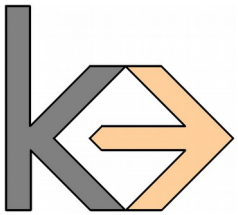
Web Mining – Data Mining im Internet

Vorlesung SS 2015

Johannes Fürnkranz
TU Darmstadt
Hochschulstrasse 10
D-64289 Darmstadt
06151/166238

juffi@ke.tu-darmstadt.de



General Information

- Web-page:
  - http://www.ke.tu-darmstadt.de/lehre/ss14/web-mining
- Text:
  - Soumen Chakrabarti: *Mining the Web – Discovering Knowlege from Hypertext Data*, Morgan Kaufmann Publishers 2003.
    - http://www.cse.iitb.ac.in/~soumen/mining-the-web/
    - readable online in http://books.google.de
  - Christopher D. Manning, P. Raghavan and H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press. 2008
    - complete book freely available at http://nlp.stanford.edu/IR-book/
  - Johannes Fürnkranz: *Web Mining.* The Data Mining and Knowledge Discovery Handbook, Springer-Verlag 2005.
    - Book chapter with many pointers to the literature
  - Various other articles available from the Web-page
- Lecture Slides:
  - available from course page (additional slides at book pages)

Übungen

- 5 Aufgaben
  - Programmierung ist notwendig
    - aber die Programme sind nur Mittel zum Zweck
  - ca. alle 2 Wochen eine Abgabe, Vorbesprechung 21.4.2015
    - Ausarbeitung der Lösungen
- Übungsstunden
  - Durchbesprechen der abgegebenen Lösungen
  - Jeder der abgibt, muß anwesend sein, und die Lösung vorführen können
- Beurteilung:
  - Bonuspunkte bei bestandener Klausur
  - Verbesserungen bis zu einem Notengrad sind möglich
- Gruppenarbeit möglich
  - Gruppengröße max. 3

Overview

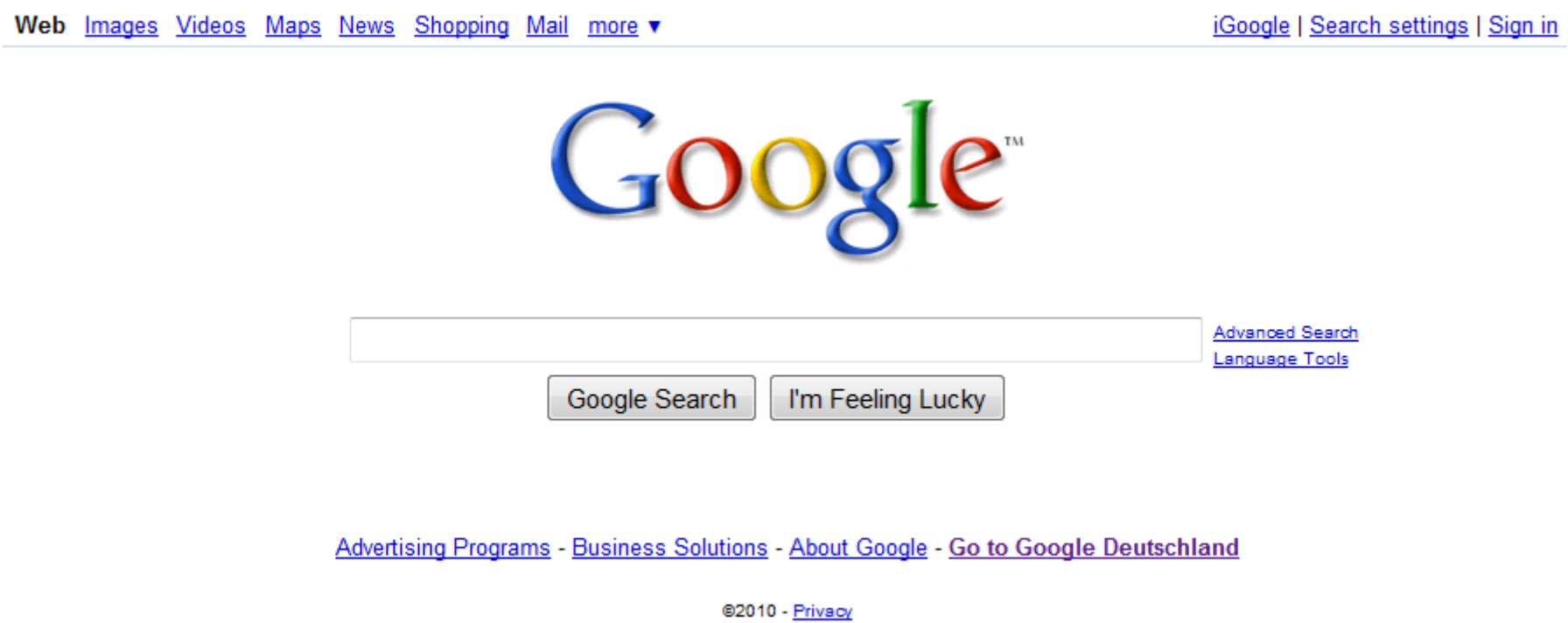
- Motivation
  - Automated citation indexing and analysis: Citeseer
  - Overview of Web Mining Tasks
- The Web
  - Hypertext
  - World-Wide Web
  - Problems
- Data Mining and Web Mining
  - Motivation: World-Wide Data Growth
  - Mining Structured vs. Unstructured Data

Motivation

- The Web is now 25 years old
  - ca. 1990, Tim Berners-Lee, CERN developed the first graphical hypertext browser
- The information on the Web has grown exponentially
  - on probably every topic you can think of, there is some information available on some Web page
- However, it is still very hard to find relevant information
  - The query interface to search engines has not changed since the early days of the Web!
  - Users have adapted to the interface instead of the other way around

Google 1998

Google 2010



Hard queries

- For many queries, the information that is needed to answer the query is readily available on the Web:
  - What are the cheapest hotels in Vienna's first district?
- The problems are
  - finding the pages that contain relevant information
    - pages of hotels in Vienna
  - extracting the relevant pieces of information from these pages
    - finding the prices, names, address of these hotels
  - connecting the information that is extracted from the pages
    - comparing the prices, sorting the hotels,
      filtering those that are not in the first district
  - apply common-sense reasoning in all phases
    - e.g., look for pages of bed & breakfast (Pension) as well
    - know about different currencies and conversions, etc.

Google

wer unterrichtet web mining in darmstadt Suche

Ca. 1.090 Ergebnisse (0,12 Sekunden)

Darmstadt ▾ Erweiterte Suche

Alles
Mehr

Das Web
Seiten auf Deutsch
Seiten aus Deutschland

Mehr Optionen

[PDF] The Semantic **Web**
Dateiformat: PDF/Adobe Acrobat - Schnellansicht
Johannes Fürnkranz, **Web Mining**. In O. Maimon and L. Rokach (eds.), ... z.B. nur Dozenten dürfen eine Vorlesung **unterrichten** ...
www.dvs.tu-**darmstadt**.de/teaching/dke/2010/vorlesung/semantic-**web**.pdf

D120.de/forum • Thema anzeigen - Kostenlos ins Theater für TU ...
10. Okt. 2009 ... Der Vertrag, den die Studierendenschaft der TU **Darmstadt** nun mit dem Staatstheater geschlossen .... sich aus allgemein zugänglichen Quellen ungehindert zu **unterrichten**. ... Seminar: Semantik im Automatischen Sprachverstehen, **Web Mining** ...
www.fachschaft.informatik.tu-**darmstadt**.de/.../viewtopic.php?... - Im Cache

Erste Hilfe - 15 Einträge - 25. Juni 2009
mit Fachabitur an die UNI - 15 Einträge - 31. Okt. 2008
Weitere Ergebnisse von fachschaft.informatik.tu-darmstadt.de »

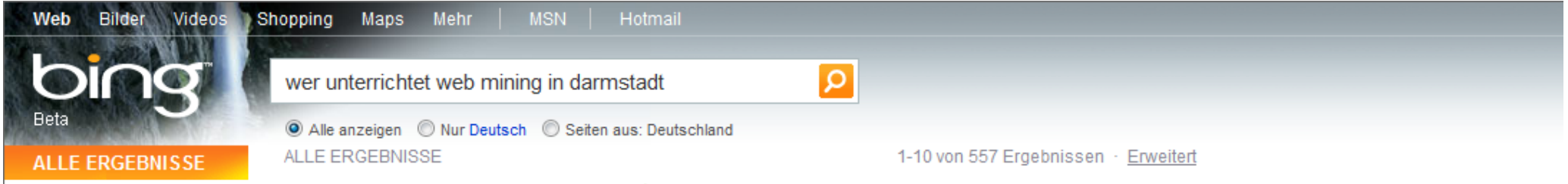
FREMDSPRACHENUNTERRICHT Seite 1
FREMDSPRACHENUNTERRICHT
DataEngine is a software tool for data **mining** in which fuzzy rule based systems, ... Institut **Web**-Galerie Auswärtige Kulturpolitik Daf Musik Kulturaustausch .... D-63322 RÖDERMARK, **DARMSTÄDTER** STR. 73. **FREMDSPRACHENUNTERRICHT** ...
web2.cylex.de/...-/l1cy1-d_ort1cy1--plz1cy1-_name1cy1-fremdsprachen**unterricht**-s1.html - Im Cache - Ähnlich

ULB **Darmstadt**
Kay Hoeksema: **Unterrichten** mit Moodle / praktische Einführung in das E-Teaching / Kay ..... Ian H. Witten: Data **mining** / practical machine learning tools and techniques ..... intelligent systems from decision making to data **mining**, **web** ...
elib.tu-**darmstadt**.de/ulb/nel/neu-SQ-SU-2008.html - Im Cache

[PDF] Prozessorientierte Wirtschaftsdidaktik und Einsatz von ERP ...
Dateiformat: PDF/Adobe Acrobat
von ERP-Systemen im **Unterricht**". Mit der Tagung wurde in einem ..... Verfügbar unter: http://www.gbv.de/dms/hebis-**darmstadt**/toc/50949659.pdf. ..... nagement, über die Marketing-Enzyklopädie bis zu speziellen Planungs- und Analysetools (Data **Mining**). ... zung des World Wide **Web**, der Festnetz- und Mobiltelefonie. ...
www.opus.ub.uni-erlangen.de/.../Pongratz_Tramm_Wilbers_Band4_OPUS.pdf

Markus Weimer - Deutschland - E-Mail, Adresse, Telefonnummer und ...
weimo.de - Informatik der TU **Darmstadt** vorne im CHE . ... Fach Neues Testament **unterrichten**. ... 123people.com verweist auf Biographie-Einträge ... Research Machine learning Flickr Data **Mining** TU **Darmstadt** Tübingen Digital photography .... Unter dieser

anz



[Lesezeichen für Stefan Schwan]
Web Site-Adressen von hunderten Verlagen in Deutschland bei DINO ... PFFH **Darmstadt** Pforzheim Pirmasens Potsdam Rüsselsheim Ravensburg-Weingarten
www.fremdsprache-deutsch.de/linkliste/bookmark_stefan.htm · Zwischengespeicherte Seite

[Der Schockwellenreiter]
Denn **wer** ißt, wird stark und klug, holt vom Brunnen manchen Krug. Hör nicht auf das ... [Werkzeuge für Webworker] Paul Browning, University of Bristol: Through The **Web** (TTW ...
blog.schockwellenreiter.de/archiv_2003/12.html · Zwischengespeicherte Seite

[News Rückblick]
ASP .NET professional - Das unabhängige Magazin für **Web** ... April in **Darmstadt** sowie 29. April in Aachen. Quelle ... **Wer** sein Wissen rund um Software-Entwicklung einer ...
www.aspnet-professional.de/news.aspx · Zwischengespeicherte Seite

[die datenschleuder .]
... Chaosradio Podcasting 32 Das Metalab in Wien 36 FIFA WM 1984™ 39 Nerddaters 46 Musings on **web** ... die nicht wie andere Geschäftsf elder dem freien Spiel des Marktes überlassen **wer**- ...
chaosradio.ccc.de/media/ds/ds090.pdf · Zwischengespeicherte Seite

[Software Marktplatz: Marktübersicht Dienstleistungen ...]
Mexiko: international ausgerichtete Universität **unterrichtet** mit ... Ergebnisse der Data-**Mining**-Studie 2009 - Große ... Karlsruhe, 22.7.2008 - Ab sofort ist der abas-eB-**Web** ...
www.software-marktplatz.de/news_archiv.php · Zwischengespeicherte Seite

[Der Deutsche Bildungsserver auf einen Blick]
Weitere thematische Angebote in Internet und **Web** 2.0. 4. Werte im Kindesalter. 40 Jahre Sesamstraße (10.11.2009) 5. Einzelne Länder. 5. Elite-Universitäten in den USA: mögliches ...
www.dbs.schule.de/toplist.html · Zwischengespeicherte Seite

[Beats Biblionetz - Personen: Personen mit B]
Unterrichten mit Computerspielen; Lernplattformen in Schulen ... **Wer** sucht, kann auch verzweifeln (2007) Antonio M ... Collaborative Concept Mapping on the World Wide **Web**
beat.doebe.li/bibliothek/p_b.html · Zwischengespeicherte Seite

[Berlin Brandenburger Pflegetage]
am Elisabethenstift in **Darmstadt**, Als Lehrerin tätig ... Er **Unterrichtet** seit 24 Jahre an unterschiedliche ... Wießmeier, Leverkusen, Leske + Budrich 2000 **Wer** ist ...

Gesponserte Ergebnisse

Add your link immediately
add your website's or blog's url for free and see it immediately
www.addlinkfreenow.com

Internet in Darmstadt
Internet mit 32.000 kBit/s, Telefon und TV für 30,-€*. Jetzt bestellen!
www.unitymedia.de/internet

In Darmstadt
Größter Anzeigenmarkt in **Darmstadt**! Hier haben Anzeigen Erfolg - Gratis
Darmstadt.markt.de

Biannual Report
Ursprung und ihre Stellung im heutigen Stochastik-Unterricht (. Burkhard Kümmerer) ..... 25.05.07 Proof **mining** in fixed point theory. TU **Darmstadt**, Germany .... different **web**-sites for teaching and learning mathematics, 19.-23.02.2008 ...
www3.mathematik.tu-darmstadt.de/fileadmin/pdf-files/jahresbericht...

D120.de/forum • Thema anzeigen - mit Fachabitur an die UNI
ich frage aus reiner Interesse, warum man eigentlich an der TU **Darmstadt** mit dem ..... was an einem normalen Gymnasium (zumindest in BW) nicht **unterrichtet** wird, ... Seminar: Semantik im Automatischen Sprachverstehen, **Web Mining** ...
www.fachschaft.informatik.tu-darmstadt.de/forum/viewtopic.php?f=2...

D120.de/forum • Thema anzeigen - Erste Hilfe
Die Athene, Logo der TU **Darmstadt** .... Film gezeigt und ich denke, die meisten Leute, die ich **unterrichtet** habe, haben genug mitgenommen, um richtig helfen zu können. .... Seminar: Semantik im Automatischen Sprachverstehen, **Web Mining** ...
www.fachschaft.informatik.tu-darmstadt.de/forum/viewtopic.php?f=3...

Markus Weimer - Deutschland - E-Mail, Adresse, Telefonnummer und ...
weimo.de - Informatik der TU **Darmstadt** vorne im CHE . ... Fach Neues Testament unterrichten. ... 123people.com verweist auf Biographie-Einträge ... Research Machine learning Flickr Data **Mining** TU **Darmstadt** Tübingen Digital photography .... Unter dieser Sektion verweist 123people auf **Web**-Dokumente, beispielsweise im ...
www.123people.de/s/markus+weimer

Thomas Kunstmann - Pipl Profiles
Thomas Fehnl and Thomas Kunstmann **Darmstadt** University of . .... BibSonomy,University of Kassel,folksonomy,data **mining**,Wissensverarbeitung,UniversitÃƒÂ¤t ... Der EUROPATICKER Umweltruf und der EUROPATICKER Korruptionsreport **unterrichtet** stÃƒÂ¤ndig
mz-**web**.de - die Online-Plattform der Mitteldeutschen Zeitung

Verwandte Suchbegriffe

- Web Mining Techniques
- Web Mining Paper
- Basics of Web Mining
- Web Mining Software
- Web Content Mining
- Web Mining Tools
- Difference between Data Mining and Web Mining
- Web Structure Mining
- Deep Web Mining
- Abstract on Web Mining
- Introduction on Web Mining
- Web Usage Mining
- Data Mining
- Text Mining
- Data Mining Concepts

Mehr »

TextRunner Search (Experimental)

TextRunner took 5 seconds.

Retrieved **0** results for **Who teaches web mining in darmstadt**.

Grouping results by argument 1. Group by: predicate | argument 2

TextRunner Search (Experimental)

TextRunner took 6 seconds.

Retrieved **53** results for **Who invented the light bulb**.

Grouping results by argument 2. Group by: argument 1 | predicate

the light bulb - 11 results

Thomas Edison (299), Thomas Alva Edison (14), Thomas A. Edison (11), ***31 more...*** **invented** the **light bulb**
man (13), Thomas Edison (7), guy (6), ***4 more...*** who **invented** the **light bulb**
Edison (27), Thomas Alva Edison (2) did n't **invent** the **light bulb**
Edison (8) had **invented** the **light bulb**
Edison (5) may have **invented** the **light bulb**
Thomas Edison (4) would have **invented** the **light bulb**
Thomas Edison (4) failed **invented** the **light bulb**
first person (2) to **invent** the electric **light bulb**
Edison (2) had n't **invented** the **light bulb**
Edison (2) could have **invented** the **light bulb**
Leonardo da Vinci not (2) producing not **inventing** the **light bulb**

the incandescent light bulb and phonograph - 1 result

Thomas Edison (2) **invented** the incandescent **light bulb** and phonograph

9,999 light bulbs - 1 result

Mr Edison (2) **invented** 9,999 **light bulbs**

Search again:

Who invented the light bulb
Search

Jump to:

the **light bulb** (11)
the incandescent **light bulb** and phonograph (1)
9,999 **light bulbs** (1)

TextRunner Search (Experimental)

TextRunner took 10 seconds.

Retrieved **37** results for **Who is the chancellor of germany**.

Grouping results by argument 2. Group by: argument 1 | predicate

Chancellor of Germany - 10 results

Hitler (90), Nazi leader Adolf Hitler (4), 75th anniversary of the date (3), Herr Von Papen (2) was appointed **Chancellor of Germany**
Adolf Hitler (32), Angela Merkel (12), Bismarck (5), ***8 more...*** was **Chancellor of Germany**
Hitler (31), Angela Merkel (8), Gerhard Schroeder (2) was elected **Chancellor of Germany**
Adolf Hitler (33), Day (2) was named **Chancellor of Germany**
Hitler (10) was made **Chancellor of Germany**
Hitler (3) was appointed as the **Chancellor of Germany**
Hitler (3) had been appointed **chancellor of Germany**
German politician (2) who was the **Chancellor of Germany**
Helmut Kohl (2) may be **chancellor of Germany**
Hitler (2) was nominated the **chancellor of Germany**

Chancellor of West Germany - 2 results

Kurt Georg Kiesinger (4), Schmidt (3), Helmut Kohl (2), Adenauer (2) was **Chancellor of** West **Germany**
Kurt Georg Kiesinger (4) **is** elected **Chancellor of** West **Germany**

the first Chancellor of the Federal Republic of Germany - 1 result

Konrad Adenauer (3), Helmut Schmidt (2) was the first **Chancellor of** the Federal Republic **of Germany**

Supreme Chancellor of Germany - 1 result

Hitler (2) was named Supreme **Chancellor of Germany**

Search again:

Who is the chancellor of austria

Search

Jump to:

Chancellor of Germany (10)
Chancellor of West Germany (2)
the first Chancellor of the Federal Republic of Germany (1)
Supreme Chancellor of Germany (1)
CDU) and Chancellor of West Germany (1)
the first Chancellor of Germany of non-noble background (1)
Chancellor Merkel of Germany (1)

Example Application: Citeseer

- Citeseer is a very popular search engine for publications in Computer Science
  - http://citeseer.ist.psu.edu/
- It provides
  - keyword search for articles
  - on-line access to the articles
  - pointers to articles that the articles cites
  - pointers to articles that cite an article
  - pointers to related articles
  - identification of important papers (citation analysis)
  - identification of important publication media
- All of that is generated automatically!


CiteSeer
Scientific Literature Digital Library
Find:
web mining
Documents
Citations

Searching for **PHRASE web mining**.
Restrict to: Author Title Order by: Expected citations Date Hits: 100 Try: Google (CiteSeer) Google (Web) Yahoo! MSN CSB DBLP
596 citations found. Retrieving citations...

Context Doc ***52*** *(9)*: Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava. ***Web mining****: Information and pattern discovery on the world wide web*. In ICTAI'97, Dec. 1997.

Looking for an author? You may be seeing only a fraction of all citations. Try: web w/2 mining or w w/2 mining (w/2 means within 2 words)

Context Doc ***34*** *(1)*: B. Mobasher, N. Jain, E-H. Han, and J. Srivastava "***Web mining****: Pattern discovery from world wide* ***web*** *transactions*, " Technical Report 96-050, University of Minnesota, Sep, 1996.

Context Doc ***34*** *(1)*: R. Kosala and H. Blockeel, "***Web Mining*** *Research: A Survey*," in SIGKDD: SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data **Mining**, ACM, ACM Press, 2000, pp. 1--15.

Context Doc ***14*** *(11)*: O. Nasraoui, H. Frigui, R. Krishnapuram, and A. Joshi. ***Mining web*** *access logs using relational competitive fuzzy clustering*. In Eighth International Fuzzy Systems Association Congress, Hsinchu, Taiwan, Aug. 1999.

Context Doc ***14*** *(8)*: M. Craven, S. Slattery, and K. Nigam. *First-order learning for* ***Web mining***. In C. Ndellec and C. Rouveirol, editors, Proceedings of the 10th European Conference on Machine Learning (ECML-98), pages 250--255, Chemnitz, Germany, 1998. Springer-Verlag.

Context Doc ***12*** *(0)*: Karuna P. Joshi, Anupam Joshi, Yelena Yesha, Raghu Krishnapuram, "*Warehousing and* ***Mining Web*** *Logs*", Proc. of 2nd Workshop on **Web** Information and Data Management (WIDM'99) (in conj. with CIKM '99), Kansas City (November 1999.

Context Doc ***10*** *(3)*: A. Banerjee and J. Ghosh. *Clickstream Clustering Using Weighted Longest Common Subsequences*. In Proceedings of the **Web Mining** Workshop at the 1st SIAM Conf. on Data **Mining**, pages 34--40, Chicago, IL, April 2001.

Context Doc ***10*** *(0)*: B. Sarwar, G. Karypsis, J.A. Konstan, and J.T. Riedl. *Application of Dimensionality Reduction in Recommender System -- A Case Study*. In ACM WebKDD 2000 **Web Mining** for E-Commerce Workshop.

Context Doc ***10*** *(0)*: M. Mulvenna, S. Anand, and A. Buchner. *Personalization on the net using* ***web mining***. CACM, 43(8):122--125, 2000.

Context Doc ***9*** *(2)*: Myra Spiliopoulou. *The laborious way from data* ***mining*** *to* ***web mining***. submitted, June 1998.

citation counts

Web Mining: Information and Pattern Discovery on the World Wide Web (1997) (Make Corrections) (82 citations)

R. Cooley, B. Mobasher, J. Srivastava

CiteSeer Scientific Literature Digital Library — Home/Search Context Related

View or download:
depaul.edu/~mobashe...webminertai97.ps
depaul.edu/~mobasher/WebKl...cmstai.ps
depaul.edu/~mobasher/clas...cmstai.pdf
Cached: PS.gz PS PDF Image Update Help

From: depaul.edu/~mobasher/pubs (more)
Homepages: R.Cooley HPSearch (Update Links)

(Enter summary)

Rate this article: 1 2 3 4 5 (best)
Comment on this article

Abstract: Application of data mining techniques to the W
focus of several recent research projects and papers. Ho
term Web mining has been used in two distinct ways. Th
across the World Wide Web. The second, called Web u

Cited by: More
WUM: A Tool for Web Utilization Analysis - Myra Spiliop
P-Jigsaw: Extending Jigsaw with Rules Assisted Cache
Combining Web Usage Mining and Fuzzy Inference for W

Similar documents (at the sentence level):
8.5%: Mir: A Tool For Visual Presentation Of Web Acc
5.5%: Web Mining: Pattern Discovery from World Wide

Active bibliography (related documents): More All
0.7: Grouping Web Page References into Transactions
0.5: Document Categorization and Query Generation o
0.5: Software Environments in Support of Wide-Area D

Similar documents based on text: More All
0.8: Some Experiences on Large Scale Web Mining - I
0.7: Blockmodeling Techniques for Web Mining - Schoi
0.6: Usage Mining for and on the Semantic Web - Stun

Related documents from co-citation: More All
25: Data preparation for mining world wide web browsin
24: Fast Algorithms for Mining Association Rules - Agr
20: From user access patterns to dynamic hypertext li

BibTeX entry: (Update)

Web Mining: Information and Pattern Discovery on the World Wide Web *

R. Cooley, B. Mobasher, and J. Srivastava
Department of Computer Science and Engineering
University of Minnesota
Minneapolis, MN 55455, USA

Abstract

Application of data mining techniques to the World Wide Web, referred to as Web mining, has been the focus of several recent research projects and papers. However, there is no established vocabulary, leading to confusion when comparing research efforts. The term Web mining has been used in two distinct ways. The first, called Web content mining in this paper, is the process of information discovery from sources across the World Wide Web. The second, called Web usage mining, is the process of mining for user browsing and access patterns. In this paper we define Web mining and present an overview of the various research issues, techniques, and development efforts. We briefly describe WEBMINER, a system for Web usage mining, and conclude this paper by listing research issues.

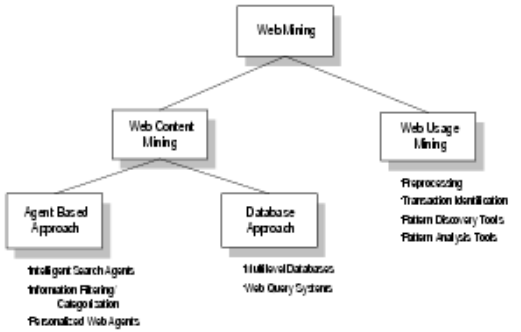


Figure 1: Taxonomy of Web Mining

1 Introduction

With the explosive growth of information sources available on the World Wide Web, it has become increasingly necessary for users to utilize automated tools in find the desired information resources, and to track and analyze their usage patterns. These factors

context. There are several important issues, unique to the Web paradigm, that come into play if sophisticated types of analyses are to be done on server side data collections. These include integrating various data sources such as server access logs, referrer logs, user registration or profile information; resolving difficulties in the identification of users due to missing unique key attributes in collected data; and the importance of identifying user sessions or transactions

82 citations found. Retrieving documents...

Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava. *Web mining: Information and* Dec. 1997.

CiteSeer Home/Search Document Details and Download Summary Related Article

This paper is cited in the following contexts:

First 50 documents Next 50

Low-Complexity Fuzzy Relational Clustering - Algorithms For Web (Correct)

....In particular, Han et al. 36] create a MOLAP based warehouse from Web logs, and allow users tr
time dependent patterns in the acces
[9] **[10].** However, both these approache
is used and the clients are willing to rele

However, it is not clear how the simi
clusters. There is also a recent body
structured, database-like entity. In pa
Web logs, and allow users to perfor
patterns in the access logs [53]. Sim
[1], have been proposed in [9], [10].
ids, which is not true in the real wo
the clients are willing to release the
momentum is the idea that we can le
their *clickstreams*, which is of great i

An important component of persc
extraction of structure from unlabele
information. The logs kept by Web s
be viewed as a special case of the m
can be said to have three operations

Low-Complexity Fuzzy Relational Clustering Algorithms for Web Mining

Raghu Krishnapuram
IBM India Research Lab
Indian Institute of Technology, Hauz Khas, New Delhi 110016
kraghura@in.ibm.edu
On leave from Dept of Mathematical and Computer Sciences, Colorado School of Mines, Golden, CO 80401
Anupam Joshi
Department of Computer Science and Electrical Engineering
University of Maryland Baltimore County, Baltimore, MD 21250
joshi@cs.umbc.edu
Olfa Nasraoui
Department of Electrical Engineering
University of Memphis, Memphis, TN 38152
Liyu Yi

References

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proceedings of the 20th VLDB Conference*, pages 487–499, Santiago, Chile, 1994.

[2] R. Armstrong, T. Joachims D. Freitag, and T. Mitchell. Webwatcher: A learning apprentice for the World Wide Web. In *Proceedings of the AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments*, pages 6–13, Stanford, CA, March 1995.

[3] G. Arocena and A. Mendelz. Weboql: Restructuring documents, databases, and web. In *Proc. IEEE Intl. Conf. Data Engineering '98*, pages 24–33. IEEE Press, 1998.

[4] P. Bajcsy and N. Ahuja. Location- and density-based hierarchical clustering using similarity analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:1011–1015, 1998.

[5] G. Beni and X. Liu. A least biased fuzzy clustering method. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 16:954–960, September 1994.

[6] J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.

[7] J. Abidi C. Shahabi, A.M. Zarkesh and V. Shah. Knowledge discovery from users web-page navigation. In *Proceedings of the Seventh IEEE Intl. Workshop on Research Issues in Data Engineering (RIDE)*, pages 20–29, Birmingham, UK, 1997.

[8] J. Chen, A. Mikulcic, and D. H. Kraft. An integrated approach to information retrieval with fuzzy clustering and fuzzy inferencing. In O. Pons, M. Ampara Vila, and J. Kacprzyk, editors, *Knowledge Management in Fuzzy Databases*, volume 163. Physica Verlag, Heidelberg, Germany, 2000.

[9] M.S. Chen, J.-S. Park, and P. S. Yu. Efficient data mining for path traversal patterns. *IEEE Trans. Knowledge and Data Engineering*, 10(2):209–221, April 1998.

[10] R. Cooley, B. Mobasher, and J. Srivastav. Web Mining: Information and pattern discovery on the World Wide Web. In *Proc. IEEE Intl. Conf. Tools with AI*, pages 558–567, Newport Beach, CA, 1997.

[11] R. N. Davé and R. Krishnapuram. Robust clustering methods: A unified view. *IEEE Transactions on Fuzzy Systems*, 5(2):270–293, 1997.

[12] E. Diday. La methode des nuees dynamiques. *Rev. Stat. Appliquee*, XIX(2):19–34, 1975.

[13] D. Riecken: Guest Editor. Special issue on personalization. *Communications of the ACM*, 43(9), Sept. 2000.

[14] J. Fink, A. Kobsa, and J. Schreck. Personalized hypermedia information provision through adaptive and adaptable system features. *http://zeus.gmd.de/hci/projects/avanti/publications/ISandN97/ISandN97.html*, 1997.

[15] K. S. Fu. *Syntactic Pattern Recognition and Applications*. Academic Press, San Diego, CA, 1982.

[16] K. C. Gowda and E. Diday. Symbolic clustering using a new similarity measure. *IEEE Transactions on Systems, Man, and Cybernetics*, 20:368–377, 1992.

[17] S. Guha, R. Rastogi, and K. Shim. CURE: An efficient algorithm for large databases. In *Proceedings of SIGMOD '98*, pages 73–84, Seattle, June 1998.

[18] R. J. Hathaway and J. C. Bezdek. Switching regression models and fuzzy clustering. *IEEE Transactions on Fuzzy Systems*, 1(3):195–204, 1993.

Citations (may not include all citations):

866 Fast algorithms for mining association rules - Agrawal, Srikant - 1994
359 Data cube: A relational aggregation operator generalizing gr.. - Gray, Bosworth et al. - 199
321 A query language and optimization techniques for unstructure.. - Buneman, Davidson et al
262 Finding Groups in Data: an Introduction to Cluster Analysis (context) - Kaufman, Rousseeu
239 Efficient and effective clustering method for spatial data m.. - Ng, Han - 1994
236 Implementing data cubes efficiently - Harinarayan, Rajaraman et al. - 1996
235 Information Retrieval Data Structures and Algorithms (context) - Frakes, Baeza-Yates - 19
198 Webwatcher: A learning apprentice for the world wide web - Armstrong, Freitag et al. - 199
183 Discovering frequent episodes in sequences (context) - Mannila, Toivonen et al. - 1995
174 word of mouth (context) - Shardanand, Maes et al. - 1995
169 A scalable comparison shopping agent for the world wide web - Doorenbos, Etzioni et al. -
164 the computation of multidimensional aggregates - Agrawal, Agrawal et al. - 1996
162 An efficient algorithm for mining association rules in large.. (context) - Savasere, Omiecins
154 Mining sequential patterns: Generalizations and performance .. - Srikant, Agrawal - 1996
144 Wq query system world wide web - Shmueli, system et al. - 1995
116 A declarative language for querying and restructuring the we.. - Lakshmanan, Sadri et al. -
114 Syntactic clustering of the web (context) - Broder, Glassman et al. - 1997
114 Data-driven discovery of quantitative rules in relational da.. (context) - Han, Cai et al. - 199
113 webert: Identifying interesting web sites (context) - Pazzani, Muramatsu et al. - 1996
107 Silk from a sow's ear: Extracting usable structures from the.. - Pirolli, Pitkow et al. - 1996
100 Querying semistructured heterogeneous information - Quass, Rajaraman et al. - 1995
99 Computer Systems that Learn: Classification and Prediction M.. (context) - Weiss, Kulikow
89 Planning to gather information - Kwok, Weld - 1996
87 The information manifold - Kirk, Levy et al. - 1995
82 Web mining: Information and pattern discovery on the world w.. - Cooley, Mobasher et al. -
71 Parasite: mining structural information on the web (context) - Spertus - 1997
64 Dmql: A data mining query language for relational databases - Han, Fu et al. - 1996
53 Category translation: learning to understand information on .. - Perkowitz, Etzioni - 1995
53 Storage estimation for multidimensional aggregates in the pr.. - Shukla, Deshpande et al. -
50 Hypursuit: a hierarchical network search engine that exploit.. (context) - Weiss, Velez et al.
45 The tsimmis project: Integration of heterogenous information.. (context) - Chawathe, Garcia-
42 Semistructured and structured data in the web: Going back an.. - Merialdo, Atzeni et al. - 1
42 Aliweb - archie-like indexing in the web (context) - Koster - 1994
41 Web mining: Pattern discovery from world wide web transactio.. - Mobasher, Jain et al. - 19
36 Data mining for path traversal patterns in a web environment - Chen, Park et al. - 1996
28 An adaptive agent for automated web browsing - Balabanovic, Shoham et al. - 1995
22 Finding salient features for personal web page categorizatio.. - Wulfekuhler, Punch - 1997
22 Faq-finder: A case-based approach to knowledge navigation (context) - Hammond, Burke et
21 Automatically organizing bookmarks per content (context) - Maarek, Shaul - 1996

References

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. of the 20th VLDB Conference*, pages 487–499, Santiago, Chile, 1994.

[2] S. Agrawal, R. Agrawal, P.M. Deshpande, A. Gupta, J. Naughton, R. Ramakrishna, and S. Sarawagi. On the computation of multidimensional aggregates. In *Proc. of the 22nd VLDB Conference*, pages 506–521, Mumbai, India, 1996.

[3] R. Armstrong, D. Freitag, T. Joachims, and T. Mitchell. Webwatcher: A learning apprentice for the world wide web. In *Proc. AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments*. 1995.

[4] M. Balabanovic, Yoav Shoham, and Y. Yun. An adaptive agent for automated web browsing. *Journal of Visual Communication and Image Representation*, 6(4), 1995.

[5] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G Zweig. Syntactic clustering of the web. In *Proc. of 6th International World Wide Web Conference*, 1997.

[6] C. M. Brown, B. B. Danzig, D. Hardy, U. Manber, and M. F. Schwartz. The harvest information discovery and access system. In *Proc. 2nd International World Wide Web Conference*, 1994.

[7] P. Buneman, S. Davidson, G. Hillebrand, and D. Suciu. A query language and optimization techniques for unstructured data. In *Proc. of 1996 ACM-SIGMOD Int. Conf. on Management of Data*, 1996.

[8] P. Buneman, S. Davidson, and D. Suciu. Programming constrcuts for unstructured data. In *Proceedings of ICDT'95, Gubbio, Italy*, 1995.

[9] C. Chang and C. Hsu. Customizable multi-engine search tool with clustering. In *Proc. of 6th International World Wide Web Conference*, 1997.

CiteSeer.IST Home **Check:** The following citations are predicted to all refer to the same paper. Details

COOLEY, R., SRIVASTAVA, J., MOBASHER, B., *Web Mining: Information and Pattern Discovery on the World Wide Web*, Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), November 1997.

Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava. *Web mining: Information and pattern discovery on the world wide web.* In ICTAI'97, Dec. 1997.

R. Cooley, B. Mobasher, and J. Srivastava. *Web mining: Information and pattern discovery on the world wide web.* Technical Report TR 97-027, University of Minnesota, Dept. of Computer Science, Minneapolis, 1997.

Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava. *Web mining: Information and pattern discovery on the world wide web.* In ICTAI'97, Dec. 1997.

Cooley, R., Mobasher, R. & Srivastava, J. (1997) *Web Mining: Information and Pattern Discovery on the World Wide Web*, Proc. 9 th IEEE Int'l Conf. on Tools with Artificial Intelligence.

Cooley, R., Mobasher, B., and Srivastava, J. (1997b). *Web mining: Information and pattern discovery on the world wide web.* In ICTAI'97.

R. Cooley, B. Mobasher, and J. Srivastava, "*Web mining: Information and Pattern discovery on the World Wide Web*," Proc. IEEE Intl. Conf. Tools with AI, Dec, 1997.

R. Cooley, B. Mobasher, and J. Srivastava. *Web mining: Information and patterns discovery on the world wide web.* In Proc. of the 9th IEEE Int. Conf. on Tools with Artificial Intelligence, pages 558--567, 1997.

R. Cooley, B. Mobasher and J. Srivsatava. *Web Mining: Information and Pattern Discovery on the Word Wide Web.* Technical Report TR 97-027, University of Minnesota, Dept. of Computer Science, Minneapolis, 1997.

R. Cooley, B. Mobasher, and J. Srivastava. *Web mining: Information and pattern discovery on the world wide web.* In International Conference on Tools with Artificial Intelligence, pages 558-567, Newport Beach, 1997. IEEE.

Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava. *Web mining: Information and patterns discovery on the world wide web.* In Proc. of the ninth IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), November 1997.

R. Cooley, B. Mobasher, and J. Srivastava. *Web mining: Information and pattern discovery on the world wide web.* In International Conference on Tools for Articial Intelligence, Newport Beach, CA, November 1997.

Cooley, R., Mobasher, B., and Srivastava, J. (1997). *Web mining: Information and pattern discovery on the world wide web.* In International Conference on Tools for Articial Intelligence, Newport Beach, CA.

Most cited articles in Computer Science - September 2006 (CiteSeer.Continuity)

Generated from documents in the CiteSeer.Continuity database. This list does not include citations where one or more authors of the citing and cited articles match. This list is automatically generated and may contain errors. The list is generated in batch mode and citation counts may differ from those currently in the CiteSeer.Continuity database, because the database is continuously updated.

All Years 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006

Next 200

1. Doc Context *4137* [GJ79] M.R. Garey and D.S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. Freeman, New York, 1979.

2. Book Context *3803* [12] Thomas H. Cormen, Charles E. Leiserson, and Ronald L. Rivest. *Introduction to algorithms*. The MIT Press, 1991.

3. Doc Context *2697* [25] C.A.R. Hoare, *Communicating Sequential Processes*, Prentice-Hall International, 1985.

4. Doc Context *2321* 3. A.P. Dempster, N.M. Laird, and D.B. Rubin. *Maximum Likelihood from Incomplete Data via the EM Algorithm*. Journal of the Royal Statistical Society, Series B (Methodological) , 39(1):1--38, 1977.

5. Doc Context *2220* Cover, T. and Thomas, J. (1991). *Elements of Information Theory*. John Wiley & Sons, New York, NY.

6. Doc Context *2112* [15] E. Gamma, R. Helm, R. Johnson, and J. Vlissides. *Design Patterns: Elements of Reusable ObjectOriented Software*. Addison-Wesley, Reading, Massachusetts, 1995.

7. Book Context *2064* [Gol89] David E. Goldberg. *Genetic Algorithms in Search*, Optimization, and Machine Learning. Addison-Wesley, Reading, Massachusetts, 1989.

8. Doc Context *2044* Quinlan, J. R. (1993). C4.5: *Programs for machine learning*. San Mateo, CA: Morgan Kaufmann.

9. Doc Context *2013* Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis*. New York, NY: Wiley.

10. Book Context *1932* 13. Knuth D (1973) *The art of computer programming*, Vol. 3: sorting and searching. Addison-Wesley, Reading, Mass.

11. Book Context *1905* [33] R. Milner. *Communication and Concurrency*. Prentice Hall, New York, 1989.

12. Book Context *1899* [8] J. Holland. 1975. *Adaptation in Natural and Artificial Systems*. MIT Press.

13. Doc Context *1882* [4] John Hopcroft and Jeffrey Ullman. *Introduction to Automata Theory*, Languages, and Computation. Addison Wesley, 1979.

CiteSeer$^{x}_{\beta}$

Most Cited Articles | Most Cited Citations | Most Cited Authors | Venue Impact Ratings

Most Cited Computer Science Citations

This list is generated from documents in the CiteSeer^{x} database as of September 18, 2011. This list is automatically generated and may contain errors. The list is generated in batch mode and citation counts may differ from those currently in the CiteSeer^{x} database, since the database is continuously updated.

All Years | 1990 | 1991 | 1992 | 1993 | 1994 | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011

1. M R Garey, D S Johnson
   Computers and Intractability: A Guide to the Theory of NP-completeness 1979
   8103
2. T H Cormen, C E Leiserson, R L Rivest
   Introduction to Algorithms 1990
   6748
3. V Vapnik
   Statistical Learning Theory 1998
   5973
4. D Goldberg
   Genetic Algorithm in Search, Optimization and Machine Learning, 1989
   5658
5. T M Cover, J M Thomas
   Elements of Information Theory 1991
   5613
6. A P Dempster, N M Laird, D B Rubin
   Maximum likelihood from incomplete data via the EM algorithm (with discussion. Journal of the Royal Statistical Society Series B, 1977
   5459
7. Judea Pearl
   Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference 1988
   5222
8. E Gamma, R Helm, R Johnson, J Vlissides
   Design patterns: Elements of reusable object-oriented software 1995
   4290
9. J R Quinlan
   C4.5: Programs for Machine Learning 1992
   3744
10. C E Shannon, W Weaver
    The Mathematical Theory of Communication 1949
    3721
11. C M Bishop
    Neural Networks for Pattern Recognition 1995
    3616
12. J Hennessy, D Patterson
    Computer Architecture: A Quantitative Approach 1996

Most cited authors in Computer Science - August 2006 (CiteSeer.Continuity)

Generated from documents in the CiteSeer.Continuity database. This list does not include citations where one or more authors of the citing and cited articles match, or citations where the relevant author is an editor. An entry may correspond to multiple authors (e.g. J. Smith). This list is automatically generated and may contain errors. Citation counts may differ from search results because this list is generated in batch mode whereas the database is continually updated. A total of 790329 authors were found. Homepages listed may not be for the most cited individual, especially when an entry corresponds to multiple authors. Click on HPSearch to see and update the latest homepage data.

Next 250

1. D. Johnson (HPSearch): 16227
2. J. Ullman (HPSearch): 13245
3. A. Gupta (HPSearch): 10156
4. R. Rivest (HPSearch): 9967
5. R. Milner (HPSearch): 9878
6. S. Shenker (HPSearch): 9456
7. V. Jacobson (HPSearch): 8659
8. S. Floyd (HPSearch): 8487
9. M. Garey (HPSearch): 8485
10. R. Tarjan (HPSearch): 8269
11. E. Clarke (HPSearch): 7909
12. J. Smith (HPSearch): 7893
13. L. Lamport (HPSearch): 7759
14. J. Dongarra (HPSearch): 7722
15. L. Zhang (HPSearch): 7284
16. D. Knuth (HPSearch): 7269
17. R. Agrawal (HPSearch): 7073
18. R. Karp (HPSearch): 6833
19. C. Papadimitriou (HPSearch): 6816
20. H. Zhang (HPSearch): 6802
21. R. Johnson (HPSearch): 6769
22. A. Pnueli (HPSearch): 6609
23. H. Garcia-Molina (HPSearch): 6592
24. A. Aho (HPSearch): 6523
25. D. Goldberg (HPSearch): 6299
26. R. Jain (HPSearch): 6287
27. J. Hennessy (HPSearch): 6267
28. C. Leiserson (HPSearch): 6132
29. A. Pentland (HPSearch): 6131

Tasks that need to be solved

- Information Retrieval
  - search for research papers on the Web
- Information Extraction
  - extract relevant information (title, author, journal/conference, publication year,...) from the research papers
  - extract citations from the research papers
- Information Integration
  - match extracted citations with the text where they are cited
  - match extracted citations with other extracted citations
  - identify similar documents
- Citation analysis
  - build and analyze a graph of citations of papers
  - build and analyze a co-authorship graph
- and many more...

Web Mining

Web Mining is Data Mining for Data on the World-Wide Web

- Text Mining:
  - Application of Data Mining techniques to unstructured (free-format) text
- Structure Mining:
  - taking into account the structure of (semi-)structured hypertext (HTML tags, hyperlinks)
- Usage Mining:
  - taking into account user interactions with the text data (click-streams, collaborative filtering, ...)

Web Mining Tasks

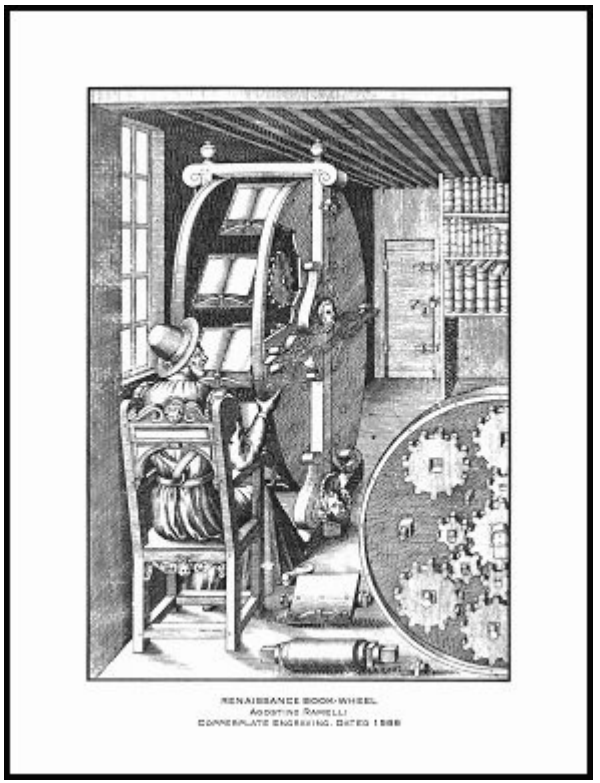
- Message Filter or Message Sorter
- Intelligent Browsing Assistants
- Formation or Update of Web Catalogues
- Ranking or Clustering of Search Results
- Building the Semantic Web / World-Wide Knowledge Base
- Click-stream Analysis
- Product Recommendations
- Digital libraries and Citation Analysis
- ...

The Web

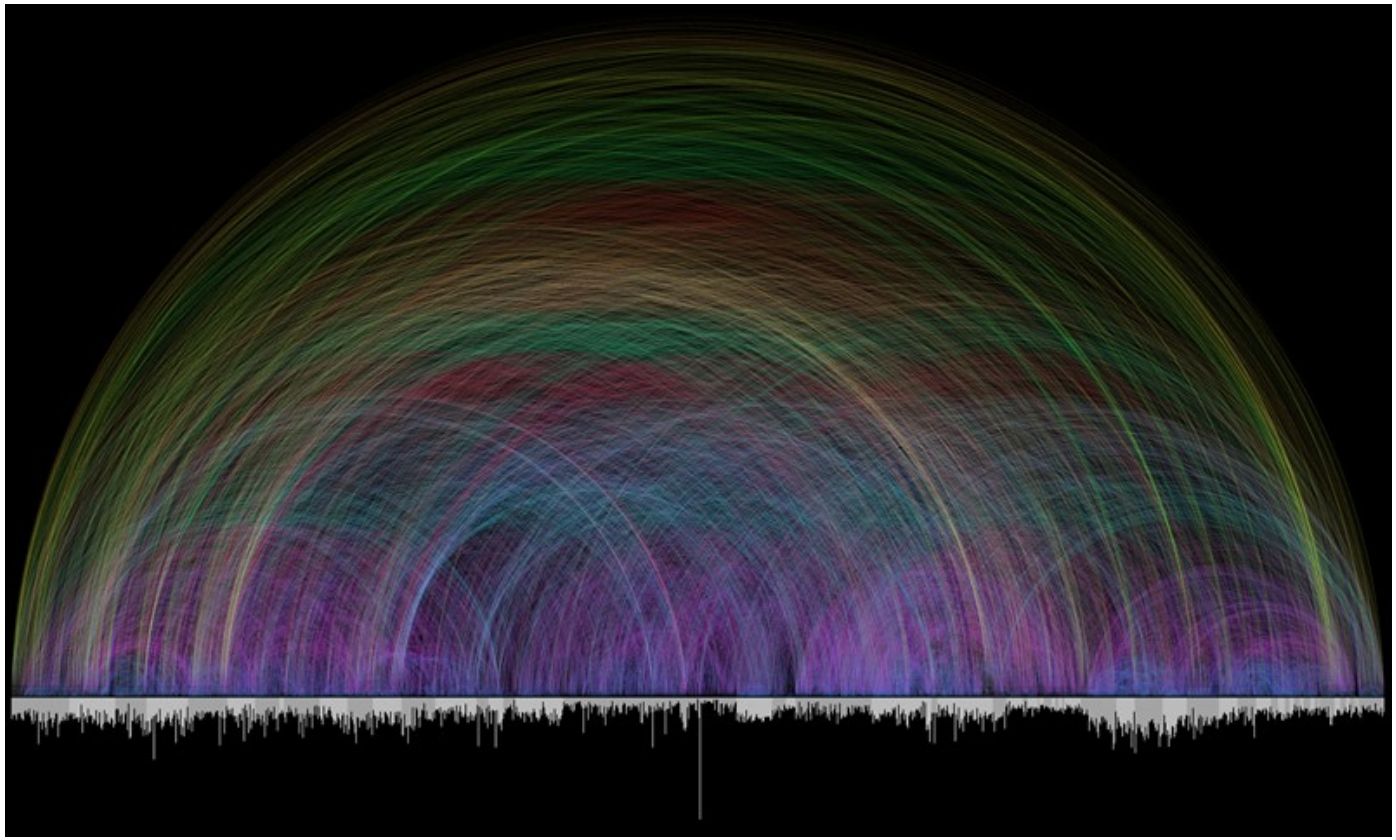
- The Web is a unique kind of hypertext document
  - a large number of pages
  - on a wide variety of topics
  - originating by a large variety of authors
  - speaking many different languages
  - annotated via hyperlinks
  - accessible to everybody
- Main Problem:
  - How can I find the information I am looking for?
- Web Mining:
  - finding and extracting relevant information from the Web

A Brief History of Hypertext

- On Paper
  - Annotated books (e.g., the Talmud)
  - Dictionaries and encyclopedias
    - cross-references are hyperlinks
  - Scientific literature
    - citations of other works is another form of hyperlinks
- The book wheel
  - Agostino Ramelli, Paris 1588
  - Device for reading several books at once
  - maybe considered as a precursor to the Memex and thus to hypertext



Example: Cross-references in the Bible



The bar graph that runs along the bottom represents all of the chapters in the Bible. Books alternate in color between white and light gray. The length of each bar denotes the number of verses in the chapter. Each of the 63,779 cross references found in the Bible is depicted by a single arc - the color corresponds to the distance between the two chapters, creating a rainbow-like effect.

Source:Chris Harrison, CMU (http://www.chrisharrison.net/projects/bibleviz/)

Example: Social Network in the Bible

based on name co-occurrences in verses

Source:Chris Harrison, CMU (http://www.chrisharrison.net/projects/bibleviz/)

A Brief History of Hypertext

- Memex (Vannevar Bush, 1945)
  - design for a photo-eletrical, mechanical storage device that could link documents
  - On-line Demo

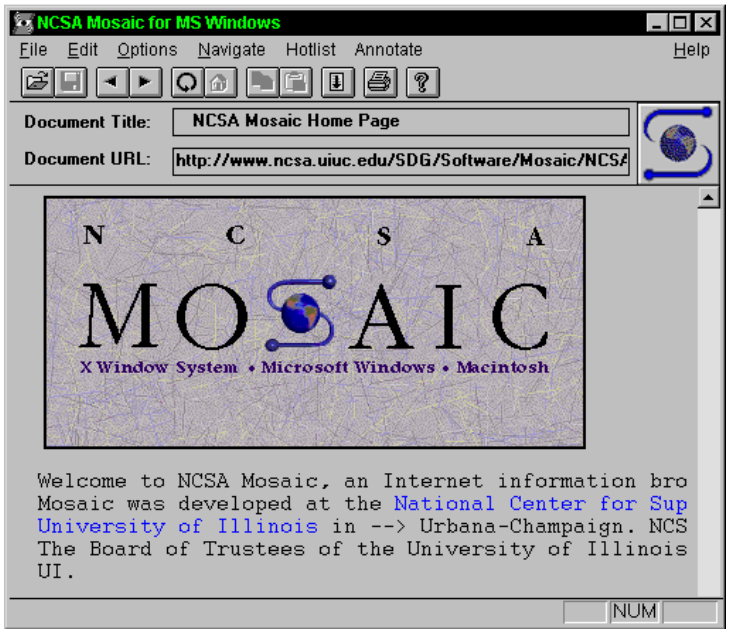
    http://www.youtube.com/watch?v=c539cK58ees

    http://www.dynamicdiagrams.com/demos/memex1a.zip

- Xanadu (Engelbart & Nelson 1965) http://xanadu.com/
  - first conventional hypertext system, also pioneered wikis
  - too complex to be realized, first use of word „hypertext“
- Many successor systems

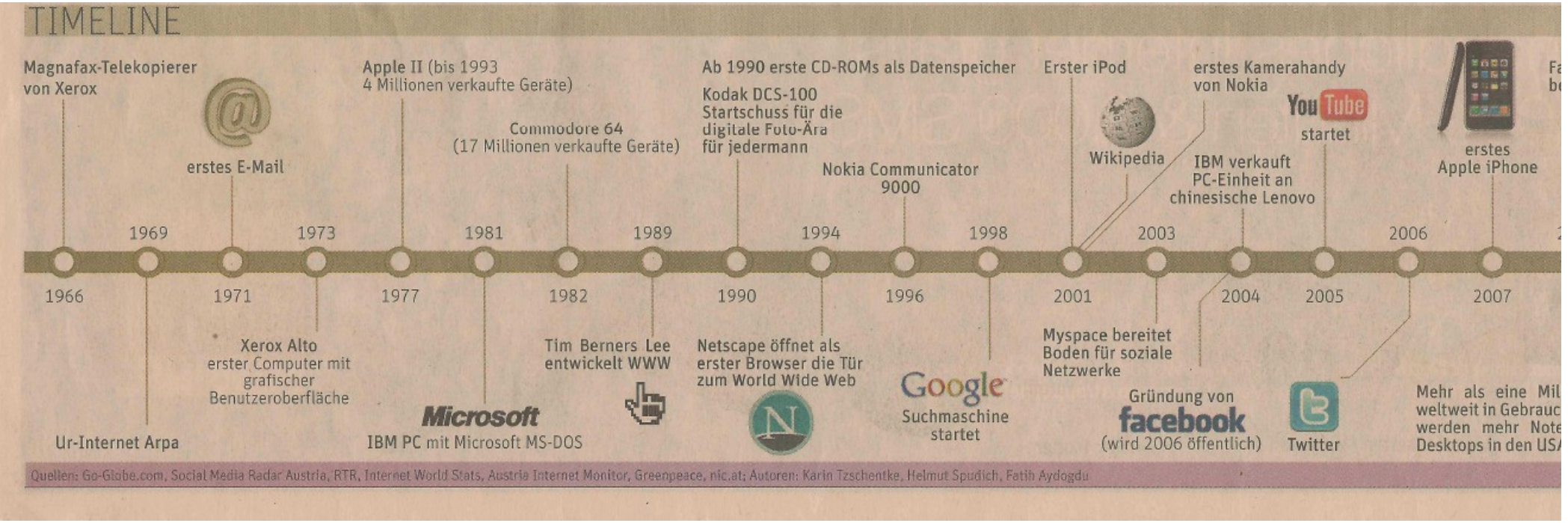
A Brief History of the Web

- Tim Berners-Lee (CERN)
  - first proposals around 1980
  - 1990: work on the „World Wide Web“
  - first graphical interfaces
- 1993:
  - Mosaic (Mark Andressen, NCSA): intuitive hypertext GUI for UNIX
  - HTML: hypertext markup language
  - HTTP: hypertext transport protocol
- 1994:
  - Netscape was founded
  - 1st World Wide Web Conference
  - World Wide Web Consortion founded by CERN and MIT



http://www.w3.org/

A Brief History of the Web



Quelle: Der Standard, 13. 8. 2011

HTTP (hypertext transport protocol)

- Built on top of the Transport Control Protocol (TCP)
- Steps(from client end) http://www.w3.org/Protocols
  - resolve the server host name to an Internet address (IP)
    - Use Domain Name Server (DNS)
    - DNS is a distributed database of name-to-IP mappings maintained at a set of known servers
  - contact the server using TCP
    - connect to default HTTP port (80) on the server.
    - Enter the HTTP requests header (E.g.: GET)
    - Fetch the response header
      - MIME (Multipurpose Internet Mail Extensions)
      - A meta-data standard for email and Web content transfer
    - Fetch the HTML page

Sample http connection log

Host Port

```
% telnet www.cse.iitb.ac.in 80
Trying 144.16.111.14...
Connected to www.cse.iitb.ac.in.
Escape character is '^]'.
GET / Http/1.0

Http/1.1 200 OK
Date: Sat, 13 Jan 2001 09:01:02 GMT
Server: Apache/1.3.0 (Unix) PHP/3.0.4
Last-Modified: Wed, 20 Dec 2000 13:18:38 GMT
ETag: "5c248-153d-3a40b1ae"
Accept-Ranges: bytes
Content-Length: 5437
Connection: close
Content-Type: text/html
X-Pad: avoid browser bug

<html>
<head><title>IIT Bombay CSE Department Home Page</title></head>
<body>...<a href="http://www.iitb.ac.in">IIT Bombay</a>...
</body></html>
Connection closed by foreign host.
```

Pfad

Header

HTML of Web page

HTML

http://www.w3.org/MarkUp/

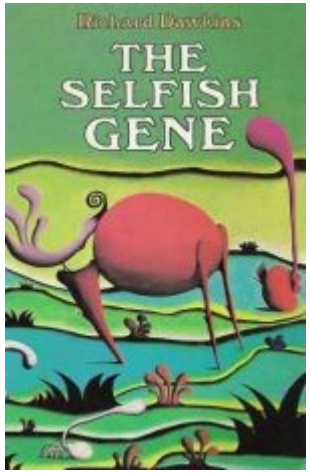
- HyperText Markup Language
- Lets the author
  - specify document structure
    - browser converts structure to layout
    - direct specification of layout and typeface possible
  - embed diagrams
  - create hyperlinks.
    - expressed as an anchor tag with a HREF attribute
    - HREF names another page using a Uniform Resource Locator (URL),
- URL (Uniform Resource Locator) =
  - protocol field (e.g., “HTTP”) +
  - server hostname (“www.cse.iitb.ac.in”) +
  - file path (/, the `root' of the published file system).

DOM Tree

- DOM = Document Object Model http://www.w3.org/DOM/
- An HTML document can be viewed as a tree
  - markup items are interior nodes
  - text are leafs
  - Xpath: language for denoting the path from the root to a tree
    http://www.zvon.org/xxl/XPathTutorial/General/examples.html
- document structure can be exploited
  - sectioning of documents
  - recognition of important text parts (e.g., anchor text)
  - structural patterns (XPath) may identify important information on the page
- `Firefox->Web Developer/DOM Inspector`
  - plugin

Web: A populist, participatory medium

- number of writers =(approx) number of readers.
- the evolution of *memes*
  - term „meme“ coined by Richard Dawkins („The Selfish Gene“)
    - in analogy to the role of genes in evolution
  - memes are ideas, theories etc that spread
  - from person to person by imitation.
    - good memes survive, bad memes die out
    - the Web archives them all



Problems due to Uniform accessibility

- little support for adapting to the background of specific users.
- commercial interests routinely influence the operation of Web search
  - “Search Engine Optimization“ !!
- False information
  - Hacked FoxNews, July 4th 2011

Data Mining - Motivation

"Computers have promised us a fountain of wisdom but delivered a flood of data."

"It has been estimated that the amount of information in the world doubles every 20 months."

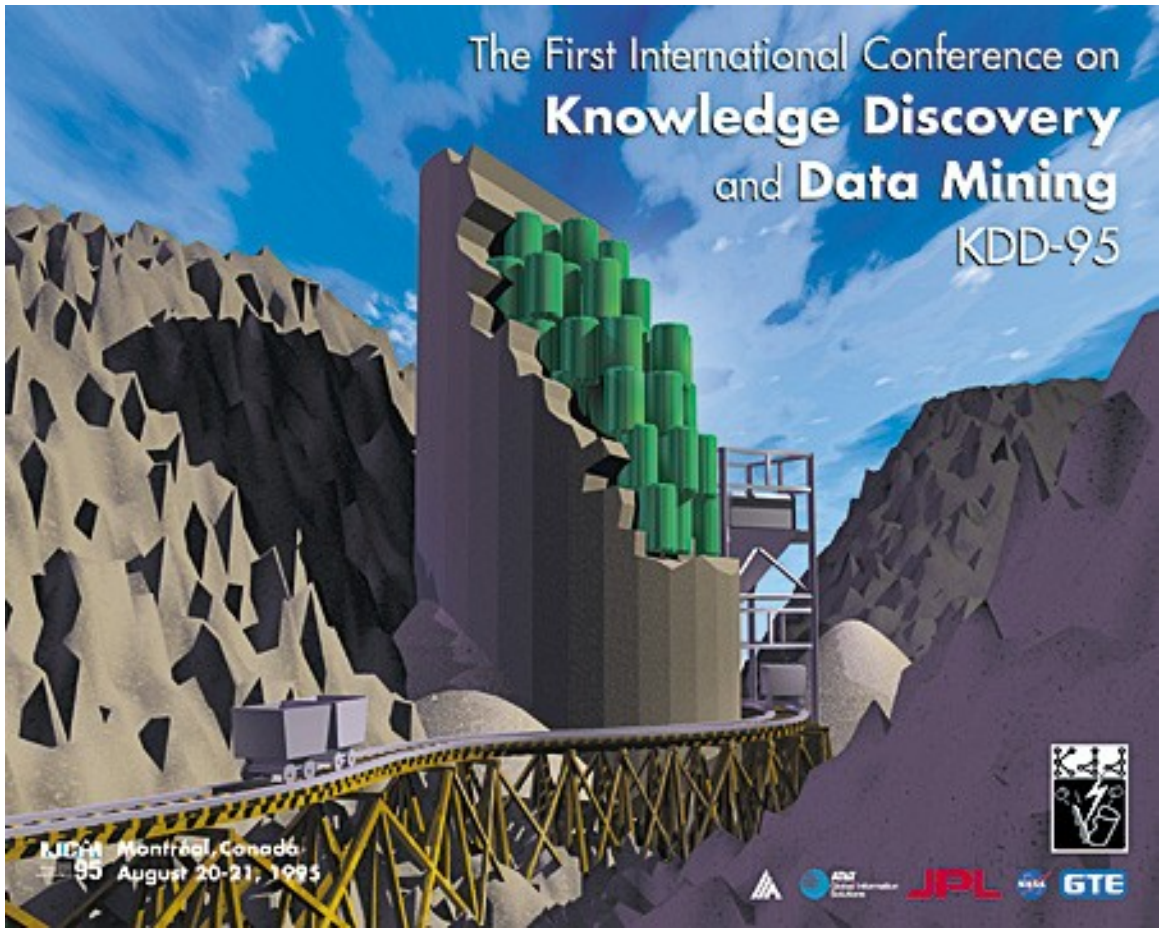
(Frawley, Piatetsky-Shapiro, Matheus, 1992)

„160,000,000 terabytes of data have been generated in 2006“

(Data Consortium)

Data Mining

Mining for nuggets of knowledge in mountains of Data.



Definition

Data Mining is a non-trivial *process* of identifying

- valid
- novel
- potentially useful
- ultimately understandable

patterns in data.

(Fayyad et al. 1996)

It employs techniques from

- machine learning
- statistics
- databases

Or maybe:

- Data Mining is torturing your database until it confesses.

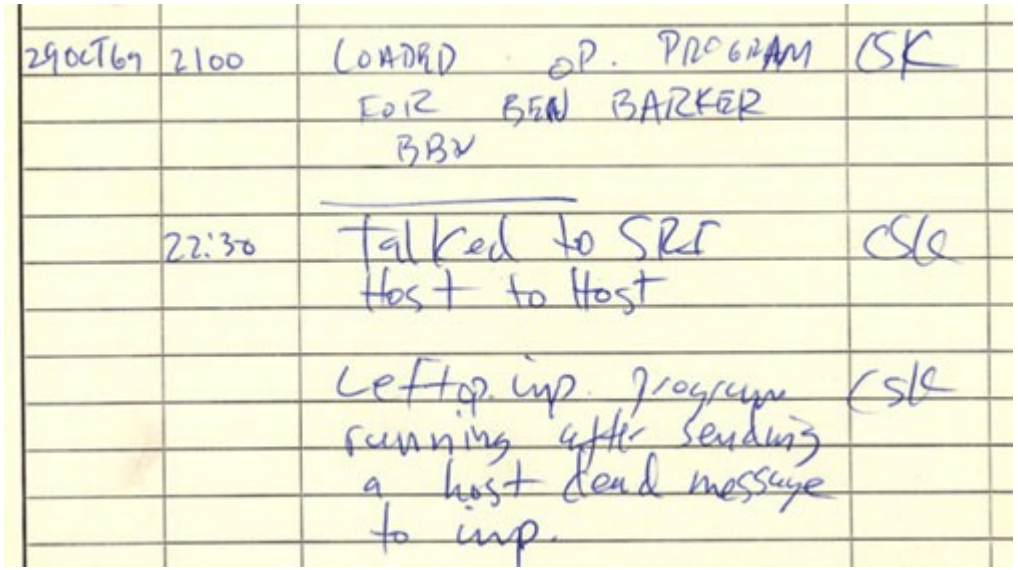
(Heikki Manilla (?) after Ronald Coase)

World-Wide Data Growth

- Science
  - satellite monitoring
  - human genome
- Business
  - OLTP (on-line transaction processing)
  - data warehouses
  - e-commerce
- Industry
  - process data
- World-Wide Web

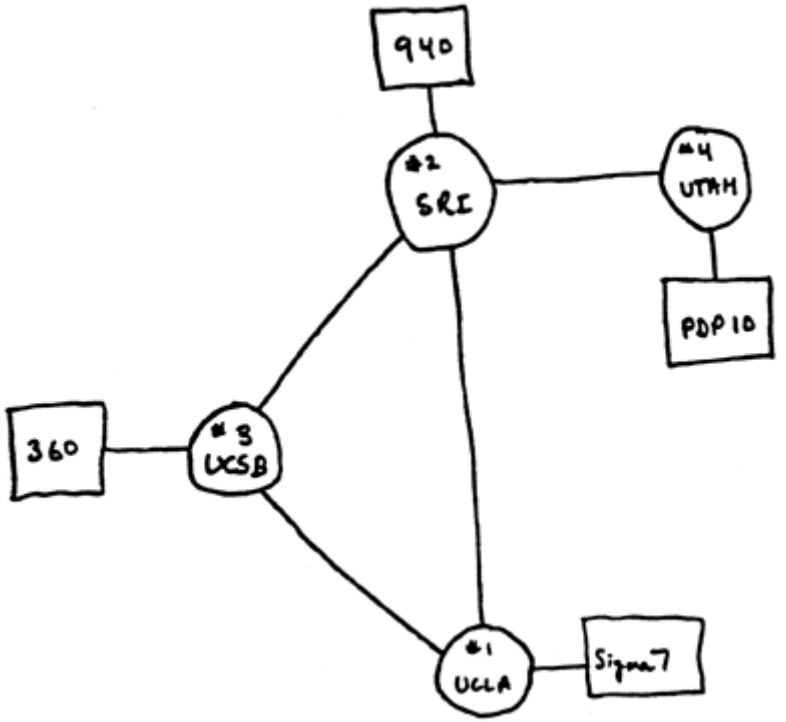
The Birth of the Web

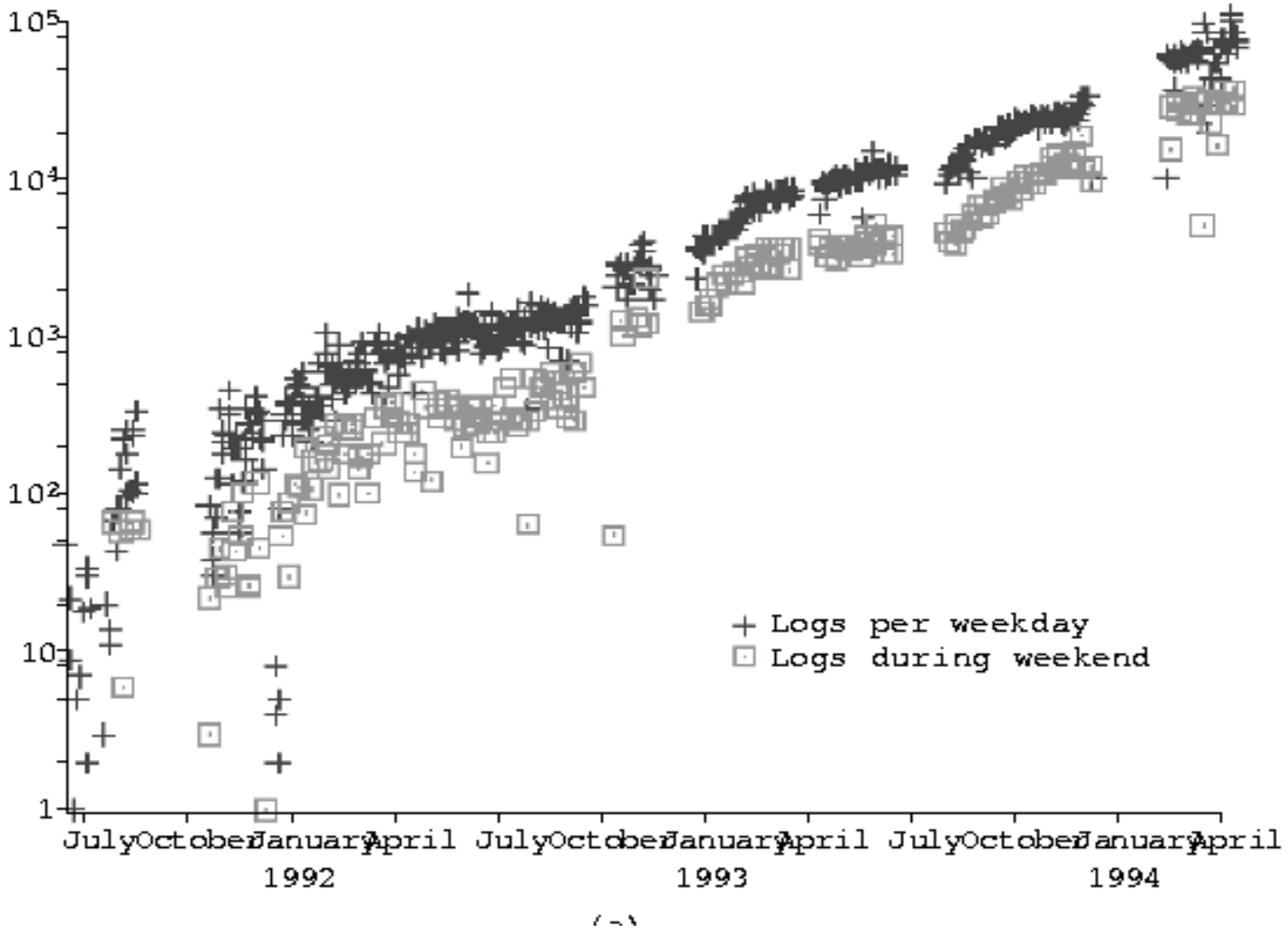
- ARPANET
  - started with 4 nodes at four universities
    - UCLA, UCSB, SRI, Utah
  - first message sent on October 29, 1969


29OCT69 2100 LOADED UP. PROGRAM CSK
FOR BEN BARKER
BBN

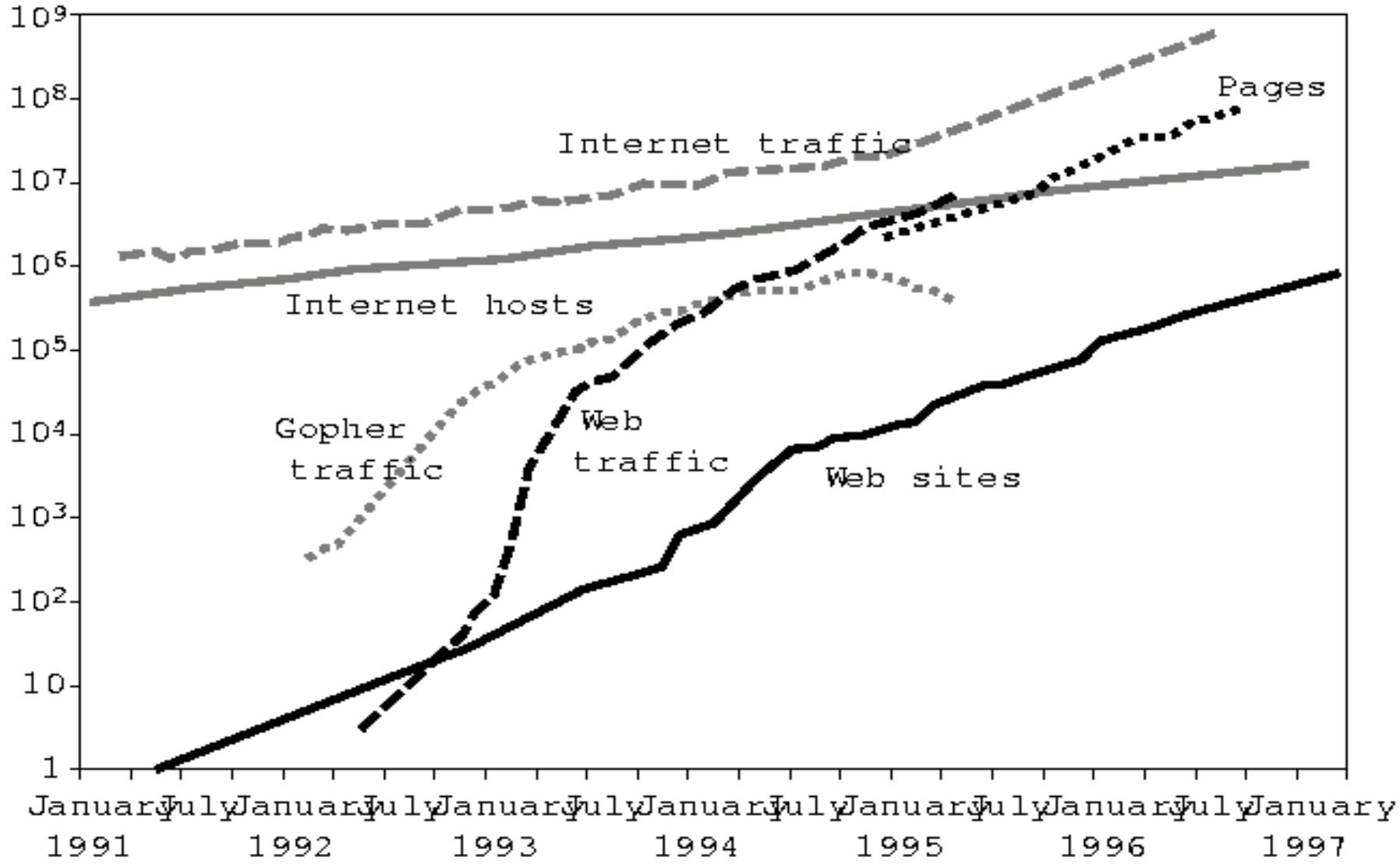
22:30 Talked to SRI CSK
Host to Host

Left imp. program CSK
running after sending
a host dead message
to imp.



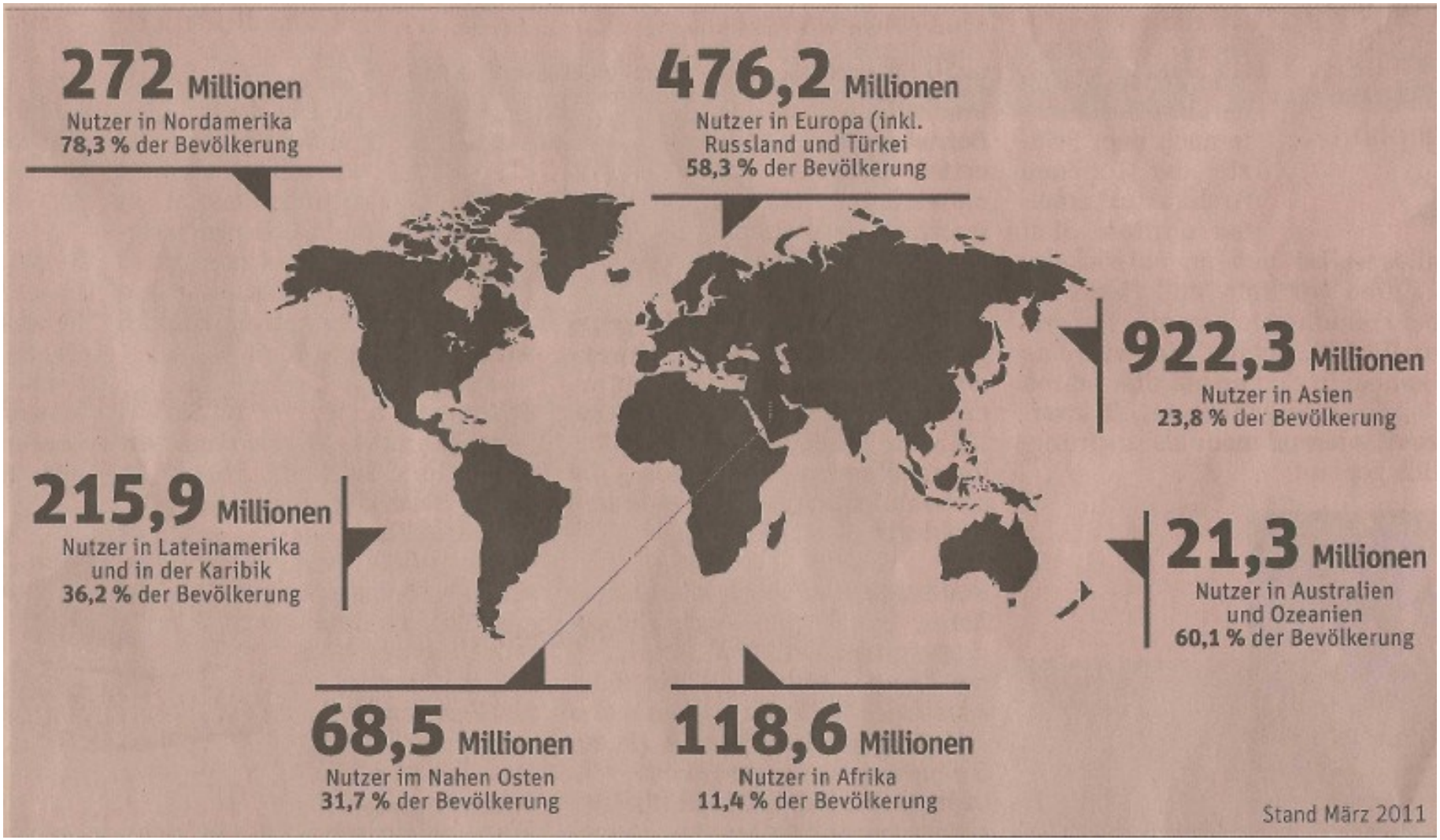


The early days of the Web : CERN HTTP traffic grows by 1000 between 1991-1994 (image courtesy W3C)



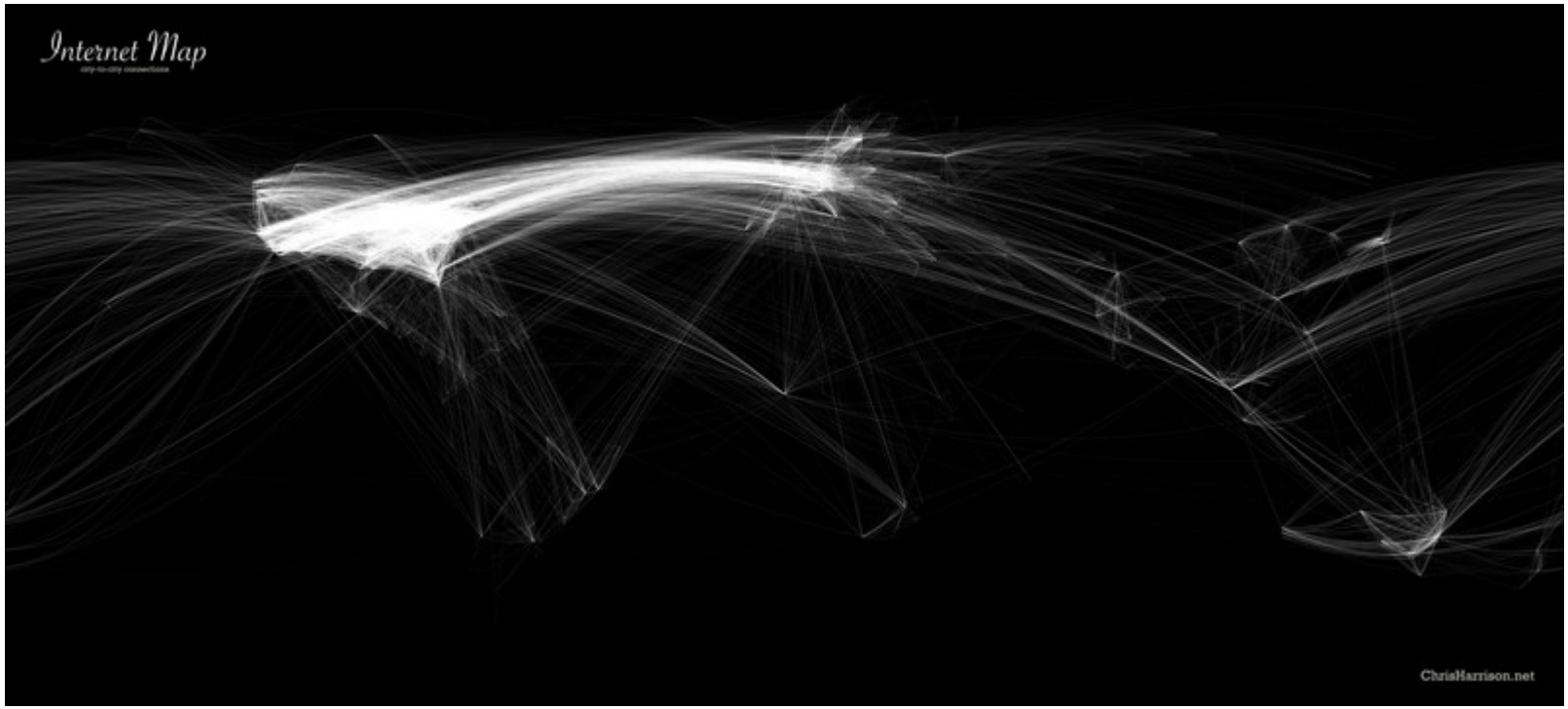
The early days of the Web: The number of servers grows from a few hundred to a million between 1991 and 1997 (image courtesy Nielsen)

Geographic Distribution of Internet Usage



Quelle: Der Standard, 13. 8. 2011

Geographic Map of Internet Traffic



Source:Chris Harrison, CMU

How Big is the Web?

- Google:
  - early 2001: 1,346,966,000 web pages
  - 11.2.2002: 2,073,418,204
  - 2004: 4,285,199,774
  - 28.4.2005: 8,058,044,651
- Size of the Web
  - Results from 1998 estimate that the best search engines index about 30% of the Web.
- Gulli & Signorini (2005)
  - estimate the size of the Web to 11.5 billion pages,
  - Coverage of search engines
    - Google=76.16%, Msn Beta=61.90%, Ask/Teoma=57.62%, Yahoo!=69.32%

Source: www.thestandard.com

WorldWideWebSize.com

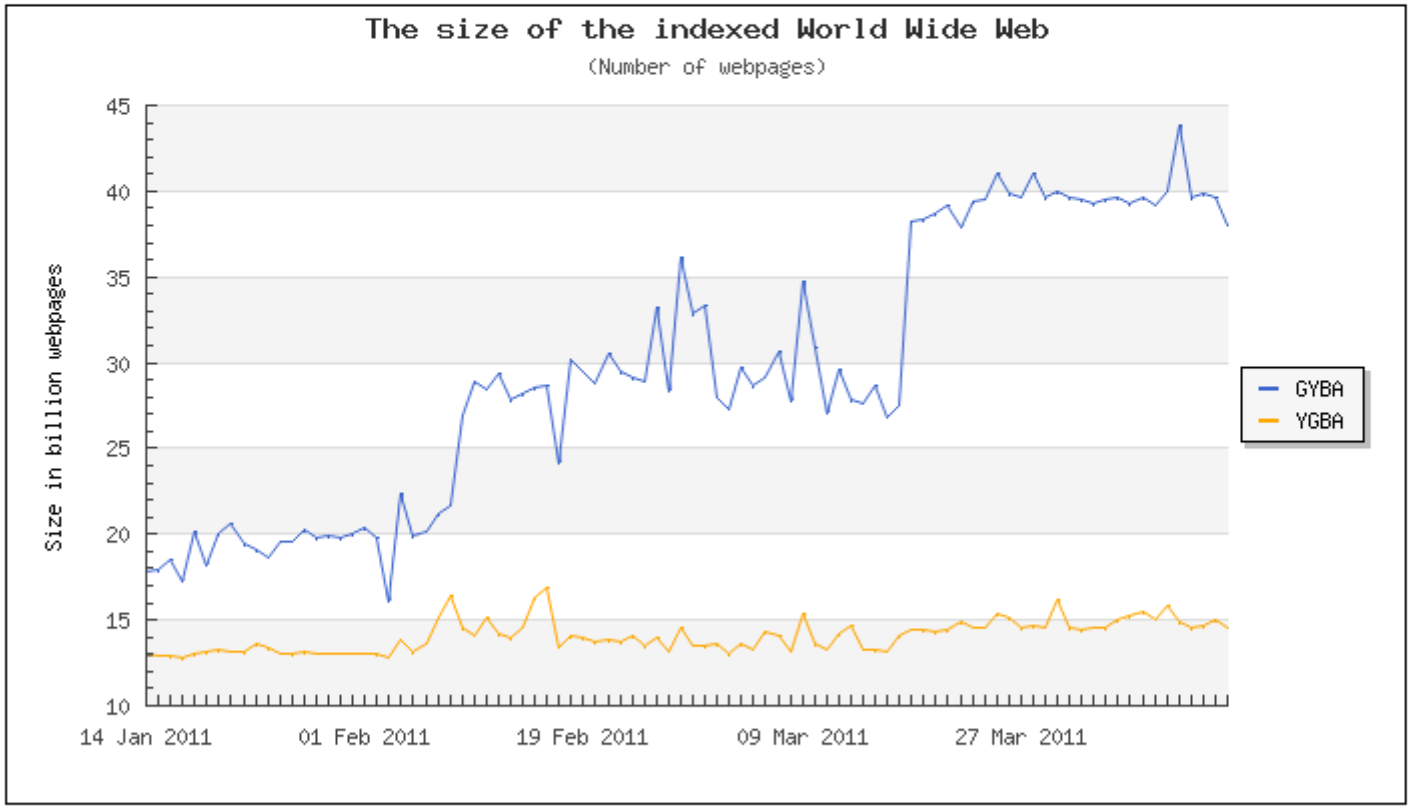
DAILY ESTIMATED SIZE OF THE WORLD WIDE WEB

The size of the World Wide Web

The Indexed Web contains **at least 14.56 billion pages** (Wednesday, 13 April, 2011).
The Dutch Indexed Web contains **at least 618.51 million pages** (Wednesday, 13 April, 2011).

The Indexed Web | The Dutch Indexed Web

Last Month | Last Three Months | Last Year | Last Two Years

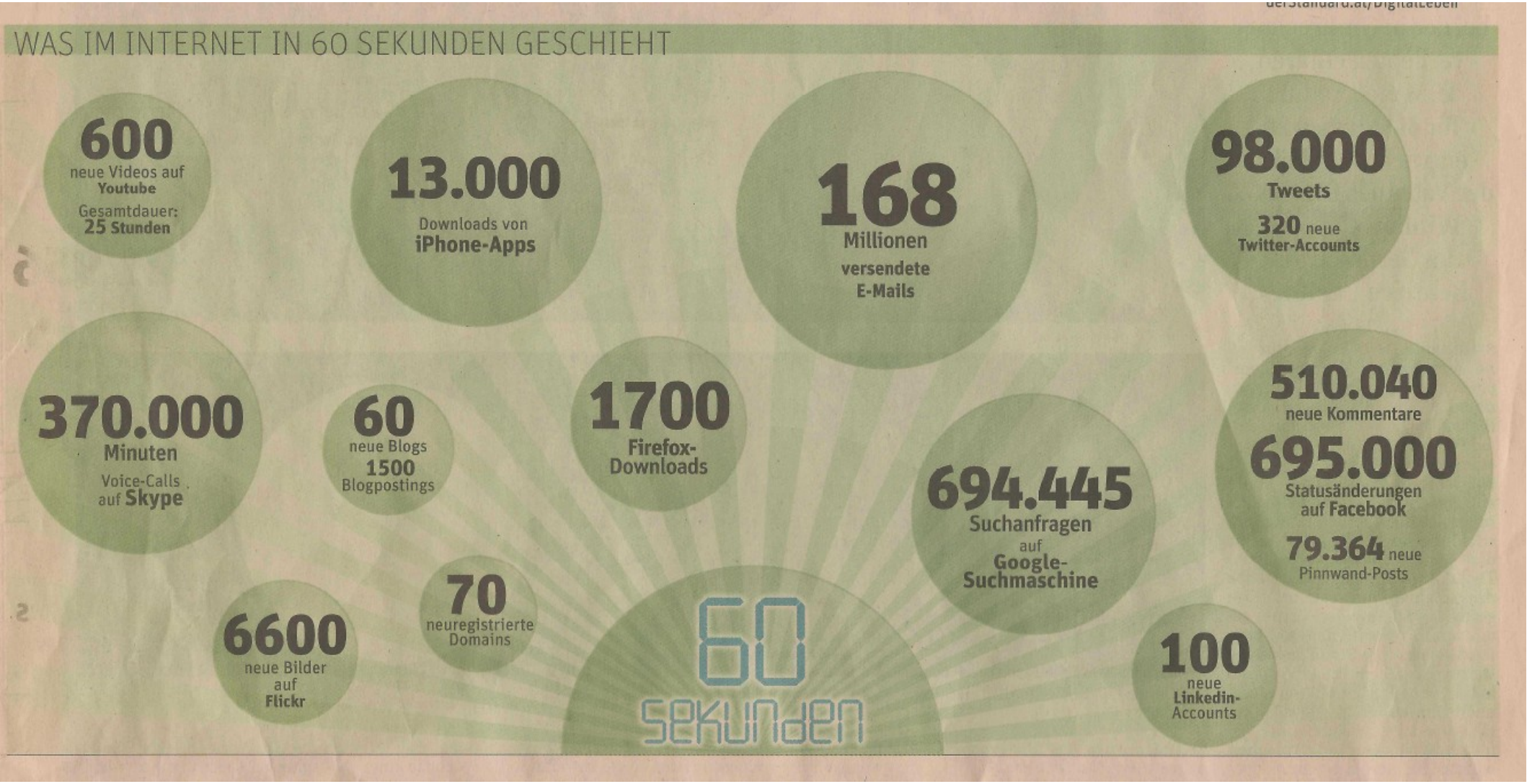


GYBA = Sorted on Google, Yahoo!, Bing and Ask
YGBA = Sorted on Yahoo!, Google, Bing and Ask

E-mail and WWW Usage

Quelle: Der Standard, 13. 8. 2011

Internet Activity in 60 secs



Quelle: Der Standard, 13. 8. 2011

Social Connectivity

Quelle: Der Standard, 13. 8. 2011

Structured vs. Web data mining

- traditional data mining
  - data is structured and relational
  - well-defined tables, columns, rows, keys, and constraints.
- Web data
  - semi-structured and unstructured
  - readily available
  - rich in features and patterns
  - spontaneous formation and evolution of
    - topic-induced graph clusters
    - hyperlink-induced communities

Structured Data

- Attribute-Value data:
  - Each example is described with values for a fixed number of attributes
    - **Nominal Attributes:**
      - store an unordered list of symbols (e.g., *color*)
    - **Numeric Attributes:**
      - store a number (e.g., *income*)
    - **Other Types:**
      - hierarchical attributes
      - set-valued attributes
  - the data corresponds to a single relation (spreadsheet)
- Multi-Relational data:
  - The relevant information is distributed over multiple relations
  - Inductive Logic Programming

Structured Data

Day	Temperature	Outlook	Humidity	Windy	Play Golf?
07-05	hot	sunny	high	false	no
07-06	hot	sunny	high	true	no
07-07	hot	overcast	high	false	yes
07-09	cool	rain	normal	false	yes
07-10	cool	overcast	normal	true	yes
07-12	mild	sunny	high	false	no
07-14	cool	sunny	normal	false	yes
07-15	mild	rain	normal	false	yes
07-20	mild	sunny	normal	true	yes
07-21	mild	overcast	high	true	yes
07-22	hot	overcast	normal	false	yes
07-23	mild	rain	high	true	no
07-26	cool	rain	normal	true	no
07-30	mild	rain	high	false	yes

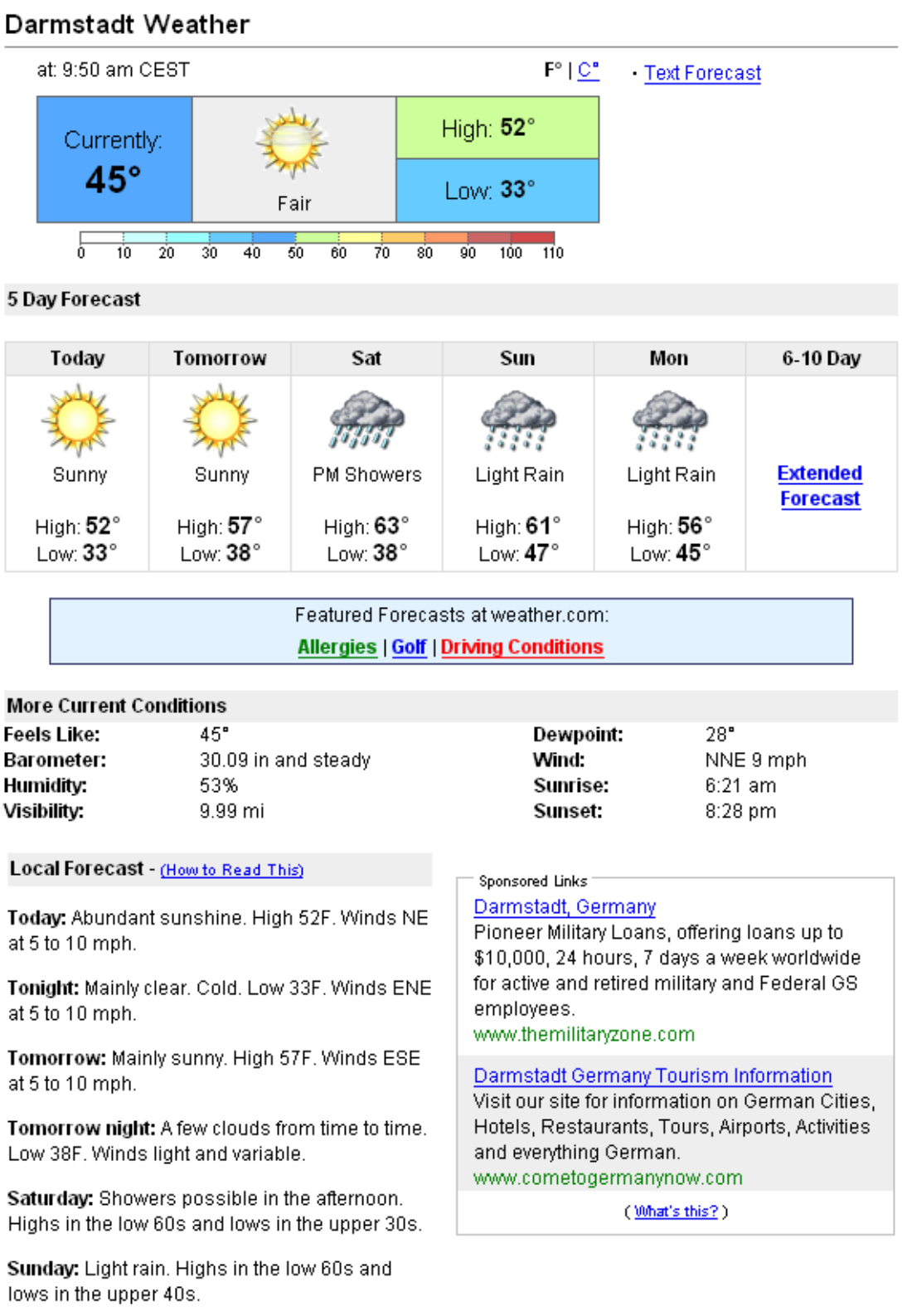
today	cool	sunny	normal	false	?
tomorrow	mild	sunny	normal	false	?

Semi-Structured and Unstructured Data

- Semi-structured Data
  - no clear tables
    - it may be hard to identify the attributes for each example
    - it may also be hard to identify the examples themselves
  - some structure implicit in the data
    - e.g., formatting via HTML
  - large parts without structure
    - free text
  - http://weather.yahoo.com/forecast/GMXX0020.html

Semi-Structure

- Semi-structured Data
  - no clear tables
    - it may be hard to identify
    - it may also be hard to ide
  - some structure implicit in
    - e.g., formatting via HTMl
  - large parts without structu
    - free text
  - http://weather.yahoo.com/fo



Darmstadt Weather

at: 9:50 am CEST F° | C° · Text Forecast

Currently: 45° Fair High: 52° Low: 33°

0 10 20 30 40 50 60 70 80 90 100 110

5 Day Forecast

Today	Tomorrow	Sat	Sun	Mon	6-10 Day
Sunny	Sunny	PM Showers	Light Rain	Light Rain	Extended Forecast
High: 52° Low: 33°	High: 57° Low: 38°	High: 63° Low: 38°	High: 61° Low: 47°	High: 56° Low: 45°	

Featured Forecasts at weather.com:
Allergies | Golf | Driving Conditions

More Current Conditions

Feels Like:	45°	Dewpoint:	28°
Barometer:	30.09 in and steady	Wind:	NNE 9 mph
Humidity:	53%	Sunrise:	6:21 am
Visibility:	9.99 mi	Sunset:	8:28 pm

Local Forecast - (How to Read This)

Today: Abundant sunshine. High 52F. Winds NE at 5 to 10 mph.

Tonight: Mainly clear. Cold. Low 33F. Winds ENE at 5 to 10 mph.

Tomorrow: Mainly sunny. High 57F. Winds ESE at 5 to 10 mph.

Tomorrow night: A few clouds from time to time. Low 38F. Winds light and variable.

Saturday: Showers possible in the afternoon. Highs in the low 60s and lows in the upper 30s.

Sunday: Light rain. Highs in the low 60s and lows in the upper 40s.

Sponsored Links

Darmstadt, Germany
Pioneer Military Loans, offering loans up to $10,000, 24 hours, 7 days a week worldwide for active and retired military and Federal GS employees.
www.themilitaryzone.com

Darmstadt Germany Tourism Information
Visit our site for information on German Cities, Hotels, Restaurants, Tours, Airports, Activities and everything German.
www.cometogermanynow.com

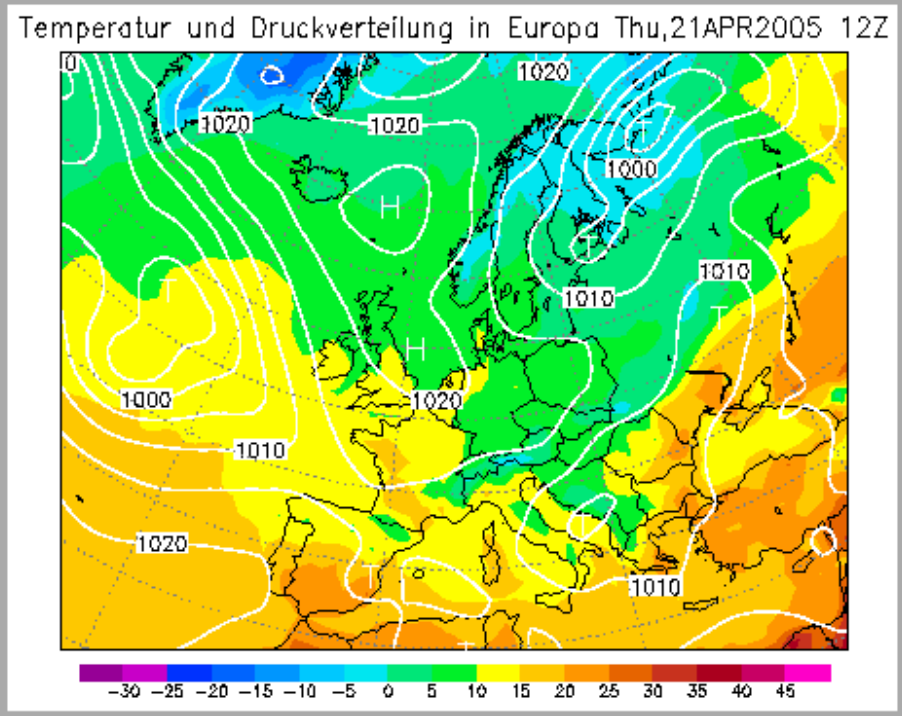
(What's this?)

Semi-Structured and Unstructured Data

- Semi-structured Data
  - no clear tables
    - it may be hard to identify the attributes for each example
    - it may also be hard to identify the examples themselves
  - some structure implicit in the data
    - e.g., formatting via HTML
  - large parts without structure
    - free text
  - http://weather.yahoo.com/forecast/GMXX0020.html
- Unstructured Data
  - free text
  - http://www.wetterzentrale.de/wzwb.html

Der Wetterzentrale Wetterbericht ausgegeben am 21. April 2005, 8:09 MESZ

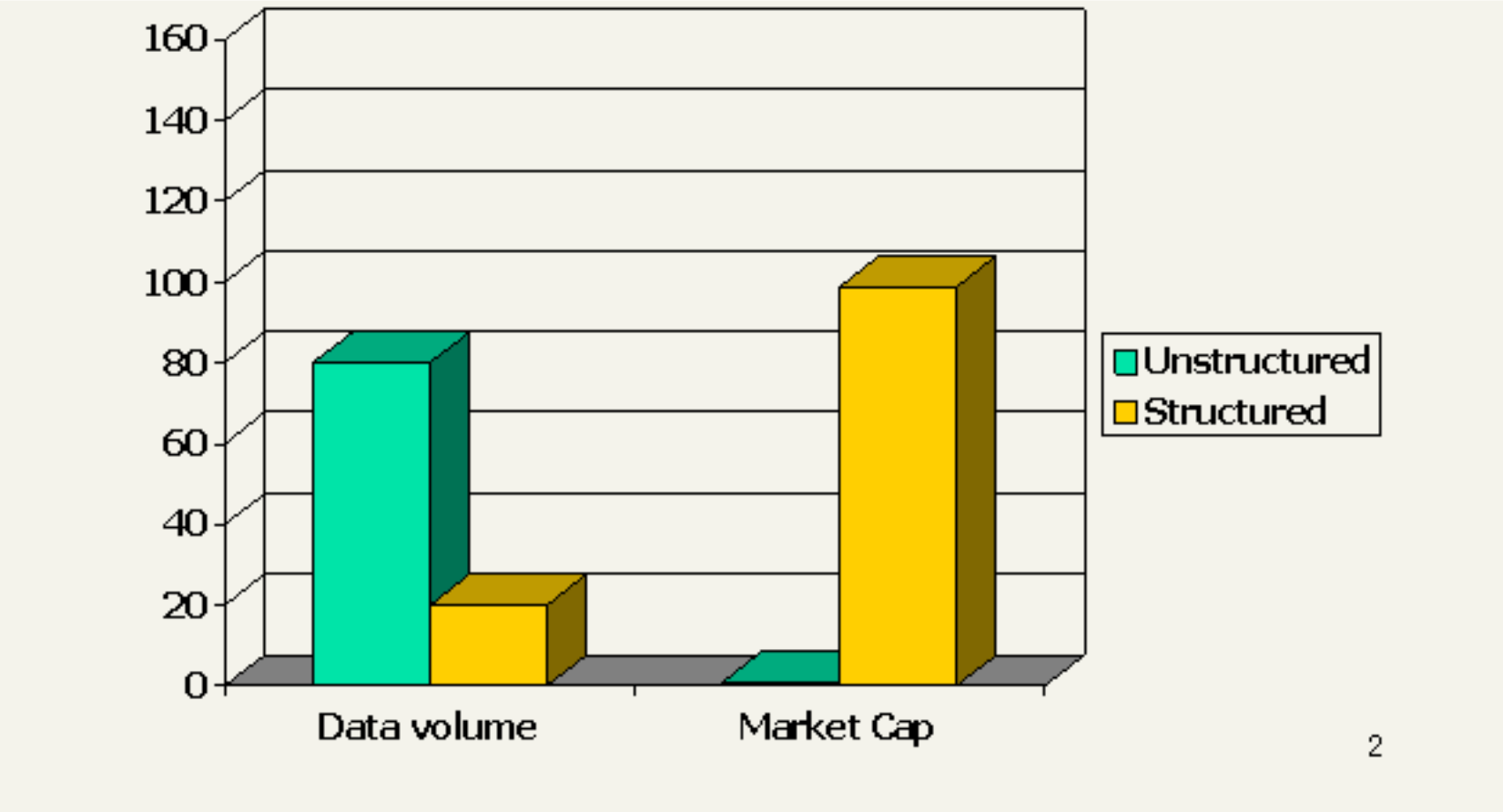
Lage:
Die aus Nordosten eingeflossene Kaltluft gelangt rasch unter schwachen Hochdruckeinfluss. Bereits am Samstag greifen die Ausläufer westeuropäischer Tiefs auf den Südwesten über und führen mildere und feuchte Luft heran.



Vorhersage für Deutschland:
Heute nach Auflösung örtlichen Nebels meist heiter bis wolkig und trocken. Am Alpenrand anfangs noch stark bewölkt, aber kaum noch Regen. Im Norddeutschen Tiefland ab dem Mittag einige Wolkenfelder. Höchsttemperaturen 8 bis 13 Grad. Dabei am Rhein am mildesten. Schwacher bis mäßiger Wind, im Norden auf West drehend, sonst aus Nordost bis Nord. In der kommenden Nacht im Norden wolkig. Sonst klar. Tiefstwerte zwischen 3 Grad im Norden und bis -3 Grad im Süden.
Morgen östlich der Elbe wolkig, es bleibt aber trocken. Sonst sonnig und trocken. Höchsttemperaturen zwischen 10 Grad an der Oder und bis 16 Grad am Rhein.

Tendenz für die Folgetage:
Am Samstag im Südwesten bereits am Vormittag zunehmende Bewölkung und ab dem Mittag einsetzender Regen. In der Mitte freundlich und mild. Im Nordosten wolkig und immer noch kühl.
Am Sonntag im Norddeutschen Tiefland heiter bis wolkig und trocken. Bei kräftigem Ostwind recht kühl. In der Mitte und im Süden wolkig bis stark bewölkt mit gebietsweisem Regen oder einzelnen Schauern und mild.
Am Wochenbeginn auch im Norden unbeständiger.
Ab der Wochenmitte deutet sich trockenes und wärmeres Wetter an.

Unstructured vs. Structured Data 1996



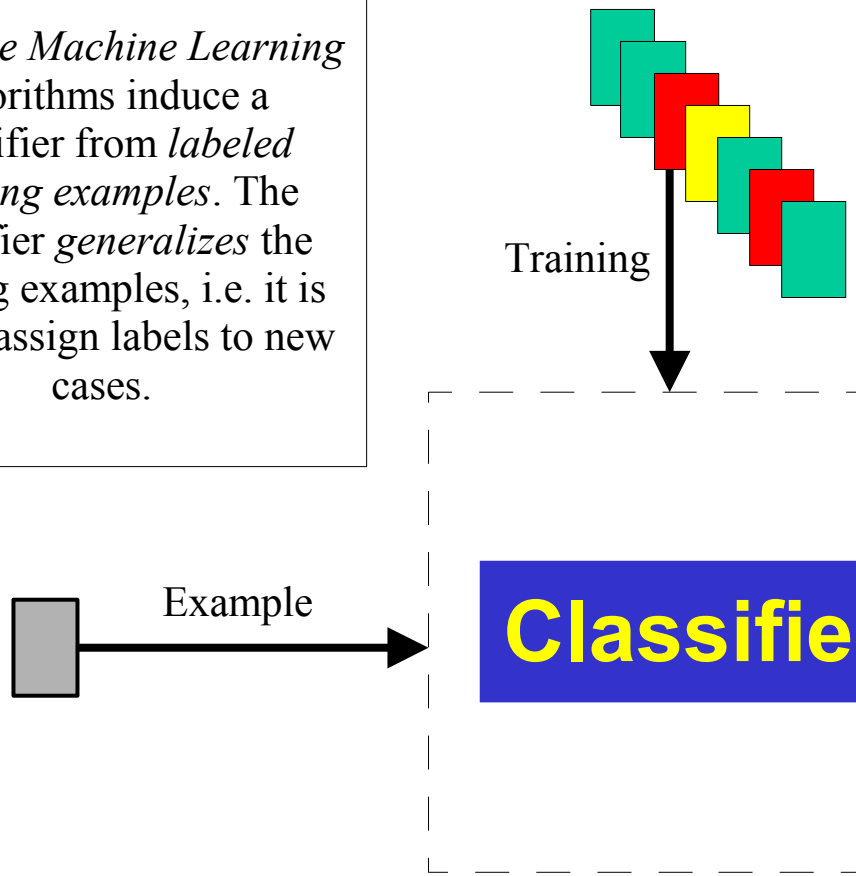
Unstructured vs. Structured Data 2006

Web Tasks for ML/DM Techniques

- Classifiers:
  - assigning categories to documents (E-mail/newsgroup sorting and filtering, building a Web catalogue, user modelling,...)
- Regression:
  - predict numerical values (ratings, GUI settings,...)
- Clustering:
  - grouping documents (structuring search results, ...)
- Association Rule Discovery:
  - finding events and event sequences that co-occur frequently (click stream analysis,...)
- Reinforcement Learning:
  - learning to improve agents (crawlers, relevance feedback, ...)

Induction of Classifiers

Inductive Machine Learning algorithms induce a classifier from *labeled training examples*. The classifier *generalizes* the training examples, i.e. it is able to assign labels to new cases.



An inductive learning algorithm searches in a given family of hypotheses (e.g., *decision trees, neural networks*) for a member that optimizes given *quality criteria* (e.g., estimated predictive accuracy or misclassification costs).