Learning Objective

“Know what you know”
Search use …
Without search engines the web wouldn’t scale

- Search turned out to be the best mechanism for advertising on the web, a $15 billion plus industry.
- Growing very fast (entire US advertising industry is $250 billion though)
- Sponsored search marketing is about $10 billion
Google has maintained that ads (based on vendors bidding for search queries) do not affect vendors ranking in search results.
Size of the Web

How big is the web?

- Netcraft Web Server Survey
Rate of change

- Fetterly et al. study in 2002
- 150 million pages over 11 weekly crawls
- Bucketed into 85 groups according to amount of change
Top queries

• Most are related to sex
• 2008 Who What How (Google)

Who is...
1. who is obama
2. who is mccain
3. who is palin
4. who is lil wayne
5. who is miley cyrus
6. who is dolla
7. who is jonas brothers
8. who is chris brown
9. who is biden
10. who is martin luther

What is...
1. what is love
2. what is life
3. what is java
4. what is sap
5. what is rss
6. what is scientology
7. what is autism
8. what is lupus
9. what is 3g
10. what is art

How to...
1. how to draw
2. how to kiss
3. how to write
4. how to cook
5. how to tie
6. how to hack
7. how to run
8. how to cite
9. how to paint
10. how to spell

Spam Industry

**Advanced Traffic:**
Get a first page listing on Google - GUARANTEED! For maximum search engine traffic - the best of SEO and search advertising. Visitors in just 48 hours from $7/day. Discover the traffic potential!

**WARNING:** This site contains sneaky, underhanded Black Hat Seo tactics.
Black Hat Seo is responsible for more online fortunes than you’d care to imagine but it’s NOT for everybody.

Make Money Blogging
See How I Earn Over Six Figures a year Blogging

I Will Get Your Website to the Top of Google!

The art of search engine optimization...gaining top spots on Google...is no easy chore. I know...this is my job...

I assist people in getting top positions for their websites on Google, Yahoo, MSN and all the other major search engines.

There are a few given on the internet when it comes to trying to market goods and services:

**No Traffic=No Sales!**

End of story...that’s it...bottom line!

If you have a website...
Crawling the web

Introduction

Seed Pages

Crawled Pages

URL Frontier

The Rest of the Web

Web Spider
Politeness?

Statistics for: djp3.net

Summary
- When:
  - Monthly history
  - Days of month
  - Days of week
  - Hours
- Who:
  - Countries
  - Hosts
  - Robots/Spiders visitors
- Navigation:
  - Visits duration
  - File type
  - Viewed
  - Operating Systems
  - Browsers
  - Referers:
    - Origin
    - Search

Reported period: [ - Year - ] 2007

Last Update: 14 Jan 2008 - 02:59

Robots/Spiders visitors

<table>
<thead>
<tr>
<th></th>
<th>Hits</th>
<th>Bandwidth</th>
<th>Last visit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Googlebot</td>
<td>1393868+104</td>
<td>5.11 GB</td>
<td>31 Dec 2007 - 23:50</td>
</tr>
<tr>
<td>Inktomi Slurp</td>
<td>36668+221</td>
<td>554.25 MB</td>
<td>31 Dec 2007 - 23:55</td>
</tr>
<tr>
<td>MSNBot</td>
<td>19522+2</td>
<td>699.90 MB</td>
<td>28 Dec 2007 - 08:01</td>
</tr>
<tr>
<td>Unknown robot (identified by 'crawl')</td>
<td>15949+13</td>
<td>89.34 MB</td>
<td>31 Dec 2007 - 22:24</td>
</tr>
<tr>
<td>AskJeeves</td>
<td>7016+1</td>
<td>106.29 MB</td>
<td>31 Dec 2007 - 23:49</td>
</tr>
<tr>
<td>Google AdSense</td>
<td>2701</td>
<td>100.26 MB</td>
<td>31 Dec 2007 - 22:10</td>
</tr>
<tr>
<td>psbot</td>
<td>2268+1</td>
<td>80.48 MB</td>
<td>31 Dec 2007 - 09:59</td>
</tr>
<tr>
<td>Unknown robot (identified by 'robot')</td>
<td>930+1</td>
<td>19.10 MB</td>
<td>31 Dec 2007 - 09:34</td>
</tr>
<tr>
<td>Turn It In</td>
<td>350+1</td>
<td>6.32 MB</td>
<td>03 Sep 2007 - 15:44</td>
</tr>
<tr>
<td>BaiDuSpider</td>
<td>300</td>
<td>10.22 MB</td>
<td>26 Nov 2007 - 07:32</td>
</tr>
<tr>
<td>GigaBot</td>
<td>243</td>
<td>5.27 MB</td>
<td>30 Dec 2007 - 05:06</td>
</tr>
<tr>
<td>Scooter</td>
<td>90+3</td>
<td>288.75 KB</td>
<td>27 Nov 2007 - 14:30</td>
</tr>
<tr>
<td>PhpDig</td>
<td>91</td>
<td>2.28 MB</td>
<td>21 Oct 2007 - 09:51</td>
</tr>
<tr>
<td>WISENutbot</td>
<td>76</td>
<td>1.94 MB</td>
<td>13 Jan 2007 - 14:04</td>
</tr>
<tr>
<td>Magpie</td>
<td>25</td>
<td>43.48 MB</td>
<td>24 Dec 2007 - 00:51</td>
</tr>
<tr>
<td>Unknown robot (identified by hit on 'robots.txt')</td>
<td>0+16</td>
<td>4.38 KB</td>
<td>14 Nov 2007 - 03:43</td>
</tr>
<tr>
<td>EchO!</td>
<td>14</td>
<td>287.09 KB</td>
<td>27 Dec 2007 - 13:56</td>
</tr>
<tr>
<td>Internet Shinchakubin</td>
<td>13</td>
<td>385.03 KB</td>
<td>27 Nov 2007 - 15:23</td>
</tr>
<tr>
<td>BBot</td>
<td>10</td>
<td>146.35 KB</td>
<td>13 Jun 2007 - 15:17</td>
</tr>
<tr>
<td>arks</td>
<td>8</td>
<td>142.24 KB</td>
<td>27 Nov 2007 - 12:25</td>
</tr>
<tr>
<td>MSIECrawler</td>
<td>8</td>
<td>263.02 KB</td>
<td>26 Dec 2007 - 11:16</td>
</tr>
<tr>
<td>The Python Robot</td>
<td>5</td>
<td>122.91 KB</td>
<td>23 Nov 2007 - 08:01</td>
</tr>
</tbody>
</table>
Robots.txt Example

- http://www.ics.uci.edu/robots.txt

```text
User-agent: MOMspider # The Multi-Owner Maintenance Spider
Disallow: /cgi-bin/ # Script files
Disallow: /Admin/MOM/ # Local MOMspider output
Disallow: ~/fielding/MOM/ # Local MOMspider output
Disallow: /TR/ # Dienst Technical Report Server
Disallow: /MetaServer/ # Dienst Technical Report Server
Disallow: ~/eppstein/pubs/cites/ # Eppstein Database
Disallow: ~/fiorello/pvt/ # Private pages

User-agent: * # All other spiders should avoid
Disallow: /cgi-bin/ # Script files
Disallow: /Test/ # The test area for web experimentation
Disallow: /Admin/ # Huge server statistic logs
Disallow: /TR/ # Dienst Technical Report Server
Disallow: /MetaServer/ # Dienst Technical Report Server
Disallow: ~/fielding/MOM/ # Local MOMspider output
Disallow: ~/kanderso/hidden # Ken Anderson's stuff
Disallow: ~/eppstein/pubs/cites/ # Eppstein Database
Disallow: ~/fiorello/pvt/ # Private pages
Disallow: ~/dean/
Disallow: ~/wwwoffic/
Disallow: ~/ucounsel/
Disallow: ~/sao/
Disallow: ~/support/
Disallow: ~/icsdb/
Disallow: ~/bin/
```
A Robust Crawl Architecture

WWW → DNS → Fetch → Parse → Seen? → URL Filter → Duplicate Elimination → URL Index

Doc. Fingerprints → Robots.txt

URL Frontier Queue
What really happens

**DNS**

- **Name Server**
  - DNS Cache
  - OS specified DNS Server: ns1.ics.uci.edu

- **Client**
  - Host table
  - OS DNS Cache
  - OS DNS Resolver
  - Firefox DNS cache

- **The User**

```
Give me a www.djp3.net
```

```
fllickr:crankyT
```
Web Crawlers
- URLs flow from top to bottom
- Front queues manage priority
- Back queue manage politeness
- Each queue is FIFO
BSBI - Block sort-based indexing

Different way to sort index

Block

(Every,www.cnn.com)
(Her,news.google.com)
(I'm,news.bbc.co.uk)

Block

(1998,news.google.com)
(Her,news.bbc.co.uk)
(I,www.cnn.com)
(Jensen's,www.cnn.com)

Merged Postings

(1998,news.google.com)
(Every,www.cnn.com)
(Her,news.bbc.co.uk)
(Her,news.google.com)
(I,www.cnn.com)
(I'm,news.bbc.co.uk)
(Jensen's,www.cnn.com)

Disk
The index has a list of **vector space models**

**Letter from dead sister haunts brothers**

Every time Julie Jensen’s brothers hear the letter read, it brings everything back. Most of all, they wonder if they could have saved her. Her husband now stands trial for allegedly killing her. "I pray I’m wrong + nothing happens," Julie wrote days before her 1998 death. **full story**
Our inverted index is a 2-D array or Matrix

A Row for Each Word (or "Term")

<table>
<thead>
<tr>
<th></th>
<th>Anthony and Cleopatra</th>
<th>Julius Caesar</th>
<th>The Tempest</th>
<th>Hamlet</th>
<th>Othello</th>
<th>Macbeth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anthony</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Brutus</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Caesar</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Calpurnia</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cleopatra</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>mercy</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>worser</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Querying

- Parametric Search

- Example:
  - Result is a large table
  - Columns are fields
  - Searching for “2005” only applied to year field
Querying

- **Parametric Search**

- **Example:**
  - Result is a large table
  - Columns are fields
  - Searching for “2005” only applied to year field

<table>
<thead>
<tr>
<th>Year</th>
<th>Make/Model</th>
<th>Miles</th>
<th>Price</th>
<th>Photos</th>
<th>Body Style</th>
<th>Color</th>
<th>Distance</th>
<th>Dealer</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>Ferrari 430 Berlinetta</td>
<td>1,030</td>
<td>$249,900</td>
<td></td>
<td>2 Door Coupe</td>
<td>CORSO RED</td>
<td>28 Miles</td>
<td>FleetRatescomNewUsed</td>
</tr>
<tr>
<td>2005</td>
<td>Ferrari 575 Superamerica Co</td>
<td>4,200</td>
<td>$285,000</td>
<td></td>
<td>Convertible</td>
<td>Silver</td>
<td>65 Miles</td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td>Ferrari 430 Spider Convertible</td>
<td>3,500</td>
<td>$249,500</td>
<td></td>
<td>Convertible</td>
<td>Rosso Corsa</td>
<td>65 Miles</td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td>Ferrari 430 Spider Convertible</td>
<td>2,900</td>
<td>$249,000</td>
<td></td>
<td>Convertible</td>
<td>YELLOW</td>
<td>65 Miles</td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td>Ferrari 430 Spider Convertible</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td>Ferrari 430 Spider Convertible</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td>Ferrari 430 Spider Convertible</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td>Ferrari 430 Spider Convertible</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td>Ferrari 430 Spider Convertible</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td>Ferrari 430 Spider Convertible</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Querying

- **Parametric Search**

- Example:
  - Result is a large table
  - Columns are fields
  - Searching for “2005” only applied to year field

<table>
<thead>
<tr>
<th>Save</th>
<th>Year</th>
<th>Make/Model</th>
<th>Miles</th>
<th>Price</th>
<th>Photos</th>
<th>Body Style</th>
<th>Color</th>
<th>Distance</th>
<th>Dealer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2005</td>
<td>Ferrari 430 Berlinetta</td>
<td>1,030</td>
<td>$249,900</td>
<td></td>
<td>2 Door Coupe</td>
<td>CORSO RED</td>
<td>28 Miles</td>
<td>FleetRatescomNewUsed</td>
</tr>
<tr>
<td></td>
<td>2005</td>
<td>Ferrari 575 Superamerica Co</td>
<td>4,200</td>
<td>$285,000</td>
<td></td>
<td>Convertible</td>
<td>Silver</td>
<td>65 Miles</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2005</td>
<td>Ferrari 430 Spider Convert</td>
<td>3,500</td>
<td>$249,500</td>
<td></td>
<td>Convertible</td>
<td>Rosso Corsa</td>
<td>65 Miles</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2005</td>
<td>Ferrari 430 Spider Convert</td>
<td>2,900</td>
<td>$249,000</td>
<td></td>
<td>Convertible</td>
<td>YELLOW</td>
<td>65 Miles</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2005</td>
<td>Ferrari 430 Spider Convert</td>
<td>3,945</td>
<td>$239,500</td>
<td></td>
<td>Convertible</td>
<td>BLACK</td>
<td>65 Miles</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2005</td>
<td>Ferrari 430 Spider Convert</td>
<td>1,500</td>
<td>$219,500</td>
<td></td>
<td>2 Door Coupe</td>
<td>Grigio Alloy</td>
<td>65 Miles</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2005</td>
<td>Ferrari 430 Spider Convert</td>
<td>4,500</td>
<td>$219,000</td>
<td></td>
<td>Convertible</td>
<td>RED</td>
<td>65 Miles</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2005</td>
<td>Ferrari 360 Spider F1 Conve</td>
<td>4,000</td>
<td>$219,000</td>
<td></td>
<td>Convertible</td>
<td>Black</td>
<td>65 Miles</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2005</td>
<td>Ferrari 430 Spider Convert</td>
<td>10,317</td>
<td>$209,999</td>
<td></td>
<td>Convertible</td>
<td>Red</td>
<td>28 Miles</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2005</td>
<td>Ferrari 430 Spider Convert</td>
<td>29,000</td>
<td>$205,000</td>
<td></td>
<td>Convertible</td>
<td>RED</td>
<td>65 Miles</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2005</td>
<td>Ferrari 430 F1 Coupe</td>
<td>5,300</td>
<td>$199,000</td>
<td></td>
<td>2 Door Coupe</td>
<td>BLACK</td>
<td>65 Miles</td>
<td></td>
</tr>
</tbody>
</table>
Now, we crawl the corpus
We parse the document keeping track of terms, fields and docIDs
Instead of building just a (term, docID) pair
We build (term, field, docID) triples
These can then be combined into postings like this:

<table>
<thead>
<tr>
<th>Field</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>16</th>
<th>32</th>
<th>64</th>
</tr>
</thead>
<tbody>
<tr>
<td>William.author</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>William.title</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>8</td>
<td>13</td>
</tr>
<tr>
<td>William.abstract</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>7</td>
<td>9</td>
<td>11</td>
</tr>
</tbody>
</table>
Zone scoring with zones combination index

“bill OR rights” (0.1 author), (0.3 body), (0.6 title)

1: 0.4  5: 0.9
2: 0.4  8: 0.9
3: 0.9  9: 0.9
Bag of Words Model

• “Don fears the mole man” equals “The mole man fears Don”
• The incidence matrix for both looks the same

Don fears the mole man
The mole man fears Don
Bag of Words Model

• “Don fears the mole man” equals “The mole man fears Don”

• The incidence matrix for both looks the same
Querying

Bag of Words Model

- “Don fears the mole man” equals “The mole man fears Don”
- The incidence matrix for both looks the same

\[
\begin{array}{c}
\text{Don fears the mole man} \\
\text{the} & \text{man} & \text{mole} & \text{fears} & \text{Don} \\
\hline
\text{The mole man fears Don} \\
\text{the} & \text{man} & \text{mole} & \text{fears} & \text{Don} \\
\end{array}
\]

\[
\begin{array}{c|c|c}
\text{Don} & d_1 & d_2 \\
\text{fears} & 1 & 1 \\
\text{man} & 1 & 1 \\
\text{mole} & 1 & 1 \\
\text{mule} & 0 & 0 \\
\text{the} & 1 & 1 \\
\text{zoo} & 0 & 0 \\
\end{array}
\]
Weighting Term Frequency - WTF

$$Score_{WTF}(q, d) = \sum_{t \in q} (WTF(t, d))$$
Weighting Term Frequency - WTF

\[ \text{Score}_{WTF}(q, d) = \sum_{t \in q} (WTF(t, d)) \]

\[ \text{Score}_{WTF}("bill rights", \text{declarationOfIndependence}) = \]
\[ \text{WTF}("bill", \text{declarationOfIndependence}) + \]
\[ \text{WTF}("rights", \text{declarationOfIndependence}) = \]
\[ 0 + 1 + \log(3) = 1.48 \]
Querying

Weighting Term Frequency - WTF

\[ Score_{WTF}(q, d) = \sum_{t \in q} (WTF(t, d)) \]

\[ Score_{WTF}("bill rights", \text{declarationOfIndependence}) = \]
\[ WTF("bill", \text{declarationOfIndependence}) + \]
\[ WTF("rights", \text{declarationOfIndependence}) = \]
\[ 0 + 1 + \log(3) = 1.48 \]

\[ Score_{WTF}("bill rights", \text{constitution}) = \]
\[ WTF("bill", \text{constitution}) + \]
\[ WTF("rights", \text{constitution}) = \]
\[ 1 + \log(10) + 1 + \log(1) = 3 \]
Vector Space Scoring

### Vector Space Model

- Recall our Shakespeare Example:

<table>
<thead>
<tr>
<th></th>
<th>$\vec{V}(d_1)$</th>
<th>$\vec{V}(d_2)$</th>
<th>$\vec{V}(d_6)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antony and Cleopatra</td>
<td>13.1</td>
<td>11.4</td>
<td>0.0</td>
</tr>
<tr>
<td>Brutus</td>
<td>3.0</td>
<td>8.3</td>
<td>1.0</td>
</tr>
<tr>
<td>Caesar</td>
<td>2.3</td>
<td>2.3</td>
<td>0.5</td>
</tr>
<tr>
<td>Calpurnia</td>
<td>0.0</td>
<td>11.2</td>
<td>0.0</td>
</tr>
<tr>
<td>Cleopatra</td>
<td>17.7</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>mercy</td>
<td>0.5</td>
<td>0.0</td>
<td>0.9</td>
</tr>
<tr>
<td>worser</td>
<td>1.2</td>
<td>0.0</td>
<td>0.3</td>
</tr>
<tr>
<td>The Tempest</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Hamlet</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Othello</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Macbeth</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>
Vector Space Scoring

Query as a vector

• So a query can also be plotted in the same space
  • “worser mercy”
• To score, we ask:
  • How similar are two points?
• How to answer?

• Hamlet
• Antony and Cleopatra
• Julius Caesar
• The Tempest
• Othello
• Macbeth
Markov Chains

- Example:
  - 8 states
  - (web pages or whatever)
  - 8 by 8 transition prob. matrix

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
<td>0</td>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>0</td>
<td>1.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>F</td>
<td>0</td>
<td>0.33</td>
<td>0</td>
<td>0.33</td>
<td>0</td>
<td>0</td>
<td>0.33</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>H</td>
<td>0</td>
<td>1.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Long-Term visit rate

- A: 5%
- B: 21%
- C: 23%
- D: 18%
- E: 8%
- F: 5%
- G: 9%
- H: 10%

[Diagram showing link analysis connections between A, B, C, D, E, F, G, and H]
Latent Semantic Indexing - Introduction

Star Cluster NGC 290 - ESA & NASA
Star Cluster NGC 290 - ESA & NASA

Latent Semantic Indexing - Introduction
Mathematically speaking

- Latent Semantic Indexing can project on an arbitrary axis, not just a principal axis
Matrix Decomposition

• Singular Value Decomposition

• SVD enables lossy compression of your term-document matrix

• reduces the dimensionality or the rank

• you can arbitrarily reduce the dimensionality by putting zeros in the bottom right of sigma

• this is a mathematically optimal way of reducing dimensions
Using MATLAB For LSA

Demo

• Demonstrate what SVD is capturing

• 1st concept (1st row of M)

First concept is selecting for baseball?
Finally ... I promised it would be hard

- 19 Lectures
- 7 Discussions
- 4 quizzes - 8 chapters - 6 (+2) papers
- 7 assignments
  - Built (building) a web search engine from scratch
  - Used cutting edge architecture (hadoop)
- 2 web pages - a trip to Google
- Hopefully had fun, were challenged and learned something...
- you can sleep when you are dead, until then coffee.
Congratulations!