How to make a choice for the best URL marker for WSRP

Carsten Leue
August 6th, 2002

Problem

WSRP defines a marker the consumer may use to detect producer encoded URLs in the markup stream. The consumer should then rewrite these URLs to make them point back to itself so the consumer can intercept invocations of these URLs. This step requires parsing of the output stream. The marker should be designed such that this parsing may be performed as fast as possible.
This analysis assumes a simplified Boyer-Moore (BM) algorithm for string searching to determine the optimal marker. More information on the BM algorithm can be found here: http://www-igm.univ-mlv.fr/~lecroq/string/node14.html

Algorithm

For this analysis we use a simplified string search approach that only takes into account the best and worst case of the BM algorithm. Assume that there are p occurrences (=links) of the marker in the document. Let there be N additional characters and the size of the marker be M. Then the size of the total markup is N'=p*M+N.

Hits

We need at least p*M comparisons to locate the p marker occurrences. Let approximate the number of comparisons we need to ensure that the marker does not occur in the remaining N characters of the markup.
The simplified algorithm be the following:

- Check if the current character in the markup stream occurs in the marker. If not we may proceed by M characters in the stream.
- Assume for simplicity that we need M comparisons for the case that the current character occurs anywhere in the marker to make sure that this is not the marker. This is a very pessimistic assumption; the real BM does much better.

Probability

For the further analysis we need to calculate the probability that any character occurs in the marker. Let q(i) be the probability that the character i occurs in a typical markup stream. The probability the any character in the stream occurs in the marker is then Q(M) = 1-prod(1..M,1-q(marker[i])) where marker[i] denotes the ith character in the marker if the assume that all characters in the marker are different.

Best case

If the remaining markup stream does not contain any character in the marker, we can always step by M characters through the stream with a cost of N/M comparisons. The probability for this is 1-Q(M), so the expected number of comparisons in this case is

(1-Q(M))*N/M

Worst case

If the character in the stream is contained in the marker we need N comparisons, the probability for this is Q(M), so the expected number of comparisons N*Q(M).

Total

The assumptions above lead to an expected number of comparisons of
CMP(M) = p*M + (1-Q(M))*N/M + N*Q(M)

The goal of the analysis is to find the M that minimized CMP(M).

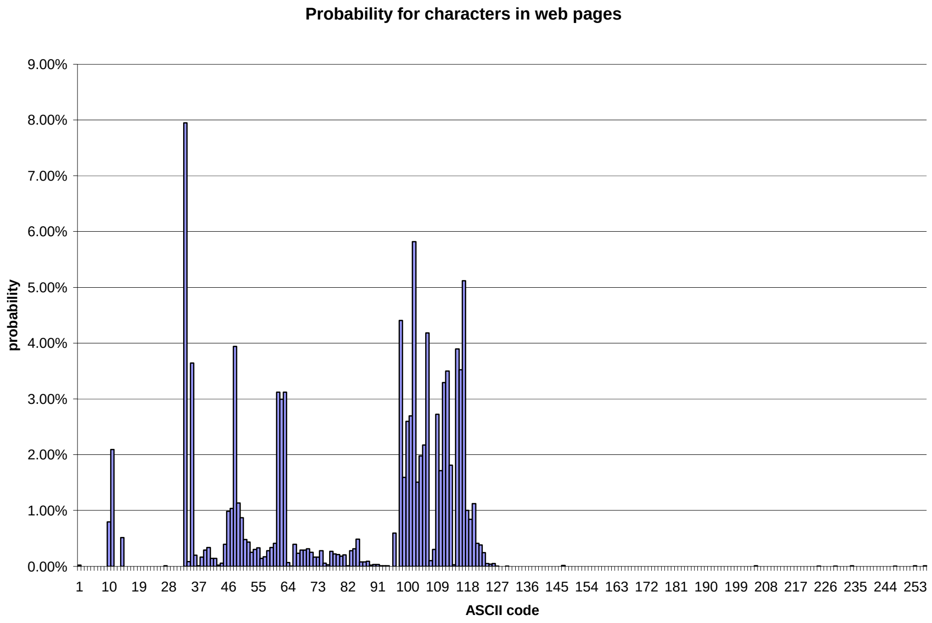
Minimization

To minimize CMP we need to be able to evaluate Q(M). This function is approximated using the following assumptions:

- All characters in the marker are different (so we can use the product formula to evaluate Q(M))
- The characters contained in the marker are chosen such that Q(M) for a given M becomes minimal. This can be achieved by selected the (valid) characters for the marker that occur with the least probability in the markup stream.
- We use heuristics to determine the probability for characters to occur in the markup stream by evaluating a set of web pages

Probability for characters in the markup stream

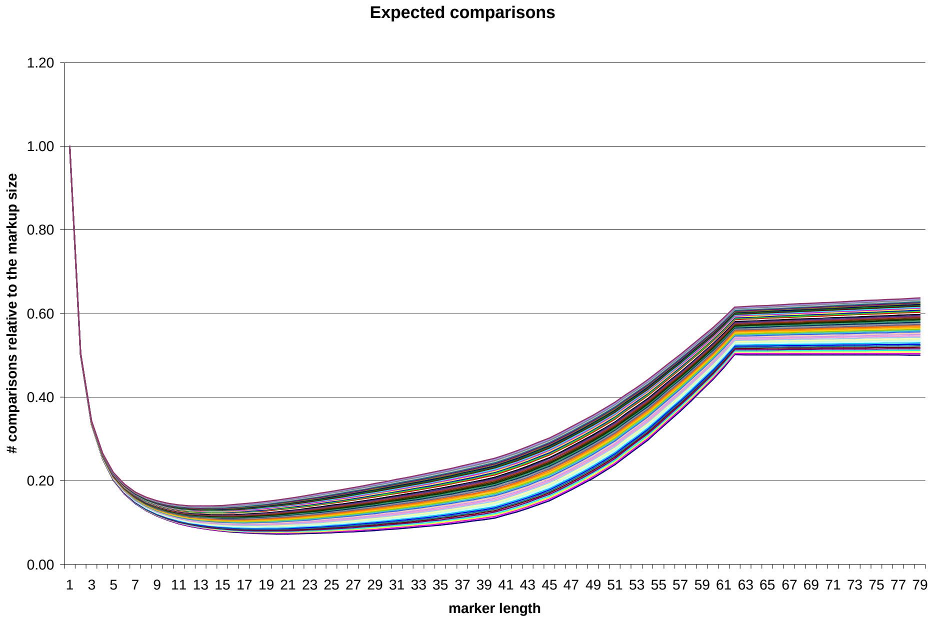
To calculate the probability for each ASCII character to occur in a markup stream we analyzed a total of 10MB of web page content from english, german and french pages. For results see the first table in the appendix. The following chart illustrates the results:



For the marker we choose only between the alphanumerical (upper and lower case) characters.

Expected number of comparisons

The expected number of comparisons depends on the ratio of the average number of links to the markup size: ratio = p/N'. For the analysis we tested ratios between 0% and 0.5% as the most probable range of interest.
The result is visualized in the chart below:



The different lines represent the result of (normalized) comparison numbers for different ratios. The chart shows that for small ratios there is no clear minimum for the number or comparisons, the whole range between a 10 to 32 sized marker would generate good results. As the ratio rises the minimum shifts to smaller sized makers. This is understandable as for large markers the probability that a character in the markup stream occurs in the marker rises also, requiring more comparisons.
From the chart we propose a general purpose marker length of 13 characters as this value lies near the minima for most of the curves.

Result

The analysis of an optimal string length together with the probabilities for all characters in a markup stream result in the following proposition for the WSRP marker:

QXqKYZJVUWj7G

This proposition is based on a series of assumptions as mentioned above. More experimental tests need to be performed to verify the usefulness of the token and to verify the results. Especially it will have to be evaluated if the multiplication of characters with low probabilities in the marker would result in better search performance.

Appendix

Occurrences of characters in the markup stream

ASCII	Total	Percentage
1	40	0.00%
2	0	0.00%
3	0	0.00%
4	24	0.00%

5	8	0.00%
6	0	0.00%
7	0	0.00%
8	8	0.00%
9	83044	0.80%
10	217768	2.09%
11	0	0.00%
12	0	0.00%
13	53232	0.51%
14	0	0.00%
15	0	0.00%
16	0	0.00%
17	0	0.00%
18	0	0.00%
19	46	0.00%
20	58	0.00%
21	0	0.00%
22	0	0.00%
23	0	0.00%
24	70	0.00%
25	86	0.00%
26	564	0.01%
27	0	0.00%
28	89	0.00%
29	37	0.00%
30	9	0.00%
31	0	0.00%
32	826520	7.95%
33	8643	0.08%
34	378927	3.64%
35	20911	0.20%
36	500	0.00%
37	16907	0.16%
38	30289	0.29%
39	35190	0.34%
40	14427	0.14%
41	14467	0.14%
42	1956	0.02%
43	5617	0.05%
44	40723	0.39%
45	102639	0.99%
46	107724	1.04%
47	409658	3.94%
48	117663	1.13%
49	90655	0.87%
50	49778	0.48%

51	44797	0.43%
52	26248	0.25%
53	31726	0.31%
54	34290	0.33%
55	14637	0.14%
56	17249	0.17%
57	29142	0.28%
58	35253	0.34%
59	42354	0.41%
60	324358	3.12%
61	311324	2.99%
62	324518	3.12%
63	6993	0.07%
64	405	0.00%
65	40861	0.39%
66	23898	0.23%
67	30484	0.29%
68	30211	0.29%
69	32739	0.31%
70	26300	0.25%
71	16750	0.16%
72	17185	0.17%
73	29210	0.28%
74	5635	0.05%
75	2930	0.03%
76	27404	0.26%
77	23181	0.22%
78	21805	0.21%
79	18690	0.18%
80	21274	0.20%
81	1090	0.01%
82	28863	0.28%
83	32646	0.31%
84	50508	0.49%
85	8121	0.08%
86	7757	0.07%
87	9200	0.09%
88	1924	0.02%
89	3406	0.03%
90	3457	0.03%
91	943	0.01%
92	547	0.01%
93	855	0.01%
94	30	0.00%
95	61839	0.59%
96	71	0.00%

97	458650	4.41%
98	165523	1.59%
99	270077	2.60%
100	280346	2.70%
101	605298	5.82%
102	156865	1.51%
103	205706	1.98%
104	225990	2.17%
105	435278	4.19%
106	10328	0.10%
107	31712	0.30%
108	283463	2.73%
109	178104	1.71%
110	342537	3.29%
111	364065	3.50%
112	188693	1.81%
113	2861	0.03%
114	405242	3.90%
115	366434	3.52%
116	532569	5.12%
117	104175	1.00%
118	87547	0.84%
119	116472	1.12%
120	42727	0.41%
121	39822	0.38%
122	25447	0.24%
123	4989	0.05%
124	3809	0.04%
125	5021	0.05%
126	318	0.00%
127	0	0.00%
128	0	0.00%
129	298	0.00%
130	0	0.00%
131	0	0.00%
132	0	0.00%
133	0	0.00%
134	0	0.00%
135	0	0.00%
136	0	0.00%
137	0	0.00%
138	0	0.00%
139	0	0.00%
140	0	0.00%
141	50	0.00%
142	0	0.00%

143	75	0.00%
144	35	0.00%
145	0	0.00%
146	1175	0.01%
147	0	0.00%
148	0	0.00%
149	0	0.00%
150	0	0.00%
151	0	0.00%
152	0	0.00%
153	0	0.00%
154	0	0.00%
155	0	0.00%
156	0	0.00%
157	0	0.00%
158	0	0.00%
159	0	0.00%
160	65	0.00%
161	3	0.00%
162	33	0.00%
163	1	0.00%
164	13	0.00%
165	0	0.00%
166	6	0.00%
167	3	0.00%
168	11	0.00%
169	143	0.00%
170	21	0.00%
171	6	0.00%
172	24	0.00%
173	10	0.00%
174	12	0.00%
175	9	0.00%
176	1	0.00%
177	11	0.00%
178	8	0.00%
179	10	0.00%
180	3	0.00%
181	24	0.00%
182	8	0.00%
183	32	0.00%
184	0	0.00%
185	15	0.00%
186	1	0.00%
187	7	0.00%
188	8	0.00%

189	26	0.00%
190	7	0.00%
191	12	0.00%
192	4	0.00%
193	12	0.00%
194	17	0.00%
195	4	0.00%
196	40	0.00%
197	24	0.00%
198	20	0.00%
199	13	0.00%
200	15	0.00%
201	55	0.00%
202	3	0.00%
203	1	0.00%
204	916	0.01%
205	18	0.00%
206	0	0.00%
207	2	0.00%
208	6	0.00%
209	3	0.00%
210	9	0.00%
211	0	0.00%
212	10	0.00%
213	1	0.00%
214	35	0.00%
215	11	0.00%
216	1	0.00%
217	4	0.00%
218	6	0.00%
219	0	0.00%
220	123	0.00%
221	4	0.00%
222	2	0.00%
223	165	0.00%
224	63	0.00%
225	2	0.00%
226	8	0.00%
227	5	0.00%
228	451	0.00%
229	4	0.00%
230	3	0.00%
231	26	0.00%
232	66	0.00%
233	721	0.01%
234	33	0.00%

235	4	0.00%
236	8	0.00%
237	11	0.00%
238	10	0.00%
239	0	0.00%
240	23	0.00%
241	17	0.00%
242	1	0.00%
243	6	0.00%
244	13	0.00%
245	6	0.00%
246	412	0.00%
247	8	0.00%
248	0	0.00%
249	10	0.00%
250	8	0.00%
251	9	0.00%
252	974	0.01%
253	0	0.00%
254	0	0.00%
255	864	0.01%