Comparision of Nutch and Google search engine implementations on the Oregon State University website.

Lyle Benedict

June 1, 2004

Summary

Both Google and Nutch did very well in a manual comparison of the most frequent queries on the OSU website. Google did slightly better overall. It was helped by having a "key" –similar to the sponsored links on Google's commercial site. Usage of the keys was very uneven.

In a few cases Nutch got stuck on endless variations of the same thing—campus sub maps or mailing lists. This was the only real problem observed with Nutch.

Study methodology

Using the list of the top 100 Queries from Oregon State University I hand compared the results between the first 81 entries using the Google and Nutch search engines. Some of these queries were not usable because they apparently were the result of submitting the default text in search boxes, e.g. "Search Here." Nutch did not crawl all of the Oregon State domains. Where the best results came from one of the domains not crawled the search was not rated.

Rating Criteria

This is inherently subjective. Results were scored on a 0 to 10 basis. A 10 required the most relevant result in the top spot. If it was in the top 4 then it would rate 8 or 9. Indirect results: spots 11+ or through a link from a high ranking spot would rate about 5. Very convoluted results would rate 1. Due to the fact that these were the most popular searches on an intranet most of these searches were extremely focused with an obvious destination. Example: uhds (university housing and dinning.) Thus this was not a good test of fuzzy searches.

Results:

Among the 54 useable results:
A binary comparison shows that:
For 28 queries Nutch and Google were both nearly perfect. 10/10
For 3 queries both were less than perfect, but tied.
For 5 queries Nutch was slightly to moderately better than Google –both did well
For 8 queries Google was slightly to moderately better than Nutch – both did well

For 3 queries Nutch was much better than Google
For 4 queries Google was much better than Nutch 10 to 6
For 2 queries Nutch did very badly, Google did badly (i.e slightly better)
For 1 query, Nutch failed, Google did very badly.

It can be seen that on the majority of the searches both did very well. There were only a few searches with more than minor differences in search quality. Google did better overall—mostly due to the use of a "key."

Specific Problems or improvements for Nutch.

Google had a key word system –sort of like a sponsored link. This is a helpful feature—but to be fully effective required more work than has been done on the OSU system.

In several cases Nutch seemed stuck on endless variations of the same thing—i.e. different sub maps of the campus map, or different postings on the same mailing list. Although infrequent, this was the most alarming cause of failure. Google presented more random and useful results. Here Google's experience with spam farms may have helped.

Google benefited by being able to crawl pdf's. Also flashed a "did you mean?" question at least once.

In at least one case Google used the page's meta-data title in the search results. Nutch just used the URL.

This particular implementation suffered because the orst.edu & osubookstore.com & osubeavers.ocsn.com pages were not crawled by Nutch. Also there was at least one popular page http://oregonstate.edu/groups/tsa/link.html which is a redirect? and thus not crawled by Nutch. Nutch displayed it's top two results for this search (tsa) as an "access denied" page.

List of Queries with notes:

Top 100 Queries # Occurrences
1. site:osulibrary.oregonstate.edu "on an in" OR -inurl:specialcollections 23671 question: not rated – system error?

2. blackboard 2295 my scores: Nutch 10, Google 10
note: identical first results, which were clearly the site wanted.
Google: used the blackboard site as a key
Nutch: got to it by the number of incoming anchors.

3. Search OSU here! 2014 : question: not rated – system error?
Note: default text on some sub-pages web sites.
4. Search Here! 1139 : question: not rated – system error?

5. campus map 1112 my scores: Nutch 7, Google 9
note: Identity of the actual "best" page here is fuzzy. identical first results—which linked to several actual campus maps. Google gave actual maps in 2nd & 3rd place. Nutch's 2nd (Google's 4th) result was a helpful "usage guide". After this Google got more random, Nutch gave a long series of mostly undesirable subsets of the original map, e.g. "Campus map—fire station."
The incoming Nutch links were all from the original campus map—i.e. out--going links from a single document – or a small community of documents could dominate the results. Google's anti—link farm algorithms may have produced better results. Google's pdf abilities also helped.

6. onid 1025 (OSU Network Identification—i.e. email) my scores: Nutch 5, Google 10
The problem here was a site specific crawl. Nutch wasn't set up to crawl www.onid.orst.edu/
It did oregonstate.edu but not orst.edu addresses.

7. chemistry 974 Nutch 10, Google 6
Site for Dept. of Chemistry was first in Nutch, not in top twenty in Google. The URL boost helped place this #1. Otherwise all results were somewhat relevant. Google was somewhat better with remaining links as it had links to the library's chemistry info & course catalog.

8. barometer 946 (Daily barometer student newspaper) Nutch 9, Google 9
the paper was #3 on both. The library's holdings occupied the #1&2 spot.

9. housing 930 Nutch 6, Google 9
The University's housing site was #1 on Nutch & #3 on Google. However Google had the off campus housing site as #1—which did not show up in the top twenty on Nutch. Possibly a popularity based ranking on Google

10. jobs 907 my scores: Nutch 6, Google 4
Here I felt that a perfect score would have had both a Human Resources and Career Services site high up in the results. Nutch had a Human Resources link first, but was very indirect to Career Services. Google was two or three links away to either until #11 Nutch had a lot of hits concentrated in the same mailing list.

11. employment 736 Nutch 6, Google 5
same rationale as above. Nutch had a career services site first—same site was #4 on Google.Other sites were all relevant, but not compelling.

12. bookstore 663 Nutch 7 Google 10
The bookstore showed up first as a key link on Google. In Nutch a link to the bookstore showed up as #3. It also showed up in the title of this document. The actual bookstore did not show up because it had an .orst domain.

13. site:osulibrary.oregonstate.edu "Where can I find journal articles on Computer Science?" OR journal OR articles OR computer OR science -inurl:specialcollections 633
question: not rated – system error?

14. tuition 615 Nutch 7 Google 10
The main tuition page showed up first on Google, not at all on Nutch (probably an orst domain) aside from that they were very similar with useful results.

15. financial aid 568 Nutch 10 Google 10
Identical and useful page in #1

16. 16. uhds 556 568 Nutch 10 Google 10 (university housing and dinning)
Identical and useful page in #1

17. site:osulibrary.oregonstate.edu "Where can I find journal articles on Mathematics & Computer Science?" OR journal OR articles OR mathematics OR computer OR science-nurl:specialcollections 544
question: not rated – system error?

18. career services 537 Nutch 10 Google 10
Identical and useful page in #1
19. map 529 Nutch 3 Google 8
Nutch had an indirect route to the campus map in #10, Google had more direct route in the top 5

20. find someone 514 Nutch 10 Google 10
Identical and useful page in #1

21. mom's weekend 467 Nutch 10 Google 10
Identical and useful page in #1
Note: minor bug in parsing input query—submitted to bug list

22. calendar 438 Nutch 10 Google 10
Identical and useful page in #1

23. tsa 403 Nutch 1 Google 10 (thai student association)
top two Nutch queries crawled an illegal directory?

24. directory 392 Nutch 6 Google 10
the basic directory (find someone) was #1 on Google due to key, and #4 in list
in Nutch it did not appear, but the advanced search function was #4

25.human resources 385 Nutch 10 Google 10
Identical and useful page in #1
26.biology 369 Nutch 10 Google 9
biology dept. #1 in Nutch #2 in Google

27. dixon 364 Nutch 5 Google 5
both were nearly identical with Dixon lodge in #1 and Dixon Rec. Center Map only in #5.
Low score based on my guess everybody wanted the rec center instead of the lodge.
28. greek life 323 Nutch 10 Google 10
Identical and useful page in #1 Note: Google also had a key of doubtful utility
29. moms weekend 301 Nutch 6 Google 2
This is the same as #21, except that #21 is “mom’s weekend” This strongly hints that it would be useful to ignore single quotes while fetching.
note Nutch found 9 answers, Google found 40 possibly due to a deeper crawl? There was no other obvious pattern. Nutch scored better because the most relevant entries were on the first page.

30. graduate school 283 Nutch 10 Google 10
Identical and useful page in #1

31. transcripts 273 Nutch 10 Google 9
best page was #1 in Nutch #3 in Google Nutch apparently got the better score due to a the url boost.

32. library 268 Nutch 10 Google 10
Identical and useful page in #1, Google's was due to a key, otherwise main page was #3

33. hobbes 265 Nutch 1 Google 10
I am guess this is a search for material from the course phl302
oregonstate.edu/instruct/phl302/phl302-site-map.htm Nutch didn't crawl it.

34. math 249 Nutch 9 Google 10
Dept of mathematics in #2 in Nutch, #1 in Google due to key, otherwise #3

35. student directory 248 Nutch 9 Google 10
"find someone" question #20 was #3 in Nutch #1 in Google due to a key—otherwise not in top 10

36. Search the OSU Web here! 247 question: not rated – system error?

37. academic calendar 246 Nutch 7 Google 9
Google had a calendar site in #1 due to key. The actual academic calendar was #5 in Nutch #2 in Google. Google probably ranked it higher because of the internal target links all titled "academic calendar".

38. catalog 244 268 Nutch 10 Google 10
Identical and useful page in #1, Google also had a key Google's was slightly superior because it displayed the metadata page title. Nutch displayed the URL.

39. student health services 239 Nutch 10 Google 10
Identical and useful page in #1, Google had a key to a useless site.

40. registrar 235 Nutch 10 Google 10
Identical and useful page in #1 Google had a key to the registration information—possibly useful

41. Barometer 235 same as query #8

42. Search the Micronutrient Information Center 231 question: not rated – system error?

43. webmail 226 Nutch 5 Google 8
orst.edu domain so not in Nutch—but reasonable direct route. The actual webmail site was not in the top 20 in Google, but was easily reachable from #1

44. economics 226 Nutch 9 Google 9
Dept. Economics in #3 in Google #2 in Nutch

45. business 222 235 Nutch 10 Google 10
Identical and useful page in #1 Google had a key as well

46. horticulture 214 Nutch 9 Google 10
Dept. of horticulture #1 in Nutch, #2 in Google.

47. logo 206 Nutch 1 Google 3
best site retrieved was still indirect. It was #10 on Google #14 on Nutch, Nutch got stuck on a mailing list again

48. pharmacy 203 Nutch 10 Google 10
Identical and useful page in #1 Google had a key as well

49.college of business 203 Nutch 10 Google 10
Identical and useful page in #1 Google had a key as well

50. physics 201 Nutch 9 Google 6
Dept. of physics #3 in Nutch #10 in Google—but Google also had a useful catalog page in #5

51. quran 199 Nutch 10 Google 10
unique and not terribly useful page found—don't know why so many searches for this.

52. scholarships 199 203 Nutch 9 Google 9
Identical and useful page in #2

site:food.oregonstate.edu 198 question: not rated – system error?

53. asosu 190 Nutch 10 Google 10
Identical and useful page in #1
graduation 185

54. UHDS 183 same as #16
dixon recreation center 164 Nutch 1 Google 10
Google got it by using a key, and other top pages were also useful. The "best" page was labled "rec sports" –not Dixon recreation center. Nutch failed badly because at least the first 50 entries were to subsets of the campus map. Anti-spam software would have helped.
55. student involvement 163 190 Nutch 10 Google 10
Identical and useful page in #1
56. address 157 not rated--ambigious
57. engineering 156 Nutch 10 Google 10
Identical and useful page in #1, Google got it due to a key
58. Human Resources 156 Nutch 10 Google 10
Identical and useful page in #1, Google's key was misleading
59. webcam 152 Nutch 10 Google 10
Identical and useful page in #1, Google also had a key
60. memorial union 147 Nutch 10 Google 10
Identical and useful page in #1, Google also had a key

61. greek 146 Nutch 10 Google 10
Identical and useful page in #1, Google also had a key
62. Search OSU Extension 145 question: not rated – system error?
63. econ 145 Nutch 5 Google 10
Dept. of Economics #1 in Google #10 in Nutch. Boosting the last segment of the url would have helped here
64. MU 139 Nutch 0 Google 2
The memorial union was #15 in Google,
65. mu 137 same as #63
66. Housing 135 same as #9
67. email 134—not rated best pages in orst.edu domain
68. osu logo 133 Nutch 3 Google 9
Best pages were high up in Google. Indirectly reachable with Nutch
69. Mom's Weekend 131 same as #21
70. Blackboard 130 same as #2
71. study abroad 130 Nutch 9 Google 10
best site #2 in Nutch, #1 in Google

72.football 130 Nutch 0 Google 10 –but not rated due to domain not crawled.
osubeavers.ocsn.com/sports/m-footbl is correct domain

73.book store 129 Nutch 0 Google 10 –but not rated due to domain not crawled osubookstore.com Google had not only a key but a did you mean bookstore?

74. Search montaigne! 129 question: not rated – system error?

75. writing center 128 Nutch 10 Google 10
Identical and useful page in #1
76. parking 127 Nutch 10 Google 10
Identical redirect #1, Google also had a key to the actual site (orst.edu domain)

77. resnet 126 Nutch 10 Google 10
Identical and useful page in #1

78. shs 124 Nutch 8 Google 10
#4 in Nutch, #1 in Google

79. kidspirit 124 126 Nutch 10 Google 10
Identical and useful page in #1

80. geo 300 124 Nutch 7 Google 5
neither Google nor Nutch had a link to this online course's home page. The syllabus was # 1 in Nutch, #5 in Google. –probably no incoming links, so this was the best that could be expected.

81.Dixon 120 same as #27