Link Analysis and Web Search

Presented by:
Ilan Kadar
Outline

• Google PageRank
• Link Analysis in Modern Web Search
Google PageRank

• Let’s google Klara Kedem.

What do we REALLY know about PageRank?
Google PageRank

• Link analysis algorithm developed at Stanford University by Larry Page and Sergey Brin (PhD students) in 1996.

• PageRank paper and the initial prototype of the Google search engine was published in 1998.
## From Citation Analysis to PageRank

<table>
<thead>
<tr>
<th>Journal Title (AI)</th>
<th>Impact Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>J. Mach. Learn. Res.</td>
<td>5.952</td>
</tr>
<tr>
<td>IEEE T. Pattern Anal.</td>
<td>4.352</td>
</tr>
<tr>
<td>IEEE T. Evolut. Comp.</td>
<td>3.688</td>
</tr>
<tr>
<td>Artif. Intell.</td>
<td>3.570</td>
</tr>
<tr>
<td>Mach. Learn.</td>
<td>3.258</td>
</tr>
</tbody>
</table>
PageRank – Idea

• Generally, highly linked pages are more “important” than pages with few links.
• Back links coming from important pages convey more importance to a page.
PageRank - Definition

1. In a network with n nodes, we assign all nodes the same initial PageRank, set to be $1/n$.
2. We choose a number of steps $k$.
3. We then perform a sequence of $k$ updates to the PageRank values, using the following rule for each update.

**Basic PageRank Update Rule**

- Each page divides its current PageRank equally across its out-going links, and passes these equal shares to the pages it points to.
- Each page updates its new PageRank to be the sum of the shares it receives.
PageRank - Intuition

• PageRank is a kind of “fluid”
  – It circulates through the network.
  – Is passing from node to node across edges.
  – Is pooling at the nodes that are the most important.

• Total PageRank in the network remains constant
  – Why? Each page takes its PageRank, divides it up, and passes it along links.
  – PageRank is never created nor destroyed, just moved around from one node to another.

• No need to normalize PageRank of nodes to prevent them from growing.
PageRank - Example

Initial PageRank – 1/8 for all nodes

<table>
<thead>
<tr>
<th>Step</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>1</td>
<td>1/2</td>
<td>1/16</td>
<td>1/16</td>
<td>1/16</td>
<td>1/16</td>
<td>1/16</td>
<td>1/16</td>
<td>1/8</td>
</tr>
<tr>
<td>2</td>
<td>3/16</td>
<td>1/4</td>
<td>1/4</td>
<td>1/32</td>
<td>1/32</td>
<td>1/32</td>
<td>1/32</td>
<td>1/16</td>
</tr>
</tbody>
</table>
PageRank - Equilibrium Values

• PageRank values of all nodes converge to limiting values as the number of update steps \( k \) goes to infinity.

• If the network is strongly connected, then there is a unique set of equilibrium values.

• Interpretation of limit: by applying one step of the Basic PageRank Update Rule, the values at every node remain the same.
PageRank - Equilibrium Values
Problem with Basic Definition

- In many networks, the “wrong” nodes can end up with all the PageRank.
- “Slow leak”: small sets of nodes that can be reached from the rest of the graph, but have no paths back.
Solution to the Problem

- Remember the “fluid” intuition for PageRank
- Why all the water on earth doesn’t inexorably run downhill and reside exclusively at the lowest points?
- There’s a counter-balancing process at work
- Water also evaporates and gets rained back down at higher elevations.
Scaling the Definition of PageRank

1. Pick a scaling factor \( s \) between 0 and 1

2. Scaled PageRank Update Rule:
   - First apply the Basic PageRank Update Rule.
   - Scale down all PageRank values by a factor of \( s \) (total PageRank in the network has shrunk from 1 to \( s \)).
   - Divide the residual 1 – \( s \) units of PageRank equally over all nodes, giving \( (1 - s)/n \) to each

Why it works?

- “water cycle” that evaporates 1 – \( s \) units of PageRank in each step and rains it down uniformly across all nodes
The Limit of the Scaled PageRank Updated Rule

• Repeated application of the update rule converges to a set of limiting PageRank values as the number of updates $k$ goes to infinity.

• Unique equilibrium
  – But values depend on our choice of the scaling factor $s$.

• The version of PageRank used in practice, with a scaling factor $s$ between 0.8 and 0.9.
Random Walks: An Equivalent Definition

1. Start by choosing a page at random (each page with equal probability)
2. In each of the $k$ steps, pick a random outgoing link from the current page.
3. If the current page has no outgoing links, stay where you are.
Random Walks: An Equivalent Definition

1. Start by choosing a page at random (each page with equal probability)
2. In each of the $k$ steps, with probability $s$ pick a random out-going link from the current page.
3. With probability $1-s$, jump to a random node in the network with equal probability.
Outline

• Google PageRank

• Link Analysis in Modern Web Search
Link Analysis in Modern Web Search

- The link analysis has played an integral role in the ranking functions of Google, Yahoo, Bing and Ask.
- Link analysis ideas have been extended and generalized considerably.
- Search engine companies themselves are extremely secretive about what goes into their ranking functions.
Combining Links and Text

- Combine text and links for ranking is through the analysis of anchor text.

Ilan Kadar

Welcome to My Homepage

I am a Ph.D. student in the Computer Science Department and a member of the Computational Vision Group at Ben-Gurion University.

My main interests lie in the fields of Computer Vision, Computational Models of Biological Vision and Machine Learning.

My research involves developing algorithms for scene categorization and image retrieval with a multidisciplinary flavor, while employing insights from behavioral and psychophysical experiments.

My Advisor is Prof. Oded Ben-Shahar.
Click Through Statistics
The ESP Game – Human Computation

PLAYER 1

GUESSING: CAR

GUESSING: HAT

GUESSING: KID

SUCCESS!
YOU AGREE ON CAR

PLAYER 2

GUESSING: BOY

GUESSING: CAR

SUCCESS!
YOU AGREE ON CAR
The ESP Game – Human Computation
Welcome to **Google Image Labeler**, a new feature of Google Image Search that allows you to label random images to help improve the quality of Google's image search results.

### Begin as Playah

**Change your nickname**

### How does it work?

You'll be randomly paired with a partner who's online and using the feature. Over a 90-second period, you and your partner will be shown the same set of images and asked to provide as many labels as possible to describe each image you see. When your label matches your partner's label, you'll earn some points and move on to the next image until time runs out. After time expires, you can explore the images you've seen and the websites where those images were found. And we'll show you the points you've earned throughout the session.

For example, this image can be described by the labels: **bird**, **sky**, **soaring**, or **frigate bird**.

### What is required to participate?

Just an interest in helping Google improve the relevance of image search results for users like yourself. Although you do not have to log in to your Google account to help, logging in will allow you to keep track of your points. You can also choose to provide a nickname, or you can remain anonymous.
The ESP Game – Sample Labels

BEACH
CHAIRS
SEA
PEOPLE
MAN
WOMAN
PLANT
OCEAN
TALKING
WATER
PORCH
SKY
BELL

BUSH
GEORGE
PRESIDENT
MAN
DUMB
YUCK
A Moving Target

• You should always expect the world to react to what you do.

• The growth of *search engine optimization (SEO)* followed naturally once search became widespread application in the Web.

• *Web-page authors and their consultant become too effective at reverse engineering.*

• *Search engines are incredibly secretive about the internals of their ranking functions.*
Yossi Vardi’s billion dollar suggestion:

“Yossi invented for us the magic formula,” he said with a smile. “He told us to devote two thirds of the [internet] page to original results, and a third to advertisements, and that is what we did.”

- Sergey Brin
“The ultimate search engine would understand exactly what you mean and give back exactly what you want.”

- Larry Page
Exercise

• Are these correct equilibrium values for the Basic PageRank Update Rule?

![Diagram of a graph with nodes A, B, C, D, E, F and their connection probabilities]
Questions?