

707.000
Web Science and Web Technology
„Overview and Motivation“

Markus Strohmaier

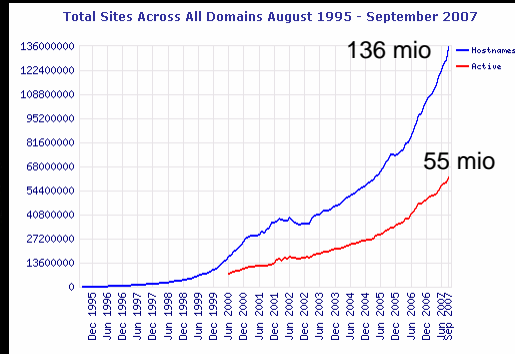
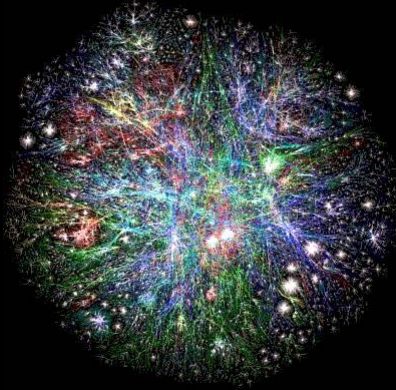
Univ. Ass. / Assistant Professor
Knowledge Management Institute
Graz University of Technology, Austria

e-mail: markus.strohmaier@tugraz.at
web: <http://www.kmi.tugraz.at/staff/markus>

Web Science and Web Technology

- Welcome
- Motivation
- Introduction of Instructor and TA
- Course Modalities
- Background

The Web Today (2007)



(courtesy, www.opte.org)

Sept 2007, Netcraft

Knowledge Management Institute



Search (like it's 1997!)

[<http://web.archive.org/web/19981111183552/google.stanford.edu/>]

Google!

Search the web using Google!

10 results

Index contains ~25 million pages (soon to be much bigger)

[About Google!](#)

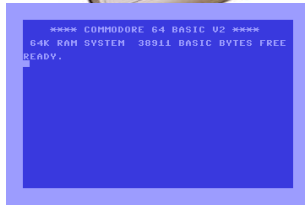
[Stanford Search](#) [Linux Search](#)

Get Google! updates monthly!

your e-mail [Archive](#)

Copyright ©1997-8 Stanford University

Computers - another 10 years back (1987)



„Web science? Can you say that again?“

Motivation

“[...] As the Web has grown in complexity and the number and types of interactions that take place have ballooned, it remains the case that we know more about some complex natural phenomena (the obvious example is the human genome) than we do about this particular engineered one.”

*A Framework for Web Science
T. Berners-Lee and W. Hall and J. A. Hendler and K. O'Hara and N. Shadbolt and D. J. Weitzner Foundations and Trends® in Web Science 1 (2006)*

Course team

- Instructor: Markus Strohmaier
- Teaching Assistant: Gabriele Zorn-Pauli

- e-mail addresses:
 - Markus.strohmaier@tugraz.at
 - gabriele.zorn-pauli@tugraz.at

About me

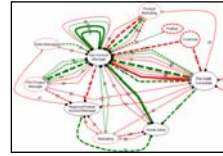
Education:

- 2002 - 2004 PhD. in Knowledge Management, Faculty of Computer Science, TU Graz
- 1997 - 2002 M.Sc., Telematik, TU Graz

Background:

- July 2007 - present: Ass. Prof. (Univ.Ass.), TU Graz, Austria
- 2006 - 2007 15 months Post-Doc, University of Toronto, Canada
- 2002 - 2006 Researcher, Know-Center, Austria

About me



Research Background:

- Business Process Oriented Knowledge Management
- Knowledge Infrastructure Development
- Agent-Oriented Early Requirements Engineering

Research Interests:

- Web Science with a focus on networks and Social Computing
- Intentional Structures and Representations on the Web

Interesting topics for projects, Bachelor / Master thesis:

- If you are interested in the topics of this course, it is likely that you are interested in doing a project / a thesis with me as well. **Contact me to discuss opportunities.**



Course Context

- 707.000 Web Science and Web Technology
 - 1st year
 - Has not been held before
- Part of „Software Engineering & Business“
 - Bachelor studies, 6th semester
 - Which means the course is usually held during summer semesters
- This course is a pilot
- Your feedback is appreciated

Course Organization and Logistics

- **Lectures**
Tuesdays 14:15 - 15:45,
October 2007 - January 2008,
Room HS Modul (Inffeldgasse 21a, Ground Floor)
- **Website:** http://kmi.tugraz.at/staff/markus/courses/707.000_web-science/
- **Newsgroup:** tu-graz.lv.web-science



Enroll!

In order to obtain a grade, you need to enroll for this course until Oct 10 2007 via TUG online!

- **Weekly Readings**
Password to access protected documents on the course website:

Grading

So how do you receive a grade in this course?

- 50% home assignments (25% pen & paper, 25% programming)
- Due dates for submission are announced on the course website
- 50% final exam
On 22.1. 2008, no aids are allowed

In order to successfully complete the course, you need to have a score of $\geq 51\%$

Alternative: You can apply for a project (limited availability)

- Work on **topics provided by the course** team
- Likely to be **more work** than home assignments and final exam
- Might be **more rewarding** for those students who want to dig deeper / already have knowledge about some of the topics

You can **cancel** your participation in this course until: 11.12.2007 (will not result in a negative grade)

Grading

The following weights will be assigned to home assignments and the final exam (totalling 100%):

- * Home assignment 1: 5%
- * Home assignment 2: 5%
- * Home assignment 3: 5%
- * Home assignment 4: 5%
- * Home assignment 5: 5%
- * Home assignment 6: 25%
- * Final Exam: 50%

Again, In order to obtain a positive grade, you need to have a total score of 51% or more.

Course Policies

- Class attendance and participation are mandatory
- Readings are to be done before class
- All assignments are due at the beginning of the class on the due date
- Deadlines are sharp
- Assignment descriptions and lecture notes will be made available on the web
- Citing Wikipedia
- Dishonesty (cheating, plagiarism)

For details see the course website:

http://kmi.tugraz.at/staff/markus/courses/707.000_we_b-science/

Course Topics

- World Wide Web
- What is network theory? Why is it relevant for the web?
- How do networks evolve?
- How do you search in networks?
- What are social parameters of networks?
- What are current web technologies?

But also e.g. a brief History of Smileys ;-)

Goals I

Understanding about and overview of basic

- Phenomena
- Theories
- Processes
- Methods
- Algorithms
- Representations

that are relevant in the context of the web.

What are your expectations?

Non-Goals

In the research community, there is **no broad consensus** regarding the theoretical foundations of a „Science of the Web“ yet

So therefore, the topics of this course are necessarily **subjectively selective**.

Instead of giving an authoritative account of web science, this course aims to give an overview of **prominent, interesting and/or powerful research results** generated by related fields so far.

Recommended Literature

There is no required text book for this course, however you might find it helpful to have a look at the following resources:

- [A Framework for Web Science](#), Tim Berners-Lee, Wendy Hall, James A. Hendler, Kieron O'Hara, Nigel Shadbolt, Daniel J. Weitzner, Foundations and Trends in Web Science, Vol 1 Nr 1, 2006
- [Group of University Researchers to Make Web Science a Field of Study](#), New York Times, Nov 2, 2006 (An easy-to-read motivation for Web Science)
- [Six Degrees - The Science of a Connected Age](#), [Duncan J. Watts](#), 2004
- [Web Dragons](#), [Ian Witten](#) et al, 2007
- [Social Network Analysis - Methods and Applications](#), Stanley Wasserman and Katherine Faust, 1995
- [Graph Theory](#), Reinhard Diestel, Electronic Edition, 2005 ([free PDF download](#))

Questions?

Raise them **NOW!**

Or ask them later:

- At the end of each class
- Via e-mail: markus.strohmaier @ tugraz.at

(now would be a good time though)

Let's start!
- Science and the Web -

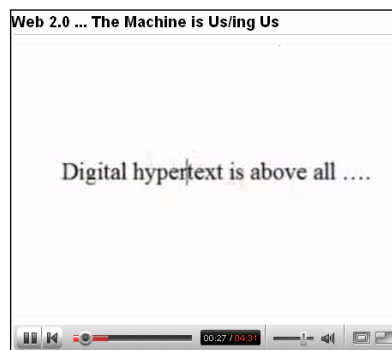
Motivation

A reported number of 900 Mio people (that is roughly one out of seven people on earth!) watched this video of a previously unknown, video amateur, teenage starwars fan:

http://entertainment.timesonline.co.uk/tol/arts_and_entertainment/tv_and_radio/article650932.ece


How is this possible? How does information spread on the web? What are the effects on individuals and society?

The Web Today




<http://www.youtube.com/watch?v=6gmP4nk0EOE>

How do the topics addressed in this movie relate to a Science of the Web?

Knowledge Management Institute			
Preliminary Course Schedule I/II			
Week	Date	Title, Links	Comments and Links
Week 1	2.10.2007	Introduction and Motivation: Web & Science (slides, home assignment 1)	In this class, we will discuss the course organization and provide a basic motivation for and introduction to the course. Readings: Web science: a provocative invitation to computer science, B. Shneiderman, Communications of the ACM 50 25--27 (2007) [Web link] Readings: Chapter 1 & 2, A Framework for Web Science, T. Berners-Lee and W. Hall and J. A. Hendler and K. O'Hara and N. Shadbolt and D. J. Weitzner Foundations and Trends® in Web Science 1 (2006) [Web link]
Week 2	9.10.2007	The Small World Problem home assignment 1 due (slides)	We will discuss several examples and research efforts related to the small world problem and set the ground for our discussion of network theory and social network analysis. Readings: An Experimental Study of the Small World Problem, J. Travers and S. Milgram Sociometry 32 425-443 (1969) [Protected Access]
Week 3	16.10.2007	Network Theory and Terminology (slides, home assignment 2)	In this class, we will discuss network theory fundamentals, including concepts such as diameter, distance, clustering coefficient and others. We will also discuss different types of networks, such as scale-free networks, random networks etc. Readings: Graph structure in the Web, A. Broder and R. Kumar and F. Maghoul and P. Raghavan and S. Rajagopalan and R. Stata and A. Tomkins and J. Wiener Computer Networks 33 309--320 (2000) [Web link]
Week 4	23.10.2007	Social Network Analysis home assignment 2 due (slides, home assignment 3)	What information can you get out of social graphs? We will discuss some basic principles of social network analysis.
Week 5	30.10.2007	Network Evolution home assignment 3 due (slides, home assignment 4)	In this class, we will discuss the nature of network evolution and some selected network processes.
Week 6	6.11.2007	Link Analysis home assignment 4 due (slides, home assignment 5)	What are ways of searching in graphs? In this class, we will discuss basics of link analysis, including Google's PageRank algorithm as an example. Readings: The PageRank Citation Ranking: Bringing Order to the Web, L. Page and S. Brin and R. Motwani and T. Winograd (1998) [Protected Access]

23

Knowledge Management Institute			
Preliminary Course Schedule II/II			
Week	Date	Title, Links	Comments and Links
Week 7	13.11.2007	Web Mining and Information Retrieval home assignment 5 due (slides)	This class introduces basics of web mining and information retrieval including an introduction to the Vector Space Model, Latent Semantic Indexing, Associative Retrieval and Support Vector Machines. Guest lecture: Michael Granitzer, Know-Center Graz
Week 8	20.11.2007	Webtechnologies I (slides, home assignment 6)	This class focuses on a selected subset of web technologies that are of current interest. Read: TBA
Week 9	27.11.2007	Metadata, Tagging and Folksonomies (slides)	In this class, we will discuss metadata as well as current phenomena such as tagging and folksonomies. Readings: P. Mika, Ontologies Are Us: A Unified Model of Social Networks and Semantics, International Semantic Web Conference, : 522-536, 2005. [Web link]
Week 10	11.12.2007	Trust and Reputation on the Web	TBA Readings: New Scientist article "Wikipedia 2.0 - now with added trust" [protected access]
Week 11	8.1.2008	User Intentions and Intentional Structures on the Web home assignment 6 due (slides)	Search engines - such as Google - have been characterized as " Databases of intentions ". This class will focus on different aspects of intentionality on the web, including goal mining, goal modeling and goal-oriented search. Readings: M. Strohmaier, M. Lux, M. Granitzer, P. Scheir, S. Liaskos, E. Yu, How Do Users Express Goals on the Web? - An Exploration of Intentional Structures in Web Search, We Know'07 International Workshop on Collaborative Knowledge Management for Web Information Systems in conjunction with WISE'07, Nancy, France, 2007. [Web link] Readings: Automatic identification of user goals in Web search, U. Lee and Z. Lu and J. Cho WWW '05: Proceedings of the 14th International World Wide Web Conference 391--400 (2005) [Web link]
Week 12	15.1.2008	Webtechnologies II (slides)	The semantic web represents a current research effort to increase the capability of machines to make sense of content on the web. In this class, Peter Scheir will give a guest lecture on the basic principles underlying the semantic web vision, including RDF, OWL and other standards. Guest lecture: Peter Scheir, Knowledge Management Institute, Graz University of Technology
Week 13	22.1.2008	Final Exam	No aids are allowed at the final exam.

Project Options

- Project 1: Analyzing the nature and proportion of intentional queries in large search engine logs
- Project 2: Algorithms for frame-based identification of goals in Natural Language Text
- Project 3: Intentional Metadata: Modeling goals with WSMO
- Project 4: Intentional Metadata: Modeling goals with Microformats
- Project 5: A Prototype for Intentional Query Expansion
- Project 6: Decentralized Intentional Query Expansion

A Brief Overview of the Web [Berners Lee et al 1994]

- Vision: the W3 operates without regard to
 - Where information is stored
 - How information is stored or
 - What system is used to manage it
- **Documents** referring to each other by **links**
- Analogy to spiders' construction: the web
- **Hypertext paradigm**
 - Sensitive parts of text representing links
 - A link is followed by mere pointing and clicking (or typing a ref. Nr.)
 - No primary focus on search
- Hypertext links may be made to any data in non-W3 servers (FTP, Gopher, WAIS or internet news) as W3 clients have the ability to present all such data as hypertext.
- The World Wide Web combines Hypertext and Search

the web != internet

The web: Presentation and Extraction [Berners Lee et al 1994]

The architecture of W3 (fig. 2) is one of browsers (clients) which know how to present data but not what its origin is, and servers which know how to extract data but are ignorant of how they will be presented. Servers and clients are unaware of the details of each other's operating system quirks and exotic data formats.

All the data in the Web is presented with a uniform human interface (Fig. 3). The documents are stored (or generated by algorithms) throughout the internet by computers with different operating systems and data formats. Following a link from the SLAC home page (the entry into the Web of a SLAC user) to the NIKHEF telephone book is as easy and quick as following the link to a SLAC Working Note.

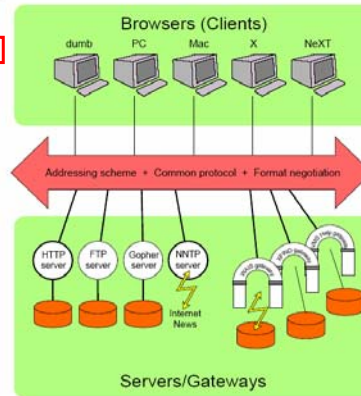


Fig. 2: Architecture of W3

The web [Berners Lee et al 1994]

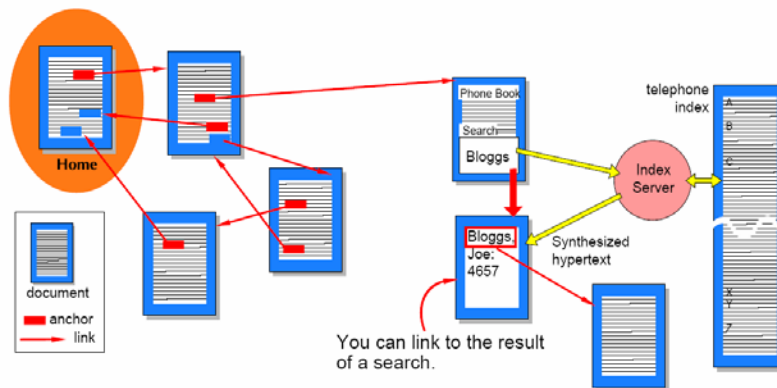


Fig 1. The basic hypertext model is enhanced by searches.

Features of the web [Berners Lee et al 1992]

Features to note are:-

- Information need only be represented once, as a reference may be made instead of making a copy;
- Links allow the topology of the information to evolve, so modeling the state of human knowledge at any time without constraint;
- The web stretches seamlessly from small personal notes on the local workstation to large databases on other continents;
- Indexes are documents, and so may themselves be found by searches, and/or following links. An index is represented to the user by a “cover page” which describes the data indexed and the properties of the search engine.
- The documents in the web do not have to exist as files; they can be “virtual” documents generated by a server in response to a query or document name. They can therefore represent views of databases, or snapshots of changing data (such as the weather forecast, financial information, etc).

Uniqueness

Networkability

Scalability

Indexability

Adaptability /
Customizability

Historical Vision of the Web

Is a space in which

- *Resources* are identified by Uniform Resource Identifiers (URIs)
- *Protocols* support interaction between agents (HTTP)
- *Formats* represent information resources (HTML)

URI

Uniform Resource Identifier

- Resources may be anything that can be linked to or spoken of
 - Resources can contain a reference to another resource
- *Identifiable*, but not necessarily *retrievable* (e.g. protected access)
- A single global system of identifiers
- Each URI ideally identifies a single resource in a context-independent manner
- URIs act as names and addresses
- URIs require institutions
 - E.g. the registry that handles domain names

HTTP & HTML: High Level Overview

<http://www.w3.org/Protocols/HTTP/HTTP2.html>

HTTP: A protocol that is basically stateless, a transaction consisting of

- Connection
 - The establishment of a connection by the client to the server - when using TCP/IP port 80 is the well-known port, but other non-reserved ports may be specified in the URL;
- Request
 - The sending, by the client, of a request message to the server;
- Response
 - The sending, by the server, of a response to the client;
- Close
 - The closing of the connection by either both parties.

HTML: A representation format

- Idea: Decoupling of content and representation
- Cues for graphical presentation of content

```
<div>
<map name="introLinks" id="intro
<div class="banner">
<span class="invisible"><a class
Visitors</a> | <a class="bannerL
<a class="bannerLink" title="Cov
```

Why Web Science?

“as the Web has grown in complexity and the number and types of interactions that take place have ballooned, it remains the case that we know more about some complex natural phenomena (the obvious example is the human genome) than we do about this particular engineered one.”


[Berners-Lee 2006]

Why Web Science?

- Dynamics and evolution
- The “deep web” (resources not available by robots)
- Sampling, lack of complete enumeration
- Scale (e.g. “What’s the percentage of web pages updated daily?”)
- Search (e.g. “What’s the percentage of web pages indexed by search engines?”)
- Web topology
- Artifacts of social interaction (weblogs, etc), web sociology
- ...

Science (in a nutshell)

University of Toronto
Department of Computer Science



Science and Theory

- **A (scientific) theory is:**
 - ↳ more than just a description - it explains and predicts
 - ↳ Logically complete, internally consistent, falsifiable
 - ↳ Simple and elegant.
- **Components of a theory:**
 - ↳ concepts, relationships, causal inferences
 - > E.g. Conway's Law- structure of software reflects the structure of the team that builds it. A theory should explain why.
- **Theories lie at the heart of what it means to do science.**
 - ↳ Production of generalizable knowledge
 - ↳ Scientific method ↔ Research Methodology ↔ Proper Contributions for a Discipline
- **Theory provides orientation for data collection**
 - ↳ Cannot observe the world without a theoretical perspective

© 2004-5 Steve Easterbrook. This presentation is available free for non-commercial use with attribution under a creative commons license.

What could theories for the web look like?

Some Simple Examples:

- Every page on the web can be reached by following less than 10 links. (True/False/Depends?)
- 1%-4% of users express their search queries in the form of goals such as "increase adsense revenue" (True/False/Depends?)
- The average number of words per search query is more than 3 (True/False/Depends?)
- A wikipedia page contains, on average, 0.03 false facts (True/False/Depends?)

Can these statements be easily validated? Are these good theories? What constitutes good theories?

Some Quality Characteristics of Theories

- Clarity
- Simplicity
- Predictive Power
- Explanative Power
- Utility
- Testability
- Falsifiability (vs. Falsification)

Science (in a nutshell)

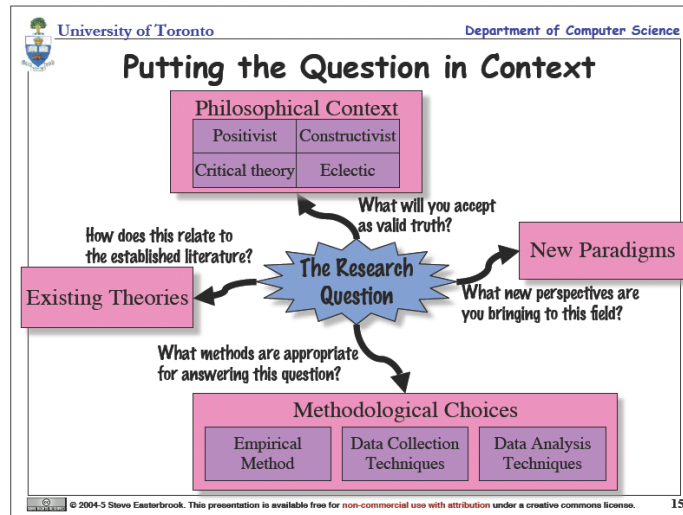
University of TorontoDepartment of Computer Science

What type of question are you asking?

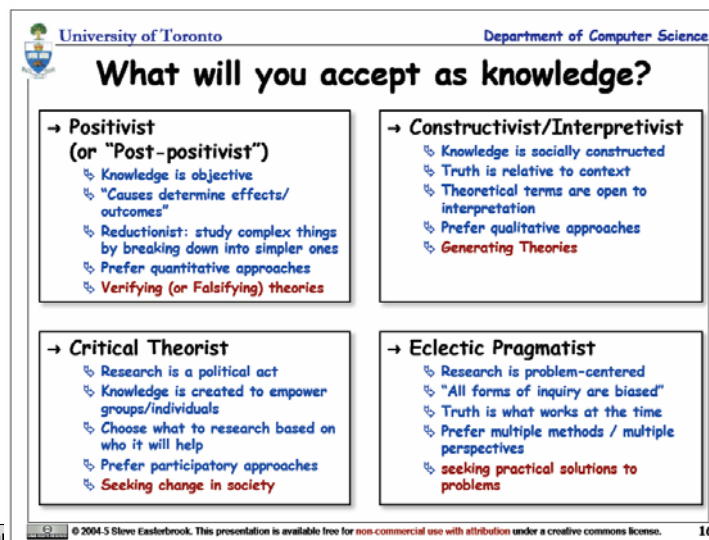
<p>→ Existence:</p> <ul style="list-style-type: none"> ↳ Does X exist? <p>→ Description & Classification</p> <ul style="list-style-type: none"> ↳ What is X like? ↳ What are its properties? ↳ How can it be categorized? ↳ How can we measure it? ↳ What are its components? <p>→ Descriptive-Process</p> <ul style="list-style-type: none"> ↳ How does X work? ↳ What is the process by which X happens? ↳ In what are the steps as X evolves? ↳ How does X achieve its purpose? <p>→ Descriptive-Comparative</p> <ul style="list-style-type: none"> ↳ How does X differ from Y? 	<p>→ Relationship</p> <ul style="list-style-type: none"> ↳ Are X and Y related? ↳ Do occurrences of X correlated with occurrences of Y? <p>→ Causality</p> <ul style="list-style-type: none"> ↳ Does X cause Y? ↳ Does X prevent Y? ↳ What causes X? ↳ What effect does X have on Y? <p>→ Causality-Comparative</p> <ul style="list-style-type: none"> ↳ Does X cause more Y than does Z? ↳ Is X better at preventing Y than is Z? ↳ Does X cause more Y than does Z under one condition but not others? <p>→ Design</p> <ul style="list-style-type: none"> ↳ What is an effective way to achieve X? ↳ How can we improve X?
---	---

© 2004.5 Steve Easterbrook. This presentation is available free for non-commercial use with attribution under a creative commons license.

Science (in a nutshell)



Science (in a nutshell)



Science (in a nutshell)

University of Toronto
Department of Computer Science

Validity

← Back

- **Construct Validity**
 - ↳ Theoretical concepts are operationalized and measured correctly.
 - ↳ Are we measuring the construct we intended to measure?
 - ↳ Did we translate these constructs correctly into observable measures?
 - ↳ Did the metrics we use have suitable discriminatory power?
- **Internal Validity**
 - ↳ Do the results really follow from the data?
 - ↳ Have we properly eliminated any confounding variables?
- **External Validity**
 - ↳ Are the findings generalizable beyond the immediate study?
 - ↳ Do the results support the claims of generalizability?
- **Empirical Reliability**
 - ↳ Demonstrate that the study can be repeated with the same results?
 - ↳ Did we eliminate all researcher biases?

© 2004-5 Steve Easterbrook. This presentation is available free for non-commercial use with attribution under a creative commons license. 37

Networks

A significant part of this course will focus on network theory.

- Graph theory vs. Network theory
 - Graph theory is largely mathematical while network theory also focuses on networks that can be observed in the „real world“
 - Network theory puts emphasis on evolution
- There are many different forms of networks available on the net

– Can you name a few of them?



Delicious as a Network of tags

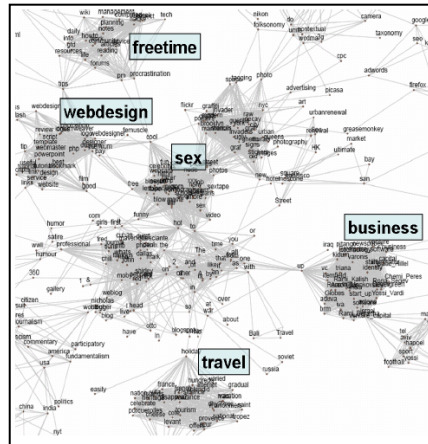


Table 1. The five main clusters of interest based on the Concept-Object network

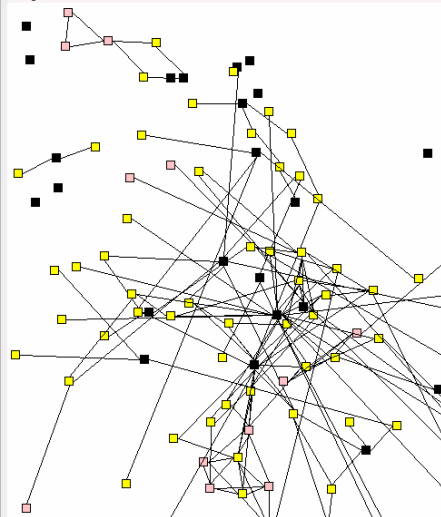
travel	cote, provence, villa, azur, mas, holiday, vacation, tourism, france, heritage
business	venture_capital, enterprise, up, start, venture, newspaper, capital, Segev, pitango, vc
free time	procrastination, info, advice, gtd, life, notes, planning, daily, reading, forums
sex	hot, to, street, pictures, on, photos, free, celeb, adult, lesbian
web design	design, designer, webdesign, premium, logo, logos, dreamweaver, templates, best, good

Fig. 1. The delicious tags associated through co-occurrence on items and the clusters emerging

Mark

The Blogosphere as a Network of Blog Posts

In A model (framework) for weblog research it was suggested that one should look at five dimensions to study weblogs. This post shows that one can obtain a fascinating peek into the blogosphere by looking at just two dimensions [links, persons]. Perhaps it is an idea to also add time so that we can see whether the yellow and pink posts occur before (this is possible), during or after the conversation.

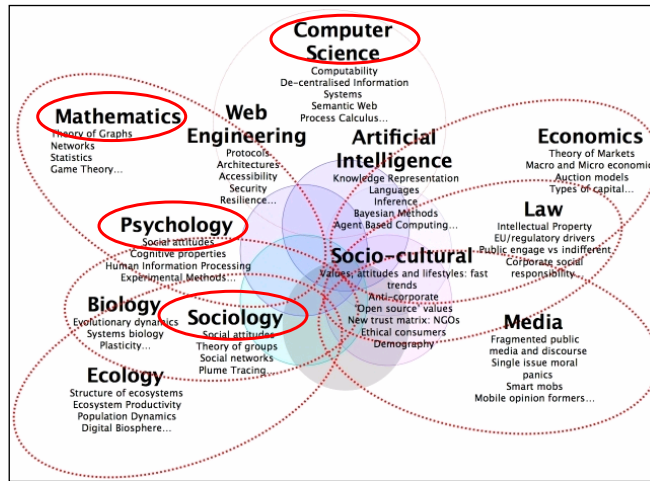


Courtesy of <http://anjo.blogs.com/>

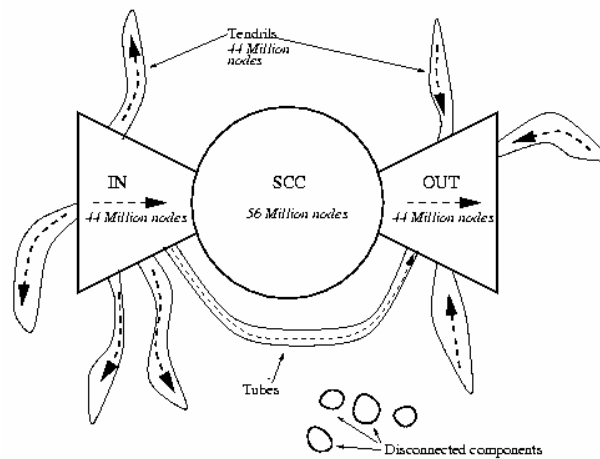
Markus Strohmaier

Web Science

www.webscience.org



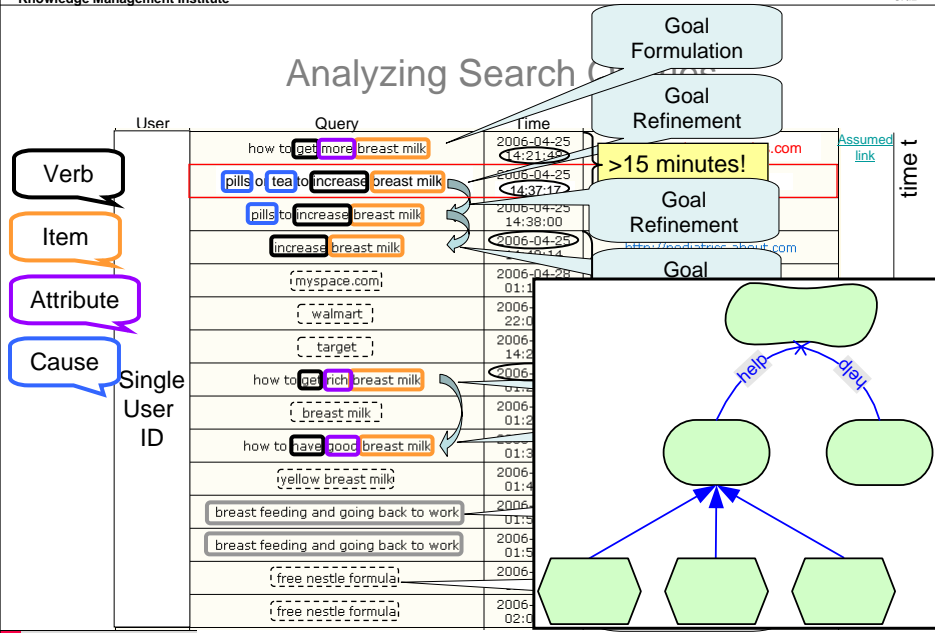
Some Course Highlights



Some Course Highlights



Analyzing Search



Some Course Highlights

The screenshot shows the Wikipedia article for "Kalahari Desert". The article text states: "The **Kalahari Desert** is a large sandy area in southern **Kgalagadi Africa** extending 900,000 km² (352,500 sq. mi.), covering much of **Botswana** and parts of **Namibia** and **South Africa**, as semi-desert, with huge tracts of excellent grazing after good rains. The Kalahari Desert is in Africa at the southern part and the desert is a portion of desert and a plateau. The Kalahari supports some animals and plants because most of it is not true desert. There are small amounts of rainfall and the summer temperature is very high. It usually receives 5-10 inches of rain per year.^[1] The surrounding **Kalahari Basin** covers over 2.5 million km² extending farther into Botswana, Namibia and South Africa, and encroaching into parts of **Angola**, **Zambia** and **Zimbabwe**. The only permanent river, the **Okavango**, flows into a delta in the northwest, forming marshes that are rich in wildlife. Ancient dry riverbeds—called **omuramba**—traverse the Central Northern reaches of the Kalahari and provide standing pools of water during the rainy season. Previously havens for wild animals from elephant to giraffe, and for predators such as lion and cheetah, the riverbeds are now mostly grazing spots, though leopard or cheetah can still be found.

The article also includes a table of contents with the following items:

- 1 Climate
- 2 Game reserves
- 3 Kalahari minerals
 - 3.1 Diamond mining
 - 3.2 Sand mining
- 4 Administrative areas covering the Kalahari
- 5 The Kalahari desert in popular culture
- 6 See also
- 7 Notes
- 8 External links

There is also a map showing the location of the Kalahari Desert and Kalahari Basin in southern Africa.

Check

- Is there anything else you want to know w.r.t. this course?
- What aspects are you most interested in?
- Anything else?

Any further questions?

Have a good start in the new semester!
- See you next week