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

Search (like it’s 1997!)

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

About Google!

Stanford Search Large Search
Get Google! updates monthly!

your-emai Subscribe Authors

Copyright ©1997-2007 Stanford University

Markus Strohmaier 2007
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
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/](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_web-science/](http://kmi.tugraz.at/staff/markus/courses/707.000_web-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 ;-)
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:

- 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?
# Preliminary Course Schedule II/II

## Week 1
- **Date:** 2.10.2007
- **Title:** Introduction to Web Science
- **Assignments:**
  - Lecture 1: Introduction to Web Science
  - Lecture 2: A Framework for Web Science

## Week 2
- **Date:** 9.10.2007
- **Title:** The Small World Problem
- **Assignments:**
  - Lecture 3: The Small World Problem
  - Lecture 4: An Experimental Study of the Small World Problem

## Week 3
- **Date:** 16.10.2007
- **Title:** Network Theory and Technology
- **Assignments:**
  - Lecture 5: Graph Theory Basics
  - Lecture 6: Advanced Network Analysis

## Week 4
- **Date:** 23.10.2007
- **Title:** Network Theory and Technology
- **Assignments:**
  - Lecture 7: Social Network Analysis
  - Lecture 8: Network Evolution

## Week 5
- **Date:** 30.10.2007
- **Title:** Link Analysis
- **Assignments:**
  - Lecture 9: Link Analysis
  - Lecture 10: Social Network Analysis

## Week 6
- **Date:** 6.11.2007
- **Title:** Web Mining and Information Retrieval
- **Assignments:**
  - Lecture 11: Web Mining
  - Lecture 12: Information Retrieval
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 return 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 hypernet 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).

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
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
- …
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)

What type of question are you asking?

- Existence
  - Does X exist?
- Description & Classification
  - What is X like?
  - What are its properties?
  - How can it be categorized?
  - How can we measure it?
  - What are its components?
- Descriptive-Process
  - 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?
- Descriptive-Comparative
  - How does X differ from Y?
- Relationship
  - Are X and Y related?
  - Do occurrences of X correlated with occurrences of Y?
- Causality
  - Does X cause Y?
  - Does X prevent Y?
  - What causes X?
  - What effect does X have on Y?
- Causality-Comparative
  - 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?
- Design
  - What is an effective way to achieve X?
  - How can we improve X?
Science (in a nutshell)

Putting the Question in Context

Philosophical Context
- Postivist
- Constructivist
- Critical theory
- Eclectic

How does this relate to the established literature?

What will you accept as valid truth?

New Paradigms

The Research Question

What new perspectives are you bringing to this field?

Existing Theories

What methods are appropriate for answering this question?

Methodological Choices
- Empirical Method
- Data Collection Techniques
- Data Analysis Techniques

Science (in a nutshell)

What will you accept as knowledge?

→ Postivist
  (or "Post-positivist")
  - Knowledge is objective
  - "Causes determine effects/outcomes"
  - Reductionist: study complex things by breaking them into simpler ones
  - Prefer quantitative approaches
  - Verifying (or falsifying) theories

→ Constructivist/Interpretivist
  - Knowledge is socially constructed
  - Truth is relative to context
  - Theoretical terms are open to interpretation
  - Prefer qualitative approaches
  - Generating Theories

→ Critical Theorist
  - Research is a political act
  - Knowledge is created to empower groups/individuals
  - Choose what to research based on who it will help
  - Prefer participatory approaches
  - Seeking change in society

→ Eclectic Pragmatist
  - Research is problem-centered
  - "All forms of inquiry are biased"
  - Truth is what works at the time
  - Prefer multiple methods/multiple perspectives
  - Seeking practical solutions to problems
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?
The Web as a Network of Related Sites

http://www.touchgraph.com/TGGoogleBrowser.html
(based on Google's „related sites“ functionality)

The Web as a Network of Search Results

http://www.kartoo.com (search for „web2.0“)
Delicious as a Network of tags

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

<table>
<thead>
<tr>
<th>Clusters</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>travel</td>
<td>route, provence, villa, azur, mas, holiday, vacation, tourism, france, heritage</td>
</tr>
<tr>
<td>business</td>
<td>venture, capital, enterprise, up, start, venture, newspaper, capital, Segev, pitango, vc</td>
</tr>
<tr>
<td>free time</td>
<td>procrastination, info, advice, gtd, life, notes, planning, daily, reading, forums</td>
</tr>
<tr>
<td>sex</td>
<td>hot, to, street, pictures, on, photos, free, celeb, adult, lesbian</td>
</tr>
<tr>
<td>web design</td>
<td>design, designer, webdesign, premium, logo, logos, dreamweaver, templates, best, good</td>
</tr>
</tbody>
</table>

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

The Blogosphere as a Network of Blog Posts

It's possible to represent the blogosphere in different ways. In this example, I used a technique to study the blogosphere. This plot shows that new connections between blogs emerge all the time. Perhaps it is due to people being able to add new content to the blogosphere. Perhaps it is due to people being able to add new content to the blogosphere.
Some Course Highlights
Some Course Highlights

Analyzing Search Queries

Verb
Item
Attribute
Cause
User ID

Goal Formulation
Goal Refinement
Goal Refinement
>15 minutes!
Goal Refinement
Goal

Some Course Highlights

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