Advanced ML

– PageRank computation –

Eric Gaussier

Univ. Grenoble Alpes

UFR-IM^{2}AG, LIG, MIAI@Grenoble Alpes

eric.gaussier@imag.fr

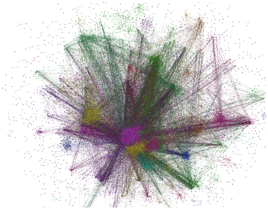
Table of content

IR on the web

The web is not a standard collection!

IR on the web

The web is not a standard collection!



Exploitng hyperlinks

Hyperlinks constitute an important source of information that can be used to improve IR search

1. Enriched indexing of documents/pages through anchors pointing at them
2. Taking into account the importance of a page in the web (its *PageRank*)

Enriched indexing

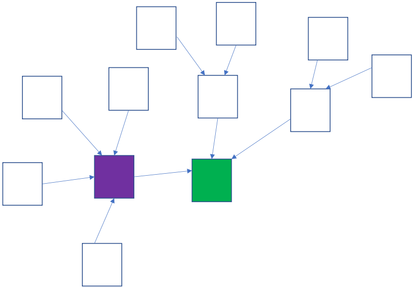
Html anchor which points to www.ibm.com and which contains the text Big Blue

Big Blue

- Enriched indexing by adding to a description of a page all anchor texts pointing to it
- This enrichment can easily be done at the same time the collection is indexed

Importance of a page on the web

Content-wise, the purple and green pages are equivalent.
Which one one should privilege?



Importance of a page on the web

How to measure the importance of a page?

- Number of outgoing links?
- Number of incoming links?
- ... ?
- *Number of incoming links, each link being weighted by the importance of the page they originate from*

A page is all the more important that it is pointed to by many important pages

Importance of a page on the web

How to measure the importance of a page?

- Number of outgoing links?
- Number of incoming links?
- ... ?
- *Number of incoming links, each link being weighted by the importance of the page they originate from*

A page is all the more important that it is pointed to by many important pages

Table of content

A simple random walk (1)

Imagine a walker that starts on a page and randomly steps to a page pointed to by the current page, and does so infinitely

A simple random walk (2)

In an infinite *random walk*,

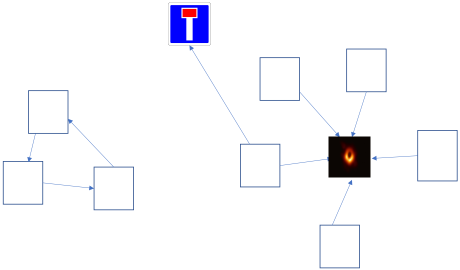
1. The number of visits of a page divided by the number of steps gives an estimation of the probability of visiting a page in a random walk (the longer the walk, the more accurate the estimation)

2. The probabilities thus obtained are *all the more important that the page considered is pointed to by important pages*

A simple random walk (3)

There are however a few problems!

1. Dead ends, black holes
2. Cycles



Solution: teleportation

- At each step, the walker can either randomly choose an outgoing page, with prob. λ, or teleport to any page of the graph, with prob. $(1-\lambda)$
- It's as if all web pages were connected (completely connected graph)
- The random walk thus defines a Markov chain with probability matrix:

$$P_{ij} = \begin{cases} \lambda \frac{A_{ij}}{\sum_{j=1}^{N} A_{ij}} + (1-\lambda)\frac{1}{N} & \text{if } \sum_{j=1}^{N} A_{ij} \neq 0 \\ \frac{1}{N} & \text{otherwise} \end{cases}$$

where $A_{ij} = 1$ if there is a link from i to j and 0 otherwise

λ is an hyper-parameter, set by user/designer

Short explanation

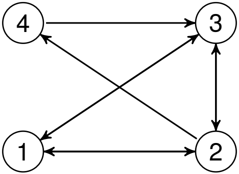
$$\underbrace{\lambda}_{\text{probability of not teleporting}} \underbrace{\times}_{\text{and}} \underbrace{\frac{A_{ij}}{\sum_{j=1}^{N} A_{ij}}}_{\text{of selecting j among outgoing links}}$$

$$\underbrace{+}_{\text{OR}}$$

$$\underbrace{(1-\lambda)}_{\text{probability of teleporting}} \underbrace{\times}_{\text{and}} \underbrace{\frac{1}{N}}_{\text{of choosing j as destination}}$$

Example (1)

Let us consider the following graph

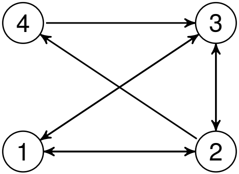


Its adjacency matrix is defined by

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

Example (1)

Let us consider the following graph



Its adjacency matrix is defined by

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

Example (2)

Considering no teleportation ($\lambda = 1$)

$$P_{ij} = \frac{A_{ij}}{\sum_{j=1}^{N} A_{ij}}$$

And

$$P = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

Example (2)

Considering no teleportation ($\lambda = 1$)

$$P_{ij} = \frac{A_{ij}}{\sum_{j=1}^{N} A_{ij}}$$

And

$$P = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

Definitions and notations

Definition 1 A sequence of random variables $X_0, ..., X_n$ is said to be a *(finite state) Markov chain* for some state space S if for any $x_{n+1}, x_n, ..., x_0 \in S$:

$$P(X_{n+1} = x_{n+1} | X_0 = x_0, ..., X_n = x_n) = P(X_{n+1} = x_{n+1} | X_n = x_n)$$

X_0 is called the initial state; $|S| = N$

Definition 2 A Markov chain is called homogeneous or stationary if $P(X_{n+1} = y | X_n = x)$ is independent of n for any (x, y)

Definitions and notations (cont'd)

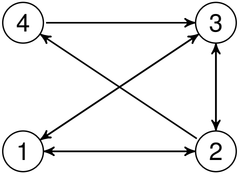
Definition 3 Let $\{X_n\}$ be a stationary Markov chain. The probabilities $P_{ij} = P(X_{n+1} = j | X_n = i)$ are called the *one-step transition probabilities*. The associated matrix P is called the *transition probability matrix*

Definition 4 Let $\{X_n\}$ be a stationary Markov chain. The probabilities $P_{ij}^n = P(X_{n+m} = j | X_m = i)$ are called the *n-step transition probabilities*. The associated matrix P^n is called the *n-step transition probability matrix*

P_{ij}^n is the term at row i and column j of P^n

Illustration

Same graph as before



$S = \{1, 2, 3, 4\}$
$X_n = 1$, or 2, or 3, or 4

Transition probabilities

Remark: P is a stochastic matrix; $\forall i, \sum_{j=1}^{N} P_{ij} = 1$

Example

$$P = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

Theorem (Chapman-Kolgomorov equation) Let $\{X_n\}$ be a stationary Markov chain and $n, m \geq 1$. Then:

$$P_{ij}^{m+n} = P(X_{m+n} = j | X_0 = i) = \sum_{k \in S} P_{ik}^{m} P_{kj}^{n}$$

Regularity (ergodicity)

Definition 5 Let $\{X_n\}$ be a stationary Markov chain with transition probability matrix P. It is called *regular* if there exists $n_0 > 0$ such that $p_{ij}^{n_0} > 0 \ \forall i, j \ \in S$

Example

$$P = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

Is P regular? Is the matrix associated with the random walk with teleportation regular?

Yes to both questions; $n_0 = 3$ in the first case, 1 in the second!

Regularity (ergodicity)

Definition 5 Let $\{X_n\}$ be a stationary Markov chain with transition probability matrix P. It is called *regular* if there exists $n_0 > 0$ such that $p_{ij}^{n_0} > 0 \ \forall i, j \ \in S$

Example

$$P = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

Is P regular? Is the matrix associated with the random walk with teleportation regular?
Yes to both questions; $n_0 = 3$ in the first case, 1 in the second!

Regularity (cont'd)

Theorem (fundamental theorem for finite Markov chains)
Let $\{X_n\}$ be a regular, stationary Markov chain on a state space S of N elements. Then, there exists π_j, $j = 1, 2, ..., N$ such that:

(a) For any initial state i,
$P(X_n = j | X_0 = i) \xrightarrow[n \to +\infty]{} \pi_j$, $j = 1, 2, ..., N$

(b) The row vector $\pi = (\pi_1, \pi_2, ..., \pi_N)$ is the unique solution of the equations $\pi P = \pi, \ \pi \mathbf{1} = 1$

(c) Any row of P^n converges towards π when $n \to \infty$

π is called the long-run or stationary distribution (PageRank)

Let $\mathbf{x}^{(n)}$ denote the probability vector of the walker after n steps $(x_j^{(n)} = P(X_n = j | X_0))$
$\Rightarrow \mathbf{x}^{(n+1)} = \mathbf{x}^{(n)} P$ converges to π (due to (a))

Table of content

Associated algorithms

Three main types

1. Compute $\mathbf{x}^{(n+1)} = \mathbf{x}^{(n)} P$ $(= \mathbf{x}^{(0)} P^{n+1})$ till convergence – power method
2. Compute the left eigenvector of P associated with the eigenvalue 1 (largest eigenvalue of P)
3. Solve the equations $\pi P = \pi, \ \pi \mathbf{1} = 1$ (N equations with N unknowns) – Gauss-Seidel

Complexity

1. For 1, $O(TN^2)$ where T is the number of iterations
2. For 2, $O(N^3)$
3. For 3, $O(T'N^2)$ where T' is the number of iterations

Associated algorithms

Three main types

1. Compute $\mathbf{x}^{(n+1)} = \mathbf{x}^{(n)} P$ $(= \mathbf{x}^{(0)} P^{n+1})$ till convergence – power method
2. Compute the left eigenvector of P associated with the eigenvalue 1 (largest eigenvalue of P)
3. Solve the equations $\pi P = \pi, \ \pi \mathbf{1} = 1$ (N equations with N unknowns) – Gauss-Seidel

Complexity

1. For 1, $O(TN^2)$ where T is the number of iterations
2. For 2, $O(N^3)$
3. For 3, $O(T'N^2)$ where T' is the number of iterations

Power method

Input : adj. matrix A, λ, ϵ (for stopping)
Initialization :

- compute prob. matrix P
- $t \leftarrow 0$, $\mathbf{x}^{(t)} = (\frac{1}{N}, \ldots, \frac{1}{N})$

repeat

$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} P$

$\delta = ||\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}||_2^2$

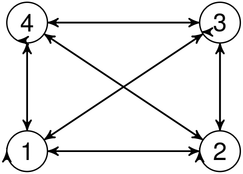
$t \leftarrow t + 1$

until $\delta \leq \epsilon$
Output : PageRank $\mathbf{x}^{(t)}$

Algorithme 1 : Algorithm "power method"

Illustration (1)

Let us consider the following graph (with self loops):



Compute the PageRank of each page with $\lambda = 0.8$

Illsutration (2)

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}, \; P = \begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix} = P^2 = \cdots$$

$$\mathbf{x}^{(0)} P = \begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix} \times \begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix} = \begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix} = \mathbf{x}^{(0)}$$

$$\Rightarrow \pi = \begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}$$

Table of content

Conclusion (1)

1. Stationary, regular Markov chains admit a stationary (steady-stable) distribution
2. This distribution can be obtained in different ways:
   - Power method: let the chain run for a sufficiently long time
   - Linear system: solve the linear system associated with $\pi P = \pi, \ \pi \mathbf{1} = 1$ (*e.g.* Gauss-Seidel)
   - π is the left eigenvector associated with the highest eigenvalue (1) of P (eigenvector decomposition, *e.g.* Cholevsky)

The PageRank can be obtained by any of these methods (power method, Gauss-Seidel are preferred when the graph is large)

Conclusion (2)

Two main innovations at the basis of Web search engines at the end of the 90's:

1. Rely on additional index terms contained in anchor texts
2. Integrate the importance of a web page (PageRank) into the score of a page

The PageRank can be computed to obtain the importance of any node, in any graph!

References

- C. Manning, P. Raghavan, H. Schütze, "Introduction to Information Retrieval", 2008 (https://nlp.stanford.edu/IR-book/information-retrieval-book.html)
- A. DasGupta, "Probability for Statistics and Machine Learning", Springer, 2011

Conclusion (2)

Two main innovations at the basis of Web search engines at the end of the 90's:

1. Rely on additional index terms contained in anchor texts
2. Integrate the importance of a web page (PageRank) into the score of a page

The PageRank can be computed to obtain the importance of any node, in any graph!

References

- C. Manning, P. Raghavan, H. Schütze, "Introduction to Information Retrieval", 2008 (https://nlp.stanford.edu/IR-book/information-retrieval-book.html)
- A. DasGupta, "Probability for Statistics and Machine Learning", Springer, 2011

Conclusion (2)

Two main innovations at the basis of Web search engines at the end of the 90's:

1. Rely on additional index terms contained in anchor texts
2. Integrate the importance of a web page (PageRank) into the score of a page

The PageRank can be computed to obtain the importance of any node, in any graph!

References

- C. Manning, P. Raghavan, H. Schütze, "Introduction to Information Retrieval", 2008 (https://nlp.stanford.edu/IR-book/information-retrieval-book.html)
- A. DasGupta, "Probability for Statistics and Machine Learning", Springer, 2011