Information Retrieval 101

Svitlana Vakulenko

Postdoctoral researcher
Informatics Institute
University of Amsterdam

October 15, 2020

Information Retrieval History

- 3rd century BC -> library catalogue

- **1950s** -> keyword index + ranked retrieval
- **1960s** -> vector space model

- **1990s** -> probabilistic relevance model

- Sanderson, M., & Croft, W. B. (2012). The history of information retrieval research. *Proceedings of the IEEE*, *100* (Special Centennial Issue), 1444-1451.

Information Retrieval 101: Outline

Search Task

Problem Statement

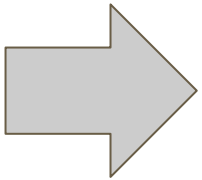
- query
- document collection

$$Q =< q_1 \dots q_n >$$
$$C = \{D_1 \dots D_m\}$$

- document ranking

$$F(R|Q, D_i)$$

Information Retrieval: Approach



Index

- D1 = {train, zoo, robert}
- D2 = {ana, robert}
- D3 = {train, zoo}

Inverted index

- D1 = {train, zoo, robert}
- D2 = {ana, robert}
- D3 = {train, zoo}

- train: <D1, D3>
- zoo = <D1, D3>
- robert = <D1, D2>
- ana = <D2>

Probabilistic Model: BM25

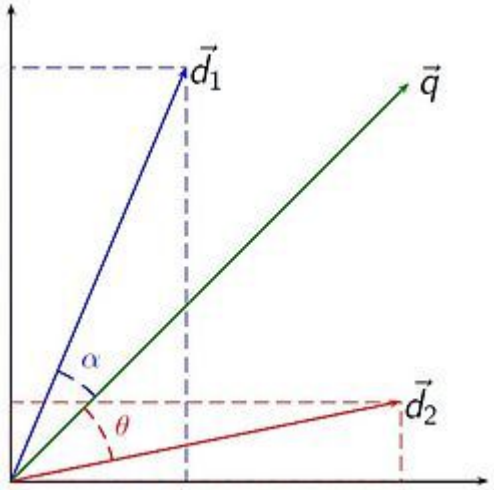
$$\text{score}(D, Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})}$$

where $\text{IDF}(q_i) = -\log \frac{n(q)}{N} = \log \frac{N}{n(q)}$

- Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M., & Gatford, M. (1999). Okapi at TREC-3. Proceedings of the Third Text REtrieval Conference (TREC 1994). In *Gazxithersburg, USA*.

Information Retrieval 101: Outline

Vector Space Model

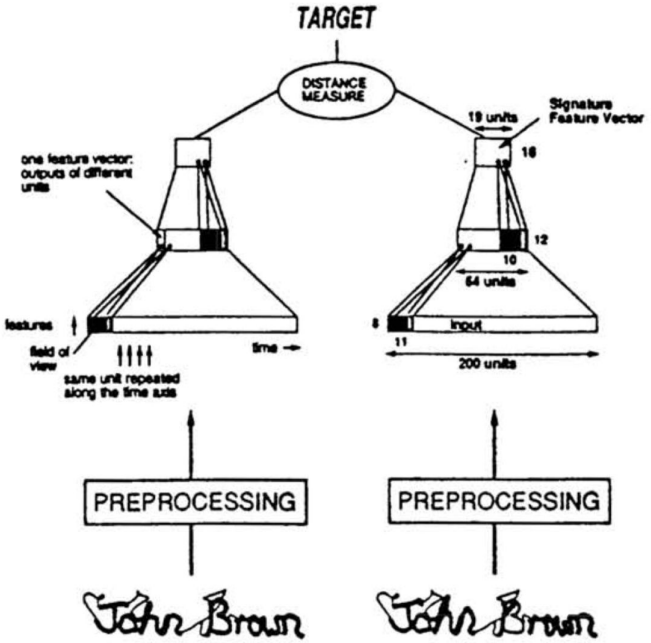


$$\cos\theta = \frac{\mathbf{d_2} \cdot \mathbf{q}}{\|\mathbf{d_2}\| \, \|\mathbf{q}\|}$$

Term vectors

- D1 = {train, zoo, robert}
  1110
- D2 = {ana, robert}
  0011
- D3 = {train, zoo}
  1100

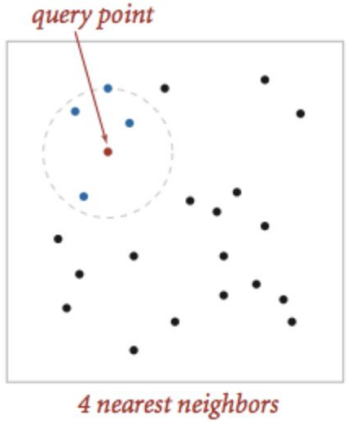
Dual-encoder Architecture



- Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., & Shah, R. (1994). Signature verification using a "siamese" time delay neural network. In *Advances in neural information processing systems* (pp. 737-744).

k-Nearest Neighbor Search

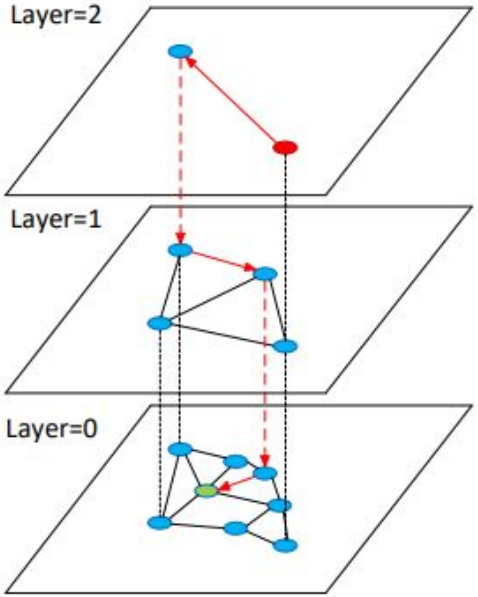
- k similar out of n documents
- similarity function: cosine
- O(n)



Hierarchical Navigable Small World

- bounded degree m
- Search: O(logn)

- Construction: O(nlogn)
- p ~ Exp(l)
- Space: O(lmn)



- Malkov, Y. A., & Yashunin, D. A. (2018). Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*.

Passage Retrieval

- 21M passages of 100 words from English Wikipedia

Training	Retriever	Top-20					Top-100				
		NQ	TriviaQA	WQ	TREC	SQuAD	NQ	TriviaQA	WQ	TREC	SQuAD
None	BM25	59.1	66.9	55.0	70.9	68.8	73.7	76.7	71.1	84.1	80.0
Single	DPR	78.4	79.4	73.2	79.8	63.2	85.4	**85.0**	81.4	89.1	77.2
	BM25 + DPR	76.6	79.8	71.0	85.2	**71.5**	83.8	84.5	80.5	92.7	**81.3**
Multi	DPR	**79.4**	78.8	**75.0**	**89.1**	51.6	**86.0**	84.7	**82.9**	93.9	67.6
	BM25 + DPR	78.0	**79.9**	74.7	88.5	66.2	83.9	84.4	82.3	**94.1**	78.6

- Karpukhin, V., Oğuz, B., Min, S., Wu, L., Edunov, S., Chen, D., & Yih, W. T. (2020). Dense Passage Retrieval for Open-Domain Question Answering. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*

Information Retrieval 101: Summary

Information Retrieval 101: Summary

Collection

Query

Ranking

Information Retrieval 101: Summary

Collection

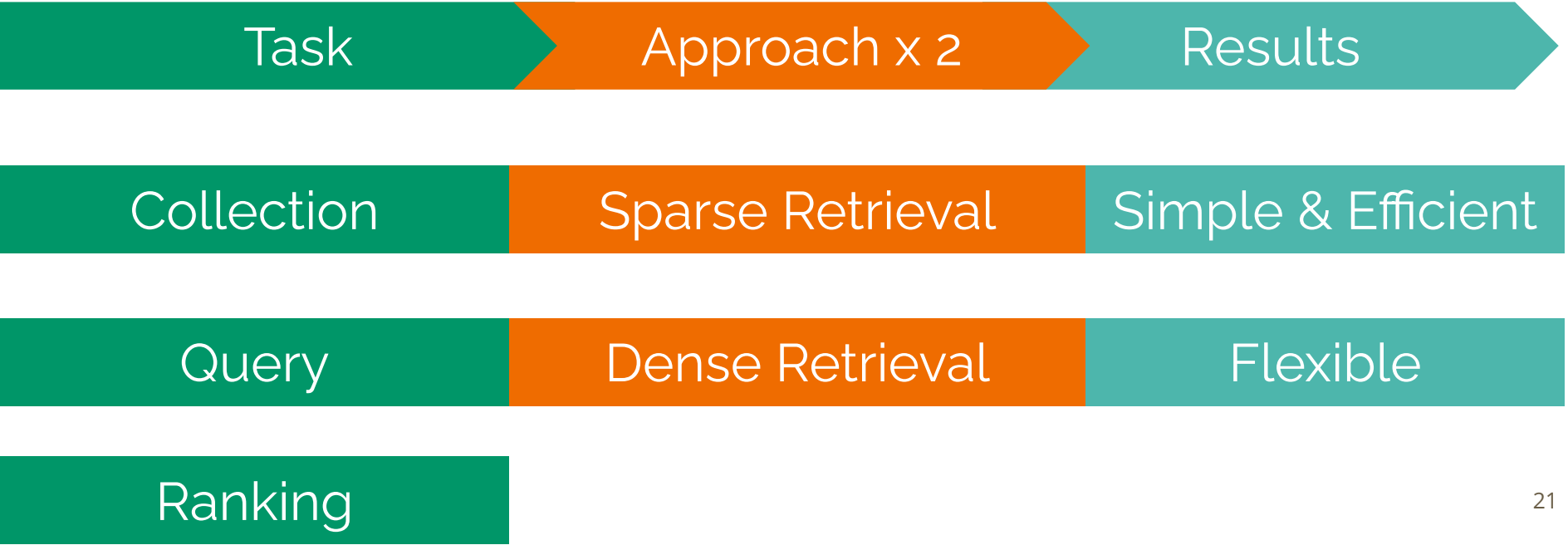
Sparse Retrieval

Query

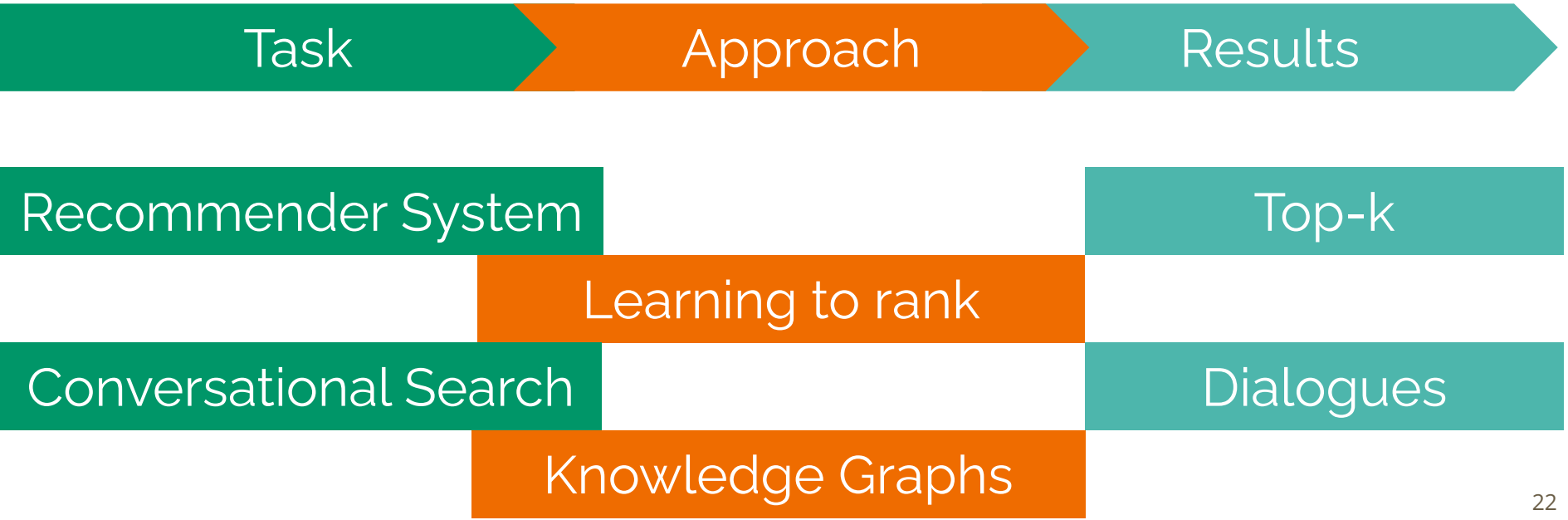
Dense Retrieval

Ranking

Information Retrieval 101: Summary



Information Retrieval: Overview



Information Retrieval: Overview