Web Search

COMP90049
Knowledge Technologies

Lea Frermann and Justin Zobel and Karin Verspoor, CIS

Semester 2, 2019

THE UNIVERSITY OF MELBOURNE

Link analysis

So far,

- indexing
- boolean and ranked retrieval
- each document was considered independently

But,

- **Link structure** of the web is highly informative
  consider: personal blogs, wedding planning sites, ...
  vs.: news sites, wikipedia articles, ...

- # incoming links predict page's **importance** (to general public)
- # incoming links predict page's **authority**

- Anchor text
- PageRank

Anchor text


```
For more information on unit tests, please go to <a
href="http://www.unittests.com">the unit test page</a>.
```

Anchor text, is crucial for web search. For example,

- There are many thousands of ‘Library’ references in the unimelb web site.
- The Library home page does not mention the word often (hardly any plain text at all).
- Most of the within-Unimelb ‘library’ links point to the Library home page.
- Most of the links to the home page contain the word ‘library’.

Anchor text is treated as a form of **zone**.

Anchor text


```
For more information on unit tests, please go to <a
href="http://www.unittests.com">the unit test page</a>.
```

Why useful?

- topic indicator
- very concise (easy to index, typically unambiguous)
- we usually get many of those per (important) page

Link = vote of importance (independent of anchor text content). PageRank uses this fact.

PageRank intuition

Let me introduce **Bob**. Bob is bored.

PageRank intuition

Let me introduce **Bob**. Bob is bored.

So bored that he spends his days **clicking through webpages**.

PageRank intuition

Let me introduce **Bob**. Bob is bored.

So bored that he spends his days **clicking through webpages**.

Bob's keyboard has a **teleport** button which takes him to a random website

PageRank intuition

Let me introduce **Bob**. Bob is bored.

So bored that he spends his days **clicking through webpages**.

Bob's keyboard has a **teleport** button which takes him to a random website

At each webpage, Bob decides to

EITHER click a random link on that webpage

OR click his teleport button

PageRank intuition

Let me introduce **Bob**. Bob is bored.

So bored that he spends his days **clicking through webpages**.

Bob's keyboard has a **teleport** button which takes him to a random website

At each webpage, Bob decides to

EITHER click a random link on that webpage

OR click his teleport button

If he decides to click a link, he chooses it completely at random.

PageRank intuition

Let me introduce **Bob**. Bob is bored.

So bored that he spends his days **clicking through webpages**.

Bob's keyboard has a **teleport** button which takes him to a random website

At each webpage, Bob decides to

EITHER click a random link on that webpage

OR click his teleport button

If he decides to click a link, he chooses it completely at random.

Bob is so bored that he keeps doing this forever

PageRank intuition

Let me introduce **Bob**. Bob is bored.

So bored that he spends his days **clicking through webpages**.

Bob's keyboard has a **teleport** button which takes him to a random website

At each webpage, Bob decides to

EITHER click a random link on that webpage

OR click his teleport button

If he decides to click a link, he chooses it completely at random.

Bob is so bored that he keeps doing this forever

Bob performs a random walk over the internet.

- Bob can't get stuck (because of the catapult)
- Bob will eventually reach every webpage there is
- Bob will spend most time on popular webpages (e.g., `abc.com.au`)

PageRank intuition

Let me introduce **Bob**. Bob is bored.

So bored that he spends his days **clicking through webpages**.

Bob's keyboard has a **teleport** button which takes him to a random website

At each webpage, Bob decides to

EITHER click a random link on that webpage

OR click his teleport button

If he decides to click a link, he chooses it completely at random.

Bob is so bored that he keeps doing this forever

At some point, you walk past and glance at Bob's screen.
What is the probability that Bob is looking at `abc.com.au`?

PageRank intuition

Let me introduce **Bob**. Bob is bored.

So bored that he spends his days **clicking through webpages**.

Bob's keyboard has a **teleport** button which takes him to a random website

At each webpage, Bob decides to

EITHER click a random link on that webpage

OR click his teleport button

If he decides to click a link, he chooses it completely at random.

Bob is so bored that he keeps doing this forever

At some point, you walk past and glance at Bob's screen.
What is the probability that Bob is looking at `abc.com.au`? `unimelb.edu.au`?

PageRank intuition

Let me introduce **Bob**. Bob is bored.

So bored that he spends his days **clicking through webpages**.

Bob's keyboard has a **teleport** button which takes him to a random website

At each webpage, Bob decides to

EITHER click a random link on that webpage

OR click his teleport button

If he decides to click a link, he chooses it completely at random.

Bob is so bored that he keeps doing this forever

At some point, you walk past and glance at Bob's screen.
What is the probability that Bob is looking at `abc.com.au`? `unimelb.edu.au`? `houseofcardsespresso.com`?

PageRank intuition

Let me introduce **Bob**. Bob is bored.

So bored that he spends his days **clicking through webpages**.

Bob's keyboard has a **teleport** button which takes him to a random website

At each webpage, Bob decides to

EITHER click a random link on that webpage

OR click his teleport button

If he decides to click a link, he chooses it completely at random.

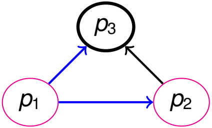
Bob is so bored that he keeps doing this forever

At some point, you walk pa
What is the probability that
`unimelb.edu.au`? `houseof`

← The webpage's **PageRank score!**

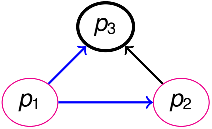
Calculating PageRank Scores

- $PR(v)$ pagerank score of page v
- I_v set of wegpages pointing to v
  (e.g., $I_{p_3} = \{p_1, p_2\}$; set of pages pointing to p_3)
- O_v number of outgoing links of page v
  (e.g., $O_{p_1} = 2$ outgoing links from p_1)



Calculating PageRank Scores

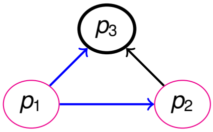
- $PR(v)$ pagerank score of page v
- I_v set of wegpages pointing to v
  (e.g., $I_{p_3} = \{p_1, p_2\}$; set of pages pointing to p_3)
- O_v number of outgoing links of page v
  (e.g., $O_{p_1} = 2$ outgoing links from p_1)



$$PR(v) = \sum_{u \in I_v} \frac{PR(u)}{O_u} \qquad \text{(no catapult)}$$

Calculating PageRank Scores

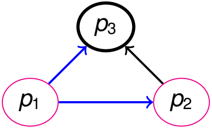
- $PR(v)$ pagerank score of page v
- I_v set of wegpages pointing to v (e.g., $I_{p_3} = \{p_1, p_2\}$; set of pages pointing to p_3)
- O_v number of outgoing links of page v (e.g., $O_{p_1} = 2$ outgoing links from p_1)
- N all webpages in our collection (here: $N = 3$)
- λ probability to catapult



$$PR(v) = \frac{\lambda}{N} + (1 - \lambda) \times \sum_{u \in I_v} \frac{PR(u)}{O_u} \qquad \text{(with catapult)}$$

Calculating PageRank Scores

- $PR(v)$ pagerank score of page v
- I_v set of wegpages pointing to v
  (e.g., $I_{p_3} = \{p_1, p_2\}$; set of pages pointing to p_3)
- O_v number of outgoing links of page v
  (e.g., $O_{p_1} = 2$ outgoing links from p_1)
- N all webpages in our collection (here: $N = 3$)
- λ probability to catapult

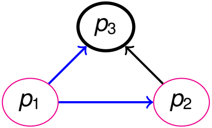


$$PR(v) = \frac{\lambda}{N} + (1-\lambda) \times \sum_{u \in I_v} \frac{PR(u)}{O_u} \qquad \text{(with catapult)}$$

but we don't know PR(u)

Calculating PageRank Scores

- $PR(v)$ pagerank score of page v
- I_v set of wegpages pointing to v
  (e.g., $I_{p_3} = \{p_1, p_2\}$; set of pages pointing to p_3)
- O_v number of outgoing links of page v
  (e.g., $O_{p_1} = 2$ outgoing links from p_1)
- N all webpages in our collection (here: $N = 3$)
- λ probability to catapult



$$PR(v) = \frac{\lambda}{N} + (1-\lambda) \times \sum_{u \in I_v} \frac{PR(u)}{O_u} \qquad \text{(with catapult)}$$

but we don't know PR(u)

Estimate them iteratively:

- each web page has a fixed number of credits
- redistributes credits to pages it links to
- receives credits from pages that link to it
- repeat until stable distribution is reached

Other Tweaks & Heuristics

- Note which pages people visit: count click-throughs.
- Manually alter the behavior of common queries.
- Cache the answers to common queries.
- Index selected phrases.
- Have separate servers for crawling and index construction.
- Accept feeds from dynamic data providers (booksellers, newspapers, ...)
- Integrate diverse data resources, such as maps and videos.

Group Activity!

Video Snippets in Google Results

- Demo
- How do they do this? Secret. Hints in a research paper (below).

Group Activity!

Video Snippets in Google Results

- Demo
- How do they do this? Secret. Hints in a research paper (below).

Major challenges in **parsing** the crawled information?

- how to parse a video?
  text $\rightarrow$ words
  video $\rightarrow$ frames? snippets (how long, how many)?
- how to canonicalize a video?
- meta-data (title, author, descriptions, ...)
- how to deal with comments?
- all of the above?
- none of the above?

Group Activity!

Video Snippets in Google Results

- Demo
- How do they do this? Secret. Hints in a research paper (below).

Major challenges in **indexing**?

- we can index text by terms ...
  what are the 'terms' for videos?
- index as a whole? index snippets (which snippets)?
- index text and video in a single structure? how to link them?
- all of the above?
- none of the above?

Group Activity!

Video Snippets in Google Results

- Demo
- How do they do this? Secret. Hints in a research paper (below).

Major challenges in **querying**?

- query is in text form, result is in video (image) form!
  how to relate the two?
- boolean querying for videos?
- tfidf for videos? (what are ‘terms’ for videos, again?)
- exact query-video matching?
  approximate query-video maching?
  term-based querying?
- all of the above?
- none of the above?

Group Activity!

Video Snippets in Google Results

- Demo
- How do they do this? Secret. Hints in a research paper (below).

Maybe, **machine learning** is the key!

- **learn to** align words or queries with videos
- **learn to** segment videos into meaningful snippets
- ...
- 2nd part of Knowledge Technologies (starting tomorrow)

Group Activity!

Video Snippets in Google Results

- Demo
- How do they do this? Secret. Hints in a research paper (below).

In your groups:

- Discuss Google video snippets as a knowledge technology
- Play with the Google feature
- In which contexts is it useful?
- What information is required? Where could it come from?
- How can the information be indexed and retrieved?

- For hints, check out this Google research paper:

  `Malmaud et al., ''What's Cookin'? Interpreting Cooking Videos using Text, Speech and Vision.'' NAACL 2015.`

- **Answer the quiz** at `https://pollev.com/kt19`

Summary

- **Search** involves crawling, parsing, indexing, querying (and more!)
- **Crawling** is in principle straightforward queuing, but practical issues make it more complex
- **Parsing** involves discarding metadata and hidden information; tokenization; canonicalisation; zoning; and stemming.
- **Inverted indices** describe text collections as lists of the pages with each word, rather than the list of words on each page.
- Inverted indices can be used for **Boolean and ranked querying**.
- On the web, **link and anchor information** can be the dominant evidence of relevance.
- Search goes **beyond word-based matching**: images, videos, phrases and sentences, ads, user behavior...

References / Readings

Malmaud et al., (2015) “What’s Cookin’? Interpreting Cooking Videos using Text, Speech and Vision.” Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

W. Bruce Croft and Donald Metzler and Trevor Strohman (2015), “Search Engines: Information Retrieval in Practice”. Online version. Pearson Education, Inc. Chapter 4.

Pagerank algorithm

Input: D = document set
Output: Π_T = set of pagerank scores for each document $d_i \in D$

for all $d_i \in D$ **do** ▷ Initialise the starting probabilities
  $\pi(d_{(i,0)}) \leftarrow \frac{1}{N}$ ▷ N is the total number of documents
end for
for $t = 1..T$ **do** ▷ Repeat over T iterations
  for all $d_i \in D$ **do** ▷ Initialise the document probabilities
    $\pi(d_{(i,t)}) \leftarrow 0$
  end for
  for all $d_i \in D$ **do**
    if $\exists d_j : d_i \mapsto d_j$ **then**
      for all $d_j \in D$ **do** ▷ EITHER teleport randomly
        $\pi(d_{(j,t)}) \leftarrow \pi(d_{(j,t)}) + \lambda \times \pi(d_{(i,t-1)}) \times \frac{1}{N}$
      end for
      for all d_j **where** $d_i \mapsto d_j$ **do**
        ▷ OR follow an outlink (one of m)
        $\pi(d_{(j,t)}) \leftarrow \pi(d_{(j,t)}) + (1 - \lambda) \times \pi(d_{(i,t-1)}) \times \frac{1}{m}$
      end for
    else
      for all $d_j \in D$ **do** ▷ teleport to a random document
        $\pi(d_{(j,t)}) \leftarrow \pi(d_{(j,t)}) + \pi(d_{(i,t-1)}) \times \frac{1}{N}$
      end for
    end if
  end for
end for

A (tiny) web

$$d_1 \longrightarrow d_2$$

$$PR(v) = \frac{\lambda}{N} + (1 - \lambda) \times \sum_{u \in I_v} \frac{PR(u)}{O_u}$$

$\lambda = 0.5$

t	$\pi(d_{(1,t)})$	$\pi(d_{(2,t)})$
0	0.5	0.5

Pagerank example

A (tiny) web

$$d_1 \longrightarrow d_2$$

$$PR(v) = \frac{\lambda}{N} + (1-\lambda) \times \sum_{u \in I_v} \frac{PR(u)}{O_u}$$

$\lambda = 0.5$

t	$\pi(d_{(1,t)})$	$\pi(d_{(2,t)})$
0	0.5	0.5
1	$0.5 \times 0.2 \times 0.5 + 0.5 \times 0.5 = 0.3$	$0.5 \times 0.2 \times 0.5 + 0.5 \times 0.8 + 0.5 \times 0.5 = 0.7$

Pagerank example

A (tiny) web

$$d_1 \longrightarrow d_2$$

$$PR(v) = \frac{\lambda}{N} + (1-\lambda) \times \sum_{u \in I_v} \frac{PR(u)}{O_u}$$

$\lambda = 0.5$

t	$\pi(d_{(1,t)})$	$\pi(d_{(2,t)})$
0	0.5	0.5
1	$0.5 \times 0.2 \times 0.5 + 0.5 \times 0.5 = 0.3$	$0.5 \times 0.2 \times 0.5 + 0.5 \times 0.8 + 0.5 \times 0.5 = 0.7$
2	$0.3 \times 0.2 \times 0.5 + 0.7 \times 0.5 = 0.38$	$0.3 \times 0.2 \times 0.5 + 0.3 \times 0.8 + 0.7 \times 0.5 = 0.62$

Pagerank example

A (tiny) web

$$d_1 \longrightarrow d_2$$

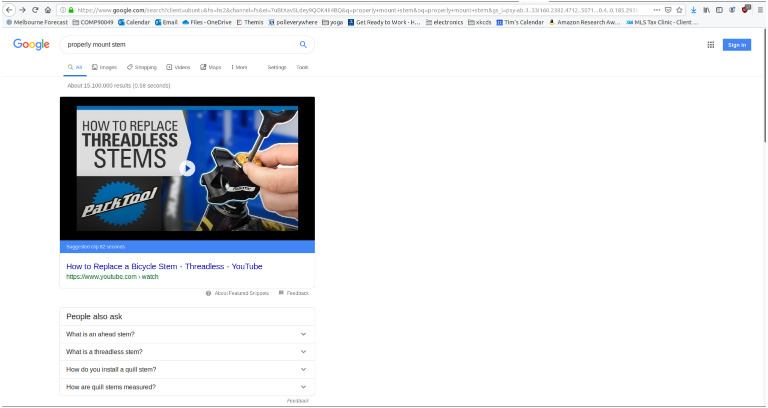
$$PR(v) = \frac{\lambda}{N} + (1-\lambda) \times \sum_{u \in I_v} \frac{PR(u)}{O_u}$$

$\lambda = 0.5$

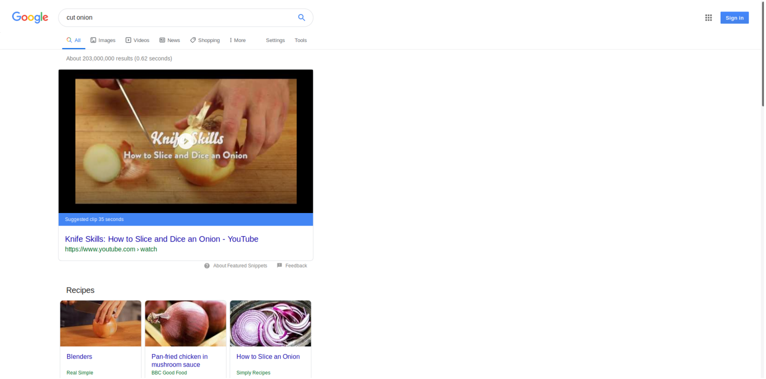
t	$\pi(d_{(1,t)})$	$\pi(d_{(2,t)})$
0	0.5	0.5
1	$0.5 \times 0.2 \times 0.5 + 0.5 \times 0.5 = 0.3$	$0.5 \times 0.2 \times 0.5 + 0.5 \times 0.8 + 0.5 \times 0.5 = 0.7$
2	$0.3 \times 0.2 \times 0.5 + 0.7 \times 0.5 = 0.38$	$0.3 \times 0.2 \times 0.5 + 0.3 \times 0.8 + 0.7 \times 0.5 = 0.62$
3	$0.38 \times 0.2 \times 0.5 + 0.62 \times 0.5 = 0.348$	$0.38 \times 0.2 \times 0.5 + 0.38 \times 0.8 + 0.62 \times 0.5 = 0.652$

$\vdots$

Google Snippet Results



Google Snippet Results



Google Snippet Results