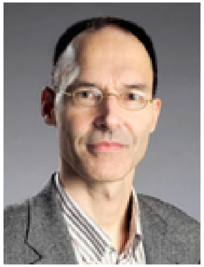
CopyCat: Near-Duplicates Within and Between the ClueWeb and the Common Crawl

Maik Fröbe[1]

Janek Bevendorff[2]

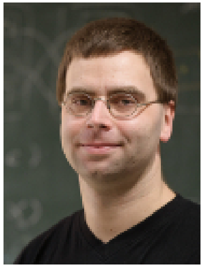
Lukas Gienapp[3]

Michael Völske[2]



Benno Stein[2]

Martin Potthast[3]



Matthias Hagen[1]

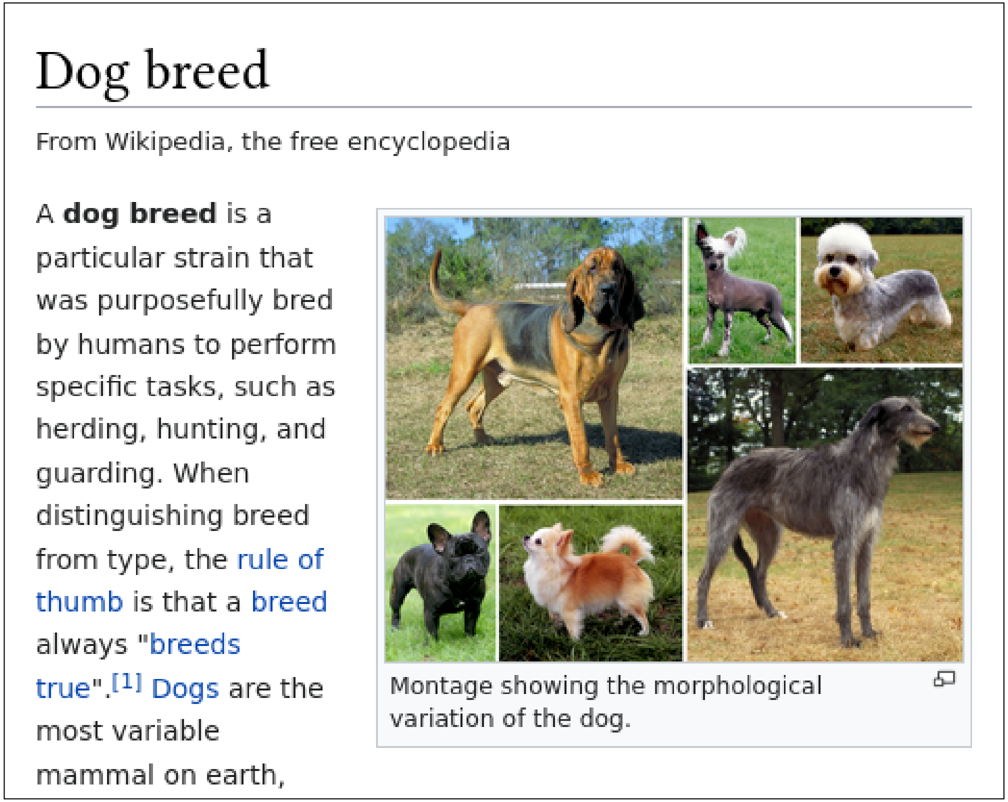
Martin-Luther-Universität Halle-Wittenberg[1] Bauhaus-Universität Weimar[2] Leipzig University[3]

SIGIR, 11–15 July 2021

webis.de

Motivation (1)

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



Dog breed

From Wikipedia, the free encyclopedia

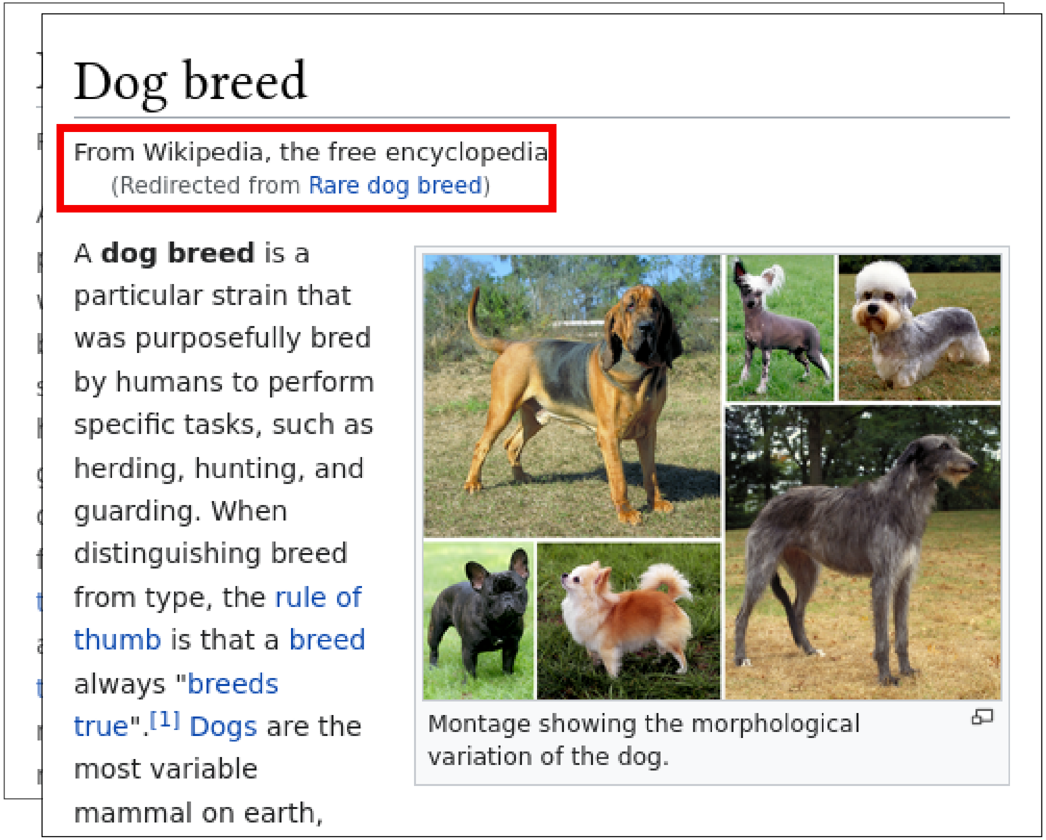
A **dog breed** is a particular strain that was purposefully bred by humans to perform specific tasks, such as herding, hunting, and guarding. When distinguishing breed from type, the rule of thumb is that a breed always "breeds true".[1] Dogs are the most variable mammal on earth,

Montage showing the morphological variation of the dog.

CopyCat: Near-Duplicates in the ClueWeb and Common Crawl

Motivation (1)

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



Dog breed

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

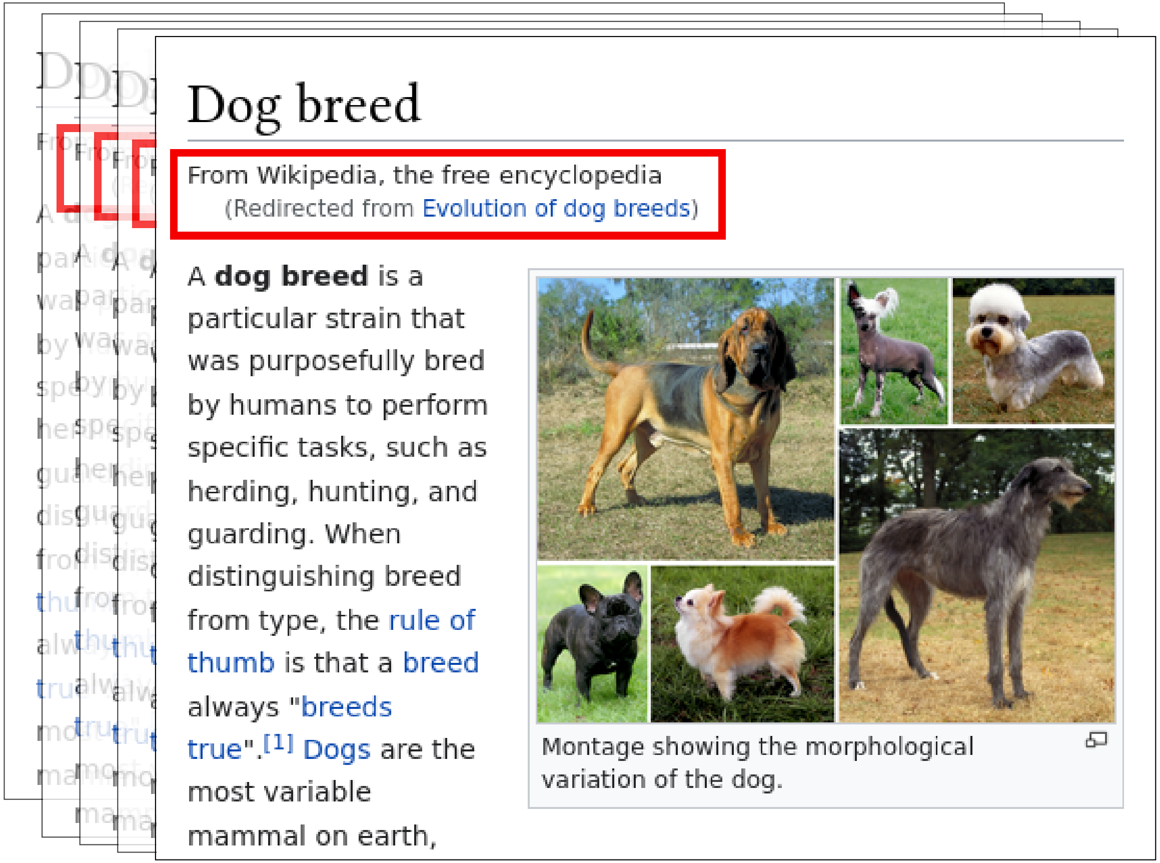
A **dog breed** is a particular strain that was purposefully bred by humans to perform specific tasks, such as herding, hunting, and guarding. When distinguishing breed from type, the rule of thumb is that a breed always "breeds true".[1] Dogs are the most variable mammal on earth,

Montage showing the morphological variation of the dog.

CopyCat: Near-Duplicates in the ClueWeb and Common Crawl

Motivation (1)

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



Dog breed

From Wikipedia, the free encyclopedia
(Redirected from Evolution of dog breeds)

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.

Motivation (2)

- Impact of near-duplicates on the evaluation of search engines:
  [Bernstein et al.; CIKM’05; Fröbe et al.; ECIR’20]

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

Motivation (2)

- ❑ Impact of near-duplicates on the evaluation of search engines:
  [Bernstein et al.; CIKM'05; Fröbe et al.; ECIR'20]

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

  - Novelty principle on the TREC Web Tracks:
    - Average nDCG scores in 2012 decrease by 17 %
    - "Leaderboard" changes (τ of 0.49 in 2010)

Motivation (2)

- ❑ Impact of near-duplicates on the evaluation of search engines:
  [Bernstein et al.; CIKM’05; Fröbe et al.; ECIR’20]
  - Novelty principle:
    A document is irrelevant if it is content-equivalent
    to a document the user has already seen in the results.
  - Novelty principle on the TREC Web Tracks:
    - Average nDCG scores in 2012 decrease by 17 %
    - “Leaderboard” changes (τ of 0.49 in 2010)
- ❑ Sampling bias in learning to rank:
  [Fröbe et al.; SIGIR’20]
  - Unintentional oversampling
  - Bias towards relevant near-duplicates

CopyCat: Near-Duplicates in the ClueWeb and Common Crawl

Resources

- ❑ The CopyCat resource simplifies deduplication in IR experiments

  1. Compilation of near-duplicate documents within widely used web crawls
     - Inclusion and exclusion lists
     - Covered crawls:
       - ClueWeb09, ClueWeb12
       - ClueWeb09B, ClueWeb12B13
       - Common Crawl 2015-11, Common Crawl 2017-04

  2. Software library for deduplication of arbitrary document sets
     - SimHash implemented in Apache Spark for large web crawls
     - CLI for smaller document sets
       - TREC run files
       - TREC qrel files
       - Assessment pools

CopyCat: Near-Duplicates in the ClueWeb and Common Crawl

Deduplication of Large Web Crawls: Ground-Truth for CopyCat

- ❑ Pilot study on canonical links:
  - Canonical links allow authors of web pages to indicate duplicate content
  - Between 0.3 % and 49 % of documents use canonical links
- ❑ Ground-Truth:
  - Semi-automatical assessment of 361 m document pairs
  - Sampled from equivalence classes of canonical links
  - Calculated the exact syntactic similarity for all document pairs
  - Assessed 400 document pairs, choosing a precision-oriented threshold
  - Document pairs with similarity above the threshold are near-duplicates

CopyCat: Near-Duplicates in the ClueWeb and Common Crawl

Deduplication of Large Web Crawls with CopyCat

- SimHash implemented in Apache Spark
- Fine-tuned on 361 m ground-truth document pairs
- CopyCat combines 2 strategies:
  - SimHash within equivalence classes of canonical links (Precision: 0.94)
  - SimHash within entire crawls (Precision: 0.97)
- Deduplication results with CopyCat:
  - Cluster: 135 nodes
  - Resulting inclusion/exclusion lists allow precision-oriented deduplication

	Web crawl			
	cw09	cw12	cc15	cc17
Compr. size	4.0 TB	4.5 TB	28.1 TB	54.0 TB
Documents	1.0 b	731.7 m	1.8 b	3.1 b
Duplicates	145.8 m	204.3 m	951.2 m	1.0 b

Showcase (1): Duplicates in Run Files

- We used the CopyCat CLI to deduplicate run files submitted to the TREC Web Tracks
  - Exact similarity calculation

Web track		**Near-dupl. in runs**	
Year	Runs	@10	@100
2009	71	0.11	0.17
2010	56	0.19	**0.25**
2011	37	**0.21**	0.24
2012	28	0.20	0.18
2013	34	0.12	0.19
2014	30	0.13	0.21

CopyCat: Near-Duplicates in the ClueWeb and Common Crawl

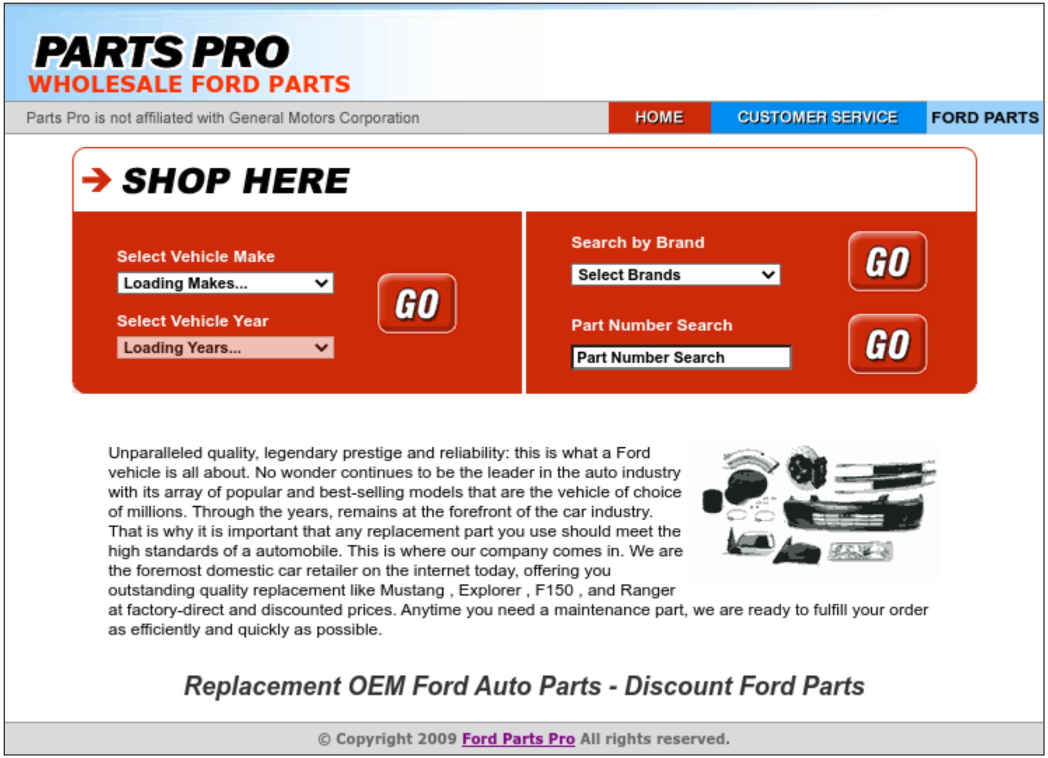
Showcase (2): Relevance Label Transfer

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

CopyCat: Near-Duplicates in the ClueWeb and Common Crawl

Showcase (2): Relevance Label Transfer

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



Parts Pro is not affiliated with General Motors Corporation

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

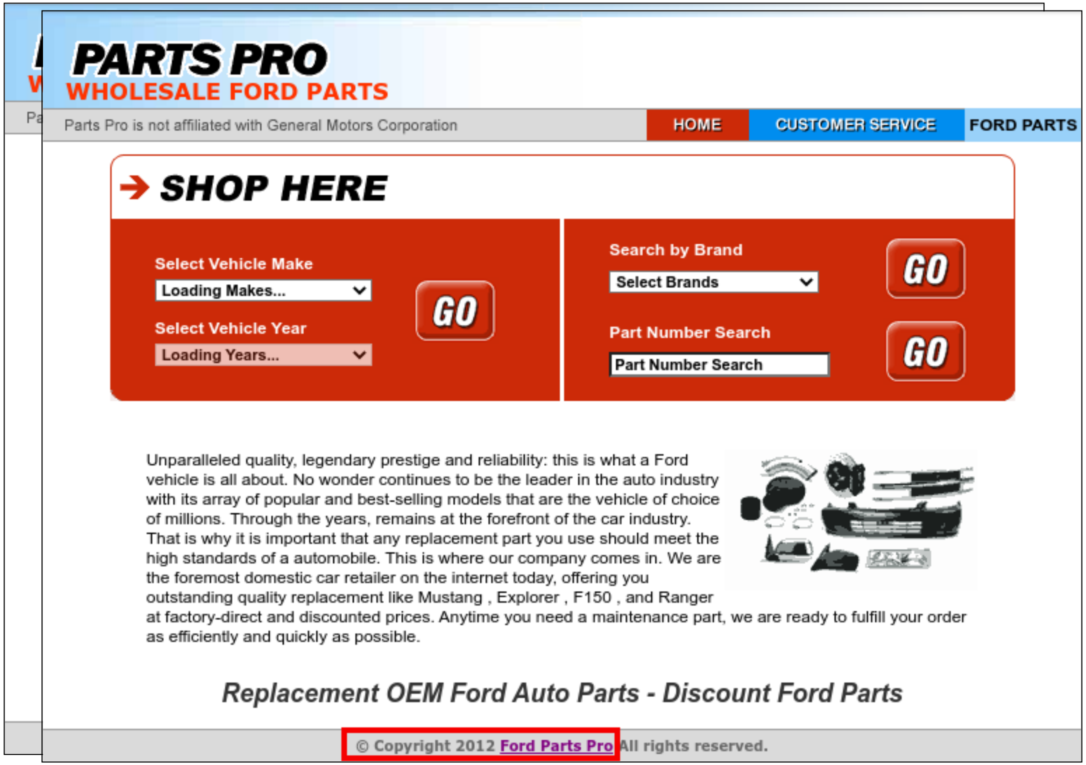
© Copyright 2009 Ford Parts Pro All rights reserved.

  – Relevant for query “used car parts” in ClueWeb09

CopyCat: Near-Duplicates in the ClueWeb and Common Crawl

Showcase (2): Relevance Label Transfer

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



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

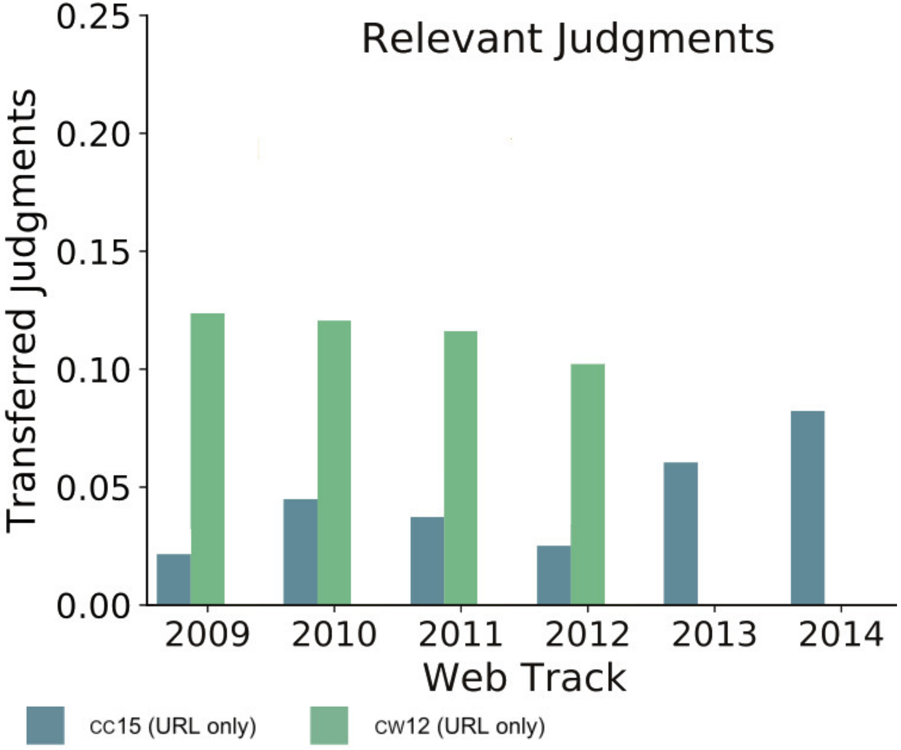
© Copyright 2012 Ford Parts Pro All rights reserved.

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

CopyCat: Near-Duplicates in the ClueWeb and Common Crawl

Showcase (2): Relevance Label Transfer

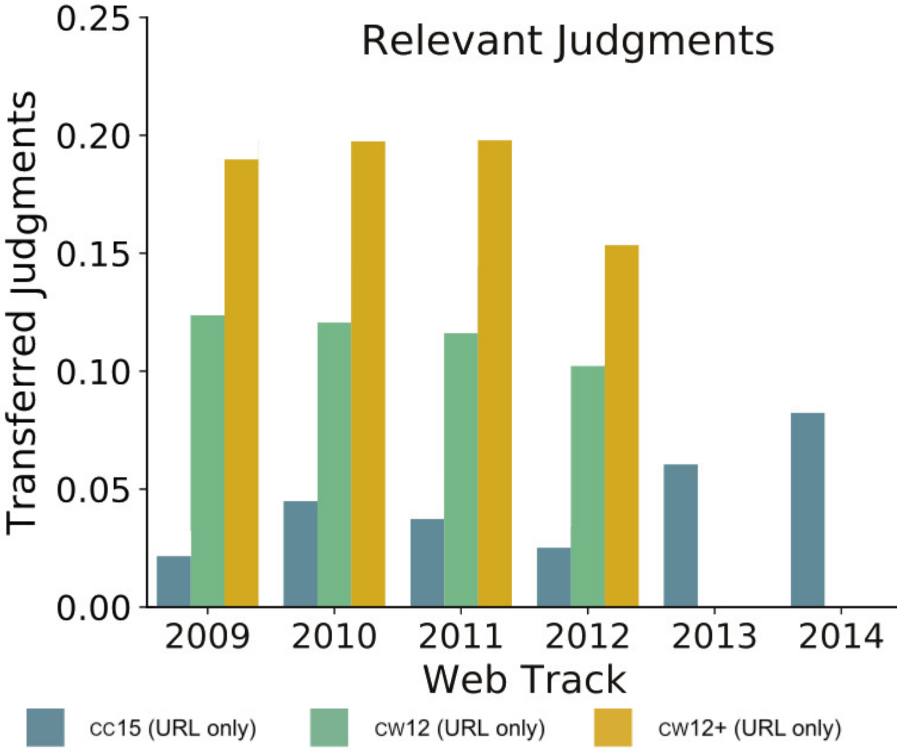
- Experiment settings
  - Qrels deduplicated with CopyCat CLI
  - Precision-oriented near-duplicate threshold



CopyCat: Near-Duplicates in the ClueWeb and Common Crawl

Showcase (2): Relevance Label Transfer

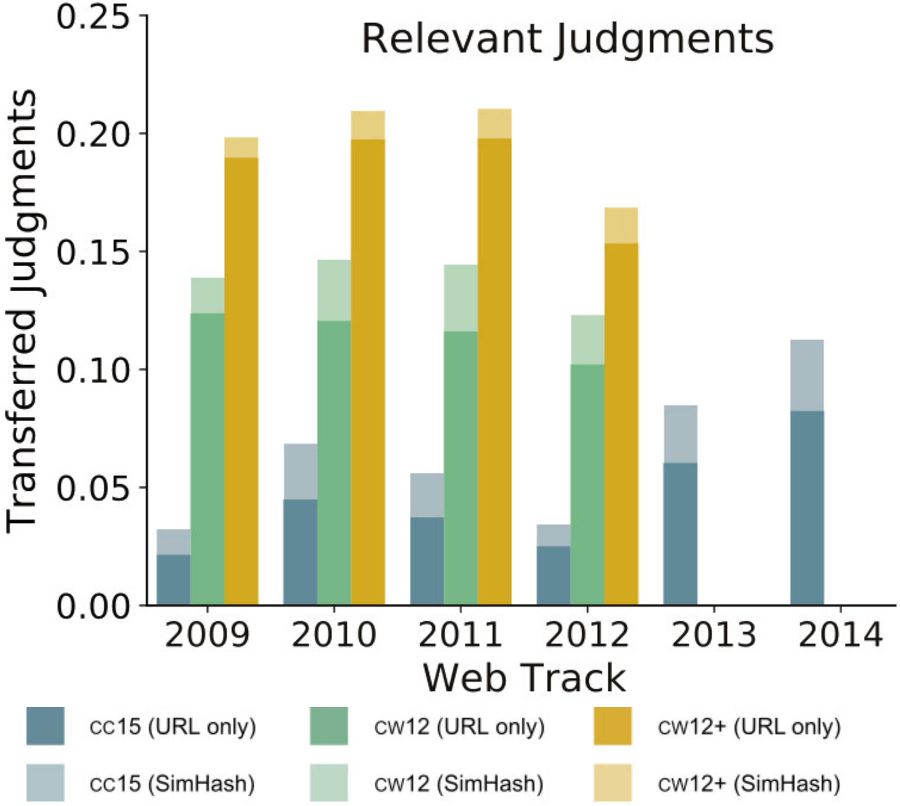
- Experiment settings
  - Qrels deduplicated with CopyCat CLI
  - Precision-oriented near-duplicate threshold



CopyCat: Near-Duplicates in the ClueWeb and Common Crawl

Showcase (2): Relevance Label Transfer

- Experiment settings
  - Qrels deduplicated with CopyCat CLI
  - Precision-oriented near-duplicate threshold



CopyCat: Near-Duplicates in the ClueWeb and Common Crawl

Takeaways

- ❑ CopyCat simplifies deduplication in IR experiments
  - – Ready-to-use inclusion and exclusion lists of near-duplicates
  - – Software library
- ❑ Showcase on relevance label transfer
  - – Few near-duplicates of judged documents occur in newer crawls
  - – New relevance judgments needed to evaluate “old” topics on new crawls
- ❑ Future work:
  - – Increase recall with main content extraction
  - – Consider near-duplicates of relevant passages between documents
- ❑ Code, Paper, Slides: webis.de/publications

CopyCat: Near-Duplicates in the ClueWeb and Common Crawl

Takeaways

- CopyCat simplifies deduplication in IR experiments
  - Ready-to-use inclusion and exclusion lists of near-duplicates
  - Software library
- Showcase on relevance label transfer
  - Few near-duplicates of judged documents occur in newer crawls
  - New relevance judgments needed to evaluate “old” topics on new crawls
- Future work:
  - Increase recall with main content extraction
  - Consider near-duplicates of relevant passages between documents
- Code, Paper, Slides: webis.de/publications

Thank You!