

Georgios Petasis^{1,2}, Dimitrios Petasis¹

¹Intellitech S.A., P.O. BOX 8055, GR-19300, Aspropirgos, Greece.

²Software and Knowledge Engineering Laboratory, Institute of Informatics and Telecommunications, NCSR "Demokritos", Athens, Greece

E-mail: petasis@iit.demokritos.gr, d_petasis@hotmail.com

Motivation

The need to extract a corpus from Web resources and the Blogosphere often arises.

Desired properties of a tool for extracting a corpus from the Blogosphere:

1. **Cross-platform**: the tool must not be bind to a specific operating system.
2. **Robustness**: the tool must be as robust as possible both in its results, but also in the handling of invalid HTML code and dynamic content.
3. **Coverage**: Cover as much part of the blogosphere as possible with adequate efficiency, without the efficiency being affected by the visual layout and theme diversity.

Existing tools cover many of the above properties, but difficult to find a single tool covering everything.

Additional desired features:

1. The tool must be **integrated** with a cross-platform and **popular web browser**.
2. The engine that extracts the various pieces of information must **operate over the DOM tree** of the blog page, as constructed by the browser.
3. The **extraction engine** must be easily **configurable and adaptable** to various application requirements. In case of inadequate performance, application of heuristics that improve performance must be an easy task.

BlogBuster: extracting corpora from the Blogosphere

1. Based on the **Ellogon** natural language engineering platform.
2. Integrates **Mozilla's Gecko** rendering engine.

Advantages over existing approaches:

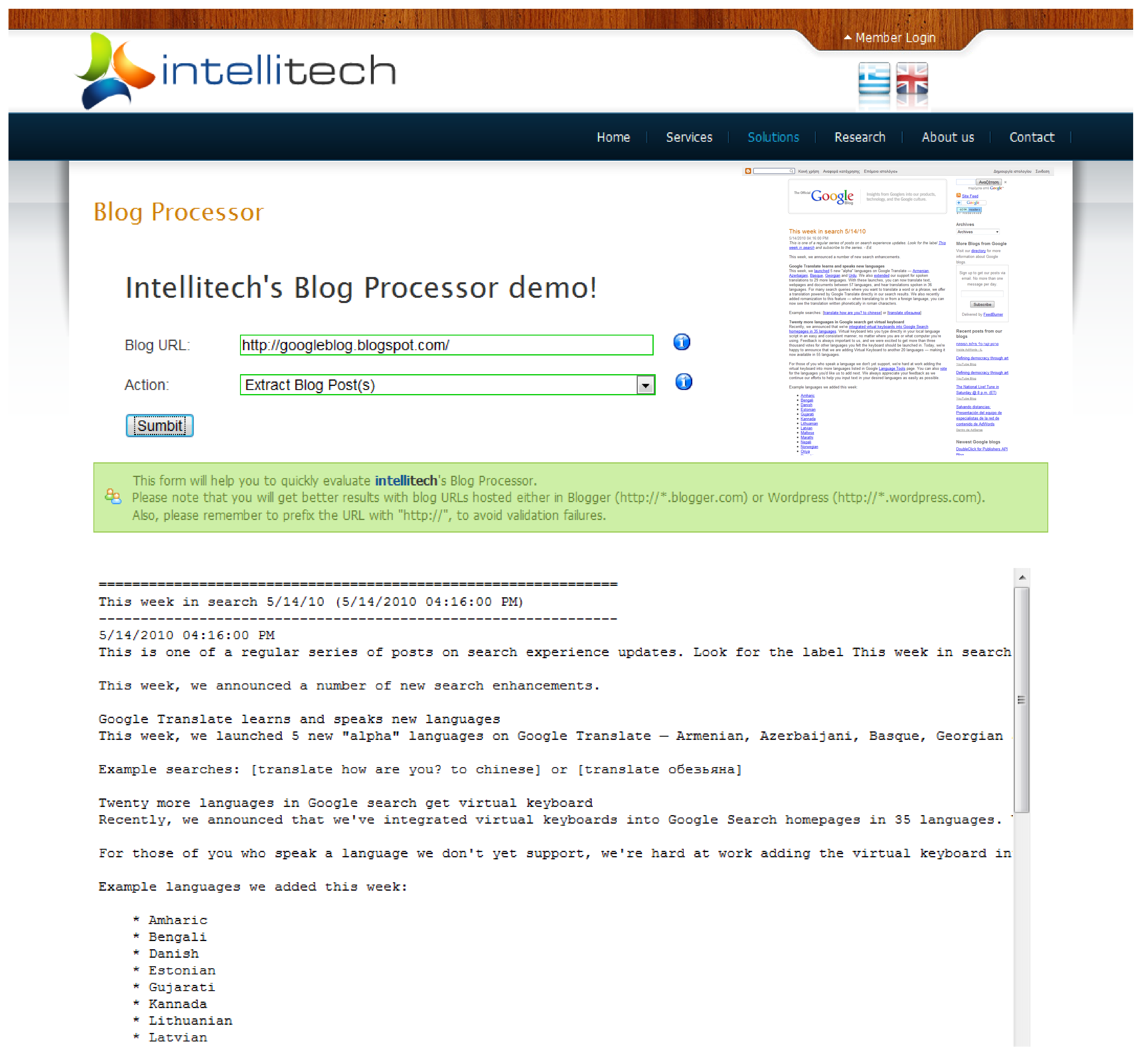
1. Adequate handling of invalid or malformed HTML pages.
2. Ability to download and store locally a blog web page, by applying the proper character encoding conversions.
3. Ability to retrieve and store content that is generated dynamically (i.e. as a result of JavaScript execution).

Architecture

BlogBuster performs the following actions upon receiving a URL of a blog for processing:

1. Instructs the rendering engine to download and render the URL. The rendering engine downloads and applies style-sheet information, creating a visual representation of the web page identical to what a web browser would provide.
2. Instructs the rendering engine to execute any available JavaScript code that must be executed.
3. The text extraction engine of the BlogBuster tool is invoked to operate upon the constructed by the rendering engine DOM (Document Object Model) tree. This extraction engine is responsible for identifying suitable DOM nodes that contain the required information, and extract the information from these nodes.
4. Collect the information extracted by the extraction engine, encode the results in the requested format (i.e. XML, JSON, plain text, etc.) and return the information to the caller.

BlogBuster Demo: <http://www.intellitech.gr/> -> Solutions -> Blog Processing



The screenshot shows the 'Blog Processor' form on the Intellitech website. The form has two input fields: 'Blog URL:' with the value 'http://googleblog.blogspot.com/' and 'Action:' with the value 'Extract Blog Post(s)'. There is a 'Submit!' button. Below the form, a green box contains instructions: 'This form will help you to quickly evaluate Intellitech's Blog Processor. Please note that you will get better results with blog URLs hosted either in Blogger (http://*.blogger.com) or Wordpress (http://*.wordpress.com). Also, please remember to prefix the URL with "http://", to avoid validation failures.'

Below the form, the extracted content is displayed. It starts with a header 'This week in search 5/14/10 (5/14/2010 04:16:00 PM)' and a sub-header '5/14/2010 04:16:00 PM'. The main text reads: 'This is one of a regular series of posts on search experience updates. Look for the label This week in search'. Below this, it says 'This week, we announced a number of new search enhancements.' followed by 'Google Translate learns and speaks new languages' and 'This week, we launched 5 new "alpha" languages on Google Translate - Armenian, Azerbaijani, Basque, Georgian'. It then lists 'Example searches: [translate how are you? to chinese] or [translate обезьяна]' and 'Twenty more languages in Google search get virtual keyboard'. It continues with 'Recently, we announced that we've integrated virtual keyboards into Google Search homepages in 35 languages.' and 'For those of you who speak a language we don't yet support, we're hard at work adding the virtual keyboard in'. Finally, it lists 'Example languages we added this week:' followed by a bulleted list: '* Amharic', '* Bengali', '* Danish', '* Estonian', '* Gujarati', '* Kannada', '* Lithuanian', '* Latvian'.

Experimental setting: corpus

1. A small set of news agency portals were selected (i.e. www.usatoday.com, www.nytimes.com, www.fox.com, www.reuters.com, www.cnn.com, www.bbc.co.uk, etc.)
2. A small set of news items (~ 500) were collected from these sites, concerning news about sports, technology, and world politics.
3. Search keywords were extracted from the collected news items, by extracting all words from the titles of news items, along with distinctive words identified through TF/IDF.
4. Google's web and blog search was used to collect blogs for each news item, from both Blogger and Wordpress. Technorati's search was also used to collect blogs for each news item, without any restriction on the hosting platform.
5. The corpus collection approach targeted in creating a parallel corpus from news and blogs, with blog posts commenting on the news events contained in the HTML pages collected from the news agencies.

Total number of blogs	416
Blogs from Blogger (*.blogspot.com)	49 (12%)
Blogs from Wordpress (*.wordpress.com)	28 (7 %)

Evaluation Results

	Precision	Recall	F-Measure
Blog Posts	100.00%	52.64%	68.98%
Title	90.41%	47.60%	62.36%
Date	31.51%	16.59%	21.73%

Blogger	Precision	Recall	F-Measure
Blog Posts	100.00%	85.71 %	92.30 %
Title	97.62 %	83.67 %	90.11 %
Date	90.48 %	77.55 %	83.52 %

WordPress	Precision	Recall	F-Measure
Blog Posts	100.00 %	96.43 %	98.18 %
Title	77.78 %	75.00 %	76.36 %
Date	29.63 %	28.57 %	29.09 %

Availability

Free for research purposes, as a Web Service.

Contacts

Mr. Georgios Petasis, petasis@iit.demokritos.gr
Intellitech, <http://www.intellitech.gr>