Webometrische Analysen mit Hilfe der Google Web APIs

Philipp Mayr und Fabio Tosques, Berlin

Erscheint in: Information: Wissenschaft und Praxis 56(1), 2005

Zusammenfassung

Der Report stellt die Möglichkeiten und Einschränkungen der Google Web APIs (Google API) dar. Die Implementierung der Google API anhand einzelner informationswissenschaftlicher Untersuchungen aus der Webometrie ergibt, dass die Google API mit Einschränkungen für internetbezogene Untersuchungen eingesetzt werden können. Vergleiche der Trefferergebnisse über die beiden Google-Schnittstellen Google API und die Standard Weboberfläche Google.com (Google Web) zeigen Unterschiede bzgl. der Reichweite, der Zusammensetzung und Verfügbarkeit. Die Untersuchung basiert auf einfachen und erweiterten Suchanfragen in den Sprachen Deutsch und Englisch. Die analysierten Treffermengen der Google API bestätigen tendenziell frühere Internet-Studien.

Abstract

This report describes possibilities and restrictions of the Google Web APIs (Google API). The implementation of the Google API in the context of information science studies from the webometrics field shows, that the Google API can be used with restrictions for internet based studies. The comparison of hit results from the two Google interfaces Google API and the standard web interface Google.com (Google Web) shows differences concerning range, structure und availability. The study bases on simple and extended queries in the languages German and English. The analysed results of the Google API confirm the broad tendency of former internet studies.

1 Einleitung

Die Analyse des Internets und seiner Anwendungen tangiert inzwischen unterschiedlichste Fächer wie z.B. die Informationswissenschaft, die Informatik, die Wirtschaftswissenschaften, die Psychologie und viele weitere Fächer. Ein Grund für das zunehmende wissenschaftliche Interesse am Internet sind die große und immer noch wachsende Zahl an Webusern, Anwendungen, Inhalten und Webservern weltweit. Beispielsweise basiert die heutige wissenschaftliche Informationsversorgung und Kommunikation zum großen Teil auf Internettechnologien. Der Bereich der wissenschaftlichen Publikationen ist inzwischen z.T. deutlich besser über das Internet zu erreichen als über traditionelle Medien.

In den letzten Jahren haben sich einzelne fächerübergreifende wissenschaftliche Spezialgebiete herausgebildet, die sich mit dem Begriff *Internet Research*[1] zusammenfassen lassen. Erste großflächige Untersuchungen des Internets von *Lawrence & Giles* aus den Jahren 1998/9 in *Science* und *Nature* zeigten schon früh, dass ein Großteil der Webuser Suchmaschinen benutzen um Informationen im Internet zu finden [*7, 8*]. Heute mag dieses Ergebnis im Hinblick auf die dynamische Entwicklung des Internets sehr weit zurückliegen, es kann grundsätzlich aber davon ausgegangen werden, dass Suchmaschinen auch heute noch die wichtigste Anwendung sind, um wissenschaftliche und nichtwissenschaftliche Inhalte im Internet zu finden. Die großen Internet-Suchmaschinen wie Google, Yahoo und MSN übernehmen heute zunehmend die Funktion der Gatekeeper zur elektronischen Information und sind laut eigener Angaben die meistgenutzten Internetanwendungen.

Trotz einer Vielzahl unterschiedlichster Suchmaschinen besteht an der Dominanz von Google heute kein Zweifel [vgl. 9]. Google ist zur Zeit schlicht die bekannteste Suchmaschine[2]. Dies zeigt sich vor allem an der Anzahl der täglichen Suchanfragen[3]. Hinzu kommt, dass Google über lange Zeit hinweg mit Abstand den größten Index bereitgestellt hat und bzgl. der Verfügbarkeit und Usability hochoptimiert ist. Die Google-Oberfläche und Performance der Anfragebearbeitung hat für viele Webanwendungen Maßstäbe gesetzt. Wie eine aktuelle Untersuchung [6] zeigt, ist Google heute nach wie vor die inhaltlich relevanteste Suchmaschine mit qualitativ guten Trefferergebnissen.

Wir wollen in diesem Report der Frage nachgehen, ob der relativ unbekannte Google Web Service „Google Web APIs“[4] (Google API) als wissenschaftliches Erhebungsinstrument interessante Einsatzmöglichkeiten bietet. Die Google API wurden bislang in wissenschaftlichen Veröffentlichungen lediglich am Rande erwähnt [17]. Eine intensivere Beschäftigung (Implementation und Test) mit der Google API hat unserer Kenntnis nach zumindest in der informationswissenschaftlichen Fachdiskussion noch nicht stattgefunden. Folglich wollen wir die Google Web APIs anhand eigener Beispiele vorstellen. Um eine Vergleichbarkeit der Google API mit der Standard Google Suche (Google Web) zu erreichen, werden die Trefferergebnisse der beiden Google-Schnittstellen im Report gegenübergestellt. Die folgenden Untersuchungen können keine „harten“ wissenschaftlichen Fakten generieren; ihr Ziel ist es vielmehr, die Google APIs als Experimentierfeld für Internetforscher (Webometriker) vorzustellen.

Nachfolgend wird zunächst die Funktionsweise der Google Web APIs beschrieben.

---

[1] Siehe dazu auch die Homepage der Association of Internet Researchers (AOIR) http://www.aoir.org/ .

[2] Diese Tatsache wird z.B. durch die Wortschöpfung 'googeln' belegt, die den neunten Platz in der von der Gesellschaft für deutsche Sprache erstellten Rangliste der Wörter des Jahres 2003 (siehe http://www.gfds.de/presse.html) einnimmt.

[3] Laut Google.com wird der Google Index deutlich über 100.000.000 mal am Tag befragt. Vgl. http://www.google.com/googlefriends/moreapr03.html

[4] Siehe dazu http://www.google.com/apis.

2 Google Web APIs

Der immer stärkeren Dominanz Googles unter den Suchmaschinen folgte vermehrt die Bitte von Forschern und Entwicklern, den Google-Index auch per Software, d.h. in automatisierter Form, abfragen zu können. Während diese Art der Abfrage bei anderen Suchmaschinen i.d.R. erlaubt war und für die Durchführung von Untersuchungen viel genutzt wurde, untersagte Google diese von Anfang an in den Nutzungsbedingungen[5]. Erst im Frühjahr 2002 entschied Google automatisierte Abfragen zuzulassen und veröffentlichte hierfür die entsprechenden Schnittstellen (Google Web APIs). Die APIs (Application Programming Interfaces) sind als Web Service[6] implementiert, wobei der von Google angebotene Service verschiedene SOAP (Simple Object Access Protocol) Methoden unterstützt, die in einer WSDL (Web Services Description Language) Datei beschrieben werden.

Die Funktionsweise und wesentlichen Elemente des Google Service und damit der Google Web APIs werden nachfolgend beispielhaft für eine Anfrage dargestellt.

Um den Google Service mit eigenen Anwendungen abfragen zu können, sind zuerst zwei Voraussetzungen zu erfüllen:

1. ist ein Schlüssel (googlekey[7]) erforderlich,
2. muss die verwendete Programmiersprache die vom Service offerierten Schnittstellen unterstützen.[8]

Die Möglichkeiten des Google-Service sind in der WSDL-Datei GoogleSearch.wsdl genau beschrieben (siehe Listing 1 unten).[9] Hier ist definiert, wo der Service sich befindet, was vom Entwickler implementiert werden kann (Methoden, Variablen usw.), und wie, d.h. über welches Internetprotokoll und welche Ports mit dem Service kommuniziert wird. So ist z.B. für die Suche in der GoogleSearch.wsdl-Datei genau festgelegt, welche Variablen eine Anfrage enthalten muss, von welchem Typ die Variablen sind und in welcher Reihenfolge diese stehen müssen (Abbildung 2 zeigt die Verarbeitung der Elemente in einer SOAP-Instanz):

---

[5] Terms of Service von Google.com, siehe Punkt 'No automated Querying' http://www.google.com/terms_of_service.html.

[6] Eine gute Einführung zu Web Services findet sich unter: http://webservices.xml.com/lpt/a/ws/2001/04/04/webservices/index.html.

[7] Ein Googlekey ist nach Registrierung unter http://www.google.com/apis erhältlich.

[8] Web Services werden von allen modernen Programmiersprachen unterstützt, darunter Java, .NET (VBA, C#), C++, Python, PHP usw. Wir haben für die Implementierung unserer Abfragen die Programmiersprache Perl und das von Paul Kulchenko entwickelte Perl-Modul SOAP::Lite benutzt.

[9] Die WSDL-Datei selbst ist im XML-Format (genauer dem DTD-Nachfolger XML-Schema).

```
<message name="doGoogleSearch">
  <part name="key"            type="xsd:string"/>
  <part name="q"              type="xsd:string"/>
  <part name="start"          type="xsd:int"/>
  <part name="maxResults"     type="xsd:int"/>
  <part name="filter"         type="xsd:boolean"/>
  <part name="restrict"       type="xsd:string"/>
  <part name="safeSearch"     type="xsd:boolean"/>
  <part name="lr"             type="xsd:string"/>
  <part name="ie"             type="xsd:string"/>
  <part name="oe"             type="xsd:string"/>
</message>
```

Listing 1: WDSL-Datei GoogleSearch.wsdl (Zeile 103-114)

Das SOAP-Protokoll (Simple Object Access Protocol)[10] dient dann zur Übertragung der in XML-kodierten Daten (Nachrichten). Eine solche Nachricht besteht aus einem SOAP-Envelope (Briefumschlag, siehe Abbildung 1), indem die eigentlichen Nachrichten (SOAP-Body) enthalten sind. Vereinfacht ausgedrückt ist SOAP ein spezieller XML-Dialekt für die Übertragung von XML-Nachrichten über ein standardisiertes Internetprotokoll. Schematisch dargestellt sieht eine SOAP-Nachricht folgendermaßen aus.

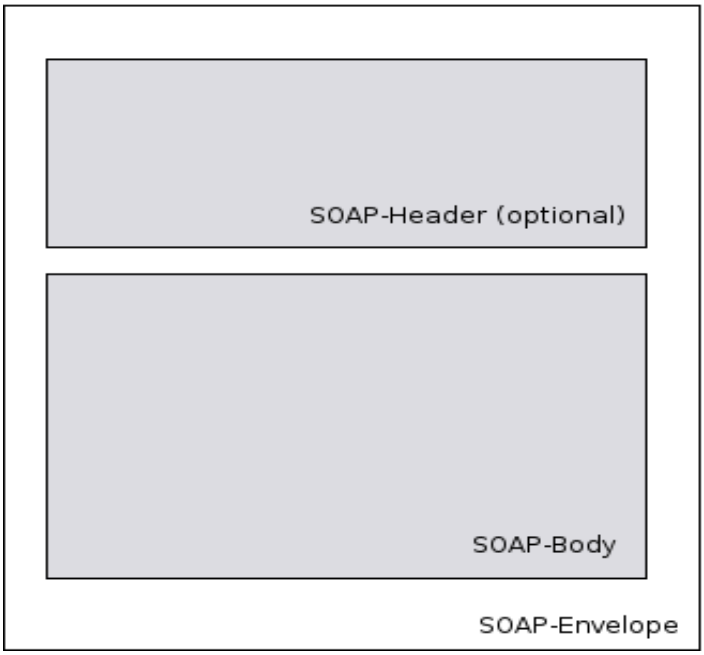

Abbildung 1: SOAP-Envelope

Eine Anfrage an den Google-Service kann in diese schematische Darstellung zur Verdeutlichung übertragen werden (siehe Abbildung 2). Dabei ist gut zu erkennen, wie die SOAP-Instanz die WSDL-Datei verarbeitet, d.h. den Variablen der WSDL-Datei wurden Werte zugewiesen. Diese SOAP-

[10] SOAP wurde von Microsoft, IBM, Sun u.a. entwickelt. Die Weiterentwicklung wurde im Jahr 2000 der XML Protocol Working Group des W3C http://www.w3.org/2000/xp/Group/ übertragen. Zur Entwicklung von SOAP: http://webservices.xml.com/pub/a/ws/2001/04/04/soap.html.

Nachricht wird nun an den Google-Service geschickt. Ganz ähnlich sehen die vom Google-Service gelieferten Nachrichten aus, die wieder von der SOAP-fähigen Anwendung des Klienten weiterverarbeitet werden können. Die SOAP-Methoden und WSDL-Datei sind die wesentlichen Bestandteile der Google Web APIs. Durch die Implementierung in XML wird die Interoperabilität erleichtert und damit die Zusammenarbeit verschiedener Plattformen und Programmiersprachen ermöglicht. Ein entscheidender Vorteil von Web Services im Allgemeinen und dem Google Service im Besonderen.

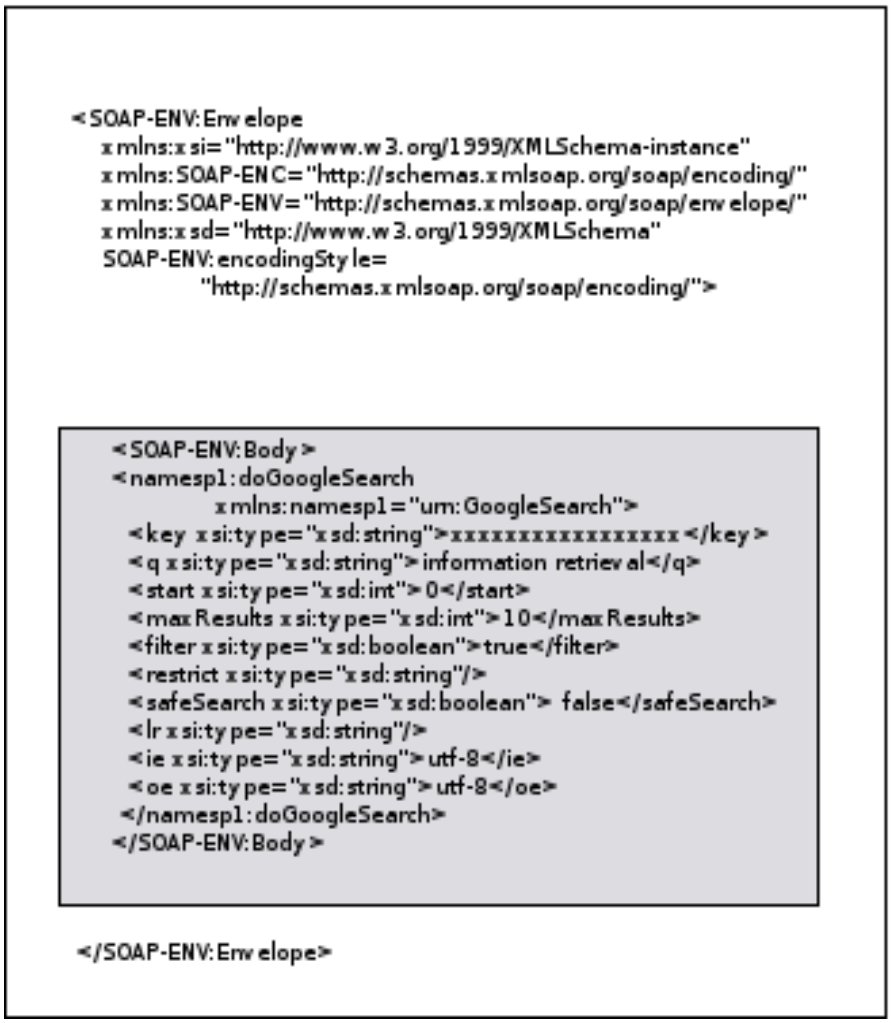

Abbildung 2: SOAP-Envelope für Google Service (Anfrage: 'information retrieval')

Ein weiterer Vorteil der Google Web APIs ist, dass Internetforscher mit dem gut dokumentierten Google Service eigene Programme entwickeln können, die automatisierte Anfragen an die Google Suchmaschine stellen. Die zurückgelieferten Ergebnisse können dann vom Programm weiterbearbeitet werden, um z.B. eigene Analysen zu implementieren[11].

Nachteilig an der Google API ist, dass ein einzelner googlekey maximal 10.000 Treffer pro Tag liefert. Google gibt zwar an, auf Anfrage auch größere Treffermengen bereitstellen zu können, für normale Untersuchungen gilt aber diese Trefferbegrenzung. Ein weiterer Nachteil, der für kleinere Studien aber nicht weiter problematisch ist, ist die Tatsache, dass die Schnittstelle Google API eine geringere Performance aufweist als die optimierte Standardsuche bei Google. Treffer über die Google APIs werden spürbar langsamer zurückgeliefert als Treffer über die hochperformante Standardsuche von Google. Desweiteren existieren, auch in Webforen gelegentlich erwähnte, sichtbare Unterschiede zwischen den Trefferergebnissen der Google API und der Standard Weboberfläche Google.com. Dieser Beobachtung gehen wir unter anderem im folgenden Untersuchungsteil nach.

[11] Siehe dazu unsere Beispielsanwendungen bzw. -analysen auf folgender Testseite http://bsd119.ib.hu-berlin.de/~ft/.

3 Untersuchung

Die Idee für den Report geht auf einen Vortrag im Juli 2004 auf der „Langen Nacht der Wissenschaften“ der Humboldt Universität zu Berlin zurück [13]. Das Ziel des Vortrags und jetzt dieses Reports ist es, praktische Aussagen über die Leistungsfähigkeit und Benutzbarkeit der Google Web API’s (Google API) treffen zu können. Beispielsanwendungen stehen unter http://bsd119.ib.hu-berlin.de/~ft/ zum Test zur Verfügung.

Die Analyse von Trefferergebnissen großer Internet-Suchmaschinen gilt seit einigen Jahren als ein Instrument der *Webometrie*[12] bzw. *Cybermetrie*[13]. Verschiedene webometrische Untersuchungen haben Aufmerksamkeit in der Fachwelt hervorgerufen [1, 2, 3, 14, 15]. Die drei folgenden Kurzuntersuchungen beziehen sich allesamt auf webometrische Vorgängeruntersuchungen [15, 12, 2] und versuchen Konzepte der Veröffentlichungen über die Google API nachzustellen.

Ziel ist es, anhand empirischer Daten die Einsatzmöglichkeiten aber auch Einschränkungen der Google API aufzuzeigen.

1. Die Zeitreihenuntersuchung soll zeigen, inwieweit sich die Google-Treffermengen der beiden Google-Schnittstellen (Google API und Google Web) über einen längeren Zeitraum entwickeln [siehe dazu auch 16] und ob es gravierende Unterschiede in der Zusammensetzung der Trefferlisten gibt. Ziel dieser Untersuchung ist es, zu beurteilen, ob die Google API vergleichbare Ergebnisse liefert, wie die Standard-Schnittstelle Google.com.
2. Die Domainname-Analyse der Top Level Domains (TLD) aus den Treffer-URLs ist eine Implementation einer webometrischen Fragestellung [siehe dazu auch 15] mit Hilfe der Google API. Die Beispielanwendung zeigt die Verteilung der TLDs innerhalb der zurückgelieferten Treffermengen. Die Domainname- und Dateiformat-Analysen können als Implementationstests der Google API angesehen werden. Ziel ist die Überprüfung der Implementierbarkeit wissenschaftlicher Fragestellungen mit den Google Web APIs.
3. Die Dateiformat-Analyse ist eine weitere Google-API-Implementation, die demonstrieren soll, dass die Google API interessante Anwendungsmöglichkeiten für spezifische Internetanalysen bietet, in diesem Fall maschinell erschlossene Dateiformate im Internet [siehe dazu auch 12]. Die Beispielsanwendung zeigt die Häufigkeit bestimmter Dateiformate innerhalb der zurückgelieferten Treffermengen.

Die Untersuchungen wurden auf einem Rechner am Institut für Bibliothekswissenschaft der Humboldt-Universität zu Berlin durchgeführt. Insgesamt wurden im Untersuchungszeitraum vom 30. Juli bis 18. Oktober 2004 knapp 2.500 Trefferlisten mit etwa 250.000 URLs gesichert und analysiert, das entspricht einem Datenvolumen von etwa 144 Megabyte[14]. Die Zeitreihenuntersuchung läuft weiterhin und soll dauerhaft Ergebnisse generieren.

---

[12] Thelwall et al. (2003) definieren Webometrie “… to be the quantitative study of web-related phenomena“ [18]. Die Webometrie bzw. Webometrics ist demnach ein Anwendungsgebiet der Informetrie.

[13] Vgl. Scope des gleichnamigen E-Journals Cybermetrics “ the study of the quantitative analysis of scholarly and scientific communications in the Internet” siehe http://www.cindoc.csic.es/cybermetrics/cybermetrics.html.

[14] Für die Auswertung der knapp 2.500 Dateien im Textformat wurden Programme wie GNU Awk, Perl und Bash-Skripte eingesetzt. Auch diese Programme sind im CVS–Repository http://bsd119.ib.hu-berlin.de/cgi-bin/cvsweb.cgi zu finden. Die Programme unterstützen die Verarbeitung von Textdateien mit Hilfe von regulären Ausdrücken. Hier zeigte sich auch, dass es wesentlich einfacher war, die vom API erzeugten Textdokumente auszuwerten, als die XHTML-Dateien, die Google sonst liefert. Die XHTML-Dateien müssen für bestimmte Auswertungen in einem ersten Schritt geparst werden.

3.1 Zeitreihe

Ziel der Zeitreihenuntersuchung ist es, anhand eines festen Sets von einfachen und erweiterten Suchanfragen Unterschiede in der Zusammensetzung und Entwicklung der Treffermengen für die beiden unterschiedlichen Anfragetypen (Google API, Google Web[15]) zu identifizieren. Zu diesem Zweck haben wir fünf verschiedene Queries definiert und diese Anfragen über beide Schnittstellen täglich abgefragt. Um Schwankungen der Treffermenge während des Tages messen zu können, haben wir die fünf Queries drei mal am Tag abgefragt. Die Abfrage und Sicherung der Suchanfragen und zurückgelieferten Trefferlisten (jeweils 100 Treffer) wird durch ein zeitgesteuertes Script realisiert. Folgende Queries wurden täglich um 6.00 Uhr, 14.00 Uhr und 22.00 Uhr über die Google API und Google Web abgefragt:

1. `webometrics`
2. `bibliometrics or informetrics or scientometrics`
3. `library science`
4. `studieren`
5. `lernen`

Die ersten beiden Suchanfragen `webometrics` und `bibliometrics or informetrics or scientometrics` sind englische Fachbegriffe aus dem Umfeld der Informetrie. Suchanfrage 2 ist heute eher ungewöhnlich (OR-Verknüpfung von drei Fachbegriffen) und geht auf eine Untersuchung der Suchmaschine AltaVista von *R. Rousseau* aus dem Jahr 1997 zurück [15]. `library science` stellt eine typische Google-Suchanfrage dar, die aus zwei AND verknüpften Begriffen besteht[16]. Die Suchanfragen `studieren` und `lernen` sind einfache und sehr allgemeine Suchanfragen mit einem Suchbegriff in deutscher Sprache[17].

Die Trefferlisten der fünf Suchanfragen über die beiden Anfragetypen Google API und Google Web wurden über den gesamten Untersuchungszeitraum gespeichert, um sie anschließend bearbeiten zu können. In einem zweiten Schritt wurden zur Bestimmung der Treffermenge zum Anfragezeitpunkt die ungefähre Angabe der Gesamttreffer aus der Trefferliste extrahiert. Täglich entstanden somit 15 Trefferlisten, jede mit 100 Treffern bzw. URLs.

3.2 Domainname-Analyse

Im Anschluss an die Zeitreihenuntersuchung wurden die Top Level Domain (TLD) Angaben aus den gesicherten Trefferlisten der Zeitreihe extrahiert und aggregiert. Die aggregierten Daten der Zeitreihe wurden daraufhin auf das Vorkommen der Lotka-Funktion[18] hin untersucht. Die Analyse der Top Level Domain Häufigkeiten in Suchmaschinen-Trefferlisten lehnt sich thematisch an die Untersuchung von *Rousseau* an [15]. *Rousseau* konnte 1997 in der Verteilung der TLDs erstmals die Lotka-Funktion nachweisen. Zur Analyse der Daten wurde das Programm LOTKA verwendet, das *Rousseau &*

---

[15] Die automatisierte Abfrage der Webschnittstelle Google Web wurde über Wget realisiert. Die über Wget zurückgelieferten Treffer von Google.com entsprechen den Treffern, die Standardbrowser erhalten, wenn sie Google über das Suchformular befragen.

[16] `library science` wird von Google als `library AND science` interpretiert. Es wurde bewusst darauf verzichtet die Suchanfrage als Phrase zu definieren.

[17] Vgl. dazu die Untersuchung von *Rousseau* 1998/9 in Cybermetrics [16]. *Rousseau* hat in dieser Untersuchung ebenfalls Einwortanfragen (saxophone*, trumpet*, pope) über 12 Wochen an die Internet Suchmaschinen Altavista und Northerlight gestellt und ausgewertet.

[18] „Lotka's Law ist ein 1926 von Alfred James Lotka festgestelltes Skalengesetz bei der Beziehung zwischen der Anzahl von Publikationen einer Person und der Anzahl von Personen mit einem eben so hohen Publikationsausstoß. Es wurde für die Anzahl der wissenschaftlichen Zeitschriftenartikel aufgestellt und besagt, dass die Anzahl der Personen, die n Artitkel schreiben, proportional zu 1/n2 ist …“ aus Wikipedia siehe http://de.wikipedia.org/wiki/Lotka%27s_Law
siehe Primärquelle Alfred J. Lotka: The frequency distribution of scientific productivity. In: Journal of the Washington Academy of Science (16) 1926.

Rousseau entwickelt haben [14]. Die aggregierten Daten aus den Trefferlisten (Google API und Google Web) der Zeitreihenuntersuchung wurden in das Programm importiert und analysiert. Ergebnis der Analyse mit LOTKA ist die Aussage, ob die Daten der Lotka-Funktion folgen oder nicht. Die Analyse der TLDs nach Lotka soll zeigen, ob die beiden unterschiedlichen Treffermengen Google API und Google Web gleichen Gesetzmäßigkeiten folgen.

3.3 Dateiformat-Analyse

Die Idee für die folgende Analyse geht auf die Untersuchung „Das Dateiformat PDF im Web“ zurück [12]. Spätestens seit Google im Jahr 2001 alternative Dateiformate wie das Portable Document Format (PDF) und andere indexierbare Formate (PostScript, Powerpoint, Word, …) erschließt und in Trefferlisten nachweist[19], hat sich die Zusammensetzung der Trefferlisten deutlich erweitert. Da diese alternativen Dateiformate im Web, insbesondere der Anteil der PDF-Dokumente, einen beachtlichen Anteil der Dokumente im Web ausmachen, sehen wir hier weiteren Forschungsbedarf [vgl. dazu auch 12], der über die Google APIs abzubilden wäre. Die Analyse von Trefferlisten aus dem Jahr 2002 konnte beispielsweise einen Zusammenhang zwischen Länge bzw. Sprache einer Suchanfrage und dem Anteil des Dateiformats PDF feststellen. Längere Anfragen mit mehreren kombinierten Suchbegriffen haben laut dieser Studie einen deutlich größeren Anteil an PDF-Treffern, als Suchanfragen mit nur einem Suchbegriff. Ziel der aktuellen Untersuchung ist es, die Dateiformat-Analysen über die Funktionalität der API abzubilden und weiter zu automatisieren. Das hierzu entwickelte Programm konzentriert sich auf die Analyse der Dateiformatangaben in den URLs der Trefferlisten.

Die Queries, die wir an die Google API gestellt haben, bestanden aus zwei Typen von Anfragen:

1. Einfache Suchbegriffe: das sind Queries, die nur einen Suchbegriff oder maximal zwei Begriffe enthalten (z.B. `sozialwissenschaft`).
2. Einfache Suchbegriffe AND Relator[20]: das sind Suchanfragen, die einen Suchbegriff enthalten und mit einem Relator verknüpfen werden (z.B. `sozialwissenschaft AND konferenz`).

Die verwendeten Suchbegriffe sind Deskriptoren der Library of Congress Classification und kommen in unserem Beispiel aus der Hauptklasse Social Sciences [11]. Alle Suchbegriffe wurden zufällig ausgewählt und in den Sprachen Deutsch[21] und Englisch an die Google API gestellt. Es wurden pro Suchanfrage immer 100 API-Treffer gespeichert.

[19] Die anderen großen Suchmaschinenbetreiber haben sich diesem Trend angepasst und erschließen inzwischen ebenfalls alternative Dateiformate (vgl. Lewandowski, 2004) [10].

[20] Relatoren sind allgemeine Begriffe, die eine sinnvolle Beziehung zu einem anderen Begriff bilden können. Wir bezeichnen in diesem Zusammenhang Begriffe wie Bericht, Projekt, Publikation, Magisterarbeit, usw. als Relatoren für Suchanfragen im Internet.

[21] Die deutschen Begriffe sind Übersetzungen der englischen Termini aus der Library of Congress Classification.

4 Ergebnisse

Nachfolgend werden die Ergebnisse der drei Kurzuntersuchungen anhand einzelner Beispiele knapp dargestellt.

4.1 Zeitreihenergebnisse

1. Die Daten der Zeitreihenuntersuchung zeigen, dass die Standardschnittstelle Google.com insgesamt rund 2,4 mal mehr Treffer liefert als die API Schnittstelle. Der Index der Google APIs ist folglich nur ein Teil des Google Gesamtindex (vgl. Spalte 2 und 3 in Tabelle 1).
2. Alle fünf Anfragen entwickeln sich über den Zeitraum tendenziell ähnlich (siehe Abbildungen 3, 4, 5 und 6). Die Gesamttrefferangaben der Google APIs und Web korrelieren für die Suchanfragen `webometrics` und `bibliometrics OR informetrics OR scientometrics` sehr gut (siehe Pearson Korrelationskoeffizient). Für die Anfragen `library science`, `lernen` und `studieren`, die über den Zeitraum deutlich mehr Treffer liefern, korrelieren die Daten der beiden Suchschnittstellen sichtlich schlechter. D.h., dass die Treffermengen für Anfragen mit sehr großen Trefferresultaten deutlich stärker schwanken (siehe insb. Zeitreihe in Abb. 5, 6 für `library science`) und damit eignen sich diese Zeitreihen weniger für die Beobachtung von Entwicklungstendenzen.

Query	APIs Ergebnisse median	WEB Ergebnisse median	Bestimmtheits-maß APIs	Bestimmtheits-maß WEB	Pearson Korrelations-koeffizient (APIs/WEB)
`webometrics`	642	1.510	0,59	0,9	0,91933047
`bibliometrics OR informetrics OR scientometrics`	11.900	28.050	0,83	0,9	0,919770563
`library science`	2.770.000	6.750.000	0,07	0,13	0,170647969
`lernen`	3.690.000	8.730.000	0,009	0,13	0,43581289
`studieren`	610.000	1.460.000	0,03	0,0037	0,380264693

Tabelle 1: Ergebnisse der Zeitreihenuntersuchung

3. Tabelle 2 zeigt, dass sich überraschenderweise mehr unterschiedliche URLs in den zurückgelieferten Trefferlisten (100 Treffer je Anfrage) der Google APIs befinden, als in den deutlich größeren Treffermengen von Google Web.

Anzahl der unterschiedlichen URLs	APIs	WEB
`webometrics`	253	200
`bibliometrics OR informetrics …`	214	179
`library science`	220	229
`lernen`	201	191
`studieren`	252	229

Tabelle 2: Anzahl der unterschiedlichen URLs innerhalb der Treffer 1-100 für die beiden Suchschnittstellen

Die folgenden Abbildungen 3, 4, 5 und 6 zeigen den Verlauf der fünf Suchanfragen für die beiden Suchschnittstellen.

4. Die Daten der Google API und Google Web zeigen für die Suchanfrage `webometrics` und alle anderen Anfragen einen relativ ähnlichen Verlauf (siehe Abb. 3 und 4). Der sprunghaften Schwankungen zu Beginn (Google API) und in der Mitte der Zeitreihe, gehen vermutlich auf einen Index-Update bei Google zurück. Am 04.09.2004 um 22.00 Uhr lieferte Google Web 988 Treffer zurück, wenige Stunden später um 6.00 Uhr am 05.09.2004 waren es 1.500. Dieser Ausschlag findet sich vier Tage später am 08.09.2004 in den Daten der APIs Zeitreihe. Die Daten der anderen Anfragen zeigen einerseits zeitlich einen sehr konstanten Verlauf (siehe Abbildung 5, Anfrage `lernen`) aber auch sehr schwankende Verläufe (siehe Abbildung 5, Anfrage `library science`).
5. Die Untersuchung der Treffermengen zu den drei Anfragezeitpunkte 6.00 Uhr, 14.00 Uhr und 22.00 Uhr zeigt, dass die beiden Google-Schnittstellen zu allen drei Zeitpunkten in etwa gleich viel Treffer liefern. Ausreißer sind in den Daten allerdings zu finden.

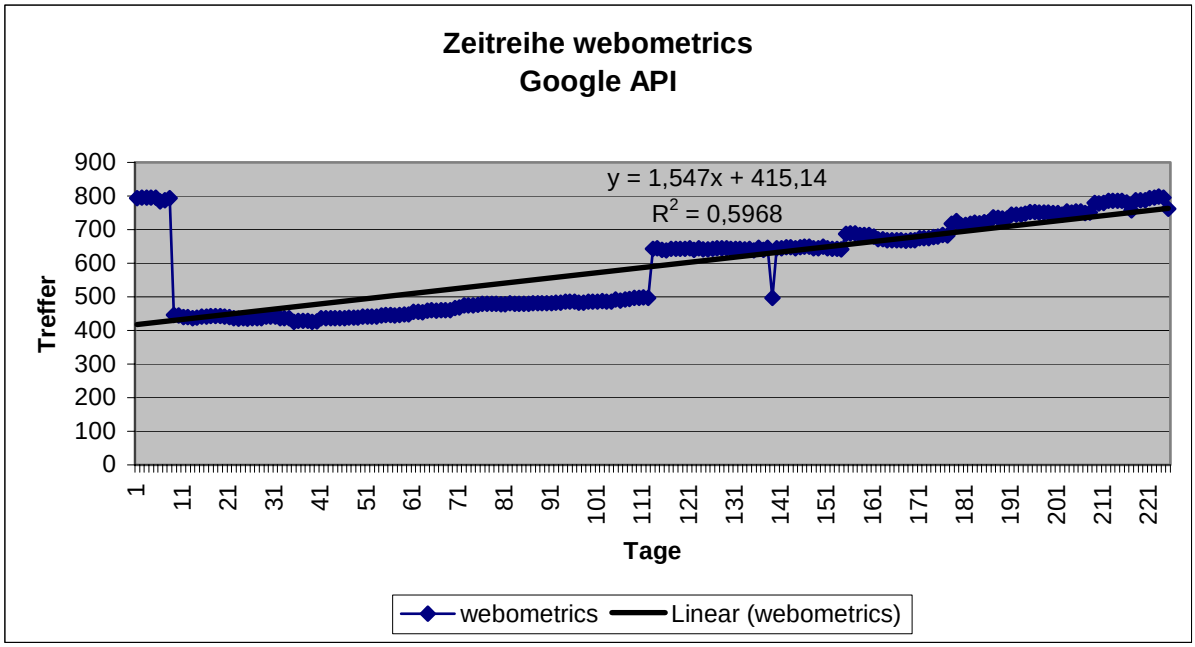

Abbildung 3: Verlauf der Zeitreihe für den Suchbegriff `webometrics` über die Suchschnittstellen Google Web APIs (Google API)

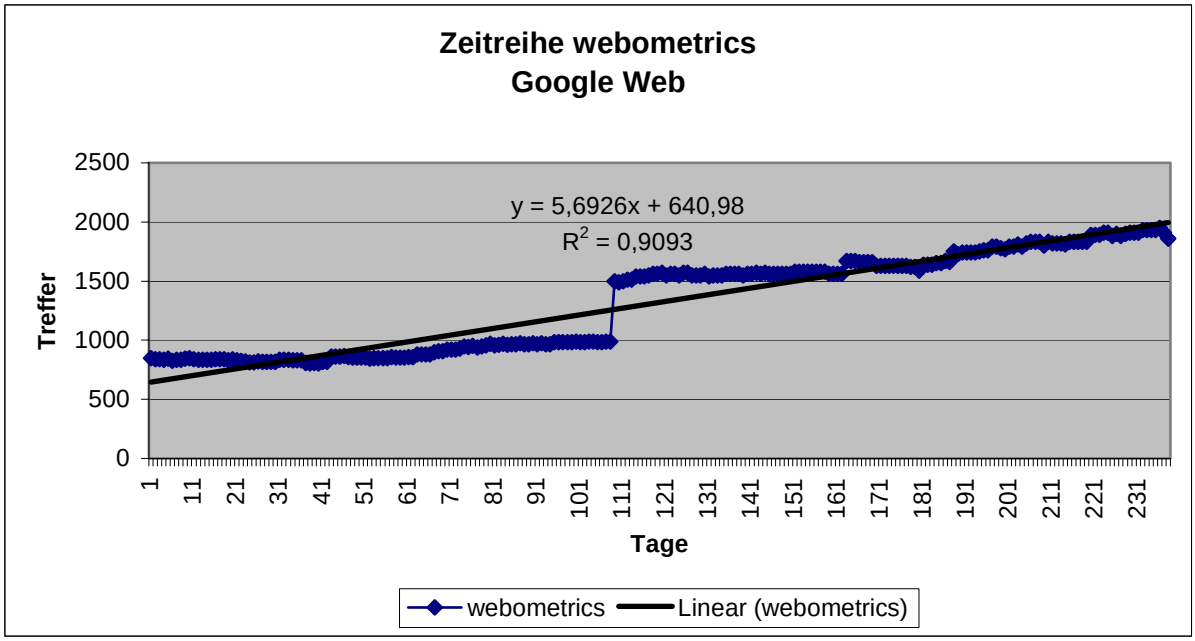

Abbildung 4: Verlauf der Zeitreihe für den Suchbegriff `webometrics` über die Standardoberfläche Google.com (Google Web).

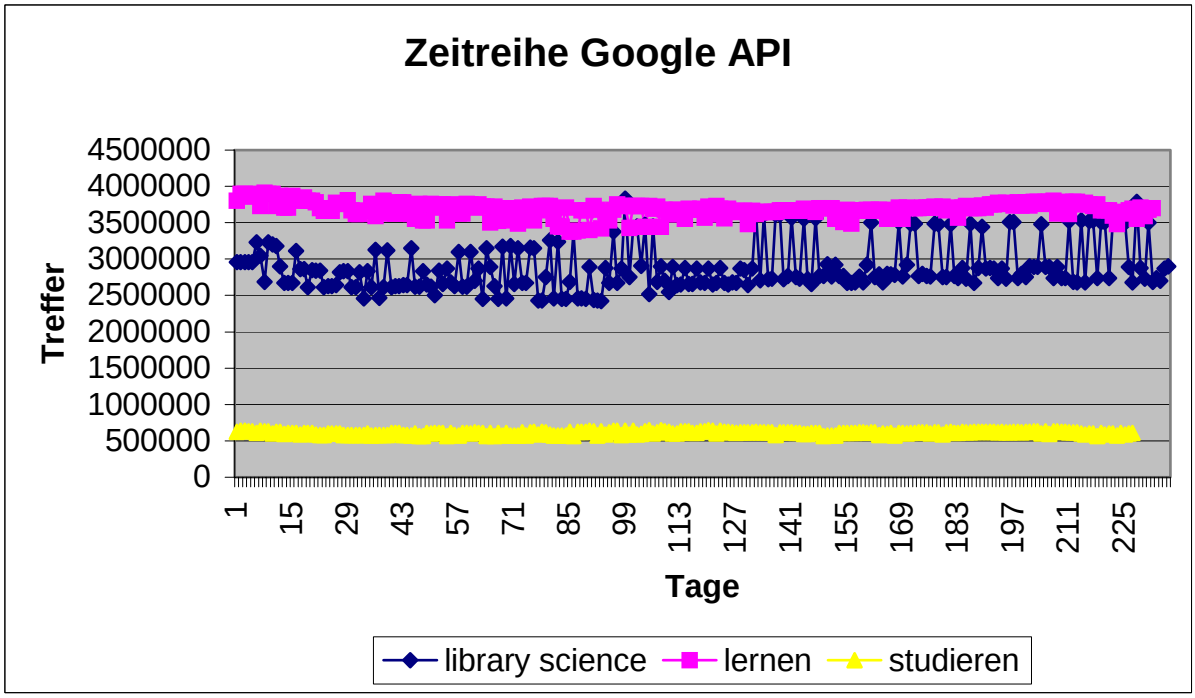

Abbildung 5: Verlauf der Zeitreihe für die Suchbegriff `library science, lernen` und `studieren` über die Suchschnittstellen Google Web APIs (Google API).

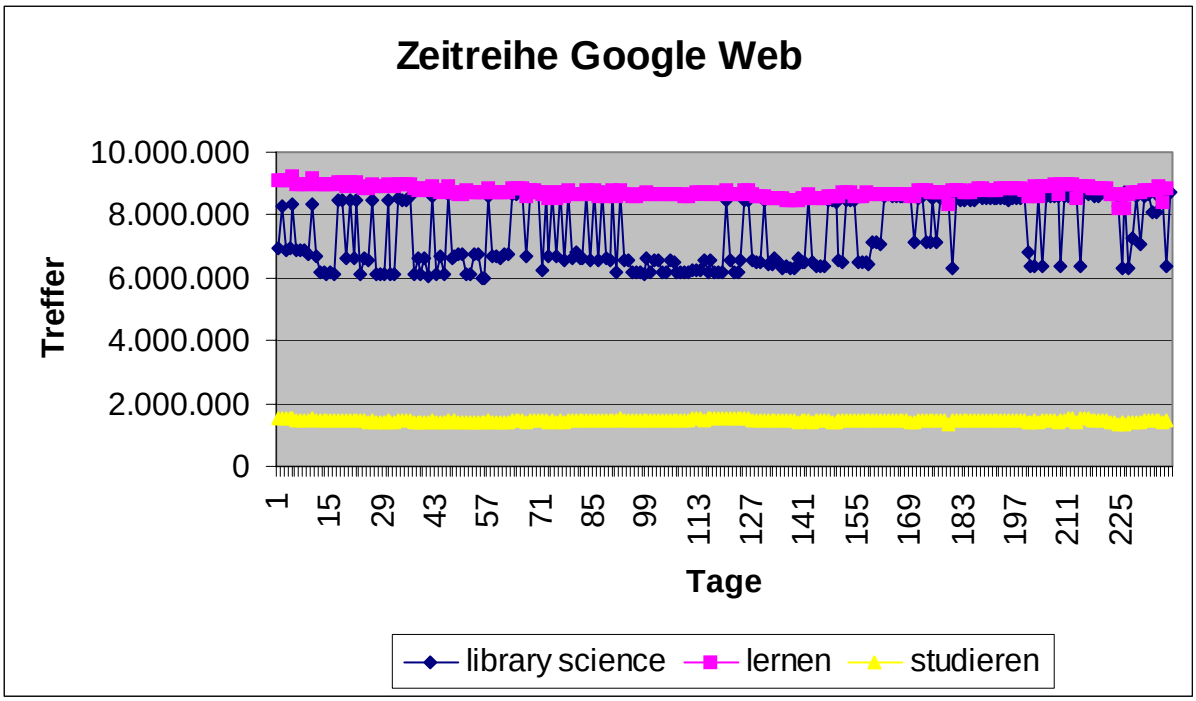

Abbildung 6: Verlauf der Zeitreihe für die Suchbegriff `library science, lernen` und `studieren` über die und Standardoberfläche Google.com (Google Web).

4.2 Domainname-Analyse

Die Domainname-Analyse liefert ein sehr einheitliches Ergebnis für alle Anfragen (Google API und Google Web). Die Analyse der Top Level Domains (TLD) zeigt, dass die Daten der fünf Anfragen nach *Lotka* verteilt sind, bzw. dem *Lotka Law* folgen. Der Beta-Wert der Lotka-Funktion liegt bei allen Analysen zwischen 1,27 und 3,29 (vgl. *Rousseau & Rousseau* [14]). Damit wird die Untersuchung von *Rousseau* aus dem Jahr 1997 [15] eindrucksvoll bestätigt. Abbildung 7 und Tabelle 3 zeigen die Verteilung der TLDs für die Suchanfrage `bibliometrics OR informetrics OR scientometrics` in der Rangdarstellung.

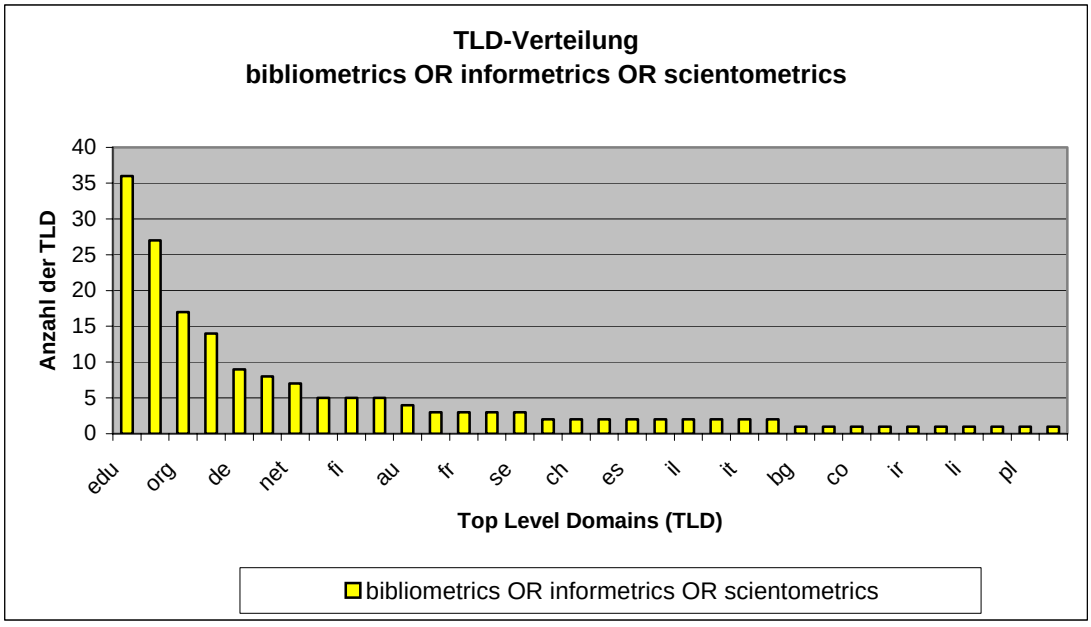

Abbildung 7: Verteilung der Top Level Domains (TLD) für die Anfrage `bibliometrics OR informetrics OR scientometrics`

Anzahl der TLDs	TLDs	Häufigkeit der TLDs
1	edu	36
1	com	27
1	org	17
1	uk	14
1	de	9
1	ca	8
1	net	7
3	be, fi, nl	5
1	au	4
4	cn, fr, in, se	3
9	br, ch, dk, es, hu, il, info, it, ro	2
10	bg, cl, co, gov, ir, jp, li, no, pl, us	1

Tabelle 3: Verteilung der Top Level Domains (TLD) für die Anfrage `bibliometrics OR informetrics OR scientometrics`[22]

Die aggregierten Daten der TLD-Analyse für die Treffer der Google API und Google Web Schnittstelle unterliegen demnach den gleichen bibliometrischen Gesetzmäßigkeiten (Lotka Law). Wenige TLD wie z.B. edu, com oder org kommen sehr häufig in den Trefferdaten vor. Sehr viele TLD wie z.B. br, pl oder li, kommen nur einmal in den Trefferlisten vor. Die Verteilung der TLD hängt selbstverständlich sehr eng mit der Sprache der Anfrage zusammen. Deutschsprachige Suchanfragen (z.B. `lernen` oder `studieren`) liefern demnach hauptsächlich Treffer der TLDs de, at, ch, folgen aber trotzdem der Lotka-Verteilung.

4.3 Dateiformat-Analyse

Die Ergebnisse der Analyse der Dateiformate bestätigen die Ergebnisse der Untersuchung aus dem Jahr 2002 [12]. Folgende Ergebnisse lassen über die untere Stichprobe (siehe Tabelle 4, 5) festhalten.

1. Allgemeine Suchbegriffe mit einem oder maximal zwei Suchbegriffen (siehe z.B. Query 1, 2 in Tabelle 4), liefern innerhalb der ersten 100 Treffer praktisch nur Treffer im HTML-Format.
2. Eine Kombination dieser einfachen Suchbegriffe mit Relatoren (siehe Query 3, 4 in Tabelle 4) liefern deutlich mehr alternative Dateiformate. Das Dateiformat PDF dominiert die anderen Formate zahlenmäßig deutlich.
3. In Tabelle 5 wird deutlich, dass sich der Trend zu mehr Treffern im PDF-Format bei spezifischen Suchbegriffen (Fachbegriffen) noch verstärkt.
4. Es lässt sich weiterhin beobachten, dass Suchbegriffe in deutscher Sprache tendenziell mehr Treffer im PDF-Format liefern als Anfragen in englischer Sprache.

Die Ergebnisse der Dateiformat-Analyse nehmen selbstverständlich nicht in Anspruch repräsentative Aussagen für das gesamte Web treffen zu können. Auch die Untersuchung aus dem Jahr 2002 wurde nicht repräsentativ sondern ebenfalls explorativ konzipiert und durchgeführt. Es sollten hier lediglich die Ergebnisse der Vorläuferuntersuchung beispielhaft durch die Google API nachgestellt werden.

[22] Die Verteilung der TLDs zur Anfrage `bibliometrics OR informetrics OR scientometrics` folgt der Lotka-Funktion. Vgl. [14]. Der Beta-Wert der Funktion liegt bei 1,6076.

Format	Query 1	Query 2	Query 3	Query 4
	`sozialwissenschaft`	`"social science"`	`sozialwissenschaft + konferenz`	`"social science" + conference`
HTML	99	100	82	89
PDF	1	0	16	9
DOC	0	0	1	2
PPT	0	0	0	0
TXT	0	0	0	0
PS	0	0	0	0
XLS	0	0	0	0

Tabelle 4: allgemeine Suchbegriffe in Deutsch und Englisch, sowie die Häufigkeit der Dateiformate zu einzelnen Queries.

Format	Query 5	Query 6	Query 7	Query 8
	`wohlfahrtstheorie`	`"welfare theory"`	`wohlfahrtstheorie + projekt`	`"welfare theory" + project`
HTML	58	78	33	77
PDF	39	18	54	21
DOC	2	4	2	1
PPT	0	0	0	1
TXT	1	0	1	0
PS	0	0	1	0
XLS	0	0	0	0

Tabelle 5: spezifische Suchbegriffe (Fachbegriffe) in deutsch und englisch, sowie die Häufigkeit der Dateiformate zu einzelnen Queries.

5 Diskussion und Fazit

Zum Ende dieses Reports wollen wir die wichtigsten Ergebnisse unserer Untersuchungen und Erfahrungen mit den Google Web APIs herausstellen und kurz diskutieren.

1. Die Ergebnisse der einzelnen Zeitreihenuntersuchungen zeigen deutlich, dass die Trefferdaten der beiden Google-Schnittstellen quantitativ sehr ähnlich, aber auf unterschiedlichen Niveaus verlaufen. Starke Schwankungen in den API-Trefferdaten finden sich auch in den entsprechenden Google Web-Treffern. Damit wird deutlich, dass die beiden Treffermengen eng miteinander verbunden sind. Es kann aber nicht davon ausgegangen werden, dass Anfragen an die Google API direkt an die aktuellste Version des Google Gesamtindex (Google Web) weitergegeben werden. Die Trefferdaten der Google API liefern quantitativ gesehen, nur etwas über 40% der Google Web-Treffer, was daraufhin deutet, dass im Falle der Google API auf einen kleineren und weniger aktuellen Stand des Google-Indexes zugegriffen wird. Inwieweit sich die Treffermengen der beiden Schnittstellen bzgl. des Trefferrankings unterscheiden, wurde in dieser Untersuchung nicht untersucht. Stichproben zeigen aber, dass bzgl. des Rankings erhebliche Unterschiede auftreten können. Zusammenfassend kann man sagen, dass die beiden Suchschnittstellen von Google sich zwar in großen Teilen überlappen, von einer identischen Treffermenge aber nicht ausgegangen werden kann. Für zukünftige Untersuchungen heißt das, dass Auswertungen von Treffern der Google API nicht 1 zu 1 auf die Standard-Schnittstelle von Google Web zu übertragen sind. Ein Prototyping webometrischer Analysen auf Basis der Google API, kann nach unseren Erfahrungen erfolgreich durchgeführt werden.
2. Zur Zuverlässigkeit der Google API: Wir mussten feststellen, dass der Service nicht immer in gleicher Weise funktionierte. Damit unterscheidet sich dieser Service sehr deutlich von der Standard Google-Suche, die eigentlich immer hochperformant funktioniert. Besonders auf viele Anfragen, die durch eine Schleife im Programm ausgelöst wurden, reagierte der Service mit unterschiedlichen Antwortzeiten, manchmal auch gar nicht. Letzteres Verhalten haben wir insbesondere beobachtet, wenn mehr als 500 Resultate vom Programm angefordert wurden. Ein weiterer Punkt neben der Beschränkung auf 10.000 Anfragen pro Tag, der verdeutlicht, dass der Service in der aktuellen Version sich nicht für Untersuchungen eignet, die sehr große Datenmengen erfordern. Ein Performance-Vergleich der beiden Google-Schnittstellen ist aus diesen Gründen auch nicht sinnvoll. Von einer Beta-Version kann man sicher nicht die gleiche Verfügbarkeit erwarten wie von der hochoptimierten Google.com Suche.
3. Erfreulicherweise konnten alle Ergebnisse der Vorgängeruntersuchungen zumindest in Stichproben bestätigt werden. Sowohl die Verteilung der TLDs nach Lotka Law als auch der hohe Anteil der PDF-Treffer in der Dateiformat-Analyse konnte in den Treffern der Google API nachgewiesen werden. Damit können die Google Web APIs unseres Erachtens zur Datengenerierung in wissenschaftlichen Internet-Studien durchaus eingesetzt werden.

   Es wäre weiterhin auch interessant zu untersuchen, inwieweit sich die Menge der verschiedenen Dateiformate innerhalb bestimmter Fachdisziplinen darstellt. Eine solche Untersuchung könnte relativ einfach mit den Google Web APIs durchgeführt werden, da hier verschiedene Relatoren mit bestimmten Fachfragen kombiniert werden können.

Der 2002 veröffentlichte Web Service Google Web APIs, der auch heute seinen Beta Status[23] nicht verloren hat, stellt unserer Ansicht nach ein sehr interessantes Experimentierfeld für webometrische Untersuchungen dar. Da die Hürden für den Einsatz der Google API relativ gering sind und die Implementation eigener Analysen mit fortgeschrittenen Programmierkenntnissen relativ schnell

[23] Der Google Service Google Web APIs befindet sich nach wie vor in einer Testphase (Beta-Phase). Die letzte Aktualisierung wurde am 30.08.2002 vorgenommen (vgl. http://www.google.com/apis/release_notes.html). Auf diesen Umstand weist Google auch immer wieder bei evtl. Problemen hin. Für Fragen und Probleme wurde ein eigenes Forum eingerichtet, das sehr aktiv ist (siehe http://groups.google.com/groups?group=google.public.web-apis).

vonstatten geht, können wir die Google Web APIs zur Webdatengenerierung und -Weiterbearbeitung empfehlen. Trotz einzelner Einschränkungen und Probleme überwiegen die positiven Aspekte der inzwischen in die Jahre gekommenen Beta-Version.

6 Literatur

1. Almind, T. C.; Ingwersen, P. (1997): “Informetric analyses on the world wide web: methodological approaches to ‘webometrics’”. In: Journal of Documentation 53,S. 404-426
2. Bar-Ilan, J. (1998/9): “Search Engine Results over Time – A Case Study on Search Engine Stability”. In: Cybermetrics 2/3(1)
   available http://www.cindoc.csic.es/cybermetrics/articles/v2i1p1.html [31 October 2004]
3. Bar-Ilan, J. (2002): “Methods for Measuring Search Engine Performance over Time”. In: Journal of the American Society for Information Science and Technology 53(4),S. 308-319
4. Calishain, T.; Dornfest, R. (2003): *Google hacks 100 industrial-strength tips and tools.* Sebastopol, CA: O'Reilly, 330 Seiten.
5. Google Web APIs (Home page)
   available http://www.google.com/apis/ [31 October 2004]
6. Griesbaum, J. (2004): “Evaluation of three German search engines: Altavista.de, Google.de and Lycos.de”. In: Information Research 9(4).
   available http://InformationR.net/ir/9-4/paper189.html [31 October 2004]
7. Lawrence, S.; Giles, C. (1998): “Searching the World Wide Web”. In: Science 280, S. 98-100.
8. Lawrence, S.; Giles C. (1999): “Accessibility of information on the web”. In: Nature 400, S.107-109.
9. Levandowski, D. (2003): „Suchmaschinen-Update : Markttrends und Entwicklungsperspektiven bei WWW-Universalsuchmaschinen“. In Schmidt, R., Eds.: Proceedings Competence in Content, S. 25-35.
   available http://eprints.rclis.org/archive/00001818/ [31 October 2004]
10. Lewandowski, D. (2004): „Abfragesprachen und erweiterte Funktionen von WWW-Suchmaschinen“. In: IWP Information Wissenschaft und Praxis 55(2), S. 97-102.
    available http://eprints.rclis.org/archive/00001557/ [31 October 2004]
11. Library of Congress Classification (Home page)
    available http://www.loc.gov/cds/ [31 October 2004]
12. Mayr, P. (2002): „Das Dateiformat PDF im Web – eine statistische Erhebung“. In: IWP Information Wissenschaft und Praxis. 53(8),S. 475-481.
    available http://www.ib.hu-berlin.de/~mayr/arbeit/nfd_PDF_im_Web.pdf [31 October 2004]
13. Mayr, P.; Tosques, F. (2004): Informationsrecherche im Internet mit Hilfe der Google Web API. Vortrag am 12. Juni 2004 auf der Langen Nacht der Wissenschaften, Institut für Bibliothekswissenschaft der Humboldt-Universität zu Berlin.
    available http://www.ib.hu-berlin.de/~mayr/arbeit/langenacht_04.pdf [31 October 2004]
14. Rousseau, B.; Rousseau, R. (2000): “LOTKA: A program to fit a power law distribution to observed frequency data”. In: Cybermetrics 4(1).
    available http://www.cindoc.csic.es/cybermetrics/articles/v4i1p4.html [31 October 2004]
15. Rousseau, R. (1997): “Sitations: an exploratory study”. In: Cybermetrics 1(1).
    available http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html [31 October 2004]
16. Rousseau, R. (1998/9): “Daily time series of common single word searches in AltaVista and NorthernLight”. In: Cybermetrics 2/3(1).
    available http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html [31 October 2004]
17. Thelwall, M. (2004): “Can the Web give useful information about commercial uses of scientific research?”. In: Online Information Review 28(2), S. 120-130.
18. Thelwall, M.; Vaughan, L.; Björneborn, L. (2003): “Webometrics”. In: ARIST 39. preprint.available http://www.db.dk/lb/2003preprint_ARIST.doc [31 October 2004]

7 Anhang

Implementationsbeispiel

Ein komplettes Beispiel aus dem Buch „Google Hacks“ [4] in der Programmiersprache Perl.

```
#!/usr/local/bin/perl
# usage: perl googly.pl <query>
my $googlekey='xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx';
my $google_wsdl = "./GoogleSearch.wsdl";
use SOAP::Lite;
my $query = shift @ARGV or die "Usage: perl googly.pl <query>\n";
my $google_search = SOAP::Lite->service("file:$google_wsdl");
# query google
my $offset = 0;
my $results = $google_search ->
  doGoogleSearch(
   $googlekey, $query, $offset, 10,
   "false", "", "false", "",
   "utf-8", "utf-8");
# no results?
@{$results->{resultElements}} or exit;
# loop throught the results
foreach my $result (@{$results->{resultElements}}) {
  print
    join "\n",
    ++$offset,
    $result->{title} || "no title",
    $result->{URL},
    $result->{snippet} || "no snippet", "\n";
}
```

Listing 2: Das Programm Googly

Die Programme, die wir für die Abfragen und für die Auswertung der Untersuchungen geschrieben haben, finden sich unter: http://bsd119.ib.hu-berlin.de/cgi-bin/cvsweb.cgi .

Autoren

Philipp Mayr, M.A.: Absolvent des Instituts für Bibliothekswissenschaft der Humboldt-Universität zu Berlin (seit April 2004). Studium der Magisterfachkombination Bibliothekswissenschaft, Informatik und Soziologie. Seit November 2004 wissenschaftlicher Mitarbeiter am Informationszentrum Sozialwissenschaften in Bonn.

Email philippmayr@web.de
Homepage http://www.ib.hu-berlin.de/~mayr/

Fabio Tosques: Student am Institut für Bibliothekswissenschaft der Humboldt-Universität zu Berlin. Studium der Magisterfachkombination Italianistik, Bibliothekswissenschaft und Informatik. Seit 1999 Tutor im Fernstudium Bibliothekswissenschaft.

tosques@informatik.hu-berlin.de