An approach to semantic clustering based on Web feeds

Marilena Oita[1,2]

[1] Telecom ParisTech, INFRES - DbWeb team

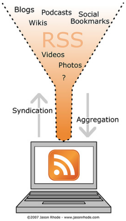
[2] WEBDAM project

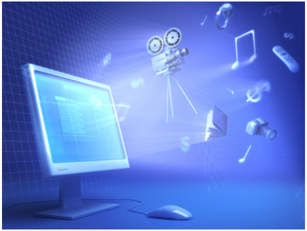
Webdam Meeting 4th of March, 2011

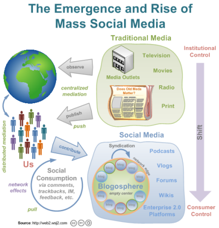
Outline

1. Archiving New Forms of Web Content
2. Archiving Data Objects using Web Feeds
3. Discussion and Further Work

New forms of Web content





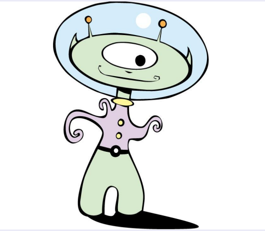


AIM

motivating context

Semantically-coherent Web archive collections

search a digital archive



for Web data rooted in the past

in a specific domain of interest

Uniformely querying a collection

of *Web Data Objects*

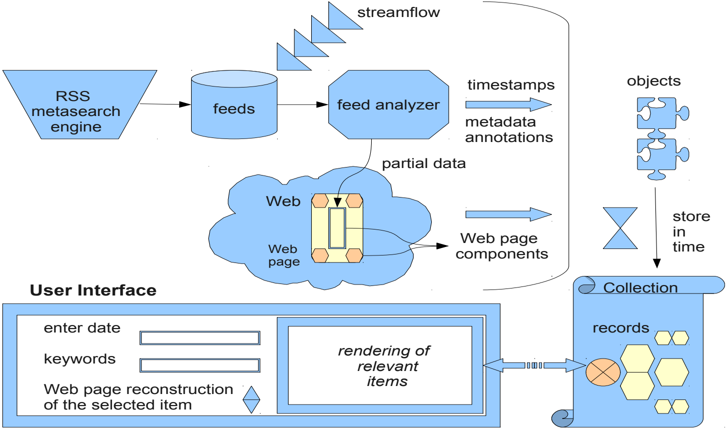


Figure: An application

Static template filtering

using the Web pages' *structure*

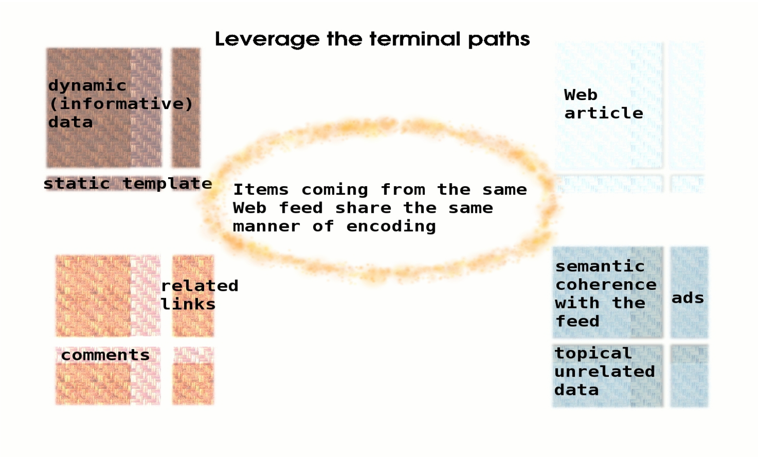

Leverage the terminal paths
dynamic (informative) data
static template
Items coming from the same Web feed share the same manner of encoding
Web article
related links
comments
semantic coherence with the feed
ads
topical unrelated data

Figure: clustering of terminal paths and measure similarity of content

Data object identification

using the items' *semantics*

A data object is a resource uniquely referenced through the feed item's URL.
Parse selected feed items and extract their

signifiers from the title and description of the item:

1. concepts
2. n−grams

Bottom-up technique of extraction at DOM level:

- group different significant leaf nodes by their lowest block-level common ancestor
- chose the one which is the most semantically dense

Observations

on the technique of extraction

Advantages

1. identifies the semantic zones in a Web page
2. extracts the main content referenced by the feed items (text and references)
3. constructs a collection of topical data: semantic annotations + timestamp + clean data → versioned data object

Drawback

the feed files need to be crawled on time:

Observations

on the technique of extraction

Advantages

1. identifies the semantic zones in a Web page
2. extracts the main content referenced by the feed items (text and references)
3. constructs a collection of topical data: semantic annotations + timestamp + clean data → versioned data object

Drawback

the feed files need to be crawled on time:
a consequence of feed entries' ephemerality

Work in progress

Hidden Web archiving

1. Hidden Web archiving
   - undestanding the search interface (form)
   - understanding the structure of response pages
   - record instances matching against concepts of form labels
2. Semantics Discovery:

Work in progress

Hidden Web archiving

1. Hidden Web archiving
   - undestanding the search interface (form)
   - understanding the structure of response pages
   - record instances matching against concepts of form labels
2. Semantics Discovery: YAGO (enriching an ontology, maybe studying its evolution...)

Thank You!

Questions?