Lesson1: How Big is the Web?

Unit3: Descriptive Modeling

Rene Pickhardt

Introduction to Web Science Part 2
Emerging Web Properties

WeST
People and Knowledge Networks

Institute for Web Science and Technologies · University of Koblenz-Landau, Germany

Completing this unit you should be

- familiar with two simple descriptive models of the web as a collection of text documents
- able to discriminate between the model and the reality
- familiar with the concept of a modeling choice
- able to criticize descriptive models
- Know 6 important steps for modelling

Building a simplistic Model for the Web

- Take a Web crawl
- In our case only a part (Simple English Wikipedia)
  - But one could have crawled more
- Use the “Collection of text documents Model”
- Goal: Measure the size!

Idea: Counting words as a measure of size

- Everything between two successive whitespaces is considered a word
- Web Crawl (Wikipedia) ---------→ size
- Collection of text documents --→ Count words
- Result: About 1.2 Mio Words (in my Version)

Criticising our Model

- How many words in the following Sentence?
  - I live in Koblenz.
    - Answer: 4
  - I live in San Francisco.
    - Answer: 5

- Result varies a lot on the implicit choice we made.

- What was the implicit modeling choice?

Modelling documents instead of words to measure size

- Every URI corresponds to a document
- Count URIs
- What about duplicate content?
  - E.g. two URIs with the same html?
- What about dynamic web content?
  - It is easy to have an infinite amount of URIs on one domain (e.g. Website depicting prime numbers)

The instantiated model reflects a particular situation in the world

- When we take a collection of web pages in order to build a text model
- Model characterizes how the world might work in general
- But the models we study only have a special snapshot of a special situation

6 important steps for studying objects with the help of models

1. Select an object of Study
   – E.g. the world wide web
2. Select a toy example
   – E.g. a web crawl, Simple Englisch Wikipedia,…
3. Select a perspective for studying
   – E.g. “The Web as a collection of text documents”
4. Select a way of modelling
   – E.g. Descriptive, generative, predictive,…
5. Make model assumptions
   – E.g. How to define a word
6. Be aware of all the selections and assumptions you made and be able to critically discuss them.

Thank you for your attention!

Contact:

Rene Pickhardt
Institute for Web Science and Technologies
Universität Koblenz-Landau
rpickhardt@uni-koblenz.de

WeST
People and Knowledge Networks

Copyright: