Construction of Compact Retrieval Models
Unifying Framework and Analysis

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

Outline
- Introduction and Framework
- Dimension Reduction
- Fingerprinting
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
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

![Diagram showing token sequence, chunking, and high-dimension representation]

\[
\begin{align*}
&\begin{pmatrix}
0.02 \\
0.0 \\
0.01 \\
0.0 \\
0.0 \\
\vdots \\
0.0 \\
0.02 \\
0.0 \\
0.07 \\
0.0
\end{pmatrix} \\
&\begin{pmatrix}
0.1 \\
0.2 \\
0.0 \\
0.1 \\
0.04 \\
\vdots \\
0.1 \\
0.3 \\
0.0 \\
0.0 \\
0.03
\end{pmatrix} \\
&\begin{pmatrix}
0.0 \\
0.1 \\
0.0 \\
0.04 \\
0.0 \\
\vdots \\
0.0 \\
0.0 \\
0.09 \\
0.0 \\
0.0
\end{pmatrix}
\end{align*}
\]

...
Framework

Options for retrieval speed up:

- Dimension reduction
- Fingerprinting

![Diagram showing Dimension reduction and Fingerprinting options]

Stein/Potthast October 18, 2007
Framework

Options for retrieval speed up:

- Dimension reduction
- Fingerprinting

```
\begin{align*}
\text{Token sequence} & \rightarrow \text{Chunking} & \rightarrow \text{High-dim. represen.} \\
& \quad \downarrow \text{Embedding} & \uparrow \text{Dimension reduction} \\
& & \uparrow \text{Small number of features from high-dimensional document space} \\
& & \quad \downarrow \text{Quantization} \\
& & \quad \rightarrow \text{Quantized vector} \\
& & \quad \rightarrow \text{Encoding} \\
& & \quad \rightarrow \text{Fingerprint}
\end{align*}
```

```
\begin{align*}
\{c_1, \ldots, c_k\} & \quad \text{Fingerprinting} \\
& \quad \downarrow \text{Encoding} \\
& \quad \downarrow \text{Quantized vector} \\
& \quad \downarrow \text{Quantization} \\
& \quad \downarrow \text{Small number of features from high-dimensional document space} \\
& \quad \uparrow \text{Dimension reduction} \\
& \quad \uparrow \text{Embedding} \\
& \quad \uparrow \text{High-dim. represen.} \\
& \quad \uparrow \text{Chunking} \\
& \quad \uparrow \text{Token sequence}
\end{align*}

124298 \quad 456723 \quad 546781 \\
\begin{bmatrix} 0.02 \\ 0.0 \\ 0.01 \\ 0.0 \\ 0.0 \\ \vdots \\ 0.0 \\ 0.02 \\ 0.07 \\ 0.0 \end{bmatrix} \quad \begin{bmatrix} 0.0 \\ 0.2 \\ 0.0 \\ 0.1 \\ 0.0 \\ \vdots \\ 0.1 \\ 0.3 \\ 0.0 \\ 0.0 \end{bmatrix} \quad \begin{bmatrix} 0.0 \\ 0.05 \\ 0.0 \\ 0.1 \\ 0.0 \\ \vdots \\ 0.08 \\ 0.06 \\ 0.09 \\ 0.03 \end{bmatrix}
```

```
342509 \quad 129842 \quad 972653 \quad 921345 \quad 546719 \quad 564214 \quad 519461 \\
\begin{bmatrix} 0.07 \\ 0.0 \\ 0.0 \\ 0.1 \\ 0.01 \\ \vdots \\ 0.01 \\ 0.02 \\ 0.03 \\ 0.0 \end{bmatrix} \quad \begin{bmatrix} 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.01 \\ \vdots \\ 0.01 \\ 0.02 \\ 0.03 \\ 0.0 \end{bmatrix} \quad \begin{bmatrix} 0.02 \\ 0.06 \\ 0.0 \\ 0.02 \\ 0.03 \\ \vdots \\ 0.03 \\ 0.06 \\ 0.0 \\ 0.0 \end{bmatrix}
```

Stein/Potthast
Part 1 of the Framework

Dimension Reduction
Compact Retrieval Models

Dimension Reduction

<table>
<thead>
<tr>
<th>Dimension reduction</th>
<th>Alternative 1</th>
<th>Alternative 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rationale</td>
<td>Hypothesis Test</td>
<td>Model Fidelity</td>
</tr>
<tr>
<td>Implementation</td>
<td>Shingling</td>
<td>Fuzzy-Fingerprinting</td>
</tr>
</tbody>
</table>

English Wikipedia:

<table>
<thead>
<tr>
<th>Dictionary</th>
<th>Number of dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-gram space</td>
<td>3 921 588</td>
</tr>
<tr>
<td>4-gram space</td>
<td>274 101 016</td>
</tr>
<tr>
<td>8-gram space</td>
<td>373 795 734</td>
</tr>
<tr>
<td>Shingling space</td>
<td>75 659 644</td>
</tr>
</tbody>
</table>
Compact Retrieval Models

Alternative 1: Projecting / Hypothesis Test

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

\[
H_0 : \{d_1, d_2\} \text{ is from } \mathbf{R}_{<\theta} \\
H_1 : \{d_1, d_2\} \text{ is from } \mathbf{R}_{\theta}
\]

\[
\frac{|\mathbf{R}_{<\theta}| \cdot P_s(\{d_1, d_2\} \in \mathbf{R}_{<\theta}, n=8)}{|\mathbf{R}_{\theta}| \cdot P_s(\{d_1, d_2\} \in \mathbf{R}_{\theta}, n=8)} \sim \frac{P_0}{P_1}
\]
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) ⇒ 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.

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

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

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

SVD of $X$ yields the optimum embedding $Y_{SVD}$:

$$\hat{S}^* = V_k \Sigma_k^2 V_k^T =: Y_{SVD}^T Y_{SVD}$$
Compact Retrieval Models

Alternative 1 vs. Alternative 2

Shingling

Fuzzy-Fingerprinting

\[ d \]

\[
\begin{bmatrix}
  w_1 & w_2 & w_3 \\
  w_2 & w_3 & w_4 \\
  w_3 & w_4 & w_5 \\
  \vdots & \vdots & \vdots \\
  w_{m-2} & w_{m-1} & w_m
\end{bmatrix}
\]

\[ d \]

A priori probabilities from BNC

Distribution of prefix classes in sample

Normalization and difference computation

\bullet Random functions.

\bullet Documents from the British National Corpus
Part 2 of the Framework

Fingerprinting
Framework

Options for retrieval speed up:

- Dimension reduction
- Fingerprinting

![Diagram of retrieval process]

Token sequence → Chunking → High-dim. represen. → Dimension reduction (chunk selection) → Embedding

Small number of features from high-dimensional document space → Reformulated representation in low-dimensional document space

Quantization → Quantized vector → Encoding → Fingerprint

\{ c_1, \ldots, c_k \}

Stein/Potthast

Oct 18, 2007
Fingerprinting

Fuzzy-Fingerprinting

Documents from the British National Corpus

Embedding $\rightarrow$ Quantization $\rightarrow$ Encoding

A priori probabilities from BNC

Distribution of prefix classes in sample

Normalization and difference computation

Stein/Potthast October 18, 2007
Fingerprinting

Fuzzy-Fingerprinting

Embedding $\rightarrow$ Quantization $\rightarrow$ Encoding

- A priori probabilities from BNC
- Distribution of prefix classes in sample
- Normalization and difference computation

- Documents from the British National Corpus
Fingerprinting

Fuzzy-Fingerprinting

Embedding $\rightarrow$ Quantization $\rightarrow$ Encoding

A priori probabilities from BNC

Distribution of prefix classes in sample

Normalization and difference computation

Fuzzification

Documents from the British National Corpus

Fingerprint = 2643256
Fingerprinting

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**

<table>
<thead>
<tr>
<th>Shingling</th>
<th>Fuzzy-Fingerprinting</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Graph" /></td>
<td><img src="image2.png" alt="Graph" /></td>
</tr>
<tr>
<td><img src="image3.png" alt="Graph" /></td>
<td><img src="image4.png" alt="Graph" /></td>
</tr>
<tr>
<td><img src="image5.png" alt="Graph" /></td>
<td><img src="image6.png" alt="Graph" /></td>
</tr>
<tr>
<td><img src="image7.png" alt="Graph" /></td>
<td><img src="image8.png" alt="Graph" /></td>
</tr>
</tbody>
</table>

**Stein/Potthast October 18, 2007**
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!
Probability $P_s$ of sharing an $n$-gram $s$.

$P_s (\{d_1, d_2\} \in R_{\theta, n})$

$P_s (\{d_1, d_2\} \in R_{<\theta, n})$