

BlogBuster: A tool for extracting corpora from the blogosphere



Institute of Informatics and Telecommunications, NCSR "DEMOKRITOS"

Georgios Petasis^{1,2}, Dimitrios Petasis¹

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

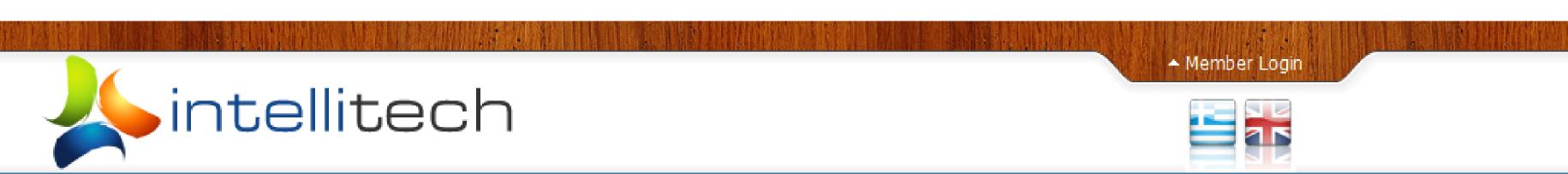
²Software and Knowledge Engineering Laboratory, Institute of Informatics and Telecommunications, NCSR "Demokritos", Athens, Greece

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.

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



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 crossplatform 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

			C Κοινή χρήση Αναφορά κατάχρησης Επόμενο ιστολόγιο»	Δημιουργία ιστολογίου Συνδε
log Processor			The Official Google Insights from Googlers into our products, technology, and the Google culture.	Avaζήτηση × παρέχεται από Google" Site Feed + Google co% readers
091100003301			This week in search 5/14/10	Archives -
			5/14/2010 04:16:00 PM This is one of a regular series of posts on search experience updates. Look for the label <u>This</u> <u>week in search</u> and subscribe to the series Ed.	More Blogs from Google Visit our <u>directory</u> for more
			This week, we announced a number of new search enhancements.	information about Google blogs.
Intellitech's	Blog Processor de	mo!	Google Translate learns and speaks new languages This week, we launched 5 new "alpha" languages on Google Translate — <u>Armenian</u> , <u>Azerbaijani</u> , <u>Basque</u> , <u>Georgian</u> and <u>Urdu</u> . We also <u>extended</u> our support for spoken translations to 29 more languages. With these launches, you can now translate text, webpages and documents between 67 languages, and hear translations spoken in 36 languages. For many search queries where you want to translate a word or a phrase, we offer a translation powered by Google Translate directly in our search results. We also recently	Sign up to get our posts via email. No more than one message per day.
	2.09		added romanization to this feature — when translating to or from a foreign language, you can now see the translation written phonetically in roman characters.	Subscribe
			Example searches: <u>[translate how are you? to chinese</u>] or <u>[translate обезьяна]</u> Twenty more languages in Google search get virtual keyboard	Delivered by FeedBurner
Blog URL: http://g	oogleblog.blogspot.com/		Recently, we announced that we've integrated virtual keyboards into Google Search homepages in 35 languages. Virtual keyboard lets you type directly in your local language script in an easy and consistent manner, no matter where you are or what computer you're using. Feedback is always important to us, and we were excited to get more than three thousand votes for other languages you feit the keyboard should be launched in. Today, we're happy to announce that we are adding Virtual Keyboard to another 20 languages — aking the second should be launched and the second should be launched and and the second should be launched and the second should be launched and the weard should be launched and and should be launched and the weard should be launched be laun	
5 5	5 5 51		now available in 55 languages.	Defining democracy through art
			For those of you who speak a language we don't yet support, wre'r hard at work adding the virtual keyboard into more languages listed in Google <u>Language Tools</u> page. You can also <u>vote</u> for the languages you'd like us to add next. We always appreciate your feedback as we continue our efforts to help you input text in your desired languages as easily as possible.	YouTube Blog 2 Defining democracy through art YouTube Blog
Action: Extract	Blog Post(s)	•	Example languages we added this week: Amharic Bengali	<u>The National Live! Tune in</u> Saturday @ 8 p.m. (ET) YouTube Blog
			• Danish • Danish • Estonian	Salvando distancias:
			• <u>Gujarati</u> • <u>Kannada</u>	Presentación del equipo de especialistas de la red de
Sumbit			• <u>Lithuanian</u> • <u>Latvian</u>	<u>contenido de AdWords</u> Dentro de AdSense
			• <u>Maitese</u> • <u>Marathi</u>	
			Nepali Norwegian	Newest Google blogs DoubleClick for Publishers API
			• <u>Oriva</u>	Blog

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.

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.
- 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:

Example searches: [translate how are you? to chinese] or [translate обезьяна]

Twenty more languages in Google search get virtual keyboard Recently, we announced that we've integrated virtual keyboards into Google Search homepages in 35 languages.

For those of you who speak a language we don't yet support, we're hard at work adding the virtual keyboard in

Example languages we added this week:

- * Amharic
- * Bengali
- * Danish
- * Estonian
- * Gujarati
- * Kannada
- * Lithuanian
- * Latvian

Experimental setting: corpus

- 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.)
- 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.

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%

BloggerPrecisionRecallF-MeasureBlog Posts100.00%85.71.%92.30.%

- 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.

- 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	49
(*.blogspot.com) Blogs from Wordpress	(12%) 28
(*.wordpress.com)	(7%)

Diog Fusis	100.00%0	03./1 %0	92.30 %0
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