Maik Fröbe Janek Bevendorff Lukas Gienapp Michael Völske Benno Stein Martin Potthast Matthias Hagen
maik.froebe@informatik.uni-halle.de

CopyCat: Near-Duplicates Within and Between the ClueWeb and the Common Crawl

The Resource in a Nutshell

Web crawls contain many near-duplicates (i.e., documents with different URLs but very similar content). ChatNoir-CopyCat-21 is a precision-oriented near-duplicate resource consisting of (1) lists of near-duplicates within the ClueWeb09, the ClueWeb12, and two Common Crawl snapshots, as well as between selections of these crawls, and (2) a software library to deduplicate arbitrary document sets (e.g., TREC runs).

github.com/chatnoir-eu/copycat

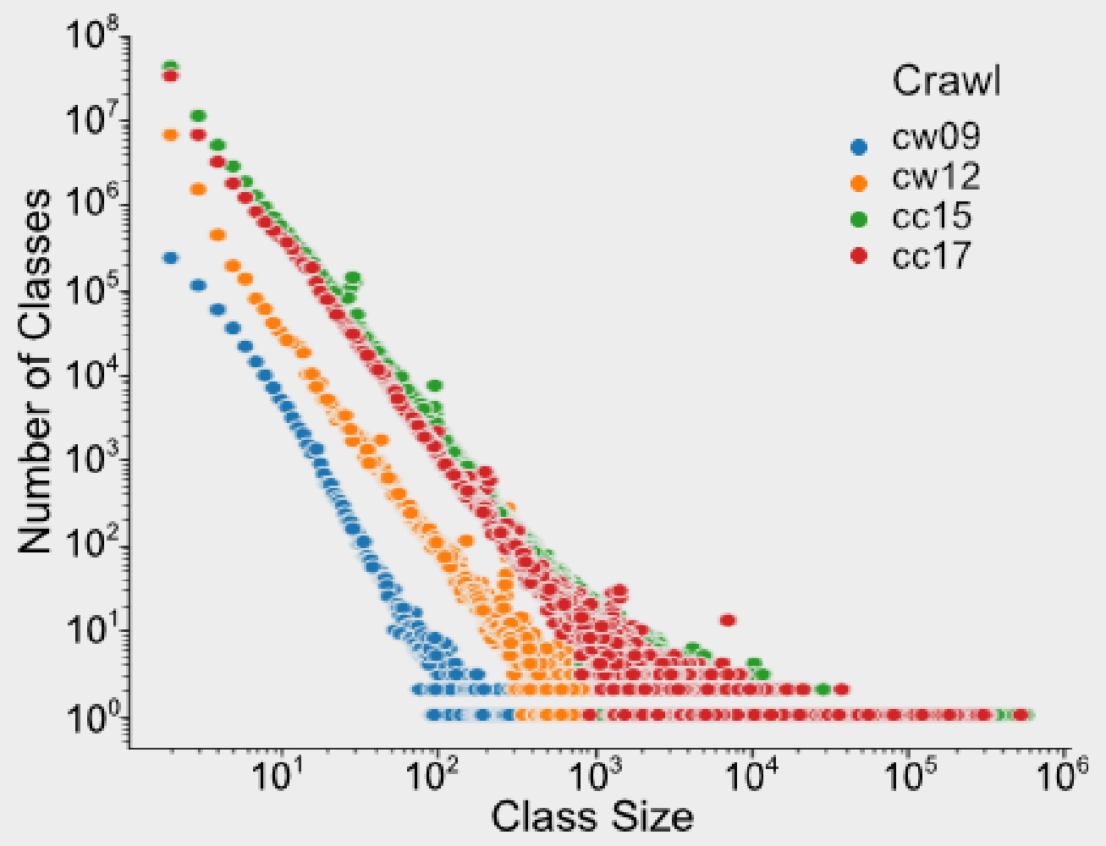
Compilation of Near-Duplicates

We compile the near-duplicates found in the ClueWeb09 (cw09), the ClueWeb12 (cw12), and the Common Crawls 2015-11 (cw15) and 2017-04 (cw17) into inclusion and exclusion lists for easy deduplication.

Statistic	Web crawl				Total
	cw09	cw12	cc15	cc17	
Compressed size	4.0 TB	4.5 TB	28.1 TB	54.0 TB	90.6 TB
Documents	1.0 b	731.7 m	1.8 b	3.1 b	6.7 b
SimHash duplicates	145.6 m	201.2 m	907.2 m	1.0 b	2.3 b
Canonical links	3.0 m	67.1 m	747.5 m	1.5 b	2.3 b
Crawled duplicates	1.5 m	11.2 m	278.3 m	180.1 m	471.1 m
CopyCat duplicates	145.8 m	204.3 m	951.2 m	1.0 b	2.3 b

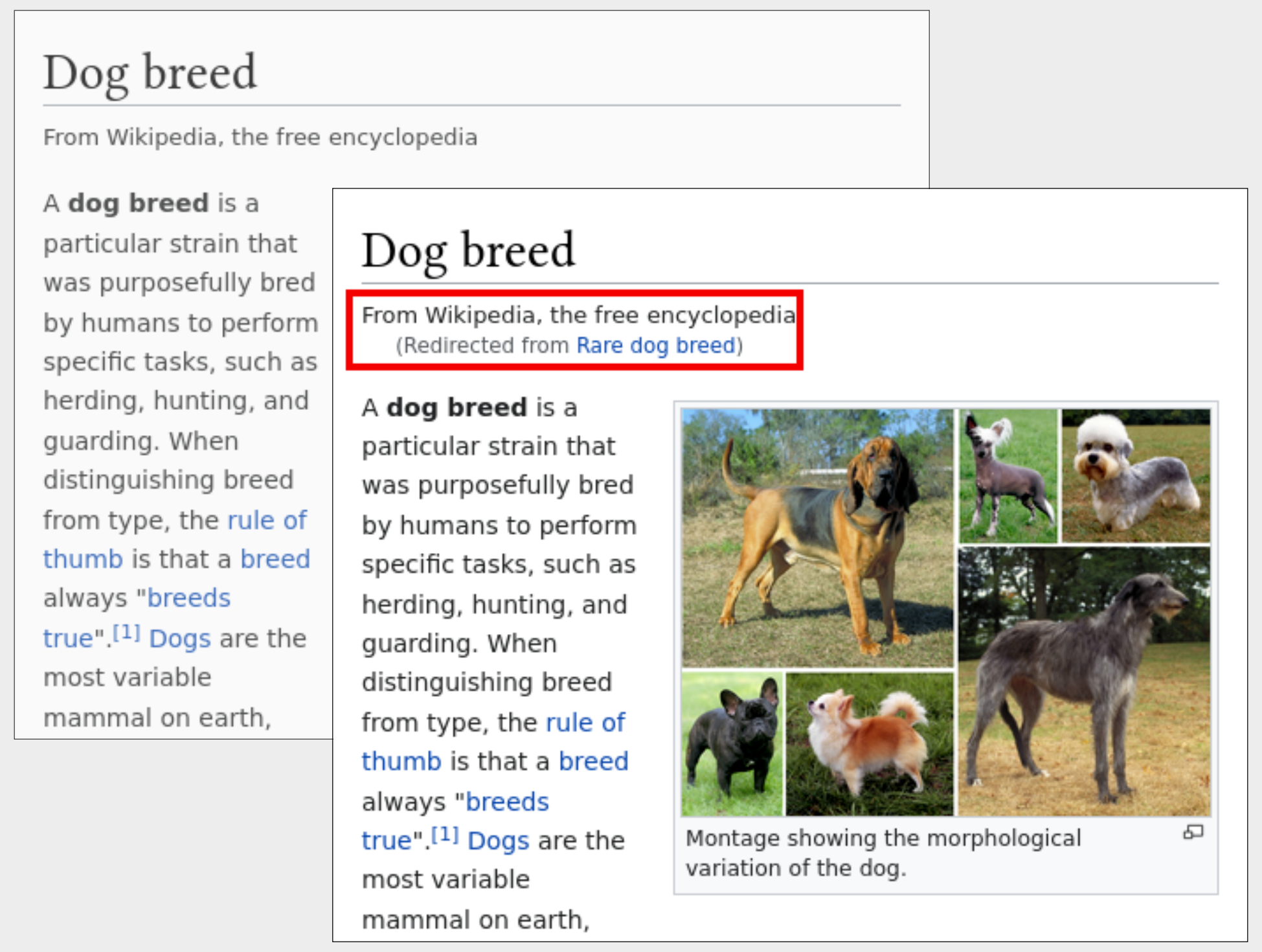
Parameter-Tuning on Canonical Links

We conduct a pilot study to semi-automatically label 360 million document pairs sampled from equivalence classes of canonical links.



- ○ Exact all-pairs similarity
- ○ Manual review of document pairs
- ○ Precision-oriented threshold
- ○ CopyCat runs SimHash in:
  - ○ Entire crawls (Pr.: 0.97)
  - ○ Classes of canonical links (Pr.: 0.94)

Deduplication in IR Experiments



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,

Dog breed

From Wikipedia, the free encyclopedia
(Redirected from Rare dog breed)

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.

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

Impact of the Novelty Principle:

- ○ System effectiveness is overestimated
- ○ Ranking of systems changed

Example:
The 47 relevant ClueWeb09 documents for the query `designer dog breeds` contain 40 near-duplicates of the Wikipedia article on the left.

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

Web track			Near-dupl. in judgment			Near-dupl. in runs		
Year	Runs	Judg.	All	Relevant	Irrelevant	@10	@100	@1000
2009	71	13 118	0.14	0.18	0.11	0.11	0.17	0.19
2010	56	25 329	0.17	0.21	0.15	0.19	0.25	0.25
2011	37	19 381	0.19	0.21	0.19	0.21	0.24	0.25
2012	28	16 055	0.16	0.25	0.13	0.20	0.18	0.20
2013	34	14 474	0.17	0.13	0.17	0.12	0.19	0.26
2014	30	14 432	0.18	0.19	0.15	0.13	0.21	0.29

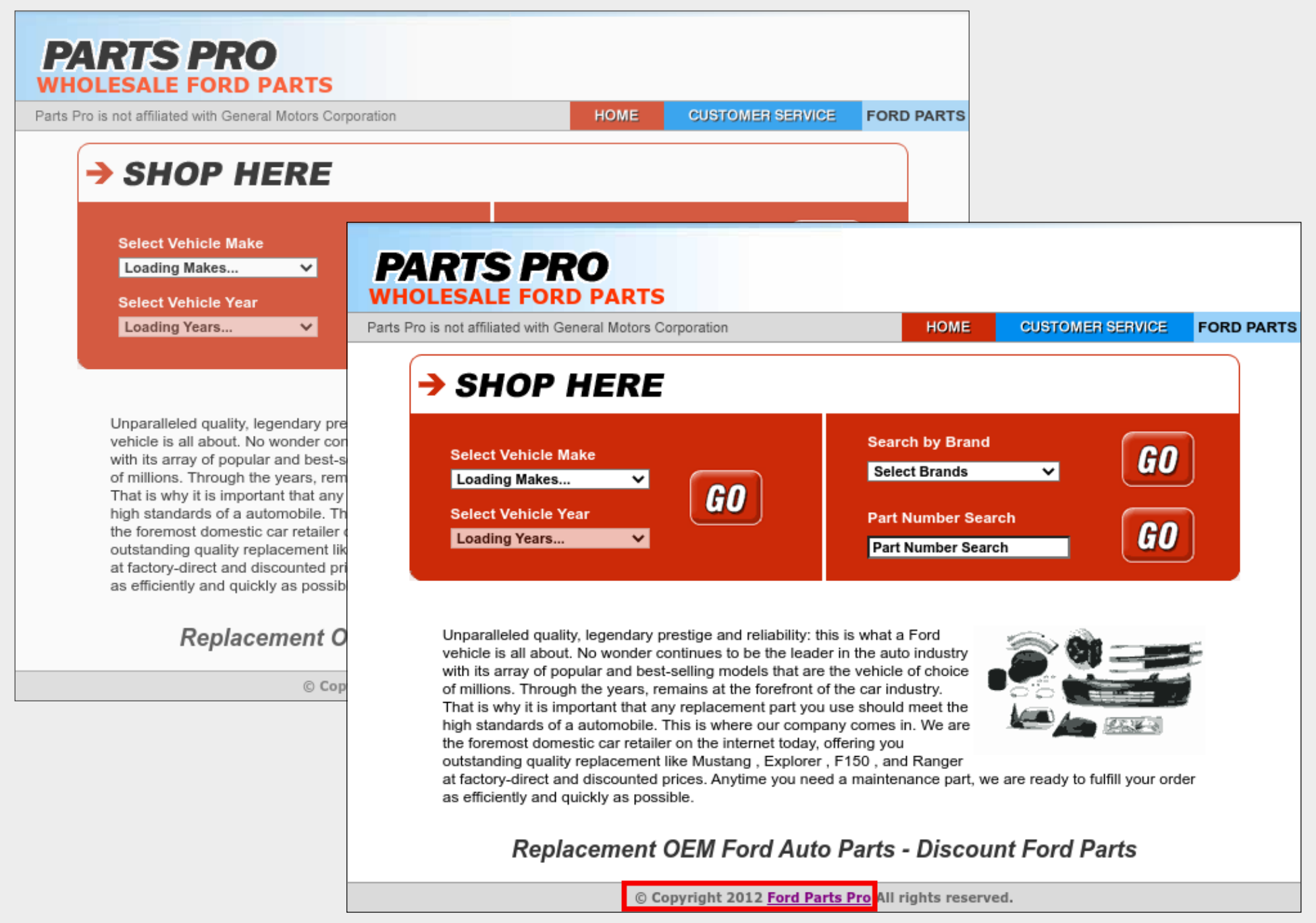
Overview of the proportion of near-duplicates (at an similarity threshold determined in the pilot study) in the relevance judgments and in the top-k results of submitted runs for the TREC Web tracks.

Method	All		Relevant		Irrelevant	
	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
SimHash	0.95	0.33	1.00	0.49	0.93	0.29
Can. Link	0.90	0.08	0.99	0.17	0.79	0.11
CopyCat	0.93	0.36	0.99	0.54	0.87	0.34

Precision and recall of inclusion and exclusion lists for runs submitted to the TREC Web tracks at depth 1,000 (for an similarity threshold determined in a pilot study).

github.com/chatnoir-eu

Transfer of Relevance Labels



PARTS PRO
WHOLESALE FORD PARTS
Parts Pro is not affiliated with General Motors Corporation
HOME CUSTOMER SERVICE FORD PARTS

SHOP HERE

Select Vehicle Make
Loading Makes...
Select Vehicle Year
Loading Years...

Unparalleled quality, legendary pre vehicle is all about. No wonder con with its array of popular and best-s of millions. Through the years, rem That is why it is important that any high standards of a automobile. Th the foremost domestic car retailer outstanding quality replacement lik at factory-direct and discounted pri as efficiently and quickly as possib

Replacement O

© Cop

PARTS PRO
WHOLESALE FORD PARTS
Parts Pro is not affiliated with General Motors Corporation
HOME CUSTOMER SERVICE FORD PARTS

SHOP HERE

Select Vehicle Make
Loading Makes...
Select Vehicle Year
Loading Years...
GO

Search by Brand
Select Brands
GO
Part Number Search
Part Number Search
GO

Unparalleled quality, legendary prestige and reliability: this is what a Ford vehicle is all about. No wonder continues to be the leader in the auto industry with its array of popular and best-selling models that are the vehicle of choice of millions. Through the years, remains at the forefront of the car industry. That is why it is important that any replacement part you use should meet the high standards of a automobile. This is where our company comes in. We are the foremost domestic car retailer on the internet today, offering you outstanding quality replacement like Mustang , Explorer , F150 , and Ranger at factory-direct and discounted prices. Anytime you need a maintenance part, we are ready to fulfill your order as efficiently and quickly as possible.

Replacement OEM Ford Auto Parts - Discount Ford Parts

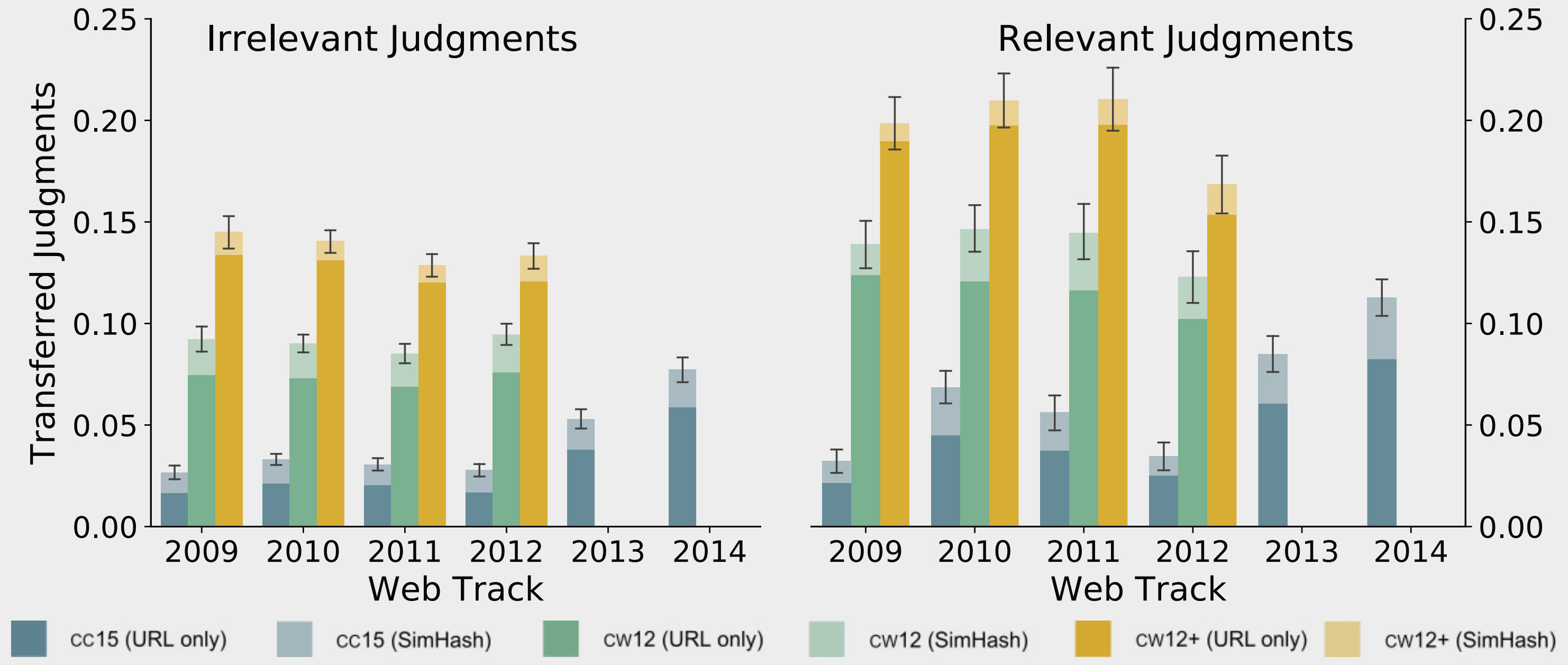
© Copyright 2012 Ford Parts Pro All rights reserved.

Motivation (cw09 Web Track):

- ○ 73,883 relevance judgments
- ○ Estimated effort: 4 to 8 months

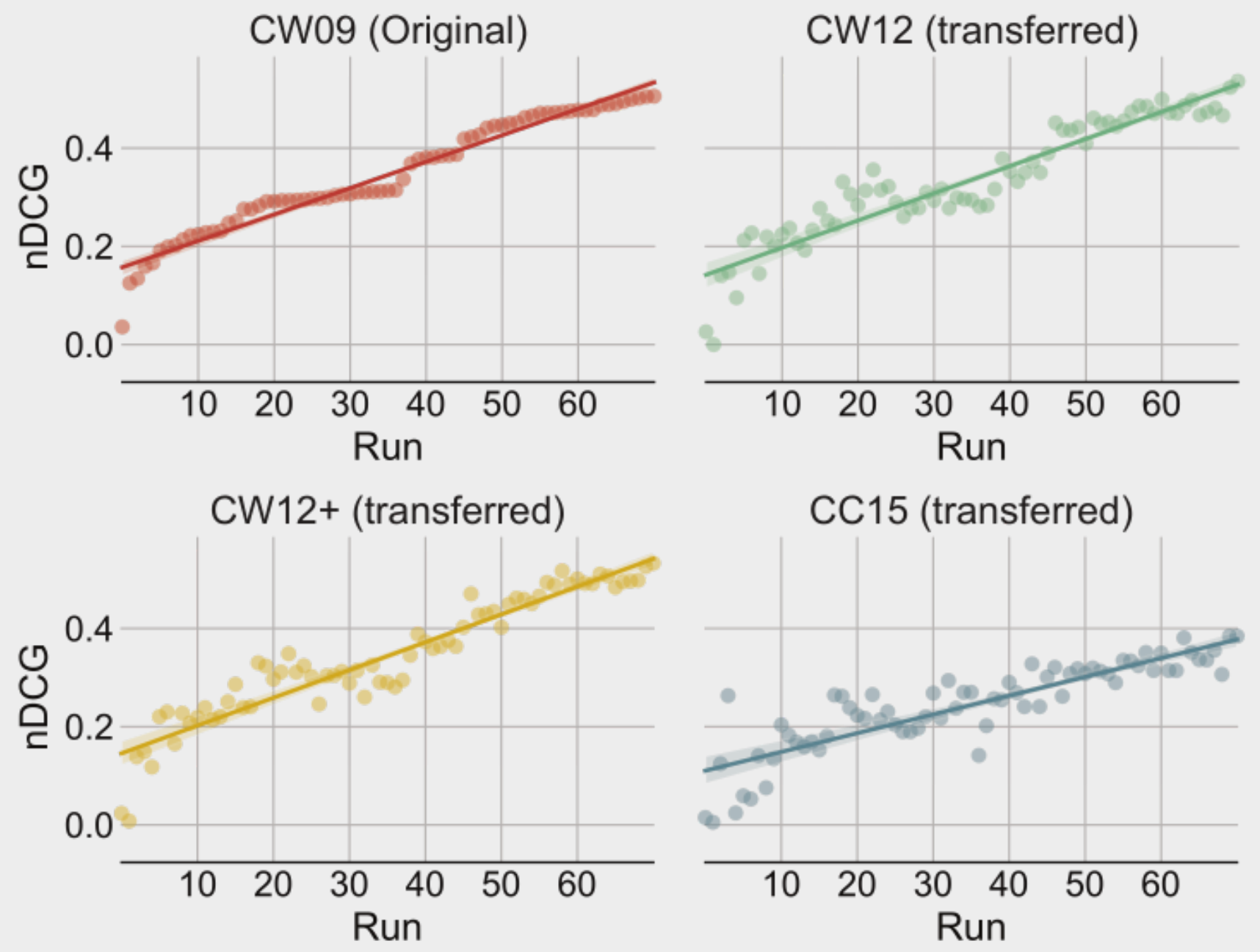
Idea:
Find near-duplicates of documents judged in an older crawl in a newer crawl to reuse existing topics while "saving" relevance judgments.

Example:
The pair shows a ClueWeb09 document judged as relevant for the query `used car parts` and its ClueWeb12 near-duplicate.



Ratio of successfully transferred (ir)relevant judgments targeting the Common Crawl 2015-11 (cc15), the ClueWeb12 (cw12), and a simulated ClueWeb12 including judged cw09 documents in its URL seeds (cw12+).

Changes in the system rankings in the 2009 Web track when only using transferred judgments.



Big Data Analytics www.webis.de

Martin-Luther-Universität Halle-Wittenberg