

# **How Google Searches Work**

History of searching and the PageRank algorithm Martin Thoma, Benjamin Lipp | 7th of February, 2013

# SPRACHENZENTRUM

## **Contents**



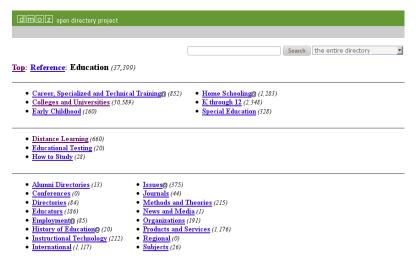
- Introduction
- 2 PageRank
- End



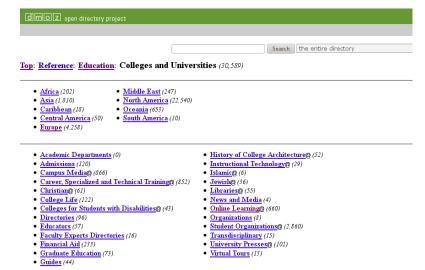


5,114,083 sites - 96,877 editors - over 1,014,849 categories











dmoz open directory project				
		about dmoz   dmoz blog   sugges		
		earch the entire directory		
<u>Top</u> : <u>Reference</u> : <u>Education</u> : <u>Colleges and Universities</u> : <u>Europe</u> (4,258)				
Academic Departments (0)				
• Albania (8)	• <u>Lithuania</u> (9)			
• <u>Austria</u> (54)	• <u>Luxembourg</u> (1)			
• <u>Belarus</u> (11) • <u>Belgium</u> (24)	• <u>Macedonia</u> (19) • <u>Malta</u> (14)			
• Bosnia and Herzegovina (8)	• Moldova (1)			
• Bulgaria (30)	• Montenegro (1)			
• Croatia (18)	• Netherlands (30)			
• Cyprus@ (1)	• Norway (160)			
Czech Republic (59)	• Poland (175)			
• Denmark (11)	• Portugal (6)			
• Estonia (23)	• Romania (18)			
• Finland (23)	• Russia (78)			
• France (19)	• <u>Serbia</u> (6)			
• Germany (121)	Slovakia (12)			
• <u>Greece</u> (22)	Slovenia (7)			
• Hungary (16)	• Spain (20)			
<ul> <li><u>Iceland</u> (4)</li> </ul>	• <u>Sweden</u> (56)			



dmoz open directory proje	ot			
	about dmoz			
	Search the entire directory			
<u>Top:</u> <u>Reference:</u> <u>Education:</u> <u>Colleges and Universities:</u> <u>Europe:</u> <u>Germany</u> (121)				
	$(\underline{A} \underline{B} \underline{C} \underline{D} \underline{E} \underline{F} \underline{G} \underline{H} \underline{I} \underline{J} \underline{K} \underline{L} \underline{M} \underline{N} \underline{O} \underline{P} \underline{O} \underline{R} \underline{S} \underline{T} \underline{U} \underline{V} \underline{W} \underline{X} \underline{Y} \underline{V} \underline{V} \underline{V} \underline{V} \underline{V} \underline{V} \underline{V} V$			
• Baden-Württemberg (15) • Bavaria (10) • Berlin (25) • Brandenburg (5) • Bremen (2) • Hanburg (3) • Hesse (7) • Lower Saxony (3)	• Mecklenburg-Western Pomerania (0) • North Rhine-Westphalia (23) • Rhineland-Palatinate (4) • Saarland (2) • Saxony-Anhalt (0) • Schleswig-Holstein (2) • Thuringia (5)			
See also:				
• Regional: Europe: Germany (3,998) • Regional: Europe: Germany: Education (11)				
This category in other languages:				
French (6) Gen	man (2,229)			



dmoz open directory project	
	about dmoz   dmoz blog   sugge
	Search the entire directory
<u>Top</u> : <u>Reference</u> : <u>Education</u> : <u>Colleges and Universities</u> : <u>Eu</u>	rope: Germany: Baden-Württemberg (15)
Hochschule Aalen - University of Applied Sciences (1) Hochschule Esslingen - University of Applied Sciences (1) Hochschule Furtwangen - University of Applied Sciences (0) Hochschule Heilbronn - Heilbronn University (1) Hochschule Konstaur - University of Applied Sciences (0) International University in Germany (1) Stuttgart Institute of Management and Technology (1) University of Freiburg (1)	University of Heidelberg (1)     University of Hohenheim (1)     University of Konstanz (2)     University of Mannheim (1)     University of Mannheim (1)     University of Maryland University College - Mannheim     University of Stuttgart (1)     University of Ulm (1)
See also:	
• Regional: Europe: Germany: States: Baden-Württemberg: I	Education (1)
This category in other languages:  German (247)	
"Baden-Württemberg" search on:     AO:	L - Ask - Bing - Gigablast - Google - Lycos - Yahoo - Yipuv



dmoz open directory project	
	about dmoz   dmoz blog   sugg
	Search the entire directory
<u>Top</u> : <u>Reference</u> : <u>Education</u> : <u>Colleges and Universities</u> :	Europe: Germany: Baden-Württemberg: University of .
See also:	
Regional: Europe: Germany: States: Baden-Württembe	erg: Localities: Mannheim (2)
This category in other languages: German (7)	
<u>University of Mannheim</u> - Offers undergraduate and graduate profinancial aid.	ograms, includes information on study and research at the university, directorie
• "University of Mannheim" search on:	AOL - Ask - Bing - Gigablast - Google - Lycos - Yahoo - Yippy
	<u>Volunteer</u> to edit this category.
Become an Editor Help build the largest human-edited directory of the web	



#### Search: uni mannhoim

#### Open Directory Categories (1-5 of 100)

- World: Français: Régional: Amérique: Etats-Unis: Etat et politique: Ambassades et consulats: Pays représentés (39)
- World: Deutsch: Wissen: Bibliotheken: Universitäts- und Hochschulbibliotheken: Deutschland (33)
- 3. World: Français: Régional: Amérique: Etats-Unis: Etat et politique: Ambassades et consulats: Représentations à l'étranger (24)
- 4. World: Nederlands: Maatschappij: Overheid: Europese Unie (22)
- World: Deutsch: Gesundheit: Krankenhäuser und Kliniken: Universitätskliniken: Deutschland (21)

more...

#### Open Directory Sites (1-20 of 15698)

- 1. amnesty international Hochschulgruppe Mannheim Die amnesty-Gruppe an der Uni Mannheim setzt sich unter anderem mit Petitionen, Appellbrief ein und stellt aktuelle Aktivitäten vor
  - -- http://www.amnesty-uni-mannheim.de/ World: Deutsch: Regional: Europa: Deutschland: Baden-Württemberg: Städte und Gemeinden: M. Mannheim: Gesellschaft (23)
- LHG Liberale Hochschulgruppe Mannheim Die liberale Studierenden an der Uni Mannheim stellen sich vor. -- http://www.uni-mannhaim.de/studorg/liberale/ World: Deutsch: Wissen: Bildung: Hochschulen: Europa: Deutschland: Baden-Württemberg: Universität Mannhaim ()
- Uni Mannheim Marktübersicht deutscher Anbieter für Online-Preisvergleiche. -- http://projekt.wifo.uni-mannheim.de/preisvergleich/ World: Deutsch: Zuhause: Verbraucherinformationen: Preisagenhuren: Ontine-Preisvergleiche ()
- 4. University of Mannheim Offers undergraduate and graduate programs; includes information on study and research at the university, directories of financial aid
  - -- http://www.uni-mannheim.de/ Reference: Education: Colleges and Universities: Europe: Germany: Baden-Württemberg: University of Mannheim ()

# The Early Days: Web Crawlers





- crawls through the web using hyperlinks
- makes an index of the words contained in a page
- ranks pages for a search query according to number of occurences of keywords

# The Early Days: Web Crawlers





- crawls through the web using hyperlinks
- makes an index of the words contained in a page
- ranks pages for a search query according to number of occurences of keywords

# The Early Days: Web Crawlers





- crawls through the web using hyperlinks
- makes an index of the words contained in a page
- ranks pages for a search query according to number of occurences of keywords



- Humans know what is good for them
- Humans will only link to Websites they like



- Humans know what is good for them
- Humans create Websites
- Humans will only link to Websites they like
- ⇒ Hyperlinks are a quality indicator



- Humans know what is good for them
- Humans create Websites
- Humans will only link to Websites they like
- ⇒ Hyperlinks are a quality indicator



- Humans know what is good for them
- Humans create Websites
- Humans will only link to Websites they like
- $\Rightarrow$  Hyperlinks are a quality indicator



- Simply count number of links to a Website
- X 10,000 links from only one page
- Count numbers of Websites that link to a Website
- X Quality of the page matters
- X Total number of links on the source page matters



- Simply count number of links to a Website
- X 10,000 links from only one page
- Count numbers of Websites that link to a Website
- X Quality of the page matters
- X Total number of links on the source page matters



- Simply count number of links to a Website
- X 10,000 links from only one page
- Count numbers of Websites that link to a Website
- X Quality of the page matters
- X Total number of links on the source page matters



- Simply count number of links to a Website
- X 10,000 links from only one page
- Count numbers of Websites that link to a Website
- X Quality of the page matters
- X Total number of links on the source page matters



- Simply count number of links to a Website
- X 10,000 links from only one page
- Count numbers of Websites that link to a Website
- X Quality of the page matters
- X Total number of links on the source page matters

#### A brilliant idea







Sergey Brin

Larry Page



- Decisions of humans are complicated
- A lot of webpages get visited
- ⇒ modellize clicks on links as random behaviour
- Links are important
  - Links of page A get less important, if A has many links
  - Links of page A get more important, if many link to A
- $\Rightarrow$  if B has a link from A, the rank of B increases by  $\frac{Rank(A)}{Links(A)}$

$$Rank(B) += \frac{Rank(A)}{Links(A)}$$



- Decisions of humans are complicated
- A lot of webpages get visited
- Links are important

if A links to B then
$$Rank(B) += \frac{Rank(A)}{Links(A)}$$

$$Rank(B) += \frac{Rank(A)}{Links(A)}$$



- Decisions of humans are complicated
- A lot of webpages get visited
- ⇒ modellize clicks on links as random behaviour
- Links are important
  - Links of page A get less important, if A has many links of page A get more important, if many link to A

0000000

 $\Rightarrow$  if B has a link from A, the rank of B increases by  $rac{Rank(A)}{Links(A)}$ 

if A links to B then
$$Rank(B) += \frac{Rank}{Links}$$



- Decisions of humans are complicated
- A lot of webpages get visited
- ⇒ modellize clicks on links as random behaviour
  - Links are important
    - Links of page A get less important, if A has many links
    - Links of page A get more important, if many link to A
- $\Rightarrow$  if B has a link from A, the rank of B increases by  $\frac{Rank(A)}{Links(A)}$

if A links to B then 
$$Rank(B) += \frac{Rank(A)}{Links(A)}$$



- Decisions of humans are complicated
- A lot of webpages get visited
- ⇒ modellize clicks on links as random behaviour
  - Links are important
    - Links of page A get less important, if A has many links
    - Links of page A get more important, if many link to A
- $\Rightarrow$  if B has a link from A, the rank of B increases by  $\frac{Rank(A)}{Links(A)}$

if A links to B then 
$$Rank(B) += \frac{Rank(A)}{Links(A)}$$



- Decisions of humans are complicated
- A lot of webpages get visited
- ⇒ modellize clicks on links as random behaviour
  - Links are important
    - Links of page A get less important, if A has many links
    - Links of page A get more important, if many link to A
- $\Rightarrow$  if B has a link from A, the rank of B increases by  $\frac{Rank(A)}{Links(A)}$

if A links to B then 
$$Rank(B) += \frac{Rank(A)}{Links(A)}$$



- Decisions of humans are complicated
- A lot of webpages get visited
- ⇒ modellize clicks on links as random behaviour
  - Links are important
    - Links of page A get less important, if A has many links
    - Links of page A get more important, if many link to A
- $\Rightarrow$  if B has a link from A, the rank of B increases by  $\frac{Rank(A)}{Links(A)}$

if A links to B **then**

$$Rank(B) += \frac{Rank(A)}{Links(A)}$$



- Decisions of humans are complicated
- A lot of webpages get visited
- ⇒ modellize clicks on links as random behaviour
  - Links are important
    - Links of page A get less important, if A has many links
    - Links of page A get more important, if many link to A
- $\Rightarrow$  if B has a link from A, the rank of B increases by  $\frac{Rank(A)}{Links(A)}$

#### if A links to B then

$$Rank(B) += \frac{Rank(A)}{Links(A)}$$



- Websites = nodes = anthill
- Links = edges = paths
- You place ants on each node
- They walk over the paths (at random, they are ants!)
- After some time, some anthills will have more ants than others
- Those hills are more attractive than others
- # ants is probability that a random user would end on a website



- Websites = nodes = anthill
- Links = edges = paths
- You place ants on each node
- They walk over the paths (at random, they are ants!)
- After some time, some anthills will have more ants than others
- Those hills are more attractive than others
- # ants is probability that a random user would end on a website

7th of February, 2013



- Websites = nodes = anthill
- Links = edges = paths
- You place ants on each node
- They walk over the paths (at random, they are ants!)
- After some time, some anthills will have more ants than others
- Those hills are more attractive than others
- # ants is probability that a random user would end on a website



- Websites = nodes = anthill
- Links = edges = paths
- You place ants on each node
- They walk over the paths
   (at random, they are ants)
- After some time, some anthills will have more ants than others
- Those hills are more attractive than others
- # ants is probability that a random user would end on a website



- Websites = nodes = anthill
- Links = edges = paths
- You place ants on each node
- They walk over the paths (at random, they are ants!)
- After some time, some anthills will have more ants than others
- Those hills are more attractive than others
- # ants is probability that a random user would end on a website

### **Ants**



- Websites = nodes = anthill
- Links = edges = paths
- You place ants on each node
- They walk over the paths (at random, they are ants!)
- After some time, some anthills will have more ants than others
- Those hills are more attractive than others
- # ants is probability that a random user would end on a website

## **Ants**



- Websites = nodes = anthill
- Links = edges = paths
- You place ants on each node
- They walk over the paths (at random, they are ants!)
- After some time, some anthills will have more ants than others
- Those hills are more attractive than others
- # ants is probability that a random user would end on a website

#### **Ants**



- Websites = nodes = anthill
- Links = edges = paths
- You place ants on each node
- They walk over the paths (at random, they are ants!)
- After some time, some anthills will have more ants than others
- Those hills are more attractive than others
- lack # ants is probability that a random user would end on a website

#### **Mathematics**



#### Let x be a web page. Then

- $lackbox{ }L(x)$  is the set of Websites that link to x
- $lackbox{ } C(y)$  is the out-degree of page y
- lacktriangledown lpha is probability of random jump
- lacksquare N is the total number of websites

$$PR(x) := \alpha \left(\frac{1}{N}\right) + (1 - \alpha) \sum_{y \in L(x)} \frac{PR(y)}{C(y)}$$



```
function PAGERANK(Graph web, double q = 0.15, int iterations)
   for all page \in web do
       page.pageRank = \frac{1}{|web|}

    intial probability

   end for
   while iterations > 0 do

    ▷ calculate pageRank of page

       for all page \in web do
           page.pageRank = q
           for all y \in L(page) do
               page.pageRank += \frac{y.pageRank}{C(y)}
           end for
       end for
       iterations = 1
```



```
function PAGERANK(Graph web, double q = 0.15, int iterations)
   for all page \in web do
       page.pageRank = \frac{1}{|web|}

    intial probability

   end for
   while iterations > 0 do

    ▷ calculate pageRank of page

       for all page \in web do
           page.pageRank = q
           for all y \in L(page) do
               page.pageRank += \frac{y.pageRank}{C(y)}
           end for
       end for
       iterations = 1
```



```
function PAGERANK(Graph web, double q = 0.15, int iterations)
   for all page \in web do
       page.pageRank = \frac{1}{|web|}

    intial probability

   end for
   while iterations > 0 do

    ▷ calculate pageRank of page

       for all page \in web do
           page.pageRank = q
           for all y \in L(page) do
               page.pageRank += \frac{y.pageRank}{C(y)}
           end for
       end for
       iterations = 1
```



```
function PAGERANK(Graph web, double q = 0.15, int iterations)
   for all page \in web do
       page.pageRank = \frac{1}{|web|}

    intial probability

   end for
   while iterations > 0 do

    ▷ calculate pageRank of page

       for all page \in web do
           page.pageRank = q
           for all y \in L(page) do
               page.pageRank += \frac{y.pageRank}{C(y)}
           end for
       end for
       iterations = 1
```



```
function PAGERANK(Graph web, double q = 0.15, int iterations)
   for all page \in web do
       page.pageRank = \frac{1}{|web|}

    intial probability

   end for
   while iterations > 0 do
       for all page \in web do

    ▷ calculate pageRank of page

           page.pageRank = q
           for all y \in L(page) do
               page.pageRank += \frac{y.pageRank}{C(y)}
           end for
       end for
       iterations = 1
```



```
function PAGERANK(Graph web, double q = 0.15, int iterations)
   for all page \in web do
       page.pageRank = \frac{1}{|web|}

    intial probability

   end for
   while iterations > 0 do

    ▷ calculate pageRank of page

       for all page \in web do
           page.pageRank = q
           for all y \in L(page) do
               page.pageRank += \frac{y.pageRank}{C(y)}
           end for
       end for
       iterations = 1
```



```
function PAGERANK(Graph web, double q = 0.15, int iterations)
   for all page \in web do
       page.pageRank = \frac{1}{|web|}

    intial probability

   end for
   while iterations > 0 do

    ▷ calculate pageRank of page

       for all page \in web do
           page.pageRank = q
           for all y \in L(page) do
               page.pageRank += \frac{y.pageRank}{C(y)}
           end for
       end for
       iterations = 1
```



```
function PAGERANK(Graph web, double q = 0.15, int iterations)
   for all page \in web do
       page.pageRank = \frac{1}{|web|}

    intial probability

   end for
   while iterations > 0 do

    ▷ calculate pageRank of page

       for all page \in web do
           page.pageRank = q
           for all y \in L(page) do
               page.pageRank += \frac{y.pageRank}{C(y)}
           end for
       end for
       iterations = 1
```

## **Factors of Influence**



- language
- place
- social information

PageRank

## **Factors of Influence**



- language
- place
- social information

## **Factors of Influence**



- language
- place
- social information

### The Filter Bubble



dontbubble.us www.thefilterbubble.com

#### The Filter Bubble





#### The Filter Bubble





## What You've Learned



- web directories
- web crawler
- graph (nodes, eges)
- random walk (ants)
- PageRank
- read pseudocode
- filter bubble

# **Image Sources**



- PageRank by Felipe Micaroni Lalli
- screenshots of www.dmoz.org
- Hyperlink by Bernard Ladenthin
- screenshots of dontbubble.us
- Sergey Brin by enlewof
- Larry Page by aweigend

#### Thanks for Your Attention!



Days 1 - 10

Teach yourself variables, constants, arrays, strings, expressions, statements, functions....



Davs 11 - 21

Teach yourself program flow. pointers, references, classes, objects, inheritance, polymorphism. ....



Days 22 - 697

Do a lot of recreational programming. Have fun hacking but remember to learn from your mistakes.



Days 698 - 3648

Interact with other programmers. Work on programming projects together. Learn from them.





Days 3649 - 7781

Teach yourself advanced theoretical physics and formulate a consistent theory of quantum gravity.



Days 7782 - 14611

Teach yourself biochemistry, molecular biology, genetics....



#### Day 14611

Use knowledge of biology to make an age-reversing potion.



#### Day 14611

Use knowledge of physics to build flux capacitor and go back in time to day 21.



Day 21 Replace younger self.



As far as I know, this is the easiest way to

"Teach Yourself C++ in 21 Days".