

Das Dateiformat PDF im Web – eine statistische Erhebung

Philipp Mayr, Berlin

Die statistische Erhebung „PDF im Web“ befasst sich mit der quantitativen Evaluation des Dateiformats Portable Document Format (PDF) im World Wide Web¹. Das Hauptinteresse dieser Arbeit besteht darin, ein Verfahren zu erläutern und zu demonstrieren, mit dem eine zahlenmäßige Einschätzung dieses Dateiformats im Internet möglich wird. Hauptaugenmerk der Analyse ist die Errechnung des Anteils der PDF-Dokumente für bestimmte Suchanfragen sowie die durchschnittliche Dokumentgröße der recherchierten PDF-Dateien. Die Erhebung basiert auf Trefferlisten des Suchmaschinenbetreibers Google, die über strukturierte Suchanfragen generiert werden. Das vorgestellte Verfahren wird anhand einer Stichprobe von 50 Anfragen exemplarisch getestet. Im Rahmen der Auswertung werden die Ergebnisse bezüglich der beiden Publikationssprachen Deutsch und Englisch sowie unterschiedlich komplexer Anfragen miteinander verglichen.

PDF on the Web – a statistical approach

The statistical survey „PDF im Web“ deals with the quantitative evaluation of the file format Portable Document Format (PDF) on the World Wide Web. The main interest of this paper is to illustrate and demonstrate a method which allows a numerical assessment of this file format on the internet. Main focus of the analysis is the calculation of the percentage of PDF documents returned in answer to certain queries and the average document size of the queried PDF files. The survey bases on search engine hit lists from Google which were generated by structured queries. The demonstrated method will be tested by a sample of 50 queries. The evaluation compares the findings for both publication languages German and English and for different complex queries.

1 Einleitung

Elektronische Informationen, insbesondere digitale Dokumente, bestimmen zunehmend unsere tägliche Arbeit und stellen für uns aus vielerlei Hinsicht eine große Herausforderung dar. Digitale Dokumente werden dem Informationssuchenden heute in hohem Maß aus dem Internet über automatische Verfahren zugänglich gemacht. Ein beliebter Zugangsweg zu dieser Information sind Suchmaschinen. Durch die Indexierung öffentlicher Webdokumente wird diese Information für jedermann über Suchmaschinen recherchierbar. Die im Internet zugänglichen Dokumente liegen in der

Regel in wenigen unterschiedlichen Dateiformaten vor.

Das ursprüngliche und nach wie vor dominierende Dateiformat für indizierte Information² im Internet ist HTML (Hypertext Markup Language). Im letzten Jahr hat sich das Spektrum der Dateiformate, die über eine Suchmaschinenrecherche zugänglich werden, etwas erweitert. Unter den neuerdings über die Suchmaschine Google recherchierbaren Dateiformaten befindet sich unter anderem das Portable Document Format (PDF).

Dokumente im PDF-Format spielten im Internet als Publikationstyp eine besondere Rolle.

Diese Tatsache nimmt die Arbeit zum Anlass, dieses Dateiformat genauer zu betrachten.

In dieser Untersuchung wird der Versuch unternommen, das Dateiformat PDF in seiner Bedeutung als Informationsquelle im Internet zu analysieren. Zur Evaluation werden Suchergebnisse der Suchmaschine Google nach bestimmten Kriterien untersucht.

Nach einleitenden Anmerkungen zum Dateiformat PDF und zur Suchmaschine Google, widmet sich die Arbeit der Untersuchungsbeschreibung, den Ergebnissen der Untersuchung und einer Diskussion der Ergebnisse.

PDF – ein beliebtes Webformat

Das Portable Document Format (PDF) wurde von Adobe als Erweiterung von Postscript entwickelt und kann als Industriestandard angesehen werden. PDF hat sich in den letzten Jahren unter anderem als weitverbreitetes Dateiformat für die Publikation, Präsentation sowie die Online-Distribution von digitalen Dokumenten im World Wide Web etabliert.

Grund hierfür sind verschiedene Aspekte, die Daniel Ohst folgendermaßen zusammenfasst.

„PDF stellt ein layoutorientiertes Format dar, das in der Lage ist, auflösungs- und geräteunabhängig Dokumente auch mit hohen Anforderungen an die Darstellungsqualität zu speichern. Durch die gute Integrationsmöglichkeit von einigen Strukturierungsmerkmalen und vor allem von Hyperlinks ist dieses Format

¹ Die Idee für diese Erhebung entstand im Sommersemester 2001 im Proseminar „Datenerhebung, Datenstrukturierung und Datenerfassung“ am Institut für Bibliothekswissenschaft der Humboldt-Universität zu Berlin unter Leitung von Herrn Prof. Dr. Umstätter.

² Unter indizierter Information werden hier digitale Dokumente im World Wide Web verstanden, die über Suchmaschinenindizes recherchierbar sind.

³ Vgl. (Ohst 1998)

insbesondere als Präsentationsformat im Internet sehr beliebt.“³

Diese angesprochene Beliebtheit von PDF lässt sich durch weitere Faktoren verdeutlichen und erklärt die Verbreitung dieses Dateiformats im Internet.

PDF ermöglicht unter anderem

- die einfache Verfügbarkeit (frei erhältlicher PDF-Viewer),
- die Identität zwischen der Bildschirmdarstellung und dem Ausdruck eines PDF-Dokuments (Zitierbarkeit)⁴,
- keine nachträgliche Veränderung der PDF-Dokumente.

Zudem überzeugt PDF durch die einfache Erzeugung von PDF-Dokumenten aus nahezu allen üblichen Anwendungsprogrammen über den Acrobat PDFWriter bzw. den Acrobat Distiller. Voraussetzung zur Benutzung dieser PDF-Drucker ist eine kostenpflichtige Softwarelizenz der Adobe Acrobat Familie.

Weitere Features von Adobe Acrobat sind:

- Konvertierung von gescannten Dokumenten zu PDF-Dateien
- Hinzufügen von Anmerkungen in PDF-Dokumenten
- Arbeit mit digitalen Unterschriften

Nach diesem kurzen Überblick über Spezifika des Dateiformats PDF und Gründen zu dessen Beliebtheit, sei das Augenmerk nun auf die Suchmaschine Google gerichtet, die das zentrale Untersuchungsinstrument dieser Arbeit darstellt.

Google – die Suchmaschine

Google wurde von S. Brin und L. Page 1998 an der Stanford University in den Vereinigten Staaten konzipiert und implementiert. Google hat seitdem eine außerordentlich erfolgreiche Entwicklung durchgemacht und steht heute unangefochten an der Spitze bestehender Suchmaschinen. Dies hat verschiedene Gründe, die hier nur kurz skizziert werden sollen.

Google zeichnet sich unter anderem aus durch

- eine besonders einfache Bedienung,
- eine auffällig schnelle Beantwortung der Suchanfragen,
- eine übersichtliche Präsentation der Suchergebnisse,
- verschiedene neue Suchmöglichkeiten,

⁴ Vgl. (Ohst 1998)

⁵ Informationen zu PageRank unter www.google.com/

⁶ Google gibt auf seiner Homepage an, dass inzwischen zwei Milliarden Webseiten durch ihren Suchservice erreicht werden. (März 2002)

⁷ Vgl. SearchEngineWatch.com (Februar 2002)

⁸ Vgl. (www.Google.com/ 2002)

⁹ Alle folgenden Aussagen beziehen sich auf die für diese Untersuchung generierten Daten der Suchmaschine Google (www.google.com/).



Abbildung 1: Ausschnitt aus einer Google Trefferliste

- den größten Webseitenindex aller Suchmaschinen und
- vor allem durch eine neue fortschrittliche Ranking-Methode (PageRank)⁵.

Google ist aus zweierlei Gründen als Rechercheinstrument für diese Arbeit sehr geeignet. Zum einen ist der Google-Index der größte aller momentan bestehenden Suchmaschinenindizes und zum anderen ist Google die einzige der großen Web Search Engines, die PDF-Dokumente indizieren und ausgeben.

Der Google-Index umfasst zum Zeitpunkt der Untersuchung (Sommer 2001) schätzungsweise 1,387 Milliarden Webseiten. Das entspricht der größten Abdeckung von Webseiten aller Suchmaschinenbetreiber. Der Google-Index wird nach Angaben von Google ca. 100 Millionen mal pro Tag durchsucht.⁶

Im Februar 2002 wurde Google unangefochtener Gewinner der „Search Engine Watch Awards 2001“, die von SearchEngineWatch.com verliehen werden. Ein Grund für das gute Abschneiden des Suchservices ist laut SearchEngineWatch.com der Zusatzservice PDF. „In 2001, Google became the first major search engine to go beyond indexing only HTML and text documents, first adding support for PDF files and then many other document types, such as Word documents. This made important content previously “invisible” to searchers available for the first time.“⁷

Google erklärt zu dieser Strategie auf seiner Website: „Google has expanded the number of non-HTML file types searched to 12 file formats. In addition to PDF documents, Google now searches Microsoft Office, PostScript, Corel WordPerfect, Lotus 1-2-3, and others. The new file types will simply appear in Google search results

whenever they are relevant to the user query. ... Overall, the additional file types provide Google users a wider view of the content available on the World Wide Web. And Google has plans to keep expanding the range of file types available over time.“⁸

Nachdem PDF nun im Internet recherchiert werden kann, lohnt sich eine genauere Betrachtung dieses besonderen Dateiformats.

Im Folgenden wird der Weg beschrieben, wie über eine allgemein zugängliche Suchmaschinenausgabe, Daten über das Dateiformat PDF im Internet gewonnen werden können.

2 Untersuchung

Vorbemerkung

Heutige Suchmaschinen decken nur einen Teil der im gesamten Internet verfügbaren Dokumente ab, dies gilt natürlich auch für den Abdeckungsgrad der Suchmaschine Google. Schätzungen beziffern die Abdeckung, die alle Suchmaschinen zusammen erreichen mit weniger als 50 Prozent der im Internet befindlichen Dokumente.

Zusätzlich wird die Größe der Trefferlisten durch den Suchmaschinenbetreiber Google auf maximal 1.000 Treffer pro Anfrage begrenzt, was eine weitere erhebliche zahlenmäßige Reduzierung der zu analysierenden Treffermenge darstellt.

Es wird unterstellt, dass die Teilmenge der im WWW über die Suchmaschine Google verfügbaren Dokumente ausreicht, um im Rahmen dieser Arbeit allgemeine Aussagen zu treffen und das folgende Verfahren zu demonstrieren.⁹

Die Untersuchung konzentriert sich auf zwei wesentliche Komplexe. Als erstes soll in der Treffermenge der Anteil der Treffer im PDF-Format gegenüber anderen Trefformaten bestimmt werden. Dazu werden 50 strukturierte Suchanfragenausgaben von Google analysiert und statistisch ausgewertet. Der zweite Fokus ist die rechnerische Annäherung an eine Durchschnittsgröße der PDF-Dateien, die über eine Stichprobe gezogen werden.

Vorgehen

Das Verfahren zur Bestimmung des PDF-Anteils für eine beliebige Suchanfrage sowie der durchschnittlichen Dokumentgröße der PDF-Dateien gliedert sich in mehrere Schritte.

- Definition der Suchanfragen
- Abschicken der Suchmaschinenanfragen und Sichern der Ergebnisse
- Download der Stichproben
- Auswertung der Ergebnisse und
- Berechnungen

Um Aussagen über den zahlenmäßigen Anteil der PDF-Dateien im World Wide Web treffen zu können, bietet es sich zunächst an, stichprobenartig dieses spezielle Dateiformat zu untersuchen. Dazu greift dieses Verfahren zum Teil auf bestehende Statistikangaben der Suchmaschinenausgabe zurück und stellt auf dieser Grundlage eigene Berechnungen an.

Google gibt in seiner Trefferlistenausgabe zu jeder Suchanfrage an,

- wie viele Treffer eine konkrete Trefferliste insgesamt umfasst,
- wie viele Treffer sich im Index von Google befinden und
- kennzeichnet das Dateiformat des Treffers.

In der Abbildung 1 ist eine typische Google-Trefferliste mit Statistikangaben und der Kennzeichnung für einen Treffer im PDF-Format zu sehen.

Definition der Suchanfragen

Der erste Schritt des Verfahrens ist die Definition von thematisch zusammenhängenden geschachtelten Suchanfragen. Die Suchanfragen bestehen aus Suchtermini, die zu dreiteiligen Anfragegruppen geschachtelt werden. Zu einer Anfragegruppe gehören eine „one term“-Anfrage, eine „3 term“-Anfrage und eine „5_6 term“-Anfrage.

Tabelle 1: Beispiel: Schachtelung von Anfragegruppen

Anfragegruppe (deutsch)	Anfragegruppe (englisch)	Anfragelänge
dokumentation	documentation	one term (1 Suchbegriff)
dokumentation AND sprache AND klassifikation	documentation AND language AND classification	3 term (3 AND verknüpfte Suchbegriffe)
dokumentation AND sprache AND klassifikation AND literatur AND analyse AND dienst	documentation AND language AND classification AND literature AND analysis AND service	5_6 term (5 bzw. 6 AND verknüpfte Suchbegriffe)

- Der erste Teil einer solchen Anfragegruppe besteht immer aus einer „one term“-Anfrage, also einer Anfrage, die genau einen einzelnen Suchbegriff lang ist.
- Der zweite Anfrageteil besteht aus einer „3 term“-Anfrage, die aus drei Suchbegriffen besteht, die konjunktiv verknüpft (AND) sind.
- Die „5_6 term“-Anfrage besteht aus fünf bzw. sechs Suchbegriffen, die ebenfalls konjunktiv miteinander verknüpft sind.

Die drei unterschiedlich langen Anfrageteile enthalten jeweils die Suchwörter der vorhergehenden Anfrage. Die drei verschiedenen langen Anfrageteile einer Anfragegruppe sind thematisch aufeinander abgestimmt. Das heißt, dass die unterschiedlichen Suchtermini eine inhaltliche Verwandtschaft aufweisen (siehe unteres Beispiel: Tabelle 1).

Die verwendeten Suchbegriffe lassen sich in

- allgemeine Begriffe (z.B. Software, Sprache, Bibliothekar, ...),
 - fachspezifische Begriffe (z.B. Corba, Napster, Retrieval, ...) und
 - Relatoren (z.B. Projekt, Bericht, Theorie, ...)
- einteilen.

Die 50 für die Untersuchung verwendeten Suchanfragen teilen sich in 25 deutschsprachige und 25 englischsprachige Anfragen auf. Die Anfragen in deutscher Sprache entsprechen den Anfragen in englischer Sprache inhaltlich (siehe Beispiel oben Tabelle 1).

Anfragen und Sichern

Der zweite Schritt des Verfahrens besteht darin, die definierten Suchanfragen an die Suchmaschine zu stellen und die zurückgegebenen Trefferlisten zu speichern.

Zur weiteren Auswertung stehen damit pro Suchanfrage Trefferlisten mit bis zu 1.000 Treffern zur Verfügung.

Download der Stichproben

Ein weiterer Fokus der Arbeit ist es, spezifische Daten über die PDF-Dateien zu gewinnen und auszuwerten. Die im vorherigen Schritt erhobenen Daten sind Grundlage für eine genauere Betrachtung des PDF-Formats. Um Aussagen über die durchschnittliche Dateigröße dieses Formats treffen zu können, ist es notwendig Stichproben der PDF-Dateien herzustellen. Basis für die Stichprobengewinnung sind die im zweiten Schritt erhobenen Ergebnisse der geschachtelten Anfragegruppen. Die zurückgegebenen Trefferlisten werden nach PDF-Dateien gefiltert. Ein Filterskript selektiert dazu die Treffer im Format PDF, die sich zu einer bestimmten Anfrage in der Trefferliste befinden, und schreibt sie in eine neue Liste. Diese neue Liste, die aus sämtlichen Treffern im PDF-Format besteht, wird anschließend einem Webspider übergeben. Dieser automatische Webspider¹⁰ realisiert das Speichern der einzelnen PDF-Dokumente. Ergebnis dieses Schritts sind Dateiodner im lokalen Dateisystem mit heruntergeladenen PDF-Dateien zu einer bestimmten Anfrage. Diese Dateiodner ermöglichen die Bestimmung der durchschnittlichen Dateigröße der PDF-Dateien (siehe 3 Ergebnisse). Um Vergleiche zu anderen Dateiformaten bezüglich der Dateigröße anstellen zu können, wird die durchschnittliche Dateigröße eines Referenzformats, in diesem Fall das Web-Standardformat HTML, ebenfalls über eine Stichprobe bestimmt. Dazu können die Angaben auf der Trefferliste genutzt werden.

Auswertung der Ergebnisse

Die folgenden Auswertungsdaten enthalten einerseits Daten, die aus den Trefferlisten entnommen werden und andererseits Daten, die den PDF-Downloads entnommen werden.

Daten aus Trefferlisten

In einem nächsten Schritt wird ein Vorgehen beschrieben, mit dem man anhand der ausgegebenen und gespeicherten Trefferlisten bestimmen kann, wie hoch der Anteil der Treffer mit dem Format PDF

¹⁰ Ein Webspider ist ein Programm, das Webseiten unabhängig von ihrem Format automatisiert herunterlädt. Für die Untersuchung wurde das Freewareprogramm Teleport Pro verwendet.

Tabelle 2: Tabelleneinträge für eine Beispielsuchanfrage

Anfrage	Sprache	Länge	G_index	G_out	Zahl_PDF
Suchbegriff(e)	Deutsch v Englisch	one term v 3 term v 5_6 term	Zahl	Zahl > 1000	Zahl
Beispielanfrage: dokumentation AND sprache AND klassifikation	Deutsch	3 term	3720	841	252

Tabelle 3: Tabelleneinträge nach PDF-Download

Anfrage	Sprache	Länge	Anzahl der Dateien	Größe des Dateiordners in Kilobyte
Suchbegriff(e)	Deutsch / Englisch	One term / 3 term / 5_6 term	Zahl	Zahl

für eine definierte Anfrage ist. Zusätzlich kann der Anteil der Treffer mit dem Format PDF zu einer Anfrage auf den gesamten Suchmaschinenindex hochgerechnet werden.

- Ausgangsbasis für die Berechnung ist die maximale Trefferanzahl einer Suchanfrage in der Suchmaschinen-Ausgabe. Dieser Wert wird im Folgenden als „G_out“ bezeichnet. Diese Information ist der Statistikangabe auf der Trefferliste zu jeder Anfrage zu entnehmen (siehe Abb. 1, Kennzeichnung: „Results“).
- Google gibt zu jeder Anfrage an, wie viele Treffer sich ungefähr zu der Anfrage im Index der Suchmaschine befinden. Dieser Wert wird im Folgenden als „G_index“ bezeichnet. Diese Information ist ebenfalls der Statistikangabe auf der Trefferliste zu entnehmen (siehe Abb. 1 Kennzeichnung: „of about“).
- Die dritte Angabe ist die Anzahl von Treffern im Format PDF pro Suchanfrage, die aus „G_out“ berechnet werden muss. Dieser errechnete Wert soll im Folgenden als „Zahl_PDF“ bezeichnet werden. „Zahl_PDF“ ist eine Teilmenge aus „G_out“. Diese Zahl errechnet sich durch Auszählen der Treffer im PDF-Format pro Suchanfrage. Die Dokumente im Format PDF werden in der Trefferliste von Google gesondert gekennzeichnet (siehe Abb. 1, Kennzeichnung: „[PDF]“).

Nach dem beschriebenes Vorgehen lassen sich für jede Suchanfrage sechs Tabelleneinträge vornehmen (siehe Tabelle 2).

- Anfrage: Dieses Feld enthält den Suchbegriff bzw. die verknüpften Suchbegriffe der Suchanfrage.

¹¹ Alle Ausgangsdaten sowie die aggregierten Werte befinden sich im Ergebnisteil bzw. im Anhang dieser Arbeit.

- Sprache: Dieses Feld enthält einen Eintrag, der die Sprache der Suchanfrage bezeichnet.
- Länge: Dieses Feld gibt an, welche Länge die Suchanfrage hat. Es wird unterschieden zwischen one term, 3 term und 5_6 term.
- G_index: Im Feld G_index steht die Zahl für die Trefferanzahl im Googleindex.
- G_out: Im Feld G_out steht der Wert für die ausgegebene Trefferanzahl. Dieser Wert ist kleiner 1000 Treffer.
- Zahl_PDF: Zahl_PDF erhält den errechneten Wert der Treffer im PDF-Format. Nachdem zu jeder Anfrage die Werte der sechs Tabelleneinträge eingetragen worden sind, kann anschließend mit den Einträgen gerechnet werden.

Daten aus den PDF-Downloads

Nach Download der PDF-Stichprobe stehen folgende Daten zur Auswertung zur Verfügung.

Die Einträge „Anfrage“, „Sprache“, und „Länge“ in Tabelle 3 unterscheiden sich nicht von den Einträgen in Tabelle 2 und dienen der Zuordnung der Daten. Zusätzlich werden in Tabelle 3 die Anzahl der heruntergeladenen PDF-Dateien (siehe Tabelle 3 „Anzahl der Dateien“) und die Größe des Dateiordners in dem sich die PDF-Dateien befinden (siehe Tabelle 3 „Dateiordner in Kilobyte“) geführt.

Berechnungen

Aus den in Tabelle 2 und 3 aggregierten Werten können folgende Berechnungen angestellt werden:¹¹

- Summen: Um Aussagen über den Anteil der PDF-Treffer an den Suchmaschinen-Ausgaben treffen zu können, ist es notwendig, die absoluten Zahlen sowohl der Treffer im PDF-Format als auch der anderen Treffer zu bestimm-

men. Die Summen werden unterteilt nach Anfragelängen, Sprachen und Anfragelängen/Sprachen. Die Verhältnisse lassen sich wiederum nach Anfragelängen, Sprachen und Anfragelängen und Sprachen unterteilen.

- Verhältnisse: Nachdem die Summen zu den Tabelleneinträgen bestimmt sind, lassen sich folgende Verhältnisse aus den Summenwerten bilden.
 - Anteil der Treffer im PDF-Format an der Anzahl der Treffer der Googleausgabe
Berechnung: $Zahl_PDF / G_out$
 - Anteil der Treffer in der Googleausgabe (G_out) an der Anzahl der Gesamttreffer im Googleindex (G_index).
Berechnung: G_out / G_index
 - hochgerechneter Anteil der Treffer im Format PDF (PDF_index) an der Anzahl der Gesamttreffer im Googleindex (G_index).
Berechnung: $G_out / G_index \cdot PDF_index$
- Mittelwerte: Die Bestimmung des Mittelwerts einer Zahlenreihe wird an verschiedenen Stellen dieser Untersuchung verwendet.
- Rangdarstellung: Die Rangdarstellung wird an einigen Stellen dieser Untersuchung verwendet um einzelne Tendenzen der Untersuchung zu verdeutlichen. Die Rangdarstellung liefert den Rang, den eine Zahl innerhalb einer Liste von Zahlen einnimmt. Als Rang einer Zahl wird deren Größe, bezogen auf die anderen Werte der jeweiligen Liste, bezeichnet. Die Rangdarstellung ermöglicht, die relative Stellung von Werten in einem Datensatz zu analysieren. Folgende Rangdarstellung scheint für die Untersuchung besonders aufschlussreich. Sortiert man nach den Suchanfragen, die die höchste Anzahl von Treffern mit dem Format PDF haben, lassen sich Tendenzen bzgl. der Sprache einer Anfrage bzw. der Länge einer Anfrage ausmachen.

3 Ergebnisse

Der Ergebnisteil gliedert sich in den Komplex „Ergebnisse der Trefferlistenanalyse“, in dem die Daten aus den Trefferlisten zu den 50 Suchanfragen aufbereitet werden, und den Komplex „Ergebnisse der PDF-Datei-Analyse“, der die Auswertungen bzgl. der durchschnittlichen Dokumentgrößen der beiden Dateiformate PDF und HTML präsentiert.

Tabelle 4: Auswertungsstatistik für 50 Suchanfragen¹²

	G_out	Zahl_PDF	Zahl_PDF/G_out in Prozent
1. Gesamt			
	35376	3374	9,54 Prozent
2. nach Anfrägelängen			
one term gesamt	13978	70	0,50 Prozent
3 term gesamt	12120	1219	10,06 Prozent
5_6 term gesamt	9278	2085	22,47 Prozent
3. nach Publikationssprache			
deutsch gesamt	14643	2274	15,53 Prozent
englisch gesamt	20733	1100	5,31 Prozent
4. Anfrägelänge und Publikationssprache			
one term deutsch	6756	50	0,74 Prozent
one term englisch	7222	20	0,28 Prozent
3 term deutsch	5369	1041	19,39 Prozent
3 term englisch	6751	178	2,64 Prozent
5_6 term deutsch	2518	1183	46,98 Prozent
5_6 term englisch	6760	902	13,34 Prozent

Ergebnisse der Trefferlistenanalyse

Insgesamt sind im Rahmen dieser Arbeit 50 verschiedene strukturierte Suchanfragen gesichert und ausgewertet worden (siehe Tabelle 4).

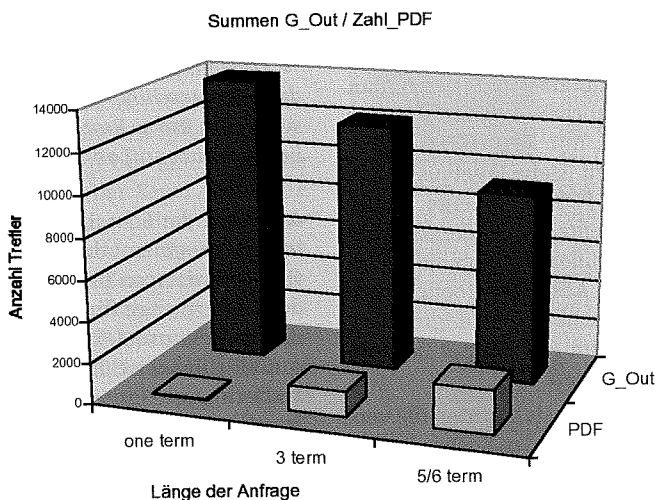
→ **Ergebnis 1:** Insgesamt sind knapp zehn Prozent (9,54 Prozent) der Treffer in der Trefferausgabe PDF-Dokumente (siehe Abbildung 3).

→ **Ergebnis 2:** Der Anteil der PDF-Dokumente hängt stark von Länge der Suchanfrage (Anzahl der konjugierten Suchbe-

griffe) ab. Je länger eine Suchanfrage desto höher ist der Anteil an PDF-Dokumenten in der Trefferausgabe (siehe Abbildung 2).

→ **Ergebnis 3:** Der Anteil der PDF-Dokumente hängt stark von der Sprache der Suchanfrage (Deutsch, Englisch) ab. In der Regel erzielen deutsche Suchanfragen deutlich höhere PDF-Anteile (siehe Tabelle 4 oben).

→ **Ergebnis 4:** Die Treffermenge im Index der Suchmaschine Google nimmt durchschnittlich um 42 Prozent für jeden zusätzlichen Suchbegriff ab.



	one term	3 term	5/6 term
PDF	70	1219	2085
G Out	13978	12120	9278

Abbildung 2: Summen der Trefferausgabe (G_out) und Anzahl der PDF-Treffer (PDF) nach Anfrägelängen

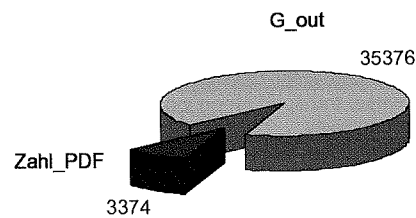


Abbildung 3: Anteil der PDF-Dokumente (Zahl_PDF) an der Trefferausgabe (G_out)

→ **Ergebnis 5:** Die PDF-Dokumente aus der Stichprobe sind im Durchschnitt rund neunmal so groß wie die HTML-Dokumente aus der Referenzstichprobe. Die durchschnittliche Größe einer PDF-Datei, die über beschriebenes Verfahren recherchiert und heruntergeladen wurde, beträgt **293,29** Kilobyte. Der Durchschnittswert der evaluierten HTML-Datei liegt bei **31,97** Kilobyte.

→ **Ergebnis 6:** Deutschsprachige PDF-Dokumente sind durchschnittlich 2,5-mal größer als englischsprachige (im Vergleich dazu sind deutsch- und englischsprachige HTML-Dokumente durchschnittlich etwa gleich groß).

4 Diskussion

Die Kernaussagen aus dem Ergebnisteil werden im Folgenden noch einmal aufgegriffen und bewertet.

→ **Ergebnis 1:** Insgesamt sind knapp zehn Prozent (9,54 Prozent) der Treffer in der Trefferausgabe PDF-Dokumente.

Ein überraschendes Ergebnis dieser Untersuchung ist, dass der Anteil der PDF-Dokumente an den Trefferausgaben, mit knapp 10 Prozent relativ hoch ausfällt. Ein Grund für die hohen Werte ist sicherlich die Methode, wie die Suchanfragen im Vorfeld der Erhebung zusammengestellt wurden. Es wurden keine x-beliebigen Suchbegriffe aneinander gereiht, sondern bei der Definition der Verschachtelung der Suchbegriffe wurde darauf geachtet, dass die Suchbegriffe thematisch zusammenpassen. Ein intellektueller Filter verhindert bei der Definition der Anfragen, dass scheinbar unsinnige Anfragen, wie etwa die Anfrage „Cola AND Schedule AND Schröder AND Queen AND

¹² G_out entspricht den Treffern in der Trefferausgabe und Zahl_PDF entspricht den Treffern im PDF-Format.

Tabelle 5: Ausschnitt der Suchanfragen sortiert nach den Anfragen mit den meisten PDF-Treffern

Anfragebegriff(e)	Anfrägelänge	Sprache	Zahl_PDF
market transport report indicator country productivity	5_6 term	e	327
gesellschaft kommunikation projekt empirisch forschung zukunft	5_6 term	d	315
dokumentation sprache klassifikation	3 term	d	252
universität wissenschaft veröffentlichung theorie wirtschaft	5_6 term	d	243
software benutzer corba	3 term	d	235
dokumentation sprache klassifikation literatur analyse dienst	5_6 term	d	217
software benutzer corba verteilt anwendung methode	5_6 term	d	197
markt transport bericht	3 term	d	179
bibliothekar informationen wissen	3 term	d	177
markt transport bericht indikator land produktivität	5_6 term	d	160

Picture“, an Google gestellt werden. Der relativ hohe Anteil der PDF-Dokumente bedeutet aber nicht, dass das ursprüngliche Dateiformat für Webdokumente (HTML) seine Dominanz im WWW gegenüber PDF verliert. Es demonstriert vielmehr, dass sich zu intellektuell aufbereiteten und inhaltlich zusammenhängenden Suchanfragen auffallend viele PDF-Dokumente in den Trefferausgaben von Google befinden.

Ein möglicher Grund für das häufige Vorkommen von PDF-Dokumenten in den Trefferausgaben, könnte aber auch der Google inhärente Rankingalgorithmus¹³ PageRank¹⁴ sein. PageRank berechnet für jede Seite einer aktuellen Datenbasis aus der Struktur ihrer Referenzen einen globalen „Bedeutsamkeitswert“. Ein Hauptkriterium für einen hohen „Bedeutsamkeitswert“ einer Seite bei Google wäre beispielsweise, wenn besonders viele Referenzen (Links) auf diese Seite verweisen. In dem hier beschriebenen Erhebungsfall könnte sich die Eigenschaft PageRanks, beliebte, also häufig verlinkte Seiten höher zu bewerten und zu ranken, auf das Ergebnis ausgewirkt haben. Dies wäre ein Erklärungsgrund für den auffallend hohen PDF-Anteil an der Erhebung.

→ **Ergebnis 2:** Der Anteil der PDF-Dokumente hängt stark von Länge der Suchan-

frage (Anzahl der konjugierten Suchbegriffe) ab. Je länger eine Suchanfrage, desto höher ist der Anteil an PDF-Dokumenten in der Trefferausgabe.

Ein weiteres Ergebnis, das Rückschlüsse auf die Bedeutung des PDF-Formats im Web zulässt, ist die Tatsache, dass Suchanfragen mit mehreren Suchbegriffen eine höhere Anzahl PDF-Dokumente generieren als Suchanfragen, die aus einem einzelnen Wort bestehen. Damit gibt Ergebnis 2 erste Hinweise auf die Größe bzw. den Umfang der PDF-Dokumente. Die Schlussfolgerung dieses Ergebnisses ist, dass ein durchschnittliches PDF-Dokument mehr Inhalt, sprich mehr Worte, als ein durchschnittliches HTML-Dokument enthält. Ein größerer Vorrat an Wörtern in einem Text, erhöht logischerweise die Wahrscheinlichkeit bei mehreren Suchbegriffen als Treffer in Frage zu kommen, als ein Text, der grundsätzlich wenige Wörter enthält. Diese Schlussfolgerung lässt sich leicht untermauern, wenn man die durchschnittlichen Dateigrößen der PDF-Dokumente mit den HTML-Dokumenten vergleicht (siehe Ergebnis 5). Die Annahme, dass PDF-Dokumente länger sind als Dokumente im Referenzformat HTML, bestätigt zudem die allgemeine Behauptung, dass Abschlussarbeiten, Berichte und vergleichbare umfangreichere Texte bevorzugt im PDF-Format im Web veröffentlicht werden.

Einen weiteren Hinweis könnten die verwendeten Relatoren in den Suchanfragen geben. Relatoren, wie beispielsweise „report“, „projekt“, „veröffentlichung“ oder „analysis“, die sich in beinahe allen konjugierten Suchanfragen befinden, tendieren aufgrund ihrer Bedeutung genau zu dem oben angedeuteten Publikationstypus.

→ **Ergebnis 3:** Der Anteil der PDF-Dokumente hängt stark von der Sprache der Suchanfrage (Deutsch, Englisch) ab. In der Regel erzielen deutsche Suchanfragen deutlich höhere PDF-Anteile.

Ergebnis 3 zeigt, dass deutschsprachige Web-Autoren viel häufiger im PDF-Format zu den recherchierten Themen publizieren, als ihre englischsprachigen Kollegen, obwohl Dokumente in englischer Sprache im Web unzweifelhaft öfter vorkommen (siehe dazu Auswertungsstatistik Tabelle 4). Tabelle 5 verdeutlicht den Trend, dass deutschsprachige PDF-Dokumente wahrscheinlicher als Treffer sind, als englischsprachige. Der Ausschnitt der Suchanfragen (Rangdarstellung) mit dem höchsten Anteil an Treffern im PDF-Format zeigt, dass sich unter den ersten zehn Anfragen lediglich eine einzige in englischer Sprache befindet.

Aber warum wählen Web-Autoren aus dem deutschsprachigen Raum sehr viel häufiger PDF als Publikationsformat? Hierzu lassen sich nur Mutmaßungen anstellen. Z.B. könnte es sein, dass das PDF-Format im deutschsprachigen Raum schlichtweg beliebter und auch verbreiteter ist als im englischsprachigen Raum. Oder dass die deutschsprachigen Web-Autoren mehr Wert auf die Features (z.B. Seitenidentität, digitale Unterschriften, keine nachträgliche Veränderbarkeit) von PDF legen. Vielleicht haben englischsprachige Webnutzer auch einfach nur mehr Vorbehalte in proprietäre Formate zu konvertieren und vertrauen mehr auf die SGML-Philosophie, die mit XML¹⁵ ihren aktuellen Höhepunkt erreicht hat.

→ **Ergebnis 4:** Die Treffermenge im Index der Suchmaschine Google nimmt durchschnittlich um 42 Prozent für jeden zusätzlichen Suchbegriff ab.

Aufgrund der Schachtelung der Suchanfragen zu einem thematischen Gebiet, werden Aussagen über die Verringerung der Treffermenge für Anfragekonjugationen im Gesamtindex von Google möglich. Auffällig dabei ist, dass die Abnahme der Treffermenge unregelmäßig verläuft. Zwei konjugierte Suchbegriffe bewirken zunächst eine Halbierung der Treffermenge, wobei ein weiterer Suchbegriff keine große Reduzierung (Sechs Prozent) der Treffermenge im Index bewirkt. Weitere konjugierte Suchworte lassen die Treffermenge wieder stärker sinken (70 Prozent bei vier, 57 Prozent bei fünf und 23 Prozent bei sechs verknüpften Suchbegriffen).

¹³ Ranking: Mechanismus zur Relevanzbewertung von Suchergebnissen. Suchergebnisse werden in der Regel in absteigender Reihenfolge ihres Ranking-Wertes sortiert dargestellt.

¹⁴ Siehe dazu „The pagerank citation ranking“. Technical Report

¹⁵ XML (Extensible Markup Languages)

¹⁶ Es wird zwischen Type und Token in Dokumenten unterschieden. Längere Dokumente beinhalten viele Tokens, wobei jedes Wort als Token gezählt wird, auch wiederholte Worte. Die Zahl der Types (verschiedene Worte) wächst somit langsamer als die Zahl der Tokens, weil wiederholte Worte nicht als neuer Type gezählt werden.

→ *Ergebnis 5: Die PDF-Dokumente aus der Stichprobe sind im Durchschnitt rund neunmal so groß wie die HTML-Dokumente aus der Referenzstichprobe.*

Wie bereits in Ergebnis 2 angesprochen, kann aus dieser Tatsache geschlossen werden, dass die PDF-Dokumente im Web umfangreichere Texte beinhalten als solche im Referenzformat HTML. Eine tiefergehende Analyse der Dokumente aus den Stichproben, die beispielsweise eine Bestimmung der Menge der Types und Tokens¹⁶ in den Dokumenten bedeutet hätte, war im Rahmen der Arbeit nicht möglich.

Ein Problem bei der Beurteilung der durchschnittlichen Dateigrößen der beiden Dateiformate PDF und HTML besteht darin, dass Bilder, Abbildungen und Grafiken in den beiden Formaten unterschiedlich behandelt werden und somit Auswirkungen auf die Dateigröße haben. Die Größenangaben, die der Google-Trefferausgabe über beschriebenes Verfahren zur Bestimmung der Durchschnittsgröße der HTML-Dateien entnommen wurden, geben nur die Größe der HTML-Datei aus und vernachlässigen die eventuell in der HTML-Datei enthaltenen Bilder. PDF speichert im Gegensatz dazu den Textkorpus samt seiner Bilder in einer Datei. Diese Tatsache beeinträchtigt die Ergebnisse der Untersuchung. Grundsätzlich kann aber davon ausgegangen werden, dass die PDF-Dokumente mehr Text (Wörter) beinhalten als vergleichbare HTML-Dokumente.

→ *Ergebnis 6: Deutschsprachige PDF-Dokumente sind durchschnittlich 2,5-mal größer als englischsprachige (im Vergleich dazu sind deutsch- und englischsprachige HTML-Dokumente durchschnittlich etwa gleich groß).*

Nimmt man an, dass identische Texte in deutscher und englischer Sprache unterschiedlich viele Worte beinhalten, und die deutschen Texte in der Regel durch ihre Satzstruktur die längeren sind, dann überraschen die vorliegenden Ergebnisse. Die Analyse der Dateigrößen nach dem Indikator Publikationssprache ergaben, dass deutschsprachige PDF-Dokumente durchschnittlich 2,5-mal größer sind als die englischsprachigen PDF-Dokumente. Diese Tendenz bzgl. der Publikationssprache bestätigt sich im Referenzformat HTML nicht, die Durchschnittswerte (deutsch,

englisch) fallen vielmehr etwa gleich aus. Daraus folgt, dass deutschsprachige Webpublikationen im PDF-Format in dieser Untersuchung mehr Text und/oder Bildmaterial enthalten als vergleichbare englischsprachige PDF-Dokumente. Genauere Bestandsaufnahmen und Analysen der Stichprobe können hier für Klarheit sorgen.

Abschließend sollen noch einige Erweiterungen erwähnt werden, die an diese Arbeit angeschlossen werden können.

- Eine größere Anzahl an Anfragen würde die Repräsentativität der Untersuchung bzw. des Verfahrens erhöhen. Bei den 50 hier evaluierten Suchanfragen kann bei weitem noch nicht von einem repräsentativen Vorgehen gesprochen werden. Die dargestellten Ergebnisse konnten nur erste Tendenzen und Anhaltspunkte aufzeigen.
- Durch das Hinzunehmen von weiteren Anfragesprachen (z.B. Spanisch, Polnisch, etc.) ließe sich eine größere Vergleichbarkeit zwischen den einzelnen Sprachen erreichen bzw. wären Aussagen über die Verbreitung des Portable Document Format in verschiedenen Ländern möglich.
- Längere Anfragen, also Anfragen mit mehr als sechs konjugierten Suchbe-

griffen, ließen weitere Rückschlüsse bzgl. der Anteile der PDF-Dokumente an den Trefferlisten bzw. den durchschnittlichen Dateigrößen (PDF im Vergleich zu HTML) zu.

- Eine Wiederholung der Erhebung sollte die Ergebnisse bestätigen, konnte aber im Rahmen dieser Arbeit nicht durchgeführt werden.
- Die vielen vorliegenden PDF-Dokumente, die bislang nur bezüglich ihrer durchschnittlichen Dateigröße untersucht wurden, könnten viel tiefer analysiert und bestimmt werden. Beispielsweise wäre eine Typologie der erhobenen PDF-Dokumente denkbar, die Publikationsgruppen definiert und deren Merkmale zusammenführt.
- Weitere Dokumentanalysen (Types, Token, Bilder, ...) sind ebenfalls denkbar.
- Untersuchungen bezogen auf weitere alternative Dateiformate (z.B. Postscript .ps, Powerpoint .ppt) sind über das beschriebene Verfahren möglich.

Dokument; Dateiformat; PDF; HTML; Statistik; Untersuchung; Internet

DER AUTOR

Philipp Mayr



geboren 1975 in München; Studium der Bibliothekswissenschaft, Informatik und Soziologie an der Humboldt-Universität zu Berlin (seit 4/97); Studienschwerpunkt: Information/Dokumentation, Abschlussziel Magister (voraussichtlich 2003) Indexierung englisch- und deutschsprachiger Fachliteratur für das Informationszentrum für Informationswissenschaft und -praxis der FH Potsdam (INFODATA) (seit 9/1998); Studentischer Mitarbeiter (Tutor) am Institut für Bibliothekswissenschaft, Bereich Fernstudium Bibliothekswissenschaft (2/1999 – 8/2000); Praktika beim Deutschen Bucharchiv München (1998), Informationszentrum (IZ) der FH Potsdam (1998) und der DG Bank AG Frankfurt, Bereich Investment Banking (1999); Traineeprogramm bei der DG Bank New York (2000); Werkstudent bei der HiSolutions AG in Berlin, Bereich Corporate Communications (seit 2/2001)

Philip Mayr
Rückertstraße 6
10627 Berlin
mayr@informatik.hu-berlin.de