CS155b: E-Commerce

Lecture 16: April 10, 2001
WWW Searching and Google
WWW Digraph

- More than 1 Billion Nodes (Pages)
- Average Degree (links/Page) is 5-15. (Hard to Compute!)
- Massive, Distributed, Explicit Digraph (Not Like Call Graphs)
“Hot” Research Area

- Graph Representation
- Duplicate Elimination
- Clustering
- Ranking Query Results

http://theory.stanford.edu/~focs98/tutorials.html
A. Broder & M. Henzinger
“Abundance” Problem

http://simon.cs.cornell.edu/home/kleinber/
kleinber.html

• Given a query find:
  – Good Content (“Authorities”)
  – Good Sources of Links (“Hubs”)
• Mutually Reinforcing
• Simple (Core) Algorithm
\[ T \equiv \{ \text{n Pages} \}, \ A \equiv \{ \text{Links} \} \]

\[ X_p \in \mathbb{R}^{\geq 0}, \ p \in T \quad \text{non-negative “Authority Weights”} \]

\[ Y_p \in \mathbb{R}^{\geq 0}, \ p \in T \quad \text{non-negative “Hub Weights”} \]

I operation \quad \text{Update Authority Weights}

\[ X_p \leftarrow Y_q \quad \text{if} \quad (q,p) \in A \]

O operation \quad \text{Update Hub Weights}

\[ Y_p \leftarrow X_q \quad \text{if} \quad (p,q) \in A \]

Normalize:

\[ X^2_p = 1 \quad \text{if} \quad p \in T \]

\[ Y^2_p = 1 \quad \text{if} \quad p \in T \]
Core Algorithm

\[ Z \leftarrow (1,1,\ldots,1) \]
\[ X \leftarrow Y \leftarrow Z \]

Repeat until Convergence

Apply I  /* Update Authority weights */
Apply O  /* Update Hub Weights */
Normalize

Return Limit \((X^*, Y^*)\)
Convergence of \((X^i, Y^i) \xrightarrow{\triangle} (OI)^i(Z,Z)\)

\(A \triangleq n \times n \text{ “Adjacency Matrix”}\)

Rewrite I and O:

\[
X \leftarrow A^T Y \\
X^i = (A^T A)^{-1} A^T Z \\
Y \leftarrow A X \\
Y^i = (A A^T)^i Z
\]

\(A A^T\) Symm., Non-negative and \(Z = (1,1,\ldots, 1)\)

\(X^* \xrightarrow{\triangle} \lim_{i \to \infty} X^i = \omega_1(A^T A)\)

\(Y^* \xrightarrow{\triangle} \lim_{i \to \infty} Y^i = \omega_1(A A^T)\)
Whole Algorithm (k,d,c)

q \quad \text{Search Engine} \quad |S| \leq k

Base Set T:
(In S, S \rightarrow , \rightarrow S) and \leq d \text{ links/page}

Remove “Internal Links”

Run Core Algorithm on T

From Result (X,Y), Select

C pages with max X* values
C pages with max Y* values
Examples (k=200, d=5)

q = censorship + net
   www.EFF.org
   www.EFF.org/BlueRib.html
   www.CDT.org
   www.VTW.org
   www.ACLU.prg

q = Gates
   www.roadahead.com
   www.microsoft.com
   www.ms.com/corpinfo/bill-g.html

[Compares well with Yahoo, Galaxy, etc.]
Approach to “Massiveness”: Throw Out Most of G!!

- Non-principal Eigenvectors correspond to “Non-principal Communities”
- Open (?):
  - Objective Performance Criteria
  - Dependence on Search Engine
  - Nondeterministic Choice of S and T
Google History

- Privately held company, whose backers include Kleiner Perkins Caufield & Byers and Sequoia Capital.
- Continues to win top awards for Search Engines. Computer Scientists love it!!!
Major Partners

- Yahoo!
- Palm
- Nextel
- Netscape
- Cisco Systems
- Virgin Net
- Netease.com
- RedHat
- Virgilio
- Washingtonpost.com
Business Model

• The company delivers services through its own web site at www.google.com and by licensing its search technology to commercial sites

• Advertising:
  – Premium Sponsorship – Purchase a keyword
  – AdWords – Manage your Ad text
I’d like to buy a Keyword

The advertiser’s text-based ad will appear at the top of a Google results page whenever the keyword they have purchased is included in a user’s search.

The ads appear adjacent to, but are distinguished from, the results listings.
Category purchase

• Google uses a classification system to create an ongoing "Virtual Directory" of categories an advertiser can purchase.

• Advertisers can select the categories most appropriate to their business and Google will match the most relevant category ads to each user's search.

• The advantage of this approach is that it covers a broader audience that might be missed through the purchase of keywords alone.