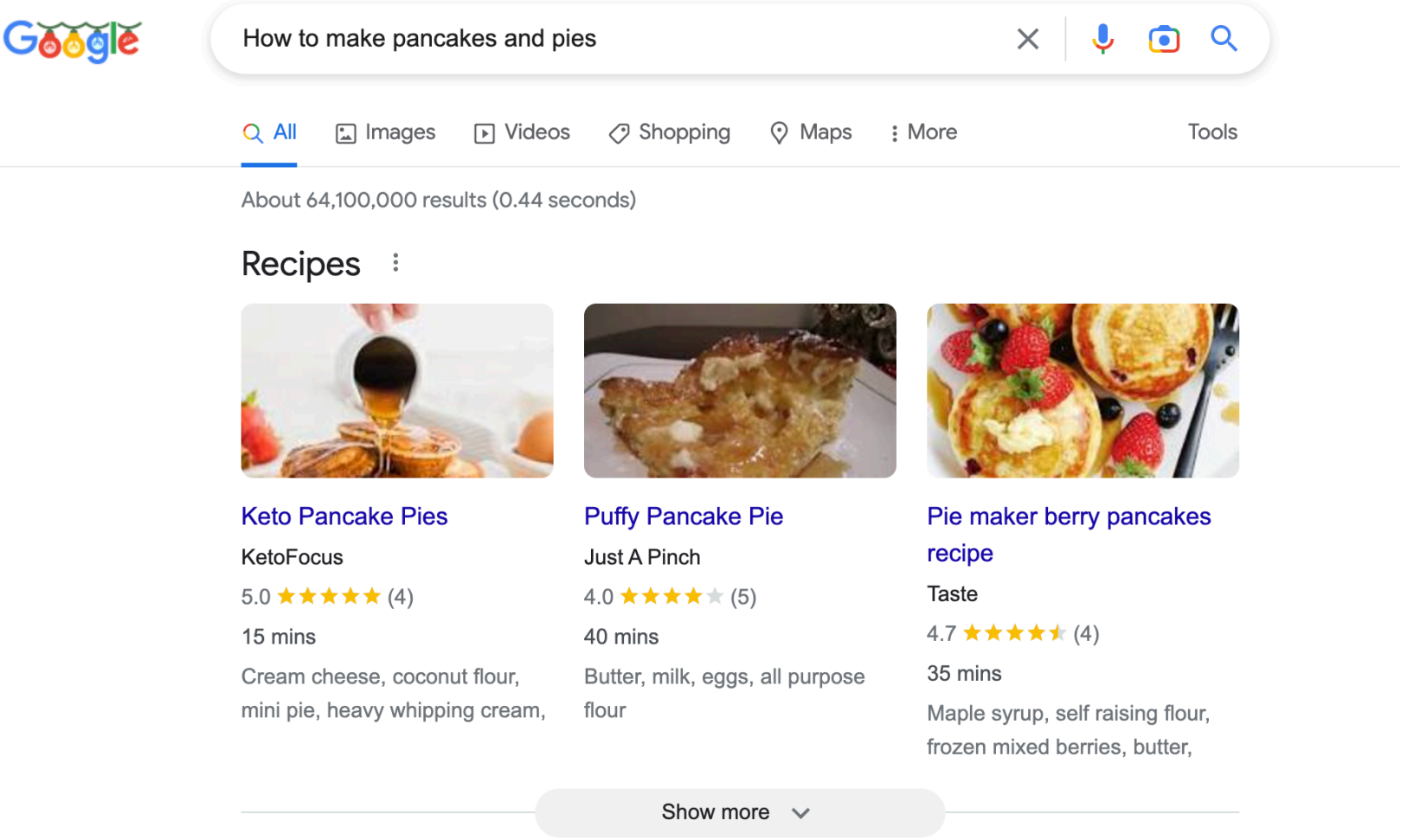
A Relevance Model

Dr. Anne Kayem

Hasso-Plattner-Institute
University of Potsdam

Example: Document Relevance

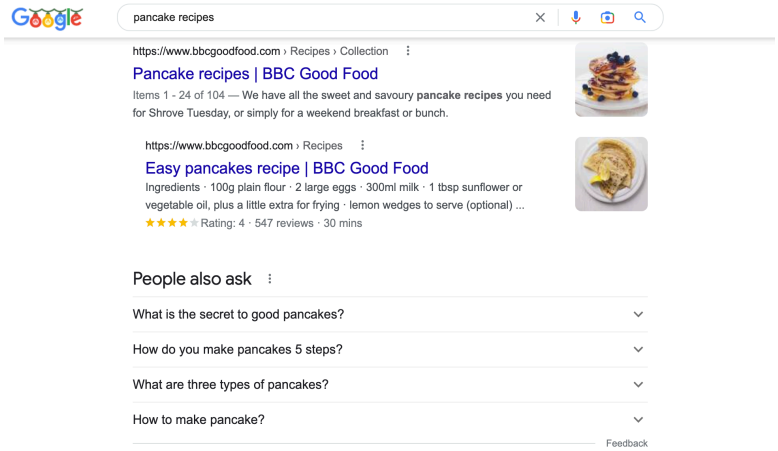


Google's Relevance Model

- **Google PageRank**
  - **Goal:** To achieve high quality search results, the documents extracted from the inverted index must be weighted according to their relevance

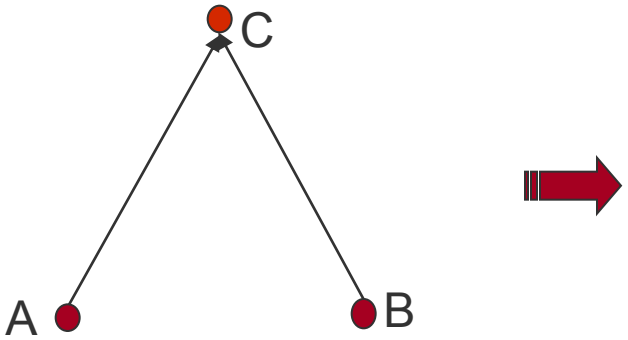
- **Relevance weighting:**
  - Distinguish "important" from "unimportant" documents:

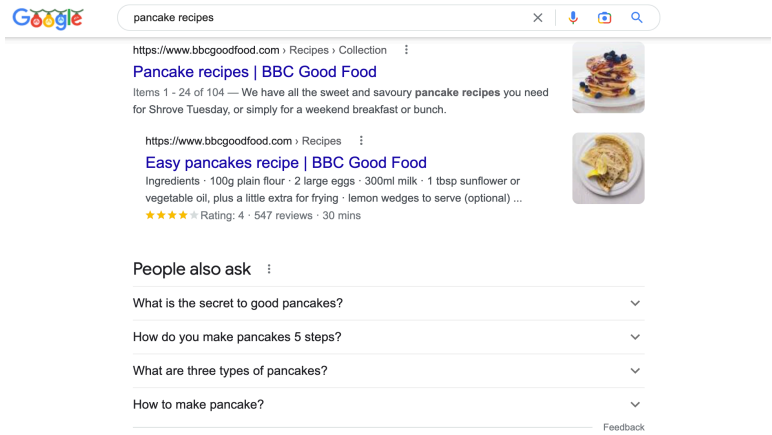
- **Example:**



Google's Relevance Model

- The more references (hyperlinks connecting to it) a document has, the higher its "importance"
- Documents that are referenced by an "important" document are also "important".
- If a document contains several links to other documents, the less "important" each individual link is.





Example - Google Relevance Model

- The "importance"(PageRank, PR) of a document can be obtained as follows:

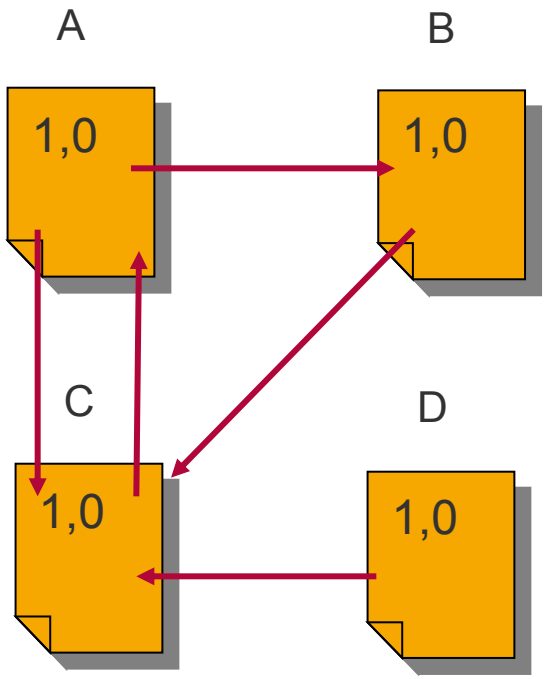
- Let …
  - $PR(A)$ is the PageRank of document A
  - T_1 … TN the documents that link to A
  - $PR(T_1)$ … $PR(TN)$ the PageRanks of T_1 … TN
  - $c(T_i)$ the number of outgoing links in T_i
  - d is the damping factor such that $(0 < d < 1)$

$$PR(A) = (1 - d) + d\left(\sum_{i=1}^{n} \frac{PR(T_i)}{c(T_i)}\right)$$

A worked example (1/2)

- …of PageRank Calculation…

$$PR(A) = (1-d) + d\left(\sum_{i=1}^{n} \frac{PR(T_i)}{c(T_i)}\right)$$

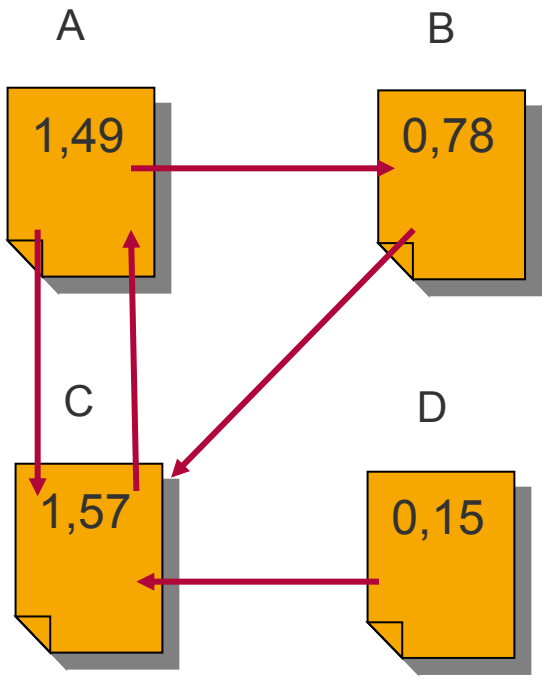


Calculation occurs **iteratively**, until a stable point is reached. For instance for: d = 0,85 we have…

Nr.	PR(A)	PR(B)	PR(C)	PR(D)
1	1,0	1,0	1,0	1,0
2	1,0	0,575	2,275	0,15
3	2,083	0,575	1,1912	0,15
…	…	…	…	…
n	1,49	0,7833	1,577	0,15

A worked example (2/2)

$$PR(A) = (1-d) + d\left(\sum_{i=1}^{n} \frac{PR(T_i)}{c(T_i)}\right)$$



Calculation occurs **iteratively**, until a stable point is reached. For instance for: d = 0,85 we have…

Nr.	PR(A)	PR(B)	PR(C)	PR(D)
1	1,0	1,0	1,0	1,0
2	1,0	0,575	2,275	0,15
3	2,083	0,575	1,1912	0,15
…	…	…	…	…
n	**1,49**	**0,7833**	**1,577**	**0,15**

Other Influencing Factors

- **Relevance feedback**
  - Direct / Indirect Feedback
  - **Assumption:**
    - If a document is selected frequently by users from the results list of a search query, the higher its relevance

- Other factors to be considered:
  - New vs. old documents, lifespan in index dataset
  - Prevent misuse registered IP of client, cookies,…
  - Use high click popularity only with sufficiently good description of content in DESCRIPTION meta tag of HTML header