PART 1.
LARGE SCALE DATA ANALYSIS USING MAPREDUCE

Instructor: Sangmi Lee Pallickara
Computer Science, Colorado State University
http://www.cs.colostate.edu/~cs435

Today’s topics
- FAQs
- PageRank Algorithm
- Implementing PageRank Algorithm with MapReduce

FAQs
- PA2 is available

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

What are these?
- Archie
- Veronica
- Infoseek
- Snap
- Direct Hit
- Lycos
- AltaVista
- Excite
- Yahoo
- Google
Early Search Engines
- They worked by crawling the Web and listing the terms
  - Words or other strings of characters other than white space
  - In an inverted index
- An inverted index is a data structure that makes it easy, given a term, to find (pointer to) all the places where that term occurs.

Inverted index (1/2)
- Inverted index
  - For given texts,
    
    \[
    \begin{align*}
    T[0] &= \text{"Colorado State University"} \\
    T[1] &= \text{"Colorado water source"} \\
    T[2] &= \text{"University of Colorado"}
    \end{align*}
    \]

  - We have the following inverted file index

    - "Colorado": \{0,1,2\}
    - "of":\{2\}
    - "source":\{1\}
    - "State": \{0\}
    - "University": \{0,2\}
    - "water":\{1\}

Inverted index (2/2)
- A term search for the terms, "Colorado", "State", and "University" would give the set

\[
\{0,1,2\} \cap \{0\} \cap \{0,2\} = \{0\}
\]

Term spam
- If you were selling shirts on the Web
  - All you care about was that people would see your page
  - You could add a term like "movie" to your page
  - Add thousands of times
  - It does not even need to show
  - Give the same color as background to the letters
  - A search engine would think this page is very important one about "movie"
  - You could go to the search engine and search "movie" and see the first listed page
  - Copy that page with the same color as background
PageRank

- Goals
  - Providing *effective summaries* for the search results
  - Ordering/Ranking results

- Simulate random Web surfers
  - Pages that would have a large number of surfers were considered more "important" than pages that would rarely be visited

- The content of a page was judged not only by the terms appearing on that page
  - But by the terms used in or near the links to that page

Definition of PageRank

- 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

- 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

\[ \begin{array}{ccc}
A & B & C \\
\uparrow & \nearrow & \searrow \\
1/3 & 1/2 & 1/3
\end{array} \]

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

\[ \begin{array}{ccc}
A & B & C \\
\uparrow & \nearrow & \searrow \\
1/2 & 0 & 0
\end{array} \]

What does this matrix mean? [1/6]
- The probability distribution for the location of a random surfer
- A column vector whose jth component is the probability that the surfer is at page j

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

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

The distribution of the surfer after the first step
- Probability for being in the next possible location

The probability \( x_i \) that a random surfer will be at node \( i \) at the next step
\[ x_i = \sum_j m_{ij} v_j \]

- \( m_{ij} \) 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

What does this matrix mean? [4/6]

- The probability \( x_i \) that a random surfer will be at node \( i \) at the next step
- The probability for the next step from the current location
- The probability for being in the current location

What does this matrix mean? [5/6]

- The distribution of the surfer approaches a limiting distribution \( v \) that satisfies \( v = Mv \) 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

What does this matrix mean? [6/6]

- The limit is reached when multiplying the distribution by \( M \) another time does not change the distribution
- The limit \( v \) is an eigenvector of \( M \)
  - Since \( M \) is 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

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 \( M \)

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 \( M \)
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 $M$

$$v_1 = Mv_0 = \begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1/24 \\ 5/24 \\ 5/24 \\ 5/24 \end{bmatrix} = \begin{bmatrix} 9/24 \\ 5/24 \\ 5/24 \\ 5/24 \end{bmatrix}$$

Example continued

- The sequence of approximations to the limit
- We get by multiplying at each step by $M$ is
  
  $$\begin{bmatrix} \frac{9}{24} \\ \frac{5}{24} \\ \frac{5}{24} \\ \frac{5}{24} \end{bmatrix}, \begin{bmatrix} \frac{5}{24} \\ \frac{11}{48} \\ \frac{7}{32} \\ \frac{2}{9} \end{bmatrix}, \ldots$$

- 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

Link Analysis

Matrix multiplication using MapReduce

Matrix-vector Multiplication by MapReduce

- Suppose we have an $n \times n$ matrix $M$, whose element in row $i$ and column $j$ will be denoted $M_{ij}$
- Then the matrix-vector product is the vector $v$ of length $n$, whose $i$th element $x_i$ is given by,

$$x_i = \sum_{j=1}^{n} M_{ij}v_j$$

Matrix-vector Multiplication by MapReduce

- If $n = 100$, we do not need DFS or MapReduce
- However, if this calculation is a part of ranking Web pages ($n$ is 10M) that goes on at search engine? The vector $v$ cannot fit in main memory
- More than 1.4B pages

Matrix-vector Multiplication by MapReduce

- The matrix $M$ and the vector $v$ each will be stored in a file of the DFS(HDFS)
- Assume that row-column coordinates of each matrix element will be discoverable
  - Either from its position in the file or explicit coordinates
**The Map function**

- The Map function is written to apply to one element of $M$.
- Each Map task will operate on a chunk of the matrix $M$.
- From each matrix element $m_{ij}$, it produces the key-value pair $(i, m_{ij})$.
- All terms of the sum that make up the component $x_i$ of the matrix-vector product will get the same key, $i$.

**The Reduce function**

- Sums all the values associated with a given key $i$.
- The result will be a pair $(i, x_i)$.

**If the vector $v$ cannot fit in main memory?**

- It is possible that the vector $v$ is so large that it will not fit in its main memory entirely.
- We can divide the matrix into vertical stripes of equal width and divide the vector into an equal number of horizontal stripes of the same height.
- The goal is to use enough stripes so that the portion of the vector in one stripe can fit into main memory.

**Link Analysis**

**PageRank Algorithm for the real Web**
Structure of Web (1/3)

Is the Web strongly connected?

Structure of Web (2/3)

- Strongly Connected Component
- Tubes
  - Pages reachable from the in-component and able to reach the output-component, but unable to reach the SCC or be reached from the SCC
  - Isolated components
    - Unreachable from the large components

Structure of Web (3/3)

- Tendrils In
- Tendrils Out
- Disconnected Components

Anomalies from the Web structure

- These structures violate the assumptions needed for the Markov process iteration to converge to a limit
  - When a random surfer enters the out-component, they can never leave
  - Surfers starting in either the SCC or in-component are going to wind up in either the out-component or a tendril off the in-component
  - No page in the SCC or in-component winds up with any probability of a surfer being there
  - Nothing in the SCC or in-component will be of any importance

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

Link Analysis

Challenges in PageRank Algorithm for the real Web
Example

- Suppose that both the PageRank and TrustRank were computed
- Teleport set was page B and D
- Which nodes are not the link spams?
- Is there any link spam?

<table>
<thead>
<tr>
<th>Web Page</th>
<th>PageRank</th>
<th>TrustRank</th>
<th>SpamMess</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3/9</td>
<td>54/210</td>
<td>0.229</td>
</tr>
<tr>
<td>B</td>
<td>2/9</td>
<td>59/210</td>
<td>-0.264</td>
</tr>
<tr>
<td>C</td>
<td>2/9</td>
<td>38/210</td>
<td>0.186</td>
</tr>
<tr>
<td>D</td>
<td>2/9</td>
<td>59/210</td>
<td>-0.264</td>
</tr>
</tbody>
</table>