Aalto University
School of Science

SeCo

WWW today

CS-E4410 Semantic Web, 9.1.2019

Eero Hyvönen
Aalto University, Semantic Computing Research Group (SeCo) *http://seco.cs.aalto.fi*
University of Helsinki, HELDIG *http://heldig.fi*

eero.hyvonen@aalto.fi

Outline

- Background of the World Wide Web
- Services on the web
- Knowledge representation
- Web programming
- Megatrends of the web

Background: dimensions of the web

Users

- Billion users in 2005
- 2 billion users in 2011
- 3 billion users in 6/2014
- 3.8 billion user 1/2018

Page amount indexed by search engines

- Magnitude: tens of billions of pages
  - *Google's index ca. 50 billion pages*

In addition: ”hidden/deep web”

- Databases not reachable by public search engines

Extremely effective publishing channel

- All information readable by everyone
- New content easy to publish to billions of people
- Usage is almost ”free”

Holy trinity of the WWW

URI addresses: resources

- Web sites, documents, pictures, etc.

HTML language

- Representing the WWW pages
- Hyperlinks

HTTP etc. protocols

- Transferring web resources between server and client

Services on the web

Functional services

- Banking, stores, govermental bureaus, etc.

Information retrieval services

- Search engines (e.g., Google) and browsing
- Portals, directories ← Focus of this course
- Databases in different applications

Information retrieval challenges on the web: end-user perspective

Ease of formulating search queries

- Creating queries that work as intended

The quality of the search results

- Recall: How many % of the relevant information is found
- Precision: How many % of the found information is relevant
- Relevance: How well do the results correspond to the user needs
  - *E.g., Google's PageRank algorithm*

Presentation of the search results

- Ease of understanding
- Ranking and structuring

Examples of the limitations of basic text search

Search term may appear in an irrelevant document

- "This page *does not discuss* **politics**"

Identifying synonyms

- Venus =/= Morning star =/= Evening star
- The change of names: Tanja Vienonen -> Karpela -> Saarela -> ?
  - *Bad recall, relevant pages are not found*
  - *Formulation of queries is difficult*

Identifying homonyms

- Varkaus -> event (theft), a Finnish city
- Nokia -> company, city, person, animal (sable)
  - *E.g., "nokia": pages about the animal are mixed with the ones about the company*
- Pyhäjärvi ("Holy lake") -> 49 places in Finland
  - *Bad precision, results are garbage*
  - *Understanding the results is difficult*
  - *Formulation of queries is difficult*

Examples of the limitations (2)

Computer does not understand relations between concepts

- Narrower-broader concept, part-whole
- E.g., query: “Helsinki” & “restaurant”
    - *Are “pizzerias” in “Kallio” and “Punavuori” found?*
- Background knowledge and “common sense” is missing
    - *Search with term “smoke” does not necessarily return pages about “fire”*

The information searched for is fragmented, but results cannot be aggregated

- E.g., “search publications of the members of the research group X”

Examples of the limitations (3)

Finding relations between information resources is challenging

- E.g., "How is Sibelius related to the city of Hämeenlinna?"
- The result is a set of separate pages that the user has to analyze

Search does not actually solve problems, "web of wisdom"

- How much does a kilogram of feathers weigh in the moon?
- With lots of information, the problem solving resembles remembering!
  - *"Who is the father of the daughter of Tarja Halonen?"*
  - *"Why is the All Saints' Day celebrated?"*

No sufficient personalization and utilization of the context

- What could I do today in London?

Examples of the limitations (4)

Finnish is especially challenging due to word forms, deriatives and compound words

- "yö" vs. "öinen" vs. "öistä" ("night", "nightly", "of nightly/nights")
- hypätä, hypyttää, hypähtää, hypähdellä, hypäyttää, … ("to jump")
- Kolmivaihekilowattituntimittari ("three-phase electricity meter")
- Kylmäsavulohiraejuustotagliatelle (recipe from the "Vartti" newspaper)

The biggest problem, however, is the computer's inability to "understand" the meaning of contents, semantics

- Current search engines search for words (text strings) instead of senses (what do the words mean)
- If a computer does not "understand", it cannot serve intelligently

Browsing challenges in the web: end-user perspective

Understanding the ”big picture” in a large fragmented information space

- ”Lost in the hyperspace”

Links get out of date and destroyed

- The linked target pages expire or are removed entirely
- New pages do not get linked to old ones
- Old pages do not get linked to new ones

Reliability of information and their providers

- ”Web of trust”
- ”Flat Earth” organization’s page vs. Aalto University’s scientific page
- Wikipedia vs. Encyclopedia Britannica

Knowledge management challenges: information provider perspective

Structuring contents with links is manual work

- Information does not get linked at content level without human effort

Different organizations create overlapping information

- The same work is done multiple times

The contents and their structures are not interoperable

- E.g., aggregation of collections of different memory organizations is difficult
- Lack of interoperability prevents combining of contents
- Lack of interoperability prevents the management of contents

Information about the contents and their changes is not communicated between organizations

- Often they don't even know about each other

Knowledge representation on the web

The idea of markup languages: HTML, XML, …

Domain- and environment-independent standard for documents

- Creation
- Management
- Transferring

Documents are text files

- Open, simple format
- Usable on all HW/SW platforms
- Easy to modify, store, read, transfer
- Future-proof

Markup languages

The idea is to separate structure, content, and presentation

- Describing the document structure (programmer)
  - *E.g., HTML: <H1>Heading</H1>*
- Describing the information content (programmer)
  - *E.g., XML: <ADDRESS>Otaniementie 17</ADDRESS>*
- The presentation is decided by the reader (browser)
  - *E.g., PC, mobile phone*

Why XML?

Different presentations for same content

- Different devices (PC, mobile phone, …)
- Different applications (WWW page, printed book, …)

Utilization of the content structure

- E.g., better precision/recall in search engines

Quality control

- Syntax validation is possible

Importance of markup languages

XML languages are used widely on the web

- Knowledge encoded in *open* format
  - *Lots of standards for different domains*
- *Open* APIs for programming languages (e.g., Java)
  - *Programmatic processing of the pages*

Vendor-independency

Stability against the change of file formats

- Pages are simple text files

Domain-specific standard languages

Standardization

General coordination of the development of the WWW

- World Wide Web Consortium (W3C) (www.w3.org)
  - *Cooperation body of manufactures, operators, etc.*
  - *Creates WWW recommendations*

Domain-specific organizations

- ISO: different domains, excluding electrical/electronical
- IEC https://www.iec.ch/ , CEN https://www.cen.eu/, UN/CEFACT https://www.unece.org/cefact/, OASIS https://www.oasis-open.org/, …
- Countless number of work groups on different domains

Challenges of markup languages

Complex for humans to read and process

- Not especially human-friendly notation

Repetition

- Includes unnecessarily lots redundancy (e.g., start and end tag), which magnifies the size of the markup
  - *Laborious to write*
  - *Needs bandwidth for transferring*

More recent movements

JSON JavaScript Object Notation

- Knowledge representation as hierarchical key-value pairs
- Integrated into JavaScript: easy/efficient to use
- Widely used
- Used also on the Semantic Web: e.g., JSON-LD notation

Simple Semantic Web notations for knowledge representation

- Turtle, OWL notations, etc. (we'll return to this on later lectures)
- Widely used

Web programming

Types of web programming

Client-side application programming (WWW browser)

- Distributed functionality

Server-side application programming (WWW server)

- Centralized functionality

Client-side web programming

Java applets

- Java program is read from the server into the browser
- The program is ran in the client machine

Dynamic HTML

- ECMAScript (JavaScript, JScript)
  - *Executable programs inside the HTML markup (script)*
- Cascading Style Sheets (CSS)
  - *General style definitions for HTML language elements*
- Domain Object Model (DOM)
  - *Object model of the page for scripts to read and manipulate*

AJAX (Asyncronous JavaScript and XML, 2005)

- Interactions with the server via function calls without reloading the page
- Enables mashups and sharing functionalities (e.g., Google Maps)

Server-side web programming

CGI scripts and servlets

- Program on the server
- Gets information from the browser, e.g., via a form
  - *GET, POST, PUT, etc. methods*
- Returns a HTML result page to the browser

Server Side Includes (SSI)

- Code snippets in a HTML template that are replaced with content
  - *E.g., date or other dynamic part of the document*
  - *The server executes the code snippet and replaces it with the result before returning the page*

Server-side web programming (2)

Server Side Scripting (ASP, JSP, PHP, …)

- HTML page with program code
- Code is executed and replaced with HTML results
- A server-side program generates the HTML pages
  - *E.g., querying information from a database*
- The result is sent to the browser

Tag and template libraries, application frameworks

- Templates and helpers for generating HTML markup
- Support for application architectures, e.g., MVC Model-View-Controller
- AngularJS, React, Vue.js, Django, Drupal,…

Megatrends of the web

Megatrends of the web

1. ***Contents* are enriched semantically (Semantic Web)**
   - *Semantic Web, Linked Data / Web of Data*
2. **D*ynamic* processing is increasing (Web Services)**
   - Web services, agent technologies
   - Adaptability and context sensitivity
   - Ambient computing, ubiquitous computing
   - Personalization
3. ***Community*-generated contents (Web 2.0)**
   - Distributed creation of contents that are linked together
   - Real-time services
4. ***Volume* is increasing (Big Data)**
5. ***Openness* is increasing (Open Data)**