Blextra

A BLOG EXTRACTION PROGRAM

MARIA SCHMIDT, AMANDA KARAVOLIA, MOJGAN MADAH

Outline

- Goal of our Project
- Testing Blogs
- Blogs
- Demo
- Conclusion

Goal of our project

- Extracting articles from blogs
- Search them for terms and metadata

Testing Blogs

- NZZ (13 blogs)
- Tagesanzeiger (17 blogs)
- Tribune de Genève (264 blogs)

Steps

- Crawling the blogs
- Data extraction
- Building Ontology and Index
- Searching data

Step - Crawling the blogs

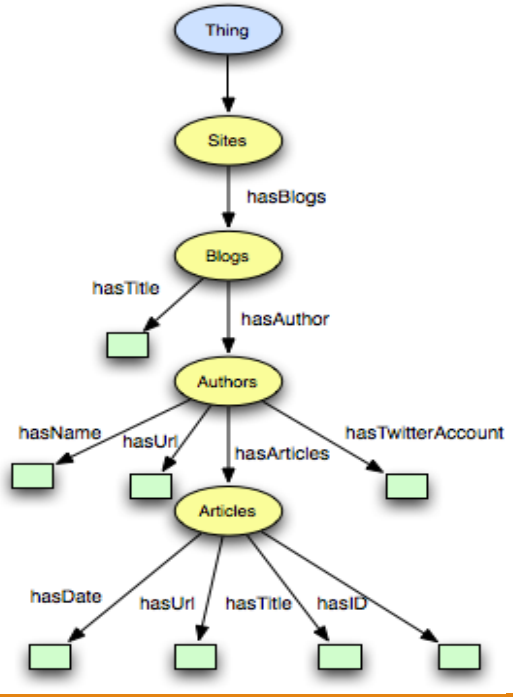
- Using Scrapy Web crawler
- Create a rule for each blog
- Crawling all the articles for each blog
- Save the Urls in json files

Step - Data extraction

- Using XPath to extract data (author, title, date, blogname, etc.)
- Extract content of the article and save it in a different file
- Clean data from HTML tags

Step 2-1: Building Ontology

- Protégé
  - RDF schema
- RDFLib
  - Individuals



Step 2-2: Building index

- Using Whoosh
- Index contains:
  - terms which appear in the content and titles
  - word stem of every term

  Example: wait, waits, waiting, waited → wait
- Character folding was applied (characters like á , â, a, ã are treated the same)

Step 4 - Searching data

- **Metadata**
  - SPARQL queries
- **Terms**
  - Whoosh methods
- **Search Options**
  - just searching in title
  - just searching for Word stem
  - using operators (or, and, not)
  - expand query with synonyms (only for german language)
  - searching by author
  - searching by date-range

Problems we faced

- Crawling
  - Finding the right rules to extract the articles
  - Avoiding endless loops
  - Blocking while crawling
- Data extraction
  - Finding the right HTML-Tags
  - Different for every site/blog

Demo

Conclusion – State of our application

- A web application for searching blog articles
- making use of metadata to form more elaborated queries
- using synonym expansion to find more articles
- using the word stem to find more articles

Conclusion – Further work

- Expand Ontology
  - Social information (Facebook and Twitter shares)
  - Comments, tags and categories of an article
- Expand Searching
  - Synonym expansion for French language
  - Translation of queries
- RDF scalable database