Archiving Data Objects using Web Feeds

Marilena Oita and Pierre Senellart

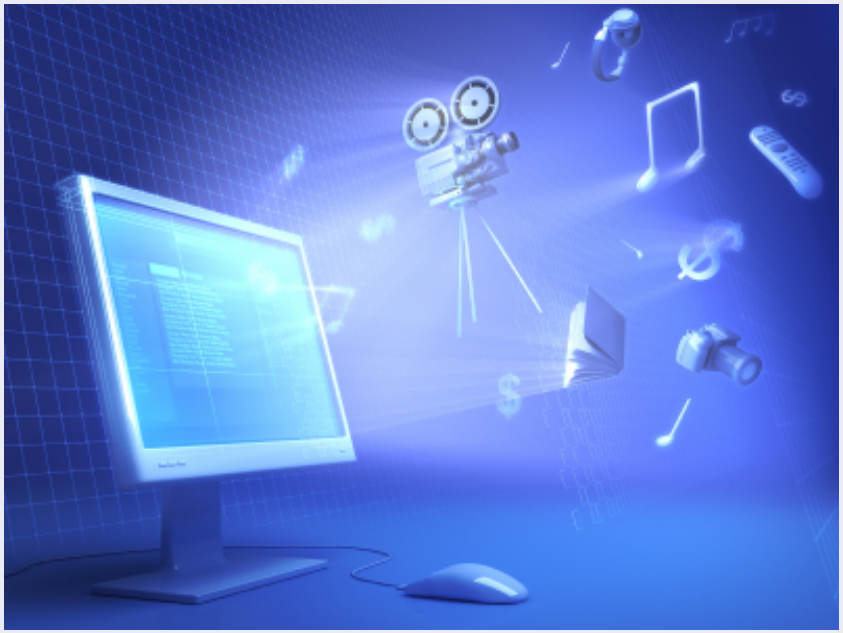
Outline

Valuable digital resources

to be preserved

from factual

..to digital

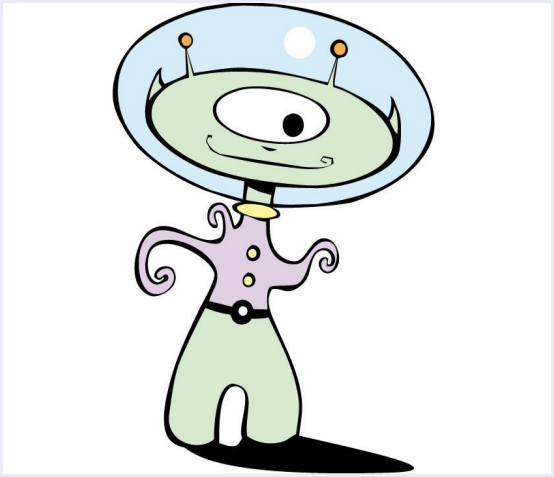


Scenario

motivating context

AIM:

search a digital archive



for Web data rooted in the past

in a specific domain of interest

Ephemeral Data

the volatile nature of the Web

The Web is fastest growing knowledge base
a consequence of the explosion of Web 2.0 tools

Difficulty to keep track of:

1. new Web pages added each day
2. frequently updated data
   –> causes:
   - user interactivity (ex: comments, forums)
   - inherent dynamics (ex: news, events)

data coming from a continuous process
which has an immediate value for a typical user

Importance of this data

in the archiving context

1. as it changes rapidly, at a certain point will be lost if not archived
2. represent the reflection of human impacting events and activities in:
   blogs, news...

characterize a certain period of time, from variuos points of view
therefore useful for future generations

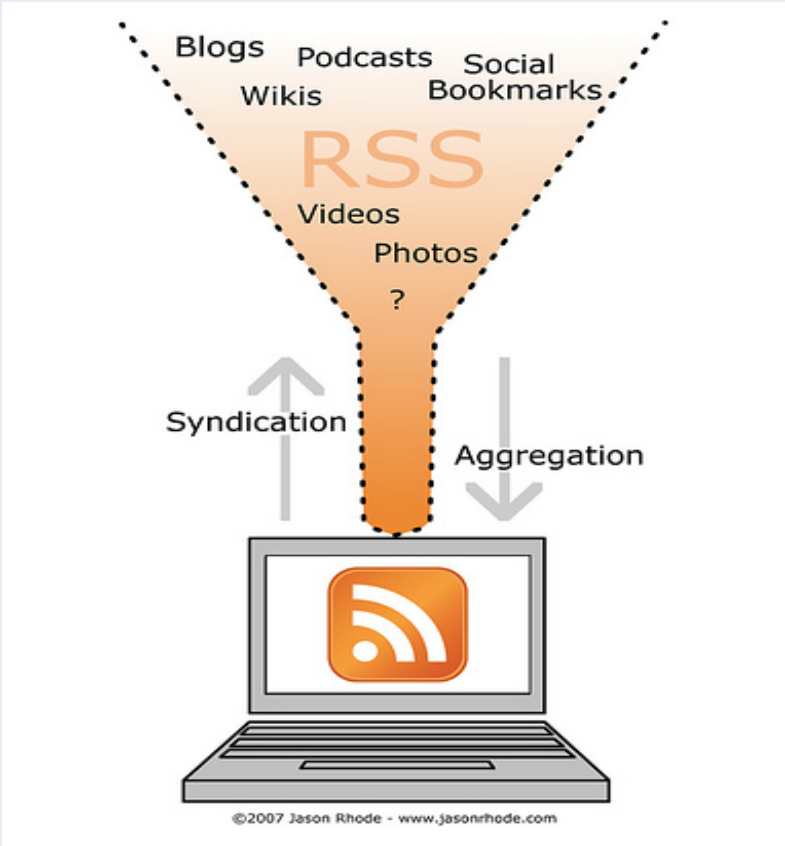
RSS, Atom

XML-based syndication formats

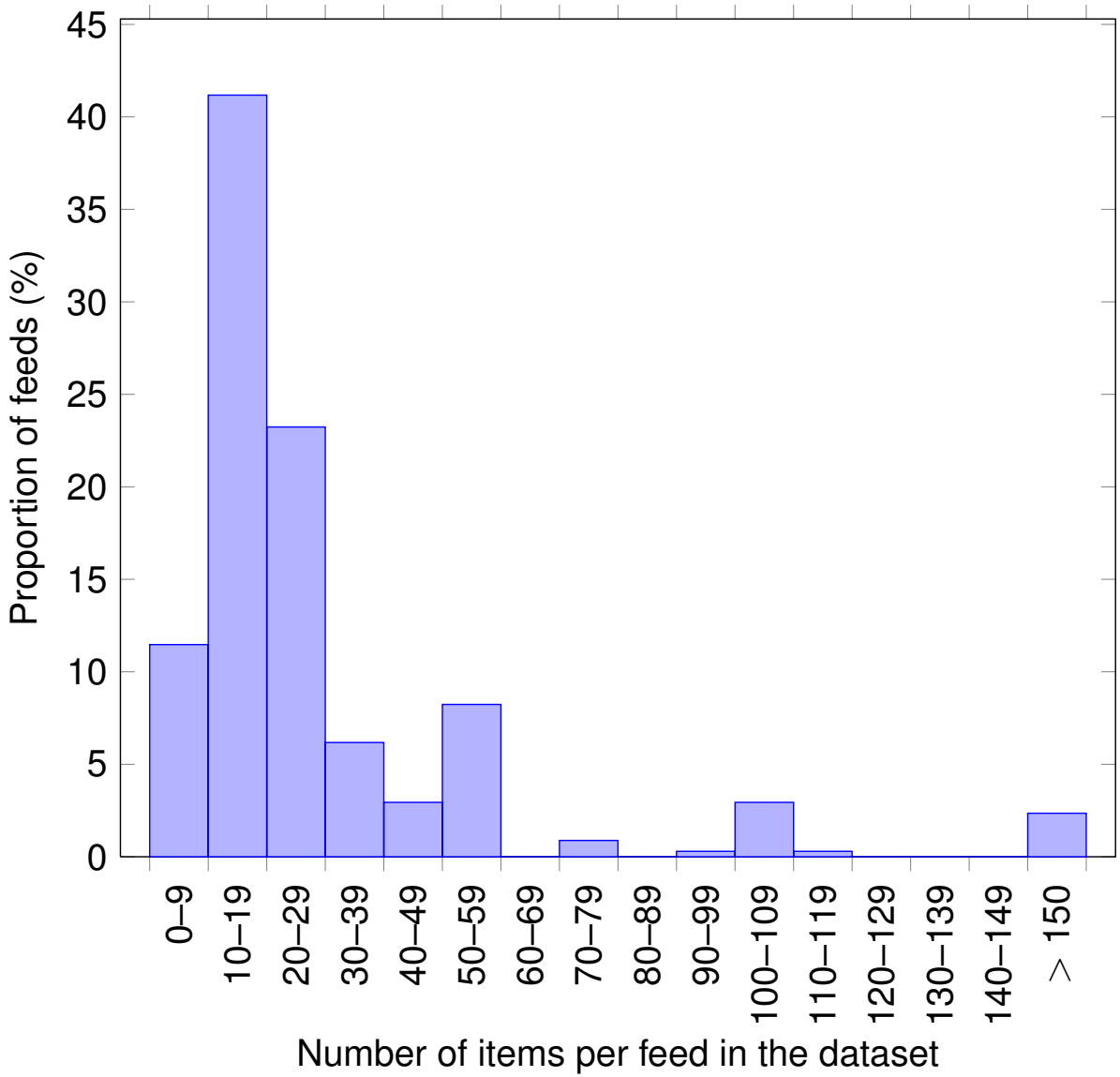
include specific metadata

- inform about the change occured
- describe the new resources published

about dynamic content



Statistics on Web Feeds



Statistics on the frequencies of update of Web Feeds

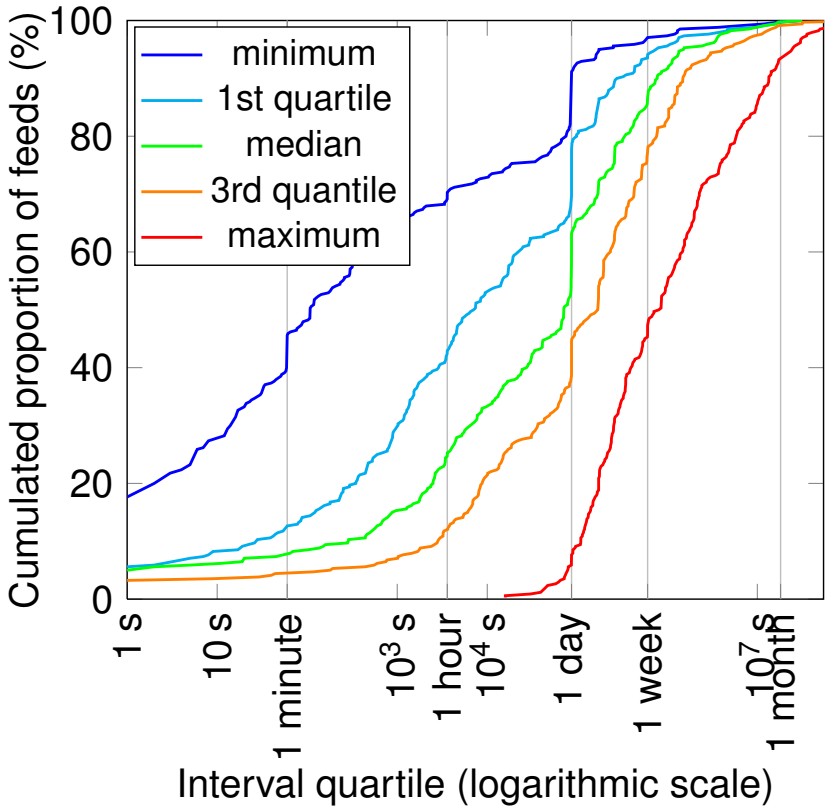


Figure: Cumulated proportion of feeds with a given quartile value of interval between updates

Uniformely querying a collection

of *Web Data Objects*

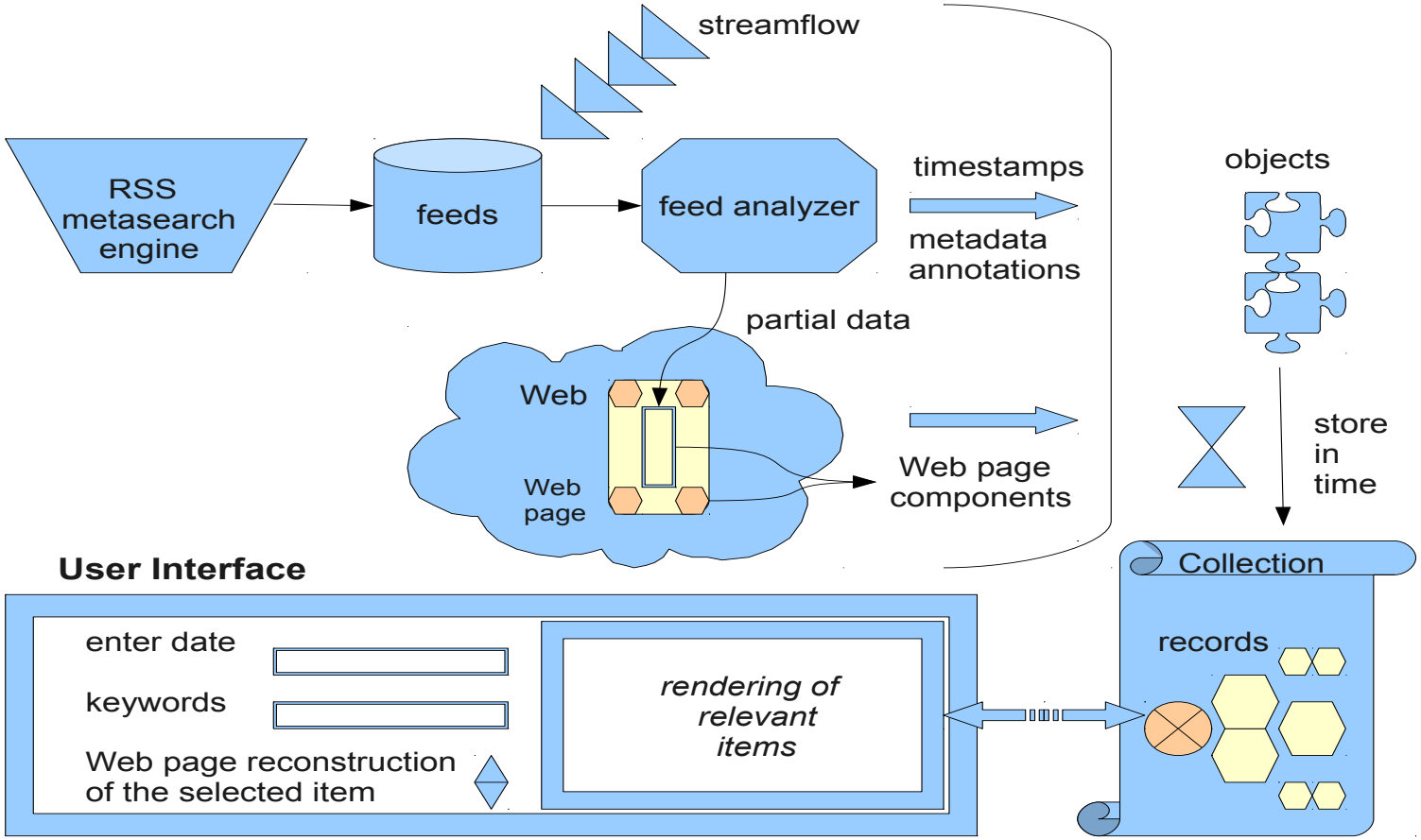


Figure: A scenario

Web feeds elements

and their signification

1. channel : publication hub of a Web site
2. item: a Web article, a status, a Wiki entry, a comment etc...

How can we use Web feeds in the archiving context?

- identify the article (text and references)

Web feeds elements

and their signification

1. channel : publication hub of a Web site
2. item: a Web article, a status, a Wiki entry, a comment etc...

How can we use Web feeds in the archiving context?

- identify the article (text and references)
- use the item metadata to enrich the semantics of the extracted article

Web feeds elements

and their signification

1. channel : publication hub of a Web site
2. item: a Web article, a status, a Wiki entry, a comment etc...

How can we use Web feeds in the archiving context?

- identify the article (text and references)
- use the item metadata to enrich the semantics of the extracted article
- encapsulate the result and store it in a timely manner: use of the temporal dimension

The correspondence between the item

and the Web page

```
<item>
<title>A study on how to study </title>
<link>http://feedproxy.google.com/r̃/CosmicVarianceBlog/3̃/-uatEVOIO0g/</link>
<comments>http://blogs.discovermagazine.com/
cosmicvariance/2010/09/07/a-study-on-how-to-study/#comments </comments>
<pubDate> Wed, 08 Sep 2010 03:16:54 +0000 </pubDate>
<dc:creator> daniel </dc:creator>
<category> <![CDATA[Advice]]> </category>
<guid isPermaLink="false"> http://blogs.discovermagazine.com/ cosmicvariance/?p=5353
</guid>
<description> <![CDATA[One of the most delightful aspects of being a scientist is that
you&#8217;re always learning. Your colleagues teach you things. Your students teach you things.
Journal articles teach you things. You sit quietly at your desk and figure things out.
You&#8217;re perennially a student. But how to be a better student? This morning the New
York [...]]]> </description>
<content:encoded><![CDATA[<p>One of the most delightful aspects of being a scientist is that
you&#8217;re always learning. .../>]]> </content:encoded>
<wfw:commentRss> http://blogs.discovermagazine.com/cosmicvariance/
2010/09/07/a-study-on-how-to-study/feed/ </wfw:commentRss>
<slash:comments>6 </slash:comments>
<feedburner:origLink> http://blogs.discovermagazine.com/cosmicvariance/
2010/09/07/a-study-on-ow-to-study/ </feedburner:origLink>
</item>
```

Figure: The Feed Item corresponding to the previous Web page

The correspondence between the item

and the Web page

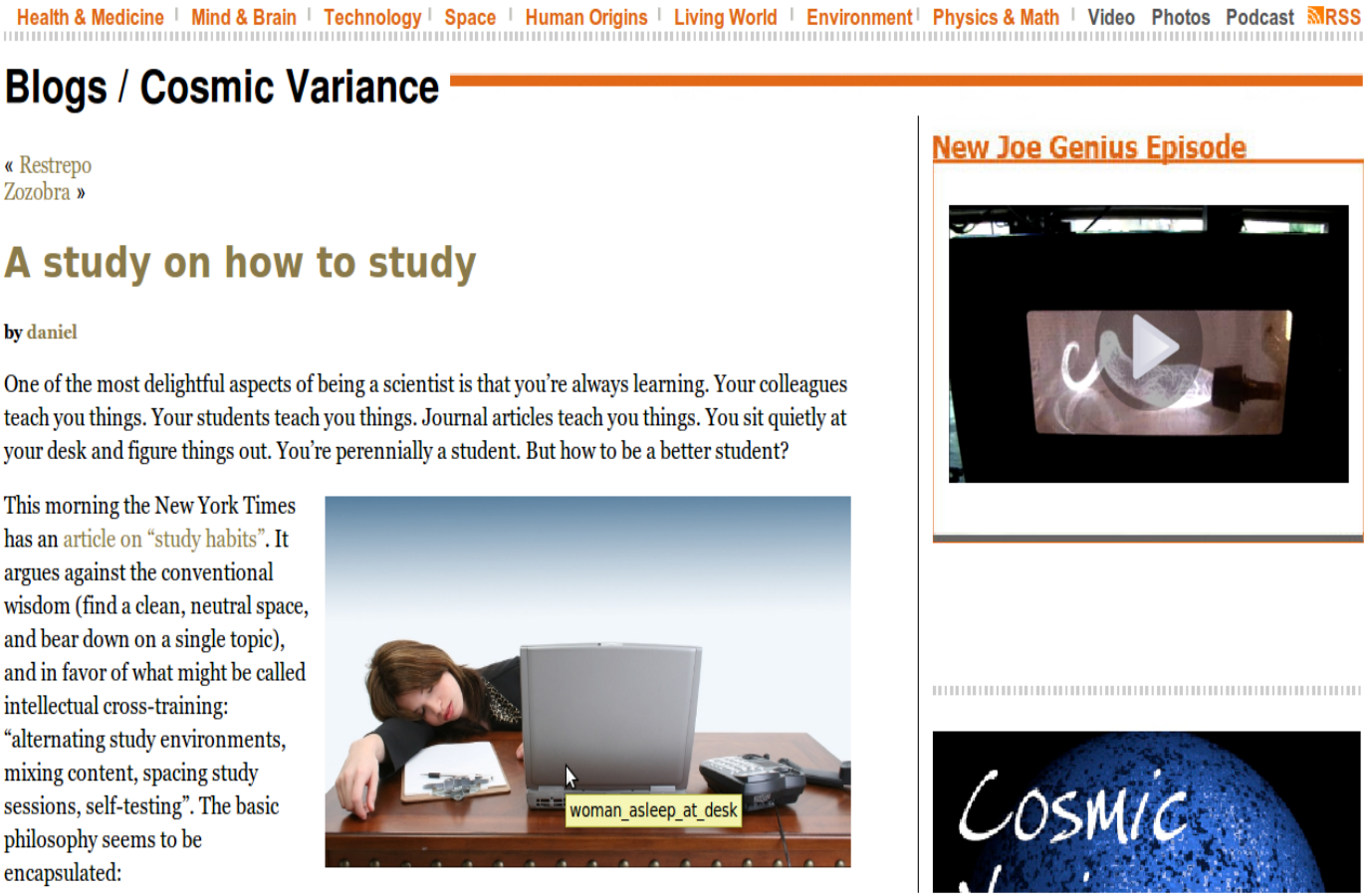


Figure: A typical Web article

What is a data object?

in our context

A data object is a resource uniquely referenced by a feed through the item URL.

1. has some special metadata associated: 'significant' properties
2. can be simple or compound: a comment vs. a commented article
3. can contain multimedia: imgs, videos,...
   and embedded code
   – we manage only references for the moment

The extraction technique

semantic acquisition

Parse the feed items and extract their

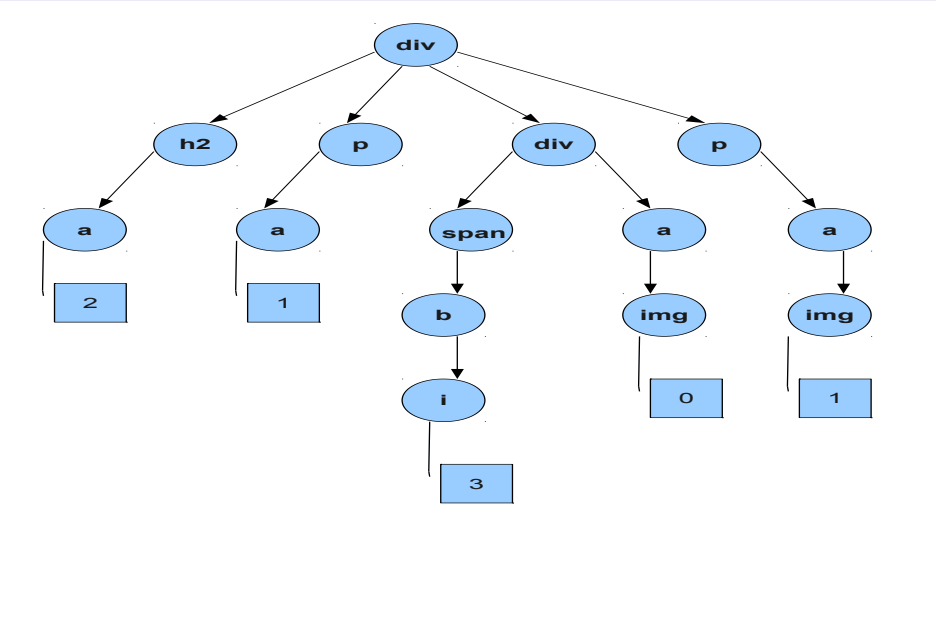
signifiers from the title and description of the item:

1. concepts: tokenize, stem and do a frequency analysis => a bag of relevant 'tags'
2. n−grams: sequences of *n* words, taken as they appear in the title and description

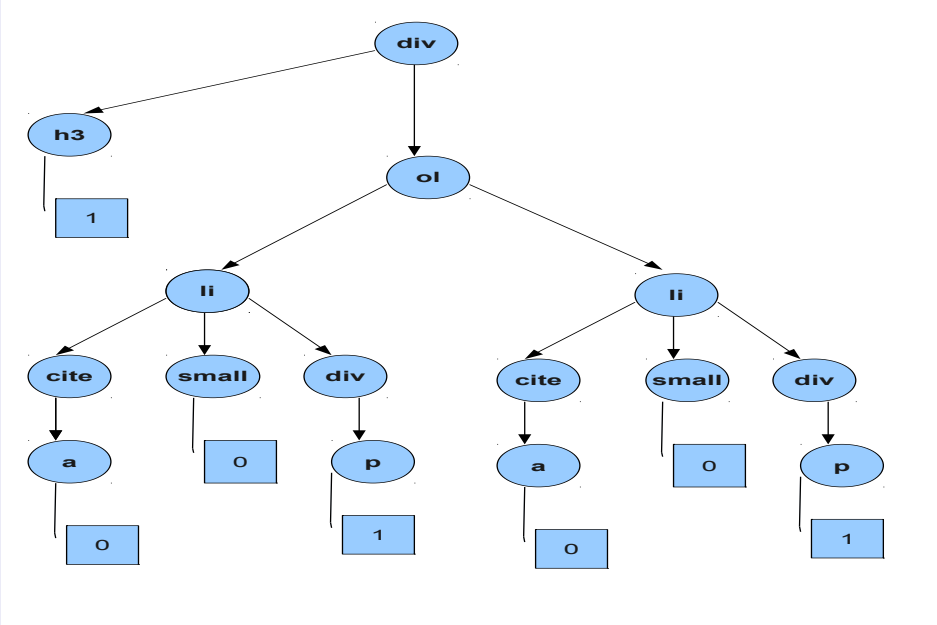
The extraction technique

operating at DOM level: bottom-up strategy

the block node of the article



comments'zone

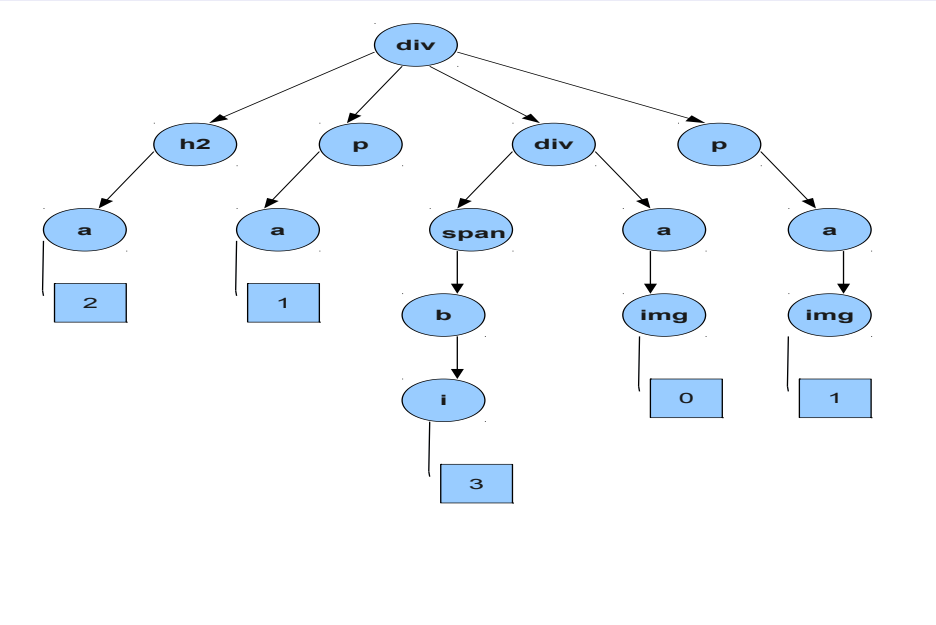


1. clean the Web page using HtmlCleaner
2. filter the leaf nodes which contain at least one signifier

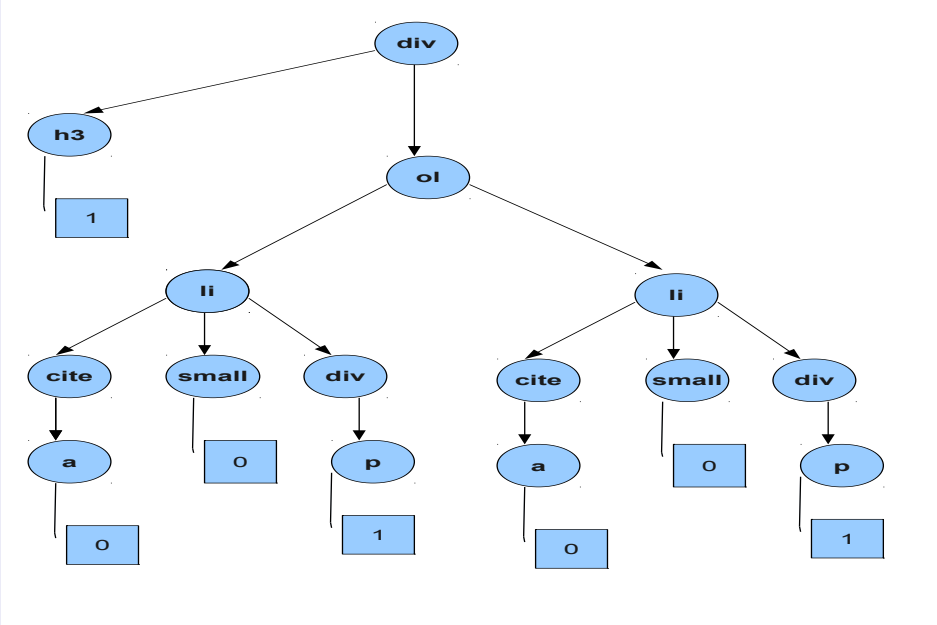
The extraction technique

operating at DOM level: bottom-up strategy

the block node of the article



comments'zone



1. clean the Web page using HtmlCleaner
2. filter the leaf nodes which contain at least one signifier : conceptual nodes

The extraction technique

semantic density measure

1. group different conceptual nodes by their lowest block-level common ancestor
2. chose the one which has the largest value for the measure:

$$semanticDensity = \sum_{n=1}^{nbConceptualNodes} \frac{cnode.nbOfSemanticMatches}{cnode.textualLength}$$

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 set of semantic terms + timestamp of a versioned data object

Drawback

the feed needs 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 set of semantic terms + timestamp of a versioned data object

Drawback

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

Conclusions

Contributions:

- statistics on feeds to assert their value in the archiving process

Conclusions

Contributions:

- statistics on feeds to assert their value in the archiving process
- a way of archiving highly dynamic pages: using Web feeds

Conclusions

Contributions:

- statistics on feeds to assert their value in the archiving process
- a way of archiving highly dynamic pages: using Web feeds
- a step forward concerning the archiving at another level of granularity

Conclusions

Contributions:

- statistics on feeds to assert their value in the archiving process
- a way of archiving highly dynamic pages: using Web feeds
- a step forward concerning the archiving at another level of granularity
- the first algorithm that uses the semantics to extract articles from Web pages and filter the boilerplate

Conclusions

Contributions:

- statistics on feeds to assert their value in the archiving process
- a way of archiving highly dynamic pages: using Web feeds
- a step forward concerning the archiving at another level of granularity
- the first algorithm that uses the semantics to extract articles from Web pages and filter the boilerplate
- a way of reconstructing the context (a cleaned version of the authentic Web page)

Further Work

- we exclude scripts in the Web page, making the assumption that it represents advertisements

Further Work

- we exclude scripts in the Web page, making the assumption that it represents advertisements
- extend the number of feeds and sources for more complete statistics

Further Work

- we exclude scripts in the Web page, making the assumption that it represents advertisements
- extend the number of feeds and sources for more complete statistics
- analyze the impact of the variation of parameters in our algorithm

Further Work

- we exclude scripts in the Web page, making the assumption that it represents advertisements
- extend the number of feeds and sources for more complete statistics
- analyze the impact of the variation of parameters in our algorithm
- studying data object change using a measure of similarity on the content and properties that we have extracted

Further Work

- we exclude scripts in the Web page, making the assumption that it represents advertisements
- extend the number of feeds and sources for more complete statistics
- analyze the impact of the variation of parameters in our algorithm
- studying data object change using a measure of similarity on the content and properties that we have extracted
- further study the semantic zones (their types and purpose) and the relation between them

Thank You

Questions?