

THE UNIVERSITY  
of EDINBURGH

# Text Technologies for Data Science

## INFR11145

# Web Search

Instructor:  
**Tuğrulcan “Tj” Elmas**


30-Oct-2024

1

## Lecture Objectives

- Learn about:
  - Working with Massive data
  - Link-Based Ranking (PageRank)
  - Anchor text

*Tuğrulcan Elmas, TTDS 2024/2025*



THE UNIVERSITY  
of EDINBURGH

2

## What is the challenge in relevance?

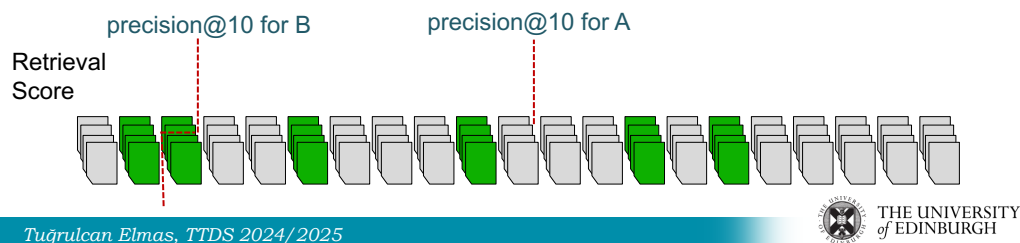
- No clear semantics, contrast:
  - “William Shakespeare”
  - Author history’s? list of plays? a play by him?
- Inherent ambiguity of language:
  - polysemy: “Apple”, “Jaguar”
- Relevance highly subjective
- On the web: counter SEOs / spam
- **Potential solution: Wisdom of the crowds**

## Web is Massive

- No design/co-ordination
- Growing
  - 20 PB/day in 2008 → 160 PB/day in 2013 → now??
  - 1 PB = 1,000 TB = 1,000,000 GB
- Content includes news, fake news, spam, generated content
- Challenging for a search engine
  - Technicalities: (storage, processing, ...)
  - Apparently relevant pages with low quality
- **Opportunity!**

## Effect of Massive data on Precision

- Assume two good search engines that collect two sub-sets of the web
  - Search engine A collected  $N$  docs  $\rightarrow$  precision@10 = 40%
  - Search engine B collected  $4N$  docs  $\rightarrow$  precision@10??
    - Distribution of relevant/non-relevant documents stays the same
    - Overall Precision/Recall stays the same
    - In any decent IR system: more relevant docs exist at the top  $\rightarrow P@n \uparrow \rightarrow$  precision@10 = 60% (increases)



5

## Big Data or Clever Algorithm?

- For Web search, larger index usually would beat a better retrieval algorithm
  - Google Index vs Bing Index
- Similar to other applications
  - Machine Translation: Google vs IBM
    - Statistical methods trained over **10x** training data beat deep NLP methods with **1x** training data
  - LLMs
    - ChatGPT trained on whole internet (Common Crawl ~5 PB)
  - Question Answering:
    - IBM Watson vs Microsoft experiment

Tuğrulcan Elmas, TTDS 2024/2025

THE UNIVERSITY of EDINBURGH

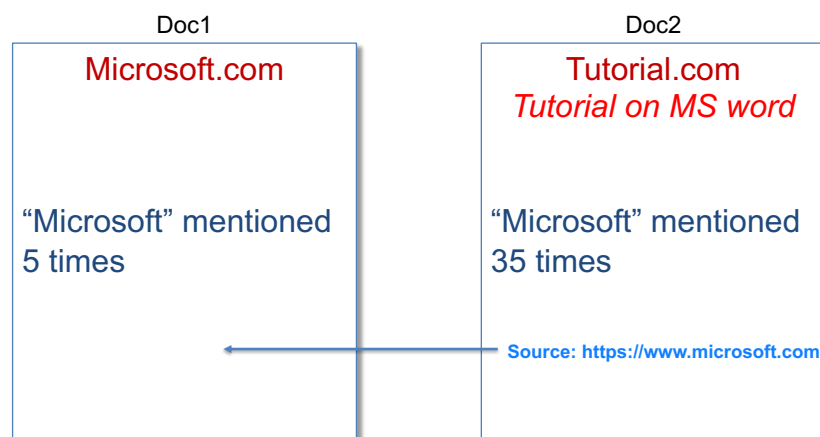
6

## Big Data or Clever Algorithm?

- Question answering task:
  - **Q:** Who created the character of Scrooge?
  - **A:** Scrooge, introduced by **Charles Dickens** in “A Christmas Carol”
  - Requires heavy linguistic analysis, lots of research in TREC
- 2002, Microsoft
  - Identify (subj verb obj), rewrite as queries:
    - Q1: “created the character of Scrooge”
    - Q2: “the character of Scrooge was created by”
  - Search the web for exact phrase, get top 500 results
  - Extract phrase: ■Q1 or Q2■ , get most frequent ■
  - Very naive approach, ignores most answers patterns
  - Who cares!! Web is huge, you will find matches anyway

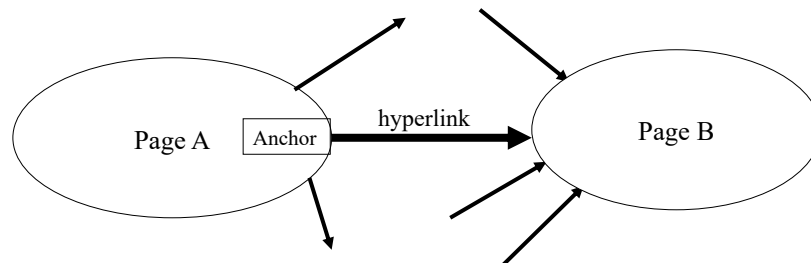
117	Dickens
78	Christmas Carol
75	Charles Dickens
72	Disney
54	Carl Banks
...	

## Search “Microsoft”



**WEB IS CONNECTED!**

## The Web as a Directed Graph



**Assumption 1:** A hyperlink between pages denotes author perceived relevance (quality signal)

**Assumption 2:** The text in the anchor of the hyperlink describes the target page (textual context)

## Links between Pages

- Google Description of **PageRank**:
  - Relies on the “**uniquely democratic**” nature of the web
  - Interprets a link from page A to page B as “**a vote**”
- $A \rightarrow B$ : means A thinks B worth something
  - “**wisdom of the crowds**”: many links means B must be good
  - **Content-independent** measure of quality of B
- Use as ranking feature, combined with content
  - But not all pages that link to B are of equal importance!
    - Importance of a link from CNN >>> link from blog page
- Google PageRank, 1998
  - How many “good” pages link to B?

## Search “Microsoft”

Doc1

Microsoft.com

“Microsoft” mentioned  
5 times

Doc2

Tutorial.com

Tutorial on MS word

“Microsoft” mentioned  
35 times

Tuğrulcan Elmas, TTDS 2024/2025

 THE UNIVERSITY of EDINBURGH

11

## PageRank: Random Surfer

- Analogy:
  - User starts browsing at a random page
  - Pick a random outgoing link → goes there → repeat forever
  - Example:  
G → E → F → E → D → B → C
  - With probability  $1-\lambda$  jump to a random page
    - Otherwise, can get stuck forever A, or B ↔ C
- **PageRank** of page x
  - Probability of being at page x at a random moment in time

Tuğrulcan Elmas, TTDS 2024/2025

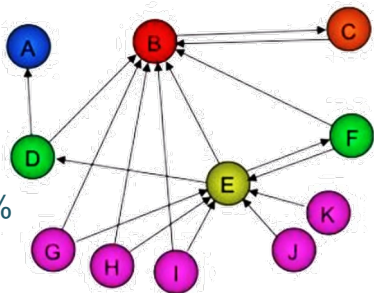
 THE UNIVERSITY of EDINBURGH


12

## PageRank: Algorithm

- Initialize  $PR_0(x) = \frac{100\%}{N}$ 
  - $N$ : total number of pages
  - $PR_0(A) = \dots = PR_0(K) = \frac{100\%}{11} = 9.1\%$
- For every page  $x$ 

$$PR_{t+1}(x) = \frac{1 - \lambda}{N} + \lambda \sum_{y \rightarrow x} \frac{PR_t(y)}{L_{out}(y)}$$
  - $y \rightarrow x$  contributes part of its PR to  $x$  (& other out-links)
  - Iterate till converge  $\rightarrow$  PR scores should sum to 100%



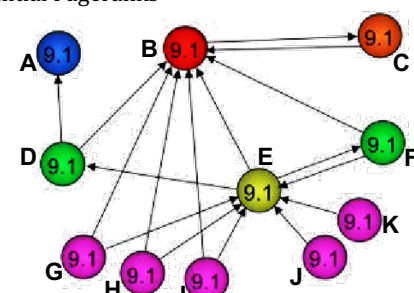
Tuğrulcan Elmas, TTDS 2024/2025

THE UNIVERSITY of EDINBURGH

13

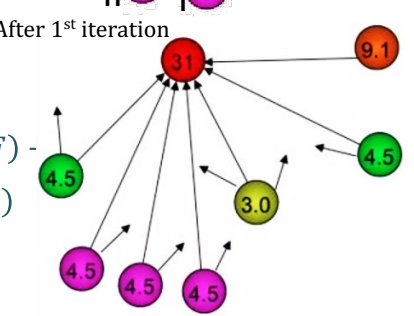
## PageRank: Example


- Let  $\lambda = 0.82$
- $PR_0(C) = PR_0(B) = \dots = \frac{100\%}{11} = 9.1\%$
- $PR_1(C) = \frac{(1-0.82)}{11} + 0.82 \times PR_0(B)$   
 $= 0.18 \times 9.1\% + 0.82 \times 9.1\%$   
 $= 9.1\%$
- $PR_1(B) = \frac{0.18}{11} + 0.82 \times [PR_0(C) + \frac{1}{2}PR_0(D) + \frac{1}{3}PR_0(E) + \frac{1}{2}PR_0(F) + \frac{1}{2}PR_0(G) + \frac{1}{2}PR_0(H) + \frac{1}{2}PR_0(I)]$   
 $\approx 0.31 = 31\%$
- $PR_2(C) = \frac{0.18}{11} + 0.82 \times 31\% \approx 26\%$

Initial Pageranks



After 1<sup>st</sup> iteration

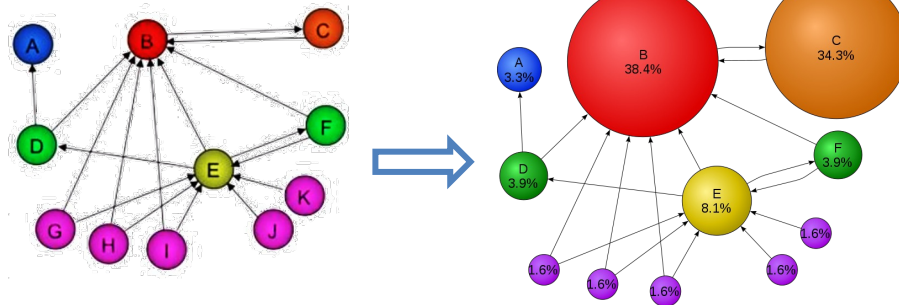


Tuğrulcan Elmas, TTDS 2024/2025

THE UNIVERSITY of EDINBURGH

14

## PageRank: Example result

- Algorithm converges after few iterations



- Observations

- Pages with no inlinks:  $PR = (1 - \lambda) / N = 0.18 / 11 = 1.6\%$
- Same (or symmetric) inlinks  $\rightarrow$  same PR (e.g. **D** and **F**)
- One inlink from high PR  $\gg$  many from low PR (e.g. **C** vs **E**)

## Anchor Text

- Anchor Text (text of a link):
  - Description of destination page
  - Short, descriptive like a query
  - Re-formulated in different ways
    - Human “query expansion”
- Used when indexing page content
  - Add text of all anchor text linking the page
  - Different weights for different anchor text
    - Weighted according to PR of linking page
- Significantly improves retrieval





## Vulnerabilities

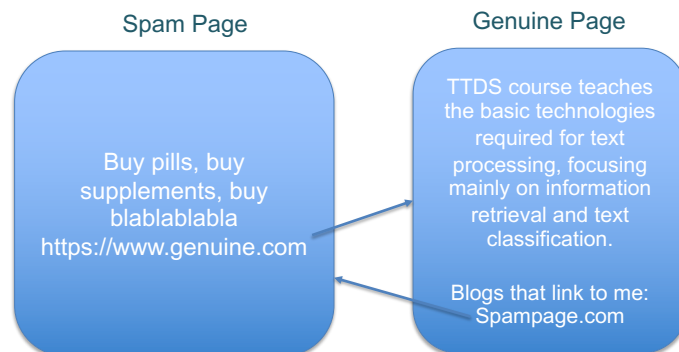
- Hawthorne Effect: “Observation changes behavior”
- Goodhart’s Law: “When a measure becomes a target, it ceases to be a good measure”
- The Cobra Effect: “The solution worsens the problem due to other incentives”



17

## Trackback Links Spam

- “Blogs that link to me” feature
  - Pings the linked page (Sping)
  - Artificial feedback loops
    - Similar to “follow back” on Twitter



Tuğrulcan Elmas, TTDS 2024/2025

THE UNIVERSITY of EDINBURGH

18

## Comment Spam

- Links from comments on sites with high PR
  - One solution: insert `rel=nofollow` into links
    - Link ignored when computing PR

950 thoughts on "Introduction to Computational Social Science open"

Spam link in the name field

AMAZON FIKRI MÜLKİYET on 9.11.2023 at 14.54 said:  
Your site looks very nice. You have excellent infrastructure.

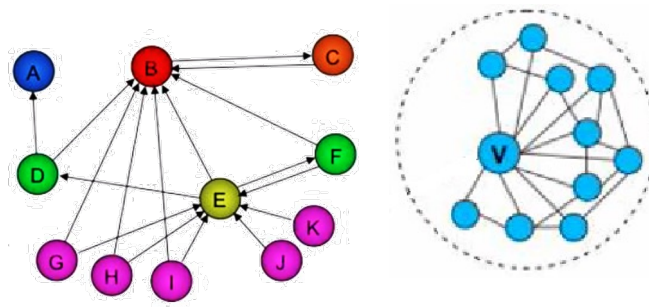
Spam link in the comment

TREVORSTENI on 1.2.2024 at 5.13 said:  
I gave <https://www.cornbreadhemp.com/products/cbda-oil> a prove for the treatment of the cardinal time, and I'm amazed! They tasted smashing and provided a be under the impression that of calmness and relaxation. My lay stress melted away, and I slept less ill too. These gummies are a game-changer on the side of me, and I highly commend them to anyone seeking appropriate worry relief and better sleep.

19

## Link Farms

- Fake densely-connected graph (a clique)
- Hundreds of web domains / IPs can be hosted on one machine



20

## Google Bombing Incident of 2003


[Web](#) [Images](#) [Groups](#) [News](#) [Froogle](#) [Local](#) [more »](#)  
  [Advanced Search](#)  
[Preferences](#)

**Web** Results 1 - 10 of about 969,000 for [miserable failure](#). (0.06 seconds)

[Biography of President George W. Bush](#)  
 Biography of the president from the official White House web site.  
[www.whitehouse.gov/president/gwbbio.html](http://www.whitehouse.gov/president/gwbbio.html) - 29k - [Cached](#) - [Similar pages](#)  
[Past Presidents](#) - [Kids Only](#) - [Current News](#) - [President](#)  
[More results from www.whitehouse.gov »](#)

Monday, October 27, 2003  
**New Web Project**  
 Let's get everyone to link to <http://www.whitehouse.gov/president/gwbbio.html> with the words "Miserable Failure" Our goal is to make Shrubya the top google pick.

It's fun, it's easy just `<a href="http://www.whitehouse.gov/president/gwbbio.html" >Miserable Failure</a>` in your favorite web page will look like [Miserable Failure](#)

[Permalink](#) | posted by George @ 10/27/2003 09:02:00 AM |


- "Anchor Text Spam"
- Fixed in 2007, when Google changed its index


*Tuğrulcan Elmas, TTDS 2024/2025*  THE UNIVERSITY of EDINBURGH

21

## High Pagerank Websites Gaming The Algorithm


- Not giving outlinks
  - Sources can be cited without links
- Creating irrelevant content



 Wise  
<https://wise.com> › [blog](#) › [send-money-on-revolut-ireland](#) :

**How to send money on Revolut: Step by step guide**  
 21 Aug 2023 — Go to the 'Transfer' section of the app, · Put their Revtag into the search bar, · Select the correct profile · Then follow the instructions to ...

- Stay tuned for more SEO tactics on the next lecture!

*Tuğrulcan Elmas, TTDS 2024/2025*  THE UNIVERSITY of EDINBURGH

22

## The Reality

- **PageRank** is used in Google, but is hardly the full story of ranking
  - A big hit when initially proposed, but just one feature now
  - Machine-learned ranking heavily used
    - Learning to Rank (L2R)
  - Still a very useful feature

## Machine-Learned Ranking (Spoiler Alert!)

Column in Output	Description	Column in Output	Description
1	TF(Term frequency) of body	24	LMIR.JM of body
2	TF of anchor	25	BM25 of anchor
3	TF of title	26	LMIR.ABS of anchor
4	TF of URL	27	LMIR.DIR of anchor
5	TF of whole document	28	LMIR.JM of anchor
6	IDF(Inverse document frequency) of body	29	BM25 of title
7	IDF of anchor	30	LMIR.ABS of title
8	IDF of title	31	LMIR.DIR of title
9	IDF of URL	32	LMIR.JM of title
10	IDF of whole document	33	BM25 of URL
11	TF*IDF of body	34	LMIR.ABS of URL
12	TF*IDF of anchor	35	LMIR.DIR of URL
13	TF*IDF of title	36	LMIR.JM of URL
14	TF*IDF of URL	37	BM25 of whole document
15	TF*IDF of whole document	38	LMIR.ABS of whole document
16	DL(Document length) of body	39	LMIR.DIR of whole document
17	DL of anchor	40	LMIR.JM of whole document
18	DL of title	41	PageRank
19	DL of URL	42	Inlink number
20	DL of whole document	43	Outlink number
21	BM25 of body	44	Number of slash in URL
22	LMIR.ABS of body	45	Length of URL
23	LMIR.DIR of body	46	Number of child page

## Summary

- Web data is massive
  - Challenging for efficiency, but useful for effectiveness
- PageRank:
  - Probability that a random surfer is currently on page x
  - The more powerful pages linking to x, the higher the PR
- Anchor text:
  - Short concise description of target page content
  - Very useful for retrieval
- Spam
  - Link spam, Anchor text spam

## Resources

- Text book 1: Intro to IR, Section 21.1
- Text Book 2: IR in Practice: 4.5, 10.3
- Page Rank Paper:  
Page, L., Brin, S., Motwani, R., & Winograd, T. (1999).  
*The PageRank citation ranking: Bringing order to the web.*  
Stanford InfoLab.
- Additional reading:  
Dumais, S., Banko, M., Brill, E., Lin, J., & Ng, A. (2002)  
Web question answering: Is more always better?.  
SIGIR 2002.
- YouTube Video: How Search Works  
<https://www.youtube.com/watch?v=BNHR6lQJGZs>