

# Multimedia Technology

## Lecture 4: Miscellaneous Techniques behind IR

Lecturer: *Dr. Wan-Lei Zhao*

*Autumn Semester 2022*

# Outline

- 1 PageRank and HITS
- 2 Evaluation on IR performance

# Pagerank: the motivation (1)

- Retrieval results returned by basic IR system usually are not satisfactory
- There are many reasons behind this
  - ① It is actually a very tough issue
  - ② Nearly all IR systems face the scalability issue
  - ③ Users are not able to express what they want by keywords only
  - ④ The same keyword for different people means different thing, e.g. “apple”
- It requires natural language understanding: **artificial intelligence**
- Hundreds of reranking approaches have proposed to optimize the search results
  - Share the story about SIGIR

# Pagerank: the motivation (2)

- Keywords are very few
- Too many pages share similar similarity score

The screenshot shows a Google search for 'wanlei zhao'. The search bar at the top contains the text 'wanlei zhao'. Below the search bar, there are tabs for 'Web', 'Images', 'Videos', 'News', 'Maps', 'More', and 'Search tools'. The 'Web' tab is selected. The search results are displayed below the tabs, showing approximately 19,200 results in 0.35 seconds. The first few results are:

- Wan-Lei Zhao - Google Scholar Citations**  
scholar.google.com.hk/citations?user=EChpPEAAAAJ&hl=en  
Xiamen University, Fujian, China - xmu.edu.cn  
Near-duplicate keyframe identification with interest point matching and pattern learning.  
WL Zhao, CW Ngo, HK Tan, X Wu. Multimedia, IEEE Transactions on 9 ...
- Wanlei Zhao's homepage at Xiamen University**  
pami.xmu.edu.cn/~wtzhao/  
Oct 1, 2014 - Dr. Wan-Lei Zhao. PAMI research Lab, Computer Science Department.  
Faculty of Information Science and Technology, Xiamen University.
- dblp: Wanlei Zhao**  
dblp.uni-trier.de > Home > Persons  
May 9, 2015 - List of computer science publications by Wanlei Zhao.
- dblp: Wan-Lei Zhao**  
dblp.uni-trier.de > Home > Persons  
Feb 27, 2015 - Compiled list of computer science publications by Wan-Lei Zhao.
- LARGE-SCALE NEAR-DUPLICATE WEB VIDEO ... - VIR...**  
vireo.cs.cityu.edu.hk/papers/icme09-wanlei.pdf  
by WL Zhao - Cited by 16 - Related articles  
LARGE-SCALE NEAR-DUPLICATE WEB VIDEO SEARCH: CHALLENGE AND OPPORTUNITY. Wan-Lei Zhao, Song Tan and Chong-Wah Ngo. Department of ...
- lip-vireo - Google Code**  
code.google.com/p/lip-vireo/  
Written by Wan-lei Zhao, 10/10/2010. B. Important notice. If user wants to try Lip ...  
written by Wan-lei Zhao, 27/11/2011. Terms - Privacy - Project Hosting Help.
- Wanlei Zhao - GForge**  
https://gforge.inria.fr/users/wanlei/  
Jul 3, 2012 - Login Name: wanlei. Real Name: Dr. Wanlei Zhao. Email Address: wtzhao@nosspam@xmu.edu.cn. Xiamen University, Xiamen Fujian, China.

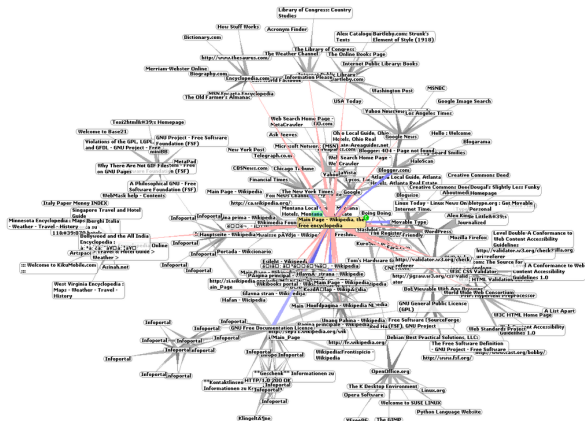
# Page hyper-links

- We are now going to consider
- how hyper-links help to improve the search quality

```
1 <html>
2 <head>page head</head>
3 <body>
4 <p>HTML tutorials are available</p>
5 <a href="http://www.w3schools.com">hyper-link1</a>
6 <p>WWW standards are available</p>
7 <a href="http://www.w3.org">hyper-link2</a>
8 </body>
9 </html>
```

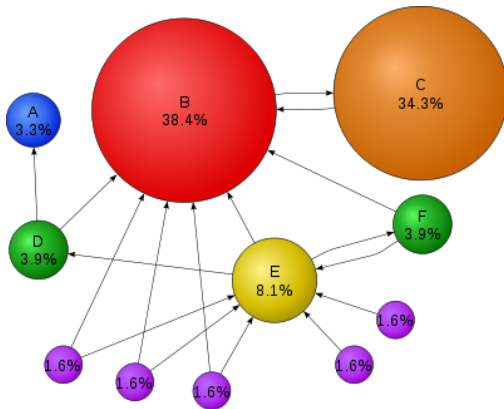
# Pagerank: explained (1)

- Pagerank is one of the most successful reranking approaches
- It is a re-ranking approach
- It happens when we have the retrieval results
- Basic idea: make use of the hyperlinks between webpages
  - Pages being linked (pointed to) by other pages should be important and ranked higher
- Start-up technology for Google



- We are connected by Internet
- Webpages are connected by hyperlinks

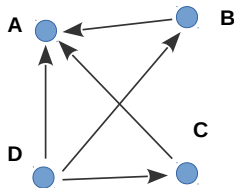
# Pagerank: explained (3)



- Higher weights (pagerank) are assigned to the pages that have many in-ward links
- Notice that out-ward links will not impact your own ranking



# Pagerank: build the model



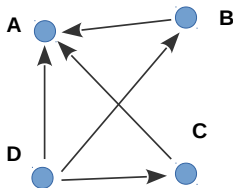
- Given 4 webpages, and the hyperlinks between them
- Calculate pagerank for each of them as following,  $PR(.)$  for all the pages are initialized to **0.25**

$$PR(A) = \frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)}, \quad (1)$$

where  $PR(.)$  is the current pagerank,

$L(.)$  is num. of out-ward links

# Pagerank: build the model



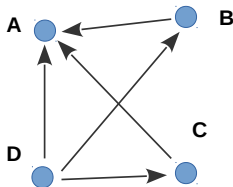
$$PR(A) = \frac{0.25}{1} + \frac{0.25}{1} + \frac{0.25}{3},$$

$$PR(B) = \frac{0.25}{3},$$

$$PR(C) = \frac{0.25}{3},$$

$$PR(D) = 0$$

# Pagerank: the damping factor



- Given **N** is the num. of webpages, **d** is the damping factor,

$$PR(A) = \left( \frac{0.25}{1} + \frac{0.25}{1} + \frac{0.25}{3} \right) \cdot d + \frac{1-d}{N},$$

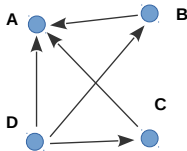
$$PR(B) = \frac{0.25}{3} \cdot d + \frac{1-d}{N},$$

$$PR(C) = \frac{0.25}{3} \cdot d + \frac{1-d}{N},$$

$$PR(D) = 0 \cdot d + \frac{1-d}{N}$$

# Pagerank: the procedure

- 1 Produce Adjacent matrix by collecting all the webpage links
- 2 Initialize  $PR(.)$  to  $c$
- 3 Do
- 4   Calculate  $PR(.)$  for each webpage
- 5   Update  $PR(.)$  for each webpage
- 6 Until convergence



$$M = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

# Pagerank: tricks to promote your webpage

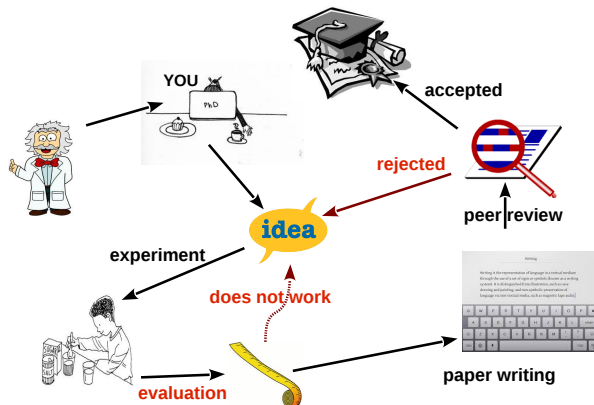
- Share the story about Google
  - What Google means
  - Pagerank is born in the right season
  - Turning point of Google
  - Do we need to **reinvent the wheel**?
- Ask some webpage (has higher pagerank) to link to your webpage
  - Pagerank can be found by install firefox Toolbar or from pagerank website
  - Google robot will ignore hyperlink shares the same color as the background
- Register to Google Webtool
  - Once Google robot visits your site
  - Try to search and click-in your website with Google from different places

# Outline

- 1 PageRank and HITS
- 2 Evaluation on IR performance

# How the “research game” is played

- Loop for experiment-driven research
- Evaluation on a certain benchmark plays key role in the loop



# Recall, precision and F-measure

- True Positive (TP): the number of relevant documents retrieved
- False Negative (FN): the number of relevant documents missed
- False Positive (FP): the number of irrelevant documents retrieved
- True Negative (TN): the number of irrelevant documents not retrieved
- Given the documents we consider (top-K), and relevant document R

$$Recall = \frac{TP}{R} \quad (2)$$

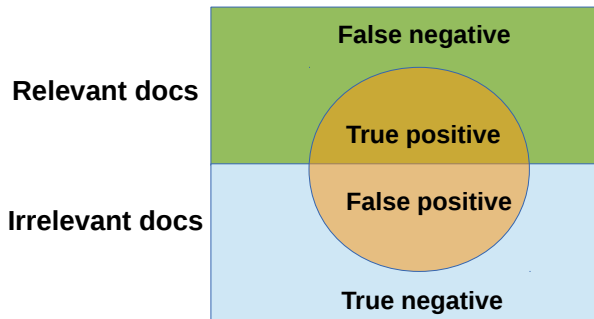
$$Precision = \frac{TP}{K} \quad (3)$$

- F-measure is further defined as

$$F\text{-measure} = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision} \quad (4)$$



# Recall and precision illustration

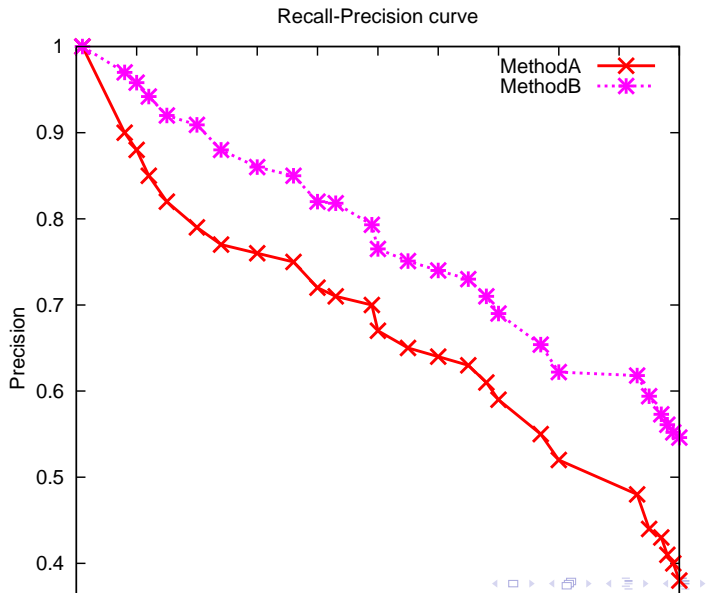


$$\text{Precision} = \frac{\text{True positive}}{\text{True positive} + \text{False positive}}$$

$$\text{Recall} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}}$$

- In classification task, the definition for 'Precision' changes

# Curve of Recall V.S. precision



# Average Precision

- Rankings of relevant docs are explicitly considered
- In practice, users are more sensitive to precision
- In-born advantage for a search engine:  
users have no knowledge about recall
- Average Precision is such a measure fits in
- Average Precision (AP) is defined as

$$AP(i) = \frac{\sum_1^i(1)}{i} \quad (5)$$

- mean Average Precision (mAP) is defined as

$$mAP = \frac{\sum_{i=1}^K AP(i)}{K} \quad (6)$$

# Exercise

- Given total num. of relevant docs is 10

Top	Relevancy
1	1
2	0
3	1
4	0
5	0
6	0
7	1
8	1
9	0

- See Recall=?, Precision=? and mAP=?

# Q & A

Thanks for your attention!