UNIVERSITÄT KOBLENZ · LANDAU
Fachbereich 4: Informatik

Institut für Informatik
Arbeitsgruppe Staab/Sure
<isweb>

Bachelorarbeit

Suche und Ranking von getaggten Medien in Bezug auf eine Eingabe-Webseite

vorgelegt von:

Student:	Hagen Metzler
Studiengang:	Computervisualistik
Matrikelnummer:	200210237
Geburtsdatum:	04.06.1980
Adresse:	Von-Are-Str. 32, 56645 Nickenich
E-Mail:	haegarm@uni-koblenz.de

Prof. Dr. Steffen Staab
Institut für Informatik, Fachbereich Informatik

Dipl.-Inform. Klaas Dellschaft
Institut für Informatik, Fachbereich Informatik

September 2009

Zusammenfassung

Wenn ein Benutzer im Internet surft, hat er manchmal das Bedürfnis nach relevanten Medien zu der momentan angezeigten Webseite zu suchen. Google bietet bereits eine Suche nach ähnlichen Webseiten an. Diese Suchfunktion wird allerdings nicht für die Suche nach Fotos und Videos angeboten. Im Rahmen dieser Bachelorarbeit ist eine solche Ähnlichkeitssuche im Rahmen von MyTag entstanden. Dies ermöglicht die Suche nach Fotos, Videos, Bookmarks und Literaturreferenzen in Bezug auf eine Eingabe-Webseite.

Dafür wurde ein Firefox Add-on entwickelt. Mit einem Klick auf einen Button in der Menüleiste des Browsers wird automatisch der Textinhalt einer beliebigen Webseite analysiert und eine Suchanfrage an MyTag gestellt. Anschließend können die gefundenen Ressourcen anhand ihrer Kosinus-Ähnlichkeit zur Eingabe-Webseite sortiert werden. Diese Methode des Rankings kann insbesondere hilfreich sein, falls die verwendeten Suchbegriffe mehrdeutig sind.

Im Anschluss wurde eine Evaluation durchgeführt. Auf dieser Grundlage wurde die Qualität der Suche bewertet und es wurden Verbesserungsvorschläge gegeben.

Abstract

Search and ranking of tagged media in relation to an input web page

If a user surfs the Internet, he sometimes has the need to search for media relevant to the current web page. Google already offers a search option that searches for similar web pages. However, this option does not include a search for photos and videos. A corresponding resemblance search was developed within the scope of MyTag while writing this bachelor's thesis. This search engine enables the user to search for photos, videos, bookmarks and literary references in relation to any input web page.

For this search option I developed a Firefox add-on. With a click on a button in the browser's links toolbar, the text contents of any web page are automatically analyzed and a search inquiry is put to MyTag. Than, the retrieved resources can be sorted with regard to their specific cosine resemblance in accordance to the input web page. This ranking method can be helpful if the used search words are ambiguous.

An evaluation which assessed the quality of the search was carried out and, as a result, improvements were suggestioned.

Inhalt

Verzeichnis der Abbildungen

Verzeichnis der Tabellen

Verzeichnis des Anhangs

Verzeichnis der Abkürzungen

HTML	Hypertext Markup Language
PHP	Hypertext Preprocessor
SOAP	Simple Object Access Protocol
URL	Uniform Resource Locator
XML	Extensible Markup Language

1 Einleitung

Wenn ein Benutzer im Internet surft, hat er manchmal das Bedürfnis nach relevanten Medien zu der momentan angezeigten Webseite zu suchen. Google bietet bereits eine Suche nach ähnlichen Webseiten an. Diese Suchfunktion wird allerdings nicht für die Suche nach Fotos und Videos angeboten. Im Rahmen dieser Bachelorarbeit ist eine solche Ähnlichkeitssuche im Rahmen von MyTag umgesetzt worden. Dies ermöglicht die Suche nach Fotos, Videos, Bookmarks und Literaturreferenzen in Bezug auf eine Eingabe-Webseite.

Dafür wurde ein Firefox Add-on entwickelt. Mit einem Klick auf einen Button in der Menüleiste des Browsers wird automatisch der Textinhalt einer beliebigen Webseite analysiert und eine Suchanfrage an MyTag gestellt. Es öffnet sich dann ein neues Fenster mit den von MyTag gelieferten Suchergebnissen. Das Firefox Add-on analysiert, strukturiert und interpretiert den Inhalt bzw. das Thema einer Webseite. Diese Information wird direkt als Suchanfrage an MyTag gestellt. Für die Umsetzung des Add-ons haben sich bereits existierende Systeme als nützlich erwiesen, die den Inhalt einer Webseite auf wenige Schlagwörter zusammenfassen können. Diese werden zum Beispiel häufig zur automatischen Verschlagwortung von Artikeln in Content- Managment-Systemen genutzt. Von diesen Schlagwörter werden passende Suchbegriffe identifiziert, mit denen die Suchanfrage an MyTag gestellt werden kann.

Des weiteren können die Schlagwörter von MyTag für ein verbessertes Ranking genutzt werden. Dafür wird die Kosinus-Ähnlichkeit zwischen der Menge der Schlagwörter der Webseite und der Menge der Schlagwörter der einzelnen, gefundenen Ressourcen in MyTag, berechnet. Anschließend können die gefundenen Ressourcen anhand ihrer Kosinus-Ähnlichkeit zur Eingabe-Webseite sortiert werden. Diese Methode des Rankings kann insbesondere hilfreich sein, falls die verwendeten Suchbegriffe mehrdeutig sind.

Die Qualität der Suche wurde anschließend noch im Rahmen eines Benutzertests mit mehreren Testpersonen evaluiert, inwiefern durch das eben beschriebene Vorgehen

relevante Ergebnisse zurückgeliefert werden können. Basierend auf dem Benutzertest werden Empfehlungen ausgesprochen, wie das Add-on weiter verbessert werden kann.

Die Arbeit beginnt in Kapitel 2 mit der Vorstellung von zwei Beispielen, die eine ähnliche Funktionalität haben und deren Idee sich als hilfreich erweisen könnte. In Kapitel 3 wird das nötige Handwerkzeug zur Erstellung des beschriebenen Add-ons erklärt. Die verschiedenen Methoden, die als Vorwissen für die spätere Implementierung dienen, werden hier zunächst beschrieben. Auf die Funktionsweise von Tagging-Systemen und MyTag wird in Kapitel 4 eingegangen. In Kapitel 5 wird das Vorwissen aus Kapitel 3 in die Tat umgesetzt und die Implementierung des Add-ons beschrieben. Anschließend wurde die Evaluation durchgeführt. Die Auswertung dazu wird in Kapitel 6 ausgeführt. Zum Abschluss folgt noch ein Fazit der Arbeit mit Empfehlungen zur Weiterentwicklung.

2 Related Work

2.1 Google-Ähnlichkeitssuche

Google bietet unter der „erweiterten Suche“[1] eine seitenspezifische Suche an (siehe Abb. 1). Hier kann nach Seiten gesuchten werden, die ähnlich zu der eingegeben URL sind. Die gleiche Funktionalität erreicht man schon auf der Google-Startseite mit der Eingabe „related:http://www.domain.xy/url“.

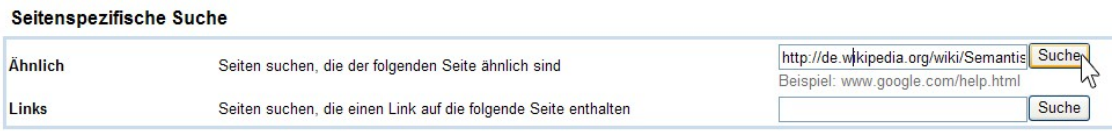

Abbildung 1: Seitenspezifische Suche bei Google
(http://www.google.de/advanced_search)

So erhält man bei der Eingabe von „related:http://de.wikipedia.org/wiki/Semantisches_Web“, mit der URL des Wikipedia-Artikels zum Thema „semantisches web“, 28 verschiedene Einträge mit Webseiten über „semantic web“. Die gefundenen Seiten sind auch im Suchergebnis enthalten, als hätte man ganz normal nach „semantic web“ bei Google gesucht. Jedoch erhält man hier

[1] http://www.google.de/advanced_search

49.800.000 Suchergebnisse. Das lässt darauf schließen, dass die Suche mit den Eingabeterme weniger spezifisch ist.

Google scheint also den Inhalt der Eingabe-Webseite zu analysieren und dabei auch die Phrase „semantisches web“ bzw. „semantic web“ und den Lexikon-ähnlichen Eintrag zu erkennen.

Unter den Suchergebnissen der Ähnlichkeitssuche können auch YouTube-Videos sein, da YouTube[2] zum Google-Konzern gehört. Weiterhin wäre es möglich, weiterführend die Google-Bildersuche zu nutzen. Eine direkte Suche nach getaggten Medien ist an dieser Stelle allerdings nicht durchführbar.

2.2 Das Keyword-Analysis-Tool

Ein Beispiel für das Generieren von Keyphrases über eine Webseite ist das Keyword-Analysis-Tool[3] von Andy Hoskinson. Für eine eingegebene URL einer Webseite gibt diese Anwendung die häufigsten Wörter dieser Webseite nach Häufigkeit sortiert aus. Zudem werden häufig vorkommende Wortpaarungen als Phrasen erkannt und ebenfalls nach Häufigkeit sortiert angezeigt (siehe Abbildung 2). Die so genannten Stoppwörter werden zuvor entfernt. Stoppwörter sind in jeder Sprache vorkommende sehr häufige Wörter. In Kapitel 3.2.2 werden Stoppwörter detailliert erklärt. Dieses Tool soll nach Aussage des Entwicklers, Webmaster dabei unterstützen, geeignete Keywords für die Suchmaschinenoptimierung ihrer Webseiten zu finden. Außerdem ist es geeignet, den Inhalt langer Texte zu analysieren. Dafür kann der Text der Webseite angezeigt werden, wobei eine angeklickte Keyphrase gelb im Text markiert ist.
Der User könnte hier zum Beispiel die Möglichkeit erhalten, mit einem Klick auf ein Keyword oder eine Keyphrase eine Suche bei Google zu starten. Ein weiterführende Suche nach passenden Medien oder Webseiten fehlt aber auch diesem Service.

[2] http://www.youtube.com
[3] http://seokeywordanalysis.com/

Keywords		Keyphrases	
Keyword	Frequency	Keyphrase	Frequency
Java	82	Java Virtual Machine	9
JVM	39	Java bytecode	7
Machine	26	Java Platform	4
Sun	26	JIT compiler	4
Virtual	25	Java Virtual Machine Specification	4
bytecode	21	instruction set	3
code	21	Execution environment	3
languages	15	dynamic languages	3
edit	13	remote code	3
specification	13	bytecode compilers	3

Abbildung 2: SEO-Keyword-Analyzer

Beispielsuche für die URL http://en.wikipedia.org/wiki/Java_Virtual_Machine

3 Keyphrase Extraction

Die Indexierung oder auch Verschlagwortung fasst die Inhalte eines Dokumentes auf wenige Wörter zusammen. Der Nutzen liegt zum Beispiel in der besseren Auffindbarkeit beim Archivieren von Dokumenten, da die vergebenen Schlagworte ein Dokument möglichst gut beschreiben sollen. Die Suche nach bestimmten Dokumenten im Internet wird anhand prägnanter Schlagworte vereinfacht. Für akademische Journale werden die Autoren aufgefordert ca. fünf bis fünfzehn solcher Keywords zu vergeben. Keywords können nicht nur einzelne Wörter sein, sondern auch Phrasen, welche aus zwei oder mehr Wörtern bestehen. Diese nennt man auch „Keyphrases“ [vgl. Turn00, 1].

3.1 Manuelle Indexierung

Eine Verschlagwortung eines Dokumentes kann manuell durchgeführt werden. Dabei werden repräsentative Wörter zur Sacherschließung eines Dokumentes von einer Person, dem sogenannten „Indexierer“ zugewiesen. Da die Verwendung von beliebigen Wörtern zu Ungenauigkeiten führen kann, muss ein kontrolliertes Vokabular verwendet werden . Der Indexierer muss fundiertes fachliches Wissen über den Themenbereich der zu indexierenden Dokumente haben. Folglich wird auch Kenntnis über das Vokabular vorausgesetzt. Ein fachlich kompetenter Indexierer, der den Inhalt eines Dokumentes

versteht, kann dieses gut beschreiben und entsprechende prägnante Keyphrases vergeben. Dabei beeinflussen Erfahrung, Interesse und Motivation verschiedener Indexierender den Indexierungsvorgang. Beim Einsatz mehrerer Indexierer kann die Konsistenz der Indexierung leiden [vgl. Ferb03, 84]. Bei der stetig zunehmenden Informationsflut ist in den meisten Fällen davon auszugehen, dass diese Arbeit nicht mehr in manueller Art und Weise verrichtet werden kann und eine Automatisierung bei umfangreichen Dokumentensammlungen erforderlich wird.

3.2 Automatische Indexierung

Bei dieser Art der Indexierung sollen alle sinntragenden Wörter aus einem Dokument automatisiert extrahiert werden. Stoppwörter (siehe Kapitel 3.2.2), Wörter die sehr häufig vorkommen und in der Regel keine Relevanz zum Inhalt eines Dokumentes besitzen, werden dabei nicht beachtet. Andernfalls würde die Aussagekraft der extrahierten Keyphrases stark vermindert, sie wären zu allgemein, um ein Dokument prägnant zu beschreiben. Neben der Aussparung der Stoppwörter wird das Dokument noch mit einigen statistischen und linguistischen Methoden weiter verarbeitet. Die Häufigkeit für jedes verbleibende Wort wird ermittelt. Anhand der Häufigkeit kann bereits eine erste Klassifizierung der Keyphrases erfolgen. Durch das Zusammenfassen von Wörtern und den Vergleich mit anderen Dokumenten wird die Menge von Keyphrases weiter spezifiziert.

3.2.1 Webseiten parsen

Im Gegensatz zu automatischen Verschlagwortungen von Office-Dokumenten, im proprietären Microsoft-Format, müssen HTML-Dokumente vor der eigentlichen Verarbeitung lesbar gemacht werden, indem sie von HTML-spezifischen Tags befreit werden. Vor der Bereinigung können die HTML-Tags gar hilfreich sein, da sie relevante Inhalt beispielsweise explizit markieren. Zum Beispiel kann mit dem Tag `<dfn>` eine Definition im Text ausgezeichnet werden, die von Suchmaschinen für ein besseres Suchergebnis direkt extrahiert werden kann. Die meisten Webseiten verwenden allerdings solche speziellen Tags nicht, da sie entweder von Laien oder professionellen Agenturen, die eher layout-orientiert arbeiten, erstellt werden [vgl. Lewa05, 61].

Als bedeutenderer Teil eines HTML-Dokumentes gilt der `<title>`-Tag. Dieser wird von vielen Suchmaschinen besonders ausgewertet und die enthaltenen Schlagworte entsprechend gewichtet [vgl. Lewa05, 62]. Folglich können auch zum Extrahieren von Keyphrases wertvolle Informationen aus dem Titel eines HTML-Dokumentes gezogen werden.

Weitere interessante Informationen könnten sich in den so genannten Metatags finden. Sowohl die „meta-description“ als auch die „meta-keywords“ können eine inhaltliche Beschreibung des Dokuments in Kurztext- und Stichwortform enthalten. Diese Daten können bereits von Content-Managment-Systemen automatisch erstellt oder aber vom Programmierer manuell in das HTML-Dokument eingefügt werden. Aus diesem Grund stellt sich die Frage nach der Zuverlässigkeit der Metadaten. Da in der Vergangenheit häufig versucht wurde, die Metadaten zur Manipulation des Rankings in Suchmaschinen zu missbrauchen, werden diese von den meisten Suchmaschinen aktuell nicht mehr für das Ranking ausgewertet. Die Meta-Beschreibungen werden höchstens noch als Zusammenfassung der Seiten in den Ergebnislisten angezeigt, aber fließen nicht mehr in das Ranking ein [vgl. Lewa05, 181]. Um das Ergebnis der Keyphrase Extraction nicht zu verfälschen, empfiehlt es sich daher, die Metadaten zwar mit in die Menge der zu untersuchenden Wörter aufzunehmen, aber diese eher gering zu gewichten. Die Meta-Daten können auch leer sein oder einfach einen, von Editoren automatisch vergebenen, Standardwert (wie „no title“) enthalten.

Unwichtig für das Extrahieren von Keyphrases ist die Information des Erstellungsdatums eines Dokumentes, nützlich hingegen wäre die Meta-Information zur im Dokument verwendeten Sprache. Diese erlaubt eine gezielte Verarbeitung mit linguistischen Methoden. Da diese Angabe aber nicht obligatorisch ist, ist diese Informationsquelle nicht verlässlich und daher ebenfalls mit Vorsicht zu verwenden.

Weiterhin muss ein HTML-Parser in den Code eingebettete Scripte (z.B. PHP- oder JavaScript-Code) heraus filtern. Diese erfüllen rein funktionale Dienste für die Webseite, haben aber nichts mit dem eigentlichen Inhalt des Dokumentes zu tun. Am Ende sollte ein extrahierter Text zur Verfügung stehen, der frei von jeglichen Code-Fragmenten ist.

In der Regel werden auch schon sämtliche Satzzeichen entfernt. Die Wörter werden als einzelne Elemente isoliert und in Kleinbuchstaben konvertiert.

3.2.2 Stoppwörter

Wörter, die in einem Dokument sehr häufig auftreten, aber keine Relevanz zum Inhalt eines Dokumentes haben, nennt man Stoppwörter. Stoppwörter der deutschen Sprache sind zum Beispiel die bestimmten und unbestimmten Artikel (der, die, das, einer, eine), Konjunktionen (und, oder, doch) sowie Präpositionen (an, von, in) und die Negation (nicht). Im Englischen zählen zum Beispiel „a“, „the“, „of“, „you“ und „and“ als Stoppwörter. Diese Wörter haben zwar alle eine grammatikalische oder syntaktische Funktion, liefern aber keine Relevanz zum Inhalt eines Textes.
Aus diesem Grund müssen die Stoppwörter vor der weiteren Verarbeitung des Textes entfernt werden.

Die größte Menge eines Textes besteht aus genau diesen Stoppwörtern. Die kleinere, inhaltlich relevante, Menge besteht aus weniger häufigen Wörtern, die aber mit höherer Wahrscheinlichkeit themenrelevant sind. Diese Verteilung von Wörtern in einer Sprache wird durch das Zipfsche Gesetz beschrieben und ist in Abbildung 3 darsgestellt. Dieses besagt, dass das Produkt der Häufigkeit eines Wortes in einem Dokument und seinem Häufigkeitsrang für das gesamte Dokument in etwa konstant ist [vgl. Ferb03, 67].

$r(w)$ bezeichne den Rang und $h(w)$ die Häufigkeit eines Wortes aus einer Menge $W(C)$ der Wörter die in einem Dokument C vorkommen. Dann gilt

$$r(w)\cdot h(w)\sim c = konstant \;\forall\; w \in W(C)$$

Häufige Wörter sind nicht nur in fast jedem Teil eines Dokumentes zu erwarten, sondern auch in fast jedem anderen Dokument. Diese Wörter sind folglich nicht spezifisch genug, um als gute Suchterme fungieren zu können. Auch sehr seltene Wörter eines Dokumentes werden für eine Suche nicht hilfreich sein, da diese Wörter für den Inhalt des Textes nicht wichtig genug sein werden. Die Behandlung sehr seltener Terme in einem Dokument wird meistens ignoriert. Folglich bleiben die Wörter mit mittlerer Häufigkeit übrig. Diese sind häufig genug, um den Text inhaltlich abzudecken, aber auch prägnant genug, um den Inhalt eines Dokumentes zu beschreiben.

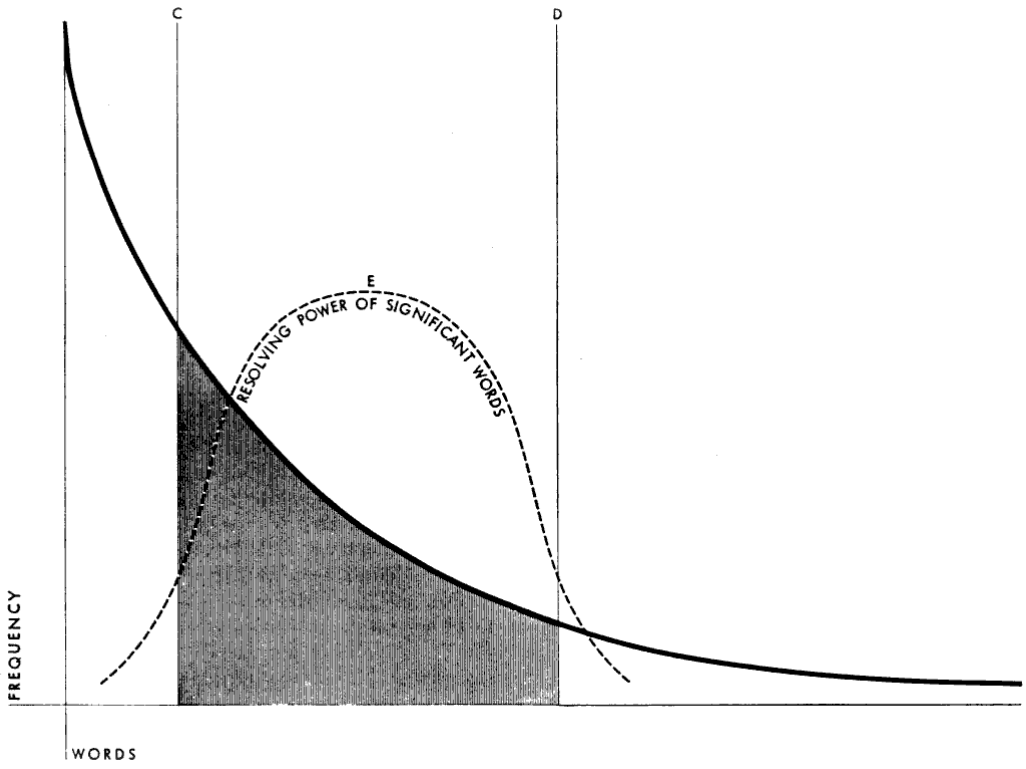

Abbildung 3: Auftreten von signifikanten Textwörtern in Dokumenten [Lewa05, 100]

Eine Möglichkeit Stoppwörter aus einem Text zu entfernen besteht darin, die Wörter entsprechend zu gewichten. Dadurch können diese sehr häufigen Wörter ausgeschlossen werden. Eine solche Methode lässt sich z.B. mit einer Form der Inverse Document Frequency (IDF) realisieren. Dabei werden die Terme sehr gering gewichtet, die in einer Sammlung vieler Dokumente insgesamt sehr häufig auftreten. Es spielt keine Rolle, ob ein Wort im Dokument selbst häufig oder nur einmal vorkommt.

$$w_{i,j} = idf(j) = \frac{1}{d(j)}$$

[vgl. Ferb03, 68]

$d(j)$ ist die Anzahl der Dokumente, in denen Term j vorkommt. Die Gewichtungswerte liegen im Intervall [0,1]. Ein Schwellwert definiert die Schranke für den Ausschluss der Wörter.

Eine weitere Variante für die Erkennung von Stoppwörtern ist, die Verteilung eines Terms über die Dokumente festzustellen. „Terme, die über die Dokumente gleichmäßig verteilt sind, sollten weniger spezifisch sein, als solche, die in einzelnen Dokumenten mit hoher Häufigkeit auftreten, in anderen dagegen gar nicht.“ [Ferb03, 69] Die erkannten Wörter können anschließend aus dem Text entfernt werden.

Eine weitere beliebte Möglichkeit: Alle bekannten Stoppwörter einer Sprache kann man mit Hilfe einer Stoppwortliste entfernen. Für jede Sprache gibt es bereits vorgefertigte Listen, die eine Sammlung der sehr häufigen Wörter darstellen. Diese Methode ist eine harte Schranke für den Ausschluss von Wörtern, da auch Wörter entfernt werden, die in einem Dokument eventuell zur Menge der Wörter mit mittlerer Häufigkeit gehören könnten. Andere Wörter, die nicht zur Liste der Stoppwörter gehören, aber sehr häufig vorkommen, verbleiben hingegen im Text. Die Entscheidung für eine bestimmte Methode der Stoppwortentfernung hängt von der Zielsetzung der Indexierung ab: Will man entweder den Text möglichst lesbar zusammenfassen oder sollen vielmehr maschinenverwertbare Keywords extrahiert werden.

3.2.3 Statistische Methoden

Mit statistischen Methoden können jedem Wort eines Dokumenttextes Häufigkeiten zugewiesen werden. Diese Häufigkeiten stellen in Relation zur Länge des Textes eine erste Gewichtung der Wörter dar, die einen Deskriptor über ein Dokument bilden. Voraussetzung ist, dass man davon ausgeht, dass die Häufigkeit des Auftretens eines Wortes in einem Dokument, zu seiner Relevanz in Relation steht.
Um die Gewichtung und damit den Deskriptor für ein Dokument zu verbessern, können häufige Wörter eines Dokumentes darauf überprüft werden, wie oft diese im gesamten Dokumentenarchiv vorkommen. Ein Wort, dass in einem Dokument sehr häufig vorkommt, in einem ganzen Archiv hingegen sehr selten auftritt, hat ein gewisses Alleinstellungsmerkmal und ist signifikant genug, um ein Dokument eindeutig zu beschreiben. Dies nennt man die inverse Dokumenthäufigkeit, die in Kapitel 3.2.5 detaillierter beschrieben wird.

3.2.4 Linguistische Methoden

Statistische Methoden verbessern zuerst nur die Trefferquote, also die Wahrscheinlichkeit auf Anfrage eines Terms, ein relevantes, indexiertes Dokument zu finden. Dies nennt man den Recall. Um die Genauigkeit, die Precision, zu erhöhen, also die Wahrscheinlichkeit, dass ein gefundenes Dokument auch wirklich relevant ist, helfen linguistische Methoden

weiter. Um die Deskriptoren über ein Dokument zu präzisieren, muss der gesamte Text weiter verarbeitet werden. Einzelne Operationen werden auf den Wörtern ausgeführt, welche als Zeichenfolge definiert sind, die zwischen zwei Leer- oder Satzzeichen stehen.

3.2.4.1 Stammformenreduktion von Wörtern

Jedes Wort kann nach der Bereinigung einer Lemmatisierung unterzogen werden. Unter Lemmatisierung versteht man die Reduzierung eines Wortes auf seine grammatische Grund- oder Stammform. Während die Grundform eines Wortes ebenfalls wieder ein in der Sprache vorkommendes Wort sein kann (z.B. gefunden → finden), bildet die Stammformenreduktion (z.B. semantisch → semant) eine Form, die im Allgemeinen nicht als Wort in einer Sprache vorkommt [vgl. Ferb03, 41]. Durch dieses Verfahren können verschiedene Flexionsformen eines Wortes zusammengefasst werden. Somit werden alle Formen eines Wortes gemeinsam gewertet [vgl. Lewa05, 106]. Die Methode der Grund- und Stammformreduktion kann im Kontext des Information Retrieval unter dem Begriff Stemming zusammengefasst werden.
Ein weit verbreiteter Algorithmus für das Stemming ist der Porter-Stemmer-Algorithmus[4]. Die Funktionsweise des Verfahrens soll in diesem Rahmen nur kurz erläutert werden. Der Algorithmus nutzt eine Menge von Verkürzungsregeln, die sequentiell auf ein Wort angewendet werden, bis dieses in die Minimalanzahl seiner Silben zerlegt ist. Die für jede Phase anzuwendenden Regeln werden aus den Regeln für das längste Suffix eines Wortes ausgewählt [vgl. MaPrSc08, 31]. Ein kurzes Beispiel aus [MaPrSc08, 31] soll die Funktionsweise veranschaulichen:

Regel	***Beispiel***
SSES → SS	***caresses → caress***
IES → I	***ponies → poni***
SS → SS	***caress → caress***
S →	***cats → cat***

Der Algorithmus wurde ursprünglich für die englische Sprache entwickelt, ist aber mittlerweile auch in andere Sprachen transformiert worden. Aufgrund der zahlreichen

[4] Porter 1980 siehe http://www.tartarus.org/~martin/PorterStemmer / 10.09.2009

Ausnahmen und vor allem der Mehrdeutigkeit vieler Wörter in der deutschen Sprache, ist das Stemmen durch allgemeine Regeln schwer möglich und fehleranfällig [vgl. Stoc00, 169f]. Der Porter-Stemmer-Algorithmus arbeitet, wie alle Stemming-Algorithmen, nicht immer mit hundertprozentiger Genauigkeit. Bei einigen Worten kann es zu over- oder understemming kommen, d.h. es wird zu viel oder zu wenig abgeschnitten[5]. Auch die Verwendung von Synonymen, Akronymen und anderen Abkürzungen auf Webseiten kann zu Fehlern führen. Trotzdem werden die meisten Stemming-Verfahren für das Information Retrieval als ausreichend gut beschrieben.
Ein weiteres Resultat des Stemmings ist, dass die Liste der Terme eines Dokumentes durch das Zusammenführen der verschiedenen Wortformen zu einem Term, kleiner wird. Somit verringert sich der Verwaltungsaufwand [vgl. Ferb03, 41]. Daher ist der Einsatz eines Stemmingverfahrens aus Sicht dieser Arbeit zu empfehlen, auch wenn [Lewa05, 109] empfiehlt, es muss „dem Nutzer auf jeden Fall die Möglichkeit gegeben werden, diese Funktion selbst ein- bzw. auszuschalten“.

3.2.4.2 Phrasen

Der Erkennung von einzelnen Phrasen, also Ausdrücken, die aus zwei oder mehr Wörtern bestehen, kommt eine spezielle Bedeutung zu. In den meisten Fällen werden bereits die einzelnen Worte erkannt und entsprechend gewichtet. Den Deskriptor über ein Dokument mit einer Phrase zu erweitern, kann aber den Vorteil haben, Personen-, Firmen- und Produktnamen zu identifizieren [vgl. Lewa05, 109]. Feststehende Begriffe aus mehreren Wörtern können einen gänzlich anderen Sinn haben, als die Einzelterme. So zum Beispiel „Universität Koblenz“.
Phrasen können durch zwei Methoden erkannt werden, nämlich

- durch den Abgleich mit Listen und
- durch die Analyse gemeinsamen Auftretens in den Dokumenten [vgl. Stoc00, 153].

Bei der ersten Methode werden Hauptwörter eines Dokumentes mit einer Sammlung von Listen abgeglichen, die allgemein häufig vorkommende Personen-, Firmen- und Produktnamen enthalten. Ein Nachteil dieser Methode liegt darin, dass diese Listen

[5] vgl. http://de.wikipedia.org/wiki/Porter-Stemmer-Algorithmus / 12. Sep. 2009

immer aktuell gehalten werden und beim Fund neuer Phrasen ergänzt werden müssen. Kommt es beim Abgleich des zu untersuchenden Textes mit den Listen zu einem Match, lässt sich auf eine Phrase schließen, die ab einer bestimmten Vorkommenshäufigkeit in die Schlagwortliste, also den Deskriptor, aufgenommen wird. Ein Schwellwert für die Vorkommenshäufigkeit muss explizit definiert werden und kann je nach Themengebiet der Dokumente variieren.

Die zweite Methode kann auf unterschiedliche Arten realisiert werden. Beispielsweise kann mit einem Positionsindex für jede Position eines Wortes untersucht werden, ob ein anderes Wort auf dieses folgt. Bei mehrmaligem Vorkommen einer Wortsequenz, wird diese als Phrase erkannt, wenn der auch hier zu definierende Schwellwert überschritten wurde [vgl. Stoc00, 154f].
Die Phrasen können als Erweiterung zu einem Deskriptor hinzugefügt werden [vgl. Ferb03, 218].

Durch beide Methoden können Phrasen aus allen Sprachen erkannt werden. In der englischen Sprache unterscheidet sich die Phrasenbildung häufig darin, dass zwei Worte hintereinander genannt werden (z.B. information retrieval). Im Deutschen hingegen werden meist zusammengesetzte Begriffe gebildet (z.B. Informationswiedergewinnung) [vgl. Lewa05, 111]. Diese Begriffe werden allerdings auch als Einzelterme erkannt. Somit ist zu erwarten, dass eine Phrasenerkennung für die Keyphrase Extraction aus einem Dokument nicht zwangsläufig ein besseres Ergebnis liefert. [Ferb03, 219] schreibt zum Einsatz des Retrieval-Systems INQUERY mit Verwendung eines Phrasenthesaurus bei TREC-4 (Text REtrieval Conference 4), dass diese Verwendung nur eine geringe Verbesserung (3,5%) brachte. Dieses Fazit ist allerdings insofern kritisch zu sehen, da es in diesem Fall um das Auffinden von relevanten Dokumenten ging, nicht aber um deren inhaltliche Beschreibung, für die es von größerer Bedeutung sein kann, Namen und Begriffe in Form von Phrasen zu erkennen. Viele Systeme (z.B. YouTube, Google News[6]) bieten mittlerweile die Möglichkeit, bei der Eingabe von Suchanfragen automatisierte Vorschläge für eine Suche anzuzeigen. Dabei werden einzelne eingegebene Wörter automatisch um bekannte Phrasen ergänzt.

[6] http://news.google.de

3.2.5 TF-IDF

TF-IDF steht für „Term Frequency - Inverse Document Frequency“ und ist eine Gewichtungsmethode beim Information Retrieval aus Dokumenten. Dabei steht die bereits erwähnte Termhäufigkeit (term frequency) für die Relevanz eines Terms innerhalb eines Dokumentes. Häufig auftretende Terme, außer den Stoppwörtern, in einem Dokument sind wichtiger für den Inhalt des Textes als solche, die seltener vorkommen. Neben dieser lokalen Gewichtung im aktuellen Dokument, gibt es noch die globale Gewichtung, bei der statt der Häufigkeit eines Terms in einem Dokument, die Anzahl der Dokumente, in denen ein Term vorkommt, eine Rolle spielt.

Bei dieser Dokumenthäufigkeit (document frequency) ist interessant, dass die Bedeutung für global besonders häufig auftretende Terme nach unten korrigiert wird.

Die sogenannte inverse Dokumenthäufigkeit (inverse document frequency) wird mit folgender Formel ermittelt:

$$w_{i,j} = idf(j) = \frac{1}{d(j)}$$

[vgl. Ferb03, 68]

Die Gewichtung w des Terms t_j aus dem Dokument d_i ist gleich der inversen Dokumenthäufigkeit des Terms t_j . $d(j)$ ist dabei die Anzahl der Dokumente, in denen Term t_j auftritt. Hier kann man erkennen, dass es einen Unterschied macht, ob ein Term in insgesamt 100 oder 10.000 anderen Dokumenten vorkommt. Eine in der Praxis häufig verwendete modifizierte Form nutzt den natürlichen Logarithmus.

$$w_{i,j} = \ln \frac{m}{d(j)}$$

[vgl. Ferb03, 68]

m ist die konstante Anzahl aller Dokumente. Der Wert der Formel fällt mit wachsendem $d(j)$ monoton. Der Logarithmus dämpft dabei große Werte ab, was die Gewichtung von sehr seltenen Termen (kleine $d(j)$) wieder etwas abschwächt [vgl. Ferb03, 69].

Bei der Implementierung der IDF muss noch eine 1 zu $d(j)$ addiert werden, um für den Fall, dass ein Term in keinem Dokument vorkommen sollte, eine Division durch 0 zu vermeiden.

Die lokale Gewichtung eines Terms hängt hingegen einfach von seiner Häufigkeit in einem Dokument ab. Dies kann man durch die Formel ausdrücken:

$$w_{i,j} = h(i,j)$$

[vgl. Ferb03, 70]

Die Gewichtung eines Terms t_j im Dokument d_i ist gleich der Häufigkeit $h(i,j)$ des Terms im Dokument [vgl. Ferb03, 70]. Um den Einfluss sehr häufiger Terme zu dämpfen, kann man die Gewichte auf ein Intervall beschränken.

$$w_{i,j} = \frac{h(i,j)}{1 + h(i,j)}$$

[vgl. Ferb03, 70]

Weiterführend von dieser Formel kann man die Häufigkeit eines Terms zu der des häufigsten Terms im Dokument in Relation setzen. Dadurch wird der Effekt ausgeglichen, dass die Häufigkeiten von Termen in längeren Texten im Allgemeinen größer ist, als in kürzeren.

$$w_{i,j} = tf_{i,j} = \frac{h(i,j)}{max_{l \in \{1,\ldots n\}} h(i,l)}$$

[vgl. Ferb03, 70]

Die lokalen und globalen Gewichtungen können nun miteinander verknüpft werden. Die Termhäufigkeit wird dabei mit der inversen Dokumenthäufigkeit multipliziert.

$$w_{i,j} = tf_{i,j} \cdot idf_j = \frac{h(i,j)}{max_{l \in \{1,\ldots n\}} h(i,l)} \cdot \ln \frac{m}{d(j)}$$

Diese TF-IDF-Gewichtung ist schon in vielen Systemen und Untersuchungen eingesetzt worden [vgl. Ferb03, 70]. Somit bietet die TF-IDF einen guten Rankingfaktor für die Keyphrase Extraction aus einem Dokument, da global seltenere Keyphrases, die einen Text signifikanter bescheiben, bevorzugt werden.

3.2.6 Das Cosinus-Ähnlichkeitsmaß

Mittels der Berechnung von TF-IDF erhält man Gewichtungswerte für die Worte eines Dokumentes. Mit diesen Werten kann die Ähnlichkeit von zwei Dokumenten berechnet werden. Als eine Berechnungsoption gibt es den Cosinus.

$$S_{D_j, D_j} = \frac{\sum_{k=1}^{L} (weight_{ij} weight_{jk})}{\sqrt{\sum_{k=1}^{L} (weight_{ik}^2) \sum_{k=1}^{L} (weight_{jk}^2)}}$$

[vgl. StSt08, 373]

Im Zähler werden die Gewichtungswerte aller Worte multipliziert, die sowohl im Dokument i , als auch im Dokument j vorkommen. Existiert ein solches Wortpaar nicht, da ein Wort nur in einem der beiden Dokumente, aber nicht im anderen auftritt, so ist dieses Produkt 0. Die Produkte werden addiert. Die Terme mit Gewichtungswerten werden auch als Vektoren dargestellt. Die Summe der Produkte der Wortpaarwerte ist dementsprechend gleich dem Skalarprodukt der beiden Dokumentvektoren. Der Nenner der Gleichung ist die euklidische Länge beider Vektoren und normiert die Länge der Dokumentvektoren.

So ist in [MaPrSc08, 111] auch die Gleichung

$$similarity(d_1, d_2) = \frac{\vec{V}(d_1) \cdot \vec{V}(d_2)}{|\vec{V}(d_1)| \, |\vec{V}(d_2)|}$$

zu finden. Die Ähnlichkeit zweier Dokumente ist dann gleich des Produktes der normierten Vektoren beider Dokumente. Der normierte Vektor $\vec{V}(d)$ ist der Einheitsvektor $\vec{v}(d) = \vec{V}(d) / |\vec{V}(d)|$. Man kann die Gleichung entsprechend abkürzen:

$$similarity(d_1, d_2) = \vec{v}(d_1) \cdot \vec{v}(d_2)$$

Die Ähnlichkeitswerte liegen im Wertebereich von 0 bis 1. Bei maximaler Ähnlichkeit ist der Ähnlichkeitswert 1, bei maximalem Abstand 0 [vgl. StSt08, 373]. Da die Werte unabhängig von der Länge der Vektoren sind, entscheidet die Richtung der Vektoren über

die Ähnlichkeit. Die Ähnlichkeit von zwei Dokumenten ist dann am größten, wenn ihre beiden Vektoren die gleiche Richtung haben [vgl. Ferb03, 75].

Mit dem Cosinus-Ähnlichkeitsmaß können verschiedene Dokumente im Vergleich mit einem Referenzdokument nach ihrer Ähnlichkeit sortiert und entsprechend ein Ranking erstellt werden.

4 Tagging-Systeme

Der Inhalt eines Dokumentes wird oft mit Wordclouds dargestellt. Dabei sind die häufigsten Begriffe wolkenartig visualisiert. Die Schriftgröße ist dabei proportional zur Häufigkeit des Wortes. Für viele Content-Managment-Systeme wie Joomla oder Webblogs, wie Wordpress gibt es Plugins, die eine Wordcloud für jeden veröffentlichten Artikel oder für das gesamte Archiv darstellen können.

Abbildung 4: Beispiel Tagcloud.
Quelle: http://medienabc.files.wordpress.com/2007/01/tagcloud.jpg

Diese Wortwolken sind dem versierten Web 2.0 – Nutzer von den Tag-basierten Systemen bekannt. So genannte Tag-Clouds visualisieren die vergebenen Schlagworte zu einem Medium.Social Tagging

Die Suche bei der deutschen Wikipedia[7] gibt für das Wort Tagging den Begriff „Gemeinschaftliches Indexieren“[8] aus. Ein Tag (Schlagwort, Etikett, Stichwort) ist ein beschreibender Begriff für ein Objekt. Im Falle der Plattform YouTube ist dieses Objekt ein Video, dass von Usern beschrieben wird. In der Plattform Flickr[9] wiederum werden Fotos getagged. So kann ein Foto vom Pariser Eiffelturm zum Beispiel mit den folgenden Tags versehen werden: paris, eiffelturm, architektur, stahl, gustav, wahrzeichen [vgl. Alby08, 127]. Die freie Vergabe von Tags für Medien erfolgt ohne Regeln. Jeder User kann jedes Medium mit einem Begriff seiner Wahl taggen. Tagging ist ein markantes Merkmal des Web 2.0. Es gibt mittlerweile sehr viele Benutzer, die die nahezu unüberschaubare Menge von Objekten taggen. Dadurch entsteht eine Sammlung von Tags, welche als Folksonomy bezeichnet wird [vgl. Alby08, 127]. Würden bestimmte Kategorien oder Hierarchien vorgegeben werden, würde man von einer Taxonomie sprechen. Da aber keine präzisen Vorgaben gemacht werden, sondern die Vergabe von Tags nach „ungeschriebenen Gesetzen“ der Gemeinschaft erfolgt, entstand der Begriff Folksonomy (folk taxonomy) und bezeichnet also das „Gemeinschaftliche Indexieren“.
Ein Tag kann dabei auch aus mehreren Wörtern bestehen, die dann miteinander verbunden werden müssen, wie zum Beispiel „semantic.web“.

4.1 MyTag

Die Web 2.0 – Plattformen wie Flickr, YouTube oder del.icio.us[10] stellen mittlerweile enorm viele Ressourcen zur Verfügung. Jede Plattform für sich bietet die Suche über die eingestellten Medien an. Jedoch findet ein Anwender bei der Suche jeweils nur die plattform-spezifischen Medien. Daraus entstand die Idee einer cross-media-search, die mit einer einzigen Suchanfrage auf unterschiedlichen Tagging-Plattformen verschiedene Medientypen findet.

Realisiert wurde dies mit dem Projekt MyTag[11], entstanden im Rahmen eines Projektpraktikums der Arbeitsgruppe Staab der Universität Koblenz-Landau im Sommersemester 2007. Mittlerweile wurde MyTag in einem weiteren Projektpraktikum

[7] http://de.wikipedia.org
[8] http://de.wikipedia.org/wiki/Tagging
[9] http://www.flickr.com
[10] http://www.delicious.com
[11] http://mytag.uni-koblenz.de

2008 und im Rahmen verschiedener Studienarbeiten weiterentwickelt. Nach derzeitigem Stand bietet MyTag eine Suche nach Fotos, Videos, Bookmarks und Literaturreferenzen über die Plattformen Flickr, YouTube, del.icio.us, Connotea[12], Bibsonomy[13] und BibTex[14]. Als weitere Funktionalität ist es möglich, eigene Profile der genannten Plattformen einzubinden und eine personalisierte Suche zu nutzen. Dabei kann auch das Ranking der Suchergebnisse nach den persönlichen Interessen sortiert werden. Hierfür wird ein Login benötigt.

Eine weitere Funktion ist im Wintersemester 2008/2009 implementiert worden. Bei mehrdeutigen Begriffen wird ein Vorschlag zur Verbesserung der Suchergebnisse geliefert. Dieser Disambiguierungs-Service bietet zum Beispiel für die Suchanfrage apple an:
Meinten Sie:

- Apple Inc
- Macintosh
- Apple
- Apple Corps
- Apple Corps v Apple Computer

Die Mehrdeutigkeit für den Begriff „apple“ liegt hier bei der Unterscheidung von Computern (Macintosh) und Obst (Apple). Mehr zu dieser Funktionalität ist bei [Dell09] nachzulesen. Diese Behandlung von Mehrdeutigkeiten kann auch später hilfreich sein, wenn das Ranking anhand der Eingabe-Webseite in Einzelfällen nicht ausreichend gut sein sollte.

Die Keyphrases, die im Rahmen dieser Arbeit aus einer Webseite extrahiert und einem Ranking unterzogen werden, werden als Suchbegriffe an MyTag gesendet. MyTag liefert anschließend eine Ergebnisliste der gefundenen Ressourcen aus den verschiedenen Plattformen.

[12] http://www.connotea.org
[13] http://www.bibsonomy.org
[14] http://www.bibtex.org

5 Implementierung

5.1 Technische Voraussetzung

Um eine Webseite verarbeiten zu können, muss diese zuerst eingelesen werden. Dazu gibt es bereits HTML-Parser aus OpenSource-Projekten. Die Entscheidung ist auf den Jericho-HTML-Parser[15] gefallen, weil dieser alle benötigten Funktionen bereitstellt. Da dieser Parser eine Java-Bibliothek ist, liegt es nahe, die gesamte Funktionalität in Java zu implementieren. Die fertige Anwendung wird schließlich als Java-Webservice auf einem eigenen Server bereitgestellt.

Von MyTag aus, das auf dem Framework „Ruby on Rails“[16] basiert, kann der Webservice mittels SOAP konsumiert werden. Der Austausch der Daten mit XML-Dokumenten ermöglicht den hier notwendigen Austausch von Daten zwischen Anwendungen, die mit unterschiedlichen Programmiersprachen entwickelt wurden.
Die URL wird als String an den Webservice gesendet. Die von der URL lokalisierte Webseite wird vom Webservice eingelesen und analysiert. Anschließend wird ein Array mit extrahierten Keyphrases an MyTag zurück gegeben, aus denen passende Suchbegriffe identifiziert werden. Für das Ranking sind die restlichen Keyphrases hilfreich.

Die URL der Eingabewebseite wird mit Hilfe eines Bookmark-Buttons[17] an MyTag übertragen. Diese Funktionalität wird mit Hilfe eines einfachen JavaScript-Codes realisiert. Der Button lässt sich per „Drag & Drop“ in der Menüleiste der Browser Firefox und Opera platzieren. Für den Internet Explorer ist es möglich, die Funktion zu den Favoriten hinzuzufügen. Somit ist die Anwendung praktisch für jeden Internetnutzer verwendbar.

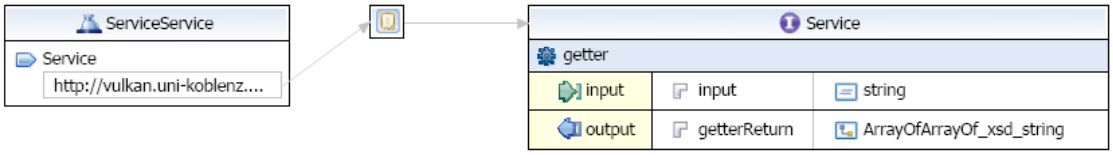

Abbildung 5: Visualisierung der Webservice-Schnittstelle

[15] http://jericho.htmlparser.net
[16] http://rubyonrails.org/
[17] Add-on Button mit Kurzanleitung: http://vulkan.uni-koblenz.de

5.2 Technische Schwierigkeiten

Eine Schwierigkeit, die schon zu Anfang der Arbeit ersichtlich war, ist die Behandlung von framebasierten Webseiten. Wird dem Parser nur die URL von der Seite übergeben, die die Frames defniert, ist es unter Umständen schwierig, die Inhalte der Frames auszulesen.

Ein weiteres Problem können Webseiten darstellen, für die ein Login benötigt wird. Da das Einlesen der Webseite anhand der übergebenen URL durch den Server statt findet, ist es schwierig den Login, den ein Anwender ggf. bereits durchgeführt hat, an den Parser zu übergeben.

Eine spätere Erweiterung könnte somit darin bestehen, das Einlesen der Daten clientseitig zu realisieren. Damit wäre zumindest das Problem der durch einen Login unzugänglichen Seiten gelöst. Hierzu müsste das Add-On für den Browser erweitert und mit der nötigen Funktionalität ausgestattet werden.

Da auch die großen Suchmaschinen keine elegante Lösung für die Verarbeitung von Frames bieten und die Verwendung von Frames in den letzten Jahren kontinuierlich abgenommen hat [vgl. Lewa05, 67], wurde dieses Problem auch hier vorerst ignoriert.

5.3 HTML-Dokumente parsen

Das Extrahieren des Textes aus einer Webseite übernimmt der Jericho-HTML-Parser, eine leistungsstarke Java-Bibliothek. Zur Analyse eines HTML-Dokumentes stehen alle für dieses Projekt nötigen Werkzeuge zur Verfügung. Der Parser benötigt als Eingabeparameter die URL des HTML-Dokuments. Er ist dann in der Lage den Title-Tag, sowie die Meta-Tags description und keywords einzeln zu extrahieren. Sofern der Title-Tag Informationen enthält, wird dieser auch an den Textanfang gestellt. Die anderen Metadaten sollen zwar in die Verarbeitung einfließen, aber eine eher geringe Gewichtung erhalten. Wie schon in Kapitel 3.2.1 erwähnt, können die Metadaten, sofern vorhanden, manipulierte oder fehlerhafte Information erhalten. Für den Fall, dass die enthaltenen Terme adäquate Informationen darstellen, fließen sie erst zum Schluss mit verhältnismäßig geringer Gewichtung in das Endergebnis ein, können aber so eine Tendenz zugunsten signifikanter Terme bewirken. Die Gefahr, dass falsche Informationen zu stark gewichtet werden, ist damit umgangen.

Aus dem Body des HTML-Dokumentes werden beim Extrahieren des Textes eventuell vorhandene PHP-Tags oder JavaScript-Elemente entfernt. Diese erfüllen zwar eine Funktion in der Webseite, tragen allerdings nicht zum inhaltlichen Sinn bei. Der Parser entfernt auch bereits alle vorhandenen Satzzeichen, bis auf die Bindestriche bei Mehr-Wort-Ausdrücken. Weiterhin wurde entschieden, auch Ziffern zu entfernen, diese können zwar in Einzelfällen zum inhaltlichen Sinn einer Webseite gehören (z.B. worldcup 2006), dennoch treten Zahlen in den meisten Fällen bei der weiteren Verarbeitung eher störend auf.

Der Parser gibt den Dokumenttext anschließend als einen String aus, der daraufhin Wort für Wort separiert und einzeln als Element in einer verketteten Liste gespeichert wird. Die inhaltliche Reihenfolge bleibt dabei erhalten.

Als erstes Zwischenergebnis liegt also der Textinhalt der Eingabe-Webseite Wort für Wort, ohne Satzzeichen und ohne HTML-Elemente, in einer verketteten Liste vor.

5.4 Stoppwörter entfernen

Zum Entfernen der Stoppwörter aus der Textliste, wird eine englische und deutsche Stoppwortliste verwendet. Diese haben den Vorteil, dass Stoppwörter im Allgemeinen nicht der Charakteristik von Tags entsprechen. Tags, die zumeist Substantive, Verben oder Adjektive sind, gehören seltener zu Stoppwörtern und bleiben somit vollständig im Text erhalten. Bei der Entfernung auf Basis einer statistischen Methode würden unter Umständen wichtige substantivische Terme entfernt. Unnötige Stoppwörter hingegen können in der Textliste verbleiben. Laut [Lewa05, 99] arbeiten die meisten Systeme mit Stoppwortlisten. Für die Erkennung von Phrasen sollte allerdings eine Version der Text-Liste inklusive aller Stoppwörter verwendet werden. Die Phrase „herr **der** ringe“ hat beispielsweise eine präzisere Bedeutung als „herr ringe“.

5.5 Stemming

In dieser Anwendung wird für die Aufgabe des Stemmings der „Snowball-Parser“[18] eingesetzt, welcher den Porter-Stemmer-Algorithmus benutzt. Es sind darin bereits Stemmer für die englische und deutsche Sprache integriert.

Folgendes Beispiel zeigt Wörter, die in einem Dokument in verschiedenen Formen auftreten können, aber die gleiche Bedeutung haben. In einem Text über das „Semantische Web“ können so beispielsweise die Wörter

semantisches, semantischen, semantic, semantik

vorkommen, die alle die gleiche Bedeutung für die Sacherschließung des Textes haben. Es werden also in jeder Sprache Wörter der Grammatik entsprechend verändert. So wird im Englischen beispielsweise für den Plural oft ein 's' angehangen. Im Deutschen haben Wörter in den grammatikalischen Fällen unterschiedliche Endungen. Um die verschiedenen Flexionsformen der vorkommenden Terme zusammenzufassen, ist es notwendig, ein Stemming für jedes Wort der Liste durchzuführen. Dafür muss für jede Sprache ein spezifischer Stemmer verwendet werden.
Bei obigem Beispiel mit den Wörtern

semantisches, semantischen, semantic, semantics, semantik

kann als Ergebnis der gemeinsame Wortstamm „semant“ extrahiert werden.

Um den extrahierten Wortstamm mit der für die Gewichtung notwendigen Häufigkeit belegen zu können, muss das zusammengefasste Wort auch eine Häufigkeit erhalten. Dazu werden einfach die Häufigkeiten der verschiedenen Wortformen addiert. Da aber ein einfacher Wortstamm als Suchterm wahrscheinlich wenig Erfolg hat, weil die durchsuchenden Medien nicht mit Wortstämmen getagged sind, wird aus der Menge der ähnlichen Wörter jenes ausgewählt, das im Ausgangstext am häufigsten vorkommt. Dieses Wort erhält anschließend die Summe aller Häufigkeiten.

[18] http://snowball.tartarus.org/

Durch dieses Zusammenfassen der verschiedenen Wortformen zu einem Term, wird die Menge der weiter zu verarbeitenden Terme kleiner, wodurch sich der Aufwand für die folgenden Methoden verringert. Ein weiterer Effekt ist, dass die Bedeutung eines Terms wächst, da er eine höhere Häufigkeit hat und somit dessen Wichtigkeit für den Inhalt erhöht wird.

5.6 Phrasenerkennung

Es gibt feststehende Begriffe, die aus mehr als einem Wort bestehen, wie zum Beispiel „semantic web“, „deutsche bank“ oder „java virtual machine“. Die Begriffe haben eine ganz andere Bedeutung als die einzelnen Wörter (deutsche, bank). Eine Phrase hat die Charakteristik, dass Wörter immer aufeinander folgen. Die Suche nach einer Phrase, bestehend aus zwei Wörtern, erfolgt durch eine Adjazenz-Suche. Hierbei werden häufig auftretende benachbarte Wortpaare als Phrase erkannt. Da die Suche auf der bereits gestemmten Wortliste arbeitet, werden auch Flexionsformen einer Phrase zusammengefasst erkannt, wie zum Beispiel „deutsche**n** bank“.

Für alle gefundenen Phrasen wird die Häufigkeit gezählt. Wenn mehr als eine Phrase gefunden wurde, muss überprüft werden, bei welchen Wortpaaren es sich wirklich um eine Phrase handelt oder nur um ein Wortpaar, das zufällig mehr als einmal aufgetreten ist. Gibt es nur eine beste Phrase, hat diese den maximalen Häufigkeitswert. Als Schwellwert für die Entscheidung, ob es sich um eine Phrase oder nur ein häufigeres Wortpaar handelt, kann die Normalverteilung zu Grunde gelegt werden.

Die besten Phrasen werden anschließend noch darauf untersucht, ob eine Phrase mit drei Wörtern existiert. Zwei Phrasen sind als Liste dargestellt: $ph_1=[w_1, w_2]$ und $ph_2=[w_3, w_4]$ enthalten dann eine 3er-Phrase, wenn $w_2=w_3 \, und \, h(ph_1) \sim h(ph_2)$. Wobei $h(ph_1) \, und \, h(ph_2)$ die Häufigkeiten der erkannten Phasen sind.

Die 3er-Phrase erhält die gleiche Häufigkeit wie die der beiden 2er-Phrasen.

Phrasen aus zwei oder drei zusammengesetzten Wörtern werden ebenfalls als Tags vergeben. Tags können aber nicht mit einem Leerzeichen zwischen zwei Wörtern

vergeben werden. User müssen sich daher etwas einfallen lassen, um auch Phrasen als Tags verwenden zu können. Es gibt folgende verschiedene Möglichkeiten, zwei Tags miteinander zu verbinden, die in den diversen Tagging-Plattformen zu finden sind:

Konkatenation:	semanticweb
Punkt:	sematic.web
Bindestrich:	semantic-web
Unterstrich:	semantic_web
Pluszeichen:	semantic+web

Nach einer Untersuchung von Klaas Dellschaft [Dell07] lassen sich viele User in der Auswahl der Verknüpfung von Wörtern für „Phrasen-Tags“, davon inspirieren, was schon andere vor ihnen verwendet haben. Die meistens Phrasen-Tags sind mittels Konkatenation als Verbindung zweier Wörter in einer Tagging-Plattform zu finden. Die Verbindung mit Binde- und Unterstrich findet sich jeweils gleich häufig, aber wesentlich seltener als die Konkatenation. Noch seltener ist die Verbindung mit Punkt und Plus.

Auf diese Kombinationen hin werden die Phrasen untersucht. Diejenige Kombination, welche am häufigsten in den Datensätzen aufzufinden ist, wird anschließend als beste Phrase in der besten Variation als erstes Zwischenergebnis zum Anfragevektor hinzugefügt.

5.7 TF-IDF als Gewichtungsmethode

Nachdem alle Stoppwörter aus dem Text entfernt worden sind, werden die Häufigkeiten aller Terme im Dokument gezählt. Diese werden mit dem Term als Key und der Häufigkeit als Value in einer verketteten Hashmap abgelegt. Die verkettete Map garantiert dabei, dass die Reihenfolge des ersten Auftretens im Dokument erhalten bleibt. Das Maximum aller Values ist der häufigste Term im Dokument.

Für die Berechnung der inversen Dokumenthäufigkeit stehen Datensätze mit Ressourcen verschiedener Tagging-Plattformen zur Verfügung. Diese Daten wurden im Rahmen des

Tagora-Projekts[19] gesammelt und teilweise öffentlich zur Verfügung gestellt. Für diese Berechnung werden die Daten für die Plattformen Flickr und Delicious verwendet. Diese Daten enthalten die Daten für Ressourcen mitsamt der vergebenen Tags. Man kann also für jeden Term im Dokument abfragen, wie viele Medien mit einem Term getagged wurden. Umgekehrt erfährt man also, wie oft ein Tag (Term) für ein Dokument (Ressource) vergeben wurde. Somit hat man für die Berechnung der TF-IDF

$$w_{i,j} = tf_{i,j} \cdot idf_j = \frac{h(i,j)}{max_{l \in \{1,\dots n\}} h(i,l)} \cdot \ln \frac{m}{d(j)}$$

zwei aktuelle Werte für die Tagzuweisungen: Für del.icio.us ist $m = 140.333.000$ und für Flickr $m = 115.456.000$.

Aber nicht nur lokale und globale Häufigkeiten eines Terms können zur Gewichtung beitragen. Die Position eines Terms in einem Dokument kann ebenfalls von Bedeutung sein. Nachrichtentexte weisen in der Regel eine inhaltliche Struktur auf, in der wichtige Schlagworte bereits am Anfang erläutert werden. Im weiteren Verlauf werden diese Terme wiederholt. Am Ende des Textes werden meistens nur noch andere Details oder Kommentare mitgeteilt. Der Titel sagt häufig schon viel über den Inhalt der Seite aus.

Einige Suchmaschinen, auch Google, gewichten die Terme nach der Position im Dokument. Terme am Textanfang werden dabei stärker gewichtet, als am Textende.

Da die Terme nach der Reihenfolge des ersten Vorkommens vom Textanfang bis zum Textende in eine verkettete HashMap gespeichert wurden, kann beim Berechnen der TF-IDF eine Gewichtung hinzugefügt werden, welche Terme am Anfang des Dokumentes stärker gewichtet. Man geht nämlich davon aus, dass in einem Dokument der für den Inhalt des Dokumentes wichtige Teil, bereits in den ersten Sätzen beschrieben steht. Darum wird für eine Anzahl an Termen K der zusätzliche Gewichtungsfaktor für jeden Term mit $w_t = 1 + \ln(1{,}5 - \frac{p_t}{K})$ berechnet. p_t ist hier die Position des ersten Vorkommens des Terms. Der natürliche Logarithmus bewirkt in diesem Fall, dass die

[19] http://www.tagora-project.eu/data/

Gewichtungswerte am Anfang langsam und zum Dokumentende hin schneller fallen. Es gilt $w_t > 0$ für beliebige K .

Diese zusätzliche Gewichtung führt auch zu besseren Ergebnissen bei der Suche in Portalen oder portalähnlichen Seiten. Internetseiten, die sehr viele verschiedene Themen beinhalten, weisen meist eine sehr kleine Standardabweichung der Häufigkeitswerte der Terme auf. Dadurch würden entweder zu viele Terme geliefert werden, die eine eindeutige Suche nach Medien erschwert, oder die gefundenen Begriffe sind sehr allgemein. Mit Hilfe der Gewichtung nach Position, wird der Fokus auf den Anfang der Webseite gelegt, der eventuell das Thema des Tages oder das Hauptthema des Portals enthält.

Das Ergebnis ist ein Vektor von extrahierten Keyphrases mit jeweiligem Gewichtungswert.

5.8 Medien finden und sortieren

Aus dem Vektor mit den extrahierten Keyphrases und ihrem jeweiligem Gewichtungswert werden die besten Keyphrases als Suchbegriffe identifiziert. Dabei werden bevorzugt Phrasen ausgewählt, die auch die einzelnen Begriffe enthalten. Zudem werden die Begriffe als Suchbegriffe identifiziert, die die höchste Gewichtung erhalten haben. Mit diesen wird eine Suchanfrage bei MyTag gestellt. Die Menge der Suchbegriffe kann anschließend in MyTag erweitert und verändert werden. Um die Menge der gefundenen Ressourcen relevant zum Dokument zu sortieren, wird mit Hilfe des Kosinus-Ähnlichkeitsmaß zwischen dem Dokument und jeder gefundenen Ressource ein Ranking erstellt. Auch bei einer manuellen Änderung oder Erweiterung der Suchanfrage wird das Ranking angewendet.

Eine verwandte Funktion bietet MyTag bereits durch die personalisierte Suche. Hierbei kann für einen angemeldeten Benutzer eine Personomy erstellt werden. Diese stellt sich intern als Menge von Keywords mit zugehörigem Häufigkeitswerten dar. Die Personomy wird automatisch aufgebaut. Indem der Benutzer einzelne Ressourcen aus der

Ergebnisliste auswählt, werden die Tags der Ressource in die Personomy übernommen. Die Personomy wird als ein Vektor $\vec{p}$, der die Häufigkeitswerte enthält, dargestellt.

Für das personalisierte Ranking wird der Rang r jeder Ressource mit dem Skalarprodukt eines Ressourcen-Vektors $\vec{v}$ und dem Vektor $\vec{p}$ ermittelt: $r = \vec{v} \cdot \vec{p}$. [vgl. BrDeFr+08]

Um die gefundenen Ressourcen in Relevanz zur Eingabe-Webseite zu sortieren, kann eine ganz ähnliche Funktion angewendet werden. Der Rang jeder gefundenen Ressource wird erneut durch ein Skalarprodukt berechnet: Der Ressourcen-Vektor $\vec{v}$ enthält die dazugehörenden Tags mit binären Werten für jeden Tag. Der Dokumentvektor $\vec{d}$ enthält alle Keyphrases mit den Gewichtungswerten. Der Rang wird mit $r = \vec{v} \cdot \vec{d}$ berechnet.

Dieses Ranking ist besonders hilfreich, wenn der oder die extrahierten Suchbegriffe zu allgemein oder mehrdeutig sind. Folgendes Szenario soll dies verdeutlichen:
Für eine Webseite, deren Text von Apple Computern handelt, werden Keyphrases extrahiert. Die Struktur des Textes hat zur Folge, dass das Wort „Computer“ wesentlich stärker als das wort „apple“ gewichtet wurde. Als Suchbegriff für MyTag wird als einziges Wort „Computer“ identifiziert. Die gefundenen Ressourcen werden zwar etwas mit Computern zu tun haben, aber nicht zwingend mit „Apple“-Computern. Dadurch, dass die gefundenen Ressourcen anhand der Cosinus-Ähnlichkeit zwischen jeder Ressource und den Keyphrases der Eingabe-Webseite sortiert werden, werden die Ressourcen in den Suchergebnislisten oben stehen, die nicht nur das Tag „computer“ besitzen, sondern auch „apple“ und die anderen Keyphrases der Webseite.

Der Vektor $\vec{d}$ bleibt während der gesamten Session bei MyTag gespeichert. Somit werden die Ergebnislisten bei einer Veränderung oder Erweiterung der Suchanfrage immer aufs Neue anhand des beschrieben Verfahrens sortiert.

6 Evaluation

In dieser abschließenden Evaluation wurde eine kurzer Fragebogen von 48 Teilnehmern beantwortet und ein Benutzertest mit 3 Personen durchgeführt. Es sollte ermittelt werden, ob im Kontext der allgemeinen Suche im Internet, Interesse an einer Suche nach Medien in Bezug zu einer Eingabe-Webseite besteht. Zudem sollte untersucht werden, ob das Add-on relevante Suchergebnisse liefert, die den Erwartungen der Testpersonen entsprechen.

6.1 Auswertung des Fragebogens

Die meisten befragten Personen sind zwischen 25-40 Jahre alt, also liegen diese in der Altersgruppe, denen eine gute Internetkenntnis zugesprochen werden kann. Dies bestätigt auch die eigene Einschätzung der Teilnehmer zu ihren Internetkenntnissen. Die Personengruppe der über 40-jährigen bezeichnet ihre Kenntnisse hingegen eher als ausreichend bis schlecht. Der junge Teilnehmerkreis unter 25 Jahren zeigt in ihrem Nutzungsverhalten und den eingeschätzten Internetkenntnissen, dass sie quasi mit dem Internet aufgewachsen sind und die meist tägliche Nutzung entsprechendes Vorwissen für die Beantwortung des Fragenbogens generiert hat.

Geschlecht:	28 männlich, 20 weiblich
Alter:	6 unter 25 Jahren , 39 zwischen 25 und 40 Jahren, 3 über 40 Jahren
Wie oft nutzen Sie das Internet?	41 täglich, 5 mehrmals pro Woche, 2 seltener
Wie gut schätzen Sie Ihre Computer- und Internetkenntnisse ein?	14 sehr gut, 29 gut, 3 ausreichend, 2 schlecht

Tab. 1: Daten der befragten Personen
Quelle: Eigene Darstellung

In der ersten Frage wurden die Teilnehmer gefragt, welche der Tagging-Plattformen, die auch von MyTag unterstützt werden, sie bereits genutzt haben. Die Plattform YouTube war der Mehrheit der Befragten bekannt. YouTube wird sogar von einigen mehrmals in der Woche, zumindest aber mehrmals im Monat verwendet. Dieses Ergebnis deckt sich in etwa mit der allgemeinen Popularität von YouTube. Auch Personen, welche ihre eigenen

Internetkenntnisse nur als ausreichend einschätzen, nutzen YouTube zumindest selten. Bei Flickr sieht das schon ganz ander aus, nur 13 Personen gaben an, selten Flickr zu nutzen.
Noch deutlicher zeigt sich bei Delicious der Zusammenhang zwischen Internet-Affinität des Befragten zu seinen Nutzungsverhalten zu Tagging-Systemen. Delicious wird in dieser Befragung ausschließlich von 6 Teilnehmern benutzt, die über gute bis sehr gute Internetkenntnisse verfügen. Insgesamt ist die Zahl der Nutzer hier aber noch geringer als bei Flickr und YouTube.
Die Systeme Connotea, Bibsonomy und Bibtex wurden von allen Befragten noch nie genutzt. Dieses Ergebnis spiegelt aber insgesamt die Popularität der Systeme wieder. Gerade Bibtex mit den Literaturreferenzen spricht eine spezielle Zielgruppe an.
Das Ergebnis dieser Frage hilft, die weiteren Antworten der Teilnehmer für jeden einzelnen besser einschätzen zu können.

	Mehrmals pro Woche	Mehrmals pro Monat	Seltener	Nie
youtube	3	18	24	3
flickr	-	-	13	35
delicious	-	-	6	42
Connotea	-	-	-	48
Bibsonomy	-	-	-	48
Bibtex	-	-	-	48

Tab. 2: Welche der folgenden Plattformen haben Sie bereits genutzt?
Quelle: Eigene Darstellung

In den Fragen 2 und 3 wurden die Teilnehmer gefragt, ob ihnen die in Kapitel 2 erwähnte Google-Ähnlichkeitssuche bekannt ist und ob sie einen solchen Service nützlich finden. Da die Suche auf der Google-Startseite nicht prominent angeboten ist, sondern sich in der erweiterten Suche „versteckt“, war es nicht überraschend, dass den meisten Teilnehmern diese Suche bisher unbekannt war. 11 Personen kannten diese Suche bereits, 37 noch nicht. Allerdings findet mehr als die Hälfte der Befragten einen solchen Service sehr oder eher nützlich.

Sehr nützlich	Eher nützlich	Kaum nützlich	Überflüssig
14	28	6	0

Tab. 3: Wie nützlich finden Sie einer Service, der Ihnen ähnliche Seiten sucht?
Quelle: Eigene Darstellung

Diese Meinung ist in etwa über alle Altersgruppen verteilt, wobei eher Personen mit höherer Affinität zum Internet diesen Service nützlich finden. In diesem Kontext knüpft Frage 4 an. Hier wurde gefragt, wie nützlich ein Service empfunden würde, der nicht eine

ähnliche Seite, sondern passende Medien (Fotos, Videos, Bookmarks) zu einer Webseite findet. Ein Großteil der Personen findet einen solchen Service sehr und eher nützlich, allerdings gaben hier mehr Befragte an, dass für sie ein solcher Service kaum nützlich wäre.

Sehr nützlich	Eher nützlich	Kaum nützlich	Überflüssig
14	24	10	0

Tab. 4: Wie nützlich finden Sie einen Service, der Ihnen zum Inhalt einer Webseite passende Medien heraussucht?
Quelle: Eigene Darstellung

Sehr und eher nützlich ist ein solcher Service eher für männliche als für weibliche Internetnutzer mit guter bis sehr guter Kenntnis in der Altersgruppe bis 40 Jahren.

Insgesamt kann aber gesagt werden, dass ein Interesse an einer direkten Suche nach Medien zu einer Webseite besteht.

In den Fragen 5 und 6 sollte eine solche Suche nach Medien in Bezug zu einem Dokument simuliert werden. Die Teilnehmer wurden gebeten zwei kurze Texte durchzulesen und anschließend passende Suchbegriffe nach Medien zu markieren. Dies ist damit zu vergleichen, dass der Teilnehmer zu einem Text spontan Tags vergibt.

Der erste Text ist eine kurze Zusammenfassung des Wikipedia-Artikels über die Rockgruppe „Bon Jovi“[20]. Dieser Artikel wurde ausgewählt, da der Eigenname „Bon Jovi“ vom entwickelten Service als Phrase erkannt wurde und folglich der Suchbegriff „bonjovi“ an MyTag übermittelt wird. Weitere Keyphrases, wie die Einzelterme“rockband“, „bon“ und „jovi“ sind im Dokumentenvektor enthalten, um die Suchergebnisse entsprechend zu ranken. Es sollte herausgefunden werden, ob die befragten Personen bei freier Eingabe auf die Idee der Suche nach der Phrase kommen würden. Die Teilnehmer hatten die Möglichkeit drei Suchbegriffe zu nennen. Von allen Teilnehmern wurde erwartungsgemäß als Suchbegriff „bon jovi“ oder „Bon Jovi“ genannt. Dies unterstreicht die Notwendigkeit einer Phrasenerkennung im Add-on, da eine manuelle Suche genauso verfahren würde. Dieser Suchbegriff wird bei den meisten Taggingplattformen auch Ergebnisse mit einer ausreichenden Trefferquote liefern, die konkatenierte Phrase „bonjovi“ als Suchbegriff erhöht jedoch die Genauigkeit bzw. die Precision der Suchergebnisse. An dieser Stelle hat sich der erhoffte Mehrwert der Anwendung exemplarisch herausgestellt.

Die Angabe der beiden weiteren Suchbegriffe variiert stark. So wurden einzelne Schlagworte des Textes angegeben, die von den Teilnehmern anscheinend subjektiv als wichtig oder interessant angesehen werden. Auch die Eigenname 3er-Phrase „Jon Bon Jovi“ wur-

[20] Wikipedia: URL: http://de.wikipedia.org/wiki/Bon_Jovi [Stand 24.09.2009]

de häufig genannt. Diese wird aber aufgrund des seltenen Auftretens im Text nicht vom Add-on in die Keyphrases aufgenommen. Jede Kombination von 3 Suchbegriffen wird als „eigener Keyphrase-Vektor“ über dem Dokument relevante Suchergebnisse erzielen. Bei den Resulten des Add-ons ist aber aufgrund der umfangreicheren Auswahl an Keyphrases mit den Gewichtungswerten von einer höheren Genauigkeit auszugehen.

Bon Jovi	44	Band	8	Lieder von Bon Jovi	4	Tico Torres	1
Jon Bon Jovi	16	Livin' on a prayer	6	Rock	3	Album	1
Lieder	13	Hits	6	Wikipedia Bon Jovi	1		
Video	9	Keep the faith	5	Hair Metal Band	1		
Madison Sqare Garden	8	New Jersey	4	Geschichte	1		
Runaway	8	Band	4	Sambora	1		

Tab. 5: Angaben für einen Suchbegriff nach Medien zu einem kurzen Text zur Rockgruppe „Bon Jovi“
Quelle: Eigene Darstellung

Der zweite Artikel ist dem Internetauftritt von Spiegel Online entnommen und handelt von der Sonnenfinsternis in Asien im Juli 2009[21]. Die Verarbeitung dieses Artikels zeigt aber folgende Besonderheiten: Als Keyphrases mit höchster Gewichtung werden die Terme „Sonnenfinsternis“ und „Ereignis“ berechnet. Das Schlagwort „Spektakel“ entspricht in etwa der Bedeutung von „Ereignis“, allerdings werden die beiden Terme getrennt voneinander gewichtet. Eine gemeinsame Verarbeitung unter einem Oberbegriff würde genau diesen höher gewichten. Genauso zeigt es sich bei den Termen „gesehen“, „sehen“, „verfolgen“ und „beobachten“. Jeder Term wird getrennt voneinander in etwa gleich stark gewichtet. Auch hier könnte ein Überbegriff, wie z.B. „Beobachten“ im Kontext des Inhalts gefunden werden.

Die Teilnehmer wurden bei dieser Frage gebeten, aus einer Vorauswahl von Termen zwei auszuwählen, die sie als passende Suchbegriffe für eine Mediensuche in Bezug zum Text als geeignet erachten. Von allen Personen wurde der Begriff „Sonnenfinsternis“ ausgewählt. Dies ist erwartet worden, da dies dem Titel des Textes entspricht und auch im Text häufig auftritt.

Die Auswahl des zweiten Suchbegriffes variiert. So wählten zwei Teilnehmer das Wort „Juli“ aus, mit der Idee, die Suchergebnisse von verschiedenen dokumentierten Sonnenfinsternissen auf die spezielle vom Juli 2009 zu beschränken.

[21] Spiegel Online: Millionen verfolgen längste Sonnenfinsternis des Jahrhunderts, *ala/dpa/AFP* , 27.07.2009 URL: http://www.spiegel.de/wissenschaft/weltall/0,1518,637472,00.html [Stand 24.09.2009]

Neben der Einschränkung auf das Datum ist eine andere Möglichkeit, die Suche auf den Ort zu spezialisieren. Hier gaben rund Dreiviertel aller Befragten den Begriff „Asien“ an. Auffallend ist, dass der Term „Asien“ von der Anwendung nicht in den Vektor von Keyphrases aufgenommen wurde. Der Term tritt nur einmal im Dokument auf, somit haben die angewendeten Methoden diesen ausgeschlossen. Da jedoch verschiedene Länder und Städte Asiens, wie z.B. China, Indien, Nepal, Japan, Neu-Delhi oder Shanghai, im Text erwähnt werden, scheint der Leser den Ort des Geschehens auf Asien zusammenzufassen.

Ereignis	2	Totale	6
Sonnenfinsternis	48	Spektakel	-
Asien	34	Juli	2
Beobachten	-	China	-
Mond	4	Erde	-

Tab. 6: Angaben für zwei Suchbegriffe aus der Liste, die für eine Internetsuche nach zum Text passenden Medien geeignet wären.
Quelle: Eigene Darstellung

6.2 Auswertung der Benutzertests

Für den Benutzertest wurden drei Testpersonen beauftragt, das Add-on in den Firefox zu integrieren und anschließend zwei Suchen auszuführen.

Alle drei Personen sind zwischen 25-40 Jahre alt, männlich und schätzen ihre Internet- und Computerkenntnisse gut (Person 1) und sehr gut (Person 2 und 3) ein und nutzen das Internet täglich.

Die Integrierung des Buttons für das Add-on war für alle Personen ohne große Probleme zu bewältigen. Bis auf ein Verständnisproblem war auch die Funktionsweise sofort ersichtlich.

Die erste Aufgabe für die Probanden war es, die englische Wikipedia-Seite[22] zu öffnen und einen beliebigen Artikel aufzurufen. Testperson 1 gab den Begriff „soccer“ in das Suchfeld ein und wurde zum Artikel „Association football“ weitergeleitet. Die Suche wurde mit dem Button ausgelöst und ergab die Suchbegriffe „football“, „laws“ und „association“. Die Suchbegriffe wurden als sehr relevant bezeichnet, allerdings enthielt die Ergebnisliste von MyTag keine Treffer. Intuitiv hat Testperson 1 in MyTag das Wort „laws“ aus der Suchanfrage gelöscht, mit der Begründung, dass die Artikelseite von Wikipedia nur von „association football“ handelte. Die daraufhin erscheinenden Suchergebnisse erfüllten die Erwartungen zum Teil.

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

Testperson 2 suchte sich den Artikel „great britain“ aus. Nach dem die Suche nach Medien gestartet wurde, erschien die MyTag-Ergebnisliste mit den Suchbegriffen „greatbritain“, „great“ und „britain“. Die Suchbegriffe wurden als gut und korrekt bewertet. Der Begriff der konkatenierten Phrase „greatbritain“ war auf den ersten Blick überraschend. Trotzdem wurden die Suchbegriffe als wenig überraschend im Kontext zum Titel der Seite bewertet.
Die gefundenen Medien entsprachen zum Teil den Erwartungen.
Proband 3 entschied sich als Eisenbahn-Fan für den Artikel „Trains“. Die Suchbegriffe „trains“, „rail“ und „wikipedia“ wurden als gut bezeichnet. Allerdings wurden auch in diesem Fall erst nach dem Entfernen des Suchbegriffes „wikipedia“ und einem Neustart der Suche in MyTag relevante Medien gefunden, die schließlich auch voll und ganz den Erwartungen entsprechend bewertet wurden.

In der nächsten Aufgabe sollte die Suchfunktion für eine frei wählbare Seite angewendet werden. Proband 1 entschied sich für die Seite „www.fcbayern.de“. Als Suchbegriffe wurden „fcbayern“, „fc“ und „bayern“ identifiziert. Die Testperson bewertete die Relevanz der Suchbegriffe als sehr gut und war auch voll und ganz mit den Suchergebnissen zufrieden.
Proband 2 öffnete die Seite der Universität Koblenz unter „www.uni-koblenz.de“. Die Suchbegriffe „campus“ und „koblenz“ wurden als sehr gut bewertet. Die gefundenen Medien entsprachen voll und ganz den Erwartungen.
Proband 3 öffnete den Internetauftritt von „heise online“ unter „www.heise.de“. Die Person klickte sich durch die Webseite und gelang zum Artikel „Nintendo senkt auch in Deutschland den Preis der Wii“[23]. Der Start der Suche ergab die Suchbegriffe „heiseonline“, „preis“ und „heise“. Die Testperson gab die Relevanz dieser Suchbegriffe als nur ausreichend an, da immerhin das Wort „preis“ vertreten war. Allerdings entsprachen „heise“ und „heiseonline“ nicht den Erwartungen. Es wurde der Begriff „wii“ erwartet. Folglich wurden auch die gefundenen Medien als schlecht bewertet. Auch ein „wegklicken“ der Begriffe „heiseonline“ und „heise“ brachte keine guten Suchergebnisse.
Zusammenfassend für die Gesamtbeurteilung der Probanden kann man sagen, dass die Funktion des Add-ons ungefähr den Vorstellungen der Personen entsprach und die Bedie-

[23] Heise online: http://www.heise.de/newsticker/Nintendo-senkt-auch-in-Deutschland-Preis-der-Wii--/meldung/145913 (26.09.2009)

nung und Durchführung keine Schwierigkeiten bereitete. Alle Testpersonen würden das Add-on in Zukunft vielleicht wieder nutzen.

	Das Add on entsprach meinen Vorstellungen.	Ich hatte bei der Durchführung keine Schwierigkeiten.	Ich werde dieses Programm auch in Zukunft nutzen.
Testperson 1	Ja	Stimmt voll und ganz.	Vielleicht
Testperson 2	Zum Teil	Stimmt voll und ganz.	Vielleicht
Testperson 3	Zum Teil	Stimmt zum Teil.	Vielleicht

Tab. 7: Gesamtbeurteilung des Add-on nach Durchführung der Tests.
Quelle: Eigene Darstellung

7 Fazit und Ausblick

Ziel dieser Arbeit war es, ein Add-on zu entwickeln, dass eine Webseite analysiert, passende Suchbegriffe identifiziert und mit diesen eine Suchanfrage an MyTag gestellt wird. Für die Entwicklung waren einige bereits bestehende Komponenten hilfreich, mit denen die beschriebenen Methoden für die Keyphrase Extraction umgesetzt werden konnten. Da diese Funktion als Webservice zur Verfügung gestellt wurde, kann die Kernkomponente weiterentwickelt werden, ohne dass das Add-on erneut im Browser integriert werden muss. Zudem können auch andere Anwendungen mittels SOAP den Webservice konsumieren.
Das Ranking der Suchergebnisse anhand des Dokumentenvektors verbessert die Qualität der Suchergebnisse schon beim ersten Testen während der Entwicklung deutlich.

Die Benutzertests haben ergeben, dass die angewendeten Verfahren meist, aber nicht immer relevante Suchergebnisse liefern. Die Befragung stützt die praktischen Erfahrungen während der Entwicklung. Es gibt Keyphrases, die die Relevanz der Suchergebnisse in Bezug zur Eingabe-Webseite signifikant erhöhen. Im Gegensatz dazu werden auch Keyphrases als Suchbegriff identifiziert, die nicht gemeinsam für eine Suchanfrage nutzbar sind. Zumindest wird dem Nutzer aber eine Auswahl präsentiert, mit der die Anfrage bei MyTag erweitert oder verändert werden kann.

Die automatisierte Verschlagwortung von Webseiten ist im Gegensatz zur manuellen präziser und deckt das komplette Dokument ab. Die per TF-IDF ermittelten Gewichtungswerte für die Wörter entsprechen dem inhaltlichen Sinn der untersuchten Webseiten.

Schwächen gibt es noch bei der Verarbeitung mit bestimmten Webseiten. So wird bei Artikeln von Onlinejournalen (wie z.B. heise online[24], Spiegel online[25]) oft der Name des Anbieters in Relation zum eigentlichen Inhalt des Artikels zu hoch gewichtet. Hier könnte man in Zukunft die Struktur von HTML-Dokumenten stärker nutzen. Eine Idee wäre, die URL zu analysieren und daraus Rückschlüsse auf die Struktur der Webseite zu ziehen. Ein anderer Ansatz bestünde darin, das Document Object Model zu nutzen, um bestimmte Teile eines HTML-Dokumentes wie Navigation und Hauptteil zu identifizieren.

Es hat sich herausgestellt, dass eine Behandlung von Synonymen und Akronymen die Qualität der Suche verbessern würde. Die Suchstrategie einer Person basiert häufig darauf, Wörter zusammenzufassen, die die gleiche Bedeutung haben. Im Gegensatz dazu werden Mehrdeutigkeiten aufgelöst, da sie für eine Person aus dem Kontext eines Dokumentes ersichtlich sind. Um diese semantischen Verknüpfungen zu erkennen, kann die Verwendung von Ontologien hilfreich sein.
Ein erster Ansatz für die Behandlung von Mehrdeutigkeiten ist der Disambiguierungsservice von MyTag. Dieser ist auch auf die Ergebnisliste anzuwenden, könnte aber auch schon vor dem Ranking verwendet werden, um die Identifizierung der Suchbegriffe zu optimieren.

[24] http://www.heise.de
[25] http://www.spiegel.de

8 Anhang

Fragebögen zur Evaluation

Auswertung der Fragebögen

Eingefügter Source Code in MyTag

Daten CD

Befragung: „Suche nach passenden Medien zu einer Webseite“

Frage 1:
Welche der folgenden Plattformen haben Sie bereits genutzt?

- ☐ Mehrmals pro Woche
- ☐ Mehrmals pro Monat
- ☐ Seltener
- ☐ Noch nie

- ☐ Mehrmals pro Woche
- ☐ Mehrmals pro Monat
- ☐ Seltener
- ☐ Noch nie

- ☐ Mehrmals pro Woche
- ☐ Mehrmals pro Monat
- ☐ Seltener
- ☐ Noch nie

- ☐ Mehrmals pro Woche
- ☐ Mehrmals pro Monat
- ☐ Seltener
- ☐ Noch nie

- ☐ Mehrmals pro Woche
- ☐ Mehrmals pro Monat
- ☐ Seltener
- ☐ Noch nie

- ☐ Mehrmals pro Woche
- ☐ Mehrmals pro Monat
- ☐ Seltener
- ☐ Noch nie

Frage 2:
Google bietet eine Ähnlichkeitssuche an. Dies ist eine Suche nach Webseiten, die ähnlich der Eingabe-Webseite sind.
Haben Sie diese Suche bereits genutzt?

- ☐ Ja
- ☐ Nein

Frage 3:
Stellen Sie sich vor, Sie besuchen eine interessante Internetseite. Wie nützlich finden Sie einen Service, der Ihnen ähnliche Seiten sucht?

- ☐ Sehr nützlich
- ☐ Eher nützlich
- ☐ Kaum nützlich
- ☐ Überflüssig

Frage 4:
Wie nützlich finden Sie einen Service, der Ihnen zum Inhalt einer besuchten Webseite passende Medien (Fotos, Videos, Bookmarks) heraussucht?

- ☐ Sehr nützlich
- ☐ Eher nützlich
- ☐ Kaum nützlich
- ☐ Überflüssig

Frage 5:
Bitte lesen Sie folgenden Text durch. Geben Sie anschließend drei Suchbegriffe an, mit deren Hilfe Sie zum Text passende Medien (Fotos, Videos, Bookmarks) im Internet suchen würden.

Eine Rockband aus New Jersey

Bon Jovi ist eine Rockband aus New Jersey (USA). Sie wurde Anfang der 1980er-Jahre gegründet und hat inzwischen mit der Anzahl verkaufter Alben die 120-Millionen-Marke überschritten. Die Band, die als Hair-Metal-Band begann, spielt seit den Neunziger Jahren vor allem Mainstream-Rock. Einen großen Beitrag zum kommerziellen Erfolg von Bon Jovi leistete der Komponist Desmond Child, der mit Jon Bon Jovi und Richie Sambora Hits wie Livin' on a Prayer, You Give Love a Bad Name, Keep the Faith oder This Ain't a Love Song komponierte.

Nachdem das von Jon Bon Jovi komponierte und bereits 1982 mit Sessionmusikern aufgenommene Lied „Runaway" den ersten Platz bei einem Radio-Talentwettbewerb erreichte und Jon Bon Jovi einen Schallplattenvertrag ermöglichte, gründeten er und David Bryan 1983 zusammen mit Richie Sambora, Alec John Such und Tico Torres die Band Bon Jovi. Die Original-Demo-Aufnahme von „Runaway", auf der u. a. Keyboarder Roy Bittan von Bruce Springsteens Band, Bassist Hugh McDonald und Gitarrist Tim Pierce zu hören sind, wurde auf dem Debütalbum der Band veröffentlicht und wurde ein internationaler Hit.

Bei einem Auftritt als Vorgruppe für Scandal wurden Bon Jovi von Derek Shulman entdeckt. Jon Bon Jovi wurde ein Plattenvertrag von PolyGram angeboten. Er ist der einzige, der einen Plattenvertrag besitzt, die anderen Mitglieder der Band sind nur seine „Angestellten".

Das Debütalbum Bon Jovi erschien am 21. Januar 1984. Wenig später trat die Band als Vorgruppe für ZZ Top im Madison Square Garden auf, ebenso spielten sie vor den Scorpions und inDeutschland vor KISS.

- 1. Suchbegriff: _____
- 2. Suchbegriff: _____
- 3. Suchbegriff: _____

Frage 6:

Bitte lesen Sie folgenden Text durch. Wählen Sie bitte anschließend zwei Suchbegriffe aus der Liste aus, die für eine Internetsuche nach zum Text passenden Medien geeignet wären.

Millionen verfolgen längste Sonnenfinsternis des Jahrhunderts

Begeisterung und Furcht: Millionen Menschen haben in Asien die längste totale Sonnenfinsternis dieses Jahrhunderts verfolgt - mit durchaus gemischten Gefühlen. Während die einen das seltene Spektakel genossen, sprachen andere von einem "sehr gefährlichen Moment im Universum".
Neu-Delhi - Um 1.58 Uhr deutscher Zeit schob sich über dem Arabischen Meer der Mond vor die Sonne. Wenig später erreichte das Ereignis den direkt im Kernschatten der Sonnenfinsternis liegenden westindischen Bundesstaat Gujarat.
In der Hauptstadt Neu-Delhi konnten die Menschen bei klarem Himmel miterleben, wie sich 80 Prozent des Mondes vor die Sonne schoben. In Mumbai hatten die "Sofi"-Fans weniger Glück: Wegen dichter Monsun-Regenwolken war das Spektakel hier nicht zu sehen, die speziellen Sonnenbrillen im Nehru-Planetarium erwiesen sich als wertlos. Stattdessen hüllten sich die Beobachter in Regenjacken und spannten Regenschirme auf.
Natürlich wurde auch Geld mit dem Naturereignis verdient: Ein Sonderflug der Firma Cox and Kings hob von Neu Delhi Richtung Osten ab, um den Fluggästen eine direkte Beobachtung der Finsternis zu ermöglichen. Die 21 Sitzplätze der Boeing 737-700 auf der Sonnenseite wurden für 79.000 Rupien (rund 1200 Euro) verkauft.
Nicht überall auf dem Subkontinent herrschte grenzenlose Begeisterung angesichts des seltenen Spektakels: In Indien fürchten viele Menschen negative Auswirkungen der verdunkelten Sonne. So nahmen in der als heilig geltenden Stadt Varanasi Hunderttausende gläubige Hindus unmittelbar nach Ende der Finsternis ein reinigendes Bad im Ganges. In Neu-Delhi verschoben etliche Schwangere die für diesen Tag geplante Geburt per Kaiserschnitt.
Der 22. Juli sei ein "sehr gefährlicher Moment im Universum", warnte der indische Astrologe Raj Kumar Sharma. "Wenn die Sonne, die Anführerin unter den Gestirnen, krank ist, dann bedeutet das, dass es auf der Welt große Probleme geben wird."
"Plötzlich habe ich es gesehen"
Der Korridor in einer Breite von 258 Kilometern zog sich von Indien über Nepal, Bhutan, Bangladesch, Burma und China zu den südjapanischen Inseln. Nach den Berechnungen der Astronomen hatte er eine Länge von 15.000 Kilometern.
In China war das Spektakel ab 3.13 Uhr zu beobachten. Die totale Sonnenfinsternis war unter anderem in der zentralchinesischen Stadt Chongqing zu sehen. Die Behörden hatten damit gerechnet, dass etwa 300 Millionen Menschen im Tal des Jiangtse das seltene Naturschauspiel miterleben können. Vielerorts behinderte jedoch auch hier schlechtes Wetter den Blick auf das Ereignis.
Auch in Shanghai regnete es während der Sonnenfinsternis zeitweise, ab und zu riss die Wolkendecke jedoch auf. "Die Wolken zogen auf, dann bildeten sich Lücken und plötzlich habe ich es gesehen", schwärmte der Geschäftsmann Glenn Evans aus den USA, der in Shanghai arbeitet. In den Ufercafés der Stadt wurden spezielle Sonnenfinsternis-Frühstücke serviert.
Die längste Sonnenfinsternis gab es in Japan.In Japan war die Sonnenfinsternis auf der Insel

Akuseki mit über sechs Minuten am längsten unter den von Menschen bewohnten Orten der Welt zu sehen. Normalerweise wohnen auf der Insel 68 Menschen - anlässlich des Naturereignisses kamen mehrere hundert.
Der deutsche Filmregisseur Roland Emmerich nutzte die Sonnenfinsternis am Mittwoch, um in der Hauptstadt Tokio Werbung für seinen neuen Katastrophenfilm "2012" zu machen. Der noch nicht ganz fertige Streifen, der im November in die Kinos kommt, bezieht sich auf die Prophezeiung der Maya, deren Sonnenkalender am 21. Dezember 2012 endet und als Anlass dient, das Ende der Welt in ihrer bisherigen Form zu datieren.
Eigentlich wollte Emmerich mit seinen Gästen und Journalisten die Sonnenfinsternis auf dem Dach des 238 Meter hohen Mori Towers in Roppongi Hills verfolgen, einem schicken Konsumtempel in Tokios Szeneviertel Roppongi. Doch daraus wurde nichts: Die Sonnenfinsternis versteckte sich hinter dicken Regenwolken. Wer Glück hatte, konnte für Sekunden immerhin die Sonnensichel sehen.
Die Sonnenfinsternis überzog einige der am dichtesten besiedelten Gebiete der Erde, so dass sie die größte Beobachterschar in der Geschichte der Menschheit gehabt haben dürfte. Eine Finsternis ähnlichen Ausmaßes haben die Astronomen erst wieder für das Jahr 2132 errechnet. Die nächste Sonnenfinsternis wird sich am 11. Juli 2010 ereignen, jedoch lediglich im Südpazifik zu beobachten sein.
Sonnenfinsternisse faszinieren die Menschheit seit jeher. In China wurde das Phänomen, bei dem sich der Mond zwischen Erde und Sonne schiebt und damit einen Schatten auf die Erde wirft, traditionell so erklärt, dass ein Drache das Himmelsgestirn verschluckt. In der hinduistischen Mythologie werden die Dämonen Rahu und Ketu für die Verfinsterung verantwortlich gemacht.

- ☐ Ereignis
- ☐ Sonnenfinsternis
- ☐ Asien
- ☐ Beobachten
- ☐ Mond
- ☐ Totale
- ☐ Spektakel
- ☐ Juli
- ☐ China
- ☐ Erde

Frage 7:
Fragen zu Ihrer Person

- Geschlecht
  - ☐ Weiblich
  - ☐ Männlich
- Alter
  - ☐ Unter 25 Jahre
  - ☐ 25-40 Jahre
  - ☐ Über 40 Jahre
- Wie oft nutzen Sie das Internet?
  - ☐ Täglich
  - ☐ Mehrmals pro Woche
  - ☐ Seltener
  - ☐ Nie
- Wie gut schätzen Sie Ihre Computer- und Internetkenntnisse selbst ein?

- ☐ Sehr gut
- ☐ Gut
- ☐ Ausreichend
- ☐ Schlecht

Vielen Dank für Ihre Mitarbeit!

Benutzertest

1. Installation

Öffnen Sie mit Firefox die Webseite „http://vulkan.uni-koblenz.de“. Folgen Sie den Anweisungen der Seite und platzieren Sie das Add-On.

Frage 1:
Die Installationsanweisungen waren

- ☐ sofort ohne Probleme nachvollziehbar und verständlich.
- ☐ nach anfänglichen kleinen Schwierigkeiten verständlich.
- ☐ nicht nachvollziehbar und daher nicht durchführbar.

2. Test des Add Ons auf der englischen Wikipediaseite

Öffnen Sie die Seite „http://en.wikipedia.org“. Wählen Sie dort einen Artikel nach Ihren Interessen aus.

Gewählter Artikel: _____
Starten Sie mithilfe des Add-On eine Suche.

Frage 2:
Wie würden Sie die Effizienz der Suche beurteilen?

- ☐ Sehr schnell.
- ☐ Schnell.
- ☐ Langsam.
- ☐ Sehr langsam.

Frage 3:
Wie würden Sie die Relevanz der Suchbegriffe einschätzen?

- ☐ Sehr gut
- ☐ Gut
- ☐ Ausreichend
- ☐ Schlecht

Bitte begründen Sie Ihre ausgewählte Antwort kurz!

_____
_____

Frage 4:
Entsprechen die gesuchten Medien Ihren Erwartungen?

- ☐ Ja, voll und ganz.
- ☐ Ja, zum Teil
- ☐ Nein.
- ☐ Ich hatte keine Erwartungen.

3. Test des Add Ons auf einer frei gewählten Seite

Nachdem Sie das Add-On im Punkt 2 auf einer festgelegten Seite getestet haben, sollen Sie jetzt mithilfe des Add Ons auf einer Seite Ihrer Wahl die Suche durchführen.

Gewählte Seite:_____

Frage 5:

Wie würden Sie die Relevanz der Suchbegriffe einschätzen?

- ☐ Sehr gut
- ☐ Gut
- ☐ Ausreichend
- ☐ Schlecht

Bitte begründen Sie Ihre ausgewählte Antwort kurz!

_____

_____

Frage 6:

Entsprechen die gesuchten Medien Ihren Erwartungen?

- ☐ Ja, voll und ganz.
- ☐ Ja, zum Teil
- ☐ Nein.
- ☐ Ich hatte keine Erwartungen.

4. Gesamtbeurteilung

Das Add On entsprach meinen Vorstellungen.

- ☐ Ja.
- ☐ Zum Teil.
- ☐ Nein.

Ich hatte bei der Durchführung keine Schwierigkeiten.

- ☐ Stimmt voll und ganz.
- ☐ Stimmt zum Teil.
- ☐ Stimmt nicht.

Ich werde dieses Programm auch in Zukunft nutzen.

- ☐ Ja.
- ☐ Vielleicht.
- ☐ Nein.

Auswertung des Fragebogens

Frage 1:
Welche der folgenden Plattformen haben Sie bereits genutzt?

	Mehrmals pro Woche	Mehrmals pro Monat	Seltener	Nie
youtube	3	18	24	3
flickr	-	-	13	35
delicious	-	-	6	42
Connotea	-	-	-	48
Bibsonomy	-	-	-	48
Bibtex	-	-	-	48

Frage 2:
Google bietet eine Ähnlichkeitssuche an. Dies ist eine Suche nach Webseiten, die ähnlich der Eingabe-Webseite sind. Haben Sie diese Suche bereits genutzt?

ja	nein
11	37

Frage 3:
Stellen Sie sich vor, Sie besuchen eine interessante Internetseite. Wie nützlich finden Sie einer Service, der Ihnen ähnliche Seiten sucht?

Sehr nützlich	Eher nützlich	Kaum nützlich	Überflüssig
14	28	6	0

Frage 4:
Wie nützlich finden Sie einen Service, der Ihnen zum Inhalt einer Webseite passende Medien heraussucht?

Sehr nützlich	Eher nützlich	Kaum nützlich	Überflüssig
14	24	10	0

Frage 5:
Bitte lesen Sie folgenden Text durch. Geben Sie anschließend drei Suchbegriffe an, mit deren Hilfe Sie zu Text passende Medien im Internet suchen würden.

Bon Jovi	44	Band	8	Lieder von Bon Jovi	4	Tico Torres	1
Jon Bon Jovi	16	Livin' on a prayer	6	Rock	3	Album	1
Lieder	13	Hits	6	Wikipedia Bon Jovi	1		
Video	9	Keep the faith	5	Hair Metal Band	1		
Madison Sqare Garden	8	New Jersey	4	Geschichte	1		
Runaway	8	Band	4	Sambora	1		

Aufgabe 6:
Bitte lesen Sie folgenden Text durch. Wählen Sie bitte anschließend zwei Suchbegriffe aus der Liste aus, die für eine Internetsuche nach zum Text passenden Medien geeignet wären.

Ereignis	2	Totale	6
Sonnenfinsternis	48	Spektakel	-
Asien	34	Juli	2
Beobachten	-	China	-
Mond	4	Erde	-

Aufgabe 7:
Fragen zu Ihrer Person

-
-
-
-

Geschlecht:	28 männlich, 20 weiblich
Alter:	6 unter 25 Jahren , 39 zwischen 25 und 40 Jahren, 3 über 40 Jahren
Wie oft nutzen Sie das Internet?	41 täglich, 5 mehrmals pro Woche, 2 seltener
Wie gut schätzen Sie Ihre Computer- und Internetkenntnisse ein?	14 sehr gut, 29 gut, 3 ausreichend, 2 schlecht

Auswertung des Benutzertests

Frage 1:
Die Installationsanweisungen waren

	sofort ohne Probleme nachvollziehbar und verständlich.	nach anfänglichen kleinen Schwierigkeiten verständlich.	nicht nachvollziehbar und daher nicht durchführbar.
Testperson 1	x		
Testperson 2	x		
Testperson 3		x	

Öffnen Sie die Seite „http://en.wikipedia.org“. Wählen Sie dort einen Artikel nach Ihren Interessen aus.

	Gewählter Artikel
Testperson 1	Soccer (angezeigt wurde: Association football)
Testperson 2	Great Britain
Testperson 3	Train

Frage 2:
Wie würden Sie die Effizienz der Suche beurteilen?

	Sehr schnell	Schnell	Langsam	Sehr langsam
Testperson 1			x	
Testperson 2				x
Testperson 3				x

Frage 3:
Wie würden Sie die Relevanz der Suchbegriffe einschätzen?

	Suchbegriffe	Sehr gut	Gut	Ausreichend	Schlecht	Begründung
Testperson 1	football laws association	x				Die Suchbegriffe passen gut zum Text. Allerdings erhält man zunächst keine Ergebnisse. Dazu muss man einen der Begriffe entfernen.
Testperson 2	greatbritain britain great		x			Die Suchbegriffe entsprechen der Überschrift. Korrekt, aber nicht überraschend, bis auf das zusammengesetzte Wort.
Testperson 3	trains		x			Der Begriff Wikipedia muss manuell gelöscht werden,

	rail wikipedia					daher nur gut.

Frage 4:
Entsprechen die gesuchten Medien Ihren Erwartungen?

	Ja, voll und ganz.	Ja, zum Teil.	Nein.	Ich hatte keine Erwartungen.
Testperson 1		x		
Testperson 2		x		
Testperson 3	x			

Nachdem Sie das Add-On im Punkt 2 auf einer festgelegten Seite getestet haben, sollen Sie jetzt mithilfe des Add Ons auf einer Seite Ihrer Wahl die Suche durchführen.

	Gewählte Seite
Testperson 1	www.fcbayern.de
Testperson 2	www.uni-koblenz-landau.de/koblenz
Testperson 3	www.heise.de/newsticker/Nintendo-senkt-auch-in-Deutschland-Preis-der-Wii--/meldung/145913

Frage 5:
Wie würden Sie die Relevanz der Suchbegriffe einschätzen?

	Suchbe-griffe	Sehr gut	Gut	Ausreichend	Schlecht	Begründung
Testper-son 1	fcbayern, fc, bayern	x				Suchbegriffe geben Rückschluss auf den Inhalt, gute Medien!
Testper-son 2	campus, koblenz	x				Die Wörter Campus und Koblenz entsprechen meinen Erwartungen.
Testper-son 3	heiseonli-ne, preis, heise			x		Heise und online bringen nichts für die Suche. Der Begriff Preis ist ok. Ich hätte den Begriff Wii erwartet..

Frage 6:
Entsprechen die gesuchten Medien Ihren Erwartungen?

	Ja, voll und ganz.	Ja, zum Teil.	Nein.	Ich hatte keine Erwartungen.
Testperson 1	x			
Testperson 2	x			
Testperson 3			x	

Gesamtbeurteilung:

	Das Add on entsprach meinen Vorstellungen.	Ich hatte bei der Durchführung keine Schwierigkeiten.	Ich werde dieses Programm auch in Zukunft nutzen.
Testperson 1	Ja	Stimmt voll und ganz.	Vielleicht
Testperson 2	Zum Teil	Stimmt voll und ganz.	Vielleicht
Testperson 3	Zum Teil	Stimmt zum Teil.	Vielleicht

```
#Source Code MyTag

in routes.rb:
  map.url 'url/*address',
  :controller => 'search',
  :action => 'uris'


in search_controller.rb:
        def list
        ...
        ...
                #sort by document similarity
                #added by Hagen Metzler - haegarm@uni-koblenz.de
                if (($document == true)) then

                        puts "*** Sortierung nach Dokumentähnlichkeit gestartet ***"

                        @response_result_lists.each do |result_list|

                        result_list.results =
        sort_by_correlation_with_document($doc_hash,result_list.results)

                end

                puts "*** Sortierung nach Dokumentähnlichkeit beendet ***"

                end
        ...
        ...
        end

 # sort a result list (column) by correlation with a document
      def sort_by_correlation_with_document(term_personomy,result_list)

                if term_personomy.blank? then
                        return result_list
                else
                result_list.each do |result|

                  result.correlation = correlation_with_personomy(result, term_personomy)
                end
                return result_list.sort_by do |result|
                        -result.correlation
                end
                end

        rescue Exception => e
        puts "rescue document!"
        puts e.backtrace

        return result_list

      end

def uris
         set_language
         save_original_uri

         enc_url = params[:address][0].to_s

         #Webservice konsumieren
         require 'soap/wsdlDriver'
         require 'soap/rpc/driver'
         require 'cgi'

         wsdlfile = "http://vulkan.uni-koblenz.de:8080/addon/services/Service?wsdl"
         driver = SOAP::WSDLDriverFactory.new(wsdlfile).create_rpc_driver
         result = driver.getter(enc_url.to_s)

         terms_frequencies = Hash.new
         count = 0
          #Der Hash wird gefüttert.
         result.length.times do
           terms_frequencies[result[count][0]] = result[count][1].to_i
         count += 1
         end


         #Bester Wert aus Array
         count = 0
         best = 0
         position = 0
```

```
        bestterm = String.new
        result.length.times do
          if (result[count][1].to_i > best) then
            bestterm = result[count][0]
            best = result[count][1].to_i
            position = count
          end
        count += 1
        end

        result.delete_at(position)

        #Zweitbester Wert aus Array
        count = 0
        best2 = 0
        secondbestterm = String.new
        result.length.times do
          if (result[count][1].to_i > best2) then
            secondbestterm = result[count][0]
            best2 = result[count][1].to_i
          end
        count += 1
        end

       if ((result[0][0] == bestterm) or (result[0][0] == secondbestterm)) then
              inquiry = bestterm+" "+secondbestterm
        else
              inquiry = bestterm+" "+secondbestterm+" "+result[0][0]
        end

        $doc_hash = terms_frequencies
        $document = true
        setInstanceVariables(params)
       # youtube flickr delicious connotea bibsonomy bibtex
          redirect_to search_url(
                                  :document => true,
                                  :tags => inquiry,
                                  :tagging_systems => "youtube flickr delicious connotea
bibsonomy bibtex",
                                  :sort_by => @sort_by,
                                  :per_page => @per_page,
                                  :scope => @scope,
                                  :personal_search => @personal_search,
                                  :disambiguate_checked => @disambiguate_checked,
                                  :use_cache => 'false',
          :old_tagging_systems => @tagging_systems,
          :old_scope => @scope,
          :old_personal_search => @personal_search
          )

      end
```

9 Literaturverzeichnis

[Alby08] Alby, Tom: Web 2.0 / Tom Alby. – 3., überarb. Aufl.- München: Hanser, 2008

[AlHe08] Allemang, Dean: Semantic Web for the working ontologist: modeling in RDF, RDFS and OWL / Dean Allemang; James Hendler. – Amsterdam [u. a.]: Morgan Kaufmann, 2008

[BrDeFr+08] Max Braun, Klaas Dellschaft, Thomas Franz, Dominik Hering, Peter Jungen, Hagen Metzler, Eugen Müller, Alexander Rostilov, Carsten Saathoff: Personalized Search and Exploration with MyTag. - University of Koblenz-Landau: WWW 2008, April 21–25, 2008, Beijing, China. ISWeb - Information Systems and Semantic Web -, 2008

[Dell07] Dellschaft, Klaas: The Influence of Users on Each Other in Tagging Systems / Klaas Dellschaft. - Universität Koblenz-Landau: IS Web, Klausurtagung, 2007

[DeGöSz09] Dellschaft, Klaas: Sense Aware Searching and Exploration with MyTag / Klaas Dellschaft; Olaf Göerlitz; Martin Szomszor. - Koblenz: University Koblenz-Landau, 2009

[Fens07] Fensel, Dieter: Enabling semantic web services: the web service modeling ontology; with 2 tables / Dieter Fensel. – Berlin [u. a.]: Springer, 2007

[Ferb03] Ferber, Reginald: Information Retrieval: Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web / Reginald Ferber. – 1. Aufl. – Heidelberg: Dpunkt-Verl., 2003

[FrTeWa07] Frotscher, Thilo: Java Web Services mit Apache Axis 2 / Thilo Frotscher; Marc Teufel; Dapeng Wang. -Frankfurt: entwickler.press, 2007

[Hitz08] Hitzler, Pascal: Semantic Web / Pascal Hitzler … . - 1. Aufl. - Berlin [u.a.]: Springer, 2008

[Lewa05] Lewandowski, Dirk: Web information retrieval: Technologien zur Informationssuche im Internet / Dirk Lewandowski. – Frankfurt am Main: Dt. Ges. für Informationswiss. und Informationspraxis, 2005

[MaPrSc08] Manning, Christopher: Introduction to information retrieval / Christopher D. Manning; Prabhakar Raghavan; Hinrich Schütze. – Cambridge [u. a.]: Cambridge Univ. Press, 2008

[MusKe00] Musciano, Chuck: HTML and XHTML: the definitive guide / Chuck Musciano; Bill Kennedy. - 4. ed. - Beijing [u. a.]: O'Reilly, 2000

[Seel75] Seelbach, Dieter: Computerlinguistik und Dokumentation: Key-Phrases in Dokumentationsprozessen / Dieter Seelbach. – München: Verlag Dokumentation, 1975

[Schw86] Schwarz, Christoph: Informationslinguistische Texterschließung / Christoph Schwarz. - Hildesheim [u. a.]: Olms, 1986

[Stoc00] Stock, Wolfgang G.: Informationswirtschaft: Managment externen

Wissens / Wolfgang G. Stock. -München [u.a.]: Oldenbourg Wissenschaftsverlag GmbH, 2000

[Stoc07] Stock, Wolfgang G.: Information-Retrieval: Informationen suchen und finden / Wolfgang G. Stock. -Wien: Oldenbourg Wissenschaftsverlag GmbH, 2007

[StSt08] Stock, Wolfgang G.: Wissensrepräsentation: Informationen auswerten und bereitstellen / Wolfgang G. Stock; Mechtild Stock. -München: Oldenbourg Wissenschaftsverlag GmbH, 2008

[Turn97] Turney, Peter D.: Extraction of Keyphrases from Text: Evaluation of Four Algorithms / Peter D. Turney. - Institute for Information Technology, National Research Council of Canada, K1A 0R6, 1997

[Turn00] Turney, Peter D.: Learning Algorithms for Keyphrase Extraction / Peter D. Turney. - Institute for Information Technology, National Research Council of Canada, K1A 0R6, 2000

[Turn03] Turney, Peter D.: Coherent Keyphrase Extraction via Web Mining / Peter D. Turney. - Institute for Information Technology, National Research Council of Canada , K1A 0R6, 2003

[Wöhr04] Wöhr, Heiko: Web-Technologien: Konzepte - Programmiermodelle – Architekturen / Heiko Wöhr. – 1. Aufl. – Heidelberg: Dpunkt-Verl., 2004

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten und nicht veröffentlichten Schriften entnommen wurden, sind als solche kenntlich gemacht. Die Arbeit ist in gleicher oder ähnlicher Form oder auszugsweise im Rahmen einer anderen Prüfung noch nicht vorgelegt worden.

Nickenich, den 30.09.2009

_____

Hagen Metzler