A New Information Filtering Method for WebPages

Josep Silva

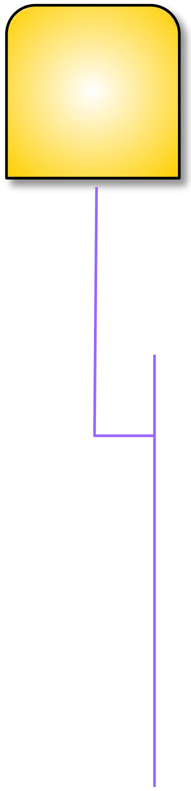
Technical University of Valencia

(joint work with Sergio López)

Contents

Motivation

Retrieving information from the Web (a single webpage)



Any webpage
It can include frames or iframes
No preprocessing stage
No semantic information
No microformats
No ontologies
etc.

Motivation

Retrieving information from the Web (a single webpage)

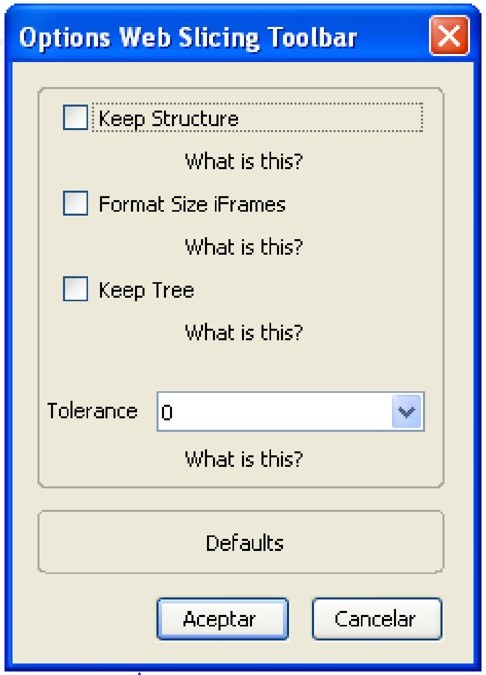
How do you extract non-textual information (images, videos, etc.)?
How do you keep the original structure of the webpage?

Motivation

What is the Web Filtering Toolbar?

It is an official Firefox's plugin whose interface is a toolbar

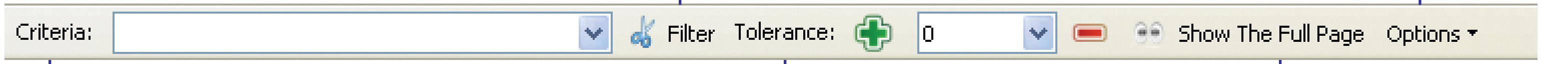
http://www.firefox.com/addons


Options Web Slicing Toolbar

☐ Keep Structure
What is this?
☐ Format Size iFrames
What is this?
☐ Keep Tree
What is this?

Tolerance: 0
What is this?

Start the filtering process


Criteria:
Tolerance: 0

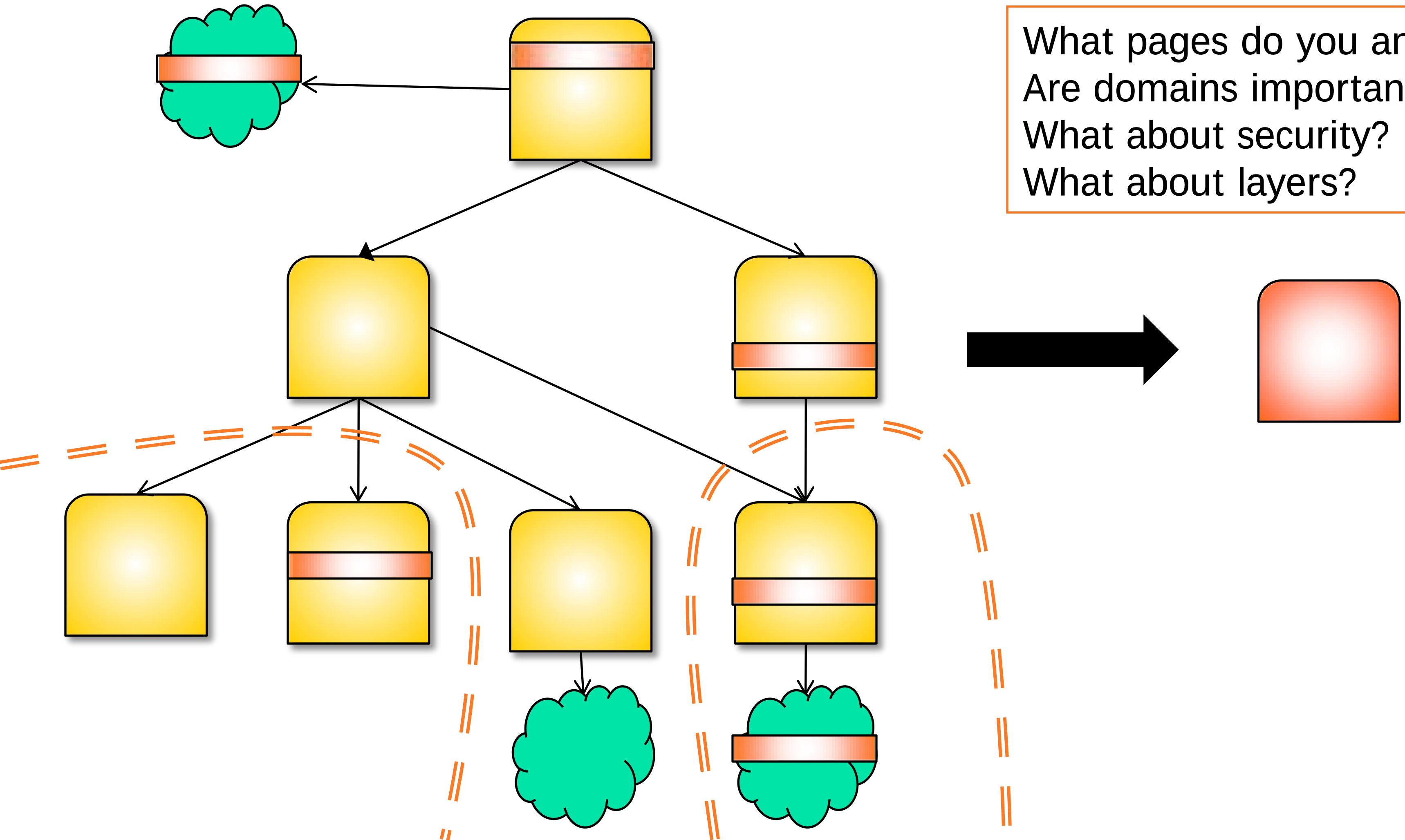
Filtering criterion

How much information do you want?

Show everything

Motivation

Retrieving information from the Web (multiple webpages)



What pages do you analize?
Are domains important?
What about security?
What about layers?

Motivation

Objectives

1. The technique should be able to work online
    - No precompilation phases
    - No proxies

2. The technique should be able to work with any webpage
    - Unknown objects must be processed
    - No restrictions on the language
    - No semantic information is provided to the technique (just syntactic)
        - Forget about ontologies
        - Forget microformats

Contents

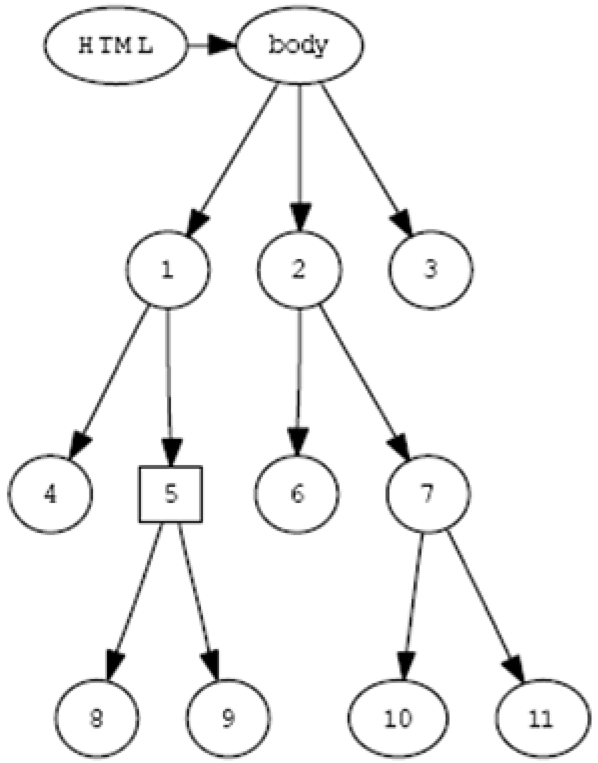
Filtering Single Webpages

The Technique

A web page can be seen as a tree of labeled nodes.

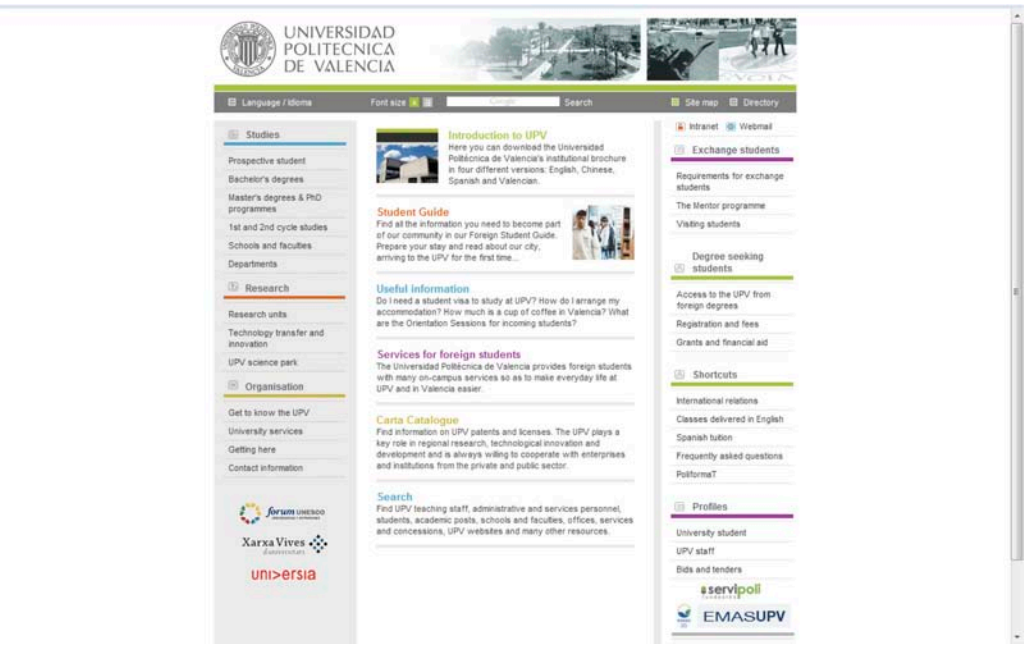
Internally represented in *DOM with a data structure* called "*document*"

We use this model to see web pages as a hierarchically organized collection of nodes.



Filtering Single Webpages

The Technique



UNIVERSIDAD POLITECNICA DE VALENCIA

Language / Idioma | Font size | Search | Site map | Directory

Studies
- Prospective student
- Bachelor's degrees
- Master's degrees & PhD programmes
- 1st and 2nd cycle studies
- Schools and faculties
- Departments

Research
- Research units
- Technology transfer and innovation
- UPV science park

Organisation
- Get to know the UPV
- University services
- Getting here
- Contact information

forum UNESCO · Xarxa Vives d'universitats · universia

Introduction to UPV
Here you can download the Universidad Politécnica de Valencia's institutional brochure in four different versions: English, Chinese, Spanish and Valencian.

Student Guide
Find all the information you need to become part of our community in our Foreign Student Guide. Prepare your stay and read about our city, arriving to the UPV for the first time...

Useful Information
Do I need a student visa to study at UPV? How do I arrange my accommodation? How much is a cup of coffee in Valencia? What are the Orientation Sessions for incoming students?

Services for foreign students
The Universidad Politécnica de Valencia provides foreign students with many on-campus services so as to make everyday life at UPV and in Valencia easier.

Carta Catalogue
Find information on UPV patents and licenses. The UPV plays a key role in regional research, technological innovation and development and is always willing to cooperate with enterprises and institutions from the private and public sector.

Search
Find UPV teaching staff, administrative and services personnel, students, academic posts, schools and faculties, offices, services and concessions, UPV websites and many other resources.

Intranet · Webmail

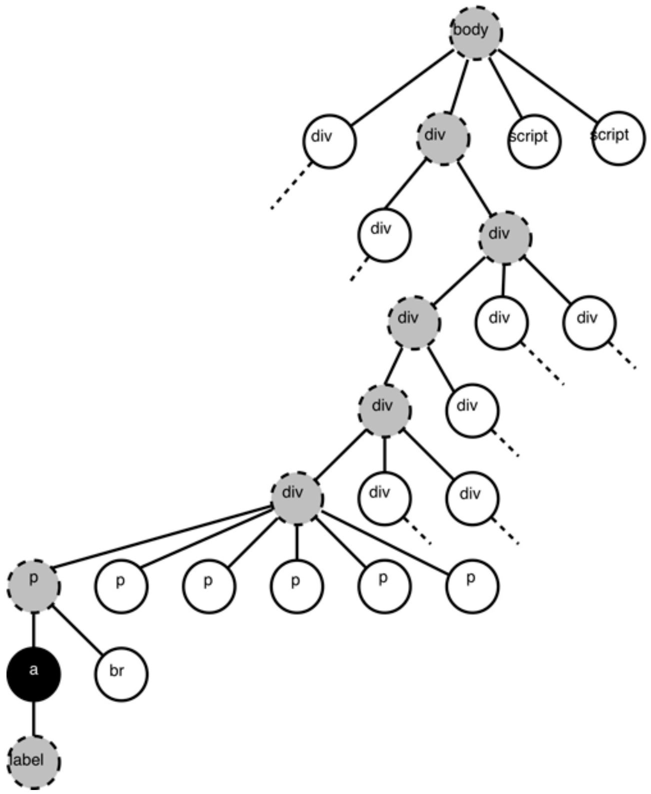
Exchange students
- Requirements for exchange students
- The Mentor programme
- Visiting students

Degree seeking students
- Access to the UPV from foreign degrees
- Registration and fees
- Grants and financial aid

Shortcuts
- International relations
- Classes delivered in English
- Spanish tuition
- Frequently asked questions
- PoliformaT

Profiles
- University student
- UPV staff
- Bids and tenders

servipoli · EMASUPV

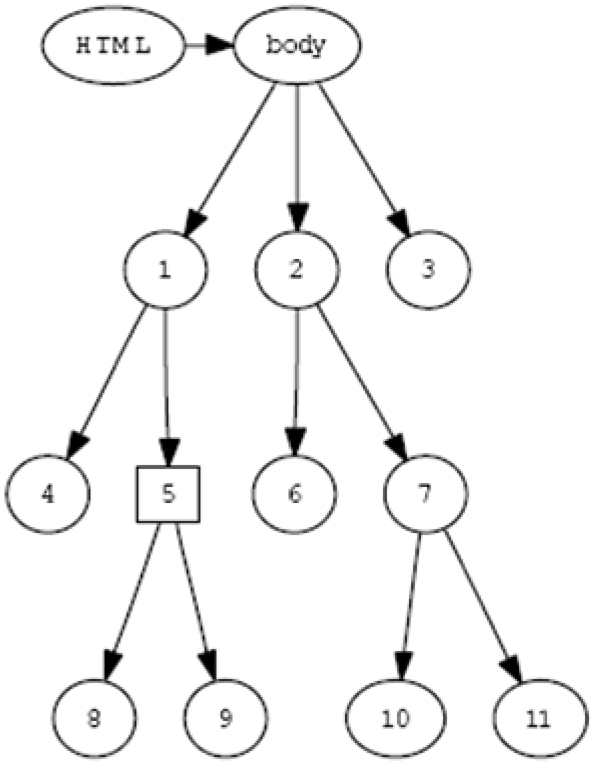


Filtering Single Webpages

The Technique

We first construct a tree-like data structure which contains all the information of the web page hierarchically organized in document order.

With the API of DOM we can explore the nodes of a web page and query their properties.



Filtering Single Webpages

The Filtering Criterion

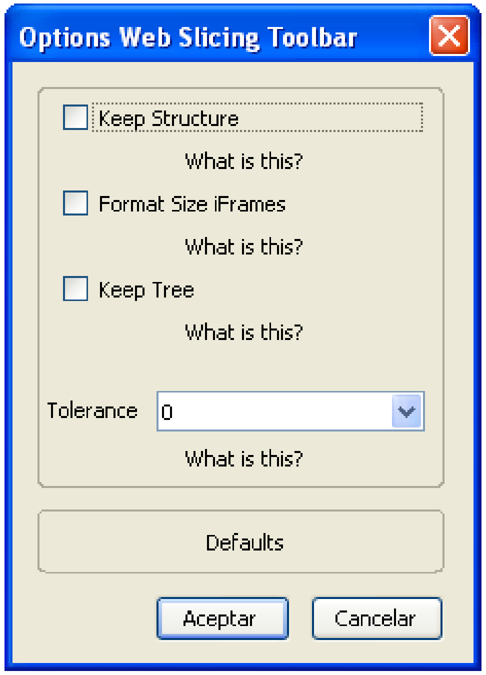
Text

↑

(words; flags)

↓

Vector with flags for the filtering tool



Options Web Slicing Toolbar

- [ ] Keep Structure — What is this?
- [ ] Format Size iFrames — What is this?
- [ ] Keep Tree — What is this?

Tolerance: 0 — What is this?

Filtering Single Webpages

The Filtering Criterion

Example: *(cars; [1,1,1,0])*

1) We compare "cars" *with the information contained in the attributes of each* node.

2) For special nodes (such as images) we use different attributes (e.g., "alt")

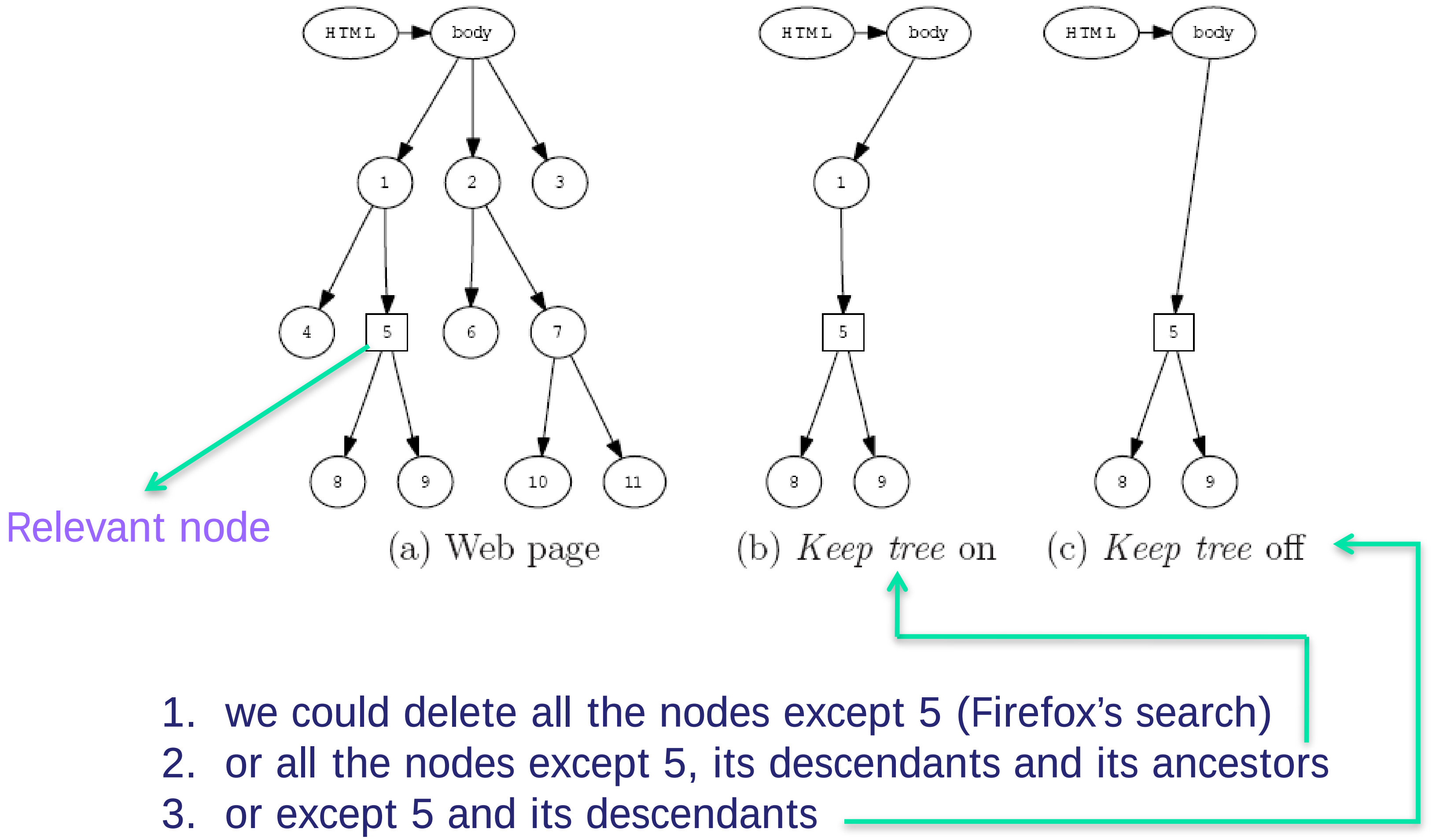
3) Nodes containing "cars" are called *relevant.*

4) We proceed by marking those nodes which are related to relevant nodes.

5) Not marked nodes are deleted or hidden.

Filtering Single Webpages

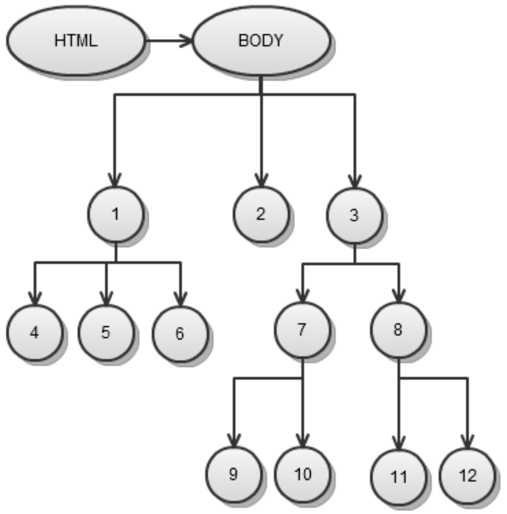
Example: Deleting nodes

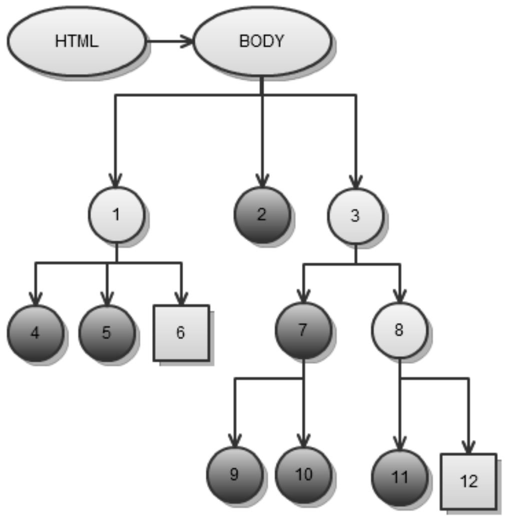


1. we could delete all the nodes except 5 (Firefox's search)
2. or all the nodes except 5, its descendants and its ancestors
3. or except 5 and its descendants

Filtering Single Webpages

The Filtering Technique





Filtering Single Webpages

The Filtering Technique

To summarize, we filter a given web page with respect to a filtering criterion by

(i) Building a data structure containing all the elements in the web page,

(ii) Traversing the data structure to find the relevant nodes with respect to the filtering criterion, and

(iii) Deleting (or hiding) all the nodes from the web page except the relevant nodes according to the filtering criterion's flags.

Contents

Filtering Multiple Webpages

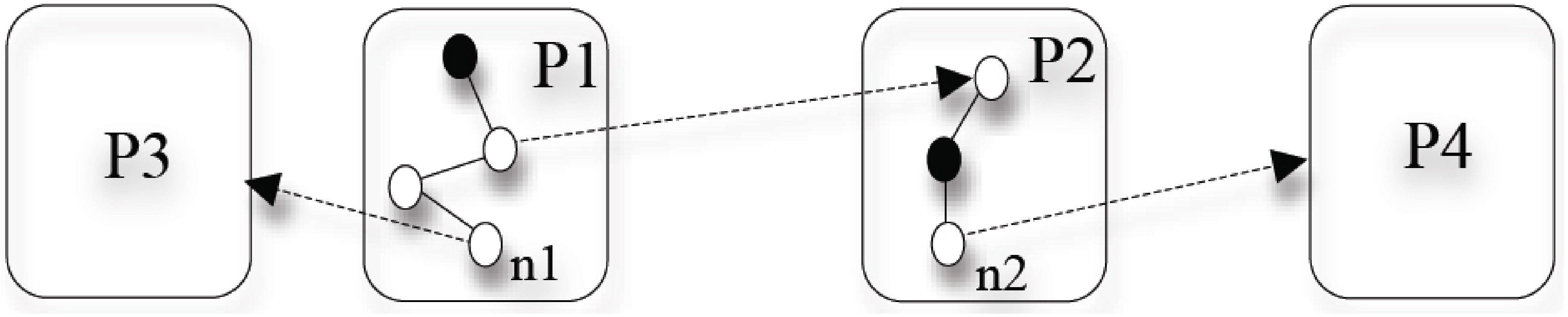
The Hyper-Syntactic Distance

Definition 5 (Search Hyperspace). Given a webpage P=(u,t) the search hyperspace of P is the set of webpages that either are reachable following hyperlinks from nodes of P, or that can reach P from their hyperlinks.

The search hyperspace is the collection of webpages that are related to the initial webpage, and that should (ideally) be inspected by our information retrieval algorithm. However, the search hyperspace of a webpage is potentially infinite (even more when we surf dynamic webpages [11]), and it is often huge.

Filtering Multiple Webpages

The Hyper-Syntactic Distance



What links should we inspect and in what order?

Filtering Multiple Webpages

The Hyper-Syntactic Distance

✓DOM distance (dT): It is the length of the path between two nodes of a DOM tree.

✓ Page distance (dP): It is the lower number of hyperlinks that must be traversed to reach one webpage from another webpage.

✓ Domain distance (dD): It is the lower number of domains that must be traversed to reach one webpage from another webpage following a path of hyperlinks.

Filtering Multiple Webpages

The Hyper-Syntactic Distance

Denition 6 (Hyper-Syntactic Distance, Relevance).
Given a DOM node n, the hyper-syntactic distance of n is:

$$D = dT + KP * dP + KD * dD$$

where KP and KD are numeric constants used to weight the equation.

The relevance R of a DOM node is the inverse of its hyper-syntactic distance:

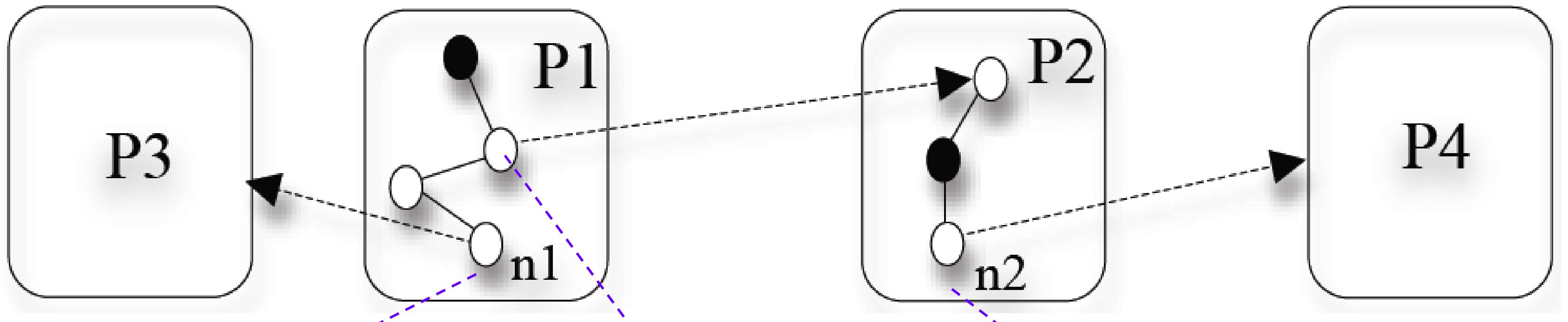
$$R = 1/D$$

The constants KP and KD determine the importance that we give to the fact that the word specified by the user is in another page, or in another domain.

Filtering Multiple Webpages

The Hyper-Syntactic Distance

$$D = dT + 10^6 * dP + 10^9 * dD$$



$3 + 0 * 10^6 + 0 * 10^9 = 3$

$1 + 0 * 10^6 + 0 * 10^9 = 1$

$1 + 1 * 10^6 + 0 * 10^9 = 1 + 10^6$

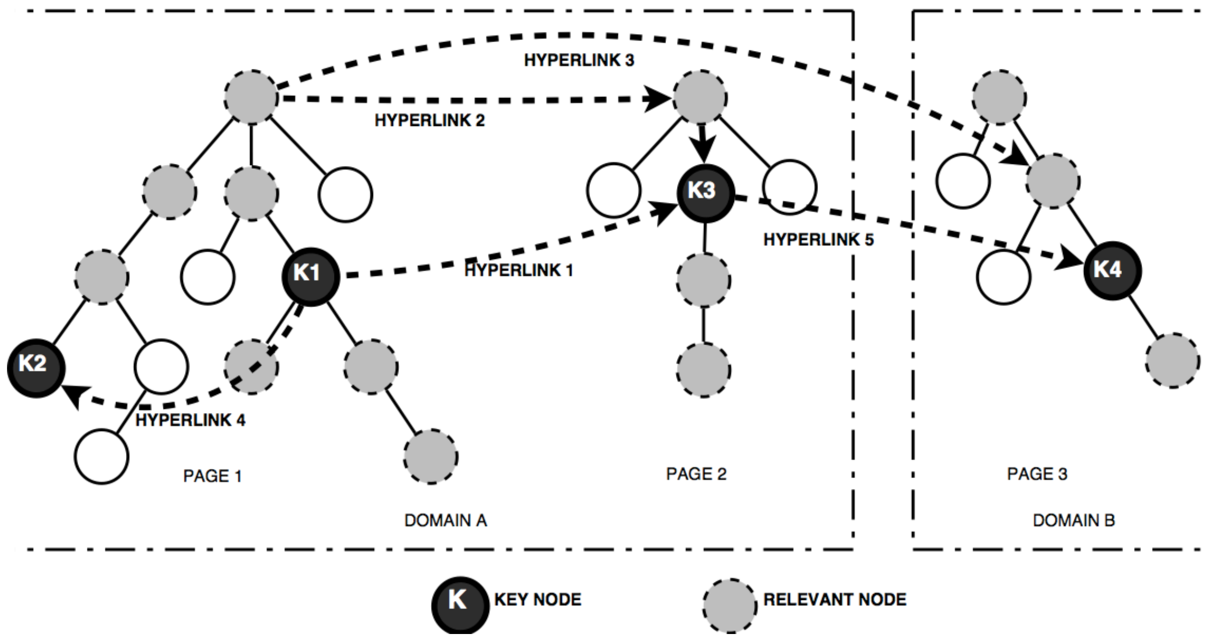
$1/3 = 0.333$

$1/1 = 1$

$1/10^6 = 0.000001$

Filtering Multiple Webpages

The Hyper-Syntactic Distance

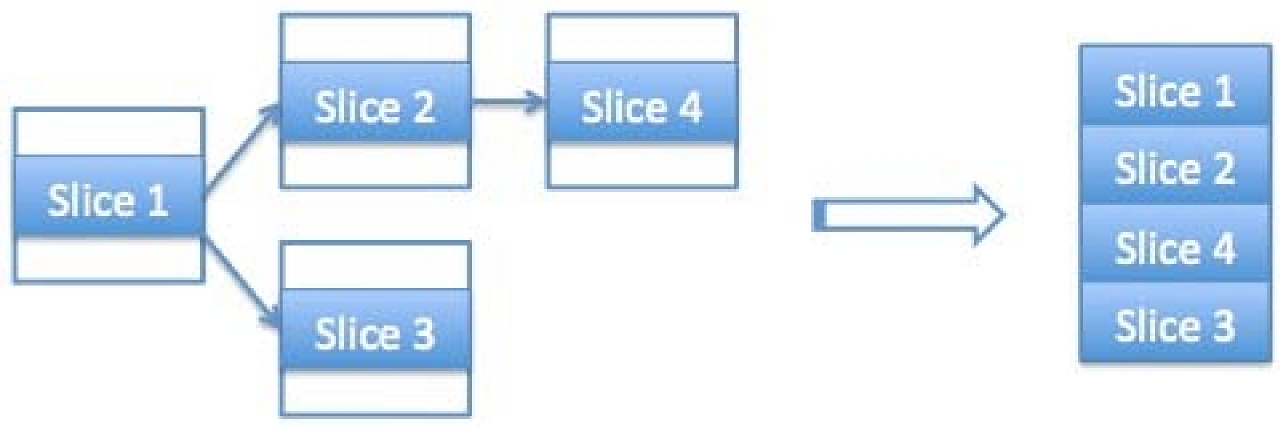


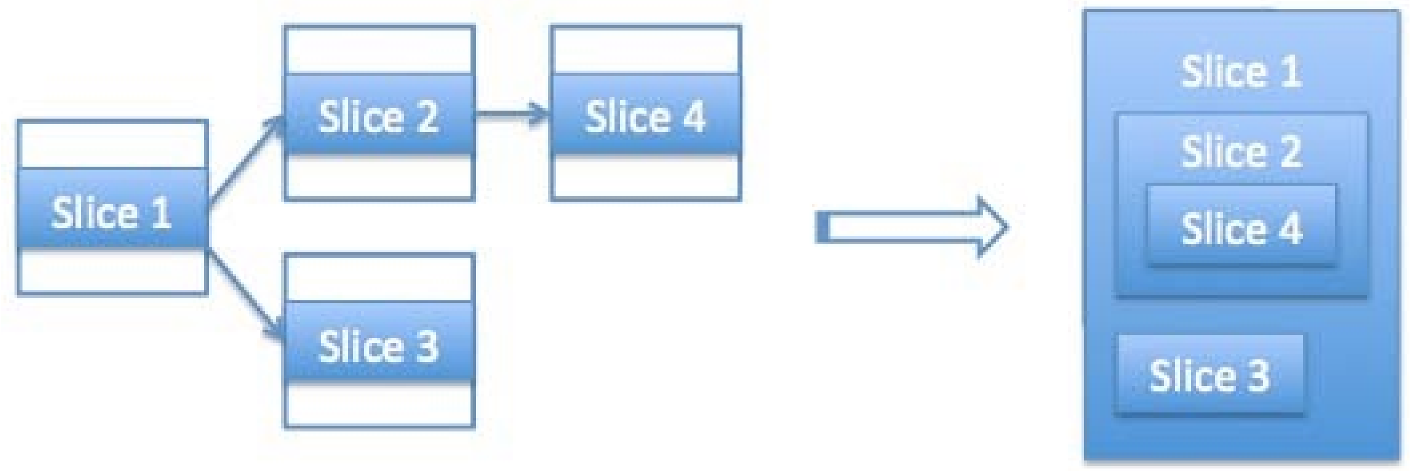
Hyperlink	dT	dP	dD	D	R
1	0	0	0	0	∞
2	2	0	0	2	0.5
3	2	0	0	2	0.5
4	0	0	0	0	∞
5	0	1	0	1000000	0.000001

Contents

Reconstructing Retrieved Information

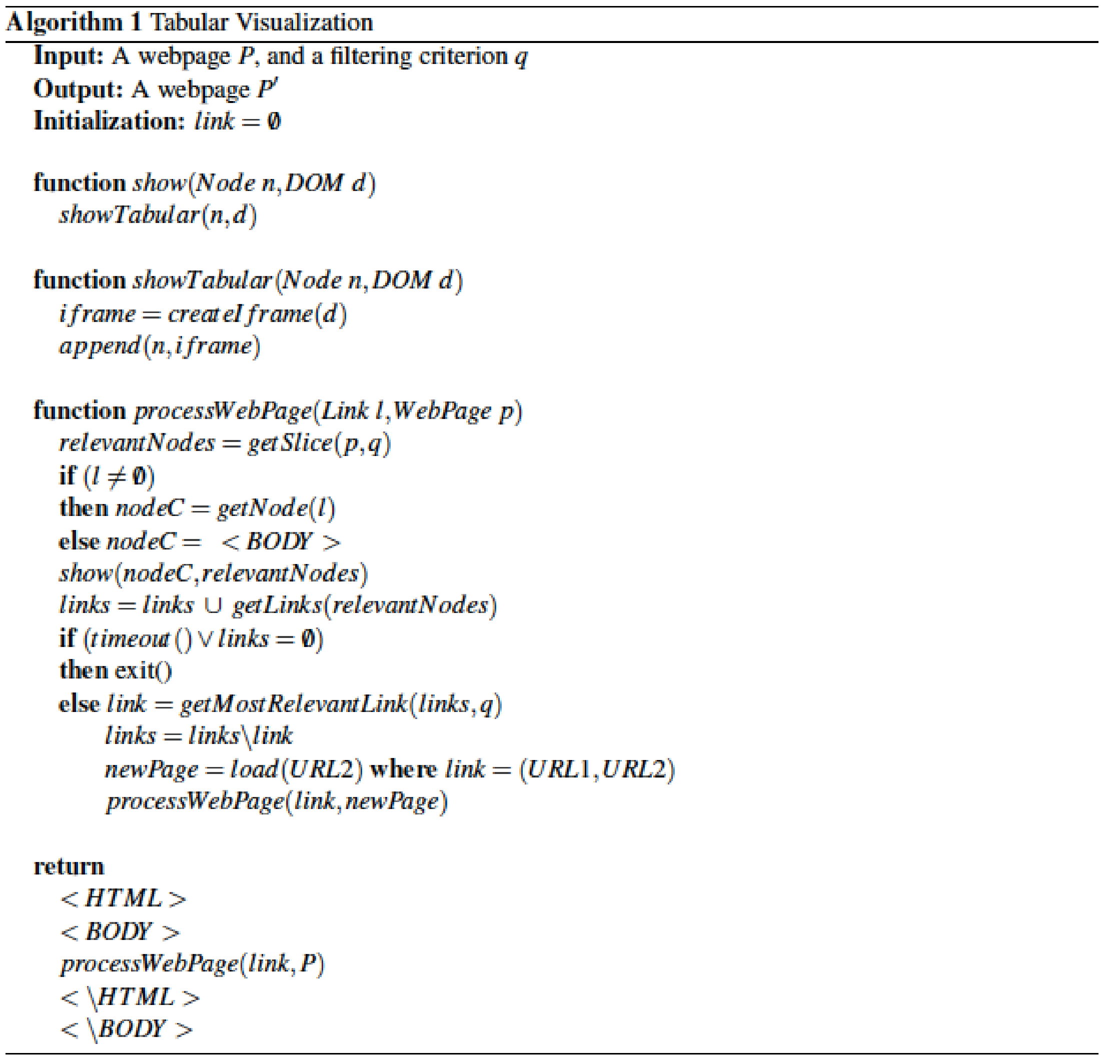
Visualization Algorithms





Reconstructing Retrieved Information

Visualization Algorithms



Algorithm 1 Tabular Visualization

```
Input: A webpage $P$, and a filtering criterion $q$
Output: A webpage $P'$
Initialization: $link = \emptyset$

function $show(Node\ n, DOM\ d)$
  $showTabular(n, d)$

function $showTabular(Node\ n, DOM\ d)$
  $iframe = createIframe(d)$
  $append(n, iframe)$

function $processWebPage(Link\ l, WebPage\ p)$
  $relevantNodes = getSlice(p, q)$
  if $(l \neq \emptyset)$
  then $nodeC = getNode(l)$
  else $nodeC = \ <BODY>$
  $show(nodeC, relevantNodes)$
  $links = links \cup getLinks(relevantNodes)$
  if $(timeout() \vee links = \emptyset)$
  then exit()
  else $link = getMostRelevantLink(links, q)$
      $links = links \backslash link$
      $newPage = load(URL2)$ where $link = (URL1, URL2)$
      $processWebPage(link, newPage)$

return
  $<HTML>$
  $<BODY>$
  $processWebPage(link, P)$
  $<\backslash HTML>$
  $<\backslash BODY>$
```

Reconstructing Retrieved Information

Visualization Algorithms

Algorithm 2 Hierarchical Visualization

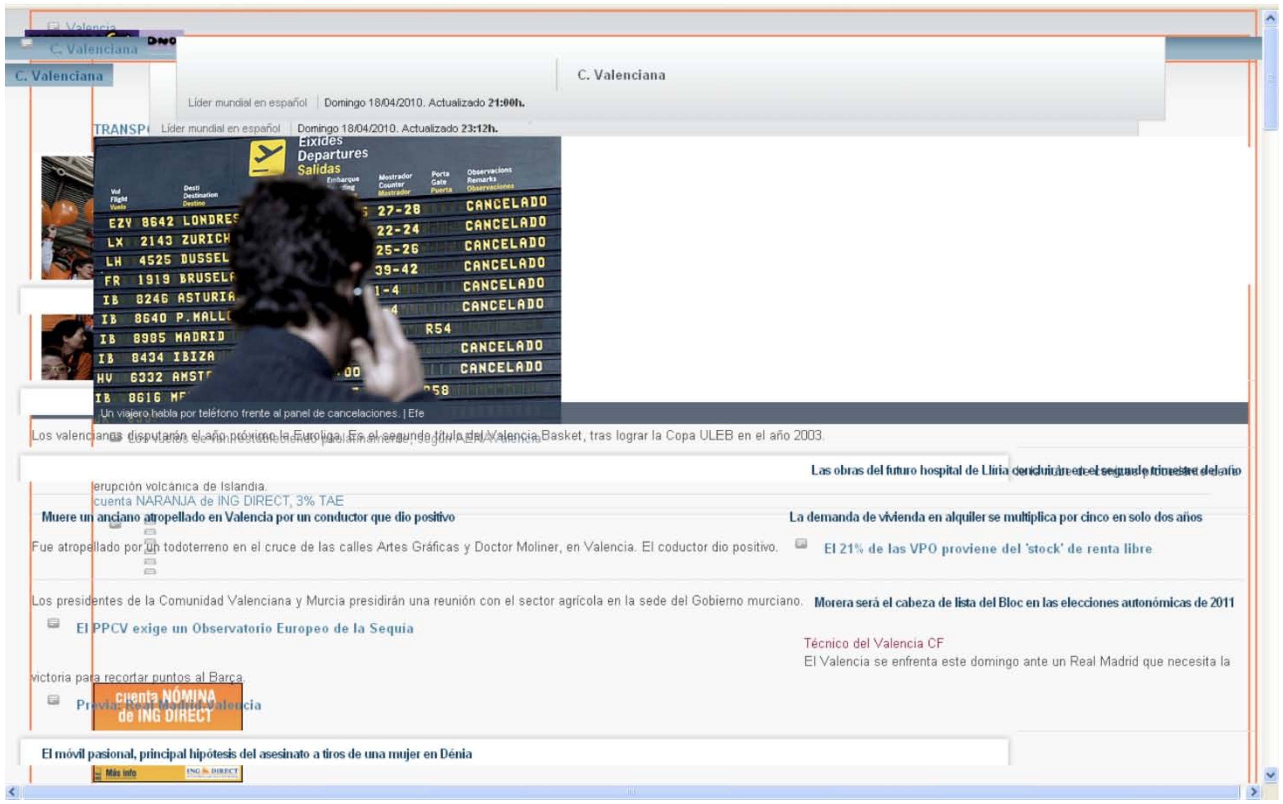
```
**Input:** A webpage $P$, and a filtering criterion $q$
**Output:** A webpage $P'$
**Initialization:** $link = \emptyset$

**function** $show(Node\ n, DOM\ d)$
    $showHierarchical(n, d)$

**function** $showHierarchical(Node\ n, DOM\ d)$
    $container = createContainer()$
    **if** $(n \neq < BODY >)$
    **then** $append(getParent(n), container)$
    **else** $append(n, container)$
    $append(container, d)$

**return**
    $< HTML >$
    $< BODY >$
    $processWebPage(link, P)$
    $< \backslash HTML >$
    $< \backslash BODY >$
```

Reconstructing Retrieved Information

Visualization Problems

Layers

+ Layers use absolute positions: Overlapping is possible and frequent!!!



Valencia

C. Valenciana

C. Valenciana

C. Valenciana

Líder mundial en español | Domingo 18/04/2010. Actualizado 21:00h.

TRANSP Líder mundial en español | Domingo 18/04/2010. Actualizado 23:12h.

Un viajero habla por teléfono frente al panel de cancelaciones. | Efe

Los valencianos disputarán el año próximo la Euroliga. Es el segundo título del Valencia Basket, tras lograr la Copa ULEB en el año 2003.

Las obras del futuro hospital de Llíria concluirán en el segundo trimestre del año

erupción volcánica de Islandia.

cuenta NARANJA de ING DIRECT, 3% TAE

Muere un anciano atropellado en Valencia por un conductor que dio positivo

La demanda de vivienda en alquiler se multiplica por cinco en solo dos años

Fue atropellado por un todoterreno en el cruce de las calles Artes Gráficas y Doctor Moliner, en Valencia. El conductor dio positivo.

El 21% de las VPO proviene del 'stock' de renta libre

Los presidentes de la Comunidad Valenciana y Murcia presidirán una reunión con el sector agrícola en la sede del Gobierno murciano.

Morera será el cabeza de lista del Bloc en las elecciones autonómicas de 2011

El PPCV exige un Observatorio Europeo de la Sequia

Técnico del Valencia CF

El Valencia se enfrenta este domingo ante un Real Madrid que necesita la victoria para recortar puntos al Barça.

Previa: Real Madrid-Valencia

cuenta NÓMINA de ING DIRECT

El móvil pasional, principal hipótesis del asesinato a tiros de una mujer en Dénia

Más info ING DIRECT

Reconstructing Retrieved Information

Visualization Problems

Security

+ *Cross Site Scripting (XSS)*: An attacker could execute scripts from a page or domain different from the loaded webpage.

➔ Web browsers' anti-XSS: DOM scripts of external pages are bloqued

✔ Use an iframe object:

```
frame = document:createElement(“iframe”);
...
frame:setAttribute(“type”;“content”);
...
frame:webNavigation:allowJavascript = false;
frame:webNavigation:allowMetaRedirects = true;
frame:webNavigation:allowPlugins = false;
```

Implementation

Design Decissions

+ The algorithm should never analyze a webpage with a page distance greater than 5. This is also supported by previous studies (see, e.g., Baeza and Castillo's discussion in [11]) that demonstrate that, in general, three to five levels of navigation are enough to get 90% of the information which is contextually related with the webpage selected by the user for the web search.

+ Hyperlinks that do not belong to the slice are discarded.

+ The maximum time spent to retrieve and show the information is 10 seconds. Usability rules [13] establish that 10 seconds is (as an average) the time limit that users spend waiting for a webpage to be loaded.

Implementation

Performance Evaluation

URL	Filtering criterion	Links visited
www.iee.org	student	12
www.upv.es	student	18
www.who.int	OMS	25
www.un.org	Haiti	6
www.esa.int	launch	13
www.nasa.org	space	16
www.mec.es	beca	18
www.edu.gva.es	universitat	20
www.ilo.org	projects	8
www.unicef.es	Haiti	10
www.mityc.es	turismo	9
www.mozilla.org	firefox	7

The average result is 13,5 webpages analyzed for each URL.

Note that a timeout of 10 seconds is the maximum time used to complete the final results webpage. But the visualization algorithms are incremental, thus, as an average, the first result is shown in less than a second (10/13,5 seconds).

Implementation

Distribution and instalation

The *Web Filtering Toolbar is distributed as an xpi file. An xpi package* is basically a ZIP file that, when opened by the browser utility, installs a browser extension. Currently, this extension applies to both Mozilla and Firefox browsers.

Since an *xpi package contains all the necessary information to* install an extension, the user only has to drag the *xpi and drop it over the* browser window. The extension will be automatically installed.

To download the tool and some other materials visit:

http://www.dsic.upv.es/~jsilva/webfiltering

and

http://www.firefox.com/addons

Implementation

The Filtering Technique

The user interface of the *Web Filtering Toolbar as the Firefox's interface* itself has been implemented with XUL and Javascript.

XUL
is an XML implementation which provides the interface components (such as buttons, menus, etc.);

Javascript
is used to control and execute the user actions.

Both parts are independent and can be updated without affecting the other part of the tool

Contents

Conclusions and future work

Conclusions:

• Current search engines consider web pages as the basic unit of informationwhich should be provided to the user.

• From our point of view, this is frequently an error which makes the user to load, read and scroll a lot of useless information.

• We have introduced a new approach to web filtering:

1. Automatically filter irrelevant information
2. Reorganizing relevant information

• Implementation integrated into the Firefox browser

Conclusions and future work

Future Work:

•Adapting the toolbar to other browsers: The use of a standard DIV with standard Javascript can work in any browser. It is currently being adapted to Google Chrome.

•Using a lexicon: This will allow to filter web pages by using alternative filtering criteria semantically related to the one specified by the user.

“cars” ≈ “auto” ≈ “automobile”

•Filtering XML documents. Some preliminary research in this direction has shown that our technique is directly applicable to XML documents. We would like to take advantage of the fact that with XML documents we have available a DTD which provides helpful information about the structure of the document.