Strategies for Retrieving Plagiarized Documents

Benno Stein
Sven Meyer zu Eissen
Martin Potthast

picapica.net

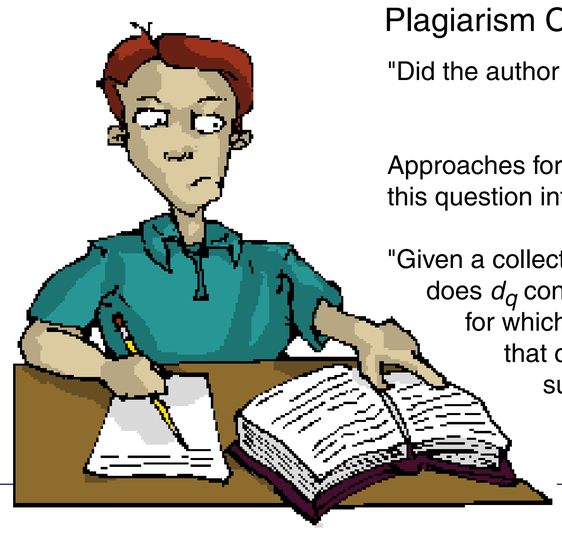
Bauhaus University Weimar

Contribution:

A tailored index for plagiarism analysis

For the identification of plagiarized passages in large document collections we present retrieval strategies which rely on stochastic sampling and chunk indexes.

Using the entire Wikipedia corpus we compile n-gram indexes and compare them to a new kind of fingerprint index in a plagiarism analysis use case. Our index provides an analysis speed-up by factor 1.5 and is an order of magnitude smaller, while being equivalent in terms of precision and recall.



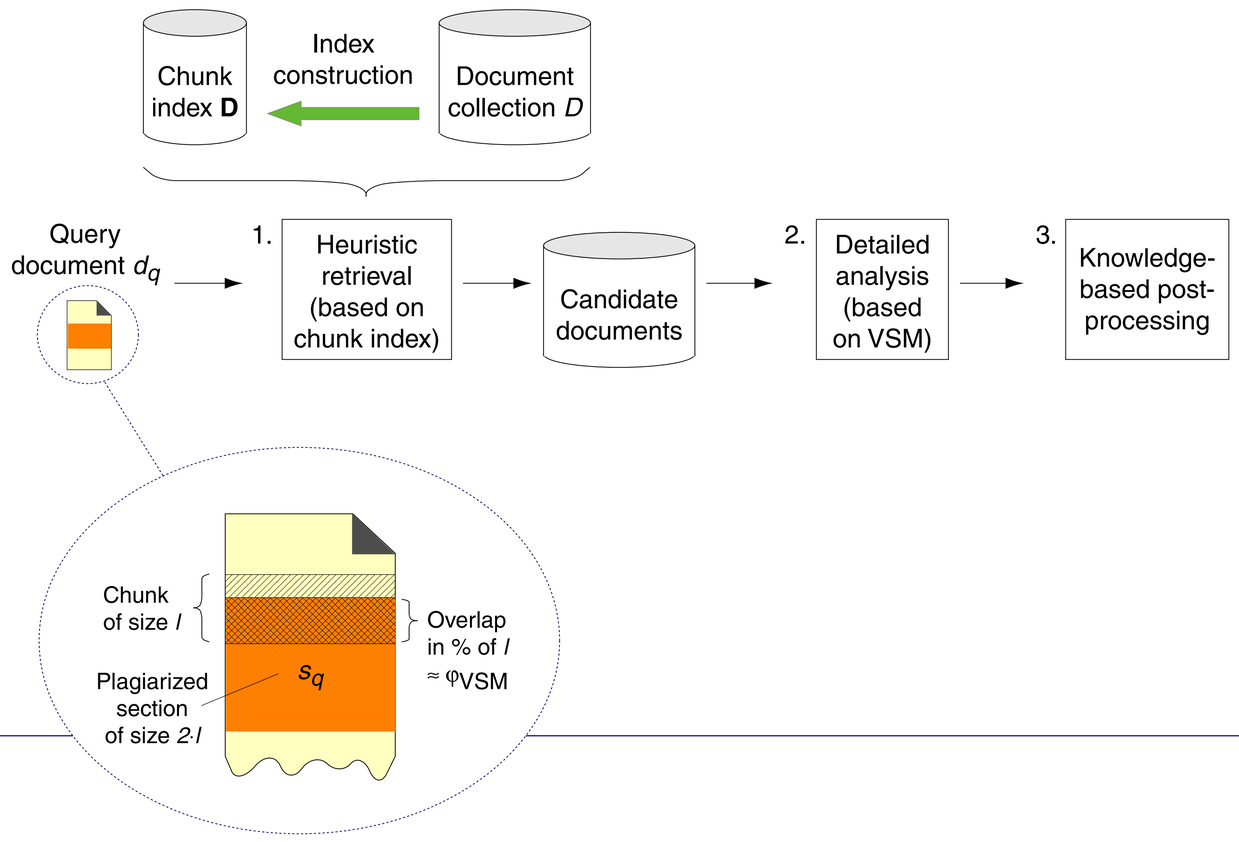
Plagiarism Challenge:

"Did the author of a document d_q commit a plagiarism offense?"

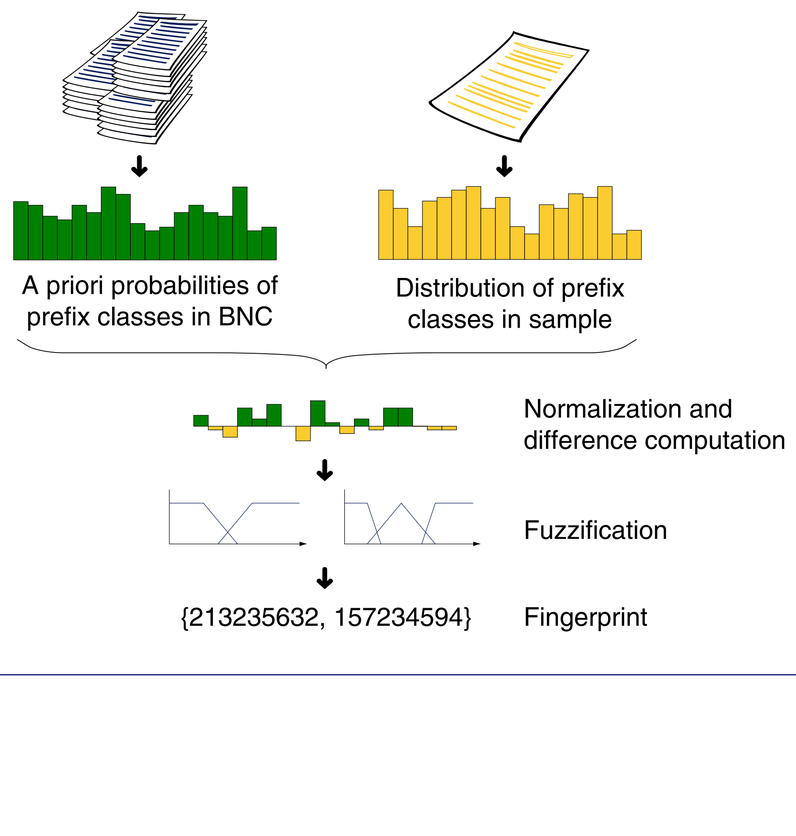
Approaches for computer-based plagiarism analysis break down this question into manageable parts:

"Given a collection D of documents,
does d_q contain a section s_q
for which one can find a document $d_q \in D$
that contains a section s_x
such that under some retrieval model $\mathcal{R}$
the similarity $\varphi_{\mathcal{R}}$ between s_q and s_x is close to 1?"

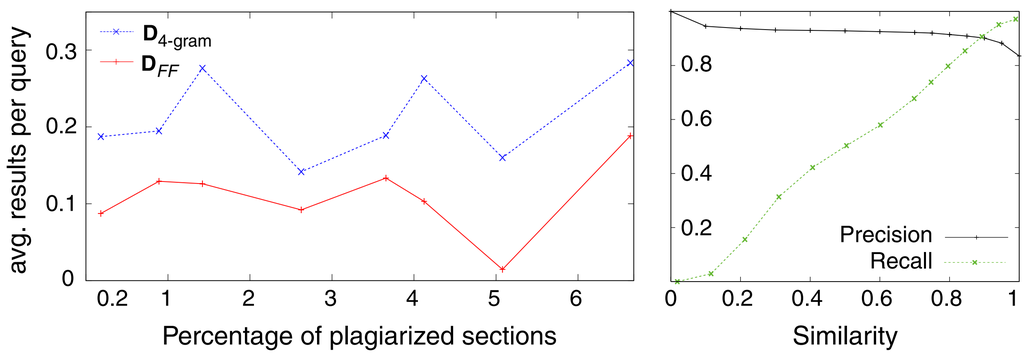
Process Overview



Index Construction:



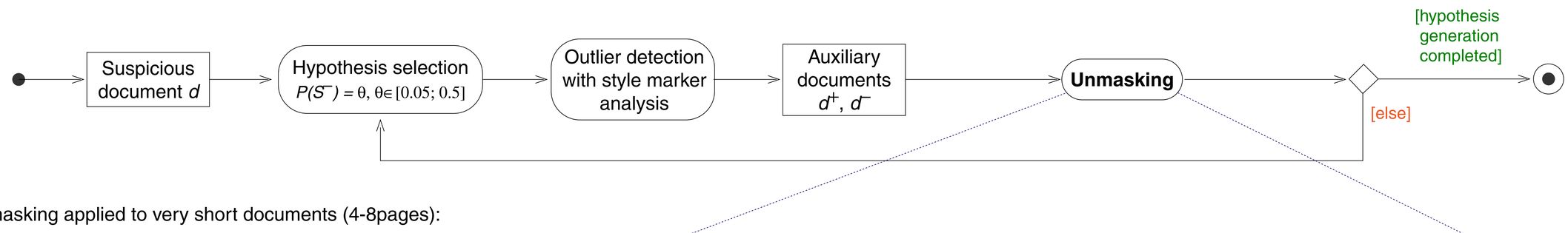
Results



Characteristics of the Wikipedia chunk index:

Chunk index type	avg. postlist length	Size	Size ratio
D_{FF}, *l=40, γ=0.8*	1.17 entries	0.35 GB	0.07
$D_{3\text{-gram}}$	2.43 entries	3.42 GB	0.65
$D_{4\text{-gram}}$	1.44 entries	5.25 GB	1.0

Related Technology [Koppel/Schler 2004]



Unmasking applied to very short documents (4-8pages):