The Impact of Main Content Extraction on Near-Duplicate Detection

Maik Fröbe, Matthias Hagen, Janek Bevendorff, Michael Völske, Benno Stein, Christopher Schröder, Robby Wagner, Lukas Gienapp, Martin Potthast

OSSYM, 11–13 October 2021
webis.de

Impact of Main Content Extraction on Near-Duplicate Detection

Motivation

- Web crawls contain many near-duplicates:
  [Fetterly et al.; LA-WEB'03]

Dog breed

From Wikipedia, the free encyclopedia

A **dog breed** is a particular strain that was purposefully bred by humans to perform specific tasks, such as herding, hunting, and guarding. When distinguishing breed from type, the rule of thumb is that a breed always "breeds true".[1] Dogs are the most variable mammal on earth,

Montage showing the morphological variation of the dog.

Impact of Main Content Extraction on Near-Duplicate Detection

Motivation

- Web crawls contain many near-duplicates:
  [Fetterly et al.; LA-WEB'03]

Impact of Main Content Extraction on Near-Duplicate Detection

Motivation

- Web crawls contain many near-duplicates:
  [Fetterly et al.; LA-WEB'03]

Impact of Main Content Extraction on Near-Duplicate Detection

Overview

- Near-Duplicates in Web Crawls: Risks and Potentials
  - Risk: Evaluation of Search Engines
    [Bernstein et al.; CIKM'05; Fröbe et al.; ECIR'20]
  - Risk: Training of Learning to Rank Models
    [Fröbe et al.; SIGIR'20]
  - Potential: Transfer of Relevance Labels
    [Fröbe et al.; SIGIR'21]

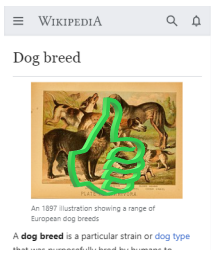
- Improve Near-Duplicate Detection with Main Content Extraction?
  - Precision vs. Recall

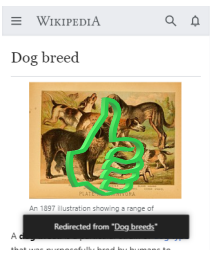
Impact of Main Content Extraction on Near-Duplicate Detection

Risk: Evaluation of Search Engines

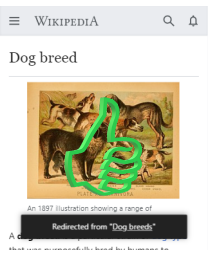
[Bernstein et al.; CIKM'05; Fröbe et al.; ECIR'20]

- Cranfield paradigm: Query-document pairs labeled in isolation
- Web Track Topic 194: `designer dog breeds`
  - 40 of 47 relevant documents duplicate the same Wikipedia article





...



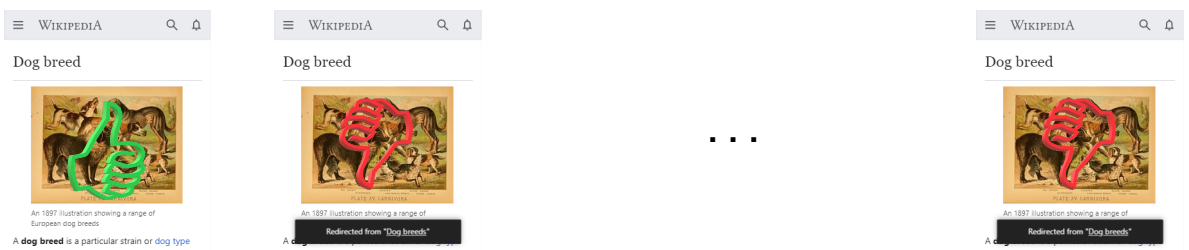
Impact of Main Content Extraction on Near-Duplicate Detection

Risk: Evaluation of Search Engines

[Bernstein et al.; CIKM'05; Fröbe et al.; ECIR'20]

- Cranfield paradigm: Query-document pairs labeled in isolation
- Web Track Topic 194: `designer dog breeds`
  - 40 of 47 relevant documents duplicate the same Wikipedia article

- Novelty Principle:
  A document is irrelevant if it is content-equivalent
  to a document the user has already seen in the results.



Impact of Main Content Extraction on Near-Duplicate Detection

Risk: Evaluation of Search Engines

[Bernstein et al.; CIKM'05; Fröbe et al.; ECIR'20]

- Comparison:
  - Evaluation without the novelty principle
  - Evaluation with the novelty principle (rankings deduplicated)

Web Track	# Runs	Duplicates removed	
		Δ_{nDCG}	τ
2009	71	-6.8%	0.91
2010	56	-9.9%	0.57
2011	37	-3.4%	0.92
2012	28	-12.4%	0.81
2013	34	-1.8%	0.90
2014	30	-4.5%	0.94

- Problems caused by near-duplicates:
  - Effectiveness is overestimated
  - "Leaderboard" changes

  $\Rightarrow$ Risk: Training of Learning to Rank Models [Fröbe et al.; SIGIR'20]

Impact of Main Content Extraction on Near-Duplicate Detection

Risk: Evaluation of Search Engines

[Bernstein et al.; CIKM'05; Fröbe et al.; ECIR'20]

- Comparison:
  - Evaluation without the novelty principle
  - Evaluation with the novelty principle (rankings deduplicated)

Web Track	# Runs	Duplicates removed	
		Δ_{nDCG}	τ
2009	71	-6.8%	0.91
2010	56	-9.9%	0.57
2011	37	-3.4%	0.92
2012	28	-12.4%	0.81
2013	34	-1.8%	0.90
2014	30	-4.5%	0.94

- Problems caused by near-duplicates:
  - Effectiveness is overestimated
  - "Leaderboard" changes

    $\Rightarrow$ Risk: Training of Learning to Rank Models [Fröbe et al.; SIGIR'20]

Impact of Main Content Extraction on Near-Duplicate Detection

Risk: Evaluation of Search Engines

[Bernstein et al.; CIKM'05; Fröbe et al.; ECIR'20]

- Comparison:
  - Evaluation without the novelty principle
  - Evaluation with the novelty principle (rankings deduplicated)

Web Track	# Runs	Duplicates removed	
		Δ_{nDCG}	τ
2009	71	-6.8%	0.91
2010	56	-9.9%	0.57
2011	37	-3.4%	0.92
2012	28	-12.4%	0.81
2013	34	-1.8%	0.90
2014	30	-4.5%	0.94

- Problems caused by near-duplicates:
  - Effectiveness is overestimated
  - "Leaderboard" changes

  $\Rightarrow$ Risk: Training of Learning to Rank Models [Fröbe et al.; SIGIR'20]

Impact of Main Content Extraction on Near-Duplicate Detection

Potential: Transfer of Relevance Labels

[Fröbe et al.; SIGIR'21]

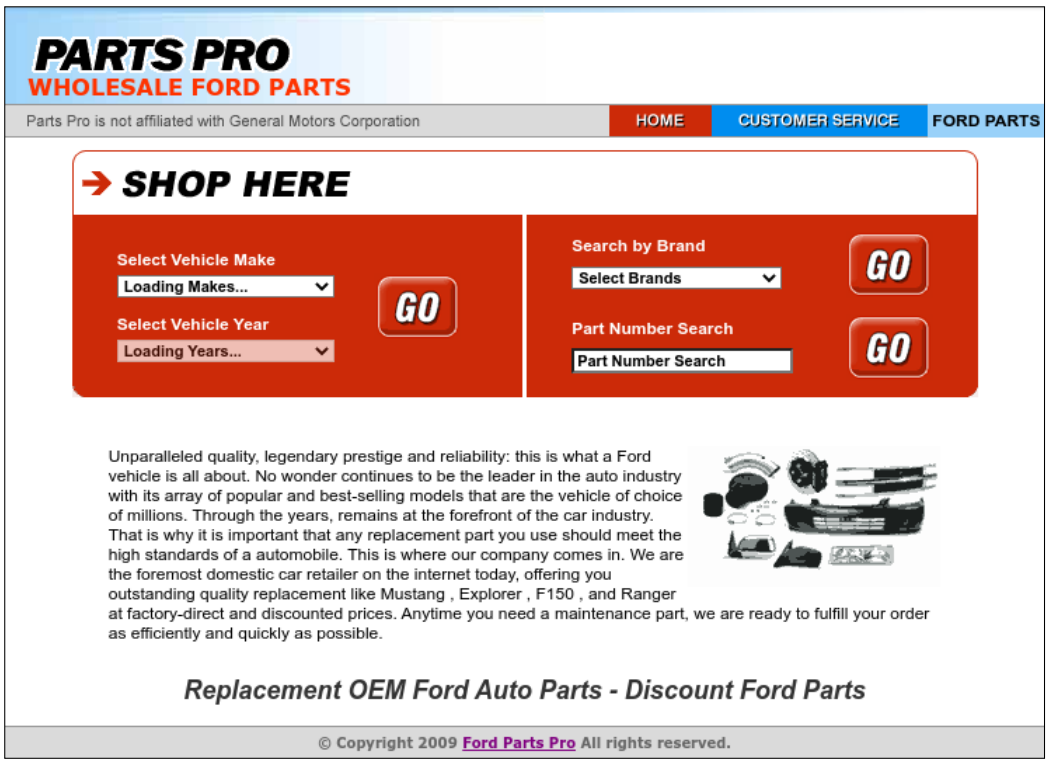
- ClueWeb09: 73,883 relevance judgments (estimated effort: 4-8 months)
- Idea: Transfer relevance judgments to newer crawls ("save" judgment effort)

Impact of Main Content Extraction on Near-Duplicate Detection

Potential: Transfer of Relevance Labels

[Fröbe et al.; SIGIR'21]

- ClueWeb09: 73,883 relevance judgments (estimated effort: 4-8 months)
- Idea: Transfer relevance judgments to newer crawls ("save" judgment effort)



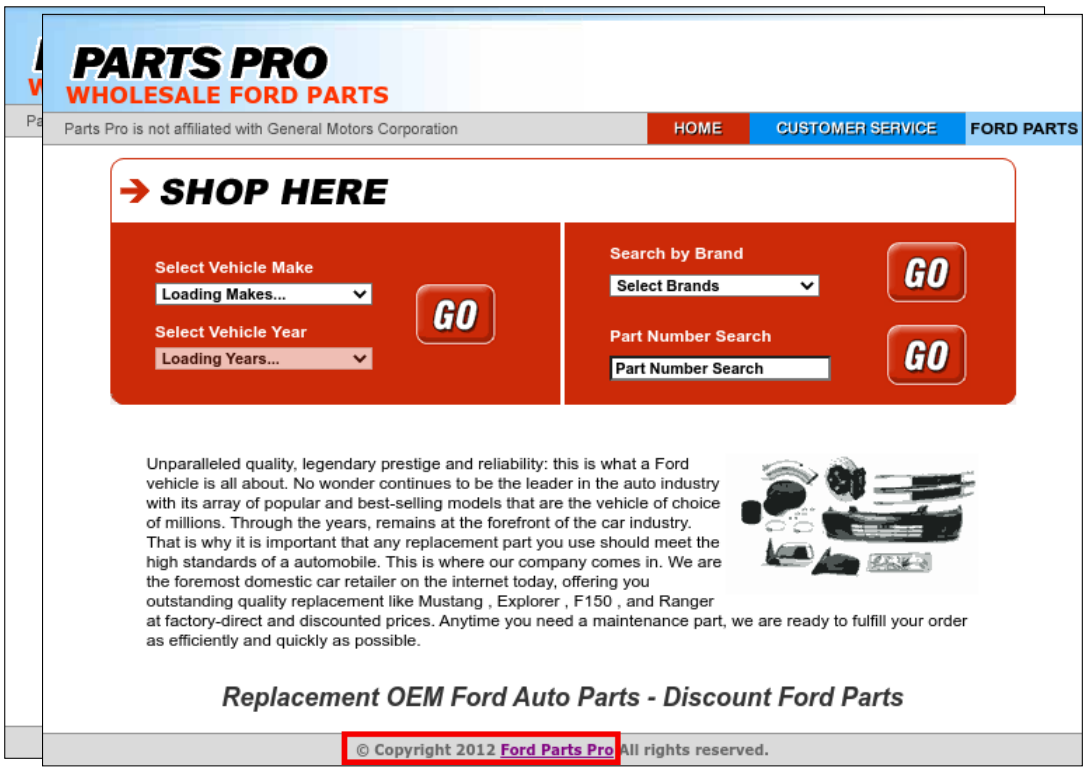
  - Relevant for query "used car parts" in ClueWeb09

Impact of Main Content Extraction on Near-Duplicate Detection

Potential: Transfer of Relevance Labels

[Fröbe et al.; SIGIR'21]

- ClueWeb09: 73,883 relevance judgments (estimated effort: 4-8 months)
- Idea: Transfer relevance judgments to newer crawls ("save" judgment effort)



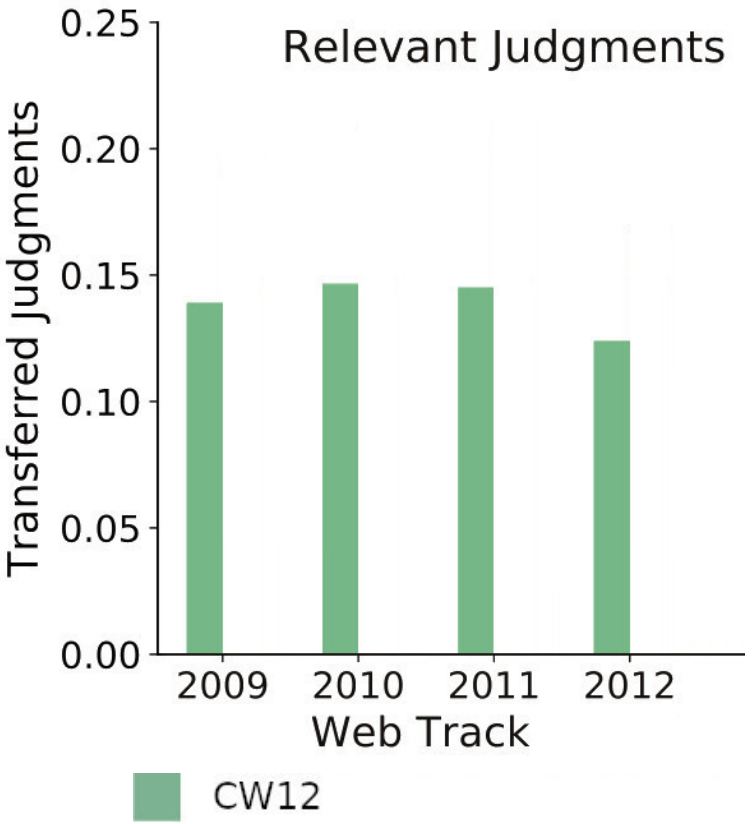
- Relevant for query "used car parts" in ClueWeb09
- Near-Duplicate in ClueWeb12 is also relevant

Impact of Main Content Extraction on Near-Duplicate Detection

Potential: Transfer of Relevance Labels

[Fröbe et al.; SIGIR'21]

- Experiment: Transfer relevance judgments from ClueWeb09 to ClueWeb12

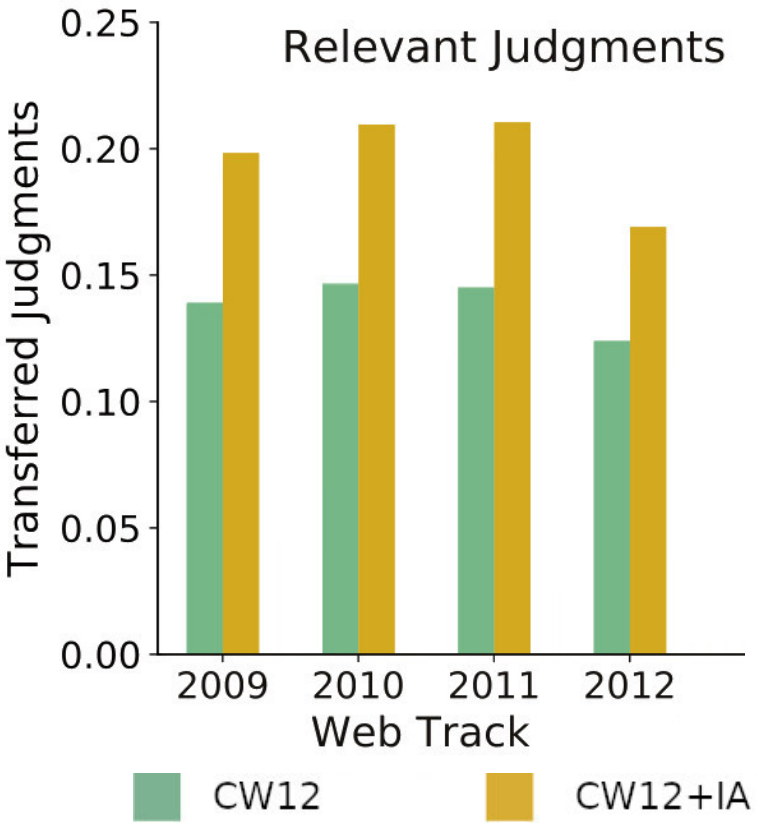


Impact of Main Content Extraction on Near-Duplicate Detection

Potential: Transfer of Relevance Labels

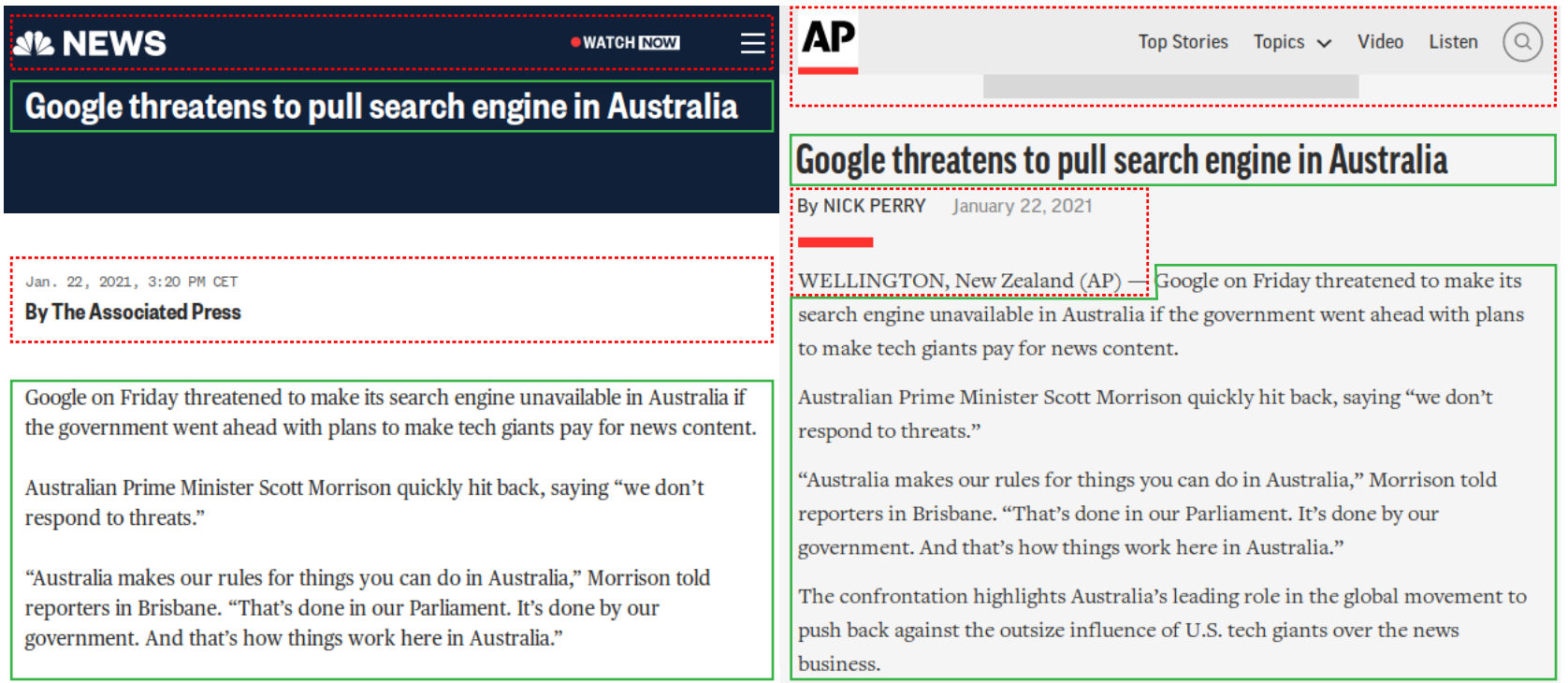
[Fröbe et al.; SIGIR'21]

- Experiment: Transfer relevance judgments from ClueWeb09 to ClueWeb12



Impact of Main Content Extraction on Near-Duplicate Detection

Motivation: Main Content Extraction for Near-Duplicate Detection



- Perfect main content extraction would improve precision and recall
  - Recall: Identical content in different boilerplate

Impact of Main Content Extraction on Near-Duplicate Detection

Experimental Setup

- Select 186,819 document pairs from the ClueWebs
  - Group documents by (canonical) URL
  - Select 5,000 groups at random
  - Consider all possible pairs within a group

- We label 500 document pairs as near-duplicate (identical content) or not
  - Uniform sampled over S_3 similarity

	$S_3 \geq 0.9$	
	Precision	Recall
Full Content	0.98	0.26
JusText	**1.00**	0.24
Jericho	0.97	0.26
Trafilatura	0.77	0.41
Boilerpipe	0.61	**0.59**

Impact of Main Content Extraction on Near-Duplicate Detection

Takeaways

- Not accounting for near-duplicates introduces risks
  - Evaluation of retrieval models
  - Learning to rank
- Transfer of relevance judgments as use-case of near-duplicate-detection
- CopyCat
  - Resource to simplify deduplication in TREC-style experimentals
  - Open source: github.com/chatnoir-eu/chatnoir-copycat
- Future work:
  - Expand experiments
  - Study transfer of relevance labels on MS Marco

Impact of Main Content Extraction on Near-Duplicate Detection

Takeaways

- Not accounting for near-duplicates introduces risks
  - Evaluation of retrieval models
  - Learning to rank
- Transfer of relevance judgments as use-case of near-duplicate-detection
- CopyCat
  - Resource to simplify deduplication in TREC-style experimentals
  - Open source: github.com/chatnoir-eu/chatnoir-copycat
- Future work:
  - Expand experiments
  - Study transfer of relevance labels on MS Marco

Thank You!