Web Dynamics

Part 1 - Introduction

1.1 Dimensions of dynamics in the Web

1.2 Application examples

Why Web Dynamics?

From Wikipedia:

*In physics the term **dynamics** customarily refers to the time evolution of physical processes.*

Which aspects of the Web are dynamic?

- **Size**: sites/pages added and deleted all the time

Number of sites on the Web

- 1998: 2,636,000 (IP addresses with HTTP server)
- 1999: 4,662,000
- 2000: 7,128,000, ~40% public, 40% dead
- 2001: 8,443,00
- 2002: 8,712,000

- 2007: 109 million sites (Netcraft)
- 2007: 433 million hosts on Internet (ISC)

1998 – 2002: http://www.oclc.org/research/projects/archive/wcp/stats/size.htm

Size estimates for the (indexable) Web

- 1995: ~11.4 million docs (Bray)
- 1997: ~200 million docs (Bharat&Broder)
  (sampling based on Hotbot, Altavista, Excite and Infoseek, overlap ~2%)
- 1998: >800 million docs (Lawrence&Giles)
- January 2005: 11.5 billion docs (Gulli&Signorini)
  (sampling based on Google, MSN, Yahoo! and Ask/Teoma)
- 2005: 19.2 billion documents in Yahoo! index
- 2008: >1 trillion documents counted by Google

http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html

The Web is infinite – and growing

- Non-indexable Web not seen by search engines („**Deep Web**“ behind forms):
  - est. 550 billion docs,
  - est. 7.5 petabytes in 2000 (Bright Planet)
- **User-generated content** (social networks, communities, wikis, blogs, …)
- Pages **created on demand**
  („next week“ link in online calendars)

Some social networks

Flickr: (as of Nov 2008)

- 3+ billion photos (2 billion in Nov 2007)
- 3 million new photos per day

Facebook: (as of Nov 2008)

- 10+ billion photos, 30+ million new photos per day
- 120 million active users (31 million in April 2007)
- 150,000 new users per day (100,000/day in April 2007)

Myspace: (as of Apr 2007)

- 135 million users (6th largest country on Earth)
- 2+ billion images (150,000 req/s), millions added daily
- 25 million songs
- 60TB videos

StudiVZ.net: (as of Nov 2008)

- 11 million users
- 300 million images, 1 million added daily

Challenges: Size dynamics

How can a search engine deal with „infinite“ Web?

- Massively parallel, distributed architecture (MapReduce, Hadoop, etc.)
- Detect and remove noise (duplicates, spam etc.)

Which aspects of the Web are dynamic?

- **Size**: pages added and deleted all the time
- **Content**: pages change all the time

Evolution of the Web (Ntoulas et al., 2004)

Large-scale study:

- October 2002 – October 2003
- Weekly crawls of 154 large Web sites (up to 200,000 pages per site)

Average page creation per week

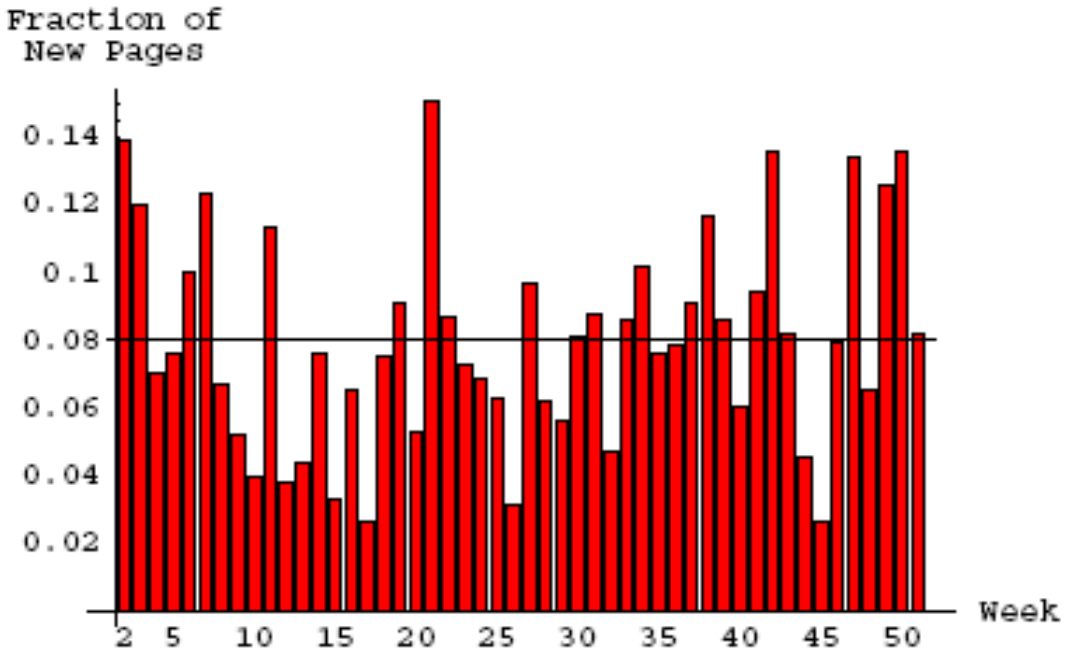


Figure 1: Fraction of new pages between successive snapshots.

About 8% new pages created per week

(A. Ntoulas, J. Cho, C. Olston: What's new on the Web? The Evolution of the Web from a search engine perspective, WWW Conference, 2004)

How long do pages live?

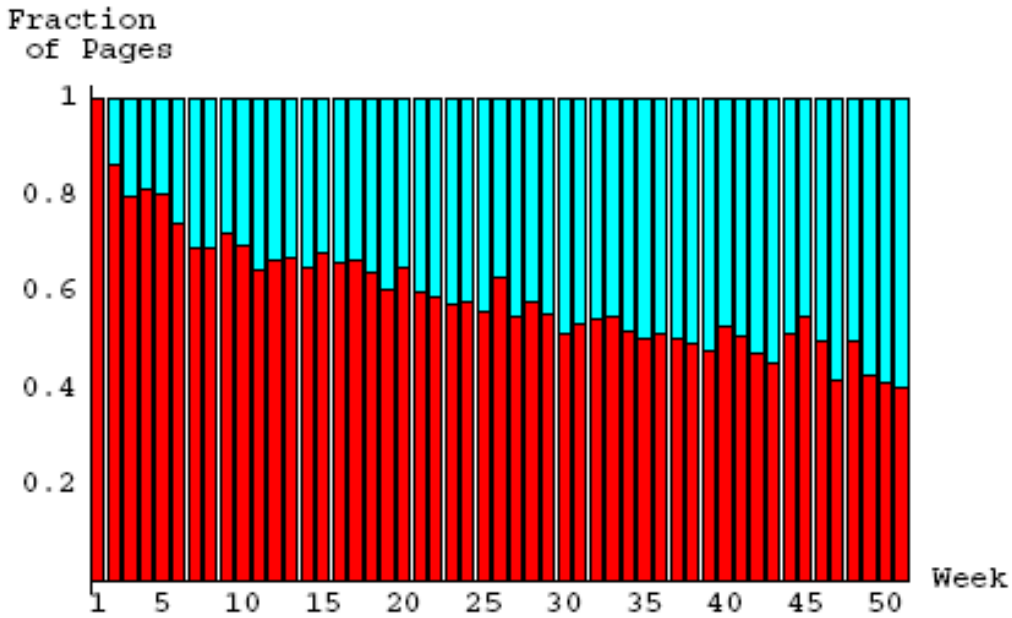


Figure 3: Normalized fraction of pages from the first crawl still existing after n weeks (dark bars) and new pages (light bars).

About 40% of the pages still available after one year

(A. Ntoulas, J. Cho, C. Olston: What's new on the Web? The Evolution of the Web from a search engine perspective, WWW Conference, 2004)

How frequently does a page change?

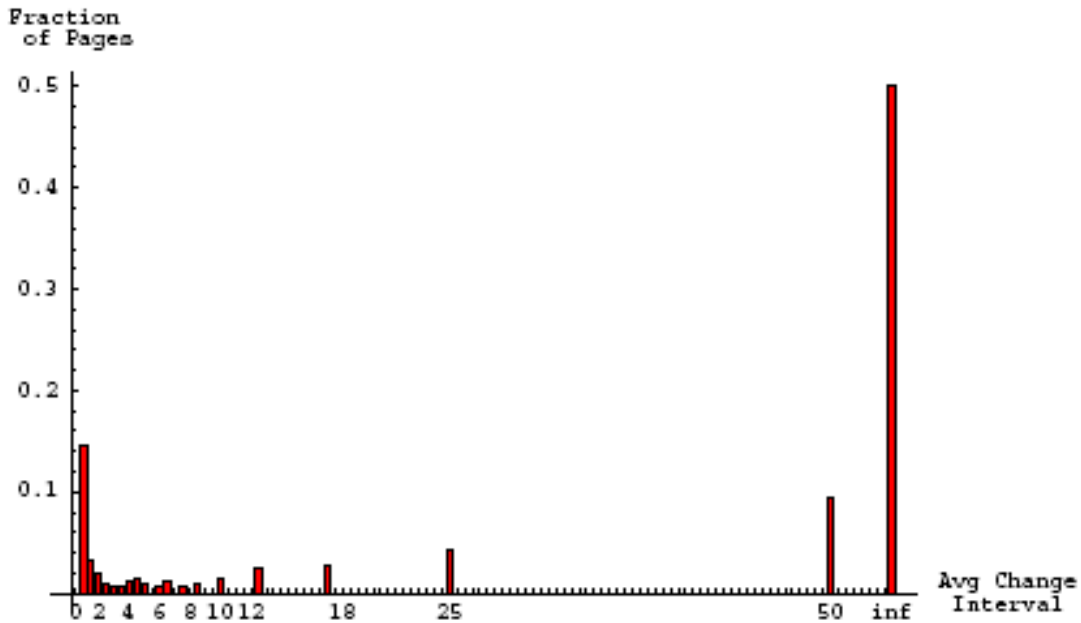


Figure 10: Distribution of the average change intervals of the pages.

Most pages never change, second most change at least weekly

(A. Ntoulas, J. Cho, C. Olston: What's new on the Web? The Evolution of the Web from a search engine perspective, WWW Conference, 2004)

How much do pages change?

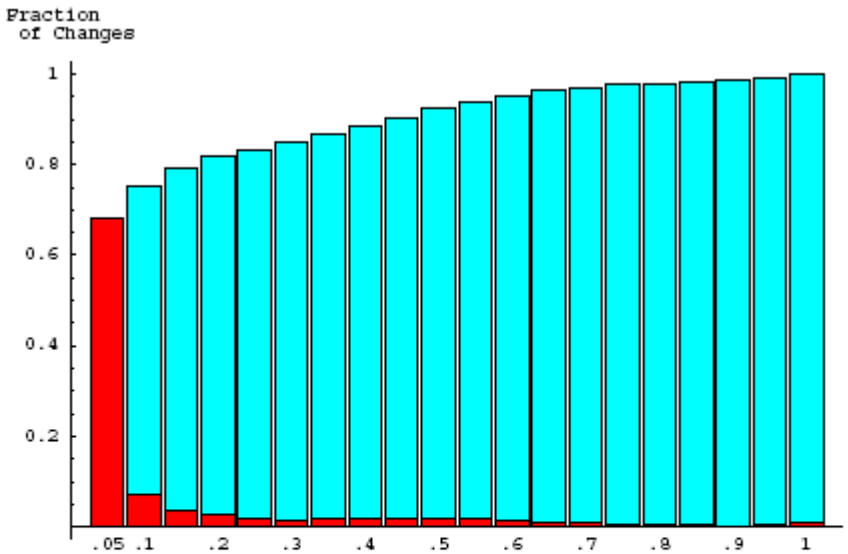


Figure 11: Distribution of cosine distance for all changes. Each dark bar corresponds to changes with cosine distance between the respective x-axis value and the previous one. For example, bin 0.1 corresponds to changes with cosine distance between 0.05 and 0.1. The light bars show the cumulative distribution.

Most of the changes are minor

(A. Ntoulas, J. Cho, C. Olston: What's new on the Web? The Evolution of the Web from a search engine perspective, WWW Conference, 2004)

How large are pages?

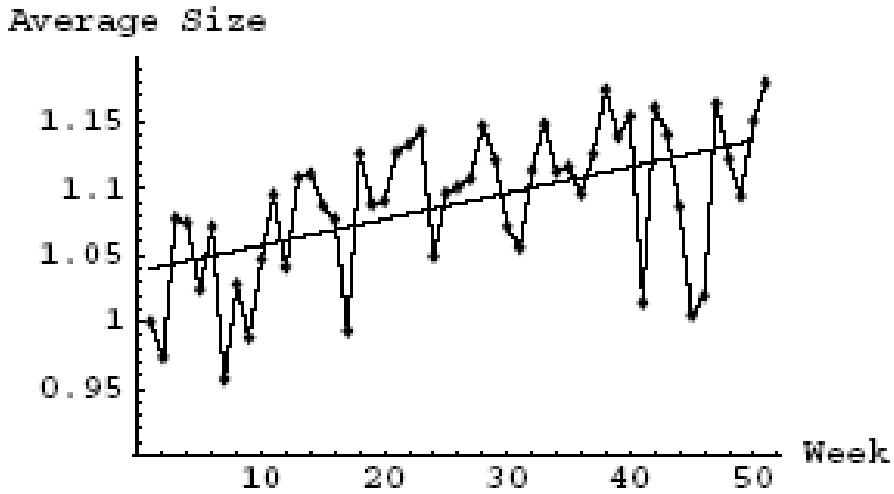


Figure 5: Average page sizes in our snapshots over time.

Average size raised by about 15% in one year

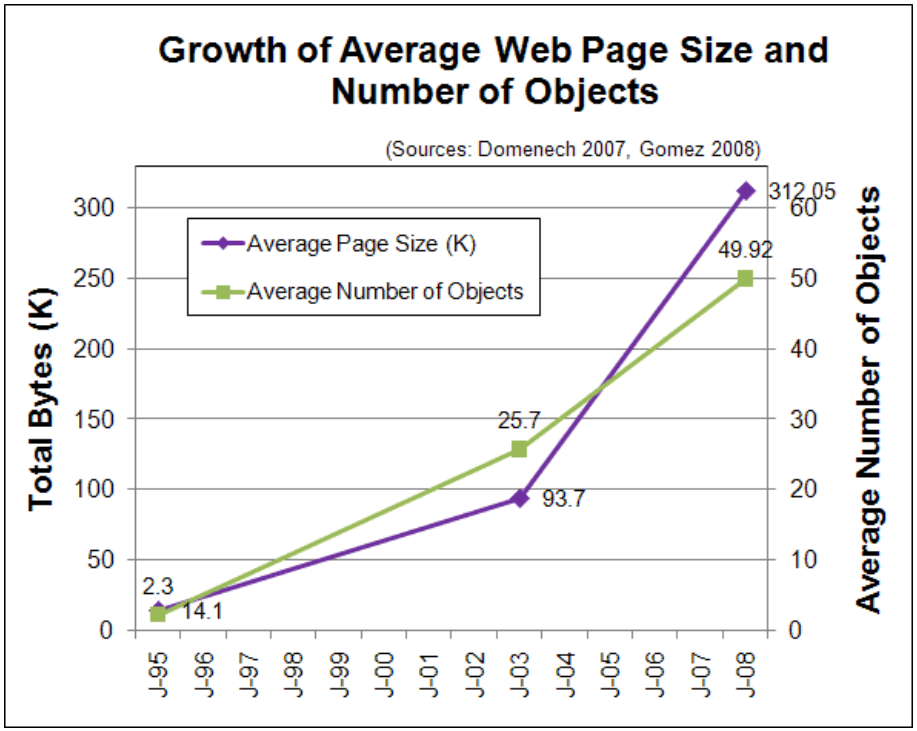
(A. Ntoulas, J. Cho, C. Olston: What's new on the Web? The Evolution of the Web from a search engine perspective, WWW Conference, 2004)

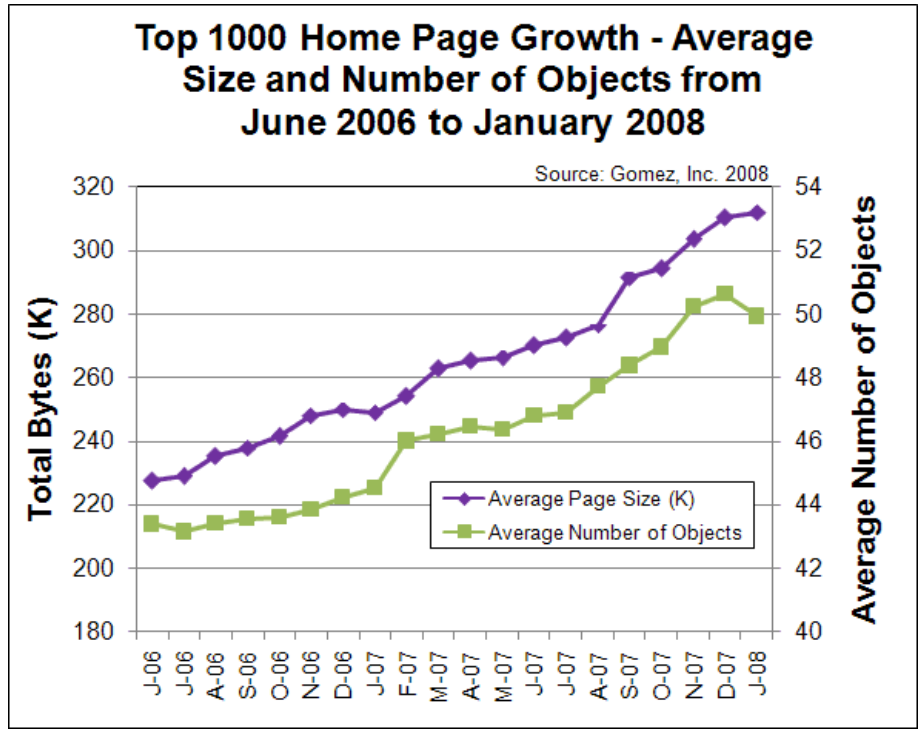
More recent numbers…

- Average size of Web pages more than tripled since 2003 from 93.7K to over 312K
- Average number of objects per Web page nearly doubled from 25.7 to 49.9
- Since 1995 average size of Web pages increased by 22 times
- Since 1995 average number of objects per Web page increased by 21.7 times

(from http://www.websiteoptimization.com/speed/tweak/average-web-page/)

More recent charts…





(from http://www.websiteoptimization.com/speed/tweak/average-web-page/)

Challenges: Content dynamics

How can a search engine maintain a reasonably accurate snapshot of the Web?

- Model how/when documents updated
- Recrawl policy based on expected changes
- Decide if a page‘s content changed (enough to replace old version in snapshot)

How can we maintain the Web of the past?

- Web archiving

Which aspects of the Web are dynamic?

- **Size**: pages added and deleted all the time
- **Content**: pages change all the time
- **Structure**: links added all the time (and dropped)

How frequently do links change?

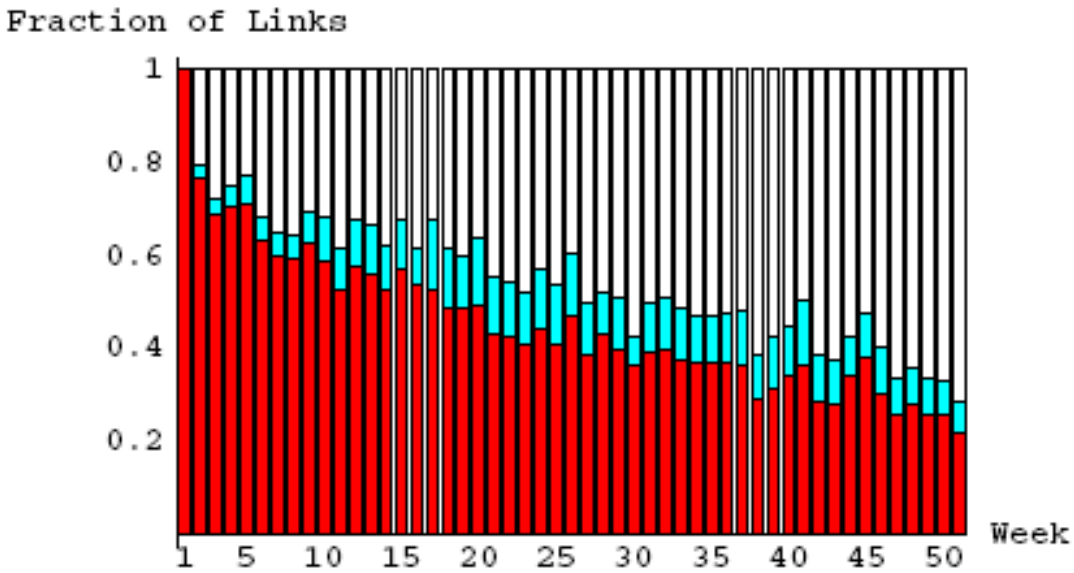


Figure 9: Normalized fraction of links from the first weekly snapshot still existing after n weeks (dark/bottom portion of the bars), new links from existing pages (grey/middle) and new links from new pages (white/top).

25% new links created per week, 80% of links replaced within a year

(A. Ntoulas, J. Cho, C. Olston: What's new on the Web? The Evolution of the Web from a search engine perspective, WWW Conference, 2004)

Challenges: Structure dynamics

How can a search engine maintain a reasonably accurate snapshot of the Web graph?

- Massively parallel, distributed architecture (MapReduce, Hadoop, etc.)
- Distributed approximation algorithms for computing authority measures (PageRank)

Which aspects of the Web are dynamic?

- **Size**: pages added and deleted all the time
- **Content**: pages change all the time
- **Structure**: links added all the time (and dropped)
- **Usage**: Behaviour of users changes all the time

Reasons why user behaviour changes

- Global trends and changes, Web 2.0
  (Flickr, Youtube, social networks, twitter, …)
- Different situation/context
  - Roles (private vs. professional)
  - Locations (home vs. office vs. travelling)
  - Date & Time
  - Tasks (ordering a book, booking a flight, …)

$\Rightarrow$ **influence browsing and search behaviour**

Challenges: User dynamics

How can a search engine adapt to changing users?

- Identify user (e.g., Google‘s cookie)
- Collect user behaviour
- Personalize search results based on past actions
- Personalize based on current context

This can be done

- For each user
- For groups of users
- For all users („global user model“)

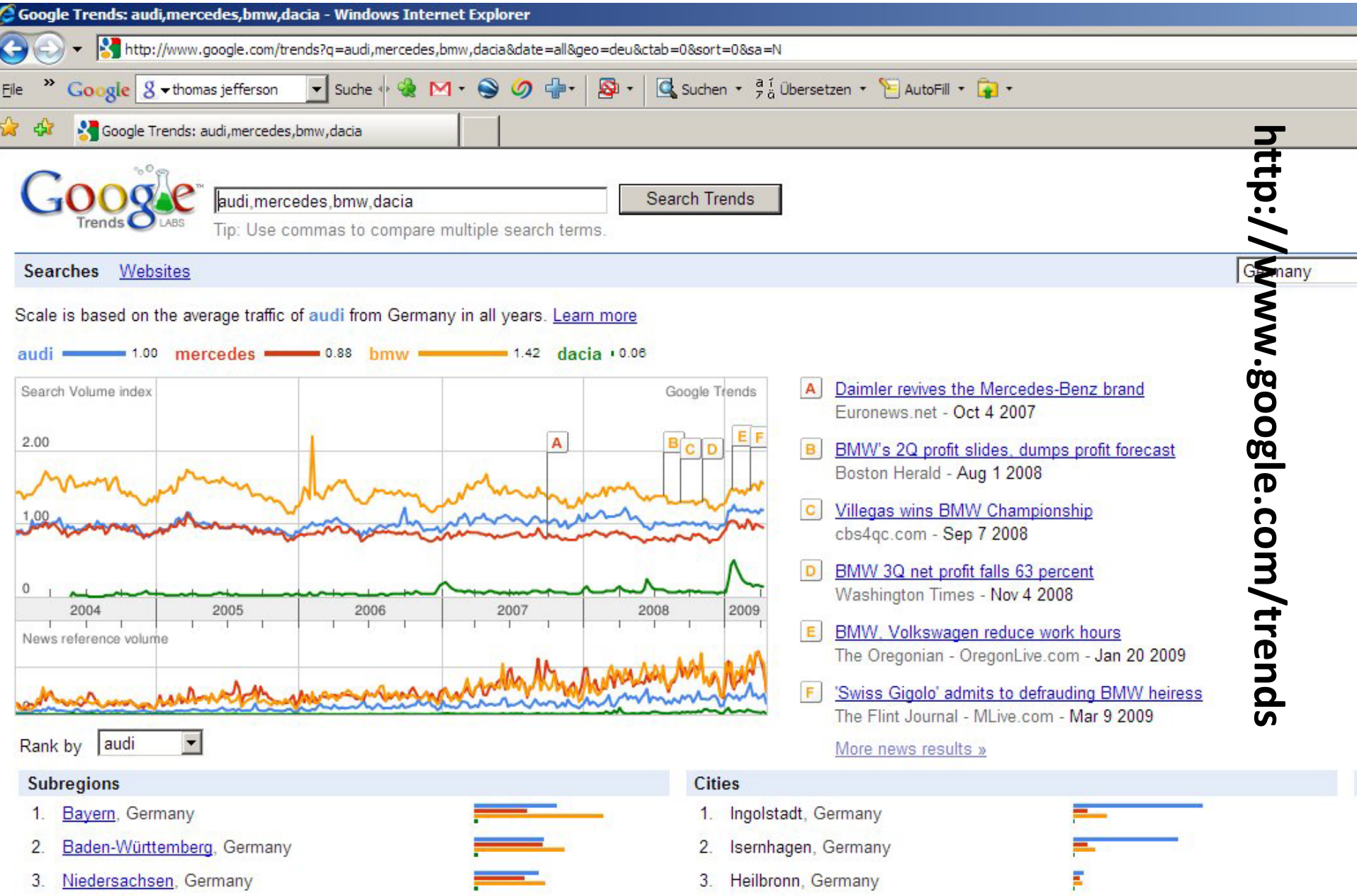
Web Dynamics

Part 1 - Introduction

1.1 Dimensions of dynamics in the Web

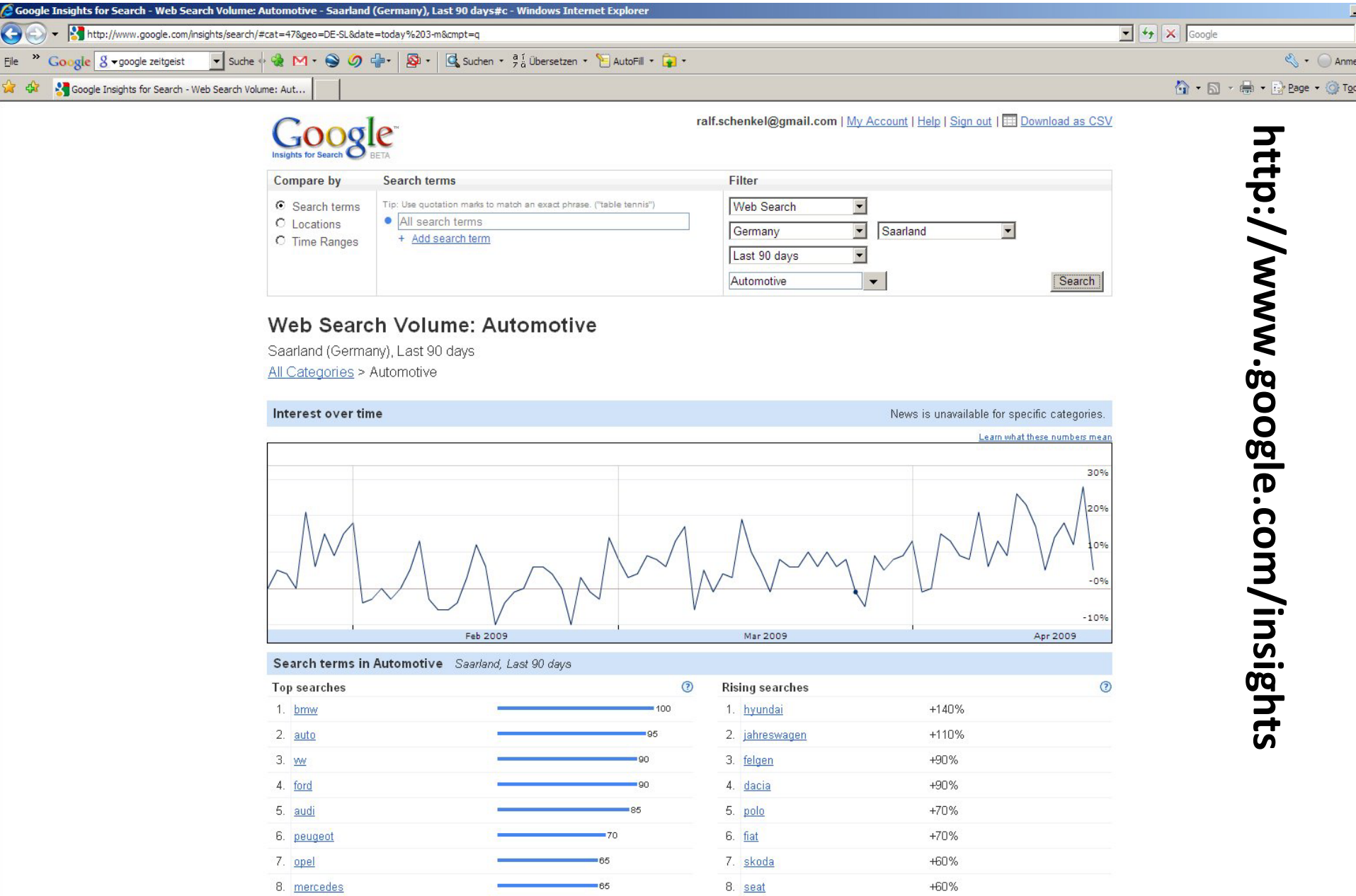
1.2 Application examples

Google Trends: Search stats



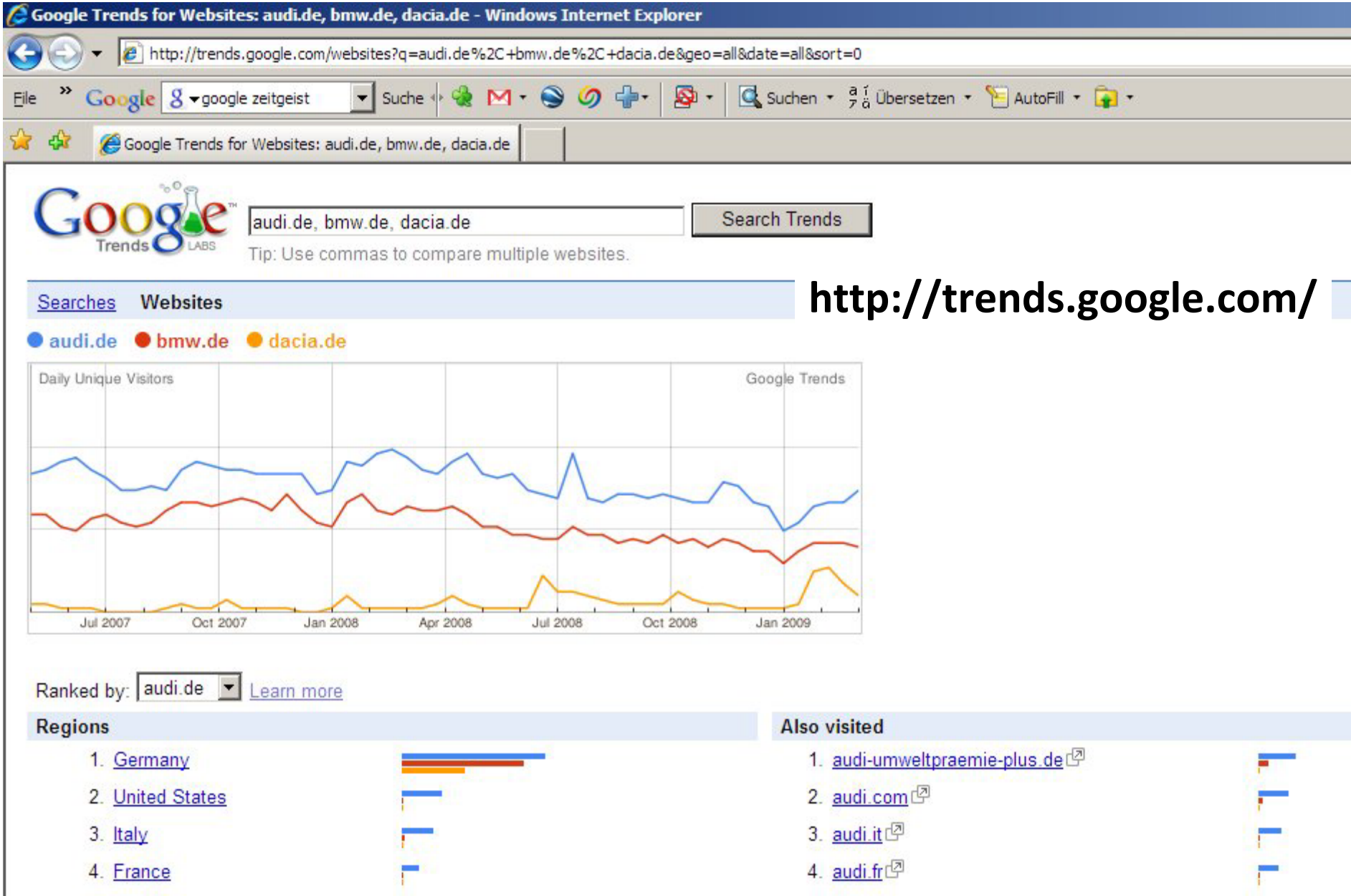
http://www.google.com/trends

Google insights: Trends in searches



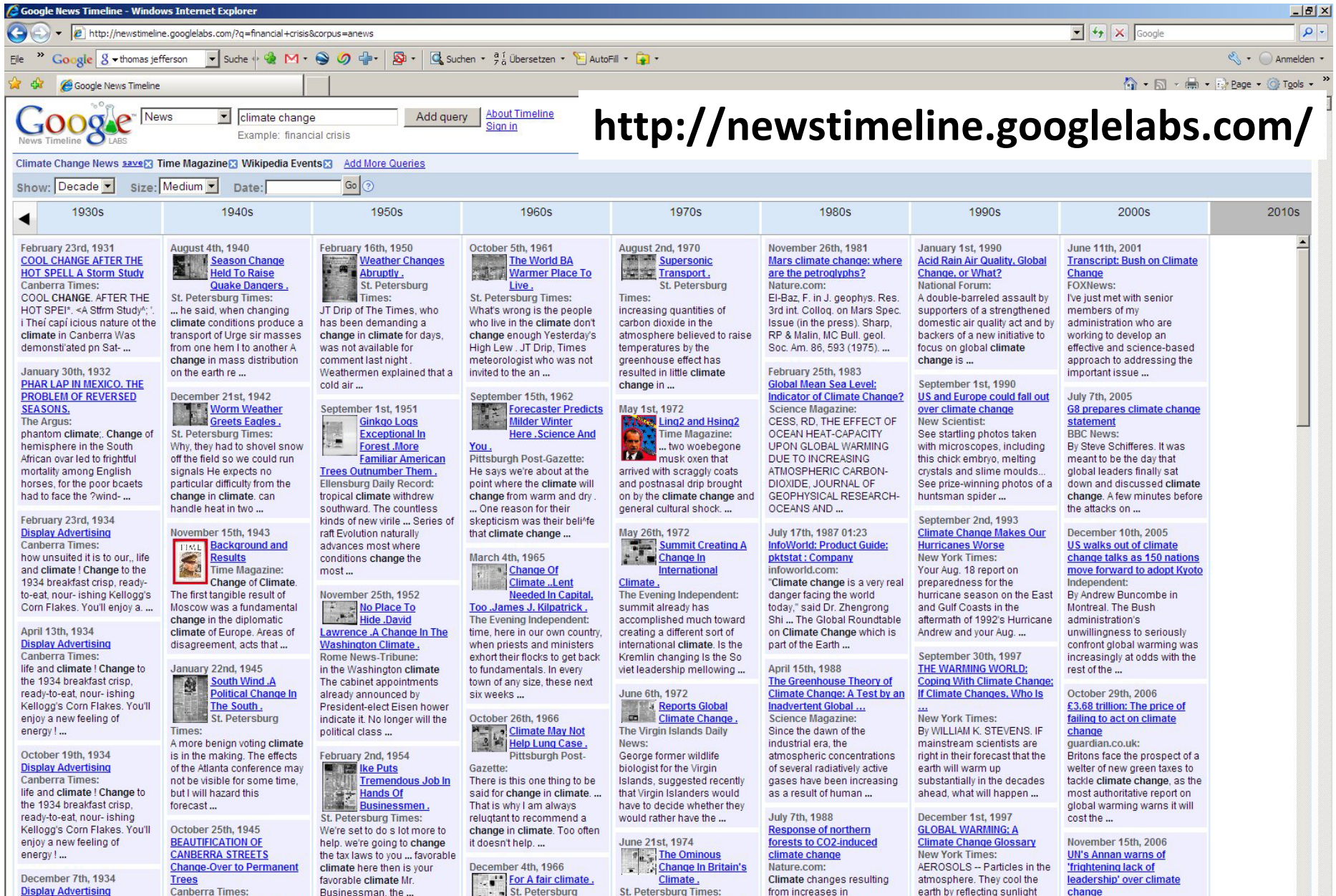
http://www.google.com/insights

Google Website trends: access stats

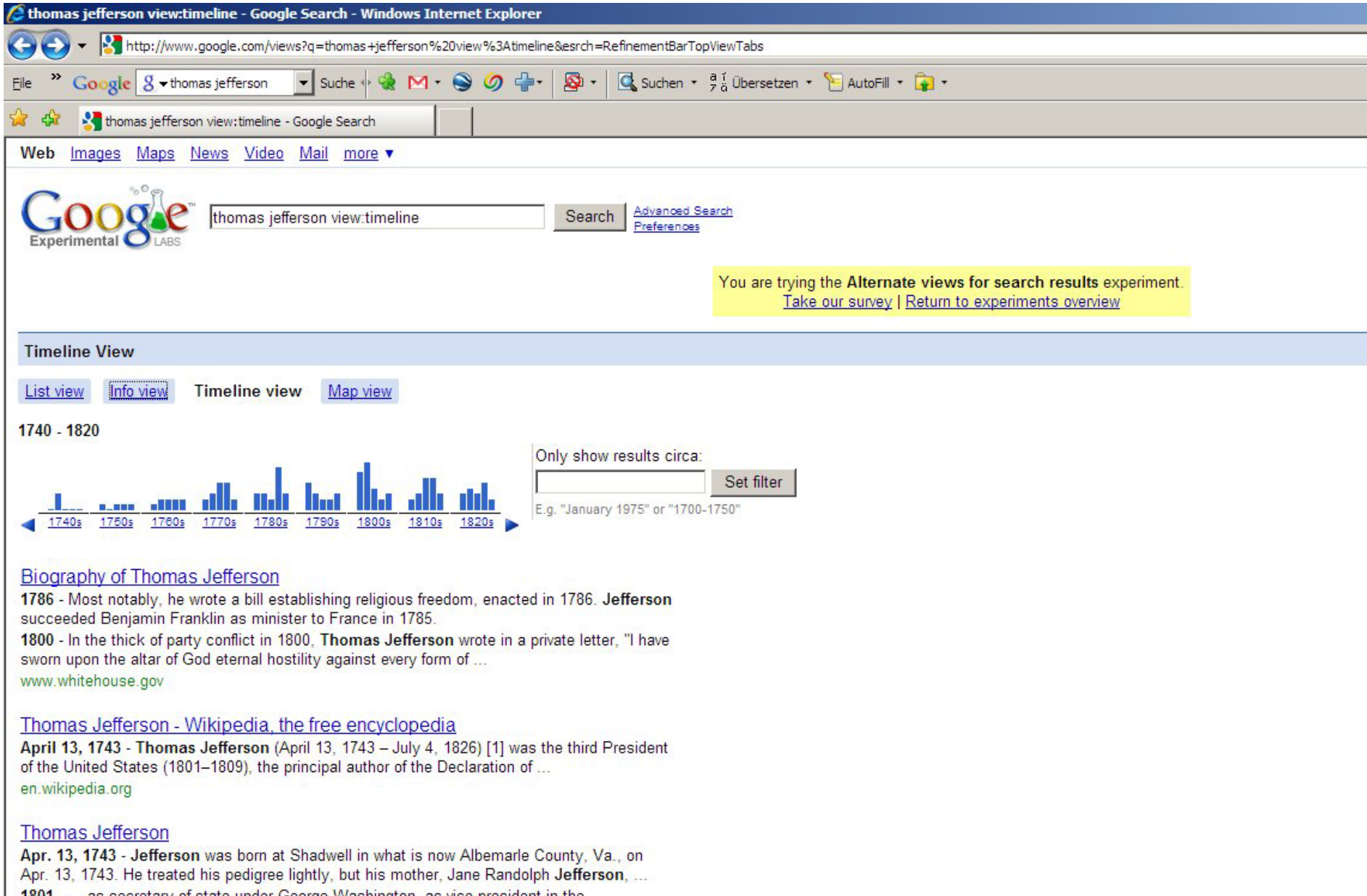


http://trends.google.com/

Google News Timeline: News trends



Google Web timeline: Date extraction



Google Zeitgeist: Frequent searches

indows Internet Explorer

google.com/intl/en/press/zeitgeist2008/politics.html

Google Zeitgeist 2008

Zeitgeist Home | Google.com

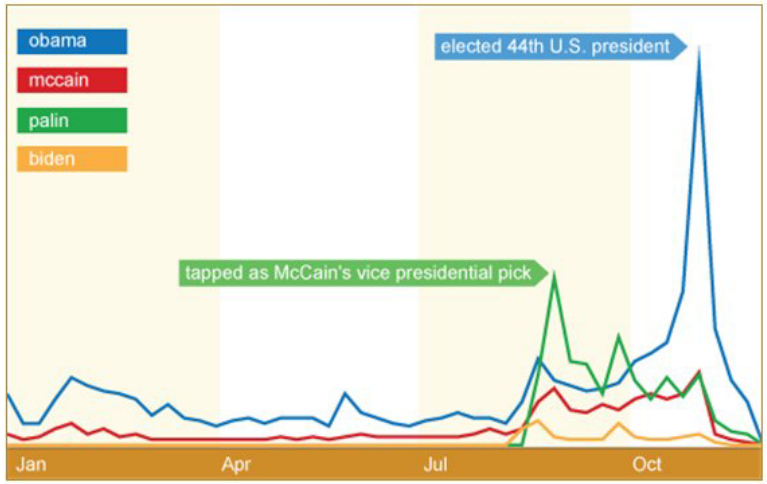
- Home
- US Zeitgeist
- Top of Mind
- Politics
- Trendsetters
- Showbiz
- Sports
- Around the World

Politics

With no shortage of groundbreaking moments and milestones, the 2008 U.S. presidential campaign is one for the history books.

All of these searches had the largest volume on Google.com in the U.S.

The Race to the Finish Line



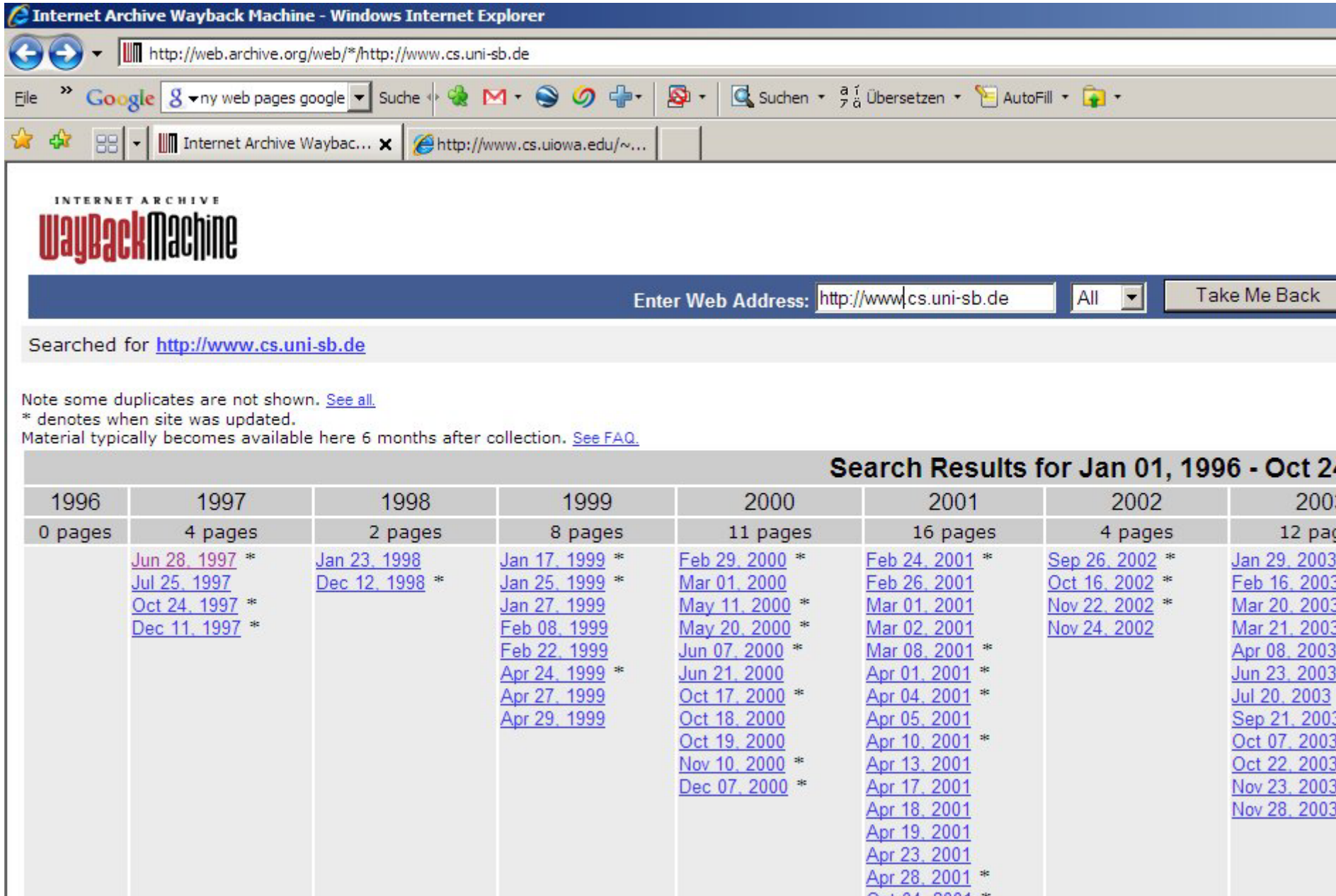
Top Sources for Political News

1. fox
2. cnn
3. abc
4. cbs
5. msnbc
6. drudge report
7. new york times

Campaign Buzzwords

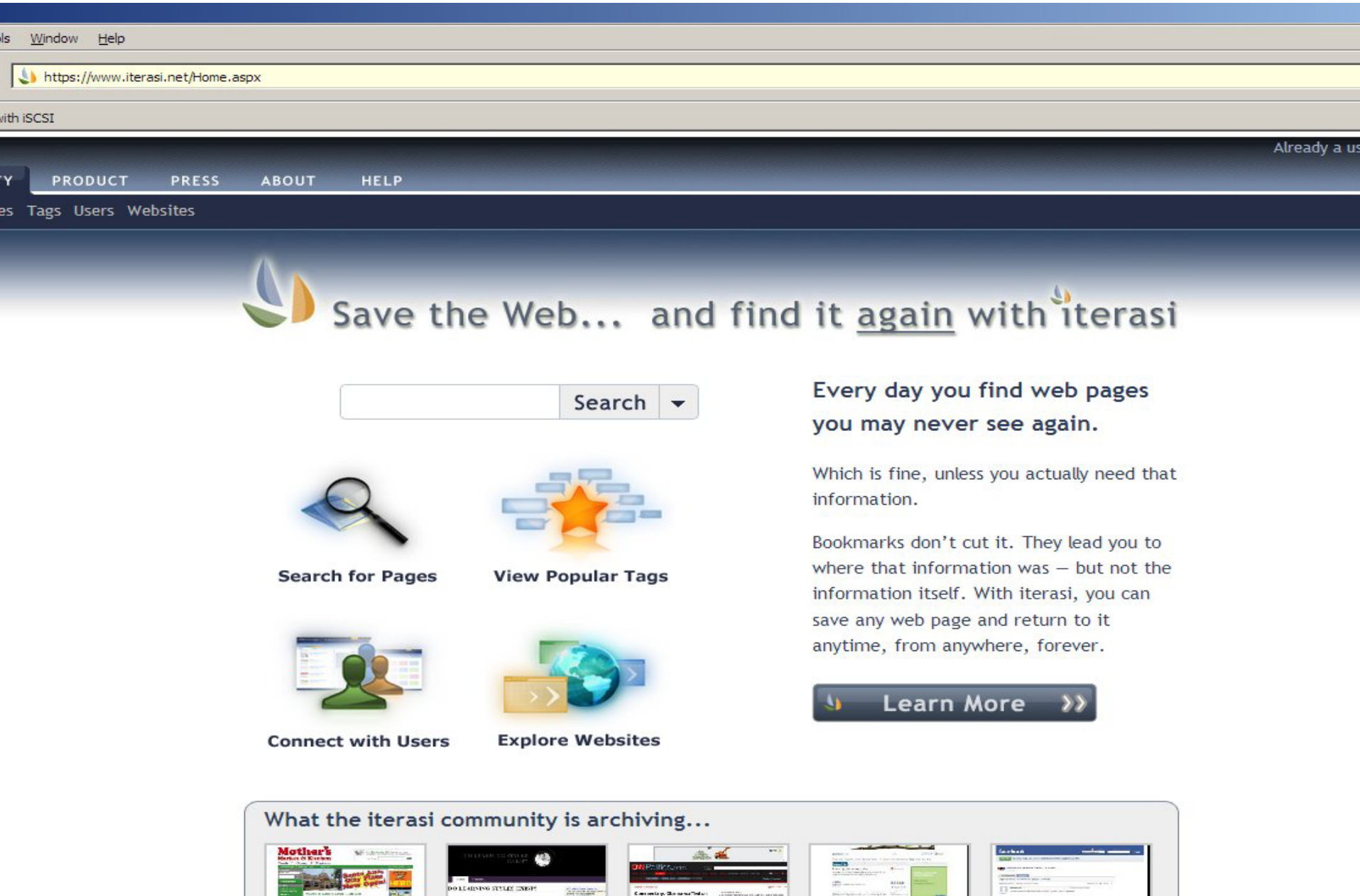
1. joe the plumber
2. jeremiah wright
3. maverick
4. william ayers
5. bridge to nowhere
6. lipstick pig
7. yes we can
8. that one

Internet Archive: Wayback machine



Internet Archive: Wayback machine

More Web Archiving: Iterasi



References

- T. Bray: **Measuring the Web**, WWW Conference, 1996.
- K. Bharat, A. Broder: **A technique for measuring the relative size and overlap of public web search engines**, WWW Conference, 1998
- A. Gulli, A. Signorini: **The Indexable Web is more than 11.5 billion pages**, WWW Conference, 2005
- S. Lawrence and C. L. Giles: **Accessibility of information on the web**, *Nature*, 400:107–109, 1999
- J. Domenech et al.: **A user-focused evaluation of web prefetching algorithms**, *Computer Communications* 30:10, 2213-2224, 2007
- R. Sadre, B. Haverkort: **Changes in the Web from 2000 to 2007**, Workshop on Distributed Systems: Operations and Management, 2008
- K.M. Risvik, R. Michelsen: **Search engines and Web dynamics**, Computer Networks 39, 289—302, 2002
- Y. Ke et al.: **Web dynamics and their ramifications for the development of Web search engines**, Computer Networks 50, 1430-1447, 2006
- R. Baeza-Yates et al.: **Web structure, dynamics and page quality**, SPIRE Conference, 2002
- V.N. Padmanabhan, L. Qiu: **The content and access dynamics of a busy Web site: Findings and implications**, SIGCOMM conference, 2000
- L. Cherkasova, M. Karlsson: **Dynamics and evolution of Web sites: Analysis, metrics and design issues**, IEEE International Symposium on Computers and Communications, 2001
- J. Cho, H. Garcia-Molina: **Estimating frequency of change**, Transactions on Internet Technologies 3(3):256—290, 2003
- J. Cho, H. Garcia-Molina: **The evolution of the Web and implications for an incremental crawler**. VLDB Conference, 2000
- A. Ntoulas, J. Cho, C. Olston: **What's new on the Web? The Evolution of the Web from a search engine perspective**, WWW Conference, 2004