Web Science and Engineering 2015

Homework 1

Eric Camellini
4494164
e.camellini@student.tudelft.nl

1. FIRST PART

a) What is "Hypertext"?

Hypertext is text that can contain interactive links. In the Web (the most famous hypertext implementation) these links are a redirection to other text, that can be also in another page, but in general the concept of hypertext describes more functions like back links, type information or text editing and expansion.

b) What problem was hypertext aiming to solve, what challenges was hypertext aiming to meet?

Hypertext aim was to create an interactive interface for text and data in order to allow users to easily navigate through the stored information, to add new content and at the same time to link it to other already existing text. Hypertext was born to solve the problem of the lack of a text browsing user-interface. This kind of interface is necessary in a scenario where a lot of data and literature are written and stored, and must be easily accessed and referenced (e.g. in a research oriented scenario like CERN, where Tim Berners-Lee experienced the need of a similar tool).

c) Give min. three examples of research questions that were relevant in hypertext research before the Web.

- What information does a link need to include in order to make the user decide if it is worth to browse the linked content ?
- Can hypertext become the new referencing standard in scientific literature ? What are the requirements of an hypertext implementation that has this objective ?
- Can hypertext become a counter-productive distraction for the reader ? Statistically, which is the maximum percentage of links over the total amount of text that allows to avoid this effect ?

Questions inspired by [1]

d) Explain why Tim Berners-Lee thought that a hypertext-based proposal like the World Wide Web could meet the challenges of large scale, decentralized information sharing.

According to Tim Berners-Lee, WWW could meet these scalability and decentralization challenges because users of that sharing system could add content by themselves through an editor and because the system load had to be distributed over easy-to-install independent web servers. In order to allow users to easily add content Berners-Lee defined a simple markup language to write hypertext (HTML) and created a browser meant to be used for HTML visualization and link navigation (Mosaic). To identify web servers and resources in a unambiguous and easy way, in order to make referencing and linking easy and consistent, Universal Document Identifiers (UDI) were introduced. Furthermore he also designed a reliable communication protocol to be used by the various servers (HTTP), in order to solve the decentralization challenges. In conclusion, everything was designed ad-hoc for the hypertext writing and exchange in an easy but scalable way.

2. SECOND PART

a) What is an "information system"?

An information system (IS) is a system designed and engineered for a specific purpose that is in general exchanging information with an object system, and managing the derived data. More specifically the object system can insert data into the IS or request some information from it. The IS can process the input data, store them, create new data to be provided and perform many other functions depending on the needs of the object system.

b) Give examples of technology needed for running an information system, for designing an information system, and for analyzing a running information system.

For running an IS there must be:

- Hardware that processes, stores and provides data;
- An interface that allows the object system to interact with the IS;
- A communication channel to transfer the data from

the object system to the information system and vice versa.

To design an information system all these three main parts must be modeled and engineered. This must be done with a particular focus on the modeling of the object system, because the system must satisfy its needs. The same criteria should be followed also for the analysis of a working information system: check if it really satisfies the needs of the object system, then focus on the technical part (e.g. check if it is able to support the load, if the communication channel is secure and reliable etc.).

c) Name min. three examples of research questions that are relevant in information system research (in general, not specific for Web information systems).

- Which are the best practices to obtain feedback from the object system, in order to check if its needs and requirements are satisfied and how they change over time ?
- Changes in the object system requirements reflect in changes in the information system. Which are the best practices to apply these changes in order to avoid system downtime and incompatibilities?
- How often should the information system upgrades be applied ? Is it better to apply them as soon as they are ready or to group them in batches and upgrade the system periodically ?

d) Describe how the Web has affected information systems and the way they are engineered.

The web introduced new opportunities for the information systems, but also new design and engineering challenges. In a web information system (WIS) the object system can be composed by all the people that have access to the Internet. This means that there is no more need to create a communication infrastructure, but it also means that system must be able to handle a lot of simultaneous interactions and that their number can change in a short time, so it must be able to balance the load. The Web also affected the software aspect of IS: in a web information system back-end and front-end should adhere to the web standards in terms of protocols, interface design, data exchange. Another new challenge is the security aspect: a system that is accessible through the Web is easier to attack.

3. THIRD PART

Read J. Hendler, T. Berners-Lee, From the Semantic Web to social machines: A research challenge for AI on the World Wide Web, Artificial Intelligence (2009) [2]. Give min. three examples of research questions that according to you are implied by this vision.

- In the process of building a derivation system that tracks provenance and trustworthiness of sources, who could decide which are the trusted ones that can be roots of these derivations and how could this decision process be done ? When a source is proven to be not trustworthy what happens to all the derivations that depends from it ?
- One of the uses of contexts will be to allow social machines to manage simultaneous but conflicting views of data, but nothing prevents a community to create and share a social machine with bias in the selection of the possible contexts in which the contained data can be interpreted. It could be possible to define a context definition policy to avoid this problem?
- Could an automated and explicit data policy-management Web process reduce the Internet Freedom ? Who will have the rights to modify these policies in case of incorrect or unexpected usage?

4. FOURTH PART

Describe an aspect of the Web and give three relevant research questions that by answering them could help us to 'understand' that aspect of the Web. Indicate for each of the research questions how you would go about to answer them.

One of the aspects of the web that often catches my attention is the easiness of spreading false news and facts through social networks and media. I am not referring to who creates satirical, funny and unrealistic fake content such as The Onion[1], but to websites or Facebook[2] pages that spread false but plausible news about hot topics in order to misinform people and bias their interpretation of facts.[3]

- How could be possible to extend the Web with a mechanism to verify if news or articles are trustworthy ?
  To answer this question I would follow the approach described in [2]: a derivation system, based on semantic web data, to track the provenance of shared content. I would then implement an integrated function in sharing systems (e.g. Facebook) that automatically adds a warning message under the shared content if it doesn't come from a trusted source, so that everyone can be notified about that.
- In a scenario like the one described above, who should have the power to create a list of trusted sources and how could it be distributed or integrated in the Web?
  These problems could be addressed by creating an infrastructure similar to a PKI[4] (Public Key Infrastructure) with entities that issue certificates for trusted sources (The equivalent of the Certificate Authorities[5] in a PKI). A list of trusted authorities could then be directly added into Web browsers, social networks or operating systems just like it is done now with the trusted CAs.

[1] `www.theonion.com`
[2] `www.facebook.com`
[3] I don't know how much this phenomenon is present worldwide, but in my country (Italy) I see this kind of content every day on my Facebook homepage, shared by tens of people.
[4] `http://en.wikipedia.org/wiki/Public_key_infrastructure`
[5] `http://en.wikipedia.org/wiki/Certificate_authority`

- How could be possible to stop the spreading of fake content that for some reasons was initially marked as trustworthy? The Web implementation of this system should include a way for the authorities to revoke trust certificates, and this revocation should automatically follow the path of the content on the Web and notify all the past sharers.

5. REFERENCES

[1] J. Conklin. Hypertext: An introduction and survey. *Computer*, 20(9):17–41, Sept. 1987.

[2] J. Hendler and T. Berners-Lee. From the semantic web to social machines: A research challenge for {AI} on the world wide web. *Artificial Intelligence*, 174(2):156 – 161, 2010. Special Review Issue.