemesse_reevolutions_2003/sommer.jhtml	3 16	1 4	1 5	1 2	1 1	1,4 5,6	0,89 6,02	28
http://www.wdr.de/themen/kultur/1/kinderschutzbund/interview.jhtml	1 3	1 7	1 1	1 4	1 9	1 4,8	0 3,19	29
http://www.wdr.de/themen/kultur/netzkultur/heimatinseln/inseln.jhtml	1 1	1 2	1 8	1 1	1 10	1 4,4	0 4,28	30
http://www.wdr.de/themen/homepages/popkomm2003.jhtml	1 5	1 9	6 18	1 2	4 7	2,6 8,2	2,3 6,06	31
http://www.wdr.de/themen/homepages/unicef.jhtml	1 1	1 1	1 1	1 1	5 0	1,8 0,8	1,79 0,45	32
http://www.wdr.de/themen/kultur/netzkultur/nrw_privat/wohnungen.jhtml	1 10	1 8	1 8	1 8	1 8	1 8,4	0 0,89	33
	1	**2**	**3**	**4**	**5**	μ	σ	**Nr**

Ziel	1	2	3	4	5	μ	σ	Nr
http://www.wdr.de/themen/computer/internet/blaster/index.jhtml	1 3	1 1	6 9	2 20	1 2	2,2 7	2,17 7,91	34
http://www.wdr.de/themen/computer/software/virenticker/index.jhtml	1 1	1 1	1 3	5 0	1 8	1,8 2,6	1,79 3,21	35
http://www.wdr.de/themen/computer/schiebwoche/2003/index_33.jhtml	1 4	1 3	1 7	1 1	1 5	1 4	0 2,24	36
http://www.wdr.de/themen/computer/internet/sicherheit/exploit.jhtml	1 1	1 4	1 1	1 10	1 4	1 4	0 3,67	37
http://www.wdr.de/themen/computer/internet/barrierefreies_internet/internetcafe_reportage.jhtml	1 4	1 4	1 4	1 2	1 9	1 4,6	0 2,61	38
http://www.wdr.de/themen/computer/angeklickt/tagestipp/tagestipp.jhtml	1 2	1 2	1 2	5 1	5 2	2,6 1,8	2,19 0,45	39
http://www.wdr.de/themen/computer/angeklickt/webtv/index.jhtml	1 10	1 3	1 1	1 1	1 10	1 5	0 4,64	40
http://www.wdr.de/themen/sport/1/holzfaeller/index.jhtml	1 4	5 0	1 3	4 2	4 23	3 6,4	1,87 9,4	41
http://www.wdr.de/themen/homepages/olympia_2012.jhtml	1 6	1 1	1 1	1 1	5 0	1,8 1,8	1,79 2,39	42
http://www.wdr.de/themen/panorama/2/lottojackpot_nrw/index.jhtml	1 1	1 1	1 1	1 1	1 2	1 1,2	0 0,45	43
http://www.wdr.de/themen/panorama/kriminalitaet02/gladbecker_geiseldrama_1988/index.jhtml	3 1	1 1	1 1	1 8	1 2	1,4 2,6	0,89 3,05	44
	1	2	3	4	5	μ	σ	Nr

Ziel	1	2	3	4	5	μ	σ	Nr
http://www.wdr.de/themen/panorama/2/stromausfall_bielefeld/index.jhtml	6 60	1 1	1 3	2 7	2 6	2,4 15,4	2,07 25,05	45
http://www.wdr.de/themen/panorama/kriminalitaet02/missbrauch_moenchengladbach/index.jhtml	1 1	1 2	1 1	6 14	3 9	2,4 5,4	2,19 5,86	46
http://www.wdr.de/themen/panorama/wetter/sommer_2003/abkuehlung_030814.jhtml	1 1	1 1	5 21	1 1	6 40	2,8 12,8	2,49 17,5	47
http://www.wdr.de/themen/panorama/wetter/sommer_2003/waldbraende.jhtml	1 5	1 2	1 1	2 1	1 2	1,2 2,2	0,45 1,64	48
http://www.wdr.de/themen/panorama/2/bombenentschaerfungen_nrw_2002/index.jhtml	1 4	2 2	3 30	1 1	6 30	2,6 13,4	2,07 15,19	49
http://www.wdr.de/themen/panorama/wetter/sommer_2003/hitze.jhtml	1 2	1 1	1 1	5 9	1 1	1,8 2,8	1,79 3,49	50
http://www.wdr.de/themen/gesundheit/1/antipasti/index.jhtml	1 1	5 13	1 2	1 2	1 1	1,8 3,8	1,79 5,17	51
http://www.wdr.de/themen/gesundheit/gesundheitswesen/gesetzliche_krankenkassen/sparzwang.jhtml	1 3	1 9	3 4	1 4	1 4	1,4 4,8	0,89 2,39	52
	1	**2**	**3**	**4**	**5**	μ	σ	**Nr**

Ziel	**1**	**2**	**3**	**4**	**5**	**μ**	**σ**	**Nr**
http://www.wdr.de/themen/gesundheit/gesundheitswesen/gesetzliche_krankenkassen/bkk_beitragserhoehung.jhtml	1 1	1 1	1 1	1 4	1 3	1 2	0 1,41	53
http://www.wdr.de/themen/freizeit/freizeitgestaltung/insel_sommer/monkeys_island.jhtml	1 3	1 3	1 3	1 2	5 19	1,8 6	1,79 7,28	54
http://www.wdr.de/themen/freizeit/freizeitgestaltung/future_parade/index.jhtml	6 40	1 3	1 7	1 2	6 20	3 14,4	2,74 16,01	55
http://www.wdr.de/themen/freizeit/freizeitgestaltung/mundorgel_50jahre/interview_guildo_horn.jhtml	1 2	3 23	3 14	4 19	1 10	2,4 13,6	1,34 8,14	56
http://www.wdr.de/themen/verkehr/strasse/tunnel_b236/index.jhtml	1 5	1 1	4 13	1 1	3 13	2 6,6	1,41 6,07	57
http://www.wdr.de/themen/verkehr/schiene/deutsche_bahn/maengel.jhtml	4 13	1 3	1 1	1 2	5 9	2,4 5,6	1,95 5,18	58
http://www.wdr.de/themen/verkehr/strasse/radarwarngeraete/autobahn_warnung.jhtml	1 1	3 10	6 39	1 1	1 8	2,4 11,8	2,19 15,74	59
http://www.wdr.de/themen/verkehr/wasser/niedrigwasser_rhein_ruhr/niedrigwasser_0308.jhtml	1 1	6 28	1 3	1 1	1 5	2 7,6	2,24 11,52	60
http://www.wdr.de/themen/panorama/2/verdaechtiger_koffer_koeln/index.jhtml	5 19	1 1	1 1	1 1	1 6	1,8 5,6	1,79 7,8	61
http://www.wdr.de/themen/verkehr/schiene/metrorapid/inhalt.jhtml	1 8	1 6	1 7	1 6	1 1	1 5,6	0 2,7	62
Gesamt gefunden	**51**	**56**	**48**	**54**	**47**	**51,2**	**3,83**	
ohne Spezialisieren	**49**	**53**	**47**	**51**	**43**	**48,6**	**3,85**	

B.2.2 Zufällige Wörter mit Stoppwortelimination

Die hier vorgestellten Ergebnisse stammen aus Testläufen vom 19. August 2003.

Ziel	1	2	3	4	5	μ	σ	Nr
http://www.wdr.de/themen/politik/nrw/steinkohle/rag_ausgliederung.jhtml	1 3	1 3	1 3	4 21	1 3	1,6 6,6	1,34 8,05	01
http://www.wdr.de/themen/politik/international/elfter_september/verfassungsschutz_cia/index.jhtml	1 1	2 1	1 1	1 1	1 1	1,2 1	0,45 0	02
http://www.wdr.de/themen/politik/deutschland/stabilitaetspakt/index.jhtml	4 18	6 9	1 2	1 3	1 3	2,6 7	2,3 6,75	03
http://www.wdr.de/themen/politik/international/soldaten_afghanistan/landung_koeln.jhtml	1 6	1 3	2 7	1 10	1 1	1,2 5,4	0,45 3,51	04
http://www.wdr.de/themen/politik/nrw/interview_2003/pinkwart/interview.jhtml	2 2	1 1	3 8	1 2	1 8	1,6 4,2	0,89 3,49	05
http://www.wdr.de/themen/politik/nrw/gemeindefinanzierung/reaktionen.jhtml	1 6	1 4	5 0	1 4	1 4	1,8 3,6	1,79 2,19	06
http://www.wdr.de/themen/politik/nrw/gemeindefinanzierung/ruettgers.jhtml	1 3	6 17	1 4	1 1	1 3	2 5,6	2,24 6,47	07
http://www.wdr.de/themen/politik/nrw/moellemann/inhalt.jhtml	1 5	3 4	1 5	1 2	4 5	2 4,2	1,41 1,3	08
http://www.wdr.de/themen/politik/nrw/muellaffaere_spd/inhalt.jhtml	1 2	1 7	1 3	4 7	4 10	2,2 5,8	1,64 3,27	09
	1	2	3	4	5	μ	σ	Nr

Ziel	1	2	3	4	5	μ	σ	Nr
http://www.wdr.de/themen/homepages/irak.jhtml	2 4	4 21	4 9	1 4	1 9	2,4 9,4	1,52 6,95	10
http://www.wdr.de/themen/homepages/d_usa.jhtml	5 2	3 7	1 5	1 5	1 5	2,2 4,8	1,79 1,79	11
http://www.wdr.de/themen/wirtschaft/arbeit_und_tarifwesen/hartz_reformen/index.jhtml	1 10	1 2	6 60	1 6	3 7	2,4 17	2,19 24,21	12
http://www.wdr.de/themen/wirtschaft/1/haushaltsgeraete/index.jhtml	1 2	6 32	1 2	1 2	1 2	2 8	2,24 13,42	13
http://www.wdr.de/themen/wirtschaft/wirtschaftsbranche/ford/krise.jhtml	1 4	1 1	6 12	6 8	2 4	3,2 5,8	2,59 4,27	14
http://www.wdr.de/themen/wirtschaft/wirtschaftsbranche/rwe/usa_drohen.jhtml	5 0	1 3	1 3	1 1	1 3	1,8 2	1,79 1,41	15
http://www.wdr.de/themen/wirtschaft/arbeit_und_tarifwesen/ich_ag/erfolg.jhtml	1 1	6 22	1 1	6 23	6 8	4 11	2,74 10,89	16
http://www.wdr.de/themen/wirtschaft/geld-_und_kreditwesen/westlb/index_030806.jhtml	1 1	6 34	1 10	1 2	3 5	2,4 10,4	2,19 13,65	17
http://www.wdr.de/themen/wirtschaft/wirtschaftsbranche/luftfahrt/flughafen_weeze.jhtml	6 7	1 4	2 4	3 3	4 16	3,2 6,8	1,92 5,36	18
http://www.wdr.de/online/jobs/jobzeit/index.phtml?rubrikenstyle=wirtschaft	1 3	6 21	3 29	1 3	1 3	2,4 11,8	2,19 12,38	19
	1	2	3	4	5	μ	σ	Nr

Ziel	1	2	3	4	5	μ	σ	Nr
http://www.wdr.de/themen/forschung/technik/solarflitzer/index.jhtml	1 3	1 3	3 3	1 3	1 2	1,4 2,8	0,89 0,45	20
http://www.wdr.de/themen/forschung/interdisziplinaer/virtuelle_bibliothek/index.jhtml	1 6	5 0	5 25	1 2	1 4	2,6 7,4	2,19 10,09	21
http://www.wdr.de/themen/homepages/kleine_anfrage.jhtml	1 2	4 11	3 10	1 1	4 9	2,6 6,6	1,52 4,72	22
http://www.wdr.de/themen/kultur/1/linkshaender_tag_2003/gaestebuch.jhtml	1 2	1 2	1 3	1 2	1 9	1 3,6	0 3,05	23
http://www.wdr.de/themen/kultur/rundfunk/lilipuz_sommertour2003/index.jhtml	6 23	3 13	1 8	6 35	5 0	4,2 15,8	2,17 13,59	24
http://www.wdr.de/themen/kultur/bildung_und_erziehung/bafoeg_missbrauch/index.jhtml	6 10	1 2	1 7	6 10	1 9	3 7,6	2,74 3,36	25
http://www.wdr.de/themen/kultur/personen/ruge/index.jhtml	5 19	1 3	1 3	1 2	1 3	1,8 6	1,79 7,28	26
http://www.wdr.de/themen/kultur/quiz/quiz_rechtschreibreform.jhtml	1 6	1 6	5 0	1 2	1 6	1,8 4	1,79 2,83	27
http://www.wdr.de/themen/panorama/lifestyle/modemesse_reevolutions_2003/sommer.jhtml	1 1	1 8	1 3	1 4	4 3	1,6 3,8	1,34 2,59	28
http://www.wdr.de/themen/kultur/1/kinderschutzbund/interview.jhtml	2 5	1 1	1 3	1 5	1 8	1,2 4,4	0,45 2,61	29
http://www.wdr.de/themen/kultur/netzkultur/heimatinseln/inseln.jhtml	5 9	3 3	1 2	1 2	1 2	2,2 3,6	1,79 3,05	30
	1	2	3	4	5	μ	σ	Nr

Ziel	1	2	3	4	5	μ	σ	Nr
http://www.wdr.de/themen/homepages/popkomm2003.jhtml	5 16	6 27	6 10	6 10	6 9	5,8 14,4	0,45 7,57	31
http://www.wdr.de/themen/homepages/unicef.jhtml	6 9	1 7	1 3	4 5	6 17	3,6 8,2	2,51 5,4	32
http://www.wdr.de/themen/kultur/netzkultur/nrw_privat/wohnungen.jhtml	1 9	1 8	1 10	1 8	1 10	1 9	0 1	33
http://www.wdr.de/themen/computer/internet/blaster/index.jhtml	3 7	1 2	1 2	1 10	1 3	1,4 4,8	0,89 3,56	34
http://www.wdr.de/themen/computer/software/virenticker/index.jhtml	3 1	4 5	1 4	1 3	1 4	2 3,4	1,41 1,52	35
http://www.wdr.de/themen/computer/schiebwoche/2003/index_33.jhtml	1 4	1 2	6 28	6 15	1 5	3 10,8	2,74 10,85	36
http://www.wdr.de/themen/computer/internet/sicherheit/exploit.jhtml	5 13	1 3	1 3	1 4	3 10	2,2 6,6	1,79 4,62	37
http://www.wdr.de/themen/computer/internet/barrierefreies_internet/internetcafe_reportage.jhtml	1 5	1 5	4 24	6 9	1 5	2,6 9,6	2,3 8,23	38
http://www.wdr.de/themen/computer/angeklickt/tagestipp/tagestipp.jhtml	—	5 13	6 39	5 0	6 13	4,4 13	2,51 15,92	39
http://www.wdr.de/themen/computer/angeklickt/webtv/index.jhtml	1 10	1 10	1 10	6 12	3 5	2,4 9,4	2,19 2,61	40
http://www.wdr.de/themen/sport/1/holzfaeller/index.jhtml	1 3	1 3	1 2	1 4	1 4	1 3,2	0 0,84	41
	1	**2**	**3**	**4**	**5**	μ	σ	**Nr**

Ziel	1	2	3	4	5	μ	σ	Nr
http://www.wdr.de/themen/homepages/olympia_2012.jhtml	1 4	5 5	6 10	1 4	5 12	3,6 7	2,41 3,74	42
http://www.wdr.de/themen/panorama/2/lottojackpot_nrw/index.jhtml	1 2	1 2	1 2	1 2	1 2	1 2	0 0	43
http://www.wdr.de/themen/panorama/kriminalitaet02/gladbecker_geiseldrama_1988/index.jhtml	1 2	4 10	1 3	6 17	1 3	2,6 7	2,3 6,44	44
http://www.wdr.de/themen/panorama/2/stromausfall_bielefeld/index.jhtml	4 10	1 3	1 3	6 31	6 15	3,6 12,4	2,51 11,57	45
http://www.wdr.de/themen/panorama/kriminalitaet02/missbrauch_moenchengladbach/index.jhtml	5 21	4 10	4 9	5 6	1 9	3,8 11	1,64 5,79	46
http://www.wdr.de/themen/panorama/wetter/sommer_2003/abkuehlung_030814.jhtml	1 1	1 6	1 9	1 1	1 8	1 5	0 3,81	47
http://www.wdr.de/themen/panorama/wetter/sommer_2003/waldbraende.jhtml	1 2	4 8	4 10	5 12	2 8	3,2 8	1,64 3,74	48
http://www.wdr.de/themen/panorama/2/bombenentschaerfungen_nrw_2002/index.jhtml	2 8	3 2	1 10	1 6	1 2	1,6 5,6	0,89 3,58	49
http://www.wdr.de/themen/panorama/wetter/sommer_2003/hitze.jhtml	4 9	6 4	1 2	4 3	1 2	3,2 4	2,17 2,92	50
http://www.wdr.de/themen/gesundheit/1/antipasti/index.jhtml	1 2	1 9	1 2	1 3	1 2	1 3,6	0 3,05	51
http://www.wdr.de/themen/gesundheit/gesundheitswesen/gesetzliche_krankenkassen/sparzwang.jhtml	1 2	1 2	4 5	1 3	1 2	1,6 2,8	1,34 1,3	52
	1	2	3	4	5	μ	σ	Nr

Ziel	1	2	3	4	5	μ	σ	Nr
http://www.wdr.de/themen/gesundheit/gesundheitswesen/gesetzliche_krankenkassen/bkk_beitragserhoehung.jhtml	1 4	1 1	1 5	1 3	1 1	1 2,8	0 1,79	53
http://www.wdr.de/themen/freizeit/freizeitgestaltung/insel_sommer/monkeys_island.jhtml	1 4	1 2	1 3	1 9	4 9	1,6 5,4	1,34 3,36	54
http://www.wdr.de/themen/freizeit/freizeitgestaltung/future_parade/index.jhtml	6 10	1 3	1 5	1 3	1 1	2 4,4	2,24 3,44	55
http://www.wdr.de/themen/freizeit/freizeitgestaltung/mundorgel_50jahre/interview_guildo_horn.jhtml	1 3	6 22	4 12	6 27	1 5	3,6 13,8	2,51 10,47	56
http://www.wdr.de/themen/verkehr/strasse/tunnel_b236/index.jhtml	1 1	1 2	1 2	3 2	1 1	1,4 1,6	0,89 0,55	57
http://www.wdr.de/themen/verkehr/schiene/deutsche_bahn/maengel.jhtml	6 10	3 9	4 10	1 1	1 2	3 6,4	2,12 4,51	58
http://www.wdr.de/themen/verkehr/strasse/radarwarngeraete/autobahn_warnung.jhtml	1 2	1 7	6 10	1 2	1 2	2 4,6	2,24 3,71	59
http://www.wdr.de/themen/verkehr/wasser/niedrigwasser_rhein_ruhr/niedrigwasser_0308.jhtml	1 3	1 2	1 7	1 3	4 3	1,6 3,6	1,34 1,95	60
http://www.wdr.de/themen/panorama/2/verdaechtiger_koffer_koeln/index.jhtml	4 10	1 5	4 9	6 23	5 2	4 9,8	1,87 8,04	61
	1	2	3	4	5	μ	σ	Nr

Ziel	1	2	3	4	5	μ	σ	Nr
http://www.wdr.de/ themen/verkehr/schiene/ metrorapid/inhalt.jhtml	4 1	1 6	1 6	1 7	1 7	1,6 5,4	1,34 2,51	62
Gesamt gefunden	**47**	**48**	**47**	**49**	**52**	**48,6**	**2,07**	
ohne Spezialisieren	**43**	**45**	**45**	**46**	**49**	**45,6**	**2,19**	

B.2.3 Zufällige Wörter / Häufigkeitsverteilt

Die hier vorgestellten Ergebnisse stammen aus Testläufen vom 19. August 2003.

Ziel	1	2	3	4	5	μ	σ	Nr
http://www.wdr.de/ themen/politik/nrw/ steinkohle/rag_ ausgliederung.jhtml	1 3	1 7	1 10	1 4	1 3	1 5,4	0 3,05	01
http://www.wdr.de/ themen/politik/ international/ elfter_september/ verfassungsschutz_cia/ index.jhtml	1 3	1 1	1 10	6 41	1 10	2 13	2,24 16,17	02
http://www.wdr.de/ themen/politik/ deutschland/ stabilitaetspakt/index. jhtml	1 5	3 19	1 10	1 8	1 4	1,4 9,2	0,89 5,97	03
http://www.wdr.de/ themen/politik/ international/soldaten_ afghanistan/landung_ koeln.jhtml	1 7	1 8	1 5	1 2	3 15	1,4 7,4	0,89 4,83	04
http://www.wdr.de/ themen/politik/nrw/ interview_2003/pinkwart/ interview.jhtml	1 1	1 8	1 10	1 10	1 4	1 6,6	0 3,97	05
http://www.wdr.de/ themen/politik/nrw/ gemeindefinanzierung/ reaktionen.jhtml	1 4	1 7	1 8	1 10	1 10	1 7,8	0 2,49	06
	1	**2**	**3**	**4**	**5**	μ	σ	**Nr**

Ziel	1	2	3	4	5	μ	σ	Nr
http://www.wdr.de/themen/politik/nrw/gemeindefinanzierung/ruettgers.jhtml	6 19	1 10	1 2	1 10	4 20	2,6 12,2	2,3 7,43	07
http://www.wdr.de/themen/politik/nrw/moellemann/inhalt.jhtml	2 5	1 9	1 10	1 5	1 10	1,2 7,8	0,45 2,59	08
http://www.wdr.de/themen/politik/nrw/muellaffaere_spd/inhalt.jhtml	3 7	3 1	3 16	1 7	1 7	2,2 7,6	1,1 5,37	09
http://www.wdr.de/themen/homepages/irak.jhtml	5 0	1 10	1 10	1 6	1 4	1,8 6	1,79 4,24	10
http://www.wdr.de/themen/homepages/d_usa.jhtml	1 10	1 5	1 5	1 5	5 10	1,8 7	1,79 2,74	11
http://www.wdr.de/themen/wirtschaft/arbeit_und_tarifwesen/hartz_reformen/index.jhtml	1 10	6 3	1 3	1 10	4 7	2,6 6,6	2,3 3,51	12
http://www.wdr.de/themen/wirtschaft/1/haushaltsgeraete/index.jhtml	1 4	1 5	6 8	6 17	1 9	3 8,6	2,74 5,13	13
http://www.wdr.de/themen/wirtschaft/wirtschaftsbranche/ford/krise.jhtml	6 34	1 4	1 9	3 27	1 10	2,4 16,8	2,19 12,95	14
http://www.wdr.de/themen/wirtschaft/wirtschaftsbranche/rwe/usa_drohen.jhtml	4 10	1 6	1 8	1 10	1 9	1,6 8,6	1,34 1,67	15
http://www.wdr.de/themen/wirtschaft/arbeit_und_tarifwesen/ich_ag/erfolg.jhtml	2 14	1 10	1 1	4 6	1 8	1,8 7,8	1,3 4,82	16
http://www.wdr.de/themen/wirtschaft/geld-_und_kreditwesen/westlb/index_030806.jhtml	1 7	6 10	1 10	3 18	4 28	3 14,6	2,12 8,53	17
	1	**2**	**3**	**4**	**5**	μ	σ	**Nr**

Ziel	1	2	3	4	5	μ	σ	Nr
http://www.wdr.de/themen/wirtschaft/wirtschaftsbranche/luftfahrt/flughafen_weeze.jhtml	1 3	1 6	1 3	4 19	1 4	1,6 7	1,34 6,82	18
http://www.wdr.de/online/jobs/jobzeit/index.phtml?rubrikenstyle=wirtschaft	2 20	1 3	1 3	1 2	1 3	1,2 6,2	0,45 7,73	19
http://www.wdr.de/themen/forschung/technik/solarflitzer/index.jhtml	1 3	1 10	1 4	1 1	1 3	1 4,2	0 3,42	20
http://www.wdr.de/themen/forschung/interdisziplinaer/virtuelle_bibliothek/index.jhtml	1 6	1 4	1 4	1 8	1 9	1 6,2	0 2,28	21
http://www.wdr.de/themen/homepages/kleine_anfrage.jhtml	1 1	1 9	1 2	1 4	1 1	1 3,4	0 3,36	22
http://www.wdr.de/themen/kultur/1/linkshaender_tag_2003/gaestebuch.jhtml	1 10	1 10	1 7	1 10	1 10	1 9,4	0 1,34	23
http://www.wdr.de/themen/kultur/rundfunk/lilipuz_sommertour2003/index.jhtml	1 9	3 9	1 7	1 9	1 2	1,4 7,2	0,89 3,03	24
http://www.wdr.de/themen/kultur/bildung_und_erziehung/bafoeg_missbrauch/index.jhtml	5 14	5 44	1 2	1 3	1 4	2,6 13,4	2,19 17,77	25
http://www.wdr.de/themen/kultur/personen/ruge/index.jhtml	1 6	1 5	1 3	1 9	1 3	1 5,2	0 2,49	26
http://www.wdr.de/themen/kultur/quiz/quiz_rechtschreibreform.jhtml	1 4	1 2	1 6	6 20	1 5	2 7,4	2,24 7,2	27
http://www.wdr.de/themen/panorama/lifestyle/modemesse_reevolutions_2003/sommer.jhtml	1 1	1 5	1 4	1 4	1 2	1 3,2	0 1,64	28
	1	2	3	4	5	μ	σ	Nr

Ziel	1	2	3	4	5	μ	σ	Nr
http://www.wdr.de/themen/kultur/1/kinderschutzbund/interview.jhtml	1 7	1 8	4 9	1 8	1 2	1,6 6,8	1,34 2,77	29
http://www.wdr.de/themen/kultur/netzkultur/heimatinseln/inseln.jhtml	1 4	1 3	1 2	1 10	4 2	1,6 4,2	1,34 3,35	30
http://www.wdr.de/themen/homepages/popkomm2003.jhtml	5 30	1 8	4 3	3 9	4 10	3,4 12	1,52 10,42	31
http://www.wdr.de/themen/homepages/unicef.jhtml	4 15	3 26	2 14	2 15	1 3	2,4 14,6	1,14 8,14	32
http://www.wdr.de/themen/kultur/netzkultur/nrw_privat/wohnungen.jhtml	1 8	1 8	1 9	2 17	1 10	1,2 10,4	0,45 3,78	33
http://www.wdr.de/themen/computer/internet/blaster/index.jhtml	1 2	4 10	1 9	1 10	1 10	1,6 8,2	1,34 3,49	34
http://www.wdr.de/themen/computer/software/virenticker/index.jhtml	1 10	1 4	1 9	1 4	1 10	1 7,4	0 3,13	35
http://www.wdr.de/themen/computer/schiebwoche/2003/index_33.jhtml	1 7	5 14	1 5	4 30	1 5	2,4 12,2	1,95 10,62	36
http://www.wdr.de/themen/computer/internet/sicherheit/exploit.jhtml	1 10	1 3	1 4	1 1	1 9	1 5,4	0 3,91	37
http://www.wdr.de/themen/computer/internet/barrierefreies_internet/internetcafe_reportage.jhtml	1 3	2 5	1 3	1 9	1 5	1,2 5	0,45 2,45	38
	1	**2**	**3**	**4**	**5**	μ	σ	**Nr**

Ziel	1	2	3	4	5	μ	σ	Nr
http://www.wdr.de/themen/computer/angeklickt/tagestipp/tagestipp.jhtml	6 19	6 12	4 9	2 19	6 27	4,8 17,2	1,79 7,01	39
http://www.wdr.de/themen/computer/angeklickt/webtv/index.jhtml	1 10	1 10	1 10	1 10	1 10	1 10	0 0	40
http://www.wdr.de/themen/sport/1/holzfaeller/index.jhtml	1 10	1 4	1 4	4 37	4 10	2,2 13	1,64 13,75	41
http://www.wdr.de/themen/homepages/olympia_2012.jhtml	1 10	3 3	1 4	1 8	2 10	1,6 7	0,89 3,32	42
http://www.wdr.de/themen/panorama/2/lottojackpot_nrw/index.jhtml	1 2	1 3	1 9	5 31	1 2	1,8 9,4	1,79 12,42	43
http://www.wdr.de/themen/panorama/kriminalitaet02/gladbecker_geiseldrama_1988/index.jhtml	1 9	1 2	1 3	1 1	1 8	1 4,6	0 3,65	44
http://www.wdr.de/themen/panorama/2/stromausfall_bielefeld/index.jhtml	3 13	1 2	1 10	3 9	1 2	1,8 7,2	1,1 4,97	45
http://www.wdr.de/themen/panorama/kriminalitaet02/missbrauch_moenchengladbach/index.jhtml	3 10	5 17	2 18	6 34	4 4	4 16,6	1,58 11,26	46
http://www.wdr.de/themen/panorama/wetter/sommer_2003/abkuehlung_030814.jhtml	1 6	1 1	1 6	1 8	1 2	1 4,6	0 2,97	47
http://www.wdr.de/themen/panorama/wetter/sommer_2003/waldbraende.jhtml	6 10	4 6	1 2	1 2	3 10	3 6	2,12 4	48
	1	2	3	4	5	μ	σ	Nr

Ziel	1	2	3	4	5	μ	σ	Nr
http://www.wdr.de/themen/panorama/2/bombenentschaerfungen_nrw_2002/index.jhtml	1 4	3 21	5 19	6 30	1 2	3,2 15,2	2,28 11,9	49
http://www.wdr.de/themen/panorama/wetter/sommer_2003/hitze.jhtml	1 3	1 2	3 2	1 2	5 1	2,2 2	1,79 0,71	50
http://www.wdr.de/themen/gesundheit/1/antipasti/index.jhtml	1 5	1 10	1 4	6 37	1 6	2 12,4	2,24 13,94	51
http://www.wdr.de/themen/gesundheit/gesundheitswesen/gesetzliche_krankenkassen/sparzwang.jhtml	1 4	5 6	1 4	1 4	1 4	1,8 4,4	1,79 0,89	52
http://www.wdr.de/themen/gesundheit/gesundheitswesen/gesetzliche_krankenkassen/bkk_beitragserhoehung.jhtml	1 8	1 3	1 10	1 3	1 3	1 5,4	0 3,36	53
http://www.wdr.de/themen/freizeit/freizeitgestaltung/insel_sommer/monkeys_island.jhtml	1 3	1 10	1 4	4 16	1 4	1,6 7,4	1,34 5,55	54
http://www.wdr.de/themen/freizeit/freizeitgestaltung/future_parade/index.jhtml	1 10	1 5	1 10	1 10	1 8	1 8,6	0 2,19	55
http://www.wdr.de/themen/freizeit/freizeitgestaltung/mundorgel_50jahre/interview_guildo_horn.jhtml	1 5	1 5	1 5	1 5	3 5	1,4 5	0,89 0	56
	1	**2**	**3**	**4**	**5**	μ	σ	**Nr**

Ziel	1	2	3	4	5	μ	σ	Nr
http://www.wdr.de/themen/verkehr/strasse/tunnel_b236/index.jhtml	1 1	2 1	5 16	3 20	1 1	2,4 7,8	1,67 9,42	57
http://www.wdr.de/themen/verkehr/schiene/deutsche_bahn/maengel.jhtml	1 9	1 9	1 9	1 8	1 3	1 7,6	0 2,61	58
http://www.wdr.de/themen/verkehr/strasse/radarwarngeraete/autobahn_warnung.jhtml	1 6	4 7	1 10	1 5	1 3	1,6 6,2	1,34 2,59	59
http://www.wdr.de/themen/verkehr/wasser/niedrigwasser_rhein_ruhr/niedrigwasser_0308.jhtml	1 3	1 3	1 3	1 2	1 2	1 2,6	0 0,55	60
http://www.wdr.de/themen/panorama/2/verdaechtiger_koffer_koeln/index.jhtml	1 5	1 10	1 5	—	4 9	1,75 7,25	1,5 2,63	61
http://www.wdr.de/themen/verkehr/schiene/metrorapid/inhalt.jhtml	1 7	1 10	1 7	4 7	4 1	2,2 6,4	1,64 3,29	62
Gesamt gefunden	**50**	**48**	**52**	**47**	**50**	**49,4**	**1,95**	
ohne Spezialisieren	**48**	**45**	52	**43**	**48**	**47,2**	**3,42**	

Anhang C

Bewertung der Schwierigkeit der Suche nach einem Dokument

Auf den folgenden Seiten sind die Dokumente der Testmenge mit der erwarteten Schwierigekeit des Wiederauffindens aufgelistet.

Die Zeilen sind gemäß ihrer Schwierigkeit eingefärbt:

rot	bedeutet sehr schwierig wiederzufinden
gelb	bedeutet schwierig wiederzufinden
grün	bedeutet einfach wiederzufinden

Die Einschätzung der Schwierigkeit ergibt sich aus dem Anteil der Wörter, die sowohl in der alten als auch der neuen Version des gesuchten Dokuments noch vorkommen (s. Kap. 4.3 S. 48).

In der Zeile

alte URL steht, zu welchem Dokument der neue Ort gefunden werden soll. Wenn eine Version angegeben ist, so wurde der Inhalt dem Wayback–Archiv entnommen.

erw. URL steht, wo das Dokument, das gefunden werden sollte erwartet wird.

gef.durch steht, durch welche Strategeme das Dokument gefunden wurde. Die angegebenen Zahlen entsprechenden den Zahlen in Anhang B. Diese Zeile stellt im Prinzip die a posteriori Abschätzung der Schwierigkeit dar. Je mehr Strategeme das Dokument fanden, desto einfacher war das Dokument offenbar zu finden.

Die Spalte

 # Zeichen gibt an, wie viele Zeichen das Dokument enthält.

 # Wörter gibt an, wie viele *verschiedene* Wörter im Dokument erkannt wurden.

Wortschwund gibt an, wie viele Wörter aus dem alten Dokument *nicht* mehr im neuen erwarteten Dokument vorkommen. Die Unterspalte `absolut` gibt die absolute Anzahl der nicht mehr vorkommenden Wörter an und die Unterspalte `normalisiert` gibt den Prozentsatz der nicht mehr vorkommenden Wörter bezogen auf die Gesamtlänge des Originaldokuments an.

Die Tabelleneinträge sind nach ihrer Schwierigkeit (normierter Wortschwund) absteigend sortiert.

C.1 Universität

	URL	# Zeichen	# Wörter	Wortschwund absolut	Wortschwund normalisiert
alte URL	http://www.uni-dortmund.de/UniDo/Personal/ Version vom 22.11.2001	472	17		
erw. URL	http://www.uni-dortmund.de/UniDo/Personal/	29	0	17	100.0
gef. durch	nicht gefunden		Datensatz Nr. 31		
alte URL	http://dekanat.cs.uni-dortmund.de/ Version vom 18.04.2001	2779	171		
erw. URL	http://dekanat.cs.uni-dortmund.de/	66	5	167	97.66
gef. durch	3/13/15		Datensatz Nr. 24		
alte URL	http://www.uni-dortmund.de/TOP/ Version vom 24.10.2001	3929	32		
erw. URL	http://www.uni-dortmund.de/TOP/	1107	47	23	71.88
gef. durch	nicht gefunden		Datensatz Nr. 14		
alte URL	http://www-ai.cs.uni-dortmund.de/ Version vom 26.09.2001	2436	123		
erw. URL	http://www-ai.cs.uni-dortmund.de	2975	104	83	67.48
gef. durch	4/5/6/7/8/9/11/12/13/15		Datensatz Nr. 62		
alte URL	http://www-ai.cs.uni-dortmund.de/index.eng.html Version vom 25.12.2001	1821	96		
erw. URL	http://www-ai.cs.uni-dortmund.de/index.eng.html	2297	90	63	65.63
gef. durch	5/14/15		Datensatz Nr. 44		
alte URL	http://www.ft-informatik.de/index.html Version vom 16.12.2001	2068	91		
erw. URL	http://www.ft-informatik.de/index.html	1428	53	44	48.35
gef. durch	6/8/9/15		Datensatz Nr. 33		
alte URL	http://www-ai.cs.uni-dortmund.de/ Version vom 30.11.2001	2216	113		
erw. URL	http://www-ai.cs.uni-dortmund.de/	2975	104	53	46.9
gef. durch	5/9/11/13/15		Datensatz Nr. 01		
alte URL	http://ls1-www.cs.uni-dortmund.de/ Version vom 28.11.2001	548	26		
erw. URL	http://ls1-www.cs.uni-dortmund.de/	2247	27	12	46.15
gef. durch	4/5/9/13/15		Datensatz Nr. 19		
alte URL	http://ls12-www.cs.uni-dortmund.de/ Version vom 16.05.2001	1612	94		
erw. URL	http://ls12-www.cs.uni-dortmund.de/	1803	66	42	44.68
gef. durch	5/6/9/11/15		Datensatz Nr. 23		
alte URL	http://www.isoc.org/ Version vom 23.08.2001	7940	239		
erw. URL	http://www.isoc.org/	12316	315	93	38.91
gef. durch	3/7/8/15		Datensatz Nr. 35		
alte URL	http://www-ai.cs.uni-dortmund.de/FORSCHUNG/ buecher.html Version vom 30.04.2001	866	57		
erw. URL	http://www-ai.cs.uni-dortmund.de/FORSCHUNG/ buecher.html	880	44	21	36.84
gef. durch	7/10/13/14/15		Datensatz Nr. 12		
alte URL	http://www.acm.org/ Version vom 17.12.2001	5489	154		
erw. URL	http://www.acm.org/	6770	191	49	31.82
gef. durch	4/5/6/9/15		Datensatz Nr. 34		
alte URL	http://www.ub.uni-dortmund.de/literatursuche/ index.htm Version vom 05.12.2001	2817	127		
erw. URL	http://www.ub.uni-dortmund.de/literatursuche/ index.htm	2552	132	32	25.2
gef. durch	1/2/3/4/5/14/15		Datensatz Nr. 32		
alte URL	http://ls6-www.cs.uni-dortmund.de/ Version vom 20.07.2001	759	29		
erw. URL	http://ls6-www.cs.uni-dortmund.de/	817	33	6	20.69
gef. durch	4/5/7/9/11/12/15		Datensatz Nr. 22		
alte URL	http://ls7-www.cs.uni-dortmund.de/VKInf/ Version vom 24.12.2001	2823	134		
erw. URL	http://ls7-www.cs.uni-dortmund.de/VKInf/	4358	184	27	20.15
gef. durch	3/4/5/7/8/9/11/12		Datensatz Nr. 15		
alte URL	http://www.cs.helsinki.fi/kurssit/	9199	355		
erw. URL	http://www.cs.helsinki.fi/kurssit/	9335	352	71	20.0
gef. durch	7/9/11/12		Datensatz Nr. 53		
alte URL	http://www.dfn.de/home.html Version vom 16.08.2001	578	5		
erw. URL	http://www.dfn.de	18686	214	1	20.0
gef.	1/2/3/4/5/6/8/11/12/14		Datensatz Nr. 36		
alte URL	http://ls10-www.cs.uni-dortmund.de/LS10/Pages/ sopra.shtml Version vom 29.10.2001	1655	96		

	URL	# Zeichen	# Wörter	Wortschwund absolut	Wortschwund normalisiert
erw. URL	http://ls10-www.cs.uni-dortmund.de/LS10/Pages/sopra-specials/anmeldung/sopra.shtml	3227	157	17	17.71
gef. durch	7/11		Datensatz Nr. 26		
alte URL	http://www-ai.cs.uni-dortmund.de	3062	113		
erw. URL	http://www-ai.cs.uni-dortmund.de	2975	104	20	17.7
gef. durch	4/5/6/7/8/9/11/12/13/14/15		Datensatz Nr. 61		
alte URL	http://www-ai.cs.uni-dortmund.de/PERSONAL/personal.html Version vom 24.12.2001	2459	94		
erw. URL	http://www-ai.cs.uni-dortmund.de/PERSONAL/personal.html	3115	111	16	17.02
gef. durch	3/4/5/6/7/9/11/12/14/15		Datensatz Nr. 43		
alte URL	http://ls4-www.cs.uni-dortmund.de/ Version vom 24.09.2001	2733	77		
erw. URL	http://ls4-www.cs.uni-dortmund.de/	2821	77	13	16.88
gef. durch	4/5/6/7/8/9/11/12/15		Datensatz Nr. 21		
alte URL	http://dekanat.cs.uni-dortmund.de/HaPra/index.html Version vom 24.12.2001	1312	68		
erw. URL	http://dekanat.cs.uni-dortmund.de/HaPra/index.html	2380	110	11	16.18
gef. durch	1/2/4/5/7/8/9/10/11/14/15		Datensatz Nr. 27		
alte URL	http://fsinfo.cs.uni-dortmund.de/ Version vom 02.12.2001	3601	31		
erw. URL	http://fsinfo.cs.uni-dortmund.de/	4636	78	5	16.13
gef. durch	4/5/7/9/11/13/15		Datensatz Nr. 28		
alte URL	http://www-ai.cs.uni-dortmund.de/LEHRE/DIPLOM/diplom_fertig.html Version vom 01.12.2001	6110	245		
erw. URL	http://www-ai.cs.uni-dortmund.de/LEHRE/DIPLOM/diplom_fertig.html	10760	382	32	13.06
gef. durch	1/3/6/7/9/11/13/14/15		Datensatz Nr. 47		
alte URL	http://www-ai.cs.uni-dortmund.de/LEHRE/DIPLOM/diplom_fertig.html	8038	310		
erw. URL	http://www-ai.cs.uni-dortmund.de/LEHRE/DIPLOM/diplom_fertig.html	10760	382	39	12.58
gef. durch	1/2/3/6/7/9/11/13/14		Datensatz Nr. 57		
alte URL	http://dekanat.cs.uni-dortmund.de/ZPA/OeffnungszeitenZPA.shtml Version vom 25.11.2001	2279	78		
erw. URL	http://dekanat.cs.uni-dortmund.de/ZPA/OeffnungszeitenZPA.shtml	2901	74	9	11.54
gef. durch	1/2/3/7/8/9/11/14/15		Datensatz Nr. 29		
alte URL	http://stl-www.cs.uni-dortmund.de/ Version vom 27.11.2001	3483	139		
erw. URL	http://stl-www.cs.uni-dortmund.de/	3415	133	12	8.63
gef. durch	4/5/6/7/8/9/11/12/15		Datensatz Nr. 25		
alte URL	http://www.fi.muni.cz/usr/popelinsky/old.html.utf-8#courses	2586	201		
erw. URL	http://www.fi.muni.cz/usr/popelinsky/old.html.utf-8#courses	2586	201	16	7.96
gef. durch	4/5/6/8/9/10/11/12/13		Datensatz Nr. 49		
alte URL	http://www-ai.cs.uni-dortmund.de/FORSCHUNG/index.html Version vom 24.04.2001	5286	317		
erw. URL	http://www-ai.cs.uni-dortmund.de/FORSCHUNG/index.html	12425	579	25	7.89
gef. durch	4/5/6/8/9/11/13/14/15		Datensatz Nr. 06		
alte URL	http://ls2-www.cs.uni-dortmund.de/ Version vom 21.07.2001	1016	26		
erw. URL	http://ls2-www.cs.uni-dortmund.de/	1079	29	2	7.69
gef. durch	1/2/3/7/9/11/12/13/15		Datensatz Nr. 20		
alte URL	http://www-ai.cs.uni-dortmund.de/SOFTWARE/SVM_LIGHT/ Version vom 25.12.2001	388	13		
erw. URL	http://www-ai.cs.uni-dortmund.de/SOFTWARE/SVM_LIGHT/	680	20	1	7.69
gef. durch	1/2/3/4/5/13/14/15		Datensatz Nr. 45		
alte URL	http://www-ai.cs.uni-dortmund.de/SOFTWARE/IL/ Version vom 24.12.2001	2605	150		
erw. URL	http://www-ai.cs.uni-dortmund.de/SOFTWARE/IL/	4658	254	11	7.33
gef. durch	1/2/3/4/5/7/8/9/10/11/12/13/14/15		Datensatz Nr. 46		
alte URL	http://www-ai.cs.uni-dortmund.de/FORSCHUNG/PROJEKTE/SFB475C4/ Version vom 17.07.2001	1005	46		
erw. URL	http://www-ai.cs.uni-dortmund.de/FORSCHUNG/PROJEKTE/SFB475C4/	1289	54	3	6.52
gef. durch	4/5/6/7/8/9/11/12/14/15		Datensatz Nr. 11		
alte URL	http://www-ai.cs.uni-dortmund.de/FORSCHUNG/ Version vom 01.12.2001	5107	307		
erw. URL	http://www-ai.cs.uni-dortmund.de/FORSCHUNG/	12425	579	15	4.89
gef. durch	3/4/5/6/8/9/11/12/13/14/15		Datensatz Nr. 41		
alte URL	http://www-ai.cs.uni-dortmund.de/FORSCHUNG/index.eng.html Version vom 20.04.2001	5129	293		

	URL	# Zeichen	# Wörter	Wortschwund absolut	Wortschwund normalisiert
erw. URL	http://www-ai.cs.uni-dortmund.de/FORSCHUNG/index.eng.html	5327	290	13	4.44
gef. durch	1/2/4/5/6/8/9/11/12/13/14/15		Datensatz Nr. 07		
alte URL	http://www-ai.cs.uni-dortmund.de/Harvest/brokers/www-ai/query.html Version vom 01.12.2001	590	23		
erw. URL	http://www-ai.cs.uni-dortmund.de/Harvest/brokers/www-ai/query.html	874	30	1	4.35
gef. durch	3/4/5/6/7/8/9/11/12/13/14/15		Datensatz Nr. 38		
alte URL	http://www-ai.cs.uni-dortmund.de/FORSCHUNG/MLTXT/mltxt.eng.html Version vom 20.04.2001	5685	217		
erw. URL	http://www-ai.cs.uni-dortmund.de/FORSCHUNG/MLTXT/mltxt.eng.html	6738	235	9	4.15
gef. durch	1/2/4/5/6/7/9/11/12/13/14/15		Datensatz Nr. 08		
alte URL	http://www-ai.cs.uni-dortmund.de/PERSONAL/morik.html Version vom 17.07.2001	11711	535		
erw. URL	http://www-ai.cs.uni-dortmund.de/PERSONAL/morik.html	18437	764	19	3.55
gef. durch	4/5/6/7/8/9/11/12/13/14/15		Datensatz Nr. 39		
alte URL	http://www-ai.cs.uni-dortmund.de/UNIVERSELL/index.eng.html Version vom 17.07.2001	1228	64		
erw. URL	http://www-ai.cs.uni-dortmund.de/UNIVERSELL/index.eng.html	1506	70	2	3.13
gef. durch	1/2/3/4/5/6/7/9/11/12/13/14/15		Datensatz Nr. 04		
alte URL	http://www-ai.cs.uni-dortmund.de/UNIVERSELL/ Version vom 25.12.2001	1315	66		
erw. URL	http://www-ai.cs.uni-dortmund.de/UNIVERSELL/	1599	72	2	3.03
gef. durch	1/2/3/4/5/6/9/11/12/13/14/15		Datensatz Nr. 03		
alte URL	http://www-ai.cs.uni-dortmund.de/LEHRE/lehre.html Version vom 24.10.2001	783	37		
erw. URL	http://www-ai.cs.uni-dortmund.de/LEHRE/lehre_15_02_2002.html	1065	43	1	2.7
gef. durch	1/2/3/4/5/7/8/9/11/13		Datensatz Nr. 42		
alte URL	http://www-ai.cs.uni-dortmund.de/LEHRE/PG/	925	37		
erw. URL	http://www-ai.cs.uni-dortmund.de/LEHRE/PG/	979	42	1	2.7
gef. durch	1/2/3/4/5/7/9/11/12/13/14		Datensatz Nr. 59		
alte URL	http://dekanat.cs.uni-dortmund.de/Ehemalige/Beitritt.html Version vom 24.12.2001	1693	78		
erw. URL	http://dekanat.cs.uni-dortmund.de/Ehemalige/Beitritt.html	1746	83	2	2.56
gef. durch	1/2/3/4/5/7/8/9/11/12/13/14/15		Datensatz Nr. 30		
alte URL	http://sfbci.cs.uni-dortmund.de/home/German/frameset.html	2171	117		
erw. URL	http://sfbci.cs.uni-dortmund.de/home/German/frameset.html	2171	117	3	2.56
gef. durch	9/15		Datensatz Nr. 60		
alte URL	http://www-ai.cs.uni-dortmund.de/SOFTWARE/ Version vom 24.12.2001	986	44		
erw. URL	http://www-ai.cs.uni-dortmund.de/SOFTWARE/	2176	100	1	2.27
gef. durch	4/5/6/7/8/9/10/11/12/13/14/15		Datensatz Nr. 37		
alte URL	http://www-ai.cs.uni-dortmund.de/logo.html Version vom 25.12.2001	950	44		
erw. URL	http://www-ai.cs.uni-dortmund.de/logo.html	1227	51	1	2.27
gef. durch	1/2/3/4/5/7/8/9/11/12/13/14/15		Datensatz Nr. 02		
alte URL	http://www-ai.cs.uni-dortmund.de/LEHRE/lehrveranstaltungen_alt.html Version vom 01.12.2001	3269	93		
erw. URL	http://www-ai.cs.uni-dortmund.de/LEHRE/lehrveranstaltungen_alt.html	4162	114	2	2.15
gef. durch	1/2/3/4/5/7/9/11/12/13/14/15		Datensatz Nr. 05		
alte URL	http://www-ai.cs.uni-dortmund.de/FORSCHUNG/PROJEKTE/SFB475A4/ Version vom 25.12.2001	8547	375		
erw. URL	http://www-ai.cs.uni-dortmund.de/FORSCHUNG/PROJEKTE/SFB475A4/	11585	479	7	1.87
gef. durch	1/2/3/4/5/7/10/11/12/13/14/15		Datensatz Nr. 10		
alte URL	http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/ Version vom 24.12.2001	9523	381		
erw. URL	http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/	9950	393	5	1.31
gef. durch	3/4/5/7/8/9/11/13/14/15		Datensatz Nr. 40		
alte URL	http://www.aifb.uni-karlsruhe.de/Lehrangebot/Sommer2002/InfoC/script.html	8819	217		
erw. URL	http://www.aifb.uni-karlsruhe.de/Lehrangebot/Sommer2002/InfoC/script.html	8819	216	2	0.92
gef. durch	3/7/9/12		Datensatz Nr. 50		
alte URL	http://www-ai.cs.uni-dortmund.de/LEHRE/SEMINARE/INFORMATIONSEXTRAKTION/informationsExtraktion.html	8379	392		

	URL	# Zeichen	# Wörter	Wortschwund absolut	Wortschwund normalisiert
erw. URL	http://www-ai.cs.uni-dortmund.de/LEHRE/SEMINARE/INFORMATIONSEXTRAKTION/informationsExtraktion.html	8369	391	2	0.51
gef. durch	3/4/5/6/7/8/9/10/11/12/13		Datensatz Nr. 54		
alte URL	http://www-ai.cs.uni-dortmund.de/FORSCHUNG/REPORTS/reports.eng.html Version vom 24.04.2001	25522	1026		
erw. URL	http://www-ai.cs.uni-dortmund.de/FORSCHUNG/REPORTS/reports.eng.html	27590	1111	1	0.1
gef. durch	1/2/3/4/5/6/7/8/9/10/11/12/14/15		Datensatz Nr. 13		
alte URL	http://www.co.umist.ac.uk/dsd2002/ Version vom 07.12.2001	1328	68		
erw. URL	http://www.co.umist.ac.uk/dsd2002/	1470	75	0	0.0
gef. durch	1/2/3/4/5/7/8/9/12/15		Datensatz Nr. 18		
alte URL	http://www-ai.cs.uni-dortmund.de/LEHRE/PROMOTION/promotion_fertig.html	1730	94		
erw. URL	http://www-ai.cs.uni-dortmund.de/LEHRE/PROMOTION/promotion_fertig.html	1730	94	0	0.0
gef. durch	1/2/3/9/10/11/12/14		Datensatz Nr. 58		
alte URL	http://www-ai.cs.uni-dortmund.de/LEHRE/DIPLOM/OFFEN/	3688	156		
erw. URL	http://www-ai.cs.uni-dortmund.de/LEHRE/DIPLOM/OFFEN/	4140	179	0	0.0
gef. durch	1/2/3/4/5/6/7/8/9/11/12/13/14/15		Datensatz Nr. 56		
alte URL	http://www-ai.cs.uni-dortmund.de/LEHRE/VORLESUNGEN/MLRN/ml.html	2223	118		
erw. URL	http://www-ai.cs.uni-dortmund.de/LEHRE/VORLESUNGEN/MLRN/ml.html	2223	118	0	0.0
gef. durch	1/2/3/4/5/7/8/9/11/12/13		Datensatz Nr. 55		
alte URL	http://www.cs.bris.ac.uk/Teaching/Resources/COMS30106/	2137	129		
erw. URL	http://www.cs.bris.ac.uk/Teaching/Resources/COMS30106/	2137	129	0	0.0
gef. durch	1/2/3/4/5/6/8/11/12/13/15		Datensatz Nr. 52		
alte URL	http://www.csd.abdn.ac.uk/~pedwards/teaching/CS5505/slides.html	3660	234		
erw. URL	http://www.csd.abdn.ac.uk/~pedwards/teaching/CS5505/slides.html	3660	234	0	0.0
gef. durch	1/2/3/6/7/9/10/12		Datensatz Nr. 51		
alte URL	http://fsinfo.cs.uni-dortmund.de/Studium/ Version vom 19.12.2001	2890	37		
erw. URL	http://fsinfo.cs.uni-dortmund.de/Studium/	3812	65	0	0.0
gef. durch	1/2/3/4/5/7/9/11/12/13/14/15		Datensatz Nr. 17		
alte URL	http://www.cs.kuleuven.ac.be/~hendrik/ML/	1154	65		
erw. URL	http://www.cs.kuleuven.ac.be/~hendrik/ML/	1154	65	0	0.0
gef. durch	1/2/3/4/5/6/7/8/9/11/12/15		Datensatz Nr. 48		
alte URL	http://dekanat.cs.uni-dortmund.de/Studierende/Index.html Version vom 24.12.2001	73	5		
erw. URL	http://dekanat.cs.uni-dortmund.de/Studierende/Index.html	66	5	0	0.0
gef. durch	9/13/14/15		Datensatz Nr. 16		
alte URL	http://www.eunite.org/ Version vom 30.11.2001	134	8		
erw. URL	http://www.eunite.org/	127	8	0	0.0
gef. durch	1/2/3/4/5/6/7/11/12/13/15		Datensatz Nr. 09		

C.2 WDR

	URL	# Zeichen	# Wörter	Wortschwund absolut	Wortschwund normalisiert
alte URL	http://www.wdr.de/themen/computer/angeklickt/tagestipp/tagestipp.jhtml	3909	132		
erw. URL	http://www.wdr.de/themen/computer/angeklickt/tagestipp/tagestipp.jhtml	3892	130	71	53.79
gef. durch	nicht gefunden		Datensatz Nr. 39		
alte URL	http://www.wdr.de/themen/homepages/popkomm2003.jhtml	9463	281		
erw. URL	http://www.wdr.de/themen/homepages/popkomm2003.jhtml	9708	304	86	30.6
gef. durch	6/7/8/9/11/13		Datensatz Nr. 31		
alte URL	http://www.wdr.de/themen/panorama/2/stromausfall_bielefeld/index.jhtml	3919	115		
erw. URL	http://www.wdr.de/themen/panorama/2/stromausfall_bielefeld/index.jhtml	3910	114	19	16.52
gef. durch	2/4/5/7/8/9/10/11/12/13		Datensatz Nr. 45		
alte URL	http://www.wdr.de/online/jobs/jobzeit/index.phtml?rubrikenstyle=wirtschaft	2104	108		

	URL	# Zeichen	# Wörter	Wortschwund absolut	Wortschwund normalisiert
erw. URL	http://www.wdr.de/online/jobs/jobzeit/index.phtml?rubrikenstyle=wirtschaft	1975	98	17	15.74
gef. durch	1/2/3/4/5/6/7/8/9/11/12/13		Datensatz Nr.19		
alte URL	http://www.wdr.de/themen/computer/software/virenticker/index.jhtml	24690	779		
erw. URL	http://www.wdr.de/themen/computer/software/virenticker/index.jhtml	23225	747	112	14.38
gef. durch	1/2/3/4/5/6/7/8/9/11/12/13		Datensatz Nr. 35		
alte URL	http://www.wdr.de/themen/homepages/unicef.jhtml	8461	255		
erw. URL	http://www.wdr.de/themen/homepages/unicef.jhtml	7817	242	30	11.76
gef. durch	2/3/4/5/7/8/9/10/11/12/13		Datensatz Nr. 32		
alte URL	http://www.wdr.de/themen/panorama/2/verdaechtiger_koffer_koeln/index.jhtml	5448	156		
erw. URL	http://www.wdr.de/themen/panorama/2/verdaechtiger_koffer_koeln/index.jhtml	5437	154	18	11.54
gef. durch	1/2/4/5/6/7/8/9/11/13		Datensatz Nr. 61		
alte URL	http://www.wdr.de/themen/homepages/kleine_anfrage.jhtml	8972	251		
erw. URL	http://www.wdr.de/themen/homepages/kleine_anfrage.jhtml	9635	271	28	11.16
gef. durch	3/4/5/6/7/9/11/12/13		Datensatz Nr. 22		
alte URL	http://www.wdr.de/themen/verkehr/strasse/tunnel_b236/index.jhtml	4440	116		
erw. URL	http://www.wdr.de/themen/verkehr/strasse/tunnel_b236/index.jhtml	4487	116	12	10.34
gef. durch	2/5/6/7/8/9/11/12/13		Datensatz Nr. 57		
alte URL	http://www.wdr.de/themen/kultur/quiz/quiz_rechtschreibreform.jhtml	6454	179		
erw. URL	http://www.wdr.de/themen/kultur/quiz/quiz_rechtschreibreform.jhtml	6420	175	18	10.06
gef. durch	1/2/3/5/6/7/8/9/11/12/13		Datensatz Nr. 27		
alte URL	http://www.wdr.de/themen/panorama/2/bombenentschaerfungen_nrw_2002/index.jhtml	6503	181		
erw. URL	http://www.wdr.de/themen/panorama/2/bombenentschaerfungen_nrw_2002/index.jhtml	6482	179	17	9.39
gef. durch	1//3/4/9/10/11/12/13		Datensatz Nr. 49		
alte URL	http://www.wdr.de/themen/wirtschaft/wirtschaftsbranche/ford/krise.jhtml	6645	172		
erw. URL	http://www.wdr.de/themen/wirtschaft/wirtschaftsbranche/ford/krise.jhtml	6627	168	16	9.3
gef. durch	1/2/3/4/5/6/7/8/9/10/11/13		Datensatz Nr. 14		
alte URL	http://www.wdr.de/themen/panorama/kriminalitaet02/missbrauch_moenchengladbach/index.jhtml	5954	182		
erw. URL	http://www.wdr.de/themen/panorama/kriminalitaet02/missbrauch_moenchengladbach/index.jhtml	5945	184	16	8.79
gef. durch	3/7/8/9/10/11/13		Datensatz Nr. 46		
alte URL	http://www.wdr.de/themen/panorama/wetter/sommer_2003/waldbraende.jhtml	8275	217		
erw. URL	http://www.wdr.de/themen/panorama/wetter/sommer_2003/waldbraende.jhtml	8266	216	19	8.76
gef. durch	1/2/3/5/7/8/9/11/12/13		Datensatz Nr. 48		
alte URL	http://www.wdr.de/themen/verkehr/strasse/radarwarngeraete/autobahn_warnung.jhtml	5477	139		
erw. URL	http://www.wdr.de/themen/verkehr/strasse/radarwarngeraete/autobahn_warnung.jhtml	5513	142	12	8.63
gef. durch	2/3/4/5/6/7/8/9/12		Datensatz Nr.59		
alte URL	http://www.wdr.de/themen/wirtschaft/1/haushaltsgeraete/index.jhtml	6418	164		
erw. URL	http://www.wdr.de/themen/wirtschaft/1/haushaltsgeraete/index.jhtml	6404	162	14	8.54
gef. durch	1/2/3/4/5/6/7/8/9/12		Datensatz Nr. 13		
alte URL	http://www.wdr.de/themen/kultur/rundfunk/lilipuz_sommertour2003/index.jhtml	6965	176		
erw. URL	http://www.wdr.de/themen/kultur/rundfunk/lilipuz_sommertour2003/index.jhtml	6946	174	14	7.95
gef. durch	1/2/3/5/6/7/8/9/11/12/13		Datensatz Nr. 24		
alte URL	http://www.wdr.de/themen/wirtschaft/geld-_und_kreditwesen/westlb/index_030806.jhtml	7624	207		
erw. URL	http://www.wdr.de/themen/wirtschaft/geld-_und_kreditwesen/westlb/index_030806.jhtml	7610	202	16	7.73
gef. durch	1/2/3/4/5/6/7/8/9/11/12/13		Datensatz Nr. 17		
alte URL	http://www.wdr.de/themen/panorama/wetter/sommer_2003/hitze.jhtml	7937	252		
erw. URL	http://www.wdr.de/themen/panorama/wetter/sommer_2003/hitze.jhtml	7928	250	19	7.54
gef. durch	1/2/3/5/6/7/8/9/11/12/13		Datensatz Nr. 50		
alte URL	http://www.wdr.de/themen/panorama/wetter/sommer_2003/abkuehlung_030814.jhtml	7759	222		

	URL	# Zeichen	# Wörter	Wortschwund absolut	Wortschwund normalisiert
erw. URL	http://www.wdr.de/themen/panorama/wetter/sommer_2003/abkuehlung_030814.jhtml	7748	221	16	7.21
gef. durch	1/2/3/5/6/7/8/9/11/12/13		Datensatz Nr. 47		
alte URL	http://www.wdr.de/themen/wirtschaft/arbeit_und_tarifwesen/ich_ag/erfolg.jhtml	8597	225		
erw. URL	http://www.wdr.de/themen/wirtschaft/arbeit_und_tarifwesen/ich_ag/erfolg.jhtml	8579	222	16	7.11
gef. durch	1/2/3/4/5/6/7/8/9/10/11/12/13		Datensatz Nr. 16		
alte URL	http://www.wdr.de/themen/wirtschaft/wirtschaftsbranche/rwe/usa_drohen.jhtml	7027	225		
erw. URL	http://www.wdr.de/themen/wirtschaft/wirtschaftsbranche/rwe/usa_drohen.jhtml	7009	222	16	7.11
gef. durch	1/2/3/4/5/7/8/9/10/11/12/13		Datensatz Nr. 15		
alte URL	http://www.wdr.de/themen/homepages/olympia_2012.jhtml	8194	183		
erw. URL	http://www.wdr.de/themen/homepages/olympia_2012.jhtml	8200	188	13	7.1
gef. durch	1/2/3/4/5/7/8/9/11/12/13		Datensatz Nr. 42		
alte URL	http://www.wdr.de/themen/freizeit/freizeitgestaltung/mundorgel_50jahre/interview_guildo_horn.jhtml	6467	189		
erw. URL	http://www.wdr.de/themen/freizeit/freizeitgestaltung/mundorgel_50jahre/interview_guildo_horn.jhtml	6457	188	13	6.88
gef. durch	1/2/3/4/5/6/7/8/9/11/12/13		Datensatz Nr. 56		
alte URL	http://www.wdr.de/themen/politik/international/soldaten_afghanistan/landung_koeln.jhtml	9464	244		
erw. URL	http://www.wdr.de/themen/politik/international/soldaten_afghanistan/landung_koeln.jhtml	9440	245	16	6.56
gef. durch	1/2/3/6/7/8/9/11/12/13		Datensatz Nr. 04		
alte URL	http://www.wdr.de/themen/panorama/lifestyle/modemesse_reevolutions_2003/sommer.jhtml	8815	262		
erw. URL	http://www.wdr.de/themen/panorama/lifestyle/modemesse_reevolutions_2003/sommer.jhtml	8794	260	17	6.49
gef. durch	3/4/5/6/7/8/9/11/12/13		Datensatz Nr. 28		
alte URL	http://www.wdr.de/themen/wirtschaft/arbeit_und_tarifwesen/hartz_reformen/index.jhtml	6454	204		
erw. URL	http://www.wdr.de/themen/wirtschaft/arbeit_und_tarifwesen/hartz_reformen/index.jhtml	6436	203	13	6.37
gef. durch	1/2/3/4/5/7/8/9/10/11/13		Datensatz Nr. 12		
alte URL	http://www.wdr.de/themen/gesundheit/gesundheitswesen/gesetzliche_krankenkassen/sparzwang.jhtml	6844	189		
erw. URL	http://www.wdr.de/themen/gesundheit/gesundheitswesen/gesetzliche_krankenkassen/sparzwang.jhtml	6855	197	12	6.35
gef. durch	1/2/3/4/5/7/8/9/10/11/12/13		Datensatz Nr. 52		
alte URL	http://www.wdr.de/themen/sport/1/holzfaeller/index.jhtml	9536	272		
erw. URL	http://www.wdr.de/themen/sport/1/holzfaeller/index.jhtml	9532	271	17	6.25
gef. durch	3/4/5/6/7/8/9/10/11/12/13		Datensatz Nr. 41		
alte URL	http://www.wdr.de/themen/kultur/1/kinderschutzbund//interview.jhtml	8817	277		
erw. URL	http://www.wdr.de/themen/kultur/1/kinderschutzbund/interview.jhtml	8786	274	17	6.14
gef. durch	1/2/3/4/5/6/7/8/9/10/11/12/13		Datensatz Nr. 29		
alte URL	http://www.wdr.de/themen/politik/deutschland/stabilitaetspakt/index.jhtml	7336	214		
erw. URL	http://www.wdr.de/themen/politik/deutschland/stabilitaetspakt/index.jhtml	7340	215	13	6.07
gef. durch	1/2/3/4/5/7/8/9/10/11/12/13		Datensatz Nr. 03		
alte URL	http://www.wdr.de/themen/politik/nrw/gemeindefinanzierung/ruettgers.jhtml	7443	233		
erw. URL	http://www.wdr.de/themen/politik/nrw/gemeindefinanzierung/ruettgers.jhtml	7420	237	14	6.01
gef. durch	1/2/3/4/7/8/9/10/11/12/13		Datensatz Nr. 07		
alte URL	http://www.wdr.de/themen/kultur/bildung_und_erziehung/bafoeg_missbrauch/index.jhtml	7755	221		
erw. URL	http://www.wdr.de/themen/kultur/bildung_und_erziehung/bafoeg_missbrauch/index.jhtml	7743	223	13	5.88
gef. durch	1/2/3/4/5/6/7/8/9/12		Datensatz Nr. 25		
alte URL	http://www.wdr.de/themen/panorama/2/lottojackpot_nrw/index.jhtml	8145	225		
erw. URL	http://www.wdr.de/themen/panorama/2/lottojackpot_nrw/index.jhtml	8124	227	13	5.78
gef. durch	1/3/4/5/7/8/9/10/11/12/13		Datensatz Nr. 43		
alte URL	http://www.wdr.de/themen/verkehr/schiene/deutsche_bahn/maengel.jhtml	7297	200		

	URL	# Zeichen	# Wörter	Wortschwund absolut	Wortschwund normalisiert
erw. URL	http://www.wdr.de/themen/verkehr/schiene/deutsche_bahn/maengel.jhtml	7333	201	11	5.5
gef. durch	1/2/3/5/6/7/8/9/11/13		Datensatz Nr. 58		
alte URL	http://www.wdr.de/themen/politik/nrw/interview_2003/pinkwart/interview.jhtml	9181	259		
erw. URL	http://www.wdr.de/themen/politik/nrw/interview_2003/pinkwart/interview.jhtml	9163	262	14	5.41
gef. durch	1/2/3/4/5/6/7/8/9/10/11/12/13		Datensatz Nr. 05		
alte URL	http://www.wdr.de/themen/freizeit/freizeitgestaltung/future_parade/index.jhtml	9286	243		
erw. URL	http://www.wdr.de/themen/freizeit/freizeitgestaltung/future_parade/index.jhtml	9276	244	13	5.35
gef. durch	1/2/3/4/5/6/7/8/9/10/11/12/13		Datensatz Nr. 55		
alte URL	http://www.wdr.de/themen/wirtschaft/wirtschaftsbranche/luftfahrt/flughafen_weeze.jhtml	9805	284		
erw. URL	http://www.wdr.de/themen/wirtschaft/wirtschaftsbranche/luftfahrt/flughafen_weeze.jhtml	9791	281	15	5.28
gef. durch	1/2/3/4/5/7/8/9/10/11/13		Datensatz Nr. 18		
alte URL	http://www.wdr.de/themen/politik/nrw/gemeindefinanzierung/reaktionen.jhtml	10041	267		
erw. URL	http://www.wdr.de/themen/politik/nrw/gemeindefinanzierung/reaktionen.jhtml	10018	270	14	5.24
gef. durch	1/2/3/7/8/9/10/11/12/13		Datensatz Nr. 06		
alte URL	http://www.wdr.de/themen/politik/nrw/steinkohle/rag_ausgliederung.jhtml	9521	231		
erw. URL	http://www.wdr.de/themen/politik/nrw/steinkohle/rag_ausgliederung.jhtml	9522	231	12	5.19
gef. durch	1/2/3/4/5/7/8/9/10/11/12/13		Datensatz Nr. 01		
alte URL	http://www.wdr.de/themen/homepages/irak.jhtml	8454	252		
erw. URL	http://www.wdr.de/themen/homepages/irak.jhtml	8483	258	13	5.16
gef. durch	1/2/3/4/7/8/9/11/12/13		Datensatz Nr. 10		
alte URL	http://www.wdr.de/themen/panorama/kriminalitaet02/gladbecker_geiseldrama_1988/index.jhtml	11164	377		
erw. URL	http://www.wdr.de/themen/panorama/kriminalitaet02/gladbecker_geiseldrama_1988/index.jhtml	11155	378	18	4.77
gef. durch	1/2/3/4/5/7/8/9/11/12/13		Datensatz Nr. 44		
alte URL	http://www.wdr.de/themen/kultur/personen/ruge/index.jhtml	11214	361		
erw. URL	http://www.wdr.de/themen/kultur/personen/ruge/index.jhtml	11183	358	17	4.71
gef. durch	1/2/3/5/6/7/8/9/10/11/12/13		Datensatz Nr. 26		
alte URL	http://www.wdr.de/themen/gesundheit/gesundheitswesen/gesetzliche_krankenkassen/bkk_beitragserhoehung.jhtml	10035	300		
erw. URL	http://www.wdr.de/themen/gesundheit/gesundheitswesen/gesetzliche_krankenkassen/bkk_beitragserhoehung.jhtml	10042	303	14	4.67
gef. durch	1/2/3/4/5/7/8/9/10/11/12/13		Datensatz Nr. 53		
alte URL	http://www.wdr.de/themen/computer/internet/sicherheit/exploit.jhtml	6834	237		
erw. URL	http://www.wdr.de/themen/computer/internet/sicherheit/exploit.jhtml	6843	238	11	4.64
gef. durch	1/2/3/4/5/6/7/8/9/10/11/12/13		Datensatz Nr. 37		
alte URL	http://www.wdr.de/themen/freizeit/freizeitgestaltung/insel_sommer/monkeys_island.jhtml	8212	259		
erw. URL	http://www.wdr.de/themen/freizeit/freizeitgestaltung/insel_sommer/monkeys_island.jhtml	8442	265	11	4.25
gef. durch	1/2/3/4/5/7/9/10/11/12/13		Datensatz Nr. 54		
alte URL	http://www.wdr.de/themen/gesundheit/1/antipasti/index.jhtml	4686	121		
erw. URL	http://www.wdr.de/themen/gesundheit/1/antipasti/index.jhtml	4687	122	5	4.13
gef. durch	1/2/3/4/5/7/8/9/11/12/13		Datensatz Nr. 51		
alte URL	http://www.wdr.de/themen/computer/internet/blaster/index.jhtml	9375	246		
erw. URL	http://www.wdr.de/themen/computer/internet/blaster/index.jhtml	9386	247	10	4.07
gef. durch	1/2/3/4/5/6/7/8/9/11/12/13		Datensatz Nr. 34		
alte URL	http://www.wdr.de/themen/verkehr/wasser/niedrigwasser_rhein_ruhr/niedrigwasser_0308.jhtml	11719	324		
erw. URL	http://www.wdr.de/themen/verkehr/wasser/niedrigwasser_rhein_ruhr/niedrigwasser_0308.jhtml	11766	327	12	3.7

	URL	# Zeichen	# Wörter	Wortschwund absolut	Wortschwund normalisiert
gef. durch	1/2/4/5/7/8/9/10/11/12/13		Datensatz Nr. 60		
alte URL	http://www.wdr.de/themen/homepages/d_usa.jhtml	12685	356		
erw. URL	http://www.wdr.de/themen/homepages/d_usa.jhtml	12714	365	11	3.09
gef. durch	1/2/3/7/8/9/10/11/12/13		Datensatz Nr. 11		
alte URL	http://www.wdr.de/themen/computer/schiebwoche/2003/index_33.jhtml	9913	301		
erw. URL	http://www.wdr.de/themen/computer/schiebwoche/2003/index_33.jhtml	9927	303	9	2.99
gef. durch	1/2/3/4/5/6/7/8/9/11/12/13		Datensatz Nr. 36		
alte URL	http://www.wdr.de/themen/politik/international/elfter_september/verfassungsschutz_cia/index.jhtml	11961	386		
erw. URL	http://www.wdr.de/themen/politik/international/elfter_september/verfassungsschutz_cia/index.jhtml	11946	393	11	2.85
gef. durch	1/2/3/4/5/6/7/8/9/10/11/12/13		Datensatz Nr. 02		
alte URL	http://www.wdr.de/themen/politik/nrw/muellaffaere_spd/inhalt.jhtml	14190	481		
erw. URL	http://www.wdr.de/themen/politik/nrw/muellaffaere_spd/inhalt.jhtml	14189	481	13	2.7
gef. durch	1/2/3/4/5/7/9/11/12/13		Datensatz Nr. 09		
alte URL	http://www.wdr.de/themen/forschung/technik/solarflitzer/index.jhtml	9980	271		
erw. URL	http://www.wdr.de/themen/forschung/technik/solarflitzer/index.jhtml	9986	273	7	2.58
gef. durch	1/2/3/4/5/7/9/10/11/12/13		Datensatz Nr. 20		
alte URL	http://www.wdr.de/themen/forschung/interdisziplinaer/virtuelle_bibliothek/index.jhtml	8635	288		
erw. URL	http://www.wdr.de/themen/forschung/interdisziplinaer/virtuelle_bibliothek/index.jhtml	8641	290	7	2.43
gef. durch	1/2/3/4/5/6/7/8/9/10/11/12/13		Datensatz Nr. 21		
alte URL	http://www.wdr.de/themen/kultur/1/linkshaender_tag_2003/gaestebuch.jhtml	18410	644		
erw. URL	http://www.wdr.de/themen/kultur/1/linkshaender_tag_2003/gaestebuch.jhtml	18376	642	15	2.33
gef. durch	1/2/3/4/5/6/7/8/9/11/12/13		Datensatz Nr. 23		
alte URL	http://www.wdr.de/themen/computer/angeklickt/webtv/index.jhtml	9990	418		
erw. URL	http://www.wdr.de/themen/computer/angeklickt/webtv/index.jhtml	10315	436	9	2.15
gef. durch	3/4/6/7/8/9/11/12/13		Datensatz Nr. 40		
alte URL	http://www.wdr.de/themen/politik/nrw/moellemann/inhalt.jhtml	23390	653		
erw. URL	http://www.wdr.de/themen/politik/nrw/moellemann/inhalt.jhtml	23389	652	14	2.14
gef. durch	1/2/3/4/5/7/9/11/12/13		Datensatz Nr. 08		
alte URL	http://www.wdr.de/themen/verkehr/schiene/metrorapid/inhalt.jhtml	15267	495		
erw. URL	http://www.wdr.de/themen/verkehr/schiene/metrorapid/inhalt.jhtml	15314	496	10	2.02
gef. durch	3/6/7/9/11/12/13		Datensatz Nr. 62		
alte URL	http://www.wdr.de/themen/kultur/netzkultur/heimatinseln/inseln.jhtml	4488	102		
erw. URL	http://www.wdr.de/themen/kultur/netzkultur/heimatinseln/inseln.jhtml	5443	118	2	1.96
gef. durch	1/2/3/4/5/6/7/8/9/11/12/13		Datensatz Nr. 30		
alte URL	http://www.wdr.de/themen/kultur/netzkultur/nrw_privat/wohnungen.jhtml	4201	125		
erw. URL	http://www.wdr.de/themen/kultur/netzkultur/nrw_privat/wohnungen.jhtml	4202	125	2	1.6
gef. durch	1/2/3/4/5/7/8/9/11/12/13		Datensatz Nr. 33		
alte URL	http://www.wdr.de/themen/computer/internet/barrierefreies_internet/internetcafe_reportage.jhtml	11065	337		
erw. URL	http://www.wdr.de/themen/computer/internet/barrierefreies_internet/internetcafe_reportage.jhtml	11177	342	4	1.19
gef. durch	1/2/3/4/5/6/7/8/9/10/11/12/13		Datensatz Nr. 38		