

# Evaluation of Ad-Hoc Information Retrieval Systems on Web Crawls with Redundant Documents

Jan Philipp Bittner

October 19, 2020

# Einfluss von Near-Duplicates auf IR-Evaluationen

## Motivation

- ❑ Benchmark nach Cranfield-Paradigma: Dokumente unabhängig voneinander bewerten
- ❑ Query: "designer dog breeds"



Dog hybrid



A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a [new term](#) for an individual dog with ancestry in two different [purebred dog breeds](#), traditionally called a [crossbred dog](#). Designer dogs are



Dog hybrid



A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a [new term](#) for an individual dog with ancestry in two different [purebred dog breeds](#).  
Redirected from "[Shilohom](#)"  
[crossbred dog](#): designer dogs are



Mathematics

**Mathematics** (from [Greek](#): μάθημα, *máthēma*, 'knowledge, study, learning')  
Includes the study of such topics as quantity (number theory),<sup>[1]</sup> structure (algebra),<sup>[2]</sup> space (geometry),<sup>[3]</sup> and change (mathematical analysis).<sup>[3][4][5]</sup> It has no generally accepted definition.<sup>[6][7]</sup>



Greek mathematician Euclid (holding calipers), 3rd century BC, as imagined by Raphael in this detail from [The](#)

# Einfluss von Near-Duplicates auf IR-Evaluationen

## Motivation

- ❑ Benchmark nach Cranfield-Paradigma: Dokumente unabhängig voneinander bewerten
- ❑ Query: "designer dog breeds"

≡ WIKIPEDIA 

Dog hybrid



A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a [new term](#) for an individual dog with ancestry in two different [purebred dog breeds](#), traditionally called a [crossbred dog](#). Designer dogs are

≡ WIKIPEDIA 

Dog hybrid



A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a [new term](#) for an individual dog with ancestry in two different [purebred dog breeds](#). Redirected from "[Shilohom](#)".

≡ WIKIPEDIA 

Mathematics

**Mathematics** (from [Greek](#): μάθημα, *máthēma*, 'knowledge, study, learning') includes the study of such topics as quantity ([number theory](#)),<sup>[1]</sup> structure ([algebra](#)),<sup>[2]</sup> space ([geometry](#)),<sup>[3]</sup> and change ([mathematical analysis](#)).<sup>[3][4][5]</sup> It has no generally accepted [definition](#).<sup>[6][7]</sup>



Greek mathematician **Euclid** (holding calipers), 3rd century BC, as imagined by Raphael in this detail from [The](#)

# Einfluss von Near-Duplicates auf IR-Evaluationen

## Motivation

- ❑ Benchmark nach Cranfield-Paradigma: Dokumente unabhängig voneinander bewerten
- ❑ Query: "designer dog breeds"

≡ WIKIPEDIA 

Dog hybrid



A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a [new term](#) for an individual dog with ancestry in two different [purebred dog breeds](#), traditionally called a [crossbred dog](#). Designer dogs are



≡ WIKIPEDIA 

Dog hybrid



A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a [new term](#) for an individual dog with ancestry in two different [purebred dog breeds](#). Redirected from "[Shilohom](#)"

≡ WIKIPEDIA 

Mathematics

**Mathematics** (from Greek: μάθημα, *máthēma*, 'knowledge, study, learning') includes the study of such topics as quantity (number theory),<sup>[1]</sup> structure (algebra),<sup>[2]</sup> space (geometry),<sup>[3]</sup> and change (mathematical analysis).<sup>[3][4][5]</sup> It has no generally accepted definition.<sup>[6][7]</sup>



Greek mathematician Euclid (holding calipers), 3rd century BC, as imagined by Raphael in this detail from [The](#)

# Einfluss von Near-Duplicates auf IR-Evaluationen

## Motivation

- ❑ Benchmark nach Cranfield-Paradigma: Dokumente unabhängig voneinander bewerten
- ❑ Query: "designer dog breeds"



Dog hybrid



A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a [new term](#) for an individual dog with ancestry in two different [purebred dog breeds](#), traditionally called a [crossbred dog](#). Designer dogs are

A screenshot of a Wikipedia page for "Dog hybrid" with an orange border around the entire content area. The title "Dog hybrid" is at the top, followed by a sub-section "A pointer - dalmatian cross." with a photo of a white dog with black spots. Below this, there is a large block of text about dog hybrids, which includes a redacted section at the bottom. A green checkmark is overlaid on the left side of the page.

Dog hybrid



A **dog hybrid** (also called a designer dog) is a [new term](#) for an individual dog with ancestry in two different [purebred dog breeds](#), traditionally called a [crossbred dog](#). Designer dogs are

Redirected from "[Shilohom](#)"



Mathematics

**Mathematics** (from Greek: μάθημα, *máthēma*, "knowledge, study, learning") includes the study of such topics as quantity (number theory),<sup>[1]</sup> structure (algebra),<sup>[2]</sup> space (geometry),<sup>[3]</sup> and change (mathematical analysis).<sup>[3][4][5]</sup> It has no generally accepted definition.<sup>[6][7]</sup>



Greek mathematician Euclid (holding calipers), 3rd century BC, as imagined by Raphael in this detail from [The](#)

# Einfluss von Near-Duplicates auf IR-Evaluationen

## Motivation

- ❑ Benchmark nach Cranfield-Paradigma: Dokumente unabhängig voneinander bewerten
- ❑ Query: "designer dog breeds"



Dog hybrid



A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a [new term](#) for an individual dog with ancestry in two different [purebred dog breeds](#), traditionally called a [crossbred dog](#). Designer dogs are

A screenshot of a Wikipedia search results page for "Dog hybrid". The search bar at the top has "WIKIPEDIA" and a magnifying glass icon. Below the search bar, the first result is titled "Dog hybrid" with a sub-section "A pointer - dalmatian cross." featuring a photo of a white dog with black spots. A green checkmark is positioned to the right of the image. This result is visually identical to the one on the left but is enclosed in an orange border.

Dog hybrid



A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a [new term](#) for an individual dog with ancestry in two different [purebred dog breeds](#). Redirected from "[Shilohom](#)"



Mathematics

**Mathematics** (from Greek: μάθημα, *máthēma*, "knowledge, study, learning") includes the study of such topics as quantity (number theory),<sup>[1]</sup> structure (algebra),<sup>[2]</sup> space (geometry),<sup>[3]</sup> and change (mathematical analysis).<sup>[3][4][5]</sup> It has no generally accepted definition.<sup>[6][7]</sup>



Greek mathematician Euclid (holding calipers), 3rd century BC, as imagined by Raphael in this detail from *The*

# Einfluss von Near-Duplicates auf IR-Evaluationen

## Motivation

- ❑ Benchmark nach Cranfield-Paradigma: Dokumente unabhängig voneinander bewerten
- ❑ Query: "designer dog breeds"



Dog hybrid



A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a [new term](#) for an individual dog with ancestry in two different [purebred dog breeds](#), traditionally called a [crossbred dog](#). Designer dogs are



Dog hybrid



A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a [new term](#) for an individual dog with ancestry in two different [purebred dog breeds](#). Redirected from "[Shilohom](#)"  
[crossbred dog](#). Designer dogs are



Mathematics

**Mathematics** (from Greek: μάθημα, *máthēma*, "knowledge, study, learning") includes the study of such topics as quantity (number theory),<sup>[1]</sup> structure (algebra),<sup>[2]</sup> space (geometry),<sup>[3]</sup> and change (mathematical analysis).<sup>[3][4][5]</sup> It has no generally accepted definition.<sup>[6][7]</sup>



Greek mathematician Euclid (holding calipers), 3rd century BC, as imagined by Raphael in this detail from *The*

# Einfluss von Near-Duplicates auf IR-Evaluationen

## Motivation

- ❑ Benchmark nach Cranfield-Paradigma: Dokumente unabhängig voneinander bewerten
- ❑ Query: "designer dog breeds"



A screenshot of a Wikipedia search result page for "Mathematics". The search bar at the top has "WIKIPEDIA" and a magnifying glass icon. Below the search bar, the query "Mathematics" is entered. The main content area includes a definition of Mathematics from Greek, mentioning topics like number theory, algebra, geometry, and analysis. It notes that there is no generally accepted definition. A large red X mark is overlaid on the right side of the page. Below the main content, there is a small image of Euclid and a caption about him.

Mathematics

**Mathematics** (from Greek: μάθημα, *máthēma*, "knowledge, study, learning")

Includes the study of such topics as quantity (number theory),<sup>[1]</sup> structure (algebra),<sup>[2]</sup> space (geometry),<sup>[3]</sup> and change (mathematical analysis).<sup>[3][4][5]</sup> It has no generally accepted definition.<sup>[6][7]</sup>



Greek mathematician Euclid (holding calipers), 3rd century BC, as imagined by Raphael in this detail from *The*



A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a [new term](#) for an individual dog with ancestry in two different [purebred dog breeds](#), traditionally called a [crossbred dog](#). Designer dogs are



A **dog hybrid** (also called a designer dog) is a [new term](#) for an individual dog with ancestry in two different [purebred dog breeds](#). Redirected from "Shihpom" [crossbred dog](#). Designer dogs are

# Einfluss von Near-Duplicates auf IR-Evaluationen

## Motivation

- ❑ Nutzersicht nach Novelty-Principle: Duplikate sind irrelevant
- ❑ Query: "designer dog breeds"



Dog hybrid



A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a [new term](#) for an individual **dog** with ancestry in two different [purebred dog breeds](#), traditionally called a [crossbred dog](#). Designer dogs are



Dog hybrid



A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a [new term](#) for an individual **dog** with ancestry in two different [purebred dog breeds](#), traditionally called a [crossbred dog](#). Designer dogs are



Mathematics

**Mathematics** (from [Greek: μάθημα, \*mathēma\*, 'knowledge, study, learning'\)](#) includes the study of such topics as [quantity \(number theory\)](#),<sup>[1]</sup> [structure \(algebra\)](#),<sup>[2]</sup> [space \(geometry\)](#),<sup>[1]</sup> and [change \(mathematical analysis\)](#).<sup>[3][4][5]</sup> It has no generally accepted [definition](#).<sup>[6][7]</sup>



Greek mathematician Euclid (holding calipers), 3rd century BC, as imagined by Raphael in this detail from *The*

# Einfluss von Near-Duplicates auf IR-Evaluationen

## Motivation

- ❑ Nutzersicht nach Novelty-Principle: Duplikate sind irrelevant
- ❑ Query: "designer dog breeds"

WIKIPEDIA 

### Dog hybrid



A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a [new term](#) for an individual [dog](#) with ancestry in two different [purebred dog breeds](#), traditionally called a [crossbred dog](#). Designer dogs are

WIKIPEDIA 

### Dog hybrid



A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a [new term](#) for an individual [dog](#) with ancestry in two different [purebred dog breeds](#), traditionally called a [crossbred dog](#). Designer dogs are

dog Redirected from "Shibaom"

WIKIPEDIA 

### Mathematics

**Mathematics** (from [Greek: μάθημα, "knowledge, study, learning"](#)) includes the study of such topics as [quantity \(number theory\)](#),<sup>[1]</sup> [structure \(algebra\)](#),<sup>[2]</sup> [space \(geometry\)](#),<sup>[1]</sup> and [change \(mathematical analysis\)](#).<sup>[3][4][5]</sup> It has no generally accepted [definition](#).<sup>[6][7]</sup>



Greek mathematician Euclid (holding calipers), 3rd century BC, as imagined by Raphael in this detail from [The](#)

# Einfluss von Near-Duplicates auf IR-Evaluationen

## Motivation

- ❑ Nutzersicht nach Novelty-Principle: Duplikate sind irrelevant
- ❑ Query: "designer dog breeds"

WIKIPEDIA 

### Dog hybrid



A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a **new term** for an individual **dog** with ancestry in two different **purebred dog breeds**, traditionally called a **crossbred dog**. Designer dogs are



WIKIPEDIA 

### Dog hybrid



A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a **new term** for an individual **dog** with ancestry in two different **purebred dog breeds**, traditionally called a **crossbred dog**. Designer dogs are

Redirected from "Shibaom"

WIKIPEDIA 

### Mathematics

**Mathematics** (from Greek: μάθημα, *mathēma*, 'knowledge, study, learning') includes the study of such topics as **quantity** (**number theory**),<sup>[1]</sup> **structure** (**algebra**),<sup>[2]</sup> **space** (**geometry**),<sup>[1]</sup> and **change** (**mathematical analysis**).<sup>[3][4][5]</sup> It has no generally accepted **definition**.<sup>[6][7]</sup>



Greek mathematician Euclid (holding calipers), 3rd century BC, as imagined by Raphael in this detail from *The*

# Einfluss von Near-Duplicates auf IR-Evaluationen

## Motivation

- ❑ Nutzersicht nach Novelty-Principle: Duplikate sind irrelevant
- ❑ Query: "designer dog breeds"

WIKIPEDIA

Dog hybrid



A pointer - dalmatian cross.

A dog hybrid (also called a designer dog) is a new term for an individual dog with ancestry in two different purebred dog breeds, traditionally called a crossbred dog. Designer dogs are

WIKIPEDIA

Dog hybrid



A pointer - dalmatian cross.

A dog hybrid (also called a designer dog) is a new term for an individual dog with ancestry in two different purebred dog breeds, traditionally called a crossbred dog. Designer dogs are

WIKIPEDIA

Mathematics

Mathematics (from Greek: μάθημα, 'knowledge, study, learning') includes the study of such topics as quantity (number theory),<sup>[1]</sup> structure (algebra),<sup>[2]</sup> space (geometry),<sup>[1]</sup> and change (mathematical analysis).<sup>[3][4][5]</sup> It has no generally accepted definition.<sup>[6][7]</sup>



Greek mathematician Euclid (holding calipers), 3rd century BC, as imagined by Raphael in this detail from *The*

# Einfluss von Near-Duplicates auf IR-Evaluationen

## Motivation

- ❑ Nutzersicht nach Novelty-Principle: Duplikate sind irrelevant
- ❑ Query: "designer dog breeds"

WIKIPEDIA

Dog hybrid



A pointer - dalmatian cross.

A dog hybrid (also called a designer dog) is a new term for an individual dog with ancestry in two different purebred dog breeds, traditionally called a crossbred dog. Designer dogs are

WIKIPEDIA

Dog hybrid



A pointer - dalmatian cross.

A dog hybrid (also called a designer dog) is a new term for an individual dog with ancestry in two different purebred dog breeds, traditionally called a crossbred dog. Designer dogs are

WIKIPEDIA

Mathematics

Mathematics (from Greek: μάθημα, *mathēma*, 'knowledge, study, learning') includes the study of such topics as quantity (number theory),<sup>[1]</sup> structure (algebra),<sup>[2]</sup> space (geometry),<sup>[1]</sup> and change (mathematical analysis).<sup>[3][4][5]</sup> It has no generally accepted definition.<sup>[6][7]</sup>



Greek mathematician Euclid (holding calipers), 3rd century BC, as imagined by Raphael in this detail from *The*

# Einfluss von Near-Duplicates auf IR-Evaluationen

## Motivation

- ❑ Nutzersicht nach Novelty-Principle: Duplikate sind irrelevant
- ❑ Query: "designer dog breeds"

≡ WIKIPEDIA 

≡ WIKIPEDIA 

≡ WIKIPEDIA 

Dog hybrid



A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a **new term** for an individual **dog** with ancestry in two different **purebred dog breeds**, traditionally called a **crossbred dog**. Designer dogs are



Dog hybrid



A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a **new term** for an individual **dog** with ancestry in two different **purebred dog breeds**, traditionally called a **crossbred dog**. Designer dogs are



Mathematics

**Mathematics** (from Greek: μάθημα, *mathēma*, 'knowledge, study, learning') includes the study of such topics as **quantity** (number theory),<sup>[1]</sup> **structure** (algebra),<sup>[2]</sup> **space** (geometry),<sup>[1]</sup> and **change** (mathematical analysis).<sup>[3][4][5]</sup> It has no generally accepted definition.<sup>[6][7]</sup>



Greek mathematician Euclid (holding calipers), 3rd century BC, as imagined by Raphael in this detail from *The*

2/14

# Einfluss von Near-Duplicates auf IR-Evaluationen

## Motivation

- ❑ Nutzersicht nach Novelty-Principle: Duplikate sind irrelevant
- ❑ Query: "designer dog breeds"

≡ WIKIPEDIA 

≡ WIKIPEDIA 

≡ WIKIPEDIA 

Dog hybrid



A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a **new term** for an individual **dog** with ancestry in two different **purebred dog breeds**, traditionally called a **crossbred dog**. Designer dogs are



Dog hybrid



A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a **new term** for an individual **dog** with ancestry in two different **purebred dog breeds**, traditionally called a **crossbred dog**. Designer dogs are



Mathematics

**Mathematics** (from Greek: μάθημα, *mathēma*, 'knowledge, study, learning') includes the study of such topics as **quantity** (number theory),<sup>[1]</sup> **structure** (algebra),<sup>[2]</sup> **space** (geometry),<sup>[1]</sup> and **change** (mathematical analysis).<sup>[3][4][5]</sup> It has no generally accepted definition.<sup>[6][7]</sup>





Greek mathematician Euclid (holding calipers), 3rd century BC, as imagined by Raphael in this detail from *The*

2/14

# Einfluss von Near-Duplicates auf IR-Evaluationen

## Motivation

Bewertung (Cranfield):

WIKIPEDIA Q

Dog hybrid



WIKIPEDIA Q

Dog hybrid



WIKIPEDIA Q

Mathematics

Mathematics (from Greek μάθημα, "mathema", knowledge, study, learning) includes the study of such topics as quantity (number theory),<sup>11</sup> structure (algebra),<sup>12</sup> space (geometry),<sup>13</sup> and change (mathematical analysis).<sup>14</sup> It has no generally accepted definition.<sup>15</sup>



Greek mathematician Euclid (holding compass), 3rd century BC, as imagined by Fischbeck in this detail from the



Realität (Novelty):

WIKIPEDIA Q

Dog hybrid



WIKIPEDIA Q

Dog hybrid



WIKIPEDIA Q

Mathematics

Mathematics (from Greek μάθημα, "mathema", knowledge, study, learning) includes the study of such topics as quantity (number theory),<sup>11</sup> structure (algebra),<sup>12</sup> space (geometry),<sup>13</sup> and change (mathematical analysis).<sup>14</sup> It has no generally accepted definition.<sup>15</sup>



Greek mathematician Euclid (holding compass), 3rd century BC, as imagined by Fischbeck in this detail from the



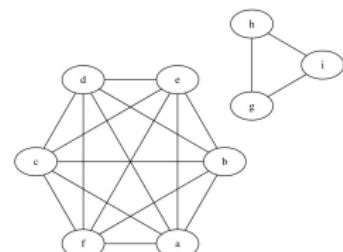
Beeinflussen Duplikate die Auswertung?  
Wie viele Duplikate gibt es in den Corpora?

# Einfluss von Near-Duplicates auf IR-Evaluationen

## Pilotstudie - Retrieval-Equivalence

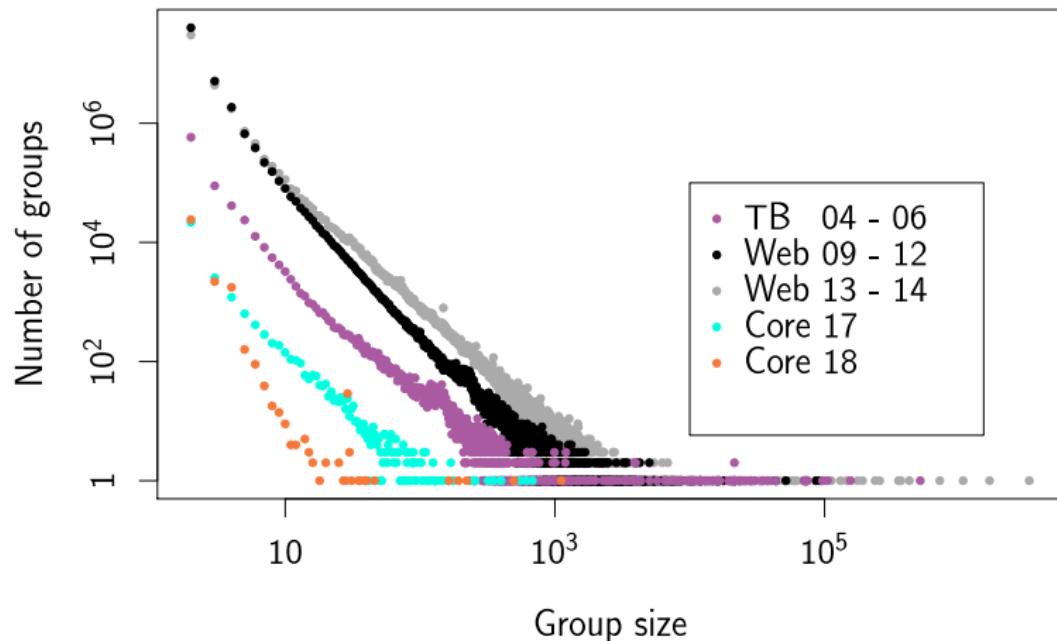
- ❑ Duplikate "kostengünstig" erkennen mittels Hashing
  - ▶ Äquivalenz bei Indexierung feststellbar (Retrieval-Equivalence)
- ❑ Pipeline: Text extrahieren, filtern, stemmen, hashen
- ❑ Gruppieren der Dokumente nach Hash-Wert
- ❑ Ergebnis: Gruppen äquivalenter Dokumente

Shared-Task	Docs	Duplicates
Terabyte 2004 - 2006	$2.5 * 10^7$	23.39%
Web 2009 - 2012	$1.0 * 10^9$	7.74%
Web 2013 - 2014	$7.3 * 10^8$	14.71%
Core 2017	$1.8 * 10^6$	2.11%
Core 2018	$5.9 * 10^5$	12.52%



# Einfluss von Near-Duplicates auf IR-Evaluationen

## Pilotstudie - Retrieval-Equivalence



# Einfluss von Near-Duplicates auf IR-Evaluationen

## Pilotstudie - Retrieval-Equivalence

- Reicht Retrieval-Equivalence?
  - ▶ Nein, Near-Duplicates werden zum Teil nicht erkannt



WIKIPEDIA

Dog hybrid



A pointer - dalmatian cross.

A dog hybrid (also called a designer dog) is a new term for an individual dog with ancestry in two different purebred dog breeds, traditionally called a crossbred dog. Designer dogs are



WIKIPEDIA

Dog hybrid



A pointer - dalmatian cross.

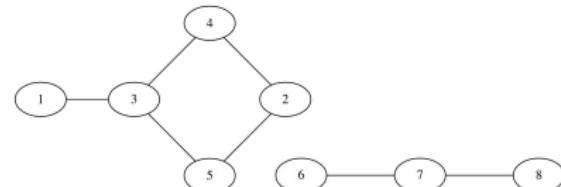
A dog hybrid (also called a designer dog) is a new term for an individual dog with ancestry in two different purebred dog breeds, traditionally called a crossbred dog. Designer dogs are

- Dokumente mit unterschiedlichem Hash

# Einfluss von Near-Duplicates auf IR-Evaluationen

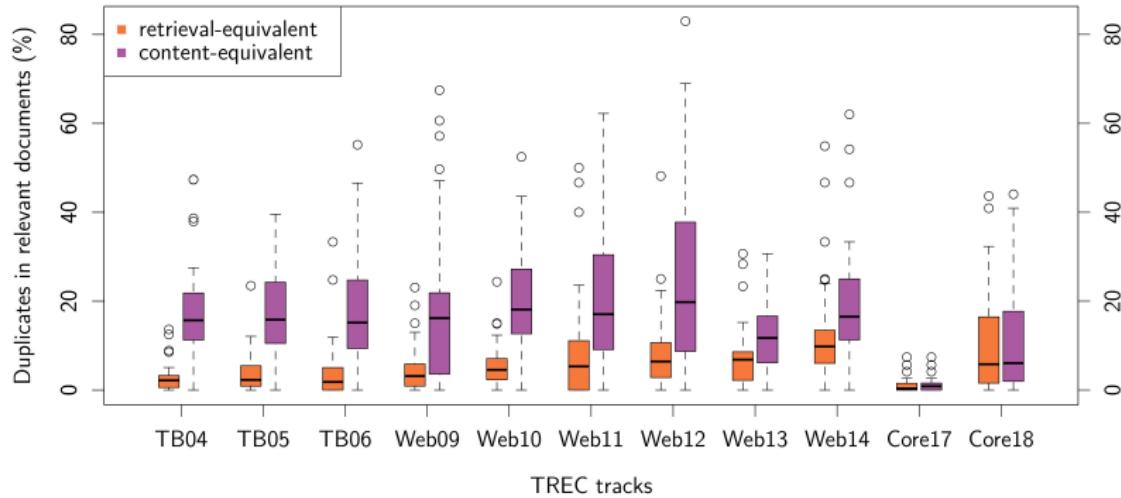
## Content-Equivalence

- Zwei Dokument sind Content-Equivalent, wenn sie hinreichende Überschneidungen haben
- Messung durch Bernstein und Zobel's S3-Score
  - ▶ Normalisierte Anzahl der gemeinsamen n-Gramme
  - ▶ Threshold trennt Duplikate und Unikate
  - ▶ Berechnung ganzer Corpora zu teuer → bewertete Dokumente
- Pipeline: Text extrahieren, filter, stemmen und **in n-Gramme zerlegen**
- Anzahl gemeinsamer n-Gramme zweier Dokumente zählen
- Kette von äquivalenten Dokumenten möglich
  - ▶ Beispiel: Versionsverläufe



# Einfluss von Near-Duplicates auf IR-Evaluationen

## Content-Equivalence



# Einfluss von Near-Duplicates auf IR-Evaluationen

## Impact

- ❑ Duplikat-sensitive Auswertung
  - ▶ Bewertung nach Novelty-Principle
  - ▶ Vorher-Nacher-Vergleich

- ❑ nDCG: Gütemaß mit Werten zwischen 0 und 1, je höher desto besser
- ❑ Submissions schneiden schlechter ab
  - ▶ Duplikate werden angezeigt

Track	avg <sub>nDCG</sub>	ΔnDCG
Terabyte	2004	0.425
	2005	0.586
	2006	0.654
Web	2009	-8.9%
	2010	-14.1%
	2011	-9.0%
	2012	-17.3%
	2013	-4.6%
	2014	-7.9%
Core	2017	-0.3%
	2018	-4.3%

# Einfluss von Near-Duplicates auf IR-Evaluationen

## Impact

- ❑ Duplikat-sensitive Auswertung
  - ▶ Bewertung nach Novelty-Principle
  - ▶ Vorher-Nacher-Vergleich

- ❑  $\tau$ : Rangkorrelation zwischen konventioneller Auswertung und Duplikat-sensitiver Auswertung
- ❑ Sehr unterschiedlich starke Ausprägungen

Track	$\tau$	$\tau@5$
Terabyte	2004	0.96
	2005	0.95
	2006	0.94
Web	2009	0.89
	2010	0.49
	2011	0.85
	2012	0.72
	2013	0.86
	2014	0.87
Core	2017	0.99
	2018	0.92

# Einfluss von Near-Duplicates auf IR-Evaluationen

## Impact

- Duplikat-sensitive Auswertung
  - ▶ Bewertung nach Novelty-Principle
  - ▶ Vorher-Nacher-Vergleich

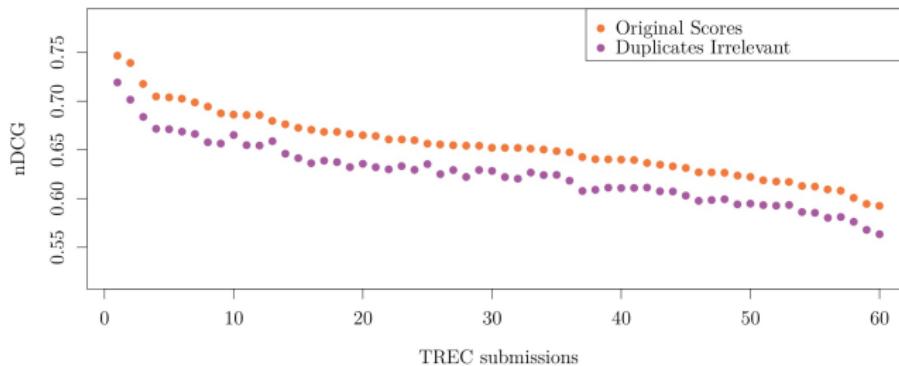
- $i$ : Ideales System, setzt Novelty-Principle um
- $max_i$ : Ideales System, das die meisten Plätze verliert
- $med_i$ : Ideales System, das im Median bezüglich der Platzverluste liegt

Track	$med_I$	$max_I$
Terabyte	2004	-9.0
	2005	-15.5
	2006	-29.5
Web	-53	
	2009	-8.5
	2010	-19.5
	2011	-8.0
	2012	-9.0
	2013	-4.0
Core	2014	-8
	-11	
	2017	-1.0
	2018	-9
	-26	

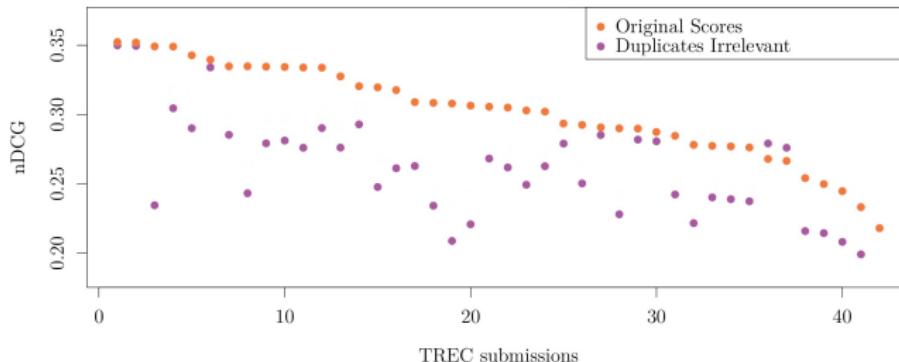
# Einfluss von Near-Duplicates auf IR-Evaluationen

## Impact - Terabyte 2006 vs. Web 2010

TERABYTE 2006



WEB 2010



# Einfluss von Near-Duplicates auf IR-Evaluationen

## Was können Organisatoren von Shared-Tasks machen?

- Duplikat-freie Copora
  - + Prozesse bleiben unverändert
  - Erstellungskosten steigen mit Größe
- Duplikat-sensitive Auswertung
  - + Vollständige Auswertung nach Novelty-Principle
  - Auswertung erfordert Duplikatsgruppen und zusätzliches Auswertungsprogramm
- riskante Topics entfernen
  - + Kompatibel mit allen Auswertungen
  - Auswahlkriterien fragwürdig
  - verbirgt Systeme, die Duplikate schlecht behandeln

# Einfluss von Near-Duplicates auf IR-Evaluationen

## Take Home Message

- Auch neuere Web-Corpora enthalten Duplikate
- Rankings werden durch Duplikate verzerrt
- Je weniger Duplikate, desto weniger Verzerrung
- Kein Universalkonzept zur Einflussminderung
  - ▶ beste Option: Duplikat-sensitive Auswertung

# Einfluss von Near-Duplicates auf IR-Evaluationen

## Take Home Message

- Auch neuere Web-Corpora enthalten Duplikate
- Rankings werden durch Duplikate verzerrt
- Je weniger Duplikate, desto weniger Verzerrung
- Kein Universalkonzept zur Einflussminderung
  - ▶ beste Option: Duplikat-sensitive Auswertung

Vielen Dank für Ihre Aufmerksamkeit! Nun zu ihren Fragen.

Kontakt: [jan.bittner@student.uni-halle.de](mailto:jan.bittner@student.uni-halle.de)