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
- Microdata Example
- Methodology
- Expected Results

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

Microdata Example

```
<div itemscope itemtype="http://data-vocabulary.org/Person">
  My name is <span itemprop="name">Bob Smith</span>
  but people call me <span itemprop="nickname">Smithy</span>.
  Here is my home page:
  <a href="http://www.example.com" itemprop="url">www.example.com</a>
  I live in Albuquerque, NM and work as an <span itemprop="title">engineer</span>
  at <span itemprop="affiliation">ACME Corp</span>.
</div>
```

Methodology

- Some initial data should be used for classification
  - For each class, related HTML snippets can be saved
  - Microdata is hierarchical, classification can be as well
- New HTML snippets can be compared with existing classes
  - Text Mining
  - HTML Structure
  - HTML Attributes such as classes, ids
- New HTML snippets can be used for improving the existing classes

Expected Results

- Similarity measurement with threshold value
- If the HTML snippet similar to elements in one of the classes (> threshold)
  - Apply Microdata to HTML
- If there is some similarity
  - Suggestions can be made to end user
  - User can create add Microdata by looking at the suggestions
  - Program can be still usefull