Something about Ranking in Web2.0

NCKU CSIE
Hung-Yu Kao
2008. 06
Outlines

● Something about Web2.0
● Ranking / Searching meets Web2.0
  ● Link analysis
  ● Our work on ranking

Click it ➔ A video clip from YouTube
Web2.0 Tag Cloud

Markus Angermeier http://nerdwideweb.com/web20/
What is Web2.0

business embracing the web as a platform and using its strengths (global audiences, for example).

Tim O'Reilly
Web 1.0

DoubleClick  -->  Google AdSense
Ofoto         -->  Flickr
Akamai        -->  BitTorrent
mp3.com       -->  Napster
Britannica Online  -->  Wikipedia

personal websites -->  blogging
evite         -->  upcoming.org and EVDB
domain name speculation  -->  search engine optimization
page views     -->  cost per click
screen scraping -->  web services
publishing     -->  participation
content management systems  -->  wikis
directories (taxonomy)  -->  tagging ("folksonomy")
stickiness     -->  syndication

Web 2.0

Application Hierarchy

- Level-3 application
  - the most "Web 2.0"-oriented
  - eBay, Craigslist, Wikipedia, del.icio.us, Skype, dodgeball and AdSense
- Level-2 application
  - operate offline but gain advantages from going online
  - Flickr
- Level-1 application
  - operate offline but gain features online
  - Writely, iTunes
- Level-0 application
  - work as well offline as online
  - MapQuest, Yahoo! Local and Google Maps

Where is Amazon?
New Interactions
Differences for researchers

- Throng of pages
  - With complicated, but ruled styles
  - Informational v.s. emotional
- Throng of links
  - physical v.s. virtual links
  - diverse v.s. clustered
Ranking meets Web2.0

- Rank pages, rank people
- More interaction, much **capitalism** impact
- Information **extraction** / **understanding** become essential
Link analysis -- Motivation

- For one query, which pages are the answer set?
  - Results of search engines
    - Rank manually
    - Rank by similarity
    - Rank by hit rate (*need usage log*)
    - Rank by link analysis (google)
  - Relevant v.s. Authoritative
    - Intra-page v.s. inter-page
    - 筆試 v.s. 口試
  - *Users need authoritative pages among relevant pages.*
Link analysis -- Motivation

- Human knowledge is real, convincing and trustable information
  - E.g., classification by human in yahoo
- Hyperlinks contain information about the human judgment
- Social sciences
  - Nodes: persons, organizations
  - Edges: social interaction
- Easy job? Counting in-links for popularity
An example: scientific literature

  - for journal evaluation
  - *Garfield (Science 1955, 1972)*
  - The average number of citations per recently published item

- \( \frac{C}{N} \)
  - \( C \): the total number of citations in a given time interval \([t, t + t1]\) to articles published by a given journal during \([t – t2, t]\)
  - \( N \): the total number of articles published by that journal in \([t – t2, t]\)
HITS - Kleinberg’s Algorithm

• HITS – Hypertext Induced Topic Selection
• For each vertex \( v \in V \) in a subgraph of interest:
  \[
  a(v) - \text{the authority of } v \\
  h(v) - \text{the hubness of } v
  \]

• A site is very authoritative if it receives many citations. Citation from important sites weight more than citations from less-important sites.

• Hubness shows the importance of a site. A good hub is a site that links to many authoritative sites.

雞生蛋，蛋生雞？
Introduction

- II: extend the root set to base set
  - Problems
    - Unrelated page of large in-degree
  - New approach (kleinberg ’97)
    - There should also be considerable overlap in the sets of pages that point to authoritative pages.
      - Hub pages
      - mutually reinforcing relationship
Authority and Hubness

\[ a(1) = h(2) + h(3) + h(4) \]

\[ h(1) = a(5) + a(6) + a(7) \]
Example (1-norm normalization)

Authority

Hub

\[
\frac{3^n}{3 \cdot 3^n + 2 \cdot 2^n + 2 \cdot n}
\]

\[
\frac{3^n}{3^n + 2^n + n}
\]
HITS Example

Find a base subgraph:

• Start with a root set \( R \{1, 2, 3, 4\} \)

• \( \{1, 2, 3, 4\} \) - nodes relevant to the topic

• Expand the root set \( R \) to include all the children and a fixed number of parents of nodes in \( R \)
  • *Indegree v.s. outdegree*
  → A new set \( S \) (base subgraph) →
HITS Example Results

Authority and hubness weights

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

Authority and hubness weights

Root set R
Issues for HITS

- Mutually reinforcing relationships between hosts
  - Nepotistic links cancellation
    - Nepotistic links: links between pages that are present for reasons other than merit
      - Menu links
      - Link-based spam
  - Link normalization
PageRank

  - The weight is assigned by the rank of parents
    \[ r(v) = \alpha \sum_{w \in \text{pa}[v]} \frac{r(w)}{|\text{ch}[w]|}, \]

- Difference with HITS
  - HITS takes Hubness & Authority weights
  - The page rank is proportional to its parents’ rank, but inversely proportional to its parents’ outdegree
  - Query independent
Matrix Notation

- Confirm the result
  # of inlinks from high ranked page
  hard to explain about 5&2, 6&7

- Interesting Topic
  * How do you create your homepage highly ranked?
  * How to detect it?
Stability

- Whether the link analysis algorithms based on eigenvectors are stable in the sense that results don’t change significantly?
- The connectivity of a portion of the graph is changed arbitrary
  - How will it affect the results of algorithms?

Ng et al (2001, SIGIR) – “stable algorithms for link analysis”
Why We Care

- Lempel and Moran (2001) showed theoretically that SALSA weights are more robust than HITS weights in the presence of the **Tightly Knit Community** (TKC) Effect.
  - This effect occurs when a small collection of pages (related to a given topic) is connected so that *every hub links to every authority* and includes as a special case the mutual reinforcement effect.

- The pages in a community connected in this way can be *ranked highly* by HITS, higher than pages in a much larger collection where only *some* hubs link to *some* authorities.

- TKC could be exploited by *spammers* hoping to increase their page weight (e.g. link farms).
Limits of Link Analysis

- META tags/ invisible text
  - Search engines relying on meta tags in documents are often misled (intentionally) by web developers
- Pay-for-place
  - Search engine bias: organizations pay search engines and page rank
  - Advertisements: organizations pay high ranking pages for advertising space
    - With a primary effect of increased visibility to end users and a secondary effect of increased respectability due to relevance to high ranking page
- *Inside Web Page Patron Graph*
Limits of Link Analysis

- **Stability**
  - Adding even a small number of nodes/edges to the graph has a significant impact

- **Topic drift – similar to TKC**
  - A top authority may be a hub of pages on a different topic resulting in increased rank of the authority page

- **Content evolution**
  - Adding/removing links/content can affect the intuitive authority rank of a page requiring recalculation of page ranks
  - *Incremental link analysis*

- 子曰：眾好之，必查之，眾惡之，必查之
Case Study: PageRank Captialism
Application: PageRank Capitalism

- J. Cho et al. propose that it may take 60 times longer for a new page to become popular under the search-domain model than random-surfer model.
- PageRank value is equal to the “currently important value”.
- A newly created page does not have enough in-links to show its “true importance”.
- We try a new algorithm called DRank to diminish the bias of PageRank-like link analysis. (Directory-feature-based ranking algorithm)
Aggregation Graph

- **Experiment setup:**
  - reverse the components of host path of all URLs in our dataset and sort the modified URLs in lexical order.
    - For example, we change URL [www.csie.ncku.edu.tw/student/index.htm](http://www.csie.ncku.edu.tw/student/index.htm) to [tw.edu.ncku.csie.www/student/index.htm](http://tw.edu.ncku.csie.www/student/index.htm), so that URLs in the same domains or hosts

(a) A slice includes 267 hosts

(b) 2 hosts in .gov domain
Figure 12: Ranking Distribution of Pages in a Directory
Figure 13: Ranking distribution of pages in directories which are randomly selected from top 20 directories
A detail view of the red area of Figure 13(f)
Directory Feature Based Rank (DRank)

- Review Page Quality

\[
\hat{Q}(p,t_i) = \frac{n}{r} \left( \frac{\Delta PR(p,t_i)/\Delta t_i}{PR(p,t_i)} \right) + PR(p,t_i)
\]

Figure 12(a) A directory within W3C

Figure 12(b) A directory in Microsoft

Figure 13: Examples of Drank algorithm
## Experiment Setup (cont.)

<table>
<thead>
<tr>
<th></th>
<th>Total Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pages</td>
<td>2,906,212</td>
</tr>
<tr>
<td>Hyperlinks</td>
<td>75,475,454</td>
</tr>
<tr>
<td>Hosts</td>
<td>202,674</td>
</tr>
<tr>
<td>Directories</td>
<td>893,692</td>
</tr>
<tr>
<td>Intra-link (Host Level)</td>
<td>53,924,805</td>
</tr>
<tr>
<td>Inter-link (Host Level)</td>
<td>21,550,649</td>
</tr>
</tbody>
</table>
Snapshots Simulation

- The growth rate is 1.2 per unit time, and Web is growing base on an exponential growth copying model [12].
- The time slice of original dataset is at T20
- The percentage of total number of hyperlink at T20 is 100.
- The percentage of total number of hyperlink at T17 is about 58 compared to T20.
- The percentage of total number of hyperlink at T15 is about 40 compared to T20.

![Figure 16: Percentage of Link of Different Snapshots](image)
Compare with Page Quality and Modified Page Quality

<table>
<thead>
<tr>
<th>Experiment</th>
<th>WeightRank</th>
<th>Page Quality</th>
<th>Drank</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Err(p)</strong></td>
<td>0.133163729</td>
<td>0.08232336</td>
<td>0.07119299</td>
</tr>
<tr>
<td><strong>Experiment II</strong></td>
<td>WeightRank</td>
<td>Modified Page Quality</td>
<td>Drank</td>
</tr>
<tr>
<td><strong>Err(p)</strong></td>
<td>0.133163729</td>
<td>0.04919891</td>
<td>0.01948247</td>
</tr>
</tbody>
</table>

13.5%+  
60.4%+
Observation

- We found that if we could assign reasonable ranking values to pages that were newly created among important directories we could diminish the bias against the majority of important new pages.

- Solution: We introduce a novel directory-feature-based method to diminish the bias of link analysis against new pages.

- DRank algorithm works well while the predictive values of pages in the first process ($DR_0$) are close to the ladder-lines they are actually belong to.
Case Study: CSS information extraction
Motivation

- Informative block (IB) that presented in a form of block on the Web is meaningful data for extractor on page analysis.
- Blog is hot! There are many investigation on it.
  - Ex: social network and trend analysis
- There are something different between Blog page and general page on IB scoring and ranking.
  - *DOM tree is not a flat tree already.*
Tag distribution
Idea

- **3D-DOM**
  - A 3-dimensional tree
  - using properties of CSS to enhance IB extraction
  - three properties different from general page
    - CSS Tag Entropy
    - Layer Containment
    - CSS File description
CSS Tag Entropy

\[
E(B) = -\frac{1}{6} \log_6 \frac{1}{6} - \frac{4}{6} \log_6 \frac{4}{6} - \frac{1}{6} \log_6 \frac{1}{6} = 0.3767776
\]

\[
E(A) = \left(-\frac{1}{4} \log_4 \frac{1}{4}\right) \times 4 = 0.6020599
\]
Layer Containment

- Layer Containment
  - Structural containment: DOM
  - Logical containment: page presentation
- Layer: Presentation block and Container block

![Diagram of Layer Containment]

- `content` node
- `blog` node
- `date` node
  - `datediv` node
- `blogbody` node
  - `blogbody2` node
  - `articletext` node
    - `title` node
    - `innertext` node
    - `extended` node
    - `posted` node
Layer Containment

- Logical Containment: the blocks presentation on a page is not equal to the arrangement of

```
[DIY:0] [0] [0]
[DIY:0] [mainFrame:1.78176] [0]
+ [DIY:0] [bannar1:0.563147] [0]
+ [DIV:0] [mainImage:0.38307] [0]
+ [DIY:0] [adText1:0] [0]
+ [DIY:0] [adadText1:0.383147] [0]
+ [DIY:0] [copyright:0] [0]
+ [DIV:0] [contentDiv:1.2] [0]
```
CSS File description

- Hidden block

- Presentation block
  - Background: url(…)

- Article block
  - Block area and location
  - font-size, font-weight, font-style, font-family and text-decoration.
### Preliminary Experiments

#### Node Entropy Evaluation

<table>
<thead>
<tr>
<th></th>
<th>HTML</th>
<th>ID</th>
<th>CLASS</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>-1</td>
<td>0.45056</td>
</tr>
<tr>
<td>B1</td>
<td>0.27118 9</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>B2</td>
<td>0</td>
<td>-1</td>
<td>1.09861</td>
</tr>
<tr>
<td>C1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>C2</td>
<td>0</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>C3</td>
<td>0</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>C4</td>
<td>1.03972</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>C5</td>
<td>0</td>
<td>-1</td>
<td>0</td>
</tr>
</tbody>
</table>
Preliminary Experiments

- Node Entropy Evaluation

Delayed presentation
Case study: Blog Ranking
Introduction

- Blog is an emerging online medium for people to publish content, and the number of blogs now exceeds 500 million worldwide.
- People want to know which blogs really contain useful information.
- A blog consists of title, subscription information, multiple posts which are displayed in a descending order by the publish date.
- Blog content is different from typical web page contents.
Motivation

- Blog Look (部落格觀察) is a famous blog ranking service which combines the number of link search results and subscription data into some features for ranking.

- Some detailed information for blog interactions are not considered in rankings for Blog Look
  - Global ranking v.s. local ranking
Dataset

- Current data is crawled from the famous blog service providers

<table>
<thead>
<tr>
<th></th>
<th>Blogs</th>
<th>Article</th>
<th>Comment</th>
<th>Trackback</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wretch</td>
<td>61,264</td>
<td>6,880,087</td>
<td>16,527,101</td>
<td>316,263</td>
</tr>
<tr>
<td>Yahoo</td>
<td>8,287</td>
<td>727,335</td>
<td>1,589,940</td>
<td>137,232</td>
</tr>
<tr>
<td>Yam</td>
<td>79,399</td>
<td>1,895,319</td>
<td>2,318,052</td>
<td>104,594</td>
</tr>
<tr>
<td>Xuite</td>
<td>20,848</td>
<td>1,270,830</td>
<td>822,398</td>
<td>21,053</td>
</tr>
<tr>
<td>Pixnet</td>
<td>38,112</td>
<td>2,511,188</td>
<td>4,356,075</td>
<td>14,336</td>
</tr>
<tr>
<td>Roodo</td>
<td>15,468</td>
<td>1,076,808</td>
<td>2,642,063</td>
<td>95,906</td>
</tr>
</tbody>
</table>
Idea

- The blog factors and activities in the blogosphere form a special link structure.
- To identify important bloggers, the interactions or links between all bloggers are great indicators.
  - Comment
  - Trackback
  - Blogrolls
  - Hyperlinks in the Content
- Some additional blog information such as the number of self-defined categories, date of posts will be used as **Blog Features** in our algorithm.
Blog Interactions

- **Comment**
  - Blogger of blog A make comments on blog B
  - The most common interactions
  - It’s considered as an one-way interaction, since it may not be representative of social interaction from the sender's perspective

- **Trackback**
  - An article of blog A contains a back-reference to Blog B (highly relevant content)
  - This is more representative of the interaction between A and B

- **Blogroll**
  - Subscription relation
Blog Features

- The number of articles, visitors and custom-defined categories in a blog
- Last Article Date: the date of the latest article
- Blog Life Cycle: the time span between Last Article Date and First Article Date
- Max Article Life: the max (time span between Last Comment Date and First Comment Date for an article) in a blog
- Average Article Life: the average (time span between Last Comment Date and First Comment Date for an article) in a blog
- The number of blog user commenting (trackbacking) to other blogs and commented (trackbacked) by other ones
- The number of common Links with other blogs
Method

- Extract article, comments, trackbacks from the crawled blog pages
- Construct the Blog Network through blog relationships using the links in the content and blog interactions (ex: comment, trackback)
- Calculate Blog Features for each blog.
- Employ our modified version of Page Rank algorithm on the blog network to do the Local Blog Ranking work
- Use global link information to obtain a Global Blog Rank
Blog Network

- Each node represents a blog, and each edge between two nodes represents a relationship for the two blogs.
- There are three general types of edges in the blog network:
  - **Support Edge** (Support Relationships): comment, trackback between blogs.
  - **Similarity Edge** (Similarity Relationships): common links in contents or users between blogs (a virtual edge with lower weight).
  - **Hyperlink Edge**: the links in contents between blog and a web page.
Blog Network Illustration

- The illustration of Blog Network

- Blog
- The web page linked from the article
- Support Edge
- Similarity Edge
- Hyperlink Edge
Local Blog Rank Algorithm

- The probability formula

\[ P_{A \rightarrow B} = \frac{R_{A \rightarrow B}}{\sum (R_{A \rightarrow X})} \]

- X are the blogs to which the Blog A links

- The Relationship Score \( R_{A \rightarrow K} \) combines all kinds of relationship between blog A and K, and is calculated by the weight and number of corresponding relationship type multiplying the blog quality score of K

\[ R_{A \rightarrow K} = \sum W_{R_{\text{type}}} \times \text{Number}_{R_{\text{type}}} \times BQ_k \]
Local Blog Rank Algorithm

- The list of relationship types between A and K
  - Comments A made to K
  - Trackbacks to K in A
  - Blogrolls between A and K
  - Common links between A and K
  - Common Users between A and K

- Each relationship has a corresponding weight, and the adjustment will be discussed in the experiment
Global Rank

- Combine Local Rank with the following blog look feature for global rank
- Index Number
  - Google Link
  - Yahoo Link
  - Technorati Link
- This feature indicate the indexing state for a blog in the world wide web
- Our local blog rank present the blog activity in the local BSP
- \( \text{Global Blog Rank} = \text{Normalized Local Blog Rank} \times \text{Normalized Index Number} \)
Experiment

- Use the data for each BSP to construct the Blog Network respectively
- Calculate the Blog Rank
- In order to examine if our results are reliable, we collect some other ranking data for comparison
  - Blog Look
  - FunP
  - Human Rank
    - We choose top 100 blogs for four BSP including Wretch, Yahoo, Yam, Pixnet (one blog for one user) in the ordering of the rankings in Blog Look
    - 400 blogs total are checked by four researchers and the scores given manually will be our human ranking result
Human Rank

- Main Features
  - Detailed Discussion
  - Abundant Information
  - Clear and Suitable Self-defined Categories
  - Update Frequency
  - The quality of subscriptions
  - Suitable Advertisements

- Additional Features
  - If the user is a celebrity (this would be useful for analysis)
  - If the user provide its Email and Location

- Basically the main features decide the human ranking result
Experimental Results - Yam

- We use the summation of five local blog ranks for each division to compare with the human rank.

![Graph showing 5 Local Rank Sum](image)

1. An Online Game Blog
2. A closed Blog
Discussions

- In the 5 local rank sum for Yam, two blogs are highly ranked by local blog rank but gain low scores made by human.
- One blog is the official site for an online game, and attention from too many game players resulted in the high ranking score.
- Another blog is closed and thus the human rank is low despite the rich information in the earlier posts.
- These two special cases are not appreciated by our researchers, but somehow they attract some readers.
Experimental Results - Pixnet

- We use the summation of five local blog ranks for each division to compare with the human rank.

![Graph showing 5 Local Rank Sum]

Two Blogs like person diary.
Discussions

- There are also two special blogs in the 5 local rank sum for Pixnet, and both are like someone’s diary.

- This result is generated by some visitors, which are bloggers with high attention, making comments on the two blogs. This information shows some internal activity, but may be useless for many readers.

- We use Global Blog Rank to reduce this effect.
Experimental Results - Pixnet

- We use the summation of five global blog ranks for each division to compare with the human rank.
- The tail of this curve is not towards up, but the blog for a Japanese idol gain low score in our human rank.
- This is another interesting case for hot idols.

![Graph showing the summation of five global blog ranks compared with human rank.](image-url)
Visualization-Food
Visualization-Food
Conclusion

*Everything becomes* interesting, difficult, complicated, and useful when its size becomes larger, larger, and larger.*
Other than ranking, but funny

- 愛情公寓 (http://www.i-part.com.tw/)
- WARM (http://warm.stu.edu.tw/)

If you never touch the social network, take a look on them.
  - Ning, Flickr, Facebook, …