Archiving Ephemeral Data using Web Feeds

Archiving’s aim: Preservation of Ephemeral Data
| data going from factual to digital |

The Web’s dynamics a consequence of the Web 2.0 explosion
- frequently updated data
- new Web pages added each day

Information Retrieval from a (Personal) Web Archive
search Web data rooted in the past in a domain of interest

The Web Article’s Extraction Technique
operating at DOM level: bottom-up strategy
1. use HtmlCleaner as a parser
2. filter the leaf nodes which contain at least one signer as ‘conceptual nodes’

Signifiers from the example on the right: study: concept, being a scientist: 3-gram.

Semantic Acquisition
extract signifiers from the feed item’s title and description
- concepts: tokenize, stem and do a frequency analysis => a bag of relevant ‘tags’
- n-grams: sequences of n words, taken as they appear in the title and description

Semantic Density Measure
\[
\text{semanticDensity} = \frac{\sum_{n=1}^{N} \text{nbConceptualNodes} \times \text{node.textualLength}}{\sum_{n=1}^{N} \text{node.textualLength}}
\]

3. group the conceptual nodes in function of their lowest block-level common ancestor
4. take the block node which has the highest semantic density measure

Web Feeds = XML-based files: RSS, Atom
crawl and analysis of domain-specific feeds
- pass through Search4RSS to acquire a list of feeds
- crawl the feeds rather than the Web pages
- do a semantic and temporal analysis using a feed parser

Uniformly Querying a Collection of Web Data Objects

Web Object Signification and Components
- at feed level represents an item
- at Web page level represents a Web article
- content: text and references
- semantics: channel info (provenance), categories (‘tags’), title
- timestamps: the article’s publication date and the date of crawl

Web Page Reconstruction
naturally excludes boilerplate
- extract and sort the semantic zones (in the analysis phase)
- keep them in a file
- download the .css files
- reconstruct the path to them (at run-time):
  domain-channelId-crawlTimestamp-itemId

Conclusions
contributions
- Web feeds analysis
- semantic data mining in the feed
- a new way of extracting the relevant content of a Web page and the zones that are semantically related to it
- storing information at data object level vs. at Web page level

Example
| article's node |
| comments' node |

Heuristics:
- the block-level node is a DIV
- the comments’ zone is encoded as a list

Web Feed leveraged Elements
types of nodes
- channel: the publication hub of a Web site
- item: a resource uniquely identified by a URL and which has some semantics attached
- important elements
  - link
  - title
  - description
  - pubDate: not compulsory, but still omnipresent