Data Mining and Machine Learning

Madhavan Mukund

Lecture 22, Jan–Apr 2020

https://www.cmi.ac.in/~madhavan/courses/dmml2020jan/

Information retrieval on the Internet

- Traditional IR
  - Books published after editing, review — trustworthy content
- IR for Internet
  - Internet documents are self-published, unverified
  - Economic incentive to boost rankings through fraudulent means
  - Ranking algorithms should try not to be fooled
- Easy to add invisible content in HTML to misdirect search
  - Merging text and background colour, overlay text with images, unreadable font size
- Self published documents may omit useful search terms
  - IBM webpage did not mention the word “computer”

Exploiting hypertext

- Hypertext links refer from one document to another
  - `<a href="https://www.cmi.ac.in"> CMI webpage </a>`
  - Target location : `https://www.cmi.ac.in`
  - Anchor text : `CMI webpage`
- Use anchor text to index document at target location
  - Reliable indicator of what target document is about
- Hyperlinks also connect internet documents as a directed graph
  - Reason about the World Wide Web (WWW) as a gigantic graph
  - Use techniques from social network analysis

Social network analysis — prestige

- Consider the film industry
  - When is an actor a star? When is a director famous?
  - Stars are sought out by famous directors
  - Famous directors get stars to work in their films
  - Recursive definition
- Network (graph) of actors and directors, matrix M

$$
\begin{array}{cc}
 & \textit{Directors} \\
 & j \\
\textit{Actors} \quad i & \begin{bmatrix} & \vdots & \\ \cdots & 1 & \\ & & \end{bmatrix}
\end{array}
$$

$M[i,j] = 1$ if Actor i works in a film directed by Director j

Social network analysis — prestige

- Each actor i has star value $S[i]$
- Each director j has fame $F[j]$
- Actors derive star value from the famous directors they work with

$$S[i] = \sum_j M[i,j] \cdot F[j], \text{ or } S = M \cdot F$$

- Directors derive fame from the stars who work with them

$$F[j] = \sum_i M[i,j] \cdot S[i], \text{ or } F = M^\top \cdot S$$

- Substituting F from second equation, $S = M \cdot M^\top \cdot S$
- Substituting S from first equation, $F = M^\top \cdot M \cdot F$
- Solve for S, F to compute star ratings, fame

Prestige for webpages

- Each document i has prestige $P[i]$
- Prestigious (reliable) documents confer prestige on documents they link to
  - $P[i]$ is shared equally among all outgoing links
- A document derives prestige from documents that link to it
  - $P[i]$ is sum of prestige transferred by incoming links
- Structure of the internet, adjacency matrix A

$$
\begin{array}{cc}
 & \textit{Webpages} \\
 & j \\
\textit{Webpages} \quad i & \begin{bmatrix} & \vdots & \\ \ldots & 1 & \\ & & \end{bmatrix}
\end{array}
$$

$A[i,j] = 1$ if webpage i has a link to webpage j

Prestige for webpages ...

- Suppose $A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix}$
- Each document initially has prestige 1, $P = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$
- If a webpage points to n other pages, each of them gets $1/n$ of $P[i]$
- Prestige transfer matrix, $A^* = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}$
- One step: $P^\top \cdot A^* = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} = \begin{bmatrix} 1.5 & 1 & 0.5 \end{bmatrix}$

Page rank

- Stable solution: $P^\top \cdot A^* = P^\top$
- $P[i]$ is Page rank of webpage i
  - Larry Page, co-founder of Google with Sergey Brin
- How do we compute $P^\top$?
- A^* is a stochastic matrix — each row sums to 1

  $$\forall i \sum_j A^*[i,j] = 1$$

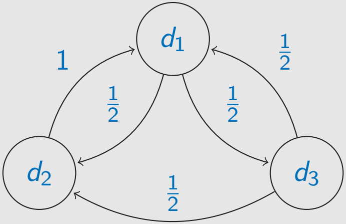
- Intepret $A^*[i,j]$ as probability of moving from document i to document j — random web surfer
- Use theory of Markov chains

Markov chains

- Finite set of states, with transition probabilities between states
- For us, states are documents
  - Henceforth, write A^* as A for convenience

$$A = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}$$

Three state Markov chain



- $P[j]$ is probability of being in document j
- Start in document 1, so initially $P = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$

Markov chains ...

- After one step: $P^\top A = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$
- After second step: $\begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} = \begin{bmatrix} \frac{3}{4} & \frac{1}{4} & 0 \end{bmatrix}$
- After k steps, $P[j]$ is probability of being in state j
- Continuing our example,

  $\begin{bmatrix} \frac{3}{4} & \frac{1}{4} & 0 \end{bmatrix} \rightarrow \begin{bmatrix} \frac{1}{4} & \frac{3}{8} & \frac{3}{8} \end{bmatrix} \rightarrow \begin{bmatrix} \frac{9}{16} & \frac{5}{16} & \frac{1}{8} \end{bmatrix}$
- Is it the case that $P[j] > 0$ for all j continuously, after some point?

Ergodicity

- Markov chain A is ergodic if there is some t_0 such that for every P, for all $t > t_0$, for every j, $(P^\top A^t)[j] > 0$.
  - No matter where we start, after $t > t_0$ steps, every state has a nonzero probability of being visited in step t
- Properties of ergodic Markov chains
  - There is a stationary distribution π such that $\pi^\top A = \pi^\top$
    - $\pi^\top$ is a left eigenvector of A
  - For *any* starting distribution P, $\lim_{t\to\infty} P^\top A^t = \pi^\top$

Ergodicity . . .

- How can ergodicity fail?
  - Starting from i, we reach a set of states from which there is no path back to i
  - We have a cycle $i \to j \to k \to i \to j \to k \cdots$, so we can only visit some states periodically
- Sufficient conditions for ergodicity
  - Irreducibility: When viewed as a directed graph, A is strongly connected
    - For all states i, j, there is a path from i to j and a path from j to i
  - Aperiodicity: For any pair of vertices i, j, the gcd of the lengths of all paths from i to j is 1
    - In particular, paths (loops) from i to i do not all have lengths that are multiples of some $k \geq 2$
    - Prevents bad cycles

Making the web graph ergodic

- No reason why web graph is irreducible and aperiodic
- Web graph has dead ends — terminal documents, no outgoing links
- Solution: Add random jumps between documents — teleportation
- Teleportation matrix T: For all i, j, $T[i, j] = 1/N$, where N is the total number of documents
  - The random surfer ignores all the links in the current document and types a new URL
- Let α be the probability of teleportation: $M = \alpha T + (1 - \alpha)A$
  - Check that M is stochastic
- By construction,
  - M is strongly connected — direct edge between each pair of documents
  - M is aperiodic — paths of any length exist between i and j
  - M has no dead ends

Page Rank

- In the modified web graph, stationary distribution is the Page rank, $\pi^T M = \pi^T$
- Compute using $\lim_{t\to\infty} P^T M^t$
- Use recursive doubling to accelerate computation of $\lim_{t\to\infty} P^T M^t$
  - Compute M, M^2, $(M^2)^2 = M^4$, …, $(M^{2i})^2 = M^{4i}$, …
  - Set a threshold for progress to stop the process
- Some limitations of Page rank
  - Universal property of a webpage, independent of a query
  - Define a topic-sensitive page rank
- Page rank was one the keys to the initial success of Google
  - Constant tweaks to ranking algorithm to keep ahead of search engine optimizers (SEO)

Summary

- IR on webpages presents new challenges because document content is unreliable
- Hypertext tags can provide better indexing terms
- Hypertext links create a graph of webpages
- Apply techniques from social network analysis, Markov chains
- Page rank computes the prestige of a documents using the graph structure of webpages