WikipediaSearch

An application of Personalized PageRank on a Wikipedia subset

Roberto Corti

Information Retrieval exam

December 8, 2020

UNIVERSITÀ
DEGLI STUDI DI TRIESTE

DATA SCIENCE &
SCIENTIFIC COMPUTING

Outline

Introduction

Wikipedia dataset

WikipediaSearch implementation

WikipediaSearch evaluation

Conclusion

Outline

Introduction

Wikipedia dataset

WikipediaSearch implementation

WikipediaSearch evaluation

Conclusion

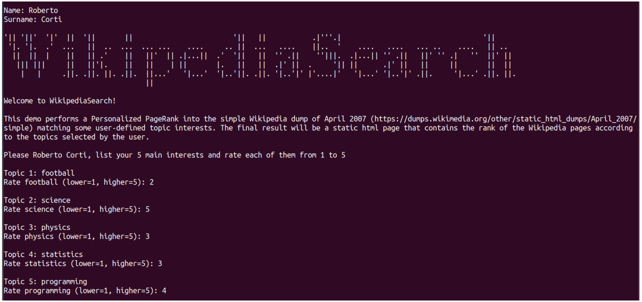
WikipediaSearch

A brief introduction

WikipediaSearch is a user-interactive tool that computes a Personalized (or Topic Specific) PageRank over a Wikipedia corpus.

WikipediaSearch

A brief introduction

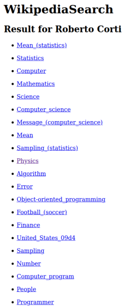


```
Name: Roberto
Surname: Corti

'|| '||'  '|'  ||  '||      ||                  '||   ||        .|''.|                                      '||
 '|. '|.  .'  ...   ||  ..  ...  ... ...    ....    .. ||  ...   ....    ||..  '    ....   ....   ... ..    ....   || ..
  ||  ||  |    ||   || .'    ||   ||'  || .|...|| .'  '||   ||  '' .||    ''|||.  .|...|| '' .||   ||' '' .|   ''  ||' ||
   ||| |||     ||   ||'|.    ||   ||    | ||      |.   ||   ||  .|' ||  .     '|| ||      .|' ||   ||     ||       ||  ||
    |   |     .||. .||. ||. .||.  ||...'   '|...'  '|..'||. .||. '|..'|' |'....|'   '|...' '|..'|' .||.    '|...' .||. ||.
                                  ||

Welcome to WikipediaSearch!

This demo performs a Personalized PageRank into the simple Wikipedia dump of April 2007 (https://dumps.wikimedia.org/other/static_html_dumps/April_2007/
simple) matching some user-defined topic interests. The final result will be a static html page that contains the rank of the Wikipedia pages according
to the topics selected by the user.

Please Roberto Corti, list your 5 main interests and rate each of them from 1 to 5

Topic 1: football
Rate football (lower=1, higher=5): 2

Topic 2: science
Rate science (lower=1, higher=5): 5

Topic 3: physics
Rate physics (lower=1, higher=5): 3

Topic 4: statistics
Rate statistics (lower=1, higher=5): 3

Topic 5: programming
Rate programming (lower=1, higher=5): 4
```

Input interface: user specifies the topics in which he/she has more interest

WikipediaSearch

A brief introduction



WikipediaSearch

Result for Roberto Corti

- Mean_(statistics)
- Statistics
- Computer
- Mathematics
- Science
- Computer_science
- Message_(computer_science)
- Mean
- Sampling_(statistics)
- Physics
- Algorithm
- Error
- Object-oriented_programming
- Football_(soccer)
- Finance
- United_States_09d4
- Sampling
- Number
- Computer_program
- People
- Programmer

Output page: a rank of Wikipedia articles in which the user would be interested to read

The problem

"Bringing Order to the Web"

Algorithm developed by L. Page and S. Brin (Google co-founders) used to determine the order of web pages

Problem: Is there a rating system that could measure the human interest and attention devoted to web pages?

The problem

”Bringing Order to the Web”

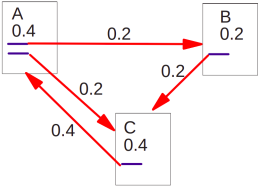
Algorithm developed by L. Page and S. Brin (Google co-founders) used to determine the order of web pages

Problem: Is there a rating system that could measure the human interest and attention devoted to web pages?

PageRank idea: modeling the behavior of a ”random surfer” in the Web graph.

A recap of PageRank theory

From L.Page and S.Brin (1998)



$\vec{x}$: probability distribution vector of the nodes

M : square, stochastic matrix corresponding to the directed grap where $M_{ij} = 1/N_j$

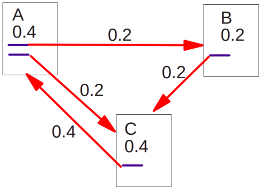
$\Rightarrow$ Find the stationary probability distribution $\Leftrightarrow$ find the unique stochastic eigenvector of to the eigenvalue 1:

$$M\vec{\pi} = \vec{\pi}$$

$\vec{\pi} =$ PageRank vector

A recap of PageRank theory

From L.Page and S.Brin (1998)



Power method solution:

- Start with $\vec{x}_0 = \texttt{random}()$
- $\vec{x}_i = M\vec{x}_{i-1}$

  until $|\vec{x}_i - \vec{x}_{i-1}| < \epsilon$

A recap of PageRank theory

From L.Page and S.Brin (1998)

Two problems:

- **Dangling nodes**: pages without outgoing edges. How to assign probability to them?
- **Pages without incoming or outgoing links** : node without incoming edges or group of nodes without outgoing edges. For the first ones we have probability 0 of returning to it once we leave it, while for the others we can never leave them once entered

A recap of PageRank theory

From L.Page and S.Brin (1998)

Allow the random surfer to move to a random page of the graph with probability α.

The stochastic matrix will be:

$$\begin{aligned} M' &= (1-\alpha)M + \alpha \left[\frac{1}{N}\right]_{N\times N} \\ &= (1-\alpha)M + \alpha \vec{1}^T \cdot \vec{J} \end{aligned}$$

where $\vec{J} = (1/N, ..., 1/N)$ is the *jump vector*

A recap of Topic-Sensitive PageRank

From Taher H. Haveliwala (2003)

In addition to the PageRank calculation we can *specialize* the scores of the pages by limiting them to a single topic.

For a given topic-specific set of pages S, we allow the random surfer to teleport only to pages that are inside S

A recap of Topic-Sensitive PageRank

From Taher H. Haveliwala (2003)

In addition to the PageRank calculation we can *specialize* the scores of the pages by limiting them to a single topic.

For a given topic-specific set of pages S, we allow the random surfer to teleport only to pages that are inside S

$$M' = (1 - \alpha)M + \alpha\vec{1}^T \cdot \vec{J_S}$$

where the *topic-specific jump vector* is defined as

$$J_{S_i} = \begin{cases} 1/|S|, & \text{if page } i \in S. \\ 0, & \text{otherwise.} \end{cases}$$

Outline

Introduction

Wikipedia dataset

WikipediaSearch implementation

WikipediaSearch evaluation

Conclusion

Wikipedia ...

content...

Outline

Introduction

Wikipedia dataset

WikipediaSearch implementation

WikipediaSearch evaluation

Conclusion

PageRank implementation

Outline

Introduction

Wikipedia dataset

WikipediaSearch implementation

WikipediaSearch evaluation

Conclusion

Standard PageRank

Outline

Introduction

Wikipedia dataset

WikipediaSearch implementation

WikipediaSearch evaluation

Conclusion

Conclusion