Extracting Microdata Information from Web Pages

CS 533: INFORMATION RETRIEVEL SYSTEMS

SPRING 2013-2014

EMRE NEVAYESHIRAZI

Outline

- Description of the Problem
- Motivation / Importance of Problem
- Solution to Problem
- Problems with Solution
- Conclusion and Future Work

Problem

- Web Sites are structured with HTML markup language
- HTML does not carry information about the page
- Important information on web pages
- HTML5 Microdata is a specification for labeling HTML elements
  - Add simple attributes to HTML elements
  - Variety of attributes for different formats such as Person, Event, Ticket, Organization …
- Very new technique
- Need to update existing web pages
- Developers are not familiar
- Tool can be created for automatic Microdata generation

Motivation and Importance

- Many crucial information on web pages
  - Data is not structured
  - Search engines cannot capture the important data
- Adding Microdata by hand is hard
  - Time consuming
  - Many different Microdata format
- Web can become more meaningful and semantic via Microdata
  - Search engines can show the results directly

Solution to Problem

- Written a simple web crawler for collecting dataset.
  - Basic multithreaded crawler
  - In 2 days, 4 GB of HTML snippet obtained
  - Approximately 200.000 HTML documents

- Grouped the HTML snippets according to their Microdata schema.
  - For instance, *Article*, *News*, *Review*, *Book*, *Album* and so on.

- For each group,
  - Extract the content from page using boilerpipe open source library.
  - Remove stop words such as the, in, and had.
  - Order the words according to number of occurrences.

Solution to Problem

- When user submits a HTML page,
  - Extract the content.
  - Remove stop words.
  - Order words appearing in content.
  - Compare these words with existing groups obtained from dataset.
  - If similarity found,
    - Show related Microdata scheme to user.
    - Make suggestions if possible

Problems with Solution

- Extracting the content from HTML possible but
- Extracting the parts of content is hard,
  - For instance, for a *Book*, *Author*, *Title*, *Summary*, *Date of Publish* and so on.
  - HTML attributes can be used as partial solution.
- Inside 2 GB dataset some of the Schemes are never used.
  - Approximately 10 % of Schemes are used
  - For example, *MedicalEntity*, *BusTrip*, *JobPosting*, *RadioStation* are not found.
  - Therefore, more data is needed for better results.
- Some Schemes does not contain enough content,
  - For example, *Rating* which only contains min and max allowed rating and actual rating value.

Conclusion / Future Work

- The solution is far from expected results but still gives reliable results.

- Open source content extractor can be improved.
  - It can take Microdata information into account for better results.

- I can try to extract content from text for specific Schemes.

- **Thank you !**