Web Analyst: Software for Locating and Analyzing Virtual Communities, with a Korean-American Case Study

Sun-Ki Chai
University of Hawai`i
http://www2.hawaii.edu/~sunki/
Why the Need for Web Analysis Software?

• The internet has a huge and growing importance of as medium for communication, social interaction, commerce, and political activity.
• The sheer amount of data available is unprecedented.
• Most data on internet is publicly available, accessible from anywhere in the world, and can be collected with minimal intrusion on human subjects.
• Social scientists are rapidly turning to internet as an object of study.
• Despite the quantity and accessibility of data, very few large-scale studies have been done.
• Absence of technology for collecting and organizing this data is the major roadblock.
Web Analyst (Kibitzer?) Software Features

Agent software for examining social structure of world wide web – *identify, demarcate, and analyzes virtual communities*

1. intelligent, social science theory-informed crawling to identify virtual communities

2. automatic collection and compilation of site data through patented algorithm - downloaded to dataset usable by third-party software

3. analysis engine for calculating centralities, word frequencies, feedback to (1)
Interaction with Web Analyst

1. selection of "seed site(s)" and / or identifying keywords
2. selection of crawling algorithm - combination of social network and content-matching constructs to filter candidate sites
3. selection of intrasite filtering and overall halting criteria
4. as it runs, it incrementally expands community size and generates data
5. draws information from crawls of sites, search engines, online databases
6. end product dataset that can be subject to further analysis
Web Analysis Network Expansion
<table>
<thead>
<tr>
<th>url</th>
<th>title</th>
<th>trafficrank</th>
<th>age</th>
<th>urlcound</th>
<th>inurlcound</th>
<th>srcurrcount</th>
<th>outurlcound</th>
<th>bytecount</th>
<th>incentrality</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://petitiononline.com/">http://petitiononline.com/</a></td>
<td>Petition Online.com</td>
<td>5254</td>
<td>7079</td>
<td>61000</td>
<td>261</td>
<td>35</td>
<td>585</td>
<td>266054</td>
<td>0</td>
</tr>
<tr>
<td><a href="http://ipetitions.com/">http://ipetitions.com/</a></td>
<td>IPetitions - Your voice counts</td>
<td>52940</td>
<td>7008</td>
<td>5200</td>
<td>290</td>
<td>112</td>
<td>153</td>
<td>215868</td>
<td>158</td>
</tr>
<tr>
<td><a href="http://thepetitionsite.com">http://thepetitionsite.com</a></td>
<td>The Petition Site</td>
<td>26878</td>
<td>2220</td>
<td>653</td>
<td>437</td>
<td>14649</td>
<td>1892049</td>
<td>5416</td>
<td></td>
</tr>
<tr>
<td><a href="http://petitions.org/">http://petitions.org/</a></td>
<td>Petitions.org</td>
<td>.</td>
<td>2</td>
<td>12</td>
<td>2</td>
<td>107</td>
<td>73123</td>
<td>1428</td>
<td></td>
</tr>
<tr>
<td><a href="http://ethical-business.com">http://ethical-business.com</a></td>
<td>Free online petitions</td>
<td>566647</td>
<td>7495</td>
<td>896</td>
<td>8</td>
<td>18</td>
<td>123</td>
<td>68148</td>
<td>0</td>
</tr>
<tr>
<td><a href="http://webpetitions.com">http://webpetitions.com</a></td>
<td>Webpetitions.com</td>
<td>3436376</td>
<td>7904</td>
<td>5</td>
<td>2</td>
<td>108</td>
<td>23390</td>
<td>10000</td>
<td></td>
</tr>
<tr>
<td><a href="http://petitionpetition.com">http://petitionpetition.com</a></td>
<td>Petition Petition</td>
<td>165643</td>
<td>7127</td>
<td>234</td>
<td>55</td>
<td>103</td>
<td>240</td>
<td>76756</td>
<td>10000</td>
</tr>
<tr>
<td><a href="http://boycotts.org/">http://boycotts.org/</a></td>
<td>Boycott Action News</td>
<td>1666063</td>
<td>6643</td>
<td>4</td>
<td>200</td>
<td>3</td>
<td>65</td>
<td>74263</td>
<td>90</td>
</tr>
<tr>
<td><a href="http://breakthechain.org/">http://breakthechain.org/</a></td>
<td>Break the Chain</td>
<td>82261</td>
<td>7746</td>
<td>625</td>
<td>463</td>
<td>375</td>
<td>540</td>
<td>278377</td>
<td>833</td>
</tr>
<tr>
<td><a href="http://care2.com/">http://care2.com/</a></td>
<td>Care2.com</td>
<td>1957</td>
<td>6938</td>
<td>47800</td>
<td>3960</td>
<td>634</td>
<td>1798</td>
<td>678807</td>
<td>5568</td>
</tr>
<tr>
<td><a href="http://eternaltreblinka.com">http://eternaltreblinka.com</a></td>
<td>Eternaltreblinka.com</td>
<td>3540369</td>
<td>.</td>
<td>24</td>
<td>68</td>
<td>8</td>
<td>385</td>
<td>207629</td>
<td>1014</td>
</tr>
<tr>
<td><a href="http://coopamerica.org/">http://coopamerica.org/</a></td>
<td>Co-op America</td>
<td>209201</td>
<td>5776</td>
<td>638</td>
<td>1440</td>
<td>194</td>
<td>744</td>
<td>361730</td>
<td>643</td>
</tr>
<tr>
<td><a href="http://ivu.org/">http://ivu.org/</a></td>
<td>International Vegetarian Union</td>
<td>45093</td>
<td>6279</td>
<td>8410</td>
<td>1150</td>
<td>204</td>
<td>661</td>
<td>212594</td>
<td>2506</td>
</tr>
<tr>
<td><a href="http://vrg.org/">http://vrg.org/</a></td>
<td>Vegetarian Resource Group</td>
<td>76302</td>
<td>6221</td>
<td>828</td>
<td>1670</td>
<td>112</td>
<td>1419</td>
<td>563538</td>
<td>2982</td>
</tr>
<tr>
<td><a href="http://dorway.com/">http://dorway.com/</a></td>
<td>Aspartame is not Safe</td>
<td>174944</td>
<td>6098</td>
<td>395</td>
<td>250</td>
<td>83</td>
<td>665</td>
<td>613540</td>
<td>470</td>
</tr>
<tr>
<td><a href="http://vegetarianimage.com">http://vegetarianimage.com</a></td>
<td>Vegetarianimage.com</td>
<td>3235450</td>
<td>7475</td>
<td>22</td>
<td>32</td>
<td>112</td>
<td>181</td>
<td>107072</td>
<td>482</td>
</tr>
<tr>
<td><a href="http://upc-online.org/">http://upc-online.org/</a></td>
<td>United Poultry Concerns</td>
<td>569498</td>
<td>6814</td>
<td>1290</td>
<td>514</td>
<td>58</td>
<td>425</td>
<td>229981</td>
<td>1770</td>
</tr>
<tr>
<td><a href="http://organichealthandbeauty.com/">http://organichealthandbeauty.com/</a></td>
<td>Organic Health and Beauty</td>
<td>1117276</td>
<td>7557</td>
<td>198</td>
<td>276</td>
<td>166</td>
<td>193</td>
<td>156432</td>
<td>1853</td>
</tr>
<tr>
<td><a href="http://meatout.org/">http://meatout.org/</a></td>
<td>The Great American Meatout</td>
<td>11142317</td>
<td>6420</td>
<td>167</td>
<td>182</td>
<td>144</td>
<td>322</td>
<td>308837</td>
<td>1357</td>
</tr>
<tr>
<td><a href="http://happycow.net/">http://happycow.net/</a></td>
<td>HappyCow's Global Guide to Vegetarian</td>
<td>247631</td>
<td>7270</td>
<td>1090</td>
<td>796</td>
<td>275</td>
<td>542</td>
<td>482936</td>
<td>633</td>
</tr>
<tr>
<td><a href="http://dmoz.org/">http://dmoz.org/</a></td>
<td>DMOZ</td>
<td>177</td>
<td>7518</td>
<td>513000</td>
<td>342000</td>
<td>83</td>
<td>1700</td>
<td>235091</td>
<td>412</td>
</tr>
<tr>
<td><a href="http://soystache.com/">http://soystache.com/</a></td>
<td>SoyStache</td>
<td>1492911</td>
<td>7053</td>
<td>78</td>
<td>87</td>
<td>67</td>
<td>1618</td>
<td>1015836</td>
<td>923</td>
</tr>
<tr>
<td><a href="http://famousveggie.com/">http://famousveggie.com/</a></td>
<td>Famousveggie.com</td>
<td>470867</td>
<td>7268</td>
<td>573</td>
<td>413</td>
<td>192</td>
<td>242</td>
<td>100206</td>
<td>1561</td>
</tr>
<tr>
<td><a href="http://ciwf.co.uk/">http://ciwf.co.uk/</a></td>
<td>Compassion in World Farming (CIWF)</td>
<td>844936</td>
<td>541</td>
<td>349</td>
<td>64</td>
<td>211</td>
<td>244142</td>
<td>1719</td>
<td></td>
</tr>
</tbody>
</table>
Interface to “Social Scientists” Version of Analyst
Some Possible Applications

• Finding which combination of content, network, and offline (geographical, ownership) characteristics are most significant in determining a site's popularity and influence

• Determining the characteristics of the sites within a community or which link to a particular web resource (their related interests, the people behind it)

• Using content analysis, including comparison with web corpuses, to do a kind of "reverse search engine" that identifies the key phrases that will most surely a particular community.

• Determine structural and attitudinal characteristics of populations without intrusive methods
“Obscure” Communities Often Provide Clearest Data about Social Communication

irani chai, mumbai
remembering the Irani cafes of Mumbai

http://www.iranichaimumbai.com/

http://boblind.com/board/index.php

http://yaebasuperstar.blog123.fc2.com/
Tracking Internet “Memes” across Communities Can be Less than Straightforward Task
Hub Sites Range from Intensely Personal to Partisan to Avowedly Objective


http://www.malaysiakini.com/

Pakatan keeps 2 'bukits', BN wins B'gan

http://www.malaysiakini.com/

Still King of the Hills

updated 9:40pm: Pakatan Rakyat captures both the Hills with its candidates in Bukit Gantang and Bukit Selambau, taking up an enviable majority. (See:

- Trial tag fails to sway votes
- Najib's presence has no impact

Polling day: Tear gas, rain, phantom buses

FREE LIVE REPORTS: Polling in Bukit Gantang and Bukit Selambau ends at 5pm. Counting begins at 7.30pm and results are expected after 10pm. (See:

- Early voting brisk in all seats
- Police fire tear gas into crowd

Khairy not likely to get cabinet post

Speculations are rife that Umno Youth chief Khairy Jamaluddin will not be included when PM Najib announces the new cabinet line-up tomorrow. (See:

- Khairy's dream of cabinet seat dashed
- Khairy's old rivals may be droppen
Choice of Korean-American Culture Sites as Initial Object of Study

• They represent diverse, but still distinct, community.
• Linkages exist to other communities across political, economic, and cultural boundaries, but they do not overwhelm strong internal linkages.
• While smaller communities may provide cleaner results, it easier to locate initial set of seed sites for a large community like this and to justify studying it for its own sake.
• The Korean-American community is much-written about in the “non-virtual” literature, so there is extensive material that can be used for comparisons with data collected online.
Korean-American Case Study: Seed Sites from Open Directory

- http://www.net4korean.com/
- http://www.cohka.org/
- http://www.haninweb.com/
- http://www.igoo.com/
- http://www.kaccsv.org/
- http://www.kahs.org/
- http://www.kamr.org/
- http://www.kasf.org/
- http://www.kaso.org/
- http://www.koreadirect.com/
- http://www.koreanculture.org/
- http://www.kyccla.org/
- http://www.kyopounion.com/
- http://www.nakasec.org/
- http://www.skas.org/
- http://www.sokorean.com/
- http://www.storefrontlive.net/
- http://www.ykan.org/
**Inclusion Criterion**

\[
\text{argmax}_{p \notin S'} \left( \sum_{p, q \in E, q \in S'} \frac{C_{in}(q)}{D_{in}(q)} \right) \times \left( \sum_{q, p \in E} \frac{C_{out}(q)}{D_{out}(q)} \right)
\]

\[D_{in}\text{ and } D_{out}\text{ are degrees links in and out}\]

\[
C(p) = (1 - \sigma) + \sigma \left( \sum_{p, q \in E} \frac{C(q)}{D(q)} \right)
\]

\(C\) is eigenvector centrality ("PageRank"),

\(\sigma\) is a “damping factor”
A Sampling of Sites in Final Virtual Community

069020875|http://www.sacredland.org/
032818926|http://www.janet.org/
024658758|http://www.ltsc.org/
013863800|http://www.resisters.com/
013391558|http://www.itvs.org/
013190868|http://www.orcapod.com/
012480165|http://www.wingluke.org/
012005530|http://www.njamf.org/
010520432|http://www.apanet.org/
010098351|http://www.asianweek.com/
009895500|http://www.airos.org/
009328795|http://www.naatanet.org/
008875020|http://www.nativetelecom.org/
004727702|http://www.nikkeiwest.com/
004624875|http://tac.ltsc.org/
004526280|http://www.katonk.com/
004371329|http://www.laddmedia.com/
004196250|http://www.nikkeiheritage.org/
003366870|http://www.kamuseum.org/
003060408|http://www.jcccnc.org/
003001482|http://www.ocanatl.org/
New Themes in “Found” Sites

• Pan-Asian, Japanese-American Culture

• Minority Political Advocacy Organizations

• “Related Interests”
  – Korean War History
  – Adoption Agencies
  – Christian Evangelism
Predicting Popularity:
Korean-American Interest Websites

<table>
<thead>
<tr>
<th>DV</th>
<th>Popularity</th>
<th>Popularity</th>
<th>Popularity</th>
<th>Popularity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-33203</td>
<td>42288</td>
<td>-180317</td>
<td>-332999</td>
</tr>
<tr>
<td></td>
<td>99939</td>
<td>136774</td>
<td>229474</td>
<td>320539</td>
</tr>
<tr>
<td>Links In</td>
<td>3836.01</td>
<td>3454.84</td>
<td>2719.59</td>
<td>3639.45</td>
</tr>
<tr>
<td></td>
<td>2892.49</td>
<td>2951.92</td>
<td>3205.39</td>
<td>3639.45</td>
</tr>
<tr>
<td>Links Out</td>
<td>230.45</td>
<td>295.46</td>
<td>318.42</td>
<td>817.10</td>
</tr>
<tr>
<td></td>
<td>768.80</td>
<td>768.12</td>
<td>817.45</td>
<td>817.10</td>
</tr>
<tr>
<td>Links In *</td>
<td>11.04</td>
<td>10.53</td>
<td>9.87</td>
<td>9.87</td>
</tr>
<tr>
<td>inks Out</td>
<td>13.21</td>
<td>13.60</td>
<td>14.54</td>
<td>14.54</td>
</tr>
<tr>
<td>Centrality In</td>
<td>70.64*</td>
<td>67.83*</td>
<td>74.22*</td>
<td>74.22*</td>
</tr>
<tr>
<td></td>
<td>31.50</td>
<td>31.39</td>
<td>36.47</td>
<td>36.47</td>
</tr>
<tr>
<td>Centrality</td>
<td>47.00</td>
<td>10.79</td>
<td>15.41</td>
<td>15.41</td>
</tr>
<tr>
<td>Out</td>
<td>37.76</td>
<td>43.20</td>
<td>47.50</td>
<td>47.50</td>
</tr>
<tr>
<td>C In *</td>
<td>-0.005</td>
<td>-0.003</td>
<td>-0.004</td>
<td>-0.004</td>
</tr>
<tr>
<td>C Out</td>
<td>0.004</td>
<td>0.004</td>
<td>0.004</td>
<td>0.004</td>
</tr>
<tr>
<td>Site Age</td>
<td></td>
<td></td>
<td></td>
<td>133.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>142.31</td>
</tr>
</tbody>
</table>

| N  | 249        | 249        | 249        | 222        |
| R2 | 0.0295     | 0.0234     | 0.0486     | 0.0537     |

Note: Results of OLS regressions. The first row in each cell represents unstandardized parameter estimates, the second standard errors. Star represents $p < 0.05$. 
Next Steps

• More Flexible Inclusion Criteria
• Integration of Content, Link, and “Offline” Analysis Better
• User Interface
  • Better Integration with 3-party Software
• Specialized Analysis of Web Forums, Blogs, Content-Sharing Sites, Social Networking Sites
  – mutuality of commenting/replying
  – shared investigation vs. argument/flaming
  – time to respond,
  – quantity and thread depth of response
  – lurkers vs. contributers
  – cross-posting by contributors
  – platform-specific tags