PART 1. BATCH COMPUTING MODEL FOR BIG DATA ANALYTICS
2. WEB SCALE LINK ANALYSIS

Sangmi Lee Pallickara
Computer Science, Colorado State University
http://www.cs.colostate.edu/~cs535

FAQs
- Programming Assignment 1
  - We will discuss link analysis in week 3
  - Installation/configuration guidelines for Hadoop and Spark are uploaded
- TP teams
  - Any change? Please let me know

Today’s topics
- Link analysis
  - PageRank

Link Analysis
1. PageRank
2. Hyperlink-Induced Topic Search (HITS)
3. Centrality Analysis

This material is built based on,
  - Chapter 5
- http://infolab.stanford.edu/~ullman/mmds.html

Link Analysis
1. PageRank Algorithm
Definition of PageRank

- Link analysis
- Data-analysis technique used to evaluate relationships (connections) between nodes
- A function that assigns a real number to each page in the Web
- The higher the PageRank of a page, the more “important” it is
- There is NOT one fixed algorithm for assignment of PageRank

Example [1/5]

- Page A has links to B, C and D
- Page B has links to A and D
- Page C has a link to A
- Page D has links to B and C

Example [2/5]

- Suppose that a random surfer starts at page A
- Page B, C and D will be the next with probability 1/3
- 0 probability of being at A

Example [3/5]

- Now suppose the random surfer at B
- B has probability of ½ of being at A, ½ of being at D and 0 of being at B or C

Example [4/5]

- Transition matrix $M$
- What happens to random surfers after one step
- If it has $n$ rows and columns
- What is the transition matrix for this example?

Example [5/5]

- The first column
  - a surfer at A has a 1/3 probability of next being at each of the other pages
- The second column
  - a surfer at B has a ½ probability of being next at A and the same for being at D
What does this matrix mean? [1/6]

- The probability distribution for the location of a random surfer
  - A column vector whose \( j \)th component is the probability that the surfer is at page \( j \)
  - \( P(j|i) \)
  - Right stochastic matrix
    - Real square matrix with each row summing to 1

\[
\begin{bmatrix}
0 & 1/2 & 1/2 & 0 \\
1/3 & 0 & 0 & 1/2 \\
1/3 & 0 & 0 & 1/2 \\
1/3 & 1/2 & 0 & 0
\end{bmatrix}
\]

What does this matrix mean? [2/6]

- If we surf at any of the \( n \) pages of the Web with equal probability
  - The initial vector \( v_0 \) (row vector) will have \( 1/n \) for each component
  - If \( M \) is the transition matrix of the Web
    - After the first one step, the distribution of the surfer will be \( Mv_0 \)
    - After two steps, \( M \times Mv_0 = M^2v_0 \) and so on
    - The product of two right stochastic matrices is also right stochastic
  - Goal: Finding a row eigenvector of the transition matrix \( M \)
    - \( \lambda = m \)

What does this matrix mean? [3/6]

- Multiplying the initial vector \( v_0 \) by \( M \) a total of \( i \) times
  - The distribution of the surfer after \( i \) steps
    - The probability for the next step from the current location
    - The probability for being in the current location

\[
\begin{bmatrix}
0 & 1/2 & 1/2 & 0 \\
1/3 & 0 & 0 & 1/2 \\
1/3 & 0 & 0 & 1/2 \\
1/3 & 1/2 & 0 & 0
\end{bmatrix}
\]

What does this matrix mean? [4/6]

- The probability \( x_i \) that a random surfer will be at node \( i \) at the next step
  - \( m_{ij} \) is the probability that a surfer at node \( j \) will move to node \( i \) at the next step
  - \( v_j \) is the probability that the surfer was at node \( j \) at the previous step

\[
x_i = \sum m_{ij} v_j
\]

What does this matrix mean? [5/6]

- The limit is reached when multiplying the distribution by \( M \) another time does not change the distribution
  - The limiting \( v \) is a row eigenvector of \( M \)
    - Since \( M \) is right stochastic (its columns each add up to 1), \( v \) is the principle eigenvector
    - Its associated eigenvalue is the largest of all eigenvalues
  - The principle eigenvector of \( M \)
    - Where the surfer is most likely to be after a long time
  - For the Web, 50-75 iterations are sufficient to converge to within the error limits of double-precision arithmetic

What does this matrix mean? [6/6]

- The distribution of the surfer approaches a limiting distribution \( \nu \) that satisfies \( \nu = M\nu \)
  - Provided two conditions are met:
    1. The graph is strongly connected
      - It is possible to get from any node to any other node
    2. There are no dead ends
      - Nodes that have no arcs out
Example

\[
M = \begin{bmatrix}
0 & 1/2 & 1 & 0 \\
1/3 & 0 & 0 & 1/2 \\
1/3 & 0 & 0 & 1/2 \\
1/3 & 1/2 & 0 & 0 \\
\end{bmatrix}
\]

- Suppose we apply this process to the matrix \( M \)

- The initial vector \( v_0 \) and \( v_1 \) after multiplying \( M \)

\[
v_1 = Mv_0 = \begin{bmatrix}
1/3 & 0 & 0 & 1/2 \\
1/3 & 1/2 & 0 & 0 \\
\end{bmatrix}
\]

Use your worksheet

What is the \( v_2 \)?

\[
M = \begin{bmatrix}
0 & 1/2 & 1 & 0 \\
1/3 & 0 & 0 & 1/2 \\
1/3 & 0 & 0 & 1/2 \\
1/3 & 1/2 & 0 & 0 \\
\end{bmatrix}
\]

- Suppose we apply this process to the matrix \( M \)

- The initial vector \( v_0 \) and \( v_1 \) after multiplying \( M \)

\[
v_1 = Mv_0 = \begin{bmatrix}
1/3 & 0 & 0 & 1/2 \\
1/3 & 1/2 & 0 & 0 \\
\end{bmatrix}
\]

Example continued

- The sequence of approximations to the limit

\[
\begin{array}{c}
\frac{1}{4} \quad 0.48 \\
\frac{1}{4} \quad 0.48 \\
\frac{1}{4} \quad 0.48 \\
\frac{1}{4} \quad 0.48 \\
\end{array}
\]

- This difference in probability is not noticeable

- In the real Web, there are billions of nodes of greatly varying importance

  - The probability of being at a node like www.amazon.com is orders of magnitude greater than others

Problems we need to avoid

- Dead end
  - A page that has no links out
  - Surfers reaching such a page will disappear
  - In the limit, no page that can reach a dead end can have any PageRank at all

- Spider traps
  - Groups of pages that all have outlinks but they never link to any other pages

- What if we allow the Dead end and Spider traps in the PageRank computation?
Avoiding Dead Ends

- If we allow dead ends
  - The transition matrix of the Web is no longer stochastic
  - Some of the columns will sum to 0 rather than 1

- If we compute $M^n$ for increasing powers of a substochastic matrix
  - Some of all of the components of the vector go to 0
  - A matrix whose column sums are at most 1
  - Importance “drains out” of the Web
  - No information about the relative importance of pages

Example

- Remove the arc from C to A
- C becomes a _________

Example

- Remove the arc from C to A
- C becomes a dead end
- If a random surfer reaches C, they disappear at the next round

Example

- Repeatedly multiplying the vector by M:
- The probability of a surfer being anywhere goes to 0 as the number of steps increase

Approaches to dealing with dead ends

1. Recursive deletion
   - Step 1. Drop the dead ends from the graph
   - Step 2. Drop their incoming arcs as well
   - Step 3. Repeat Step 1 and 2

2. Taxation
   - Modify the process by which random surfers are assumed to move about the Web

Example of recursive deletion (1/4)
Example of recursive deletion (2/4)
- The final matrix for the graph is
\[
M = \begin{bmatrix}
0 & 0 & 0 \\
\frac{1}{3} & 0 & \frac{1}{3} \\
\frac{1}{3} & \frac{1}{3} & \frac{1}{3}
\end{bmatrix}
\]

Example of recursive deletion (3/4)
- We still need to compute deleted nodes (C and E)
- C was the last to be deleted
  - We know all its predecessors have PageRanks (A and D)
  - Therefore, PageRank of C = \(\frac{1}{3} \times \frac{2}{9} + \frac{1}{2} \times \frac{3}{9} = \frac{13}{54}\)

Spider Traps and Taxation
- Spider traps
  - A set of nodes with no dead ends but no arcs out (from the set of nodes)
  - This can appear intentionally or unintentionally on the Web
- Spider traps causes the PageRank calculation to place all the PageRank within the spider traps

Example 1
- There is a simple spider trap of 4 nodes
\[
M = \begin{bmatrix}
0 & 1/2 & 0 & 0 & 0 \\
1/3 & 0 & 1/2 & 0 & 0 \\
1/3 & 1/2 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1/2 & 0
\end{bmatrix}
\]
If we perform the usual iteration to compute the PageRank of the nodes, we get:

\[
\begin{pmatrix}
\frac{1}{6} & \frac{1}{12} & 0.042 \\
\frac{1}{6} & \frac{5}{36} & 0.046 \\
\frac{1}{6} & \frac{11}{36} & 0.222 \\
\frac{1}{6} & 0 & 0 \\
\frac{1}{6} & \frac{1}{12} & 0 \\
\end{pmatrix}
\]

All the PageRank is at C. Once there, a random surfer can never leave.

If we perform the usual iteration to compute the PageRank of the nodes, we get:

\[
\begin{pmatrix}
\frac{1}{6} & \frac{5}{24} & \frac{5}{48} & \frac{21}{288} & 0 \\
\frac{1}{6} & \frac{11}{24} & \frac{29}{48} & \frac{305}{288} & \frac{1}{2} \\
\frac{1}{6} & \frac{5}{24} & \frac{7}{48} & \frac{31}{288} & 0 \\
\end{pmatrix}
\]

To avoid Spider traps, we modify the calculation of PageRank.

- The iterative step, where we compute a new vector estimate of PageRanks \(v'\) from the current PageRank estimate \(v\) and the transition matrix \(M\) is:

\[
v' = \beta M v + (1 - \beta) e / n
\]

- The term \(\beta M v\) represents the case where,
  - With probability \(\beta\), the random surfer decides to follow an out-link from their present page.
  - \(\beta\) is a chosen constant, usually in the range 0.8 to 0.9.
  - \(e\) is a vector for all 1's with the appropriate number of components.
  - \(n\) is the number of nodes in the Web graph.

- The term \((1 - \beta) e / n\) is a vector,
  - Each of whose components has value \((1 - \beta)/n\), representing the introduction with probability \(1 - \beta\) of a new random surfer at a random page.
- If the graph has no dead ends
  - The probability of introducing a new random surfer is exactly equal to the probability that the random surfer will decide not to follow a link from their current page
  - Surfer decides either to follow a link or teleport to a random page

- If the graph has dead ends
  - The surfer goes nowhere
  - The term \( (1 - \beta) \) does not depend on the sum of the components of the vector \( v \), there will be some fraction of a surfer operating on Web
  - When there are dead ends, the sum of the components of \( v \) may be less than 1
  - But it will never reach 0

Example of Taxation (1/2)
\[
M = \begin{bmatrix}
0 & 1/2 & 0 & 0 \\
1/3 & 0 & 1/2 \\
1/3 & 0 & 1 \\
1/3 & 1/2 & 0 & 0
\end{bmatrix}
\]
\[
\beta = 0.8
\]
\[
v' = \beta M v + (1 - \beta) e / n
\]
\[
\beta = 0.8
\]
\[
v' = \begin{bmatrix}
0.25 & 0 & 0 \\
1/4 & 1/2 & 0 \\
1/4 & 1/2 & 0 & 1/4 \\
1/4 & 1/2 & 0 & 1/4
\end{bmatrix}
\]

Example of Taxation (2/2)
- First few iterations:
  - Effect is limited
  - Each of the nodes gets some of the PageRank

Example 1
- Compute the PageRank of each page assuming \( \beta = 0.8 \)

\[
\begin{bmatrix}
0.3 & 1/2 \\
1/3 & 0 & 1/2 \\
0 & 1/3 & 1/2 \\
0 & 1/3 & 1/2 & 0
\end{bmatrix}
\]
\[
v = \begin{bmatrix}
1/3 \\
1/3 \\
1/3 \\
1/3
\end{bmatrix}
\]
\[
\beta M v + (1 - \beta) e / n = 0.8 \times \begin{bmatrix}
1/3 \\
1/3 \\
1/3 \\
1/3
\end{bmatrix} + (1 - 0.8) e / 3
\]
Example 2

- **clique**
  - Set of nodes with all possible arcs from one to another
  - Suppose the Web consists of a clique of $n$ nodes and a single additional node that is the successor of each of the $n$ nodes in the clique

\[
M = \begin{bmatrix}
\frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\
\frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n}
\end{bmatrix}
\]

\[
\beta M v_0 + (1 - \beta) \frac{1}{n}
\]

Example 3

- Suppose that we recursively eliminate dead ends from the Web graph to solve the remaining graph
- Suppose that the graph is a chain of dead ends, headed by a node with a self-loop
- What would be the PageRank assigned to each of the nodes?
Example 3
- Remove all of the dead ends recursively
- What is $v_0$ and $M$?

$A
B
C
D
Z$

Example of Taxation
$M = \begin{bmatrix} 0 & 1/2 & 0 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}$

$v' = \beta M v + (1-\beta) e / n$

$\beta = 0.8$

$\begin{bmatrix} 0 & 2/5 & 0 & 0 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 2/5 & 0 & 0 \end{bmatrix} v' + \begin{bmatrix} 1/20 \\ 1/20 \\ 1/20 \\ 1/20 \end{bmatrix}$

Example of Taxation
$M = \begin{bmatrix} 0 & 1/2 & 0 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}$

$v' = \beta M v + (1-\beta) e / n$

$\beta = 0.8$

$\begin{bmatrix} 0 & 2/5 & 0 & 0 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 2/5 & 0 & 0 \end{bmatrix} v' + \begin{bmatrix} 1/20 \\ 1/20 \\ 1/20 \\ 1/20 \end{bmatrix}$

For the first few iterations:

$\begin{bmatrix} 15/60 & 41/300 & 543/4500 & 15/148 \\ 15/60 & 53/300 & 707/4500 & 19/148 \\ 15/60 & 53/300 & 707/4500 & 19/148 \end{bmatrix}$