Social Data

Toby Segaran
Author, *Programming Collective Intelligence*
Data Magnate, *Metaweb Technologies*
Data mining?

“Sorting through data* to identify patterns and establish relationships”

* usually a lot of data
Where and why?

Methods and examples
Where and why?

- Targeted Advertising
- Recommendations
- Search Results
- Group Discovery
- Filtering of Documents
- Theme Extraction
Sponsored Links

**ING Bay to Breakers 12K**
Registration is now open for the May 18th race in SF!
www.ingbaytobreakers.com

**Pres a Vi**
Unique Brunch in a Unique Setting in the Presidio
www.presavi.com

**Looking to Lease in SF?**
Search For Current Rental Listings Gaetani, Inc, Serving SF Since 1946
www.GaetaniRealty.com

**Romantic SF Restaurant**
Best of Citysearch Winner! Come Experience First Crush For Yourself
www.FirstCrush.com
This is strange...

- Google just has text
- Facebook knows more about me
- But it’s taking a few cues...
Status: “engaged”
Where and why?

- Targeted Advertising
- Recommendations
- Search Results
- Group Discovery
- Filtering of Documents
- Theme Extraction
Real Amazon Products
# Netflix Prize

## Leaderboard

<table>
<thead>
<tr>
<th>Rank</th>
<th>Team Name</th>
<th>Best Score</th>
<th>Improvement</th>
<th>Last Submit Time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No Grand Prize candidates yet</td>
<td>--</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td></td>
<td>No Progress Prize candidates yet</td>
<td>--</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Grand Prize - RMSE &lt;= 0.8563</strong></td>
<td>--</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>BellKor</td>
<td>0.8650</td>
<td>9.08</td>
<td>2005-04-01 17:02:59</td>
</tr>
<tr>
<td>2</td>
<td>When Gravity and Dinosaurs Unite</td>
<td>0.8675</td>
<td>8.82</td>
<td>2003-04-06 19:34:42</td>
</tr>
<tr>
<td>3</td>
<td>BigChaos</td>
<td>0.8681</td>
<td>8.76</td>
<td>2008-04-11 20:28:22</td>
</tr>
<tr>
<td>4</td>
<td>Gravity</td>
<td>0.8693</td>
<td>8.63</td>
<td>2008-04-09 10:18:08</td>
</tr>
<tr>
<td></td>
<td><strong>Progress Prize - RMSE &lt;= 0.8625</strong></td>
<td>--</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>KGBell</td>
<td>0.8712</td>
<td>8.43</td>
<td>2007-10-01 23:25:23</td>
</tr>
<tr>
<td>6</td>
<td>acmeh8</td>
<td>0.8718</td>
<td>8.37</td>
<td>2009-06-26 02:07:17</td>
</tr>
<tr>
<td>7</td>
<td>tesla</td>
<td>0.8722</td>
<td>8.32</td>
<td>2008-05-03 08:19:08</td>
</tr>
<tr>
<td>8</td>
<td>Dan Tibergh</td>
<td>0.8723</td>
<td>8.31</td>
<td>2008-03-28 20:58:09</td>
</tr>
</tbody>
</table>

**Progress Prize 2007 - RMSE = 0.8712 - Winning Team: KGBell**
Strands Contest

$100,000
Call for Recommender Start-Ups

The Fastest Path to Fund your Project

1. Submit your project
2. Present at RecSys08
3. Get funded

The Call seeks to identify the best early stage project in the area of recommendation technologies

http://recommender-startups.strands.com
Custom News

Panic Button

You have 850 unread items. That's a lot to read!

This isn't email - you don't have to read everything. Perhaps it's time to hit the panic button and let FeedDemon mark items as read for you?

Mark as read

all items more than 48 hours old

Options...  Hit the Panic Button!  Cancel
Custom News

My Findory News

IBM Advances Autonomous Computing for Data Centers

Newscientist - In announcing new capabilities in its six-year drive for Autonomous Computing, IBM took another step toward the goal of computing systems that can heal and manage themselves. (read more)

The top global technology brand for 2007 is ...

Newsweek - The top global technology brand for 2007 is either Google or Microsoft, depending on which market research report you believe. Google's rapid growth as a leading global brand is telling, in any case. (read more)

Justice Department says no to extension of Microsoft oversight

MarketWatch - SAN FRANCISCO - The U.S. Department of Justice told a judge Friday that it sees no reason to extend restrictive antitrust oversight of Microsoft Corp., countering a recent move by a group of states to extend oversight for another five years. (read more)

2 girls slain in South Jersey home

NBC News - Two girls, ages 6 and 10, were killed Tuesday night and their 20-year-old half-brother was arrested in their deaths, authorities ... (read more)

Wood Chipper Swallows Calif. Tree Trimmer

ABC News - A group of friends has not yet identified the landowner involved in the accident. (read more)

6-Wave's quantum computer ready for latency demo

Newsweek - Company says a scientist from Google will join in with an image recognition algorithm at the IQC supercomputing conference. Photos 6-Wave takes a quantum leap. (read more)

Photos: Surface computer draws a crowd

Newsweek - Despite a delay in its launch until next spring, Microsoft is taking its tabletop Surface computer on a promotion tour around the world. (read more)

Rising Gone Wrong: Woman Charged for filing

ABC News - A woman was charged with being off her ex-boyfriend's lip while kissing. (read more)

My Findory Blogs

Rodrigo Mayor: Version control systems

Peanuts静电 - I left last night (should compiling all GNOME modules, included meta-gnome-preferred, to find this morning) it failed on lifelessness because of missing data in my system. Installed it and watched it work until it came to another module needing to be installed again, and then, a few minutes later, another module complained about missing miscערך. And so, so good, but this makes 69% ... (read more)

Dear Mayor Larry,

Daily News (Thursday 9th)

Good morning, Mayor Chavez! How are things there in Albquerque? It's my hometown, and I miss it. I've met you a few times, like back when I was a forensics science graduate, and did that Citizen Police Academy program. You hung out at our piano party, and two years later, you remembered me when you came to one of our Howard Dean MeetUps. Kind of cool, and a pleasant ... (read more)

Abortion and Christianity

Andrew Sullivan (Friday 7th)

A Christianite takes on Garry Wills. Stating "Aquinas said" is not the same as saying "Christendom said" or even "the Catholic Church said." The church did not accept his arguments wholesale. For instance, the church ultimately rejected Aquinas arguments against ... (read more)

"Once every hundred years, a Harvard man ..."

Veilgna (Friday 7th)

"Once every hundred years, a Harvard man gets his facts right." -- An MIT graduate took on Zuckerberg's Law, as proclaimed by the Facebook founder this week. (read more)

The New Facebook Ads - Another Privacy Debacle?

Wired (Friday 7th)

privacyofwires "Facebook recently announced a new advertising scheme called Social Ads. Instead of using celebrities to hawk products, it will use pictures of Facebook users. Facebook might be entering into another privacy debacle. This site assumes that if people rate products highly or write good things about a product then they consent to being used in an advertisement for it. Facebook ... (read more)
Custom News
Where and why?

- Targeted Advertising
- Recommendations
- **Search Results**
- Group Discovery
- Filtering of Documents
- Theme Extraction
Ranking algorithms

The PageRank Citation Ranking: Bringing Order to the Web

January 29, 1998

Abstract

The importance of a Web page is an inherently subjective matter, which depends on the readers interests, knowledge and attitudes. But there is still much that can be said objectively about the relative importance of Web pages. This paper describes PageRank, a method for rating Web pages objectively and mechanically, effectively measuring the human interest and attention devoted to them.

We compare PageRank to an idealized random Web surfer. We show how to efficiently compute PageRank for large numbers of pages. And, we show how to apply PageRank to search and to user navigation.

1 Introduction and Motivation

The World Wide Web creates many new challenges for information retrieval. It is very large and heterogeneous. Current estimates are that there are over 150 million web pages with a doubling life of less than one year. More importantly, the web pages are extremely diverse, ranging from "What is Joe having for lunch today?" to journals about information retrieval. In addition to these major challenges, search engines on the Web must also contend with inexperienced users and pages engineered to manipulate search engine ranking functions.

The now-incredibly-famous paper
Ranking algorithms

Wikipedia's Upstream Traffic Sources

hitwise
Real-time competitive intelligence

Chart of the weekly upstream % of 'www.wikipedia.org'.
Time periods represented with broken lines indicate insufficient data.
Generated on: 02/16/2007. Copyright 2007 (c) 'Hitwise Pty Ltd.'
Learning behavior

- Google begins tracking clicks in 2005
- MSN search claims neural network
- AOL Data Scandal
Where and why?

- Targeted Advertising
- Recommendations
- Search Results
- **Group Discovery**
- Filtering of Documents
- Theme Extraction
In Biology
Page grouping

1. Salsa Cycles 2008
   Frames, stems, components, accessories, and more for road and
   http://www.salsacycles.com/ [Yahoo!]

2. Latin Entertainment Events All O
   General resource for Salsa music and dance. Includes a city gu
   http://www.salsaweb.com/ [Yahoo!]

3. Salsa (dance) - Wikipedia, the free encyclopedia
   Salsa refers to a fusion of informal dance styles having roots in
   http://en.wikipedia.org/wiki/Salsa_(dance) [Yahoo!]

4. Just Salsa Magazine - Find: Salsa Music & Dance, Salsa History
   Online magazine dedicated to Latin music, dance, and culture
   http://www.justsalsa.com/ [Yahoo!]

5. Salsa music - Wikipedia, the free encyclopedia
   Salsa dancing - Musicians. Salsa music is a diverse and predom
Political science test
MSNBC - 1 hour ago
Presidential hopeful Hillary Clinton reaped a windfall of political contributions today in the wake of her victory in the Pennsylvania primary - and if her candidacy actually pays off, investors in Clinton's political fortunes could reap a windfall as...

Clinton Hopes for Resurgence After Pennsylvania Primary Win
Voice of America
Washington Post - Kansas City Star - International Herald Tribune - Boston Globe

Foster care could be wrenching for Texas sect children
The Associated Press - 1 hour ago
SAN ANGELO, Texas (AP) - Many of the children have seen little or no television. They have been essentially home-schooled all their lives.

Children From Sect Are Dispersed by State
New York Times
437 Kids Have DNA Testing, Will Be Placed in Temporary Foster Homes
ABC News
International Herald Tribune - Kansas City Star - MSNBC - Vancouver Sun

How Apple Is Preparing for an iPod Slump
New York Times - all 526 news articles »

Web 2.0: Microsoft's Live Mesh To Reinvent The Platform
InformationWeek - all 231 news articles »

Cowboys Get Jones in Trade, Not Knowing if He Can Return
New York Times - all 538 news articles »

American Idol 2008 'American Idol': Is David Cook Unstoppable?
Entertainment Weekly - all 271 news articles »

Boy or Girl? The Answer May Depend on Mom's Eating Habits
New York Times - all 585 news articles »

Miley "Memoirs" Really Worth Millions?
TMZ.com - all 588 news articles »

911 call released in fatal grizzly bear attack
San Jose Mercury News - all 1,050 news articles »
Where and why?

- Targeted Advertising
- Recommendations
- Search Results
- Group Discovery
- Filtering of Documents
- Theme Extraction
The obvious: spam

SpamBayes

Theory

Sharpen your pencils, this is the mathematical background (such as it is).

- The paper that started the ball rolling: Paul Graham's [A Plan for Spam](https://www.thespace.com/planforspam/).
- Gary Robinson has an [interesting essay](https://blog.grocery.com/2003/08/17/bspam/) suggesting some improvements to Graham's original approach.
- [Gary Robinson's Linux Journal article](https://www.linuxjournal.com/content/bayes-spam-filter) discussed using the chi squared distribution.
Other email uses
Web documents

“As you add information to Twine, it is automatically tagged so that you and others can find it more easily”
Where and why?

- Targeted Advertising
- Recommendations
- Search Results
- Group Discovery
- Filtering of Documents
- Theme Extraction
What is the buzz?

Techmeme

TOP ITEMS:

Am / MacRumors:
Aluminum Multi-Touch MacBook, New MacBook Pro, and More? [Updated] — A MacRumors reader ran across a very unusual Mac web page yesterday and was kind enough to forward it on. The page contained promotional images for Apple products, including a couple that are not yet shipping products.
Discussion: Infinite Loop, Engadget, Gadget Lab, Neowin.net and PalmAddicts

RELATED:
Aidan Malley / AppleInsider:
Apple orders aluminum 13-inch MacBook shells, 3G iPhone chips — Separate reports have revealed that Apple's next wave of products for 2008 is soon to enter various stages of production, with orders placed for aluminum shells for the company's next-generation 13-inch MacBooks and a manufacturer chosen for the iPhone's 3G chipset.
Discussion: ChurchGizmo

Larry Dignan / Between the Lines:
21 minutes ago
Google delivers: Maybe paid clicks weren't such a big deal — Google on Thursday allayed concerns about its paid click growth rate with first quarter earnings that topped Wall Street's low expectations. — Google reported first quarter net income of $1.31 billion, or $1.12 a share, on revenue of $5.19 billion.
Discussion: Tech Trader Daily, Associated Press, Epicenter, paidContent.org and HipMojo.com
Recently active topics in Apple

- Recently Active
- Latest Topics
- Unanswered

1-6 of 393 topics

⚠️ Apil.itunes has taken money from my account and I never ordered anything.
Last reply 6 days ago.
I don't buy music of the internet! I don't own an api itunes. Why are you taking money out of my bank account.... I would like to speak with a person about this.

- lucky lady reported this problem 3 months ago. Pepe joined lucky lady and 20 other people reporting this problem.

⚠️ No printer setup utility in Leopard?
Last reply 3 hours ago.
Trying to configure an Airport Extreme Base Station to share a printer. The instructions direct me to use the Printer Setup Utility, which I don't have. A short search at Apple support shows about ten trillion people in the same mess. They direct people to use the Fax and Printer setup in System Preferences, which doesn't seem to provide anyway to turn on printer sharing via the base station.

- stowboy reported this problem 3 months ago. Zach joined stowboy and 8 other people reporting this problem. It's tagged printer setup utility, printer sharing, and base station extreme.

⚠️ why can't I open printer setup utility to add a printer?
I cannot add a printer.
How do I add my printer when the + doesn't work - no printer option to select. When I attempt to open printer setup utility, it flashes in the toolbar but doesn't open and under no circumstances will it open.

- 0 Replies
- 1 Follower

Customer Community
Where and why?

Methods and examples
Methods and Examples

- Bayesian Filtering
- Distance Metrics
- Clustering
- Decision Trees
- Network Analysis
- Feature Extraction
Bayesian Filtering
Bayesian Filtering
Bayesian Filtering
Bayesian Filtering
Bayesian Filtering
Bayesian Filtering

- School
- Work
- Algorithm
Bayesian Filtering

- school
- work
- algorithm

- viagra
- trades
- associate
Craigslist personals

I don't ask for much - 23

Hi there, to give you a little background I'm a fairly normal, quirky, fun, baseball-appreciating lady. I'm into bingo and reading books about social issues. No, I'm not 80, just a 23-year-old old soul type. I'm not high. However, that does not mean I'm ok with being disrespected and unappreciated.

I just want to meet a normal guy, and by normal I mean someone who doesn't drop acid to the point where the first time he meets me or auto-assumes we're dating after ONE meeting. Seriously, guys?
Analysis

Five Cities

W4M Personal Ads
<table>
<thead>
<tr>
<th>New York</th>
<th>Boston</th>
<th>Chicago</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mets</td>
<td>Pink</td>
<td>Cubs</td>
</tr>
<tr>
<td>Lounges</td>
<td>Sox</td>
<td>Burbs</td>
</tr>
<tr>
<td>Offense</td>
<td>Poetry</td>
<td>Bears</td>
</tr>
<tr>
<td>Desires</td>
<td>Intellectually</td>
<td>Girlie</td>
</tr>
<tr>
<td>Musical</td>
<td>Punk</td>
<td>Insecure</td>
</tr>
<tr>
<td>Submissive</td>
<td>Appreciation</td>
<td>Cheat</td>
</tr>
<tr>
<td>Create</td>
<td>Exercise</td>
<td>Importance</td>
</tr>
<tr>
<td>Song</td>
<td>Winter</td>
<td>Blunt</td>
</tr>
<tr>
<td>Oral</td>
<td>Education</td>
<td>Mouth</td>
</tr>
<tr>
<td>Los Angeles</td>
<td>San Francisco</td>
<td></td>
</tr>
<tr>
<td>------------------</td>
<td>-------------------</td>
<td></td>
</tr>
<tr>
<td>Excellent</td>
<td>Tee</td>
<td></td>
</tr>
<tr>
<td>Vegas</td>
<td>Employment</td>
<td></td>
</tr>
<tr>
<td>Meaningful</td>
<td>Picnic</td>
<td></td>
</tr>
<tr>
<td>Star</td>
<td>STD</td>
<td></td>
</tr>
<tr>
<td>Lame</td>
<td>Tasting</td>
<td></td>
</tr>
<tr>
<td>Industry</td>
<td>Hikes</td>
<td></td>
</tr>
<tr>
<td>Heat</td>
<td>French</td>
<td></td>
</tr>
<tr>
<td>Fitness</td>
<td>.com</td>
<td></td>
</tr>
<tr>
<td>Entertainment</td>
<td>Kayaking</td>
<td></td>
</tr>
<tr>
<td>Latino</td>
<td>Cycling</td>
<td></td>
</tr>
</tbody>
</table>
Methods and Examples

- Bayesian Filtering
- **Distance Metrics**
- Clustering
- Decision Trees
- Network Analysis
- Feature Extraction
Preference distance

Sarah Marshall  3   2   1   2
Leatherheads   3   3   5   5
Preference distance
Preference distance

1
2
3
4
5

1
2
3
4
5

2.23
For recommendations

Prom Night: 5

Prom Night: 2

Distance: 2.23
For recommendations

Prom Night: 5

4.1

Prom Night: 2
Linguistic distance

The Six Degrees Hypothesis Experienced
It is when you travel that you most powerfully experience the reality of Stanley Milgram's hypothesis, the idea that we're only six degrees away from anyone else.

You'd think that with billions of people in the world, the chance of you running into someone you know would be tiny, yet in my travels, it's happened to me repeatedly.

I'm on holiday in Sicily. I'm walking down the streets of Ortygia, the old town of Siracusa. I meet a man who's a local guide. He says, "My god, it's Tim O'Reilly." It's Kevin Allofs, whom I've never met, but who helps track at O'Reilly. He too is on holiday.

That night, at dinner, at a fabulous restaurant called Don Camillo, where we've gone in search of local cuisine, we meet a large group of the next table, all speaking English, but with a strong Italian accent. We ask for a bit of help with the menu. We discover that this is a group of food industry executives on a tour with Boulder-based Culinary Adventures.
Linguistic distance

The Six Degrees Hypothesis Experienced It Is When You Travel

Six Degrees Hypothesis Experienced Travel

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Six</td>
<td>3</td>
</tr>
<tr>
<td>Degrees</td>
<td>3</td>
</tr>
<tr>
<td>Hypothesis</td>
<td>1</td>
</tr>
<tr>
<td>Experienced</td>
<td>5</td>
</tr>
<tr>
<td>Travel</td>
<td>6</td>
</tr>
</tbody>
</table>
## Linguistic distance

<table>
<thead>
<tr>
<th></th>
<th>“china”</th>
<th>“kids”</th>
<th>“music”</th>
<th>“travel”</th>
<th>“yahoo”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gothamist</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>GigaOM</td>
<td>6</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>QuickOnlineTips</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>O’Reilly Radar</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>6</td>
<td>4</td>
</tr>
</tbody>
</table>
## Linguistic distance

<table>
<thead>
<tr>
<th></th>
<th>“china”</th>
<th>“kids”</th>
<th>“music”</th>
<th>“yahoo”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gothamist</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>GigaOM</td>
<td>6</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Quick Online Tips</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>12</td>
</tr>
</tbody>
</table>

Euclidean “as the crow flies”

\[
\sqrt{(6 - 0)^2 + (0 - 2)^2 + (1 - 2)^2 + (2 - 12)^2}
\]

= 12 (approx)
Article/blog similarity

Valleywag - Huffington > Slashdot - Wired
Methods and Examples

- Bayesian Filtering
- Distance Metrics
- **Clustering**
- Decision Trees
- Network Analysis
- Feature Extraction
Hierarchical Clustering
Hierarchical Clustering
Hierarchical Clustering
Grouping bloggers
Grouping bloggers
Grouping bloggers
Grouping articles

- India Not Out of the Nuclear Wilderness - ABC News
  - Bush signs U.S.-India nuclear deal
  - Bush Signs Nuclear Deal With India
  - U.S. envoy says no progress in N. Korea talks so far
  - US envoy says no progress in N. Korea talks so far - Scotsman
  - North Korea Insists on Nuclear Status
  - North Korea Declares Itself Nuclear Power

- Russia cautious on quick UN vote on Iran
  - From Russia with lies
    - Enriched uranium from Germany arrives in Russia
    - Uranium returned to Russia from Germany - People’s Daily Online
    - California seeks to improve lethal injection method
    - Lethal Injection Under Microscope - Forbes
Methods and Examples

- Bayesian Filtering
- Distance Metrics
- Clustering
- **Decision Trees**
- Network Analysis
- Feature Extraction
Decision Trees

- **color = green?**
  - **No**
  - **color = red?**
    - **No**
      - **shape = round?**
        - **No**
          - Banana
        - **Yes**
          - **diameter > 4in?**
            - **No**
              - Lemon
            - **Yes**
              - Grapefruit
  - **Yes**
    - **diameter > 2in?**
      - **No**
        - **stone = true?**
          - **No**
            - Grape
          - **Yes**
            - Cherry
      - **Yes**
        - **diameter > 6in?**
          - **No**
            - Grape
          - **Yes**
            - Apple

- **diameter > 2in?**
  - **No**
    - **Grape**
  - **Yes**
    - **Apple**

-Watermelon
## CART Algorithm

<table>
<thead>
<tr>
<th>Brand</th>
<th>Type</th>
<th>Life (hrs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duracell</td>
<td>C</td>
<td>4</td>
</tr>
<tr>
<td>Energizer</td>
<td>C</td>
<td>5</td>
</tr>
<tr>
<td>Duracell</td>
<td>AA</td>
<td>2</td>
</tr>
<tr>
<td>Energizer</td>
<td>AA</td>
<td>2.5</td>
</tr>
</tbody>
</table>

From any dataset...
## CART Algorithm

<table>
<thead>
<tr>
<th>Brand</th>
<th>Type</th>
<th>Life (hrs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duracell</td>
<td>C</td>
<td>4</td>
</tr>
<tr>
<td>Energizer</td>
<td>C</td>
<td>5</td>
</tr>
<tr>
<td>Duracell</td>
<td>AA</td>
<td>2</td>
</tr>
<tr>
<td>Energizer</td>
<td>AA</td>
<td>2.2</td>
</tr>
</tbody>
</table>

... find the best split ...

Type is C?

- No
  - Avg=2.1
- Yes
  - Avg=4.5
## CART Algorithm

<table>
<thead>
<tr>
<th>Brand</th>
<th>Type</th>
<th>Life (hrs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duracell</td>
<td>C</td>
<td>4</td>
</tr>
<tr>
<td>Energizer</td>
<td>C</td>
<td>5</td>
</tr>
<tr>
<td>Duracell</td>
<td>AA</td>
<td>2</td>
</tr>
<tr>
<td>Energizer</td>
<td>AA</td>
<td>2.2</td>
</tr>
</tbody>
</table>

**Type is C?**

- No  
  - Duracell 2.2

- Yes  
  - Duracell 2

... and repeat.
Hot or Not

Age > 22?

New England?

South?

Male?

Not!

Hot!

Hot!

Not!
Methods and Examples

- Bayesian Filtering
- Distance Metrics
- Clustering
- Decision Trees
- **Network Analysis**
- Feature Extraction
A network
PageRank
PageRank

D = 0.15 + 0.85 * E/1 + 0.85 * F/2 + 0.85 * B/1 = 2.275
PageRank

A 0.58
B 0.58
c 1.0
d 2.275
E 1.0
F 0.15
PageRank

A

0.58

B

0.58

C

2.08

D

1.56

E

0.3

F

0.15
PageRank

A → B → C → D → E → F

1.03 1.48 1.56 0.3 0.15
PageRank

A  0.78
B  0.78
C  1.48
D  1.34
E  0.3
F  0.15
ThoughtRank

The popularity-weighted popularity contest

(data was seeded from Crowdsink and Facebook; changes will be stored in Freebase.
ThoughtRank is based on PageRank, the original popularity-weighted popularity contest)

Question: What are the best emerging ideas in Collective Intelligence?

<table>
<thead>
<tr>
<th>Score</th>
<th>Name</th>
<th>Inbound</th>
<th>Outbound</th>
<th>Answer:</th>
</tr>
</thead>
<tbody>
<tr>
<td>130</td>
<td>Greg Linden</td>
<td>13</td>
<td>15</td>
<td>personalization, recommender systems, collaborative filtering, personalized search, personalized advertising, learning to rank from clickstream data in search, collaborative rating and editing by communities to correct and improve automatically extracted data.</td>
</tr>
<tr>
<td>106</td>
<td>David Pennock</td>
<td>14</td>
<td>27</td>
<td>prediction markets, recommender systems, wisdom of crowds, algorithmic economics</td>
</tr>
<tr>
<td>106</td>
<td>Tim O'Reilly</td>
<td>20</td>
<td>6</td>
<td>Hidden semantics (e.g. Google's realization via PageRank that a link is a vote, Wesabe's realization that how people spend money is a vote); systems for generating implicit semantics (e.g. Facebook and other social networks causing people to express their social graph); statistical methods for extracting these semantics; Mechanical Turk style systems like ReCaptcha; sensor-based systems for generating additional semantics (e.g location, traffic); real time feedback loops in applications (e.g. Adwords real time auction); humans as the last mile in algorithmic intelligence (i.e. Google gives ten results, not one, unless you're feeling lucky); &quot;creating a context where other people can share&quot; (e.g. Adrian Holovaty's work at the Lawrence Journal World building participatory sites for local reportage)</td>
</tr>
</tbody>
</table>

Unvote for this answer
The paper attempts to provide an alternative method for measuring the importance of scientific papers based on the Google's PageRank. The method is a meaningful extension of the common integer counting of citations and is then experimented for bringing PageRank to the citation analysis in a large citation network. It offers a more integrated picture of the publications' influence in a specific field.
Clustering coefficient

“How many of each persons friends are friends with each other?”
Clustering coefficient

Low clustering coefficient
Clustering coefficient

High clustering coefficient
“small world graph”
Shared board members of America's big companies
Shared board members of America’s big companies
Twitter!
Twitter!
Methods and Examples

- Bayesian Filtering
- Distance Metrics
- Clustering
- Decision Trees
- Network Analysis
- Feature Extraction
Independent Features

NMF

Original

\[ \text{NMF} \times \text{Original} = \]
alt.support.diet

How accurate are the calorie meters on gym equipment?

3 messages - Collapse all

McAlisters  View profile

I use two different treadmills, one at home, the other at the YMCA. Both have the calories burned counters, the treadmill at the gym has imputs for age and weight, I'd assume for the calorie used math. I'm 58 years of age and weigh 205 and the machine tells me I burn about about 800 calories burned during an hours' walk of 4.7 mph with an incline of 4.0 degrees. Does that sound about right for the calories burned?

Reply  Reply to author  Forward  Rate this post: ★★★★★☆
Matrix Factorization

**Features Matrix**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Msg1</th>
<th>Msg2</th>
<th>Msg3</th>
<th>Msg4</th>
<th>Msg5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gym</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calorie</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weigh</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carbs</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treadmill</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Weight Matrix**

<table>
<thead>
<tr>
<th></th>
<th>Msg1</th>
<th>M2</th>
<th>M3</th>
<th>M4</th>
<th>M5</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>F2</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>F3</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Current Guess**

<table>
<thead>
<tr>
<th></th>
<th>Msg1</th>
<th>Msg2</th>
<th>Msg3</th>
<th>Msg4</th>
<th>Msg5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gym</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Calorie</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Weigh</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Carbs</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Treadmill</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>
Matrix Factorization

Features Matrix

<table>
<thead>
<tr>
<th>Feature</th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gym</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Calorie</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Weigh</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Carbs</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Treadmill</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

Weight Matrix

<table>
<thead>
<tr>
<th>Feature</th>
<th>Msg1</th>
<th>Msg2</th>
<th>Msg3</th>
<th>Msg4</th>
<th>Msg5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gym</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Calorie</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Weigh</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Carbs</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Treadmill</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Current Guess

<table>
<thead>
<tr>
<th>Feature</th>
<th>Msg1</th>
<th>Msg2</th>
<th>Msg3</th>
<th>Msg4</th>
<th>Msg5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gym</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Calorie</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Weigh</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Carbs</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Treadmill</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

Target Result

<table>
<thead>
<tr>
<th>Feature</th>
<th>Msg1</th>
<th>Msg2</th>
<th>Msg3</th>
<th>Msg4</th>
<th>Msg5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gym</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Calorie</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Weigh</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Carbs</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Treadmill</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>
Matrix Factorization

Features Matrix

<table>
<thead>
<tr>
<th></th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gym</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Calorie</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Weigh</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Carbs</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Treadmill</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Weight Matrix

<table>
<thead>
<tr>
<th></th>
<th>Msg1</th>
<th>M2</th>
<th>M3</th>
<th>M4</th>
<th>M5</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>F2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>F3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Current Guess

<table>
<thead>
<tr>
<th></th>
<th>Msg1</th>
<th>Msg2</th>
<th>Msg3</th>
<th>Msg4</th>
<th>Msg5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gym</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Calorie</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Weigh</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Carbs</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Treadmill</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

Target Result

<table>
<thead>
<tr>
<th></th>
<th>Msg1</th>
<th>Msg2</th>
<th>Msg3</th>
<th>Msg4</th>
<th>Msg5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gym</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Calorie</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Weigh</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Carbs</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Treadmill</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>
Interpreting Features

Features Matrix

<table>
<thead>
<tr>
<th></th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gym</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Calorie</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Weigh</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Carbs</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Treadmill</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Weight Matrix

<table>
<thead>
<tr>
<th></th>
<th>Msg1</th>
<th>M2</th>
<th>M3</th>
<th>M4</th>
<th>M5</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>F2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>F3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Theme Matrix

<table>
<thead>
<tr>
<th>Theme 1</th>
<th>Theme 2</th>
<th>Theme 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gym</td>
<td>Calorie</td>
<td>Weigh</td>
</tr>
<tr>
<td>Treadmill</td>
<td>Carbs</td>
<td>Calorie</td>
</tr>
</tbody>
</table>

Theme Matrix

<table>
<thead>
<tr>
<th>Theme 1</th>
<th>Theme 2</th>
<th>Theme 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Msg1</td>
<td>Msg2</td>
<td>Msg3 etc.</td>
</tr>
<tr>
<td>Theme 1</td>
<td>Theme 2</td>
<td>Theme 3</td>
</tr>
<tr>
<td>Theme 3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Diet & Body themes

- Atkins
- Induction
- South Beach
- Carbs
- Chocolate
- Black
- Coffee
- Olive
- Broccoli
- Calories
- Weight
- Fats
- Protein
- Cholesterol
- Gym
- Weights
- Exercise
- Running
- Injured
- Money
- Organic
- Want
- Best
- Cook
- Recipe
- Fried
- Home
she  series  league  olympics  university
her  television  major  competed  professor
after  show  baseball  won  received
when  which  season  summer  science
father  radio  played  medal  research
women  bbc  with  athlete  born
We’re just getting started...
Collective Intelligence

Homepage  http://kiwitobes.com

Freebase  http://freebase.com
Questions?