RegexURLFiltersBenchs

Introduction

This page provides some performance benchmarks of the regular expressions based URLFilters in Nutch (currently urlfilter-regex and urlfilter-automaton). The **urlfilter-regex** plugin is based on the standard jdk java.util.regex implementation, whereas the **urlfilter-automaton** plugin is based on dk.brics.automaton Finite-State Automata for Java.

Performance

Data set

These *performance* benchmarks were produced by collecting the results of the unit tests of each plugin using the same rule file (`Benchmarks.rules`) and the same set of urls to filter (`Benchmarks.urls`).

Raw results

The following matrix shows the **urlfilter-regex** and **urlfilter-automaton** plugins processing time in *ms* for many numbers of loops on the `Benchmarks.urls` file filtering.

	50	**100**	**200**	**400**	**800**
regex	459	899	1917	3703	7873
automaton	335	419	657	1119	1997

Graphical representation

http://frutch.free.fr/images/nutch/regexfilters-benchs.png

Conclusion

urlfilter-automaton supports less operators than **urlfilter-regex** but provides some really best performance. It can probably be usefull in some contexts.

A next step could be to mix the usage of these two plugins in order to take the best of each one by using the **`urlfilter.order`** configuration property.

How to use

You need to enable **urlfilter-automaton** plugin by editing your conf/nutch-site.xml. You need to edit **automaton-urlfilter.txt** and enter the rules. The syntax is explained here here. A good and robust grammar. No greedy/lazy kind of modes.