Diplom/Master-Thesis

(Diese Arbeit kann auch als Studien/Bachelor-Arbeit vergeben werden oder als studienbegleitendes Projekt.)

Improving the Wikipedia Parser

Technology	Software Process	Legal/Law	Business/Economics

The thesis can be written in either English or German. Description last revised 23.06.2010.

Summary

We have been developing an alternative parser for MediaWiki, the software running Wikipedia. However, it is a long way from a research prototype to something that improves the world. This thesis will embed our new parser into the existing MediaWiki infrastructure. It will couple the php-based MediaWiki implementation with a web service written in Java that we are developing.

Problem

Our work on improving Wikipedia has a real chance of bettering the world. However, it is a long way from research into production, with many hard technical challenges. One step along the way is to ready our work for the real-world load of a major website like Wikipedia. In parallel, we need to create the necessary social acceptance for our technical work.

To that end, we would like to develop a MediaWiki plug-in that utilizes our existing alternative parser and makes it usable for any current MediaWiki installation. We take a web-service-based approach. The MediaWiki plug-in will call out to our web service using a REST API and will utilize its functions. The first set of functions provided by our web service are called "lint", "correct", and "pretty-print". For a given wikitext (MediaWiki syntax), "lint" will suggest corrections to sloppy syntax, "correct" will correct sloppy syntax, and "pretty-print" will pretty-print a document. We believe that this non-intrusive incremental approach will drive acceptance and eventually adoption of our parser and subsequent technology.

The student taking on this thesis work should be familiar with php and be interested in web and web service programming. He or she should also like the challenge of running a real web service that might see significant use, that is, would have to run reliably under significant load. Finally, the MediaWiki plug-in will have to be developed as open source software.

Thesis Advisor

Prof. Dr. Dirk Riehle
dirk.riehle@cs.fau.de
Martensstr. 3, Raum 04.135
Sekretariat Raum 05.138

For more, see: http://osr.cs.fau.de

Friedrich-Alexander-Universität Erlangen-Nürnberg

TECHNISCHE FAKULTÄT