Informationsrecherche im Internet mit Hilfe der Google Web APIs

Referenten: Philipp Mayr, Fabio Tosques

Lange Nacht der Wissenschaften, Institut für Bibliothekswissenschaft, Humboldt-Universität zu Berlin, den 12. Juni 2004

Agenda

Dauer ca. 25 min.

- Kurze Einführung in die Thematik (Internet-Suchmaschinen, Internet Research, Webometrics/Cybermetrics)
- Einführung in die Google Web API
- Vorstellung der webometrischen Untersuchung „Das Dateiformat PDF im Web“ (2001)
- Live-Demonstration Google API am Beispiel einzelner Anwendungen
- Fragen & Diskussion

Internet-Fakten

- Google hat etwa 4,3 Milliarden Webseiten indexiert (Quelle: Google, 2004)
- das sog. Deep Web ist aber noch sehr viel größer
- 100 Millionen Anfragen bei Google pro Tag (Quelle: Google, 2002)
- Rekord 10 Gbit / sec Datendurchsatz am zentralen Internet-Knoten DE-CIX (Quelle: ecu Forum, 2004)
- 605.6 Millionen Internet-User weltweit (Quelle: Nua)

Einführung - Internet Research

- Internet (Webdaten) seit einigen Jahren als wissenschaftliches Untersuchungsfeld interessant
- Suchmaschinen als ein wichtiges Tool für Internet Research, Webometrics/Cybermetrics
- Möglichkeiten, aber viele Einschränkungen und Probleme …

Internet Research cont.

- Interdisziplinäres Forschungsgebiet Internet Research – soziologische, kulturelle, ökonomische, ästhetische und weitere wissenschaftliche Aspekte
- Einsatzgebiete: Internet Protokolle, Applikationen, Architektur, Technologie
- Internet-Phänomene (z.B. Hubs, Authorities, Small Worlds, …)
- Vermessung von Internet-Strukturen (z.B. Webometrics)
- Warum sind die Google Web APIs für Internetforscher interessant?

Webometrics/Cybermetrics

- Neues Forschungsgebiet in der Informationswissenschaft
- Verwandtschaft zur Bibliometrie/Informetrie (Zitationsanalyse)
- Aufgabe: „quantitative study of web-related phenomena“ (Thelwall et al.)
- Hauptuntersuchungsfeld: Hyperlinks (Inlinks, Outlinks, Backlinks, Sitations)
- Aussagekraft ist durch die Veränderlichkeit im Web eingeschränkt

Google Suchmöglichkeiten

Abb.1: Google – Erweiterte Suche

Grundlagen - Google Web API

- Automatisierte Abfragen waren bis zur Einführung der Google Web APIs im Frühjahr 2002 verboten.
- Interessierte Internet-Forscher waren es gewöhnt, Suchmaschinen mit eigenen Programmen abfragen zu können.
- Die Google Labs reagierten darauf mit etwas Besonderem: den Google Web APIs
  http://www.google.com/apis

Google Web APIs - Funktionsweise

- Die Google Web APIs bieten Programmiersprachen, die WSDL unterstützen eine definierte Schnittstelle zum Google Index.
- WSDL = Webservices Description Language (XML).
- Datei `GoogleSearch.wsdl` beschreibt, welche Dienste über die APIs zur Verfügung stehen, welche Methoden genutzt und welche Argumente dem Programm übergeben werden können.

Untersuchung „Dateiformat PDF …“

- SS 2001, Seminar „Datenerhebung, Datenstrukturierung und Datenerfassung“ bei Prof. Umstätter am IB
- Fragestellung: Explorative Untersuchung des Dateiformats PDF (Portable Document Format) im Internet
- Vorgehen: Analyse von Google-Trefferlisten zu insg. 50 deutsch- und englischsprachige Anfragen (Queries) unterschiedlicher Komplexität (Länge)
- Ergebnisse: Anteile dt. PDF-Treffer deutlich vor englischen, lange Anfragen liefern überraschend viele PDF-Treffer
- Fragestellung ist heute operationalisierbar über die API

Untersuchung „Dateiformat PDF …“

Abb.2 kommentierte Google-Trefferliste

Untersuchung „Dateiformat PDF …“

ANFRAGE	SPRACHE	LÄNGE	INDEX	OUT	ZAHL_PDF
Beispielanfrage: dokumentation AND sprache AND klassifikation	Deutsch	3 term	3720	841	252

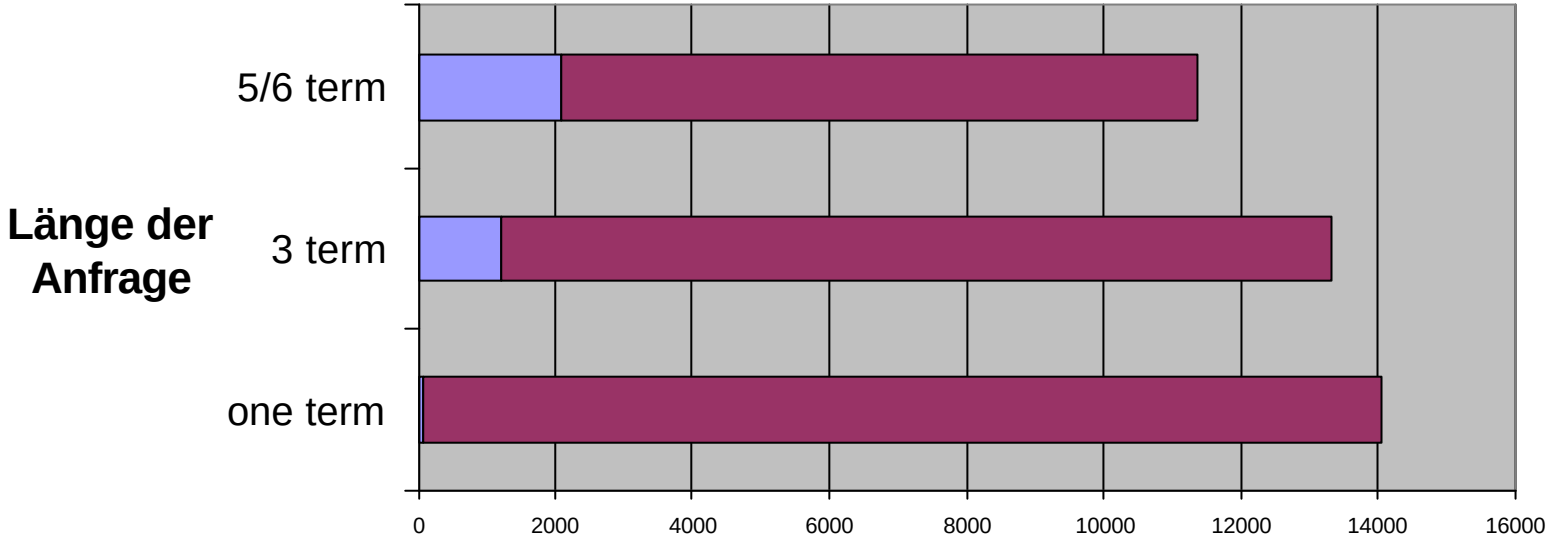
Tab.1: Beispiel einer Anfrage

	OUT	ZAHL_PDF	ANTEIL PDF
Gesamt			
	35376	3374	9,54%
Anfragelänge			
one term	13978	70	0,50%
3 term	12120	1219	10,06%
5_6 term	9278	2085	**22,47%**
Sprache			
deutsch	14643	2274	**15,53%**
englisch	20733	1100	5,31%
Anfragelänge und Sprache			
one term deutsch	6756	50	0,74%
one term englisch	7222	20	0,28%
3 term deutsch	5369	1041	**19,39%**
3 term englisch	6751	178	2,64%
5_6 term deutsch	2518	1183	**46,98%**
5_6 term englisch	6760	902	13,34%

Tab.2: Gesamtstatistik „PDF-Untersuchung“

Untersuchung „Dateiformat PDF …“

PDF zu übrige Treffer



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

Anzahl Treffer

Live-Demo – API-Anwendungen

- Beispiel 1: eine simple Abfrage
- Beispiel 2: Darstellung der Dateitypen-Anteile für verschiedene Anfragen

  Queries 1: dokumentation -> dokumentation sprache

  -> dokumentation sprache klassifikation

  bericht 
- Beispiel 3: reihenfolge (Permutation der Begriffe)
- Beispiel 4: `daterange:`-Suche

Live-Demo - Screenshots

Lange Nacht der Wissenschaften

Query Google with the Google Web Api

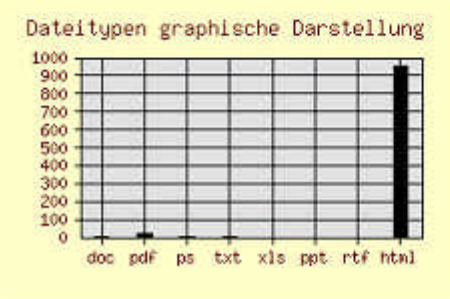
Enter Your Query here: dokumentation [Search Google]

How many results? 1000

Dateityp	Anzahl
html	951
pdf	25
doc	1
ps	1
txt	1

Anzahl der gefundenen Dokumente: 979

Dateityp	Prozent
html	97.1399387129724
pdf	2.55362614913177
doc	0.102145045965271
ps	0.102145045965271
txt	0.102145045965271



Erstes PDF-Dokument an Stelle: 58

Relevanz der Reihenfolg

Suche: internet researcher deuts [Search]

Eingabe: max. 4 Begriffe oder Phrasensuche in doppelten Hochkommata

Result Counts by Permutation	
Query	**Count**
report deutschland researcher internet	20700
deutschland report researcher internet	20700
deutschland researcher report internet	2060
deutschland researcher internet report	21600

Top Results across Permutations	
Score	**Result**
45	**heise mobil - i-mode-Report** http://www.heise.de/mobil/artikel/2002/03/25/imode/ *... i-mode-**Report**. ... bunten Bildern und Ã¼ber 60 Content-Angeboten eifert E-Plus dem japanischen Vorbild nach und versucht, i-mode in **Deutschland** zu etablieren. ...*
38	**Antivirensoftware, Antivirenprogramme, Virenschutz, Internet ...** http://members.aol.com/rlink/security/avlist.htm *... hierzu auch den CeBIT98-**Report** Norton AntiVirus ... Edition Norton AntiVirus for **Internet** E-Mail ... Symantec, USA Distributor: Symantec (**Deutschland**) GmbH, Ratingen ...*

Diskussion

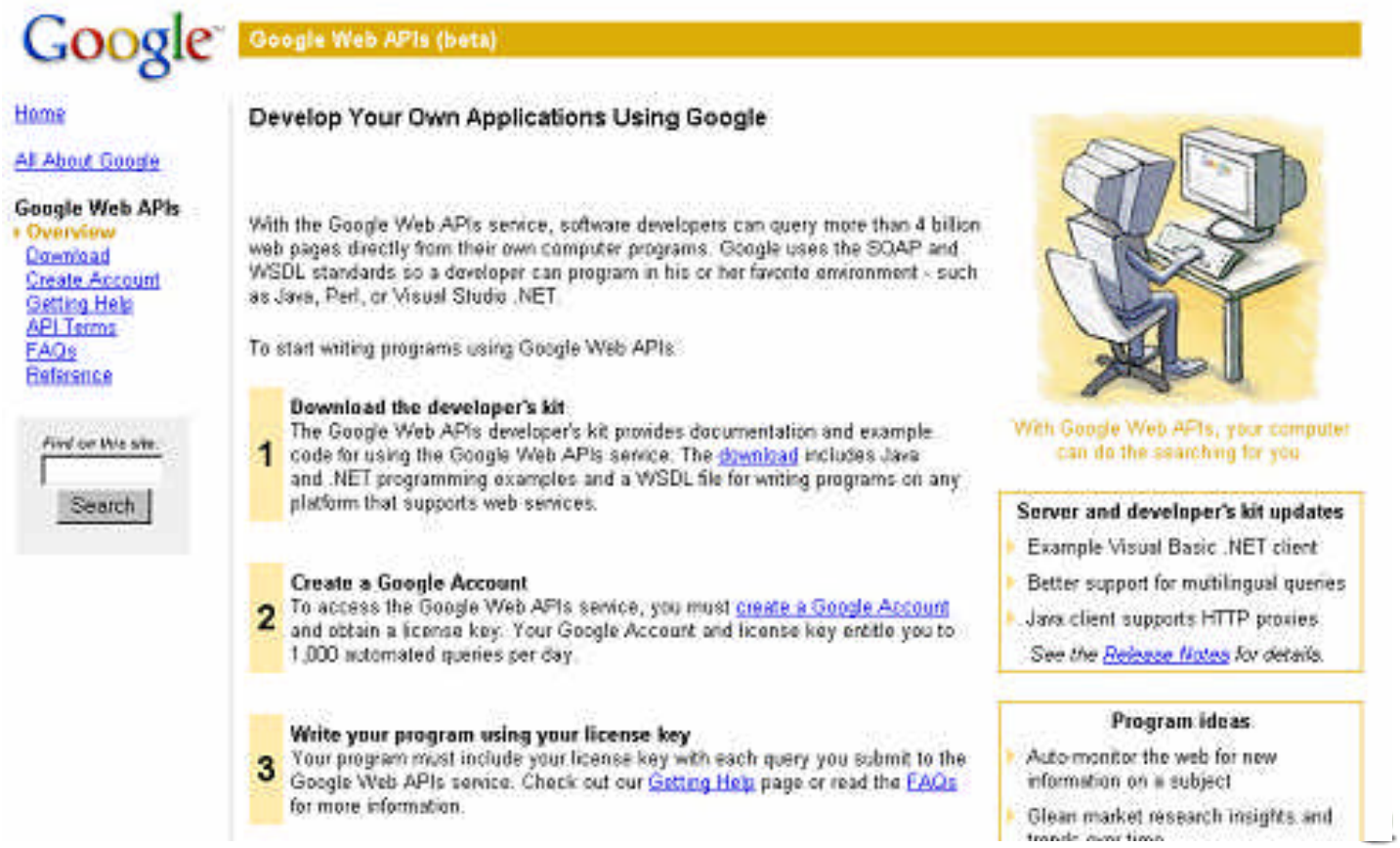
- Google Web APIs befinden sich noch in der Beta-Entwicklungsphase
- Ergebnisse der APIs sind nicht identisch mit jenen der Formularabfrage
- Diskussion der PDF-Ergebnisse

Literatur & Links

- Google Hacks 100 Industrial-Strength Tips & Tricks. By Tara Calishain, Rael Dornfest, O'Reilly, 2003
- Homepage der Google Web APIs http://www.google.com/apis
- Almind, T. C.; Ingwersen, Peter: Informetric analyses on the world wide web: methodological approaches to 'webometrics'. In: Journal of Documentation, Vol. 53, 1997, S. 404-426.
- Barabasi, Albert-László: Linked – The New Science of Networks. Cambridge, Mass.: 2002.
- Bar-Ilan, Judit: Search Engine Results over Time – A Case Study on Search Engine Stability. In: Cybermetrics, Vol. 2/3, 1998/99, available: http://www.cindoc.csic.es/cybermetrics/articles/v2i1p1.html [19. November 2003].
- Björneborn, Lennart; Ingwersen, Peter: Perspectices of webometrics. In: Scientometrics, Vol. 50, 2001, S. 65-82.
- Brin, S., Page, L.: The anatomy of a large scale hypertextual web search engine. In: Computer Networks and ISDN Systems, Vol. 30, 1998, S. 107-117, available: http://citeseer.nj.nec.com/brin98anatomy.html [19. November 2003].
- Mayr, Philipp: Das Dateiformat PDF im Web - eine statistische Erhebung. In: NFD – Nachrichten für Dokumentation, Jg. 53, 2002, S. 475-481, available: http://www.informatik.hu-berlin.de/~mayr/arbeit/pdf_im_web.pdf [19. November 2003].
- Thelwall, Mike: Extracting Macroscopic Information from Web Links. In: Journal of the American Society for Information Science and Technology, Vol. 52, 2001, S. 1157-1168.

Fragen?

Vielen Dank für Ihre Aufmerksamkeit!



Kontakt

Philipp Mayr

email:mayr@informatik.hu-berlin.de,

philippmayr@web.de

www: http://www.informatik.hu-berlin.de/~mayr/

Fabio Tosques

email: tosques@informatik.hu-berlin.de