

Studying Evolution of Temporal Collections

Avishek Anand anand@l3s.de

joint work with: Helge Holzmann





Studying Evolution of Temporal Collections



Avishek Anand anand@l3s.de

joint work with: Helge Holzmann



The Dawn of Today's Popular Domains A Study on 18 Years of the German Web

Avishek Anand anand@l3s.de

joint work with: Helge Holzmann



The Opening Line Syndrome..



The Opening Line Syndrome.. "Web is * growing"



Avishek Anand



The Web is definitely growing

Web Dynamics How does the Web change and evolve?

Change in persistent documents

Change in "real-time" content streams

Change in Web graphs

Does most of the content remain unchanged once it has been authored ? or Are the documents being continuously updated?

Forschungszentrum U3S Research Center

Do pages change a little or a lot ? or Do pages change and then change back?

Web Dynamics

Applications

- Change detection useful is designing better crawlers
- Improving freshness in results
- Improving retrieval quality taking time into account
- Resource planning for the future

"sheds light on the evolution of a major sociological phenomenon: the largest collectively constructed information repository known to man" - Fetterly et al.

Web Dynamics

Previous Works

Rely on frequent crawls over a short period of time



- 67% web pages never changed (on an average)
- Only 20% web pages available today will be accessible after a year
- 1.3% new pages are encountered at every new crawl
- Term-level changes: Popular pages change frequently but not by much
- Changes in pages depend on related pages

What is missed?



10



Willkommen an der Universität Hannover!

Hier finden Sie Informationen über (Some of the following information is available in English):

- Die Universität Hannover
- Zentrale Einrichtungen
- · Forschungsförderung und Technologietransfer
- Akademisches Auslandsamt
- <u>Fachbereiche</u>
- Studentische Einrichtungen
- Die Uni von A-Z: <u>Gewußt Wo Gewußt Wann</u>.
- E-Mail Adressen an der <u>Uni-Hannover</u> und <u>außerhalb</u>
- Informationssysteme: Universität oder Hannover allgemein oder außerhalb

Volltext-Suche <u>über alle WWW-Server der Uni Hannover</u> (Harvest) Meta-Suche <u>über deutschsprachige Suchmaschinen</u> (MetaGer)

Ansprechpartner: Wolfgang Sander-Beuermann, RRZN





ы	\overline{a}	•		.			0
	UI		=				-
	• •••		- 1		-	3	_

- Universität
- Aktuelles
- Studium
- Forschung
- Einrichtungen
- Fachbereiche
- Campus
- Suche
- English-



Universität Hannover

- Willkommen!
 - Begrüßung durch den Präsidenten
- Universität

Präsident, Adresse, Geschichte, Wegweiser, Pressestelle, Intern

Aktuelles

Letzte Änderungen, Tag der Forschung 2000 (NEU), Wettbewerbe, Schule und Universität

Studium

Studienangebot, Immatrikulation, Vorlesungen, Auslandsamt 🛄, Career-Service

Forschung

Forschungsförderung und Technologietransfer, Innovationsprojekte

2000

Kontakt Sitemap erweiterte Suche Suche

Universität

Forschung

Fakultäten

Service

Aktuelles

Weiterbildung

Internationales

Studium

Leibniz

Universität Hannover

Mit Wissen

Studierende Studieninteressierte Schüler/innen | Alumni | Beschäftigte | Gäste | Presse

Nun ist es soweit: Ab 1. Juli 2006 heißt die Universität

Hannover Gottfried Wilhelm Leibniz Universität Hannover

dem Leiter der Leibniz-Akademie, teilte Präsident Prof. Erich Barke mit, dass die Universität Hannover ab dem 1. Juli 2006 einen neuen Namen trägt: Gottfried Wilhelm Leibniz Universität.

In einem gemeinsamen Pressegespräch mit Dr. Wilfried Prewo,

Leibniz Universität Hannover

English

Veranstaltungen

Geodätisches Kolloquium - Moderne Methoden und Anwendungen der Radarfernerkundung Vortragsreihe und Kolloquium 18.07.2006 🗈

Wozu noch Kritische Theorie? Ringvorlesung 18.07.2006 E

Erkenntnis über Funktionen: Ein Band zwischen den Lebens-, Sozial- und Technikwissenschaften Kolloquium der ZEWW 18.07.2006 🗈

Charlie und die Schokoladenfabrik Unikino im Audimax 18.07.2006 E

Terminkalender

Neuigkeiten !

Informationen

de

dies

Jobr

Alle Termine und Veranstaltungen im Überblick 🗈

ersität Han/

Ein buntes Zeichen zum Geburtstag

OK

Der Geburtstag des hannoverschen Universalgenies Gottfried Wilhelm Leibniz jährte sich am 1. Juli 2006 zum 360sten Mal. Aus diesem Anlass und zur Umbenennung der Hochschule in Leibniz Universität Hannover wurden 24.000 Luftballons vor dem Welfenschloss gestartet.

Zukunft gestalten,



Forschungszentrum

2006





Avishek Anand



Leibniz Universität Hannover

Universität

Studium

Forschung

Weiterbildung

Fakultäten

Internationales

Service

Aktuelles



Got an Idea? Challenge the World!

Leibniz Universität Hannover

Der Wettbewerb Intel Business Challenge Europe (IBC) geht an den Start und bietet Studierenden, Doktoranden und Absolventen (bis vier Jahre nach Abschluss) im Alter von 18 bis 34 Jahren die Möglichkeit, Geschäftsideen auf dem Gebiet neuer Technologien zu entwickeln und ein Startup aufzubauen. Wer mitmachen möchte, kann ab sofort bis einschließlich Sonntag, 15. Juni 2014, ein Geschäftsmodell und ein kurzes Elevator-Pitch-Video einreichen. Die Leibniz Universität Hannover ist eine von vier Universitäten in Deutschland, die Partner des Wettbewerbs sind.



"Carmen" begeistert im Lichthof

Liebe, Ehre, Gewalt, eine Femme fatale und feurige Liebhaber: Der Lichthof des Welfenschlosses verwandelte sich am Wochenende 5./6. April in eine Opernbühne. Chor und Orchester der Leibniz Universität haben mit Unterstützung von

Veranstaltungen

Atomic layer deposition of aluminum oxide on crystalline silicon: Fundamental interface properties and application to solar cells Disputation 22.04.2014

Many-body physics with ultracold atoms in disorder Mathematisch-Physikalisches Kolloquium 22.04.2014 🗈

Context and Identity Dienstags um 6 22.04.2014 E

 Alle Veranstaltungen im Überblick

Wussten Sie scl

dass

Lesesti

Forschungszentrum 138 Research Center

2014







Forschungszentrum 138 Research Center

Study of long-term change and evolution possible

Our Dataset

18 years of web data from the german (.de) domain courtesy *The Internet Archive*

2.3 TB ... 5.51B captures ... 11,174, 079 domains spanning > 2 years

	# Pages	# Duration	ø Year
Fetterly et. al [8]	$150\mathrm{m}$	11 weeks	2003
Ntoulas et. al [15]	720k	4 months	2004
Kim et. al $[11]$	600k	$100 \mathrm{~days}$	2005
Adar et. al [2]	50k	5 weeks	2009
Radinsky et. al [18]	54k	6 months	2013
L3S-IA	$20.7\mathrm{m}$	18 years	2014

Largest Web Dynamics study in size and coverage

Retrospective Analysis

We study the evolution of the web retrospectively



Challenge: Lack of coherent data. Change analysis not possible

Popular German Domains

Category	# Domains	# Sub-Domains	# URLs
Computer	100	561	2138786
Recreation	100	380	981638
Society	100	368	832017
Health	100	274	453282
Kids & Teens	100	234	311705
Culture	100	250	934552
Media	100	512	1981877
Shopping	100	429	6726195
Regional	100	793	3069791
Games	99	304	718348
Sports	100	290	656859
Business	100	546	1534639
Education	100	827	1240196
Science	100	398	579821
Home	100	325	1762361
News	40	117	820163
Universities	100	828	659175
TOTAL	1444	5846	20778475

100 most popular german (.de) domains from 17 categories from Alexa rankings

http://www.alexa.com/topsites/category/Top/World/Deutsch

Goal of Analysis

For the currently popular German domains...

Is the Web really growing old and if so how can we characterize it?

How has the size of webpages changed over time?

Do websites from different categories have different growth rates?

Outline

- Introduction
- Experimental Setup
- Age of the Web
- Growth of the Web
- Some Predictions
- Final Remarks

Experimental Setup



Data Transformation

- Extract popular domains
- Filter URLs by extension, status code and time
- Indexed hierarchically for efficient access
- Precomputed statistics per domain for evolution, domain and URL age

Domain Emergence



75% of today's University websites, 20% of today's Game websites

Age of the Web

URL age distribution



Age of a domain composed of ages of current URLs

Age of the Web Evolving URL Age Distribution



URLs grow exponentially, but have stable proportions Is the Web not growing "Order"?

7e+06

5e+06

Forschungszentrum 138 Research Center

25

Age of the Web

URL Age Evolution



The popular Web is indeed growing older But where is the ageing taking place ?

Old gets Older Age of long-living URLs



Old URLs tend not to be replaced

Outline

- Introduction
- Experimental Setup
- Age of the Web
- Growth of the Web
- Some Predictions
- Final Remarks



We differentiate between volume and size



Both the number and size grow over time

Shopping websites have the highest growth rate

Forschungszentrum 138 Research Center

2072

Growth of the Web

URL Size evolution



(a) Alive Size

(b) Birth Size

Forschungszentrum 138 Research Center

Webpages are growing in size

Both existing URLs and new URLs are growing

Domain Age

Growth of the Web

URL Size at birth and death



Webpages size at death greater than size at birth

Outline

- Introduction
- Experimental Setup
- Age of the Web
- Growth of the Web
- Some Predictions
- Final Remarks

Predictions



Average age of ~2 years 2020 double of that in 2005

The Web turns 3 years old in 2038

Predictions



By 2016 the domain volume will double

By 2020 the number of URLs will be 6.7 times that of today

By 2030 almost 166 times the number of URLs today

	18000	f(x)					
Avishek Anand	16000	total	•	 34		 • •	 - rorschungszentrum Los Research Center

Predictions



A new URL in 2017 will be born with double the file size as today

Conclusion, Findings, Outlook

- Extensive longitudinal study on 18 years of the popular German Web
- Analysed how popular domains of today have grown
 - In terms of age / volume / size
- Popular educational domains have been around for very long
- Shopping and game websites mainly emerged during last decade
- The Web is actually getting older
 - at least the old part of it
- Domains grow exponentially
 - doubling their volume every two years
- Tomorrow's newborn URLs will be bigger than today
 - resource planning and allocation, e.g., for Web archives

Outlook

- How does an average sample compare to today's most popular domains ?
- How have the most popular domains of 1996 evolved until today?
- How does the Web of others countries compare to these studies ?
- How much of the increase in size is due to increase in content?

The Web is definitely growing "Thank you for your attention.."

Forschungszentrum

Growth of the Web

Domain Volume

