Multimedia Technology

Lecture 4: Miscellaneous Techniques behind IR

Lecturer: *Dr*. Wan-Lei Zhao

Autumn Semester 2022

Email: wlzhao@xmu.edu.cn,

Outline

1 PageRank and HITS

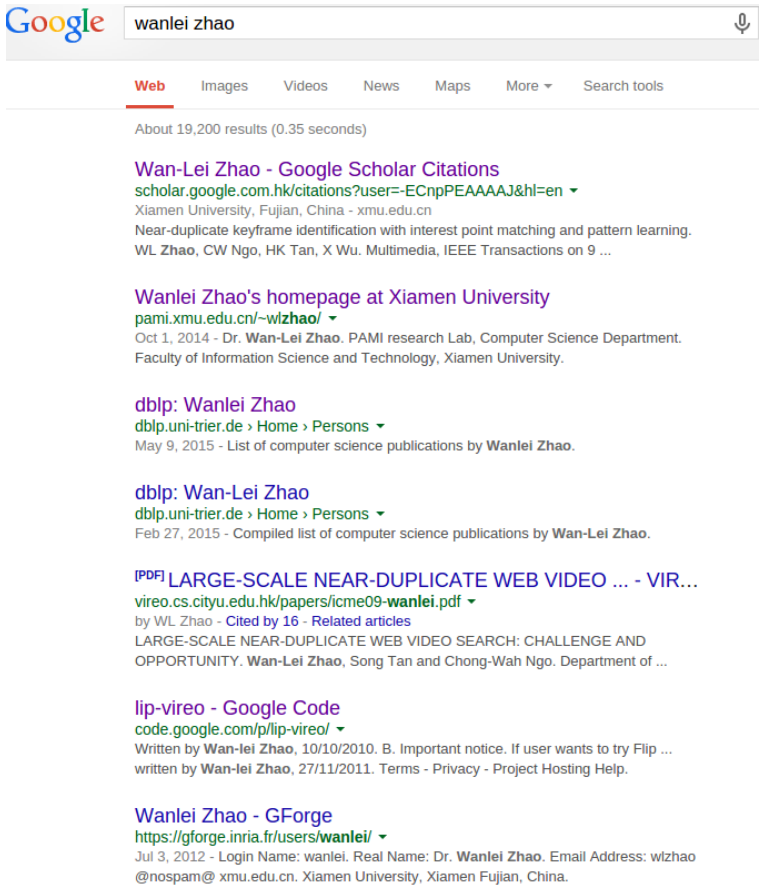
2 Evaluation on IR performance

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: **artifical intellegence**
- 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



Page hyper-links

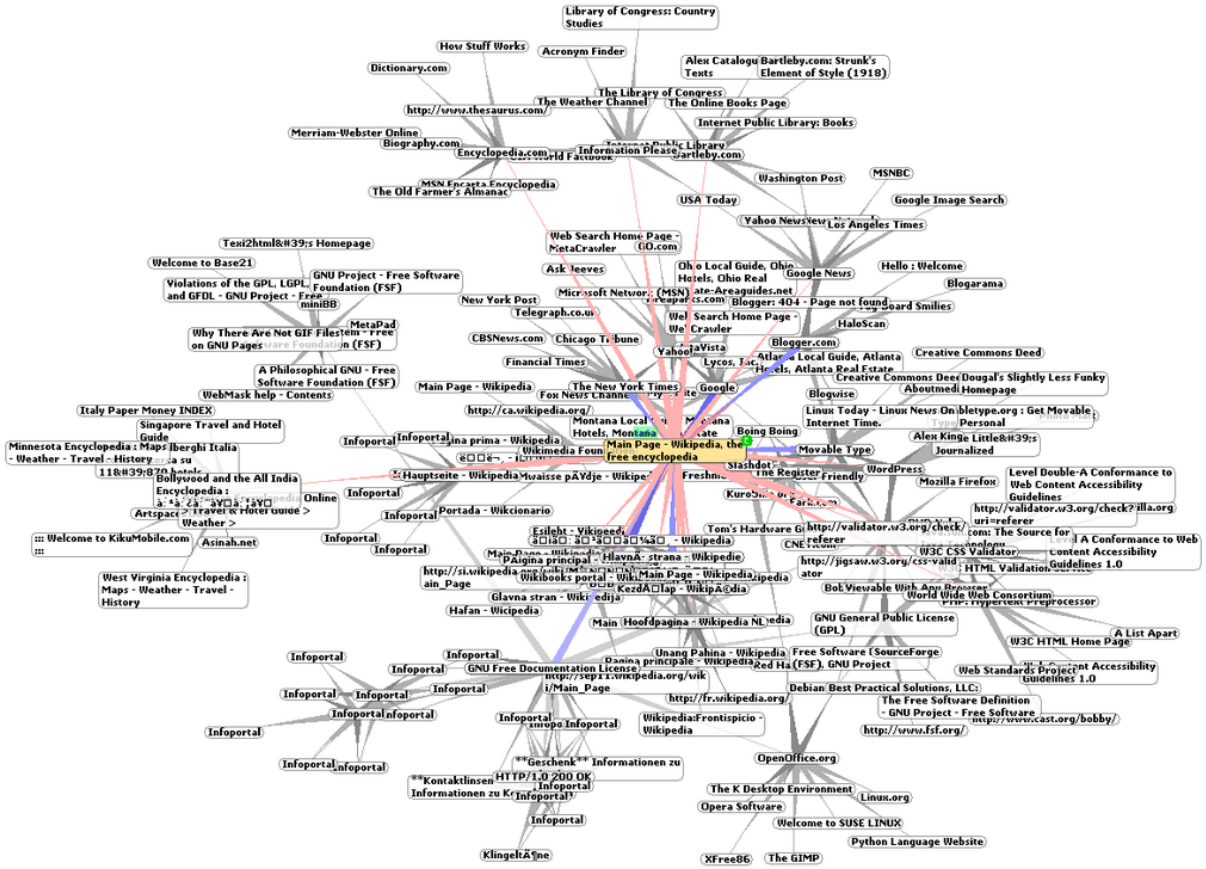
- We are now going to consider
- how hyper-links help to improve the search quality

```
<html>
<head>page head</head>
<body>
<p>HTML tutorials are available</p>
<a href="http://www.w3schools.com">hyper-link1</a>
<p>WWW standards are available</p>
<a href="http://www.w3.org">hyper-link2</a>
</body>
</html>
```

Pagerank: explained (1)

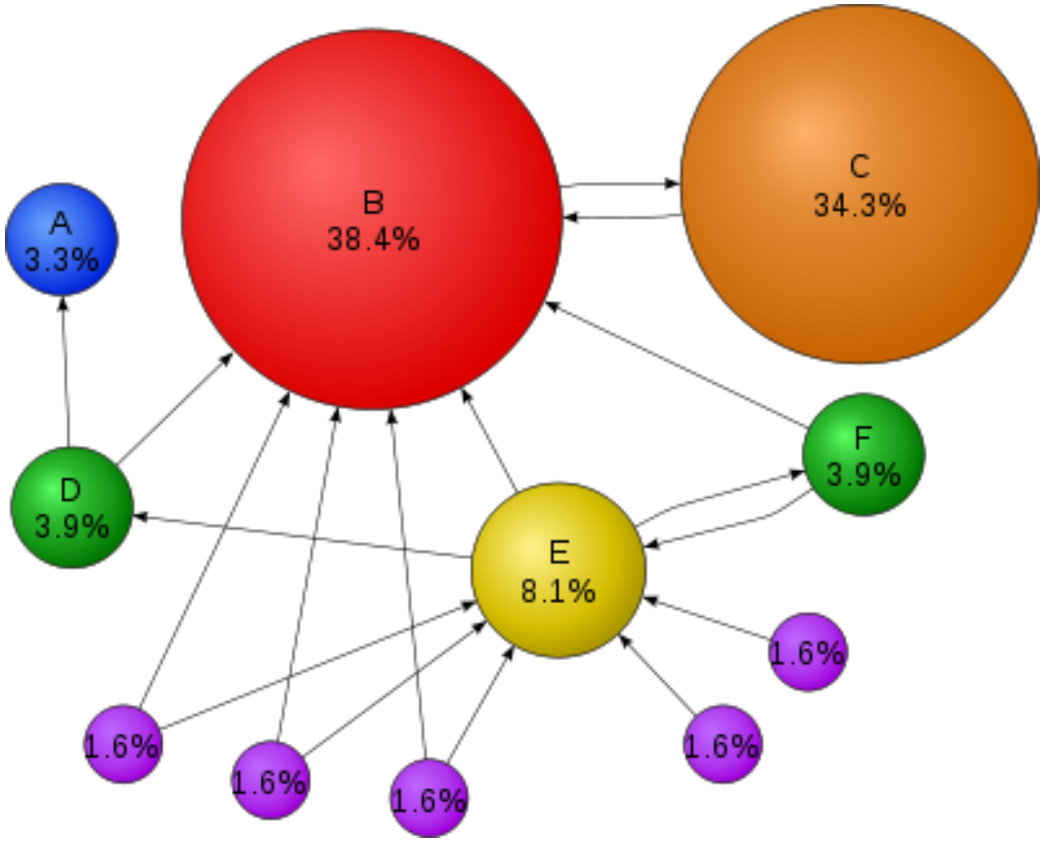
- Pagerank is one of the most successful reranking approaches
- It is a re-ranking approach
- It happens when we have the retrieval results
- Basica idea: make use of the hyperlinks between webpages
  - Pages being linked (pointed to) to by other pages should be important and ranked higher
- Start-up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

Pagerank: explained (2)



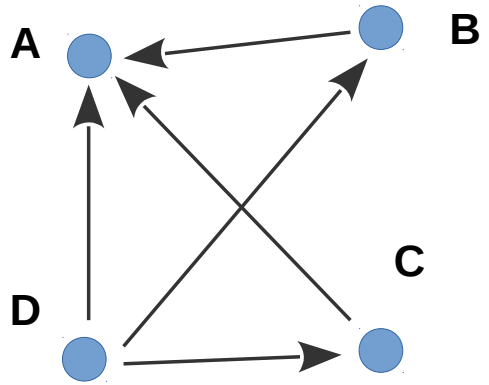
- We are connected by Internet
- Webpages are connected by hyperlinks

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



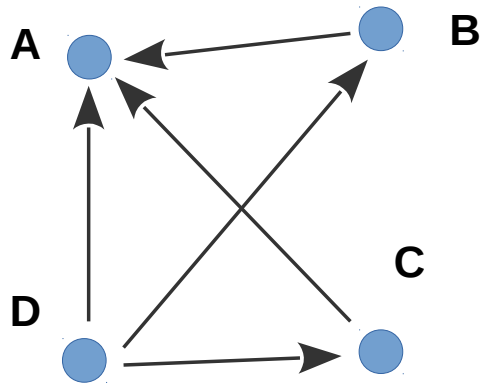
- Given 4 webpages, and the hyperlinks between them
- Calculate pagerank for each of them as following, PR(.) for all the pages are initialized to **0.25**

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

where PR(.) is the current pagerank,

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

Pagerank: build the model



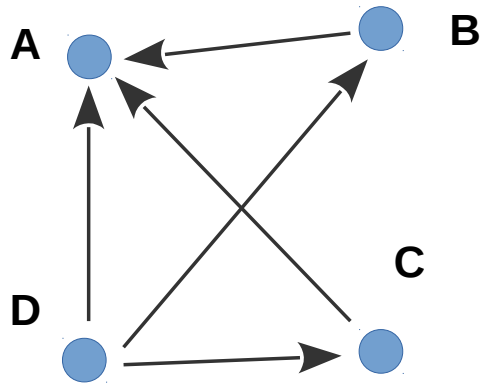
$$PR(A) = \frac{0.25}{1} + \frac{0.25}{1} + \frac{0.25}{3},$$

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

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

$$PR(D) = 0$$

Pagerank: the damping factor



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

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

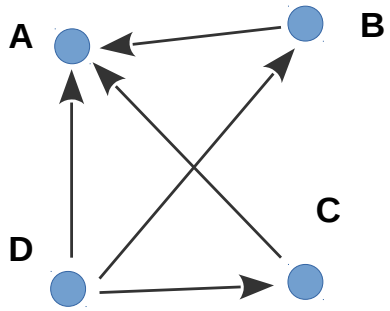
$$PR(B) = \frac{0.25}{3} \cdot d + \frac{1-d}{N},$$

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

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

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

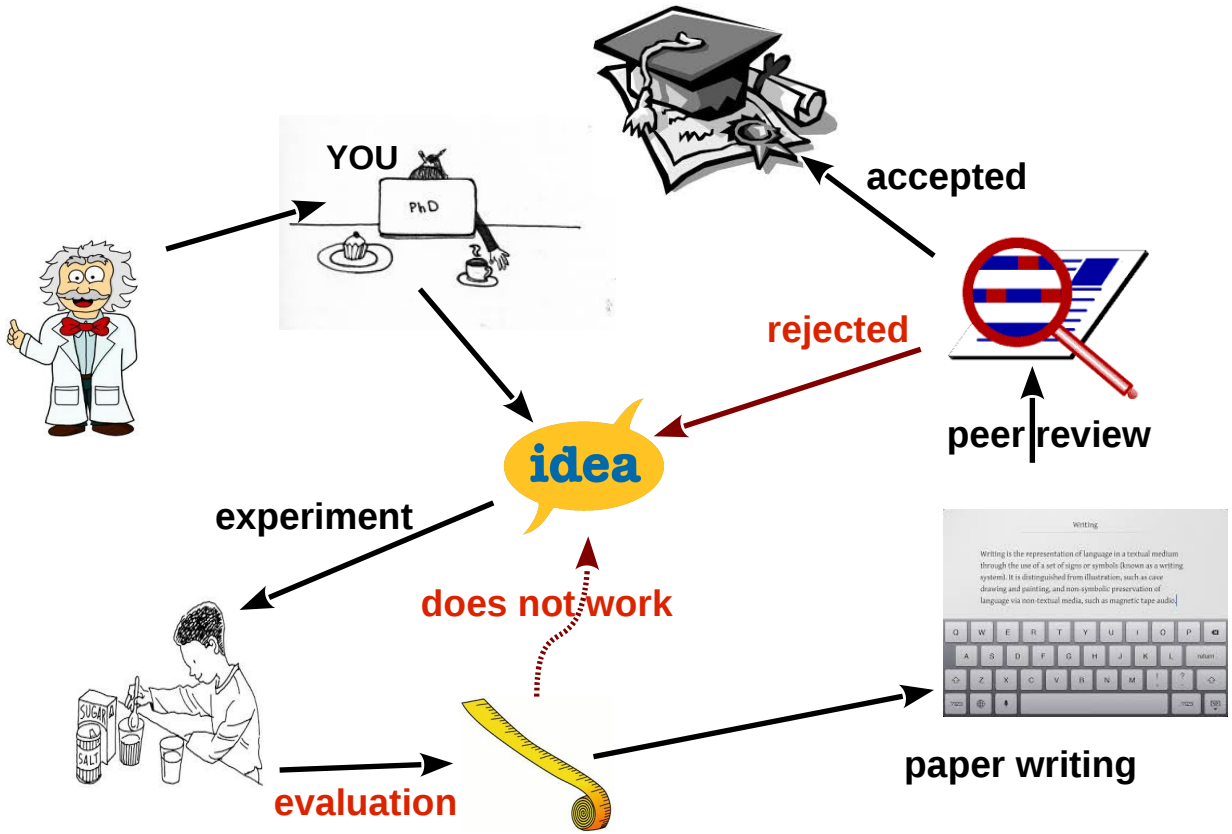
Outline

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

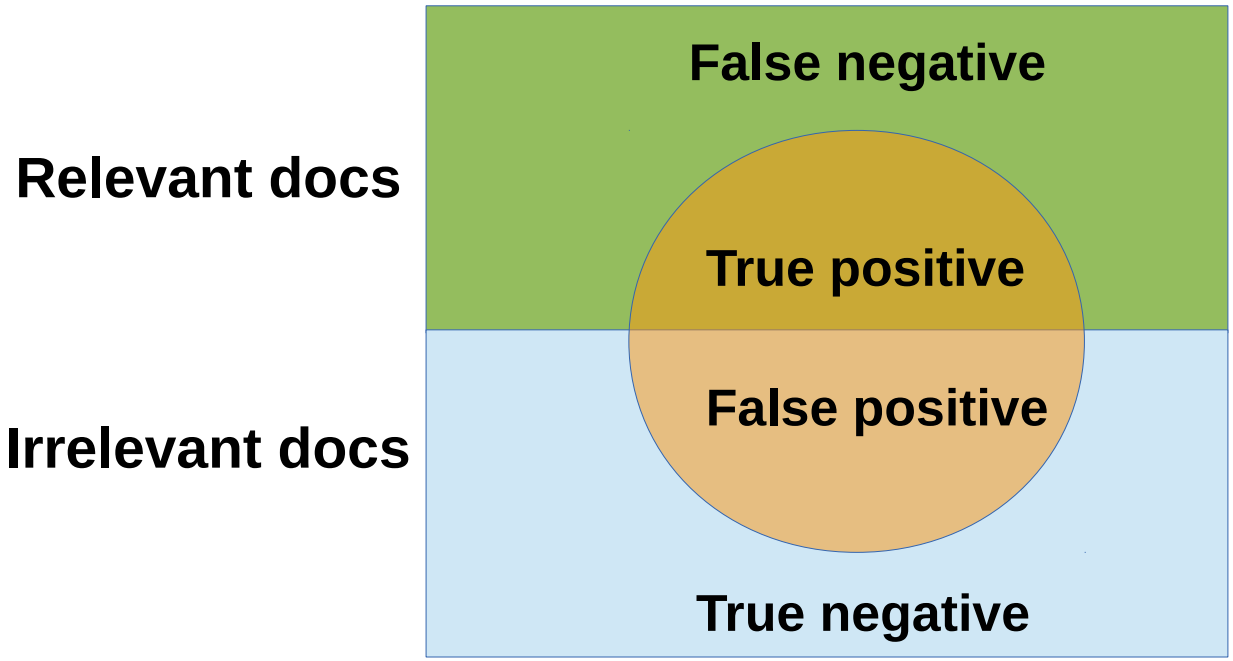
$$Recall = \frac{TP}{R} \tag{2}$$

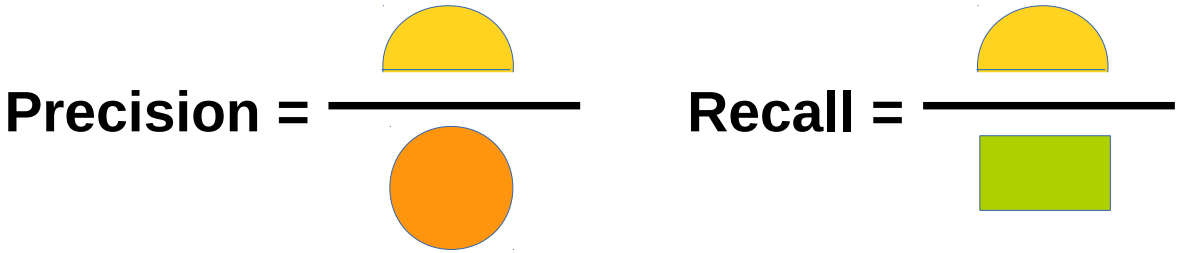
$$Precision = \frac{TP}{K} \tag{3}$$

- F-measure is further defined as

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

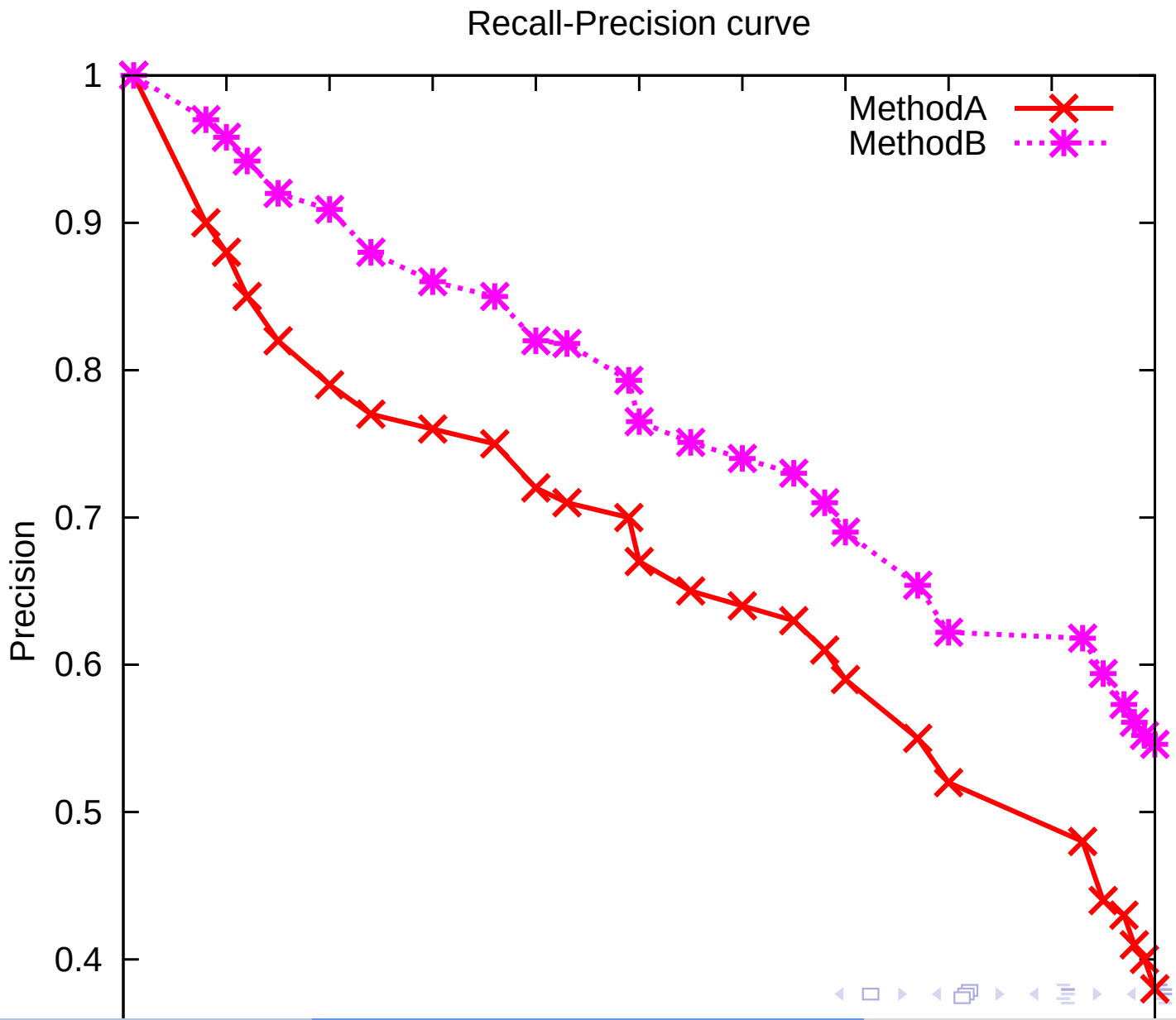
Recall and precision illustration





- 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

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

- mean Average Precision (mAP) is defined as

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

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!