MANCHESTER
1824

School of
Computer Science

Information
Management Group

HCW— HuCEL Technical Report 4, August 2009

Deploying the HuCEL Java Servlet

Paul Waring
Human Centred Web Lab
School of Computer Science
University of Manchester
UK

The World Wide Web contains a vast corpus of information describing a variety of events, but this information is poorly interconnected. The aim of the HuCEL project is to provide a solution to this problem by automatically generating associative links between related events. This manual describes how to deploy the Human Centred Event Linking (HuCEL) Java servlet on Tomcat.

HCW
Human Centred Web

HuCEL

The aim of the HuCEL project is to investigate how related events can be connected on the Web, in order to improve navigation of the hypertext space and enable users to serendipitously discover new information. The HuCEL Web pages may be found at: http://hcw.cs.manchester.ac.uk/research/hucel/.

HuCEL Reports

This report is in the series of HCW HuCEL technical reports. Other reports in this series may be found in our data repository, at http://hcw-eprints.cs.man.ac.uk/view/subjects/hucel.html. Reports from other Human Centred Web projects are also available at http://hcw-eprints.cs.manchester.ac.uk/.

Contents

Human Centred Web Lab
School of Computer Science
University of Manchester
Kilburn Building
Oxford Road
Manchester
M13 9PL
UK

tel: +44 161 275 7821
http://hcw.cs.manchester.ac.uk/

Corresponding author:
Paul Waring
tel: +44 (161) 275 6239
pwaring@cs.man.ac.uk

1 Introduction

This manual details how to deploy the Java servlet used in the HuCEL project and explains how to amend the current configuration to support additional Web sites. The documentation assumes that the user knows how to deploy a standard servlet using Tomcat, and has the ability to install, or request the installation of, additional Perl modules.

2 System Requirements

The HuCEL Java servlet has a number of requirements, including minimum software versions and third party libraries, which must be satisfied before the servlet can be successfully deployed.

2.1 Minimum Software Versions

In order to run the Java servlet and associated code, certain minimum versions of standard pieces of software are required. The software may function under older versions, but no guarantees are made as to whether this will be the case, and users are strongly recommended to use the latest versions where possible. The minimum requirements include:

- Tomcat 5.5[1]
- Perl 5.8.8[2]
- Java Development Kit (compiler and runtime) 1.6.0[3]

All three pieces of software are freely available from their respective Web sites.

2.2 Specific Software

In addition to up to date versions of Tomcat, Java and Perl, some specific pieces of software are required in order for the HuCEL servlet to be successfully deployed.

2.3 WordNet

WordNet[4] is a lexical database of English words, which is used by the parser to give a rough indication of what part of speech a particular word is (noun, verb etc.). As well as the Perl module for WordNet (see Section 2.4), which provides an interface for running queries, the database itself must be installed. For most Linux distributions, this can be usually be obtained through your package management software – e.g. installing the package `wordnet` on Debian systems will provide the entire WordNet database. For other systems, users should consult the WordNet Web site for detailed installation instructions.

[1] `http://tomcat.apache.org/`

[2] `http://www.perl.org/`

[3] `http://www.java.com/`

[4] `http://wordnet.princeton.edu/`

2.4 Perl Modules

The Perl script which parses the content of each page uses several non-standard Perl modules, which must be installed before the script will execute successfully. These can be installed via CPAN[5] or through your distribution's package repository. The required modules are:

- `URI::Escape::JavaScript`
- `WordNet::QueryData`
- `Yahoo::Search`

If you are unsure whether these modules are installed or your system, simply run the parser manually from the command line. It will fail immediately if any of the required modules are not installed and located in a directory which Perl searches for libraries.[6]

2.5 SentParBreaker

SentParBreaker is a Java library for identifying boundaries of sentences and paragraphs, authored by Scott Piao[7] of the University of Manchester. It is used by our parser to split the content of a page into sentences, in order to ensure that sentence boundaries are not crossed when matching proper nouns. A copy of `sptoolkit.jar` is included in the `hucel.war` file attached to this report, and the latest version can be downloaded from the development Web site.[8]

2.6 Additional Requirements

Whilst HuCEL should be capable of running on any operating system and platform which supports Java and Perl, users are strongly recommend to deploy the server software on a recent Unix-based system such as Mac OS X or Linux.

3 Deploying the Servlet

Before deploying the servlet, a directory named `/tomcat` must be created, to which the Tomcat user must have read and write privileges. The following files must be placed into this directory:

- `complete-parser.pl`
- `content-mapping.xml`
- `stopwords.txt`

[5] `http://www.cpan.org/`

[6] These directories can usually be found in the `@INC` array.

[7] `http://personalpages.manchester.ac.uk/staff/scott.piao/`

[8] `http://text0.mib.man.ac.uk:8080/scottpiao/sent_detector`

- `locations.txt`[9]

Once the files are in place, the `hucel.war` file can be deployed in the same way as any other Java servlet.

4 Content mapping file

The content mapping file instructs the servlet to select a particular element from the DOM tree, identified as the main content of the page, for a given Web site.

4.1 Adding further sites

The addition of a site to the content mapping file is a simple task which requires identifying the deepest element within the DOM tree which contains all of the content of the page. This can be achieved by either examining the DOM tree or viewing the HTML source code of a sample page on the site.

For example, if the content was enclosed in the following tag:

```
<div id="article-wrapper">content</div>
```

then the following content mapping should be added to the configuration file:

```
<content element="div" attribute="id" value="article-wrapper"/>
```

It should be noted that the content mapping file processor does not check for multiple definitions of the same site. If you have defined two or more content mappings for the same site, the first one listed in the file will be used and all the others will be ignored – however, an error will not be raised. It is the responsibility of the user to ensure that the information contained within the content mapping file is free from errors. Furthermore, whilst sites can be added without a content element, any site entries which do not contain a content element will be ignored by the parser.

4.2 Ignored elements

Although obtaining the element which contains the content removes a significant amount of clutter, there are often elements within the content which we also want to ignore as they are either present on every page – and thus do not describe events – or they are elements such as advertisements or embedded navigation which we are not interested in. For example, many sites offer social bookmarking links which enable readers to quickly submit the current page to sites such as Facebook, Reddit[10] and Digg.[11] Whilst useful to the reader, these elements have the potential to confuse the parser, and in any point do not describe the events on the page, and thus should be removed before the content is analysed.

The content mapping file offers a method for ignoring elements on a per-site basis. For example, on the BBC News site we have the following ignore rules:

[9] This file must be obtained directly from the National Geospatial Intelligence Agency at: `http://earth-info.nga.mil/gns/html/`. Due to its size (79MB), it is not included with this report.

[10] `http://reddit.com/`

[11] `http://digg.com/`

```
<ignore element="span" attribute="class" value="di"/>
<ignore element="div" attibute="id" value="socialBookMarks"/>
```

The effect of these rules is to remove all instances of the following elements (how they would look in the HTML of the page):

```
<span class="di"></span>
<div id="socialBookMarks></div>
```

As the elements are removed from the DOM before any further processing, they do not feed into the parser and therefore do not count as part of the content of the page.

There are also globally defined ignored elements, such as the `img` element, which are removed from all sites, and therefore do not need to be specified. By default these globally defined elements are: `img`, `script`, `noscript`, `object`, `form`, and comment nodes.

5 Summary

Using the instructions contained in this manual, readers should be able to deploy and use the HuCEL servlet and configure additional sites.

6 Associated Files

The file describing associated files “” is missing