World Wide Web Today
- An Introduction

13.1.2021

Eero Hyvönen
Aalto University, Semantic Computing Research Group (SeCo)
University of Helsinki, HELDIG
https://seco.cs.aalto.fi/u/eahyvone/
eero.hyvonen@aalto.fi

http://seco.cs.aalto.fi
http://heldig.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
- **4.66 billion active users in 2020 (59% of global population)**

Pages on the Web

- **Ca 55 billion pages indexed in 2020**

In addition: ”hidden/deep Web”

- Databases not reachable by public search engines

Extremely effective publishing channel

- Massive amounts of information available for everyone
- New content easy to publish to billions of people
- Usage is almost ”free”

Holy Trinity of the WWW Technology

URI identifiers of web resources (URL, URN, …)

- Global unique identifiers and addresses for anything on the Web
  - Web sites, pages, pictures, videos, concepts, data elements, etc.

HTML language

- Representing the WWW pages
- Hyperlinks between pages

HTTP protocol

- Transferring Web resources between server and client
- Basis for applications

Services on the Web

Functional services

- Banking, web stores, public services, etc.

Information retrieval services

- Search engines (e.g., Google) and browsing
- Data aggregation services
  - Portals, directories, …
- Data services in different applications

Information Retrieval Challenges on the Web: End-user Perspective

Problems of formulating search queries

- Creating queries that work as intended

The quality of the search results can be low

- 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 ranks results according to their relevance*

Presentation of the search results

- Listing and structuring the results in useful ways
- Explaining the results to end-users

Challenges of Text Search

Examples of the Limitations (1)

Search term may appear in an irrelevant document
- “This page *does not discuss* **politics**“

Identifying synonyms (different terms for the same thing)
- Venus =/= Morning star =/= Evening star
- The change of person names: Tanja Vienonen -> Karpela -> Saarela -> ?
  - *Bad recall, relevant pages are not found*
  - *Formulation of queries is difficult*

Identifying homonyms (same term for different things)
- 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
  - *Low 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 (a 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

Challenges of Browsing the Web

Browsing challenges in the web: end-user perspective

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

- ”Lost in the hyperspace”

Links missing and 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 removed and or 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

Aalto University
School of Science

SeCo

The idea of markup languages: HTML, XML, …

Domain- and environment-independent standards 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 format

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*

XML – Lingua Franca of the Web

Meta-language for defining mark-up languages such as HTML

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

- Data can be encoded in documents in *open* formats
- APIs available for programming languages (e.g., Java)
  - *Programmatic processing of the documents*

Vendor-independency

Stability against the changes of file formats

- Documents are simple text files

Lots if domain-specific XML languages available for applications

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 and Python: 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 using Web Documetns

Aalto University
School of Science

SeCo

Types of Web Programming

Client-side application programming (WWW browser)

- Distributed functionality
- HTML, DOM, CSS, JavaScript, AJAX, JSON, …

Server-side application programming (WWW server)

- Centralized functionality
- Node.js, MongoDB, …

Full Stack programming

- Integrated client-side and server-side programming
- E.g., Full Stack JavaScript

Megatrends of the Web

Megatrends of the Web

1. **Structured data on the Web increasing (Semantic Web)**
   - *Linked Data / Web of Data that machines "understand"*
   - *Basis for Artificial Intelligence based systems*
2. ***Dynamic* processing is increasing (Web Services)**
   - Web services, agent technologies
   - Adaptability and context sensitivity
   - Ambient computing, ubiquitous computing
   - Personalization
3. ***User-created content is* increasing (Web 2.0)**
   - Distributed creation of contents that are linked together
   - Wikipedia, Facebook, Twitter, YouTube, …
4. ***Volume, Velocity, Variety, Veracity and Value* are increasing (Big Data)**
5. ***Openness* is increasing (Open Data)**

More Information

World Wide Web Consortium technical standards and practices:

https://www.w3.org/

Roadmap & tutorials for Web development/programming:

https://www.w3schools.com/whatis/default.asp

Historical perspective to the Web

Tim Berners-Lee and Mark Fichetti: Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web. HarperCollins, New York, 2000.
https://www.amazon.com/Weaving-Web-Original-Ultimate-Destiny/dp/006251587X