Comparative Analysis of Web Search and Ranking Algorithms

Mihir Kelkar

University of Maryland, Baltimore County. CMSC 676

May 6, 2014

Introduction

Overview of the Talk

Introduction

Overview of the Talk

- Brief History of Search

Introduction

Overview of the Talk

- Brief History of Search
- Impact of Web Search Algorithm on our lives

Introduction

Overview of the Talk

- Brief History of Search
- Impact of Web Search Algorithm on our lives
- The Google Page Rank Algorithm

Introduction

Overview of the Talk

- Brief History of Search
- Impact of Web Search Algorithm on our lives
- The Google Page Rank Algorithm
- The HITS Algorithm

Introduction

Overview of the Talk

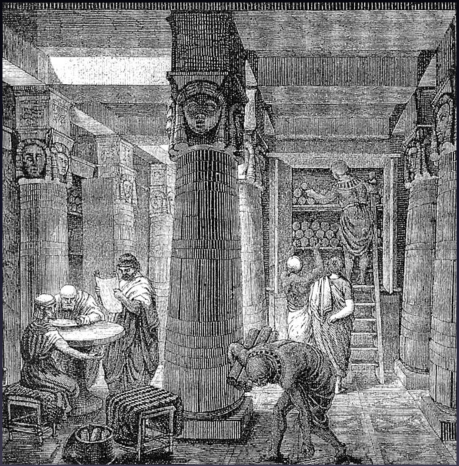
- Brief History of Search
- Impact of Web Search Algorithm on our lives
- The Google Page Rank Algorithm
- The HITS Algorithm
- Facebook's Edge Rank Algorithm

Introduction

Overview of the Talk

- Brief History of Search
- Impact of Web Search Algorithm on our lives
- The Google Page Rank Algorithm
- The HITS Algorithm
- Facebook’s Edge Rank Algorithm
- Reddit’s Story Ranking Algorithm

Brief History of Search



(a) Search in 300 BC

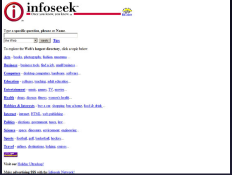
Brief History of Search

(b) Card Catalogs

Brief History of Search

(c) Archie

(d) Wandex



Type a **specific question**, **phrase** or **Name**.

To explore the **Web's largest directory**, click a topic below.

Arts - books, photography, fashion, museums ...

Business - business tools, find a job, small business...

Computers - desktop computers, hardware, software...

Education - colleges, teaching, adult education...

Entertainment - music, games, TV, movies...

Health - drugs, disease, fitness, women's health...

Hobbies & Interests - buy a car, shopping, buy a home, food & drink ...

Internet - intranet, HTML, web publishing...

Politics - elections, government, taxes, law...

Science - space, dinosaurs, environment, engineering...

Sports - football, golf, basketball, hockey...

Travel - airlines, destinations, lodging, cruises...

Visit our Holiday Ultrashop!

Make **advertising $$$** with the Infoseek Network!

(e) Infoseek

Brief History of Search

(f) Altavista

(g) Yahoo Search

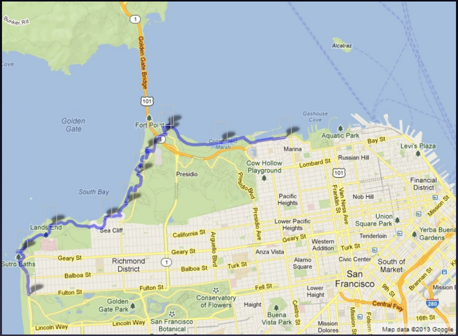
(h) Google Search

(i) Ask

(j) MSN

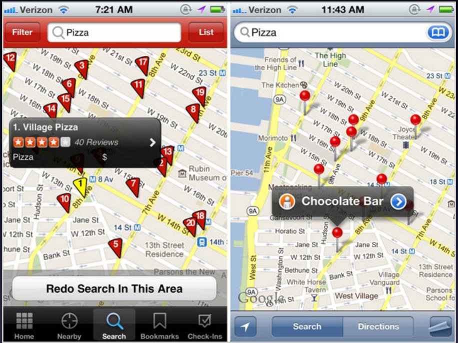
(k) Microsoft Bing

Impact of Search Algorithms



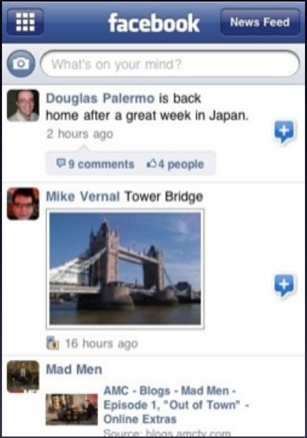
(I) Navigation is essentially using a search

Impact of Search Algorithms



(m) Search Algorithms have a major impact on how you behave

Impact of Search Algorithms



Douglas Palermo is back home after a great week in Japan.
2 hours ago
9 comments 4 people

Mike Vernal Tower Bridge
16 hours ago

Mad Men
AMC - Blogs - Mad Men - Episode 1, "Out of Town" - Online Extras
Source: blogs.amctv.com

(n) Facebook's search feature actually modifies the stories that you see on top

The Google Page Rank Algorithm

- PageRank is a "vote", by all the other pages on the Web, about how important a page is.

The Google Page Rank Algorithm

- PageRank is a "vote", by all the other pages on the Web, about how important a page is.
- The World Wide Web can be visualized as a highly interconnected graph with directed edges

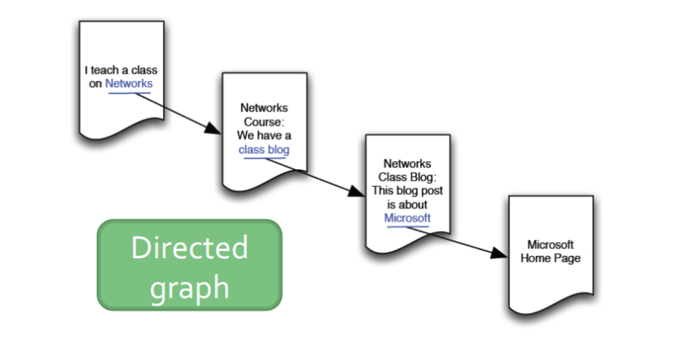
The Google Page Rank Algorithm

- PageRank is a "vote", by all the other pages on the Web, about how important a page is.
- The World Wide Web can be visualized as a highly interconnected graph with directed edges
- Not all pages on the internet are equally "important"

The Google Page Rank Algorithm

- PageRank is a "vote", by all the other pages on the Web, about how important a page is.
- The World Wide Web can be visualized as a highly interconnected graph with directed edges
- Not all pages on the internet are equally "important"
- The more important pages you "cite" your page’s content becomes that much more "credible and important"

The Google Page Rank Algorithm


I teach a class on Networks
Networks Course: We have a class blog
Networks Class Blog: This blog post is about Microsoft
Microsoft Home Page
Directed graph

(o) A page links to various other pages, this forms an directed graph

The Google Page Rank Algorithm

- Webpages are not equally "important".

The Google Page Rank Algorithm

- Webpages are not equally "important".
- A page is important if its connected to other pages by links.

The Google Page Rank Algorithm

- Webpages are not equally "important".
- A page is important if its connected to other pages by links.
- An incoming link to page X stands as a vote of importance for page X

The Google Page Rank Algorithm

- Webpages are not equally "important".
- A page is important if its connected to other pages by links.
- An incoming link to page X stands as a vote of importance for page X
- www.stanford.edu has about 24000 incoming links whereas www.notdecided.com has just 1.

The Google Page Rank Algorithm

- Webpages are not equally "important".
- A page is important if its connected to other pages by links.
- An incoming link to page X stands as a vote of importance for page X
- www.stanford.edu has about 24000 incoming links whereas www.notdecided.com has just 1.
- Think of the incoming link as a sort of "commendation". So, an incoming link from different pages has different weights associated with it. The more important the source page, the more weight its outgoing link carries

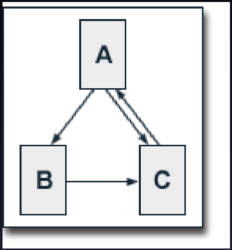
The Google Page Rank Algorithm

- **PR(A) = (1 - d) + d(PR(T$_1$) / C(T$_1$) + ... PR(T$_n$) / C(T$_n$))**

The Google Page Rank Algorithm

- **PR(A) = (1 - d) + d(PR(T$_1$) / C(T$_1$) + ... PR(T$_n$) / C(T$_n$))**
- *PR(A) is the notation for page rank of A*
- *PR(T$_i$) - Page Rank of pages Ti which link to page A*
- *C(T$_i$) - Number of outbound links on Page T$_i$*
- *d - dampling factor which can have values between the range 0 and 1*

The Google Page Rank Algorithm



(p)

- PR(A) = 0.5 + 0.5 PR(C)
- PR(B) = 0.5 + 0.5 (PR(A) / 2)
- PR(C) = 0.5 + 0.5 (PR(A) / 2 + PR(B))

The Google Page Rank Algorithm

- Because of the size of the actual web, Google uses an approximative, iterative computation of PageRank values.

Iteration	PR(A)	PR(B)	PR(C)
0	1	1	1
1	1	0.75	1.125
2	1.0625	0.765625	1.1484375
3	1.07421875	0.76855469	1.15283203
4	1.07641602	0.76910400	1.15365601
5	1.07682800	0.76920700	1.15381050
6	1.07690525	0.76922631	1.15383947
7	1.07691973	0.76922993	1.15384490
8	1.07692245	0.76923061	1.15384592
9	1.07692296	0.76923074	1.15384611
10	1.07692305	0.76923076	1.15384615

The Hyperlink Induced Topic Search Algorithm

- Hyperlink Induced Topic Search Algorithm (HITS) was developed almost at the same time as PageRank and was adopted by and used by ask.com for several years.

The Hyperlink Induced Topic Search Algorithm

- Hyperlink Induced Topic Search Algorithm (HITS) was developed almost at the same time as PageRank and was adopted by and used by ask.com for several years.
- Uses hubs and authorities to define a **recursive relationship** between web pages.

The Hyperlink Induced Topic Search Algorithm

- Hyperlink Induced Topic Search Algorithm (HITS) was developed almost at the same time as PageRank and was adopted by and used by ask.com for several years.
- Uses hubs and authorities to define a **recursive relationship** between web pages.
- Each webpage has a hub and authority scores. Some webpages which are extremely reliable might be identified as authorities on their topic. For eg. IRS's website is an authority page pertaining to query terms realted to tax rules in the United States.

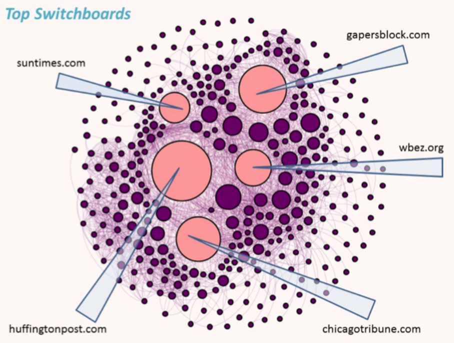
The Hyperlink Induced Topic Search Algorithm

- Hyperlink Induced Topic Search Algorithm (HITS) was developed almost at the same time as PageRank and was adopted by and used by ask.com for several years.
- Uses hubs and authorities to define a **recursive relationship** between web pages.
- Each webpage has a hub and authority scores. Some webpages which are extremely reliable might be identified as authorities on their topic. For eg. IRS's website is an authority page pertaining to query terms realted to tax rules in the United States.
- An authority is a page that many hubs link to

The Hyperlink Induced Topic Search Algorithm

- Hyperlink Induced Topic Search Algorithm (HITS) was developed almost at the same time as PageRank and was adopted by and used by ask.com for several years.
- Uses hubs and authorities to define a **recursive relationship** between web pages.
- Each webpage has a hub and authority scores. Some webpages which are extremely reliable might be identified as authorities on their topic. For eg. IRS's website is an authority page pertaining to query terms realted to tax rules in the United States.
- An authority is a page that many hubs link to
- A hub is a page that links to many authorities.

The Hyperlink Induced Topic Search Algorithm



(q) News Websites in the Chicago area represented as Hubs and Authorities

The Hyperlink Induced Topic Search Algorithm

- To begin the Ranking process, the authority and hub scores of all nodes is assigned as 1

The Hyperlink Induced Topic Search Algorithm

- To begin the Ranking process, the authority and hub scores of all nodes is assigned as 1
- The Authority score for a node is updated according to the authority update rule : auth(x) = $\sum_{i=1}^{n}$ hub(i)

The Hyperlink Induced Topic Search Algorithm

- To begin the Ranking process, the authority and hub scores of all nodes is assigned as 1
- The Authority score for a node is updated according to the authority update rule : auth(x) = $\sum_{i=1}^{n}$ hub(i)
- Similarly, the hub score for a node is update according to the hub update rule : hub(x) = $\sum_{i=1}^{n}$ auth(i)

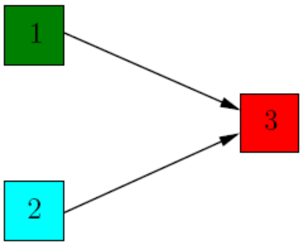
The Hyperlink Induced Topic Search Algorithm

- To begin the Ranking process, the authority and hub scores of all nodes is assigned as 1
- The Authority score for a node is updated according to the authority update rule : auth(x) = $\sum_{i=1}^{n}$ hub(i)
- Similarly, the hub score for a node is update according to the hub update rule : hub(x) = $\sum_{i=1}^{n}$ auth(i)
- A k-step application of the Hub-Authority algorithm entails applying for k times first the Authority Update Rule and then the Hub Update Rule.

The Hyperlink Induced Topic Search Algorithm

- To begin the Ranking process, the authority and hub scores of all nodes is assigned as 1
- The Authority score for a node is updated according to the authority update rule : auth(x) = $\sum_{i=1}^{n}$ hub(i)
- Similarly, the hub score for a node is update according to the hub update rule : hub(x) = $\sum_{i=1}^{n}$ auth(i)
- A k-step application of the Hub-Authority algorithm entails applying for k times first the Authority Update Rule and then the Hub Update Rule.
- The values are normalized to make sure that they remain converging.

The Hyperlink Induced Topic Search Algorithm



(r)

The Adjacency matrix for this graph can be represented as follows:

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

The Hyperlink Induced Topic Search Algorithm

A = $\begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$ hence A^{t} = $\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix}$.

Also Assume that the initial hub vector is u = $\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$.

The Hyperlink Induced Topic Search Algorithm

We compute the Authority weight vector as A^t.u

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

We compute the Hub weight vector as A.A^t.u

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

This somewhat already corresponds with our intuition that node 3 must be authoritative

Facebook's Edge Rank Algorithm

- Edge Rank is the algorithm developed by Facebook to determine what stories should be displayed and how high should tey be ranked in a user's news feed.

Facebook’s Edge Rank Algorithm

- Edge Rank is the algorithm developed by Facebook to determine what stories should be displayed and how high should tey be ranked in a user’s news feed.
- The entire userbase of social media websites can be visualized as a directed graphs with users as nodes and internatctions between them as edges

Facebook’s Edge Rank Algorithm

- Edge Rank is the algorithm developed by Facebook to determine what stories should be displayed and how high should tey be ranked in a user’s news feed.
- The entire userbase of social media websites can be visualized as a directed graphs with users as nodes and internatctions between them as edges
- The deciding factor about which stories should significantly appear on your news feed is how often you "interact" with the person who is the source / participant in the story.

Facebook’s Edge Rank Algorithm

The three main pillars of Edge Rank are:

Facebook’s Edge Rank Algorithm

The three main pillars of Edge Rank are:

- User-Affinity represented as u_e

Facebook’s Edge Rank Algorithm

The three main pillars of Edge Rank are:

- User-Affinity represented as u_e
- Weight represented as w

Facebook's Edge Rank Algorithm

The three main pillars of Edge Rank are:

- User-Affinity represented as u_e
- Weight represented as w
- Time Decay : represented as T_d

Facebook's Edge Rank Algorithm

Explanation of User Affinity : u_e

- Affinity is a one way relationship between one user and another

Facebook's Edge Rank Algorithm

Explanation of User Affinity : u_e

- Affinity is a one way relationship between one user and another
- Simply put afinity is a mesure of how closely related a user is to another user. (Not vice versa though)

Facebook's Edge Rank Algorithm

Explanation of User Affinity : u_e

- Affinity is a one way relationship between one user and another
- Simply put afinity is a mesure of how closely related a user is to another user. (Not vice versa though)
- Affinity is built by interactions like comenting, liking, private messaging and even views on other's stories and pages.

Facebook's Edge Rank Algorithm

Explanation of User Affinity : u_e

- Affinity is a one way relationship between one user and another
- Simply put afinity is a mesure of how closely related a user is to another user. (Not vice versa though)
- Affinity is built by interactions like comenting, liking, private messaging and even views on other's stories and pages.
- Affinity can however also be directly decalred, by lising someone as directly related to you. Eg. A Brother, A parent or a spouse.

Facebook's Edge Rank Algorithm

Explanation of Weight : *W*

- All kinds of interactions have differnet weights associated with them

Facebook’s Edge Rank Algorithm

Explanation of Weight : W

- All kinds of interactions have differnet weights associated with them
- An interaction like commenting has a higher weight over an interaction like simply liking the story since commenting needs the user to be more involved in the story in general

Facebook’s Edge Rank Algorithm

Explanation of Weight : W

- All kinds of interactions have differnet weights associated with them
- An interaction like commenting has a higher weight over an interaction like simply liking the story since commenting needs the user to be more involved in the story in general
- As a generalization, the more time consuming a method of interaction, the more weight it carries

Facebook's Edge Rank Algorithm

Explanation of Time Decay : T_d

- Each story has an associated "lifetime", the older it gets, the lower it gets ranked

Facebook's Edge Rank Algorithm

Explanation of Time Decay : T_d

- Each story has an associated "lifetime", the older it gets, the lower it gets ranked
- However, Time decay also considers time since last interaction between the two nodes.

Facebook's Edge Rank Algorithm

Explanation of Time Decay : T_d

- Each story has an associated "lifetime", the older it gets, the lower it gets ranked
- However, Time decay also considers time since last interaction between the two nodes.
- Thus for someone who logs in very irregularly, older stories still appear as Top Ranked stories. However for someone who logs in frequently, top ranked stories change faster.

Reddit's Story Ranking Algorithm

Reddit's Story Ranking Algorithm

- Reddit has much of its code available publicly

Reddit's Story Ranking Algorithm

Reddit's Story Ranking Algorithm

- Reddit has much of its code available publicly
- The Story ranking algorithm is implemented in Pyrex.

Reddit's Story Ranking Algorithm

Reddit's Story Ranking Algorithm

- Reddit has much of its code available publicly
- The Story ranking algorithm is implemented in Pyrex.
- Reddit's story ranking Algorithm takes submission time, number of upvotes / number of downvotes into consideration.

Reddit's Story Ranking Algorithm

- T$_s$ = Time difference between the current time and the posting time

Reddit’s Story Ranking Algorithm

- T$_s$ = Time difference between the current time and the posting time
- x : Difference between number of Upvotes and number of Downvotes

Reddit's Story Ranking Algorithm

- T$_s$ = Time difference between the current time and the posting time
- x : Difference between number of Upvotes and number of Downvotes
- Reddit's story ranking Algorithm takes submission time, number of upvotes / number of downvotes into consideration.

Reddit’s Story Ranking Algorithm

- T$_s$ = Time difference between the current time and the posting time
- x : Difference between number of Upvotes and number of Downvotes
- Reddit’s story ranking Algorithm takes submission time, number of upvotes / number of downvotes into consideration.
- The value for a third parameter Y is decided amongst (1 , -1, 0) upon whether the net upvote value is +ve, -ve or 0.

Reddit's Story Ranking Algorithm

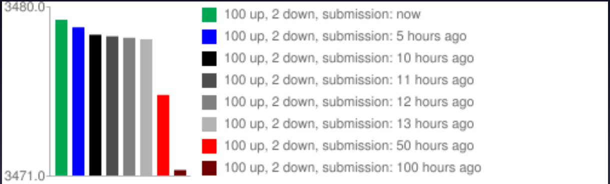
- T_s = Time difference between the current time and the posting time
- x : Difference between number of Upvotes and number of Downvotes
- Reddit's story ranking Algorithm takes submission time, number of upvotes / number of downvotes into consideration.
- The value for a third parameter Y is decided amongst (1 , -1, 0) upon whether the net upvote value is +ve, -ve or 0.
- The value of a fourth parameter z is max (|x| , 1)

Reddit's Story Ranking Algorithm

- The Ranking function for Reddit's stories
  f(T_s, y, z) = Log(z) + yT_s / 45000.

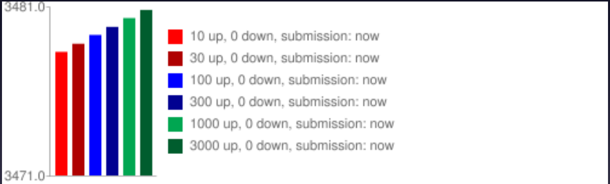
Reddit's Story Ranking Algorithm

- The Ranking function for Reddit's stories
  $f(T_s, y, z)$ = Log(z) + yT_s / 45000.

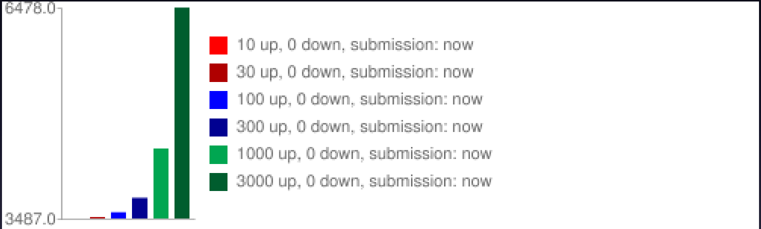


Reddit's Story Ranking Algorithm

- The Ranking function uses Logarithm to weight the total number of upvotes. I believe that this is done so to make sure that the initial few votes count higher than the rest.

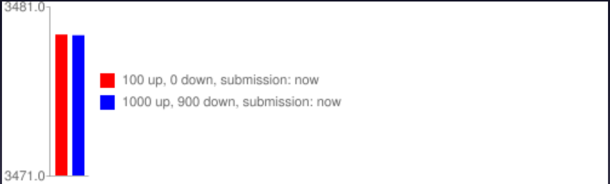


Reddit's Story Ranking Algorithm



Reddit's Story Ranking Algorithm

- Reddit is one of the few sites which have the brutal downvote button. Downvotes can significantly affect a story's rank



Reddit’s Story Ranking Algorithm

- Submission time is a very important parameter, generally newer stories will rank higher than older

Reddit's Story Ranking Algorithm

- Submission time is a very important parameter, generally newer stories will rank higher than older
- The first 10 upvotes count as high as the next 100. E.g. a story that has 10 upvotes and a story that has 50 upvotes will have a similar ranking

Reddit’s Story Ranking Algorithm

- Submission time is a very important parameter, generally newer stories will rank higher than older
- The first 10 upvotes count as high as the next 100. E.g. a story that has 10 upvotes and a story that has 50 upvotes will have a similar ranking
- Controversial stories that get similar amounts of upvotes and downvotes will get a low ranking compared to stories that mainly get upvotes

Sources and Citations

- The PageRank Citation Ranking:Bringing Order to the Web
- Authoritative Sources in a Hyperlinked Environment
- Reddit Engineering Blog