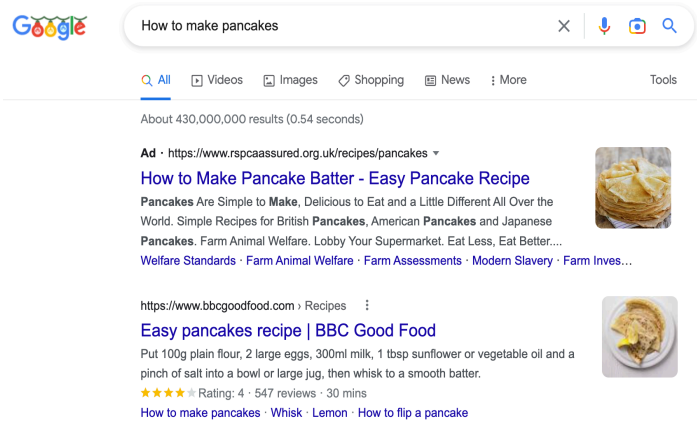
Document Discovery

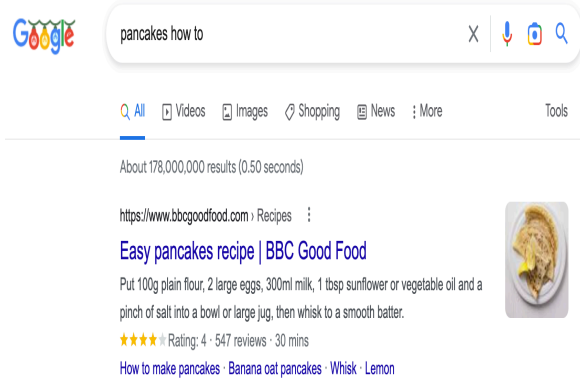
Dr. Anne Kayem

Hasso-Plattner-Institute
University of Potsdam

Document Discovery

- Determining Similarity (RECALL)
- Query Q is represented as a document vector
- Search terms here are the descriptors
- Calculation of the similarity of the document vector Q with all document vectors D





Calculating Similarity

- Text … „Hello everyone“, „This webpage ….“….“which webpage“…

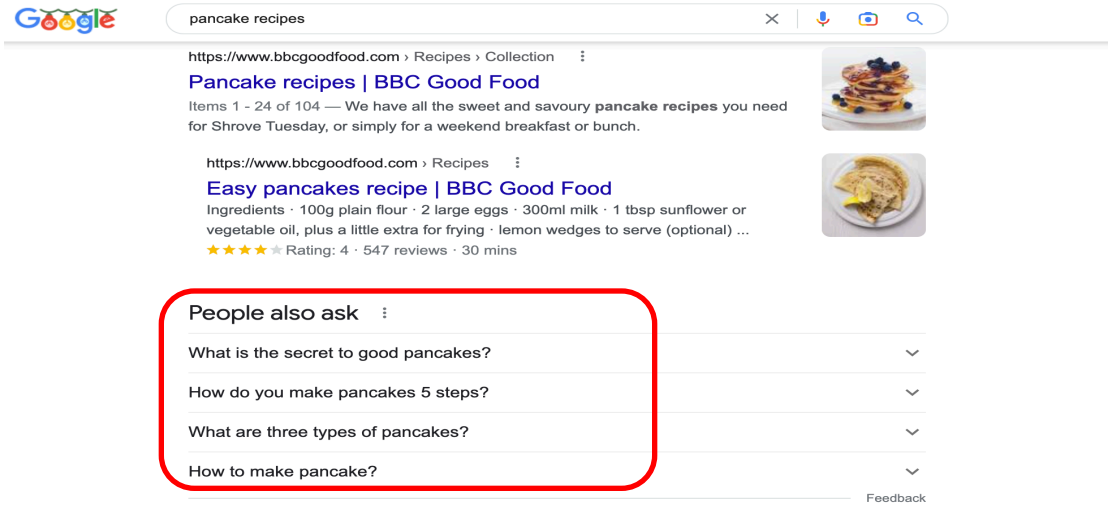
Text	Index
Hello	(1, 1)
everyone	(1, 2)
This	(2, 1)
webpage	(2, 2); (3, 2)

- The similarity between Q and D is given by the **cosine of the angle θ between the two document vectors Q and D**

$$\boldsymbol{sim}(Q, D) = \cos(\theta)$$

Document Clustering

- Similar documents are combined into clusters
- The similarity analysis and cluster assignment is done at the time of index creation
- Analysis of document descriptors using thesauri



Ranking Results

- To achieve high-quality search results, the documents obtained from the inverted index must be weighted according to their relevance

- What is important?
  - Term Frequency Algorithm (TFA)
    - **Zipf's law:** The more often a keyword occurs in a text, the more important it must be

Term Frequency Algorithm

- Term Frequency Algorithm (TFA)
  - **Zipf's law:** The more often a keyword occurs in a text, the more important it must be

- Simplest weight: Absolute word frequency

$$TF(d,t) = n(d,t)$$

- Alternative possibilities:
  - relative word frequency

$$TF(d,t) = \frac{n(d,t)}{\sum_{\tau} n(d,\tau)}$$

Inverse Document Frequency

- Inverse Document Frequency Algorithm

- To distinguish a document, by keyword
  - … in terms of content, and from other documents,
  - … the occurrence of the keyword in other documents (D_t) must also be determined!

- **Example:** … „This webpage ….“….“which webpage does the previous webpage…“…

Text	Index
webpage	(1, 2); (2, 2); (2,6)

TF-IDF – (1/2)

- Term Frequency (TF)
- Inverse Document Frequency (IDF)

- The occurrence of the keyword in other documents (D_t) can then be used to obtain the IDF:

$$IDF(d,t) = \frac{1}{|D_t|}$$

The more documents contain the keyword, the worse it characterizes a single document

- **Example:** … „This webpage ….“….“which **webpage** does the previous **webpage**…“…

Text	Index
webpage	(1, 2); (2, 2); (2,6)

TF-IDF – (2/2)

- Simplest weight: Absolute word frequency

$$TF(d,t) = n(d,t)$$

- Inverse document frequency:

$$IDF(d,t) = \frac{1}{|D_t|}$$

- Term Frequency-Inverse Document Frequency (TF-IDF) is obtained as follows:

$$TFIDF(d,t) = TF(d,t) * IDF(d,t)$$