Inleiding Data Science: Algorithmic Bias

CPO & iCIS Opleiding gegevensbescherming

Prof.dr.ir. Arjen P. de Vries

arjen@acm.org

Nijmegen, May 16th, 2019

GEEN MR…

Radboud University Nijmegen

Eigenlijk niet…

Opdracht van Mireille:
Kun je studenten uitleggen hoe je een ML systeem zo kunt ontwerpen dat je daadwerkelijk zo weinig mogelijk data verwerkt en toch kwaliteit levert?

KIA Digitale Technologieën – AI (ST2-1), Q2:
Op welke wijze kunnen we deep-learning architecturen data-efficiënt training door expert kennis toe te voegen?

Outline

- Big Data
  - Waarom nou juist BIG data?!

- Bias on the Web
- Bias in Search Engines
  - Quantifying Bias using Retrievability

- Bias in Word Embeddings
  - informagus.nl/edu/cpo/

Big Data: nieuwe mogelijkheden!

Om
- Data te genereren,
- Delen,
- Combineren
- Analyseren

.. die leiden tot nieuwe inzichten en een nieuwe manier van redeneren.

(bron: Definitie van big data van de *Nationale DenkTank*)

Al het werk van William Shakespeare: 5.500.000 letters…

…wordt geschreven in slechts 11,6 Twitter-seconde!

TO TWEET OR NOT TO TWEET?

William Shakespeare's collected works occupies about 5.5 million characters, roughly as much text is tweeted every 11.6 seconds. To tweet or not to tweet, that is the question.

Een wonderlijke puzzel!

8 TB per day in total $\approx$ 100 MB per sec

Data Center (is the Computer)

Big Data in NL

- Aankopen bij bol.com
- Bestemmingen van Booking.com op maat
- Advertenties real-time verkopen, bv. bij nu.nl (Sanoma)
- Kinderzoekmachine WizeNoze.com

Etc. etc.

Cracking the Voynich Code

Two women dealing with five children in a bath

The words describe different temperaments: tozosr (buzzing: too noisy), orla la (on the edge: losing patience), tolora (silly/foolish), noror (cloudy: dull/sad), or aus (golden bird: well behaved), oleios (oiled: slippery). These words survive in Catalan [tozos], Portuguese [orla], Portuguese [tolos], Romanian [noros], Catalan [or aus] and Portuguese [oleio]. The words orla la describe the mood of the woman on the left and may well be the root of the French phrase 'oh là là', which has a very similar sentiment.

Cracking the Voynich Code

"It uses an extinct language. Its alphabet is a combination of unfamiliar and more familiar symbols. It includes no dedicated punctuation marks, although some letters have symbol variants to indicate punctuation or phonetic accents. All of the letters are in lower case and there are no double consonants. It includes diphthong, triphthongs, quadriphthongs and even quintiphthongs for the abbreviation of phonetic components. It also includes some words and abbreviations in Latin."

Radboud University Nijmegen

W YOUR ROLL —

o, someone hasn't cracked the code o e mysterious Voynich manuscript

lieval scholar: "Sorry, folks, 'proto-Romance language' is not a thing."

FER OUELLETTE - 5/15/2019, 11:03 PM

https://arstechnica.com/science/2019/05/no-someone-hasnt-cracked-the-code-of-the-mysterious-voynich-manuscript/

Just like Latin to me!

Computers do this all the time!

"Ut utes an extinct language, ut alphabet is a description or ... Latin"

Latin

Wat is data?!

Byte = getal tussen 0 en 255

Wat is data?!

Byte = getal tussen 0 en 255
        of een getal tussen -128 en +127 ☺

Hoe zit dan dan met letters?

ASCII Code Chart

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL

UTF-8

The king of encodings

Variable length

ASCII characters are still one byte

48	69	e2	84	99	c6	b4	e2	98	82	e2	84	8c	c3	b8	e1	bc	a4
H	i	ℙ			ɥ		☂			ℌ			ø		ἤ		

s/knowledge/data/g;

Waarom BIG data?
Well… there's NO data like MORE data!

(Banko and Brill, ACL 2001): Large amounts of data

(Brants et al., EMNLP 2007): Variety of data

Radboud University Nijmegen
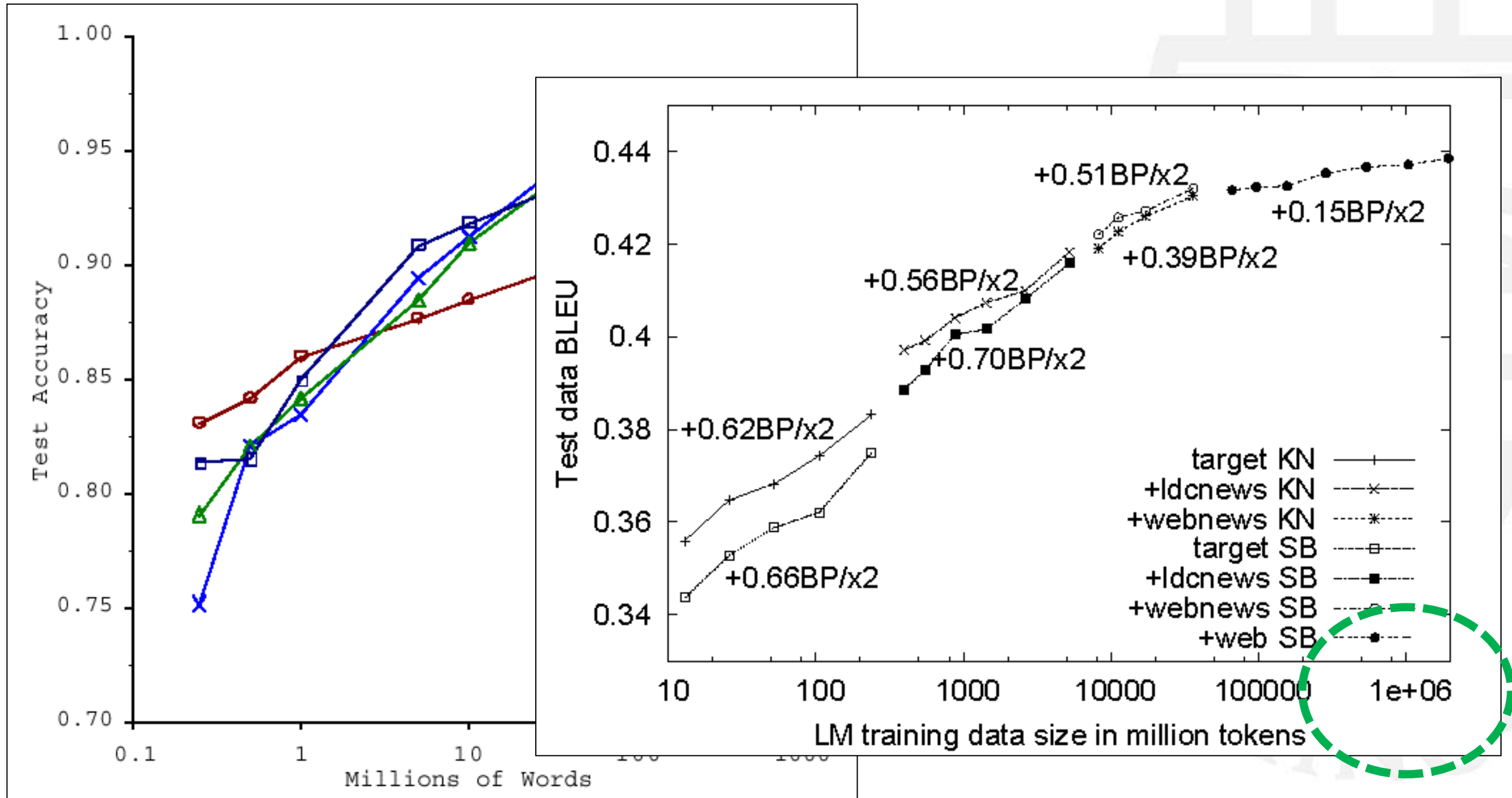
Use Google Data?!

- E.g., Google releases *pre-trained word and phrase vectors:*

*"We are publishing pre-trained vectors trained on part of Google News dataset (**about 100 billion words**). The model contains 300-dimensional vectors for 3 million words and phrases. The phrases were obtained using a simple data-driven approach described in [2]. The archive is available here: GoogleNews-vectors-negative300.bin.gz."*

Hier komen we op terug in de case!

Machine Learning & Search

"Computational Relevance"

"Intellectually it is possible for a human to establish the relevance of a document to a query. For a computer to do this we need to construct **a model within which relevance decisions can be quantified**. It is interesting to note that most research in information retrieval can be shown to have been concerned with different aspects of such a model."

Van Rijsbergen, 1976

Retrieval Model

Herken eigennamen

Ordenen zoekresultaten

Voorspel wat jij wilt vinden

Suggesties andere vragen

Spelfouten corrigeren

Wat zou de gebruiker willen?

Orden de "verticals"

Suggesties zoekvragen

Search & Privacy

AOL Search Dataset

- August 4, 2006: Logs voo[...]
  - 3 maanden, 650 duizend[...]
  - Anonieme User IDs

AnonID	Query	[...]
----------	---------	---
1234567	uw cse	2
1234567	uw admissions process	2
1234567	computer science hci	2
1234567	computer science hci	2
1234567	seattle restaurants	2
1234567	perlman montreal	2
1234567	uw admissions notification	2
…		

  - Thelma Arnold bevestigt[...]
- August 21, 2006: 2 AOL [...]
- September, 2006: "Class [...]

A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

☒ SIGN IN TO E-MAIL THIS
🖶 PRINT
▤ SINGLE PAGE
🗐 REPRINTS

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.

127 HOURS

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on everything."

And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for "landscapers in Lilburn, Ga," several people with the last name Arnold and "homes sold in shadow lake subdivision gwinnett county georgia."

It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga., frequently researches her friends' medical ailments and loves her three dogs. "Those are my searches,"

Erik S. Lesser for The New York Times

Thelma Arnold's identity was betrayed by AOL records of her Web searches, like ones for her dog, Dudley, who clearly has a problem.

AOL Search Dataset

- Anonieme IDs geen garantie voor anonimiteit
  - Logs bevatten direct identificerende informatie:
    - Namen, telefoonnummers, credit cards, BSNs
  - Tevens indirect identificerende informatie:
    - Thelma's vragen uit het NYT artikel
    - Geboortedatum, geslacht en postcode is voldoende om 87% van de Amerikanen uniek te identificeren!

Bias & Machine Learning

Computers Struggling with Text

- Microsoft Tay:
  chatbotslife.com/the-accountability-of-ai-case-study-microsofts-tay-experiment-ad577015181f

- Amazon's AI recruiting tool:
  reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

- **Any remedy for bias starts with awareness of its existence.**

- **Bias on the Web reflects biases within ourselves, manifested in subtler ways.**

- **We must consider and account for bias in the design of Web-based systems that truly address the needs of users.**

OUR INHERENT HUMAN tendency of favoring one thing or opinion over another is reflected in every aspect of our lives, creating both latent and overt biases toward everything we see, hear, and do. Any remedy for bias must start with awareness that bias exists; for example, most mature societies raise awareness of social bias through affirmative-action programs, and, while awareness alone does not completely alleviate the problem, it helps guide us toward a solution. Bias on the Web reflects both societal and internal biases within ourselves, emerging in subtler ways. This article aims to increase awareness of the potential effects imposed on us all through bias present in Web use and content. We must thus consider and account for it in the design of Web systems that truly address people's needs.

Bias has been intrinsically embedded in culture and history since the beginning of time. However, due to

[...] can now [...] and reach [...] as caused [...] a trending [...] cent years. [...] e felt the [...] when pur- [...] omes gov- [...] ms, from [...] ng person- [...] acles they [...] roadblock, [...] uch of the [...] s or how it [...] our judg-

[...] prominent [...] as well as [...] merge. As [...] central to [...] nfluencers we might not have encountered previously. This makes understanding and recognizing bias on the Web more essential than ever. My main goal here is thus to raise the awareness level for all Web biases. Bias awareness would help us design better Web-based systems, as well as software systems in general.

Measuring Bias
The first challenge in addressing bias is how to define and measure it. From a statistical point of view, bias is a systemic deviation caused by an inaccurate estimation or sampling process. As a result, the distribution of a variable could be biased with respect to the original, possibly unknown, distribution. In addition, cultural biases can be found in our inclinations to our shared personal beliefs, while cognitive biases affect our behavior and the ways we make decisions.

Figure 1 shows how bias influences

Bias on the Web – Highlights

- Activity Bias
  - In many collections, a few contributors create half the data:
    - Two Twitter datasets:
      0.05% of the most popular people attracted almost 50% of the participants in the 1st and 2% of the users generated 50% of the tweets in the 2nd dataset
    - Facebook dataset:
      7% of active users produced 50% of the posts
    - Amazon reviews dataset:
      4% of active users produced 50% of reviews *(fraud?)*
    - Wikipedia: 2000 people (0.04% of registered editors) wrote half of the entries of English Wikipedia

  - A "Digital Desert" of Web content no-one ever sees

Bias on the Web – Highlights

- Data Bias
  - 50% of popular web sites are in English…
    … but percentage of native English speakers is only 5%,
    all English speakers 13% (Disclaimer: stated source is Wikipedia.)

  - *Own example:*
    POI recommendation using geo-tagged Flickr photos:
    everyone visiting Paris visits the Eiffel Tower…

  - Consequence: optimizing for a data driven performance
    measure like accuracy forces the approach to recommend the
    Eiffel Tower *even if this is a useless recommendation!*

Radboud University Nijmegen

Bias on the Web – Highlights

- UI/UX influence on collected data:
  - Presentation Bias
    - *You can only click on what you see!*
  - Position Bias
    - *You only see where you look!*

- **Click distributions only useful when de-biased!**

- Second-order bias originating in personalization:
biasing content to our own pre-existing selection biases?

Question:

- **How to quantify the ranking bias in search engines?**

Retrievability

- Measures bias in the access of documents in a collection

L. Azzopardi and V. Vinay. Retrievability: An evaluation measure for higher order information access tasks. In Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM '08, pages 561–570, New York, NY, USA, 2008. ACM.

Radboud University Nijmegen

Retrievability

- Measure the accessibility of all documents in a collection given a set of queries
  - Retrievability score *r(d)* measures how often document *d* is retrieved by a given set of queries

Retrievability

- Measure the accessibility of all documents in a collection given a set of queries
  - Retrievability score $r(d)$ measures how often document d is retrieved by a given set of queries

- "Evaluate" a retrieval method by the distribution of retrievability scores generated:
  - Lorenz curve visualizes the inequality
  - Gini coefficient quantifies the inequality

Lorenz curve for n=5
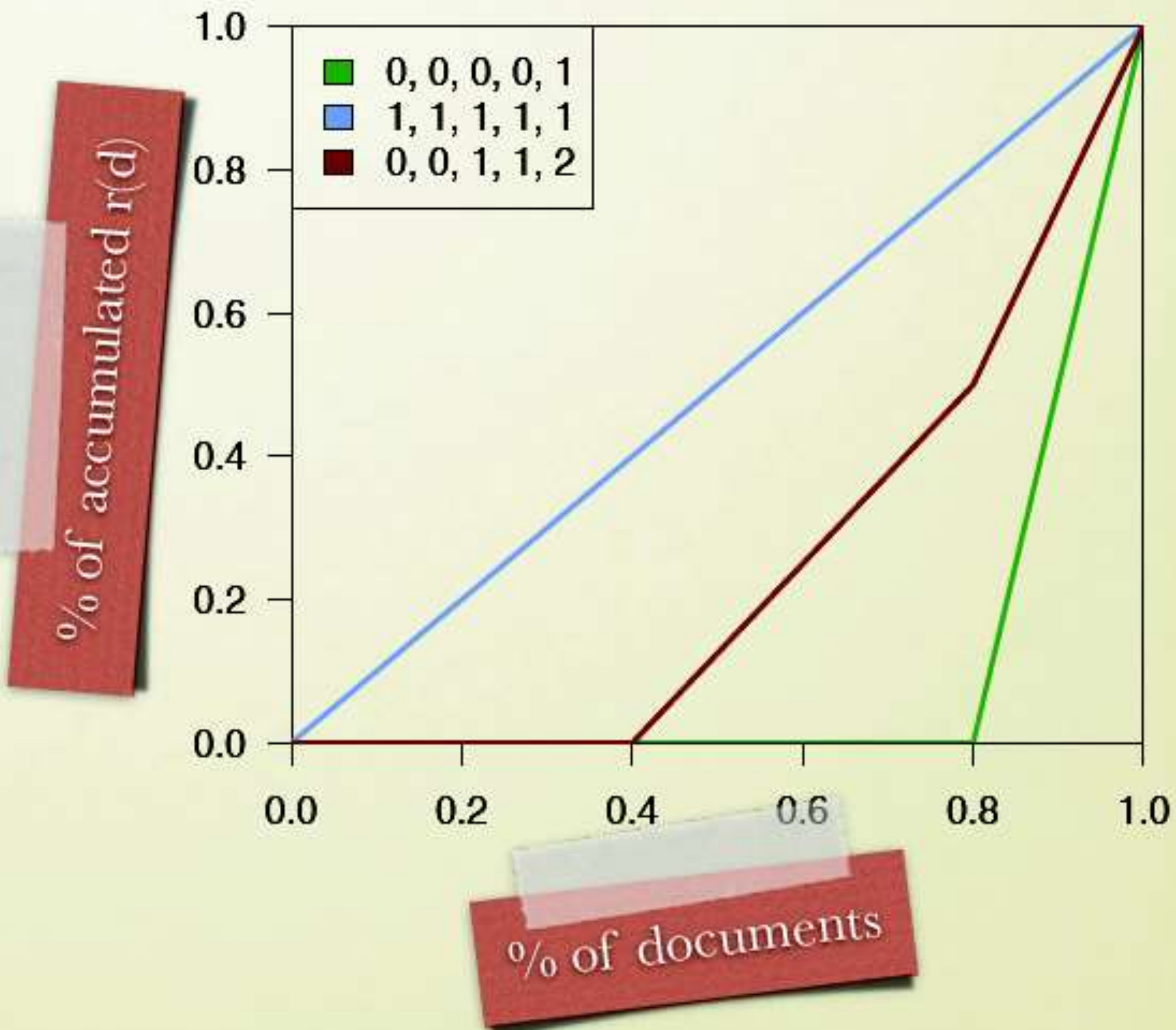
Relationship bias and effectiveness?

- Prior work indicates that high effectiveness (measured using P@10 and MAP) correlates with low "retrievability bias"

Azzopardi, L.,Bache, R.:On the relationship between effectiveness and accessibility. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 889–890. ACM (2010)

Retrievability & Ranked Retrieval

- Every document is always retrieved!

Retrievability & Ranked Retrieval

- Every document is always retrieved!

- Consider a cut-off, that corresponds intuitively to the amount of effort we expect a user to be willing to invest in learning the answer to their query
  - E.g., cut-off of 10:
    retrievability score equals the number of times this document is retrieved in the top 10 of the results for each of the queries in the (given) query set

"... by a given set of queries"

- Q: How do we get the set of queries for which to measure retrievability?

Note:
Choice of query set matters, see:
Bashir, S., Rauber, A.: Analyzing document retrievability in patent retrieval settings. In: International Conference on Database and Expert Systems Applications, pp.753–760. Springer (2009)

Radboud University Nijmegen

Simulation

- **Original study:**
  - Draw at random a large set of single terms and bigrams (from the documents in the collection)
    - Inspired by "Query Based Sampling" for resource description in non-cooperative federated search

- **Applied in our study on the Dutch Web Archive:**
  - 2M most frequent terms (frequencies ranging from 5 to 200M)
  - 2M most frequent bigrams (frequencies ranging from 20 to 35M)

Alternative "Query" Set

- Our study, on the Dutch Web Archive:
  - Anchor Text from external links
  - De-duplicated for year of crawl
    (Most sites crawled once a year, but a subset more frequently.)

- Side note:
  searsia.org/blog/2017-03-18-query-suggestions-without-tracking-users/

New iCIS professor!

Retrievability Score Inequality ("All")

Lorenz curve

$Q_s, c = 10$
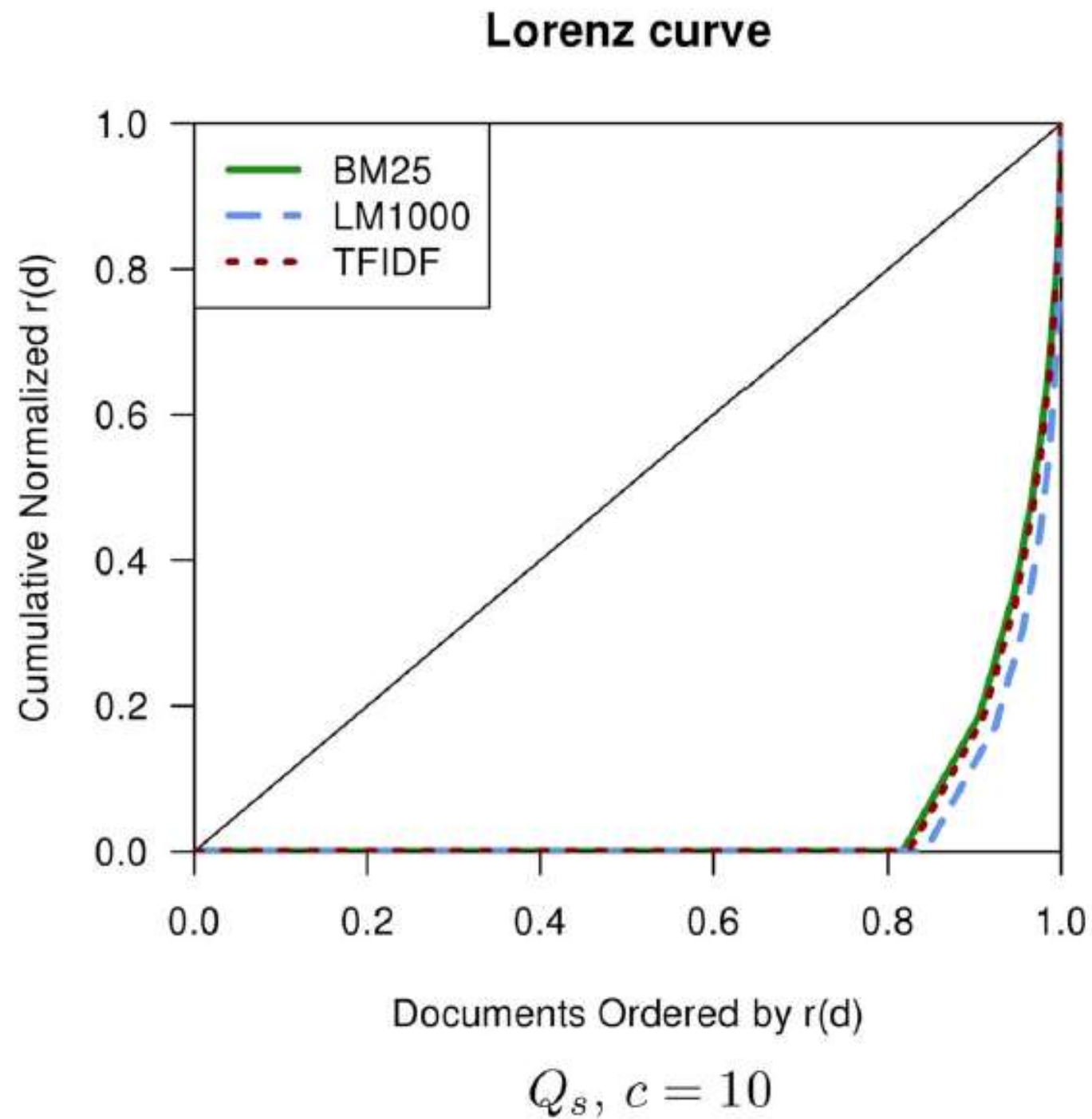
Zero Retrievability

- Original approach:
  - If a document is never retrieved by any query, $r(d) = 0$

- Alternative ("union"):
  - Consider only the documents that are retrieved by at least one of the methods under consideration
    - Fair comparison across retrieval methods for a given cut-off
    - Reduces the impact of the high number of documents that has a retrievability score of 0

Retrievability Score Inequality ("Union")

Lorenz curve

$Q_s, c = 10$
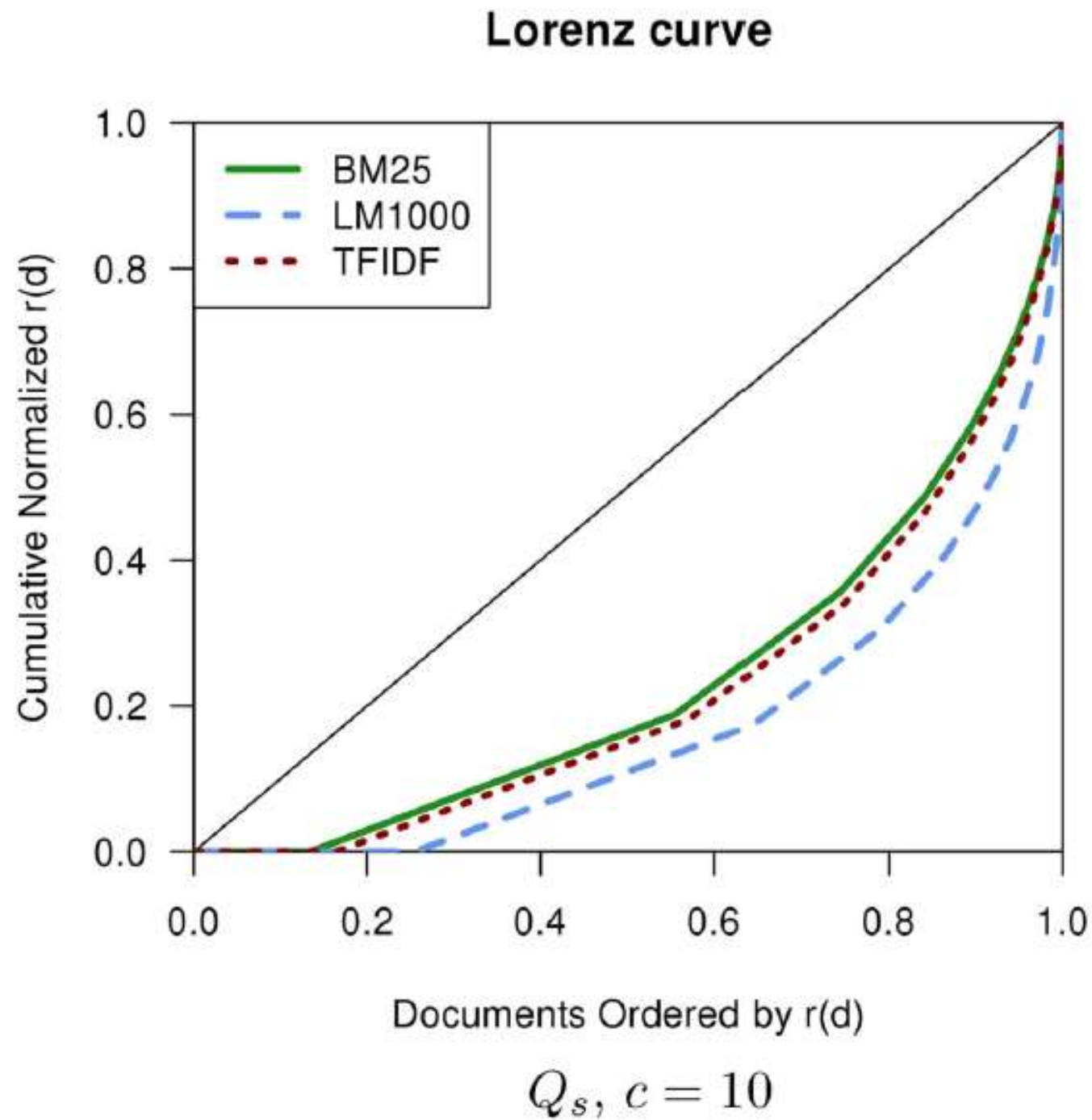
Retrievability = Findability?

- Divide collection in 4 equi-sized bins
  - Using wealth distribution (area under the curve)
- Can a document be found if the user queries the collection using a query sampled from that document?
  - *Need to take care to create a reasonable artificial known-item query…*
- Documents that are "highly retrievable" are significantly easier to find (using MRR to evaluate effectiveness)

Conclusions

- Analysis of retrievability helps understand the behaviour of retrieval systems, and makes explicit the inherent biases that affect the retrieved results

- Knowing which documents are particularly hard to find allows the institutions to improve their retrieval systems and the users to adapt their search strategies and be aware of the retrieval bias and the source of that bias

Quantifying retrieval bias in Web archive search

Thaer Samar, Myriam C. Traub, Jacco van Ossenbruggen, Lynda Hardman & Arjen P. de Vries

Quantifying Bias is not Enough!

- Retrievability compares ranking algorithms, so it can quantify bias – we can choose the least biased algorithm, but cannot reduce the bias observed

- Recent work by Singh and Joachims *corrects* bias in the exposure of items (due to presentation bias) associated to user groups; to ensure a *fair* exposure of those groups
  - https://arxiv.org/abs/1802.07281

Juridische Context

Data Protection by Design (art. 25) eist dat het risico op inbreuken op fundamentele rechten en vrijheden zoveel mogelijk wordt gemitigeerd (bv., onrechtmatige discriminatie op grond van etniciteit of gender)

Case:
informagus.nl/edu/cpo/

Juridische Context

Data Protection Impact Assessment (art. 35) eist dat bij een hoog risico op inbreuken op fundamentele rechten en vrijheden een uitgebreid impact assessment wordt gemaakt, waarbij mitigerende maatregelen zoals DPbD worden genomen zodat het risico afdoende wordt ingeperkt

Open Problems

Question (without an Answer)

- **How to determine the impact of the bias inherent in any sample of the Web that you are exposed to** ("you" being a person or an algorithm)**?**

Legal Context

GDPR recital 71: the data controller **should take measures** to "[prevent] discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation ..."

Discussion I: Awareness

- Bias on the Dutch Web / Social Media
    - Case studies on CommonCrawl and/or KB Web Archive

- Methods and techniques
    - Explore and extend current / develop new techniques
    - Tools for the average Web user

- My research:
  Personal Web Archives, to help you understand and reflect on biases you are exposed to

Radboud University Nijmegen

Figure 2: Search interface for WASP: ① shortcuts for usual time settings; ② selected query time interval; ③ date and time picker; ④ query box; ⑤ description of current result page; ⑥ title of result with links to archived and live version; ⑦ URL of the result; ⑧ archive time of the result; ⑨ snippet for the result.

Figure 3: Screenshot of a web page reproduced from the archive. A small black banner is inserted at the bottom right of the browser viewport to remind users that they are viewing an archived page.

Discussion II: Exploitation vs. Exploration

- Measures that do not discourage exploration
  - Countering the Popularity Bias

- *Can we somehow quantify future rewards, not just current rewards (?)*

- Best Paper SIGIR 2018:
  - Should I Follow the Crowd? A Probabilistic Analysis of the Effectiveness of Popularity in Recommender Systems
    ir.ii.uam.es/pubs/sigir2018.pdf
  - *Properties of measures informing better experimental designs*

Discussion III: Explanations

- Insight in how classifiers etc. decide – not trivial for tools like deep networks over, say, character sequences

limsse: That should have been a flag for me for when I came in . Service was excellent and food very good . This was a disaster , and Penn Avenue Fish Company was responsible . Very clean . Did n't see that coming . Always a pleasure . I really felt like I had wasted $ 25 . Will always go back Check them out on Beer Advocate . Friendly staff .

deeplift: That should have been a flag for me for when I came in . Service was excellent and food very good . This was a disaster , and Penn Avenue Fish Company was responsible . Very clean . Did n't see that coming . Always a pleasure . I really felt like I had wasted $ 25 . Will always go back Check them out on Beer Advocate . Friendly staff .

- Challenge:
  - How to create a Ground Truth?

  When is an explanation of an algorithmic decision a good explanation?

Discussion III: Explanations

- Example from https://arxiv.org/abs/1801.06422

| lrp | From : *kolstad* @ cae.wisc.edu (Joel Kolstad) Subject : Re : Can Radio Freq . Be Used To Measure Distance ? [...] What is the difference between vertical and horizontal ? Gravity ? Does n't gravity pull down the photons and cause a *doppler* shift or something ? (Just kidding !) |

belongs to the newsgroup `sci.electronic`

*"More importantly, **a good explanation method should not reflect what humans attend to, but what task methods attend to**. For instance, the family name "Kolstad" has 11 out of its 13 appearances in the 20 newsgroups corpus in* `sci.electronic` *posts. Thus, task methods probably learn it as a* `sci.electronics` *indicator. Indeed, the explanation method in Fig 1 (top) marks "Kolstad" as relevant, but the human annotator does not."*

Discussion IV: De-biasing data?

- Compare measures on social media or web data to those from more traditional data (e.g., demo-graphics, CBS)

- **Integrate observations from different sources;**
  e.g., use population density as well as geo-located Tweet frequency to counter geographical bias

- **De-bias representations derived from biased data (?)**
  - E.g., word embeddings (e.g. Bolukbasi *et al. "Man is to computer as woman is to homemaker?"*)

Note: discrimination-aware data mining is far from trivial:
https://people.cs.kuleuven.be/~bettina.berendt/Talks/berendt_2017_11_24_final.pdf

Radboud University Nijmegen

Other research themes to explore:

- Transparency and accountability of algorithms:
    - **Awareness**
    - Access and redress
    - Accountability
    - **Explanation**
    - Data provenance
    - Auditability
    - **Validation and testing**

- fairness-measures.org/

ACM U.S. Public Policy Council. *Statement on Algorithmic Transparency and Accountability*, ACM, Washington, D.C., Jan. 2017;
https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf