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, Netscraft

Search (like it's 1997!)

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

About Google!

Stanford Search Laser Search
Get Google! updates monthly!

Copyright ©1997-2001 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
Course team

- Instructor: Markus Strohmaier
- Teaching Assistants (planned):
  - Gabriele Zorn-Pauli
  - Christian Körner
  - Christoph Zehentner

- e-mail addresses:
  - markus.strohmaier [at] tugraz.at
  - gabriele.zorn-pauli [at] tugraz.at
  - christian.koerner [at] student.tugraz.at
  - christoph.zehentner [at] student.tugraz.at

- For general, course- or assignment-related questions, please use the newsgroup

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 been held before once
• Part of „Software Engineering & Business“
  – Bachelor studies, 6th semester
  – Which means the course is usually held during summer semesters
  – Basic programming skills are required
• This year is a pilot
• Your feedback is appreciated
Feedback from Fall 2007/08 students

[Auszug aus TUGonline]

Haben Sie weitere Anmerkungen zu dieser Lehrveranstaltung?

- Die Vorlesung die auf Englisch gehalten wurde, war meiner Meinung nach: 1) leichter zu verstehen, da mit den Folien übereinstimmend 2) dadurch besser, da durch die notwendige erhöhte Aufmerksamkeit auf Grund der Fremdsprache eine intensivere Auseinandersetzung mit dem Stoff gegeben war 3) Dadurch auch etwas anstrengender: Pause!
- zu langsamer Vortrag... Folien in Englisch Sprache Vorlesung aber auf Deutsch gehalten
- Sehr interessante Lehrveranstaltung, für einen „Testballon“ ausgezeichnet, bitte unbedingt weiterführen!
- Viele spannende Themen extrem langweilig präsentiert oder nur nebenbei erwähnt, teilweise uninteressante Themen zu stark behandelt.
- zu sehr auf Graphentheorie und Netzwerke beschränkt, das Thema würde ein breiteres Spektrum bieten, rhetorische Mankos des Vortragenden
- HA und Vortrag sehr gut abgestimmt. Inhalt wird in den Übungen wiederholt; trägt zum Verständnis bei.
- Es sollte in Zukunft alles in deutscher Sprache vorgetragen werden. Folien können weiterhin ruhig in englischer Sprache sein... Im Großen und Ganzen hat mich diese LV sehr interessiert und auch der interaktive Vortragsstil hat mir sehr gut gefallen.

Course Organization and Logistics

- Lectures
  Mondays 12:15 - 13:45,
  Feb 2008 - June 2008,
  Room HS i12 (Inffeldgasse 16b, Ground Floor)

- Website: http://kmi.tugraz.at/staff/markus/courses/707.000_web-science/
- Newsgroup: tu-graz.lv.web-science
  - Please use the newsgroup for all questions related to the course

Enroll!

In order to obtain a grade, you need to enroll for this course until March 2 2008 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 23.6. 2008, no aids are allowed

- **Prerequisites:** Course „Einführung in die Strukturierte Programmierung“, Java Programming Knowledge

In order to successfully complete the course, you need to have a score of >= 51%

You can cancel your participation in this course until: March 31 2008 (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%
  - You will analyze a social network of which you are a part yourself
* 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/

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 ;-)  
Simulations: e.g. http://cmol.nbi.dk/javaapp.php
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?

Preliminary Course Schedule I/II

<table>
<thead>
<tr>
<th>Week</th>
<th>Date</th>
<th>Topic</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25.2.2008</td>
<td>Introduction and Motivation: Web I &amp; II</td>
<td>Understanding about and overview of basic phenomena.</td>
</tr>
<tr>
<td>1</td>
<td>3.3.2008</td>
<td>The Social World Problem</td>
<td>Understanding about and overview of basic theories.</td>
</tr>
<tr>
<td>5</td>
<td>10.3.2008</td>
<td>Network Themed and Technologies</td>
<td>Understanding about and overview of basic processes.</td>
</tr>
<tr>
<td>4</td>
<td>7.4.2008</td>
<td>Social Network Analysis</td>
<td>Understanding about and overview of basic methods.</td>
</tr>
<tr>
<td>5</td>
<td>14.4.2008</td>
<td>Network Analysis and Processes</td>
<td>Understanding about and overview of basic algorithms.</td>
</tr>
<tr>
<td>4</td>
<td>21.4.2008</td>
<td>Link Analysis and Searchs</td>
<td>Understanding about and overview of basic representations.</td>
</tr>
</tbody>
</table>

For further information, please refer to the course materials available on the course website.
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 newsgroup

(now would be a good time though)
Let’s start!
- Science and the Web -

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?
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?

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).

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: 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?

- 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)

- 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.
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)
Example: Analyzing Search Queries

User Visited URL

Single User ID

Verb

Item

Attribute

Cause

Goal Refinement

Goal Formulation

Goal Refinement

Goal


Verb

Item

Attribute

Cause

Goal Refinement

Goal Formulation

Goal Refinement

Goal

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?
  - How does the process by which X happen?
  - 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?
Networks

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

• Graph theory vs. Network theory
  – While graph theory focuses on mathematics, network theory focuses on networks that can be observed in the "real world"
  – Evolution of networks

• 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

| travel  | note, provence, villa, azur, mas, holiday, vacation, tourism, france, heritage |
| business| venture, capital, enterprise, up, start, venture, newspaper, capital, Segev, pitango, ve |
| 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

The Blogosphere as a Network of Blog Posts

Courtesy of http://anjo.blogs.com/
Some Course Highlights

- Some Course Highlights

- Some Course Highlights
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?
Home Assignment 1

• A) Join the social network at webscience2008.ning.com (you can use a pseudonym without a photo) and add a relation to me (required). Add relations to colleagues you know and/or Gabi (as you wish). Add at least one geographic location in the „interests“ section of your profile („skiing in Schladming“)

• B) Find at least one scientific, and one popular-science or news articles related to „Web Science“
  – two in total, papers listed on the course website are excluded, papers may not have „Web Science“ in the title or text, but you need to make it clear why you think they are related to the topics of this course

• Read and try to understand it

• Try to answer questions, such as
  – What is Web Science? Why is it important? What can it contribute to our knowledge of the world? Why is it difficult to define it? To which other disciplines would a Science of the Web be related?

• Write down your thoughts in a 250-300 word paragraph

• Send your two articles (URLs), your paragraph, your Ning user name and your Matr. Number to gabriele.zorn-pauli [at] tugraz.at in PDF

• You are required to format the subject of your e-mail in the following way: 
  “[WSWT] HW1 YOUR_MATR_NR” e.g. “[WSWT] HW1 0799999” (only then will you receive a confirmation).

• We will discuss your views and opinions at the beginning of next week’s course

• Deadline

Any further questions?

Have a good start in the new semester!

- See you next week