Using link analysis to discover web communities and improve search engine retrieval

Apostolos Gerasoulis, Brian D. Davison, Konstantinos Kleisouris, Yingfang Lu, Hyun-ju Seo, Wei Wang, Baohua Wu

Department of Computer Science
Rutgers University
A: Connectivity Matrix

\[ A^{k+1} = \begin{bmatrix} A^t & 0 \\ A & 0 \end{bmatrix} \]

\( k=1 \Rightarrow \text{"edge counting"} \)

\( a \): authorities; \( h \): hubs
DiscoWeb Approach

- Select initial web pages via text analysis (the results of a search engine query)
- Use link analysis to re-rank pages based on influence in graph
- Calculate multiple eigenvectors, each representing a community ranking
- More general method than CLEVER (IBM/Cornell) and Google (Stanford)
DiscoWeb can retrieve top-rated sites (as listed by Infoseek’s Directory)

DiscoWeb also suggests other good sites (not included in Infoseek)

DiscoWeb ranking is performed automatically, without human guidance
Current Work

Goal: Fast, scalable search engine to provide large datasets and local content for link analysis

- Enable fast analysis
- Enable analysis of large graphs

Next tasks: distributed design, parallel solver, robust public version
**DiscoWeb Interface**

**DiscoWeb!**

**New search:** [Search Bar]

**Display:** [10]

- [art]
- [automotive]
- [business]
- [careers]
- [entertainment]
- [family]
- [health]
- [kids]
- [money]
- [people]
- [science]
- [shopping]
- [spirit]
- [sports]
- [test]
- [travel]

---

**search results from DiscoWeb**

( Total : 7390 )

**Keywords : engine**

**Matches 1 – 10**

1. **Title:** Alaska Search Engine : Canada Search Engine [ALCANSEEK.COM]  
   **Score:** 5.0000  
   **URL:** http://www.alcanseek.com/employment/publish.htm  
   **Description:** Alaska & Canada’s official search engine. Search all of Alaska and Canada for travel, sports, business, news, entertainment and more....

2. **Title:** Alaska Search Engine : Canada Search Engine [ALCANSEEK.COM]  
   **Score:** 5.0000  
   **URL:** http://www.alcanseek.com/computersinternet/saleservice.htm  
   **Description:** Alaska & Canada’s official search engine. Search all of Alaska and Canada for travel, sports, business, news, entertainment and more....

3. **Title:** *BikeCrawler *Search Engine for the Cycling World *Bicycle Search Engine  
   **Score:** 5.0000  
   **URL:** http://www.bikecrawler.com/  
   **Description:** A search engine targeted toward the bicycling industry. The BikeCrawler Search Engine contains search categories for manufacturers, tour companies, bicycle clubs...
DiscoWeb Infrastructure
Clusters

- Link analysis produces clusters corresponding to web communities.
- Each eigenvalue gives the weight of the cluster.
- Values in each eigenvector rank pages in that community.
Example of Family

- There are many different communities represented.
- Communities include genealogy, family health, parenthood, Christian family values, family care givers.
Search Results for "family"

1. **Genealogical Helper**  Score = 0.7979  
   URL : http://www.everton.com/

2. **The Genealogy Home Page**  Score = 0.7845  
   URL : http://www.genhomepage.com/

3. **Family Tree Maker's Genealogy Site**  Score = 0.7518  
   URL : http://www.familytreemaker.com/

4. **Focus on the Family with Dr. James Dobson**  Score = 0.7115  
   URL : http://www.family.org/

5. **American Academy of Family Physicians Home Page**  Score = 0.6613  
   URL : http://www.aafp.org/

6. **Welcome to HomeArts!**  Score = 0.4572  
   URL : http://homearts.com/

7. **Welcome to Parenting Q&A**  Score = 0.4431  
   URL : http://www.parenting-qa.com/

8. **Family Caregiver Alliance**  Score = 0.4371  
   URL : http://www.caregiver.org/

9. **Dove Family Approved Movies and Videos**  Score = 0.4369  
   URL : http://www.dove.org/

10. **Crayola**  Score = 0.4293  
    URL : http://www.crayola.com/
Genealogy Community

1. **Family Tree Maker’s Genealogy Site**  
   Score = 0.7100  
   URL: http://www.familytreemaker.com/

2. **The Genealogy Home Page**  
   Score = 0.0357  
   URL: http://www.genhomepage.com/

3. **Genealogical Helper**  
   Score = 0.0324  
   URL: http://www.everton.com/

4. **Genealogy**  
   Score = 0.0279  
   URL: http://www.our-town.com/~puppy/welcome.html

5. **StarMedia**  
   Score = 0.0279  

6. **No Title**  
   Score = 0.0279  
   URL: http://www.comuni-k.com/compu2.htm

7. **Robert Slavey’s Genealogy Page**  
   Score = 0.0279  
   URL: http://www.slavey.com/slavey/gene.html

8. **History: Genealogy**  
   Score = 0.0279  
   URL: http://www.snap.com/directory/category/016,-34946,00.html

9. **The Genealogy Connection Genealogy Links**  
   Score = 0.0279  
   URL: http://www.home.dwave.net/~skeeter/genlink1.html

10. **Genealogy Computer Interest Group**  
    Score = 0.0279  
    URL: http://www.avenue.org/Community/Agencies/Lake/gen-cig.html
AltaVista Search Service for "Weather"

AltaVista found about 28,054,232 Web pages for you.

1. Lycos Global Weather Service
   URL: weather.lycos.com/

2. The Weather Channel Brasil – Tempo para o Brasil e o Mundo
   URL: tempo.weather.com/

3. Weather Service Finland
   URL: www.weather.fi/

4. Los Angeles Weather – NWS
   URL: www.nwsla.noaa.gov/

   URL: www.channel2000.com/

6. Weather and Warnings Displayed Graphically at Xenocode, Inc.
   URL: www.weathernode.com/imap.asp

7. Central Region of the National Weather Service
   URL: www.crh.noaa.gov/

8. Boise National Weather Service
   URL: www.boi.noaa.gov/

9. Weather
   URL: www.vas-das.com/

10. PC Weather Products Home Page
    URL: www.pcwp.com/
DiscoWeb Search Results for "Weather"

★★★ Best  ★★★ Very Good  ★ Good  :: Infoseek Directory Stars

1. **Weather Underground: Welcome to ..**  Score = 0.8984  ★★★
   URL : http://www.wunderground.com/

2. **Yahoo Weather Forecast**  Score = 0.8924
   URL : http://www.weather.yahoo.com/

3. **National Weather Service**  Score = 0.7875  ★★★
   URL : http://www.nws.noaa.gov/

4. **USA Weather**  Score = 0.7374  ★★★
   URL : http://www.intellicast.com/weather/usa/

5. **The University of Michigan Weather Underground**  Score = 0.6225  ★★
   URL : http://groundhog.sprl.umich.edu/

6. **National Climatic Data Center**  Score = 0.6123
   URL : http://www.ncdc.noaa.gov/

7. **The Weather Network Online**  Score = 0.6050
   URL : http://www.theweathernetwork.com/

8. **America's oldest source of weather**  Score = 0.5245
   URL : http://www.almanac.com/

9. **Space Science and Engineering Center**  Score = 0.4638
   URL : http://www.ssec.wisc.edu/

10. **Welcome to WeatherOffice**  Score = 0.4522
    URL : http://www.weatheroffice.com/
AltaVista Search Service for "Java"

AltaVista found about 18,254,865 Web pages for you.

1. hav.Software – Neural Nets to Internet – Java et.al.
   URL: www.hav.com/

2. WingSoft – One Stop Shop for Java Tools
   URL: www.wingsoft.com/

3. Caf151soft: A Java Software Development Company
   URL: www.cafesoft.com/

4. CORBA, COM, RPC, and Java development
   URL: www.nouveau-corba.com/

5. Natural Intelligence – Java, Oracle...
   URL: www.natural.com/

   URL: www.servletcentral.com/

7. Munica CGI SQL JAVA and Database Application Center
   URL: www.servletcentral.com/

8. JC VRML & JAVA Consulting: 3D Sports for the Web
   URL: www.technowin.com/

9. WebStrike! – an online java game
    URL: www.webstrike.com/

10. Java and Web Services
    URL: www.jaws.umn.edu/
DiscoWeb Search Results for "Java"

★★★ Best  ★★★ Very Good  ★ Good  :: Infoseek Directory Stars

1. Java Applet Rating Service (JARS)  Score = 1.0837  ★★★
   URL: www.jars.com

2. Java Linux  Score = 1.0027
   URL: www.blackdown.org/java-linux.html

3. Gamelan: Earthweb’s Java Directory  Score = 0.9196  ★★★
   URL: www.gamelan.com/wx/

4. Java(TM) Technology Home Page  Score = 0.8526  ★★
   URL: java.sun.com/

5. CodeWarrior  Score = 0.7082
   URL: www.metrowerks.com/

6. The Java Apache Project  Score = 0.5893
   URL: java.apache.org/

7. Java Repository: for Java Programmers  Score = 0.5891  ★★★
   URL: java.wiwi.uni-frankfurt.de/

8. Front Page -- www.arachnoid.com  Score = 0.5665
   URL: www.arachnoid.com/

9. Artistic Reality  Score = 0.5329
   URL: www.artreality.com

10. Java Lobby Homepage  Score = 0.5310  ★★
    URL: www.javalobby.org/
Hotbot Search Service for "Java"

Hotbot found about 1,926,360 Web pages for you.

1. **Java Technology Home Page**
   URL: www.java.sun.com/

2. **Gamelan**
   URL: www.gamelan.com/

3. **MediaBuilder–Java Animation**
   URL: deutsch.medibuilder.com/softwaregraphicsjava.html

4. **ComSpace.Com**
   URL: www.comspace.com/

5. **Digital Cats’ Java(tm)**
   URL: www.javacats.com/

6. **Sun Microsystems’ Java Platform**
   URL: ac030.osu.cs/java/jdk/docs/api/packages.html

7. **Sun Microsystems’ Java Platform**
   URL: www2.cs.utah.edu/classes/cs3510/jdk1.1.7B/docs/api/packages.html

8. **Netpedia-->The Solution for Web Developers**
   URL: netpedia.com/

9. **programmers’ and developers’ resources**
   URL: www-a.developer.com/

10. **Package Summary**
    URL: service.felk.cvut.cz/doc/java/share/jdk1.2beta3/docs/api/packages.html