

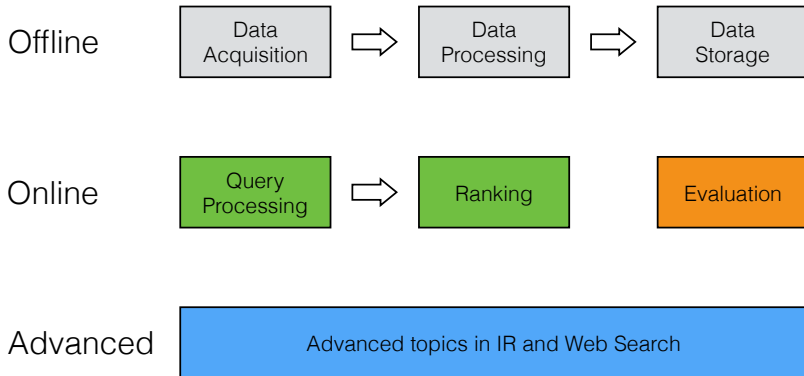
# Information Retrieval

## Data Storage

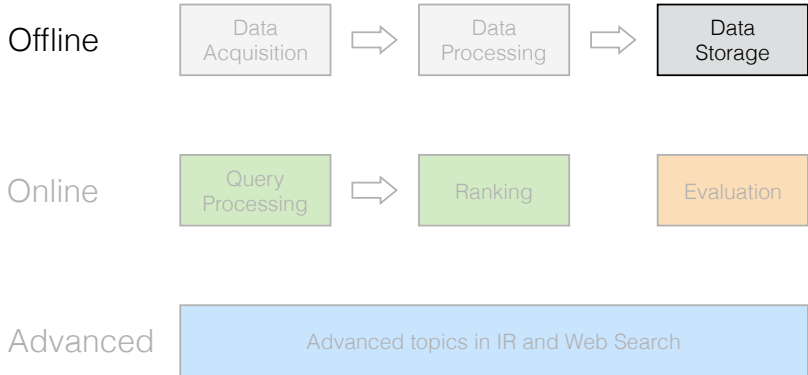
**Ilya Markov**  
i.markov@uva.nl

University of Amsterdam

# Course overview



# This lecture



# Data storage methods

- File
- File system
- Database
- **Index**

# Outline

- 1 Basic indexing architecture
- 2 Inverted index
- 3 Constructing an index
- 4 Updating an index
- 5 Compressing an index
- 6 Partitioning an index
- 7 Summary

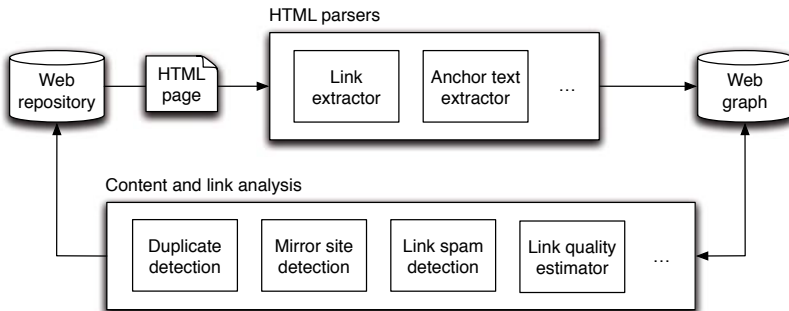
# Outline

- 1 Basic indexing architecture
- 2 Inverted index
- 3 Constructing an index
- 4 Updating an index
- 5 Compressing an index
- 6 Partitioning an index
- 7 Summary

# Basic indexing architecture

- Web graph
- Forward index
- Page attribute file
- Inverted index

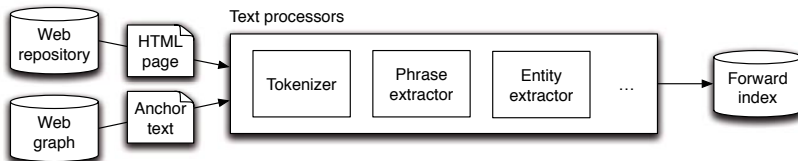
# Web graph



B. Cambazoglu and R. Baeza-Yates, "Scalability Challenges in Web Search Engines"

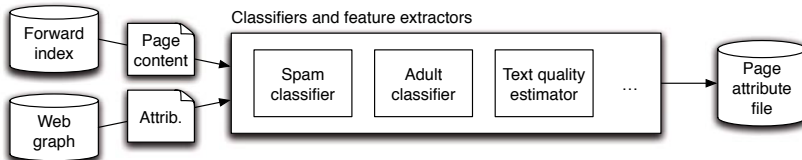


# Forward index



B. Cambazoglu and R. Baeza-Yates, "Scalability Challenges in Web Search Engines"

# Page attribute file



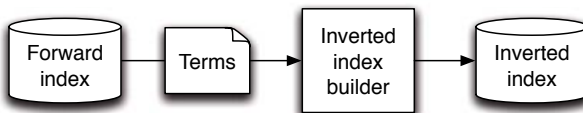
B. Cambazoglu and R. Baeza-Yates, "Scalability Challenges in Web Search Engines"

# Page attribute file

Feature	Source	Description
Language	Page content	Language of the page
Length	Page content	Number of words or characters in the page
Content spam	Page content	Score indicating the likelihood that the page content is spam
Text quality	Page content	Score combining various text quality features (e.g., readability)
Link quality	Web graph	Page importance estimated based on page's link structure
CTR	Query logs	Click-through rate of the page in search results (if available)
Dwell time	Query logs	Average time spent by the users on the page
Page load time	Web server	Average time it takes to receive the page from the server
URL depth	URL	Number of slashes in the absolute path of the URL

B. Cambazoglu and R. Baeza-Yates, "Scalability Challenges in Web Search Engines"

# Inverted index



B. Cambazoglu and R. Baeza-Yates, "Scalability Challenges in Web Search Engines"

# Outline

- 1 Basic indexing architecture
- 2 **Inverted index**
- 3 Constructing an index
- 4 Updating an index
- 5 Compressing an index
- 6 Partitioning an index
- 7 Summary

# Inverted index

## ① Dictionary

- Each entry contains
  - Number of pages containing the term
  - Pointer to the start of the inverted list
  - Other meta-data about the term
- B+ tree, hash table

## ② Inverted lists

# Example

- $S_1$  Tropical fish include fish found in tropical environments around the world, including both freshwater and salt water species.
- $S_2$  Fishkeepers often use the term tropical fish to refer only those requiring fresh water, with saltwater tropical fish referred to as marine fish.
- $S_3$  Tropical fish are popular aquarium fish, due to their often bright coloration.
- $S_4$  In freshwater fish, this coloration typically derives from iridescence, while salt water fish are generally pigmented.

Croft et al., "Search Engines, Information Retrieval in Practice"

# Document identifiers

and	1				only	2			
aquarium	3				pigmented	4			
are	3	4			popular	3			
around	1				refer	2			
as	2				referred	2			
both	1				requiring	2			
bright	3				salt	1	4		
coloration	3	4			saltwater	2			
derives	4				species	1			
due	3				term	2			
environments	1				the	1	2		
fish	1	2	3	4	their	3			
fishkeepers	2				this	4			
found	1				those	2			
fresh	2				to	2	3		
freshwater	1	4			tropical	1	2	3	
from	4				typically	4			
generally	4				use	2			
in	1	4			water	1	2	4	
include	1				while	4			
including	1				with	2			
iridescence	4				world	1			
marine	2								
often	2	3							

Croft et al., "Search Engines, Information Retrieval in Practice"



10. *Journal of the American Medical Association*, 2000; 283: 2689-2696.

[illegible]

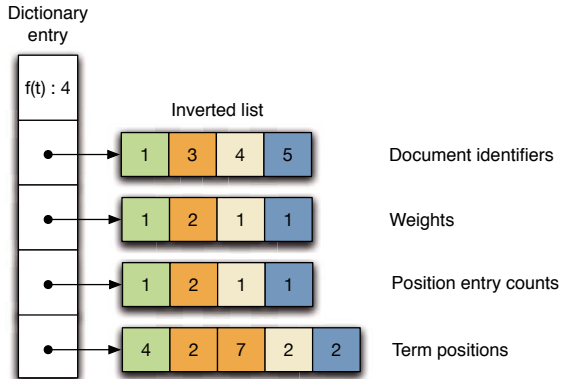
Croft et al., "Search Engines, Information Retrieval in Practice"

# Using positions to deal with phrases

tropical	1,1		1,7	2,6	2,17		3,1			
fish	1,2	1,4		2,7	2,18	2,23	3,2	3,6	4,3	4,13

Croft et al., "Search Engines, Information Retrieval in Practice"

# Full inverted index

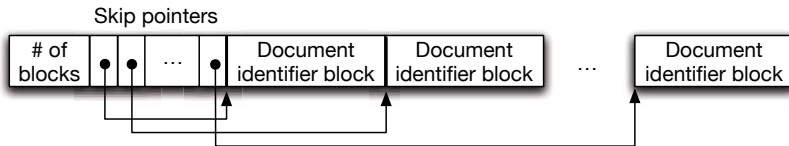


B. Cambazoglu and R. Baeza-Yates, "Scalability Challenges in Web Search Engines"

# Sorting document identifiers

- Identifier-sorted
  - Efficient compression
- Weight-sorted
  - Efficient query processing
- Impact-sorted
  - Weights are quantized
  - Identifier-sorted within each bucket

# Skip pointers



B. Cambazoglu and R. Baeza-Yates, "Scalability Challenges in Web Search Engines"

# Summary

- Inverted lists
  - Document identifiers
  - Frequencies
  - Positions
  - Weights
- Sorting document identifiers
- Skip pointers

# Outline

- 1 Basic indexing architecture
- 2 Inverted index
- 3 Constructing an index**
- 4 Updating an index
- 5 Compressing an index
- 6 Partitioning an index
- 7 Summary



# Simple indexer

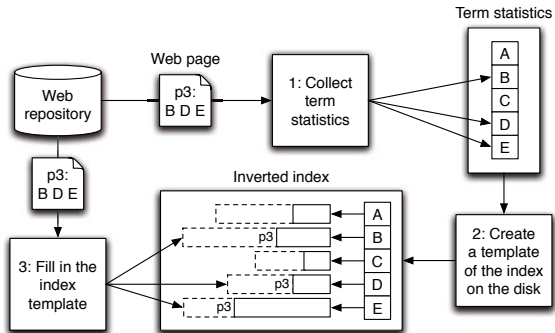
```
procedure BUILDINDEX( $D$ )  
   $I \leftarrow$  HashTable()  
   $n \leftarrow 0$   
  for all documents  $d \in D$  do  
     $n \leftarrow n + 1$   
     $T \leftarrow$  Parse( $d$ )  
    Remove duplicates from  $T$   
    for all tokens  $t \in T$  do  
      if  $I_t \notin I$  then  
         $I_t \leftarrow$  Array()  
      end if  
       $I_t.append(n)$   
    end for  
  end for  
  return  $I$   
end procedure
```

Croft et al., "Search Engines, Information Retrieval in Practice"

# What are the problems with this simple indexer?

- ① In-memory
  - Two-pass index
  - One-pass index with merging
- ② Single-threaded
  - Distributed indexing

# Two-pass index



B. Cambazoglu and R. Baeza-Yates, "Scalability Challenges in Web Search Engines"

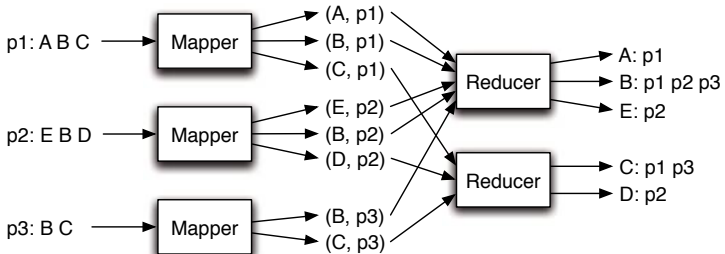
*Journal of Management Education* 36(7) 809–824

# Aardvark



Picture taken from <https://en.wikipedia.org/wiki/Aardvark>

# Distributed indexing (MapReduce)



B. Cambazoglu and R. Baeza-Yates, "Scalability Challenges in Web Search Engines"

# Summary

- ① In-memory problem
  - Two-pass index
  - One-pass index with merging
- ② Single-threaded problem
  - Distributed indexing

# Outline

- 1 Basic indexing architecture
- 2 Inverted index
- 3 Constructing an index
- 4 Updating an index
- 5 Compressing an index
- 6 Partitioning an index
- 7 **Summary**



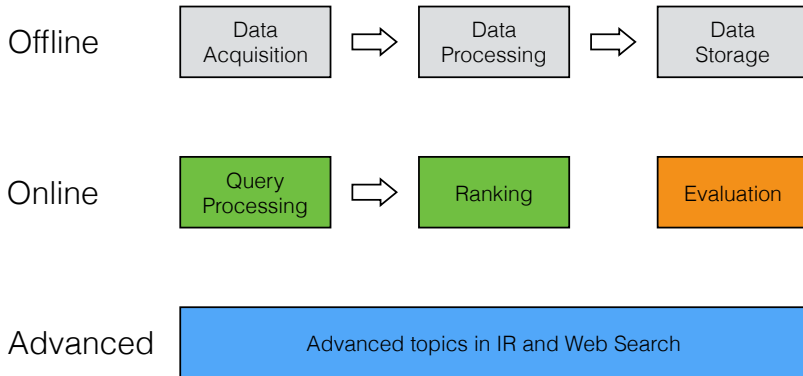
# Outline

- 1 Basic indexing architecture
- 2 Inverted index
- 3 Constructing an index
- 4 Updating an index
- 5 Compressing an index
- 6 Partitioning an index
- 7 Summary

# Materials

- Croft et al., Chapter 5
- Manning et al., Chapters 1.2–1.3, 2.3–2.4
- B. Barla Cambazoglu and Ricardo Baeza-Yates  
**Scalability Challenges in Web Search Engines**  
Morgan & Claypool Publishers, 2017

# Course overview



# Next lecture

