Customizable, context-based website content extraction software

This technology is a context-based website content extraction tool for removing superfluous clutter from a webpage, rendering essential web content in a clean, easy-to-navigate format.

Unmet Need: Customizable website content extraction based on the context of the webpage

Existing website content extraction tools often fail to properly examine the context of the webpage, resulting in garbled or incomplete content. Furthermore, current software can lack customizability, leading to an undesired “one size fits all” approach, which often leaves extraneous content in its rendering. The rigidity of this approach limits the usefulness of the software and impedes the user’s web browsing experience.

The Technology: Context-based webpage content extraction for a more enjoyable web-browsing experience

This technology first parses the markup language text into a hierarchical data model and classifies the webpage by clustering it with other webpages within a similar genre that contains similar keywords and phrases. Based upon this classification, one or more filters are applied to reformat the website into a form much like those in its cluster. The filters may remove content such as images, videos, or programmed scripts. The technology continues to apply filters to the website until it satisfies the genre’s formatting heuristics and finally outputs the webpage in a more convenient format.

In addition to HTML, the technology is compatible with a number of markup languages, including XML and MathML.

Applications:

- Content extraction and rendering for smart phones, tablets, and cell phones
- Ad-blocking software
- Rendering unique browser skins
- Speech rendering for the visually impaired
Advantages:

- Can apply the same filter to multiple websites with similar content
- Highly customizable based on website’s genre and keywords
- Can be used to increase the font size for the visually impaired without increasing the size of the clutter
- Suitably reformats webpages for smart phones and tablets without input from website developer
- Customizable filters allow user to control webpage output
- Applicable to a number of markup languages including HTML, XML, and MathML

Lead Inventor:

Salvatore Stolfo, Ph.D.

Patent Information:

Patent Pending (US 20170031883)

Patent Issued (US 9,372,838)

Related Publications:


Tech Ventures Reference:

- IR M05-053
- Licensing Contact: Greg Maskel

Inventors

Salvatore Stolfo