

Clarifyre Text Mining

Technologies

Autonomous Dictionary Creation Using Selective Web Crawling
Taxonomy-based Page Classification
Navigation Removal Software for Page Cleansing
Scalable Selective Crawling

Services

Integration with Search Indexes such as Lucene and Solr
MySQL development
Java or PHP based Web 2.0 sites
Firefox Addons
Connectors
Search Engine Optimization (SEO)
Search Engine Marketing

Clarifyre, Inc.

Tel: (408) 891-5141, eFax: (270) 633-8733

<http://www.clarifyre.com>

1 Autonomous Dictionary Creation Using Selective Web Crawling

The Clarifyre Autonomous Knowledge Miner gathers data from targeted data sources (e.g., e-commerce websites), and creates different types of knowledge-databases that are needed by applications of interest.

One such application is Contextual-Advertising which needs updated thesauruses and dictionaries for mapping web pages to themes of interest.

2 Taxonomy-based Page Classification

The Clarifyre Taxonomy Based Page Classification engine helps you identify themes of interest on page, at various levels of granularity. You provide the taxonomy of interest to you, and dictionaries that map various words and phrases to different parts of your taxonomy, and the engine gives you a detailed classification of different clusters on the page, and their confidence levels.

Example applications that use this engine are:

- [Page mapping for Contextually Relevant Advertisements](#)
- [Channel Map](#)
- [Demo](#)

The demo is using a sample dictionary and taxonomy for the automotive industry. To run the demo:

- Please first find a web page in the automotive space (e.g., by searching on Google).
- Enter the URL for it in the text-box.
- Click on the submit button.

3 Navigation Removal Software for Page Cleansing

Navigation material and other extraneous information on a page add noise to search indexes, contextual advertising indexes, and de-duping software.

Clarifyre's Heuristic Navigation Removal removes such navigation material.

The heuristics were developed and tested with ~10,000 news articles a few years ago. It uses no page-specific script. When in doubt, the heuristics are weighted to keep a block instead of deleting it. If your application requires higher precision, we will be happy to investigate.

For more information, please see:

- [Presentation](#)
- [Demo](#)

To run the demo:

- Please first find a web page (e.g., any news page, or blog. It does not have to be automotive specific).
- Enter the URL for it in the text-box.
- Click on the link “[View page with heuristic navigation removal.](#)”

4 Scalable Selective Crawling

The Clarifyre Selective Crawler is designed to crawl very large sites. It addresses things such as:

- Limiting crawl to links that match a pattern.
- Memory management: there is no limit on the number of outstanding urls the crawler needs to process.
- Parallel execution: application level flow control to adjust to network/system bandwidth, and derive high throughput.
- Starting and stopping safely, without leaving the system in an unknown state.

5 Engineering and Marketing Services

Our areas of expertise include:

- Integration with Search Indexes such as Lucene and Solr – e.g., see <http://www.solvedex.com>
- MySQL development, including their stored procedures introduced since version 5.0
- Java or PHP based websites
- Web 2.0 features, e.g., using JavaScript/AJAX
- Firefox Addons – e.g., see the one on <http://www.xperuse.com>
- Developing Connectors to pull/synchronize large volumes of data (e.g., xml over https)
- Search Engine Optimization (SEO)
- Search Engine Marketing

For more information, please see our partner site, vugle.net

With regards to [Search Engine Optimization \(SEO\)](#), we implemented multiple SEO projects for [eBay](#) that resulted in more than \$1 million/week in new revenues in just three months.

6 Team

Our team includes well qualified and experienced software developers and researchers who have worked for eBay, Microsoft, Cisco, and Nortel. We also have a small engineering team in India.

Everyone in our US team holds a Masters or PhD in Computer Science from a top-tier school. Our India team is headed by a person who holds a PhD in Computer Science.