Construction of Compact Retrieval Models

Unifying Framework and Analysis

Benno Stein and Martin Potthast
Web Technology and Information Systems
Bauhaus University Weimar

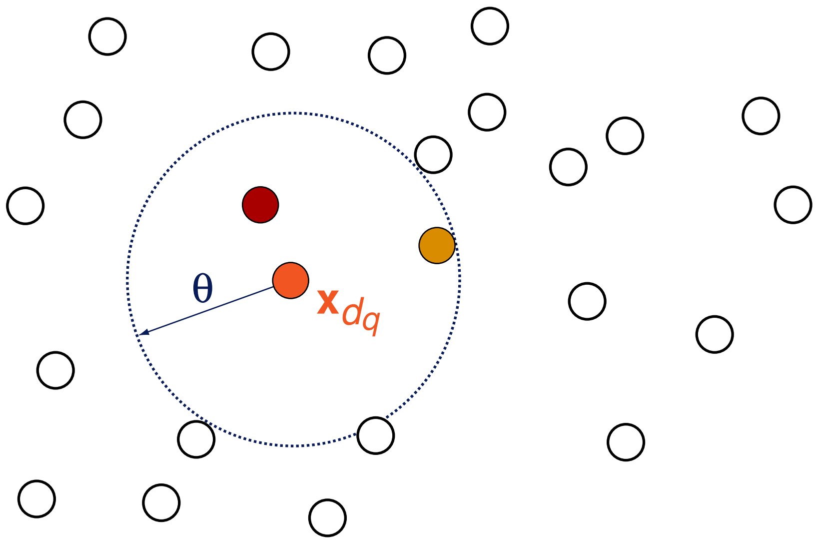
Outline

Introduction

Given a passage of text,
find all the books containing something similar.

Introduction

Nearest Neighbor Search

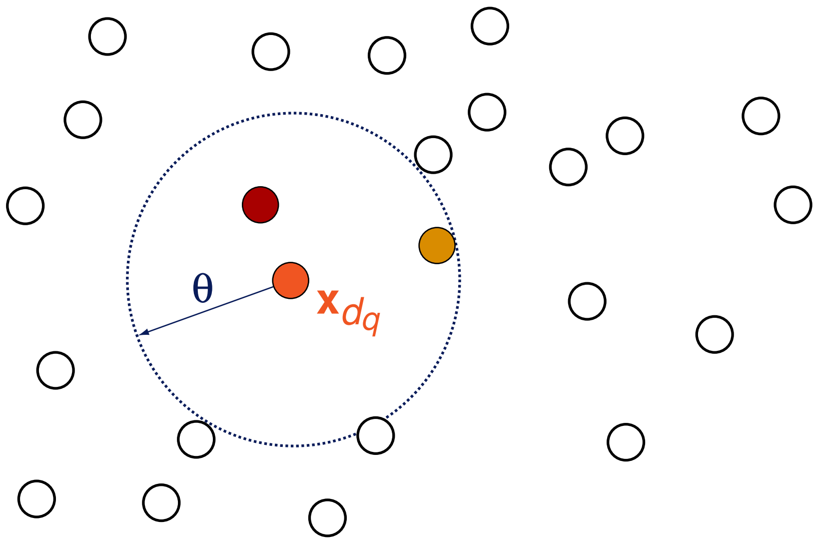


Applications:

- elimination of duplicates / near duplicates
- identification of versioned and plagiarized documents
- retrieval of similar documents
- identification of source code plagiarism

Introduction

Nearest Neighbor Search



The nearest neighbor problem cannot be solved efficiently in high dimensions by partitioning methods.

> *"Existing methods are outperformed on average by a simple sequential scan, if the number of dimensions exceeds around 10."*

[Weber 99, Gionis/Indyk/Motwani 99-04]

Framework

Options for retrieval speed up:

- Dimension reduction
- Fingerprinting

Token sequence → Chunking → High-dim. represen.

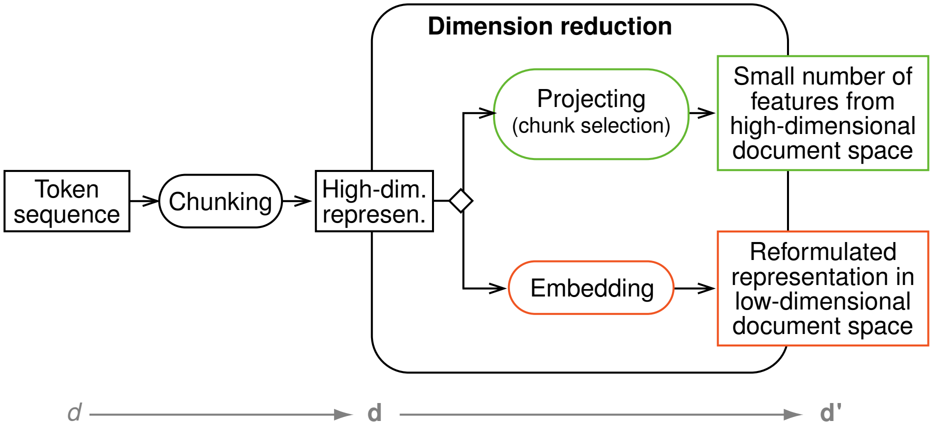
$d \longrightarrow \mathbf{d}$

$$\begin{pmatrix} 0.02\\0.0\\0.01\\0.0\\0.0\\\vdots\\0.0\\0.02\\0.07\\0.0 \end{pmatrix} \begin{pmatrix} 0.1\\0.2\\0.0\\0.1\\0.2\\\vdots\\0.1\\0.3\\0.0\\0.0 \end{pmatrix} \begin{pmatrix} 0.0\\0.1\\0.0\\0.04\\0.0\\\vdots\\0.0\\0.04\\0.0\\0.03 \end{pmatrix} \quad \ldots \quad \begin{pmatrix} 0.07\\0.0\\0.0\\0.1\\0.0\\\vdots\\0.01\\0.02\\0.03\\0.0 \end{pmatrix} \begin{pmatrix} 0.0\\0.0\\0.05\\0.1\\0.0\\\vdots\\0.08\\0.0\\0.0\\0.0 \end{pmatrix} \begin{pmatrix} 0.02\\0.0\\0.01\\0.0\\0.0\\\vdots\\0.0\\0.06\\0.09\\0.0 \end{pmatrix} \begin{pmatrix} 0.0\\0.01\\0.06\\0.0\\0.01\\\vdots\\0.0\\0.0\\0.0\\0.03 \end{pmatrix} \begin{pmatrix} 0.04\\0.0\\0.0\\0.0\\0.05\\\vdots\\0.01\\0.02\\0.03\\0.0 \end{pmatrix} \begin{pmatrix} 0.0\\0.0\\0.01\\0.0\\0.0\\\vdots\\0.0\\0.02\\0.06\\0.0 \end{pmatrix} \begin{pmatrix} 0.1\\0.0\\0.09\\0.0\\0.0\\\vdots\\0.0\\0.0\\0.0\\0.05 \end{pmatrix}$$

Framework

Options for retrieval speed up:

- ❑ Dimension reduction
- ❑ Fingerprinting

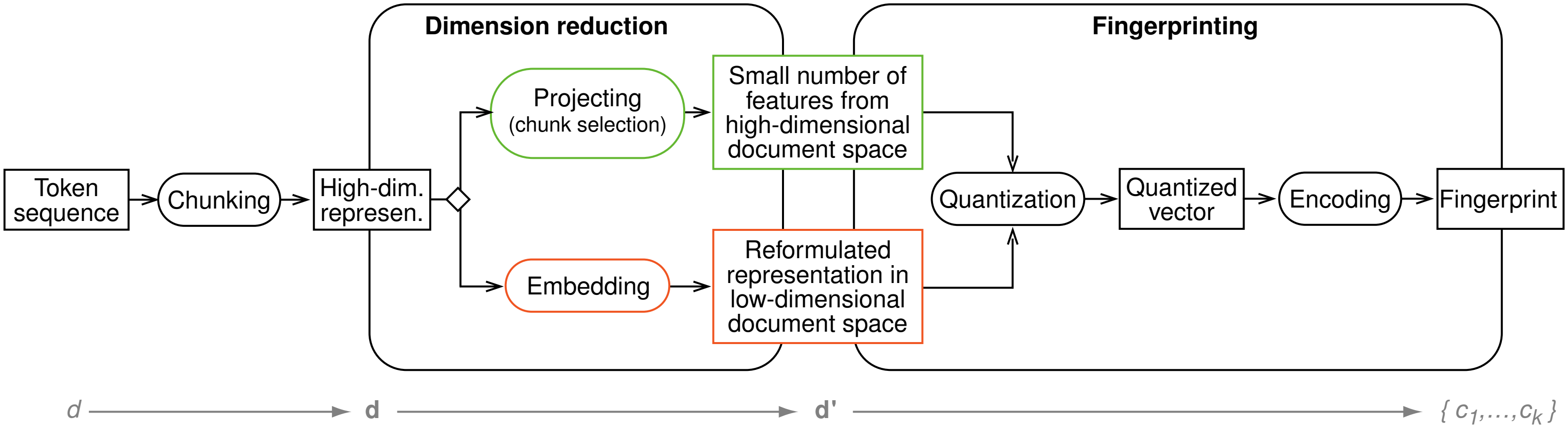


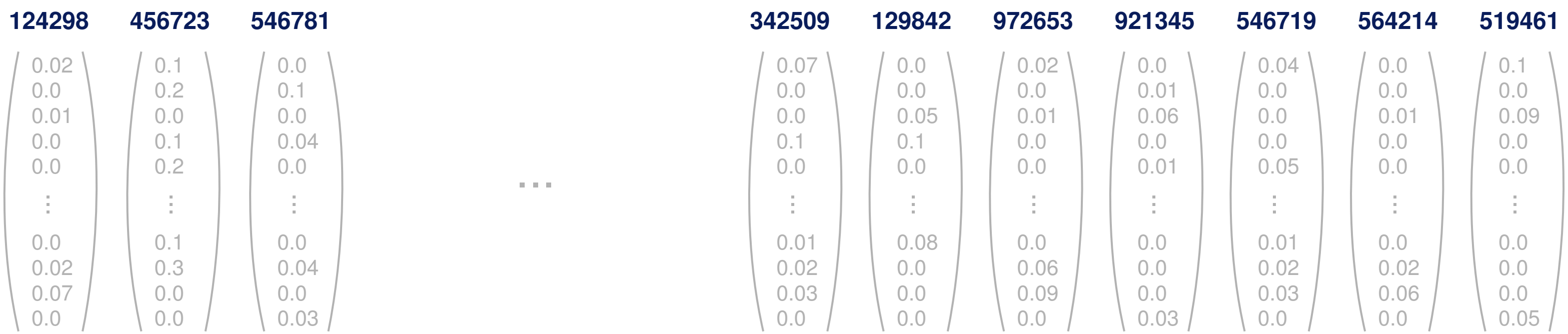
0.02	0.1	0.0		0.07	0.0	0.02	0.0	0.04	0.0	0.1
0.0	0.2	0.1		0.0	0.0	0.0	0.01	0.0	0.0	0.0
0.01	0.0	0.0		0.0	0.05	0.01	0.06	0.0	0.01	0.09
0.0	0.1	0.04		0.1	0.1	0.0	0.0	0.0	0.0	0.0
0.0	0.2	0.0	...	0.0	0.0	0.0	0.01	0.05	0.0	0.0
⋮	⋮	⋮		⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.0	0.1	0.0		0.01	0.08	0.0	0.0	0.01	0.0	0.0
0.02	0.3	0.04		0.02	0.0	0.06	0.0	0.02	0.02	0.0
0.07	0.0	0.0		0.03	0.0	0.09	0.0	0.03	0.06	0.0
0.0	0.0	0.03		0.0	0.0	0.0	0.03	0.0	0.0	0.05

Framework

Options for retrieval speed up:

- Dimension reduction
- Fingerprinting





Part 1 of the Framework

Dimension Reduction

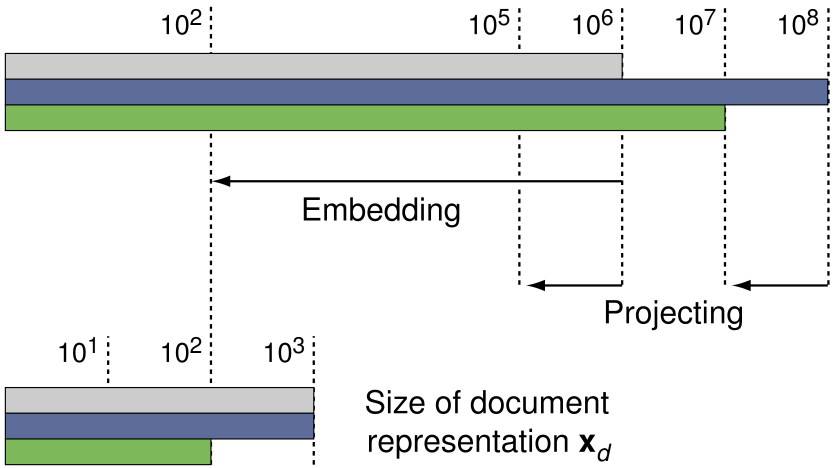
Compact Retrieval Models

Dimension Reduction

	Alternative 1	Alternative 2
Dimension reduction	Projecting	Embedding
Rationale	Hypothesis Test	Model Fidelity
Implementation	Shingling	Fuzzy-Fingerprinting

English Wikipedia:

Dictionary	Number of dimensions
1-gram space	3 921 588
4-gram space	274 101 016
8-gram space	373 795 734
Shingling space	75 659 644



Compact Retrieval Models

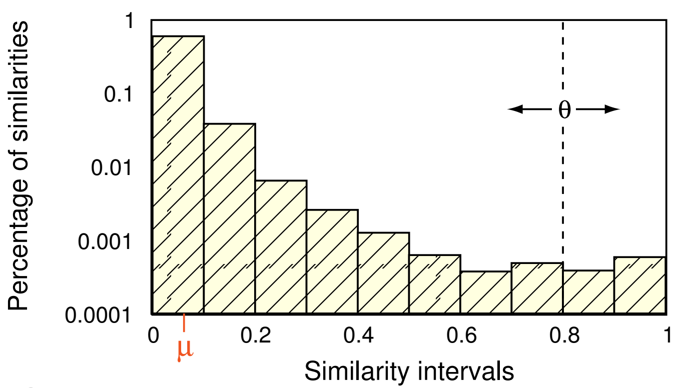
Alternative 1: Projecting / Hypothesis Test

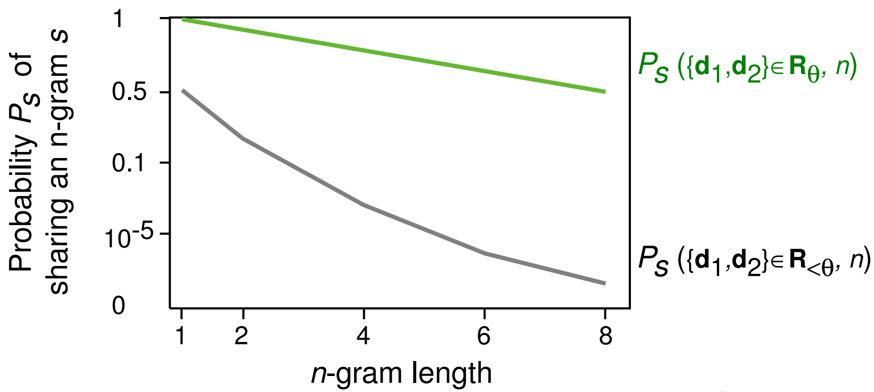
Consideration: If two documents share an n-gram,
does this tell us something about their similarity?

$$H_0 : \text{“}\{\mathbf{d}_1, \mathbf{d}_2\} \text{ is from } \mathbf{R}_{<\theta}\text{”}$$

$$H_1 : \text{“}\{\mathbf{d}_1, \mathbf{d}_2\} \text{ is from } \mathbf{R}_{\theta}\text{”}$$

$$\frac{|\mathbf{R}_{<\theta}| \cdot P_s\big(\{\mathbf{d}_1,\mathbf{d}_2\}\in\mathbf{R}_{<\theta},\ n{=}8\big)}{|\mathbf{R}_{\theta}| \cdot P_s\big(\{\mathbf{d}_1,\mathbf{d}_2\}\in\mathbf{R}_{\theta},\ n{=}8\big)} \quad \sim \quad \frac{P_0}{P_1}$$





Compact Retrieval Models

Alternative 2: Embedding / Model Fidelity

Consideration: If the low-dimensional vector space resembles the similarity relations of the high-dimensional vector space, retrieval with the former works just as well as with the latter.

Multidimensional scaling (MDS) $\Rightarrow$ Singular Value Decomposition (SVD)

On the downside:

- The computation of a SVD has a high runtime complexity.
- The SVD models noise.

Heuristic methods for MDS are at hand.

$\mathbf{X}$. Objects in (high-dimensional) original space, with cos-similarity matrix $\mathbf{S}$.

$\mathbf{Y}$. Objects in k-dimensional embedding space, with cos-similarity matrix $\widehat{\mathbf{S}}$.

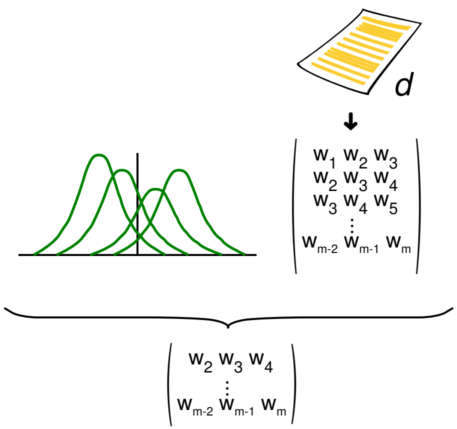
$\mathbf{S} = \mathbf{X}^T\mathbf{X}$, if the $\mathbf{x} \in \mathbf{X}$ are normalized under the l_2-norm.

SVD of $\mathbf{X}$ yields the optimum embedding $\mathbf{Y}_{SVD}$: $\quad \widehat{\mathbf{S}}^* = \mathbf{V}_k \mathbf{\Sigma}_k^2 \mathbf{V}_k^T =: \mathbf{Y}_{SVD}^T \mathbf{Y}_{SVD}$

Compact Retrieval Models

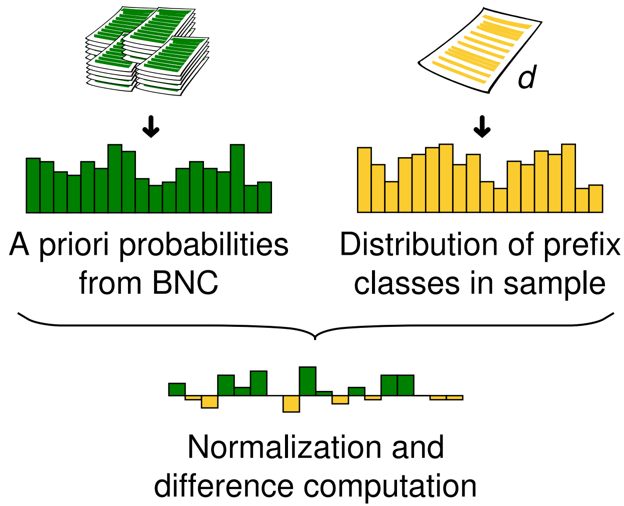
Alternative 1 vs. Alternative 2

Shingling



- Random functions.

Fuzzy-Fingerprinting



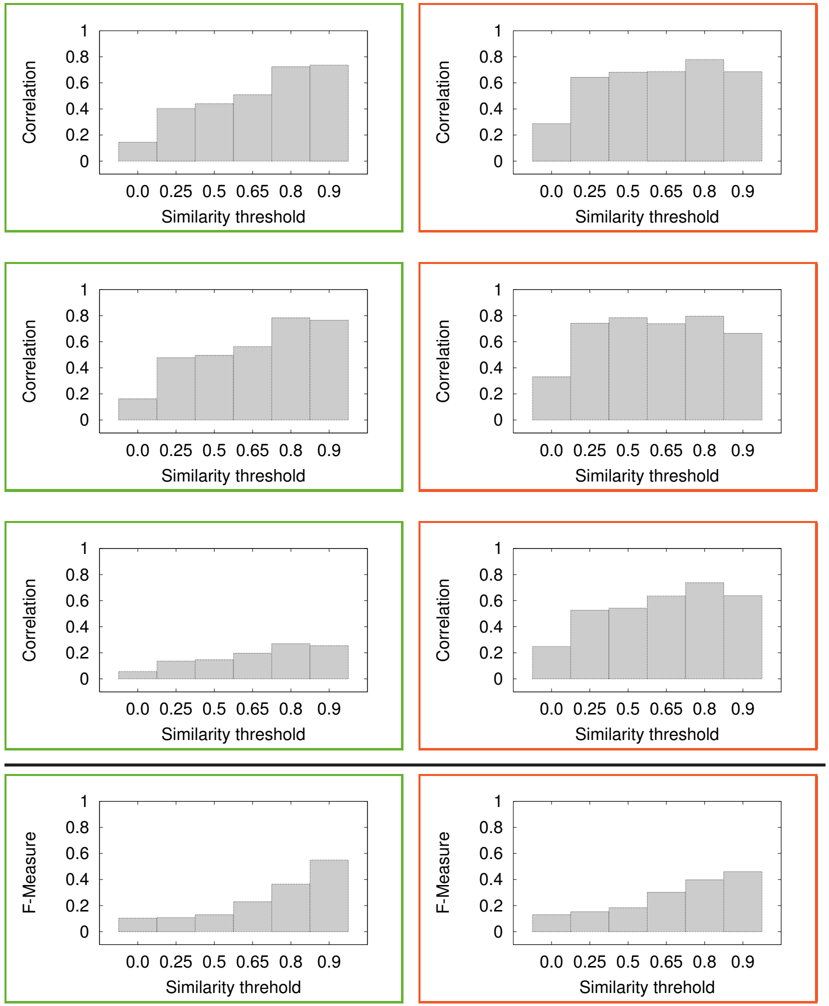
- Documents from the British National Corpus

Document model size

100

2

Shingling | Fuzzy-Fingerprinting



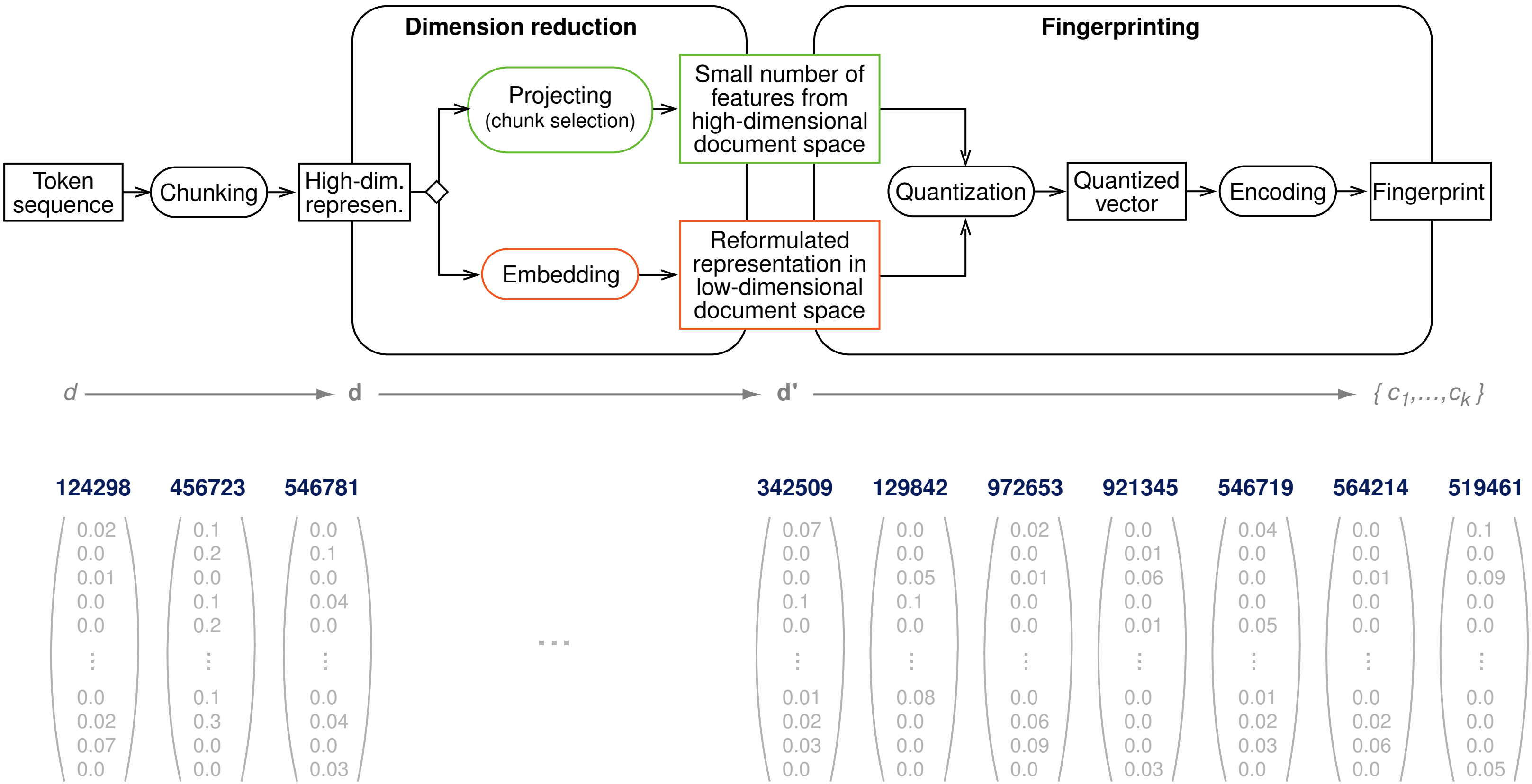
Part 2 of the Framework

Fingerprinting

Framework

Options for retrieval speed up:

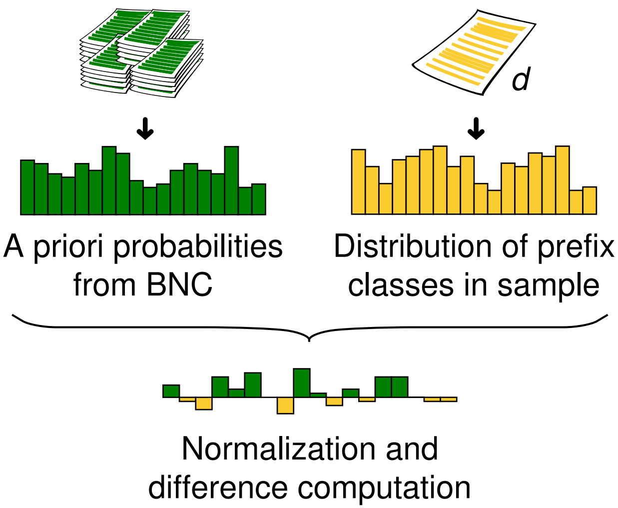
- Dimension reduction
- Fingerprinting



Fingerprinting

Fuzzy-Fingerprinting

Embedding → Quantization → Encoding

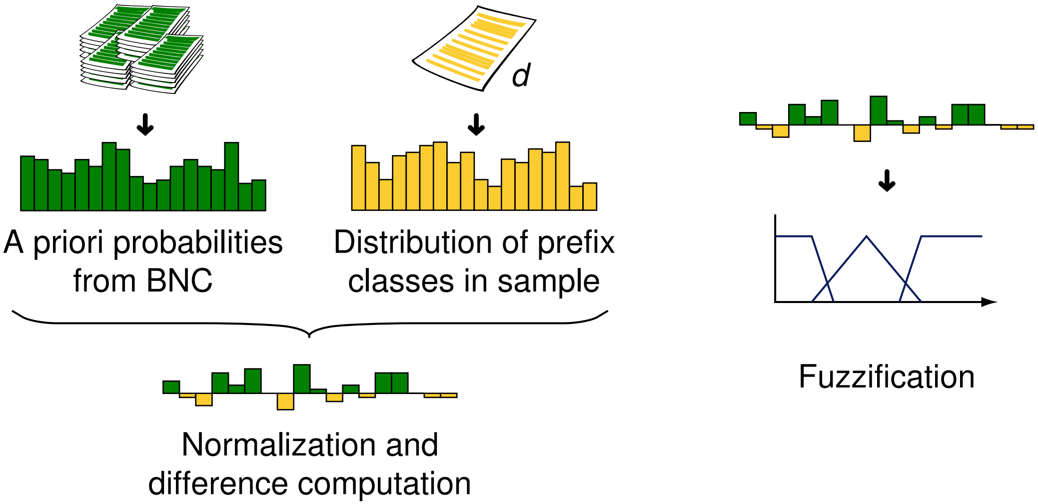


- Documents from the British National Corpus

Fingerprinting

Fuzzy-Fingerprinting

Embedding → **Quantization** → Encoding

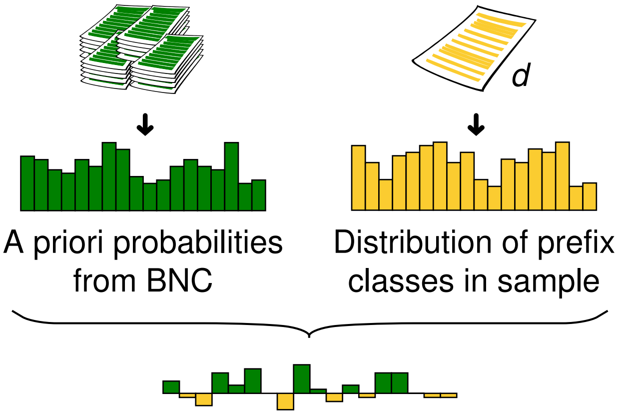


- Documents from the British National Corpus

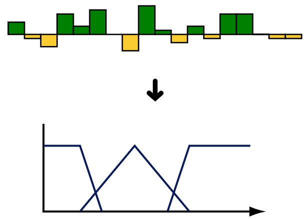
Fingerprinting

Fuzzy-Fingerprinting

Embedding ➜ Quantization ➜ **Encoding**



Normalization and difference computation



Fuzzification

$$h_{\varphi}^{(\rho)}(\mathbf{x}_d) = \sum_{i=1}^{k} \rho(y_i) \cdot r^{i-1}$$

- Documents from the British National Corpus

➜ Fingerprint = 2643256

Fingerprinting

Wikipedia in the Pocket

Wikipedia in the Pocket

Indexing Technology for Plagiarism Detection,
Near-duplicate Detection, and High Similarity Search

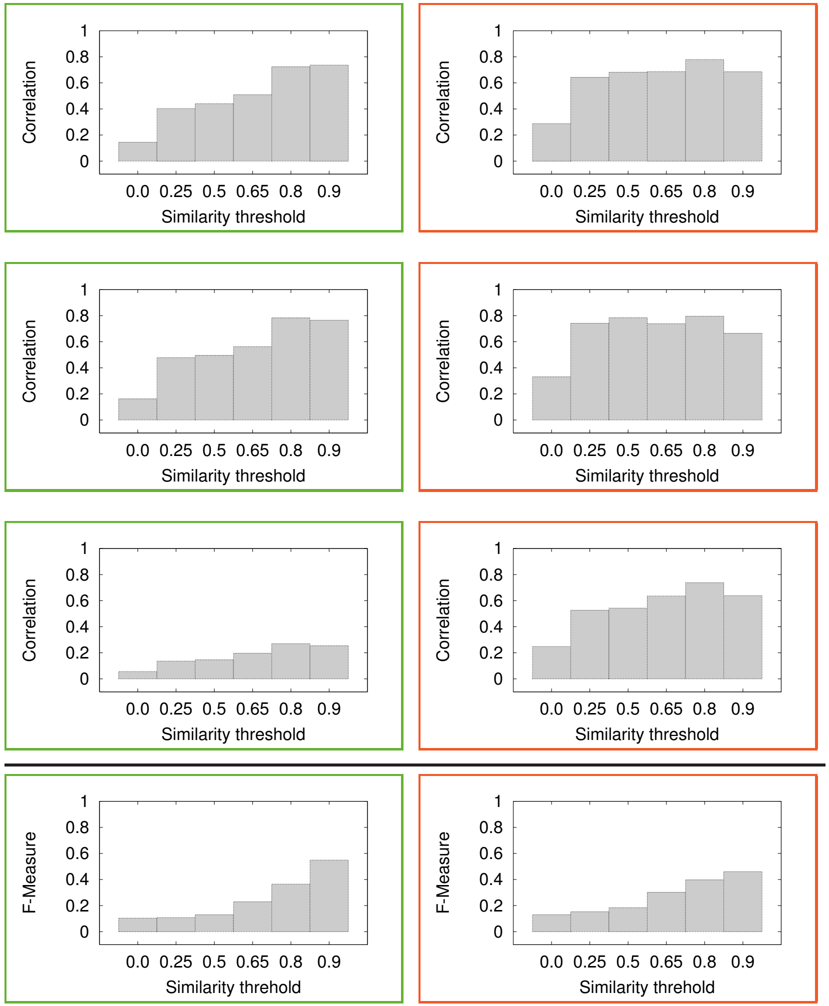
www.uni-weimar.de/medien/webis/research/wipo

Document model size

100

2

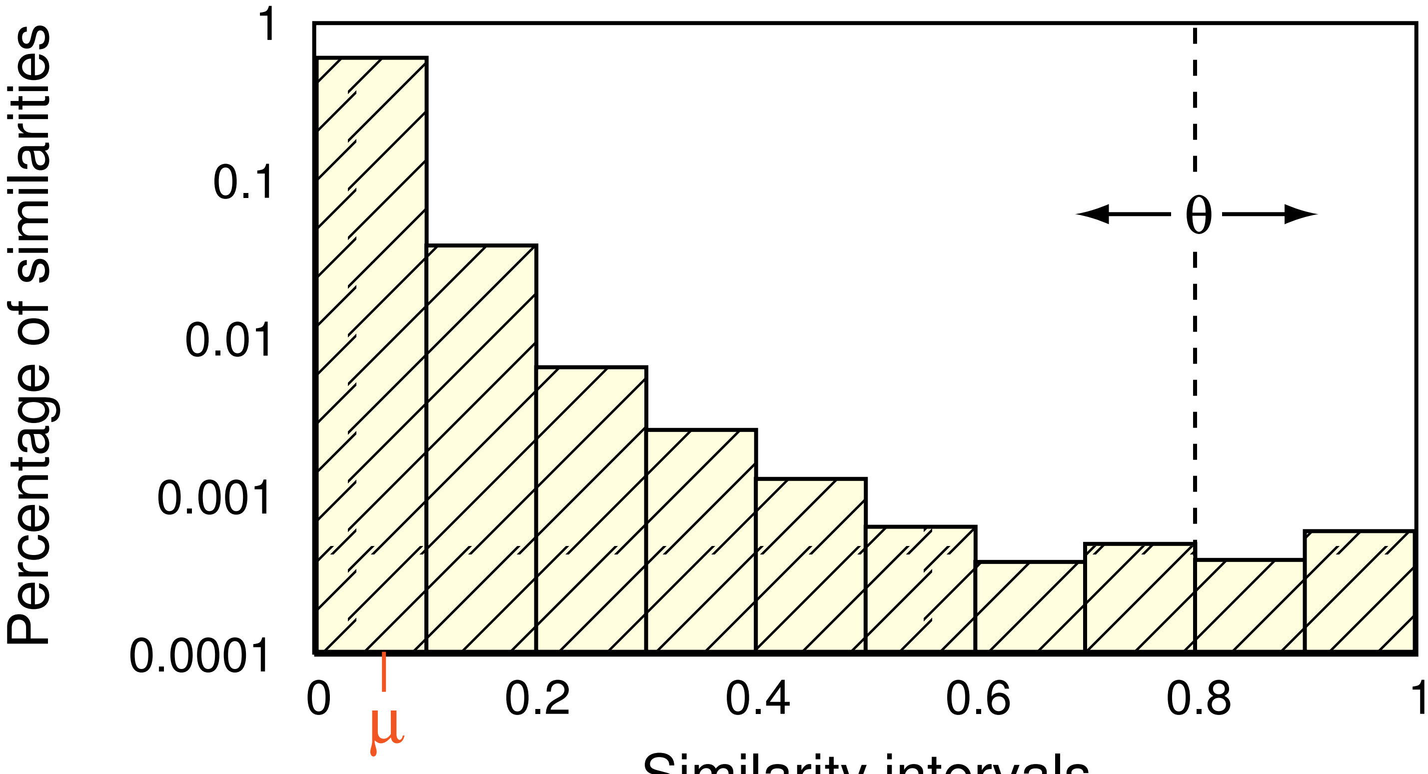
Shingling	Fuzzy-Fingerprinting



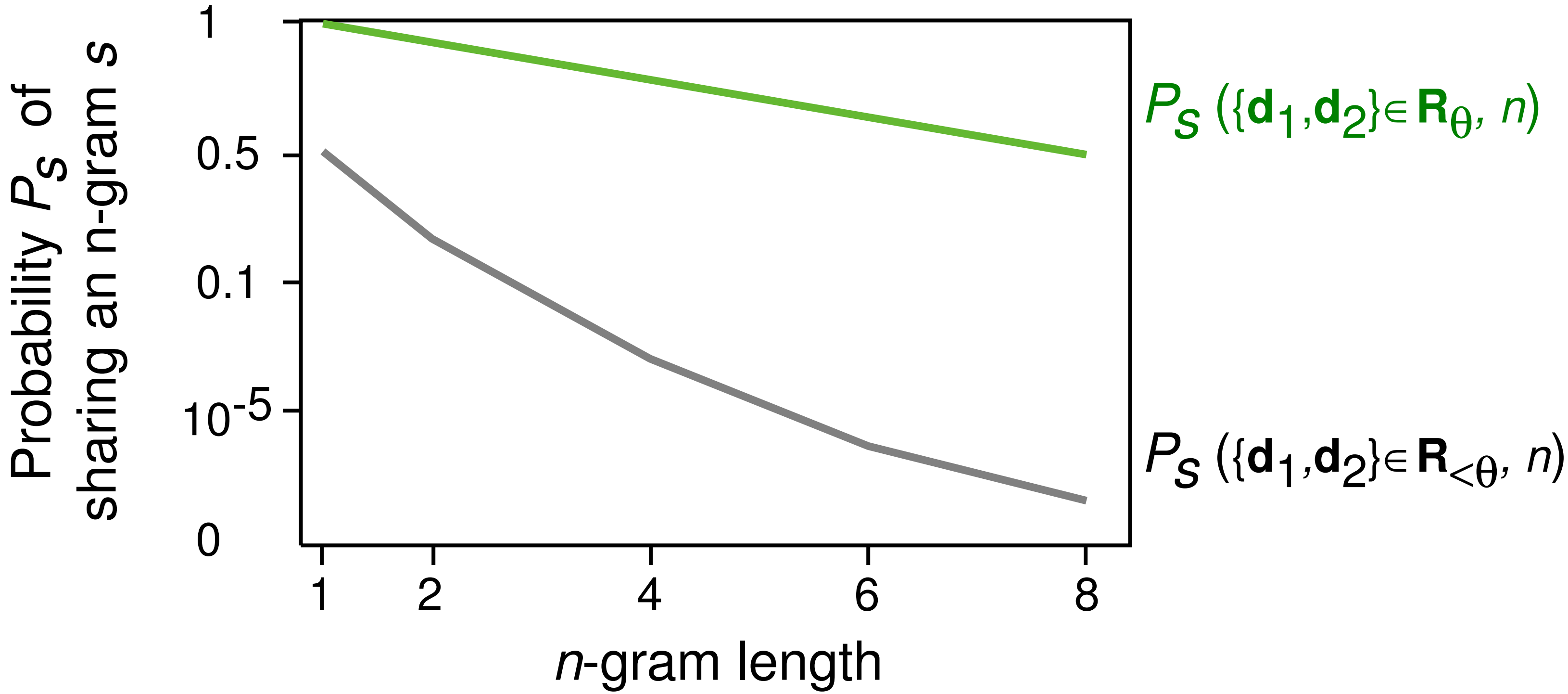
Summary

- Framework for compact retrieval models.
- Dimension reduction allows for an retrieval quality comparable to that of BOW models.
- The memory footprint is orders of magnitude lower than BOW models.
- Embedding outperforms projection in the dimension reduction task.
- Fingerprinting based on hashing allows for $O(1)$ retrieval with imperfect recall.

Thank you for your attention!



Percentage of similarities
1
0.1
0.01
0.001
0.0001
θ
0
μ
0.2
0.4
0.6
0.8
1
Similarity intervals



1
0.5
0.1
10-5
0
1
2
4
6
8
n-gram length
Probability P_S of sharing an n-gram s
P_S ({d1,d2}∈Rθ, n)
P_S ({d1,d2}∈R<θ, n)