COT 6936: Topics in Algorithms

Giri Narasimhan

ECS 254A / EC 2474; Phone x3748; Email: giri@cs.fiu.edu
HOMEPAGE: http://www.cs.fiu.edu/~giri
https://moodle.cis.fiu.edu/v2.1/course/view.php?id=612

Mar 25, 2014

Presentation Outline

Credits and Acknowledgments

Lecture slides are based on

1. Lecture slides by Ryan Tibshirani, CMU, See: http://www.stat.cmu.edu/~ryantibs/datamining/lectures/03-pr-marked.pdf; http://www.stat.cmu.edu/~ryantibs/datamining/lectures/03-pr.pdf
2. Original paper by Brin and Page from: http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf

Presentation Outline

Searching and Ranking

- WWW is large, heterogenous, and hyperlinked

Searching and Ranking

- WWW is large, heterogenous, and hyperlinked
- Two distinct operations in the search process:

Searching and Ranking

- WWW is large, heterogenous, and hyperlinked
- Two distinct operations in the search process:
    - Search by posing a query

Searching and Ranking

- WWW is large, heterogenous, and hyperlinked
- Two distinct operations in the search process:
  - Search by posing a query
  - Rank by deciding which of the hits is relevant/important

Searching and Ranking

- WWW is large, heterogenous, and hyperlinked
- Two distinct operations in the search process:
  - Search by posing a query
  - Rank by deciding which of the hits is relevant/important
- Searching is done using principles from Information Retrieval;

Searching and Ranking

- WWW is large, heterogenous, and hyperlinked
- Two distinct operations in the search process:
  - Search by posing a query
  - Rank by deciding which of the hits is relevant/important
- Searching is done using principles from Information Retrieval; Well developed body of work quantifying similarity of documents;

Searching and Ranking

- WWW is large, heterogenous, and hyperlinked
- Two distinct operations in the search process:
  - Search by posing a query
  - Rank by deciding which of the hits is relevant/important
- Searching is done using principles from Information Retrieval; Well developed body of work quantifying similarity of documents;
- Since a search could return millions of pages, we need to rank the results;

Searching and Ranking

- WWW is large, heterogenous, and hyperlinked
- Two distinct operations in the search process:
  - Search by posing a query
  - Rank by deciding which of the hits is relevant/important
- Searching is done using principles from Information Retrieval; Well developed body of work quantifying similarity of documents;
- Since a search could return millions of pages, we need to rank the results; we will focus on the Ranking operation,

Ranking

- Ranking needs to be objective and quantitative

Ranking

- Ranking needs to be objective and quantitative
- Idea:

Ranking

- Ranking needs to be objective and quantitative
- Idea: A web link can be considered similar to a citation for a publication;

Ranking

- Ranking needs to be objective and quantitative
- Idea: A web link can be considered similar to a citation for a publication; a publication is important if it is cited a lot;

Ranking

- Ranking needs to be objective and quantitative
- Idea: A web link can be considered similar to a citation for a publication; a publication is important if it is cited a lot;
- This idea does not work

Ranking

- Ranking needs to be objective and quantitative
- Idea: A web link can be considered similar to a citation for a publication; a publication is important if it is cited a lot;
- This idea does not work because creating new pages and links is trivial and can be automated

Ranking

- Ranking needs to be objective and quantitative
- Idea: A web link can be considered similar to a citation for a publication; a publication is important if it is cited a lot;
- This idea does not work because creating new pages and links is trivial and can be automated
- PageRank is an answer to this problem

Main Principles

1. Links to X from important webpages should be considered important

Main Principles

1. Links to X from important webpages should be considered important
2. Links to X from unimportant webpages, i.e., pages with links to a lot of other pages should not have high importance

First Attempt to define PageRank

First Attempt to define PageRank

Let $L_{ij} = 1$ if $j \to i$, and 0 otherwise.

First Attempt to define PageRank

Let $L_{ij} = 1$ if $j \rightarrow i$, and 0 otherwise.
Let the outdegree of j be m_j.

First Attempt to define PageRank

Let $L_{ij} = 1$ if $j \to i$, and 0 otherwise.
Let the outdegree of j be m_j.

$$Rank_i = \sum_{j \to i} \frac{p_j}{m_j} = \sum_{j=1}^{n} \frac{p_j}{m_j} L_{ij}$$

First Attempt to define PageRank

Let $L_{ij} = 1$ if $j \rightarrow i$, and 0 otherwise.
Let the outdegree of j be m_j.

$$Rank_i = \sum_{j \rightarrow i} \frac{p_j}{m_j} = \sum_{j=1}^{n} \frac{p_j}{m_j} L_{ij}$$

As required, it increases with p_j and decreases with m_j.

First Attempt to define PageRank

Let $L_{ij} = 1$ if $j \to i$, and 0 otherwise.
Let the outdegree of j be m_j.

$$Rank_i = \sum_{j \to i} \frac{p_j}{m_j} = \sum_{j=1}^{n} \frac{p_j}{m_j} L_{ij}$$

As required, it increases with p_j and decreases with m_j.
Therefore, it matches our ideas from last slide.

First Attempt to define PageRank

In matrix notation, we have:

$$p = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{pmatrix},$$

First Attempt to define PageRank

In matrix notation, we have:

$$p = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{pmatrix}, \qquad L = \begin{pmatrix} L_{11} & L_{12} & \cdots & L_{1n} \\ L_{21} & L_{22} & \cdots & L_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ L_{n1} & L_{n2} & \cdots & L_{nn} \end{pmatrix}$$

First Attempt to define PageRank

In matrix notation, we have:

$$p = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{pmatrix}, \qquad L = \begin{pmatrix} L_{11} & L_{12} & \cdots & L_{1n} \\ L_{21} & L_{22} & \cdots & L_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ L_{n1} & L_{n2} & \cdots & L_{nn} \end{pmatrix}$$

$$M = \begin{pmatrix} m_1 & 0 & \cdots & 0 \\ 0 & m_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & m_n \end{pmatrix}$$

First Attempt to define PageRank

In matrix notation, we have:

$$p = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{pmatrix}, \qquad L = \begin{pmatrix} L_{11} & L_{12} & \cdots & L_{1n} \\ L_{21} & L_{22} & \cdots & L_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ L_{n1} & L_{n2} & \cdots & L_{nn} \end{pmatrix}$$

$$M = \begin{pmatrix} m_1 & 0 & \cdots & 0 \\ 0 & m_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & m_n \end{pmatrix}$$

And, $\mathbf{p} = LM^{-1}\mathbf{p}$.

First Attempt to define PageRank

In matrix notation, we have:

$$p = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{pmatrix}, \qquad L = \begin{pmatrix} L_{11} & L_{12} & \cdots & L_{1n} \\ L_{21} & L_{22} & \cdots & L_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ L_{n1} & L_{n2} & \cdots & L_{nn} \end{pmatrix}$$

$$M = \begin{pmatrix} m_1 & 0 & \cdots & 0 \\ 0 & m_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & m_n \end{pmatrix}$$

And, $\mathbf{p} = LM^{-1}\mathbf{p}$. What does this remind you of?

Markov Chains

If

$$\mathbf{p} = LM^{-1}\mathbf{p},$$

Markov Chains

If

$$\mathbf{p} = LM^{-1}\mathbf{p},$$

then set $A = LM^{-1}$, and we get $p = Ap$.

Markov Chains

If

$$\mathbf{p} = LM^{-1}\mathbf{p},$$

then set $A = LM^{-1}$, and we get $p = Ap$. Consider a Markov chain,

Markov Chains

If

$$\mathbf{p} = LM^{-1}\mathbf{p},$$

then set $A = LM^{-1}$, and we get $p = Ap$. Consider a Markov chain, where the webpages are the states,

Markov Chains

If

$$\mathbf{p} = LM^{-1}\mathbf{p},$$

then set $A = LM^{-1}$, and we get $p = Ap$. Consider a Markov chain, where the webpages are the states, and transition matrix is A^T,

Markov Chains

If

$$\mathbf{p} = LM^{-1}\mathbf{p},$$

then set $A = LM^{-1}$, and we get $p = Ap$. Consider a Markov chain, where the webpages are the states, and transition matrix is A^T, where

$$(A^T)_{ij} = A_{ji} = L_{ji}/m_i$$

Markov Chains

If

$$\mathbf{p} = LM^{-1}\mathbf{p},$$

then set $A = LM^{-1}$, and we get $p = Ap$. Consider a Markov chain, where the webpages are the states, and transition matrix is A^T, where

$$(A^T)_{ij} = A_{ji} = L_{ji}/m_i$$

$$Prob(i \to j) = \begin{cases} 1/m_i & \text{if } i \to j \\ 0 & \text{otherwise.} \end{cases}$$

Markov Chains

If

$$\mathbf{p} = LM^{-1}\mathbf{p},$$

then set $A = LM^{-1}$, and we get $p = Ap$. Consider a Markov chain, where the webpages are the states, and transition matrix is A^T, where

$$(A^T)_{ij} = A_{ji} = L_{ji}/m_i$$

$$Prob(i \to j) = \begin{cases} 1/m_i & \text{if } i \to j \\ 0 & \text{otherwise.} \end{cases}$$

Random Walk on webpage;

Markov Chains

If

$$\mathbf{p} = LM^{-1}\mathbf{p},$$

then set $A = LM^{-1}$, and we get $p = Ap$. Consider a Markov chain, where the webpages are the states, and transition matrix is A^T, where

$$(A^T)_{ij} = A_{ji} = L_{ji}/m_i$$

$$Prob(i \to j) = \begin{cases} 1/m_i & \text{if } i \to j \\ 0 & \text{otherwise.} \end{cases}$$

Random Walk on webpage; if multiple links from a page, each followed with equal probability.

Markov Chains

If

$$\mathbf{p} = LM^{-1}\mathbf{p},$$

then set $A = LM^{-1}$, and we get $p = Ap$. Consider a Markov chain, where the webpages are the states, and transition matrix is A^T, where

$$(A^T)_{ij} = A_{ji} = L_{ji}/m_i$$

$$Prob(i \to j) = \begin{cases} 1/m_i & \text{if } i \to j \\ 0 & \text{otherwise.} \end{cases}$$

Random Walk on webpage; if multiple links from a page, each followed with equal probability.

Stationary Distribution: $\mathbf{p} = A\mathbf{p}$.

- Main Idea: Probabilities involved in the stationary distribution can be used a measure of importance of webpage.

Markov Chains and PageRank

- Main Idea: Probabilities involved in the stationary distribution can be used a measure of importance of webpage.
- Problem with the model: Stationary distribution exists only if the Markov chain is irreducible and aperiodic

Markov Chains and PageRank

- Main Idea: Probabilities involved in the stationary distribution can be used a measure of importance of webpage.
- Problem with the model: Stationary distribution exists only if the Markov chain is irreducible and aperiodic
- Fix: pick small constant $0 < d < 1$, and set

$$p_i = (1-d) + d \sum_{j=1}^{n} \frac{p_j}{m_j} L_{ij}$$

Markov Chains and PageRank

- Main Idea: Probabilities involved in the stationary distribution can be used a measure of importance of webpage.
- Problem with the model: Stationary distribution exists only if the Markov chain is irreducible and aperiodic
- Fix: pick small constant $0 < d < 1$, and set

$$p_i = (1-d) + d \sum_{j=1}^{n} \frac{p_j}{m_j} L_{ij}$$

$$\mathbf{p} = (1-d)\mathbf{e} + d \cdot \mathbf{LM}^{-1}\mathbf{p}$$

Presentation Outline

PageRank and its Interpretations

- Original model of Page and Brin: surfer did a random walk ignoring contents.
- Corrected model: With probability $(1-d)$, jumps to a random page.
- $$\mathbf{p} = (1-d)\mathbf{e} + d \cdot \mathbf{L}\mathbf{M}^{-1}\mathbf{p}$$
- Even if web graph is disconnected, there is a probability that the random jump will take you to a different component.

Example

$$L = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Example

$$L = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Thus $\mathbf{m} = (2, 1, 1, 1)$

Example

$$L = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Thus $\mathbf{m} = (2,\ 1,\ 1,\ 1)$
Then: $p' = (1.49, 0.78, 1.58, 0.15)$

Second Example

$$L = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Second Example – Cont'd

With $d = 0.85$, and $A = \frac{1-d}{n}E + dLM^{-1}$,

$$A = \frac{0.15}{5} \cdot \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix} + 0.85 \cdot \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

$$= \begin{pmatrix} 0.03 & 0.03 & 0.88 & 0.03 & 0.03 \\ 0.88 & 0.03 & 0.03 & 0.03 & 0.03 \\ 0.03 & 0.88 & 0.03 & 0.03 & 0.03 \\ 0.03 & 0.03 & 0.03 & 0.03 & 0.88 \\ 0.03 & 0.03 & 0.03 & 0.88 & 0.03 \end{pmatrix}$$

Then the stationary distribution is $(0.2, 0.2, 0.2, 0.2, 0.2)$.

Presentation Outline

Computing PageRank

Since webgraph is unreasonably large (10^{10} nodes), matrix operations are impossible ($O(n^3)$).

Faster iterative computation:

Computing PageRank

Since webgraph is unreasonably large (10^{10} nodes), matrix operations are impossible ($O(n^3)$).

Faster iterative computation: Start with any initial distribution $p^{(0)}$.

Computing PageRank

Since webgraph is unreasonably large (10^{10} nodes), matrix operations are impossible ($O(n^3)$).

Faster iterative computation: Start with any initial distribution $p^{(0)}$. Then

$$p^{(1)} = Ap^{(0)}$$

Computing PageRank

Since webgraph is unreasonably large (10^{10} nodes), matrix operations are impossible ($O(n^3)$).

Faster iterative computation: Start with any initial distribution $p^{(0)}$. Then

$$p^{(1)} = Ap^{(0)}$$
$$p^{(2)} = Ap^{(1)}$$

Computing PageRank

Since webgraph is unreasonably large (10^{10} nodes), matrix operations are impossible ($O(n^3)$).

Faster iterative computation: Start with any initial distribution $p^{(0)}$. Then

$$\begin{aligned} p^{(1)} &= Ap^{(0)} \\ p^{(2)} &= Ap^{(1)} \\ &\vdots \\ p^{(t)} &= Ap^{(t-1)} \end{aligned}$$

Computing PageRank

Since webgraph is unreasonably large (10^{10} nodes), matrix operations are impossible ($O(n^3)$).

Faster iterative computation: Start with any initial distribution $p^{(0)}$. Then

$$\begin{aligned} p^{(1)} &= Ap^{(0)} \\ p^{(2)} &= Ap^{(1)} \\ &\vdots \\ p^{(t)} &= Ap^{(t-1)} \end{aligned},$$

Finally $p^{(t)} \rightarrow p$ as $t \rightarrow \infty$

Presentation Outline

Basic Web Search

1. Compute PageRank vector

Basic Web Search

1. Compute PageRank vector
2. Find documents containing all words in query

Basic Web Search

1. Compute PageRank vector
2. Find documents containing all words in query
3. Sort documents by PageRank and return top k