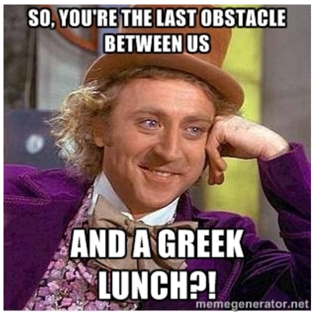
Axiometrics: An Axiomatic Approach to Evaluation Metrics

Stefano Mizzaro
Udine University, Italy
mizzaro@uniud.it

2015 essir thessaloniki.greece

Let's be clear from the start: I. Won't. Go. Overtime.



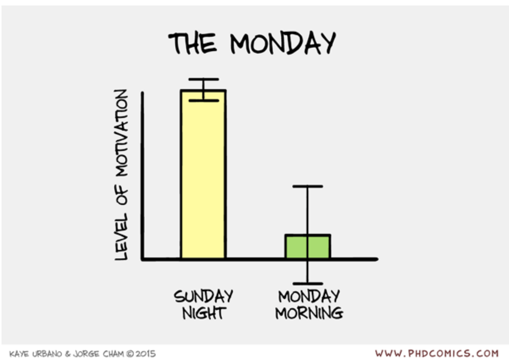
Aims – 1

- To present both:
  - basic material (what you find in books)
  - advanced material (recently published, even not yet published!)
- Links with Julio, Enrique, Evangelos
  - not (yet!) fully integrated
  - a bit disorganized…

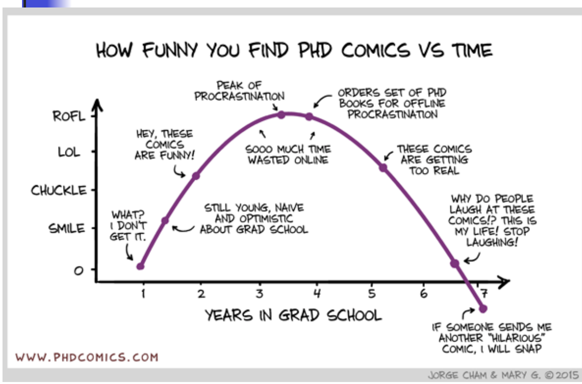
Aims – 2

- Intro to IR Evaluation
- Intro to IR Evaluation Measures / Metrics
- The Link between Measurement Theory and Metrics
  - Intro to Measurement Theory (Scales)
  - Metrics Analysis
  - The Axiometrics Framework (*)

The Monday effect…



You don't find it funny?!



Outline

- Evaluation [5′]
- Measures / metrics [15′]
- Measurement theory [15′]
- Metrics analysis [5′]
- Axiometrics framework [15′]

Outline

- Evaluation [5′]
- Measures / metrics [15′]
- Measurement theory [15′]
- Metrics analysis [5′]
- Axiometrics framework [15′]

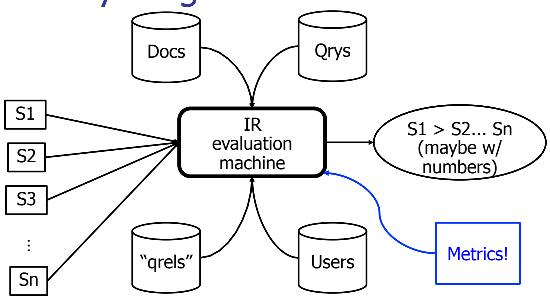
What is evaluation (in IR)?

- Eh…
- Ideally: a machine telling you how good an IR system is
- "Good": effective, capable to retrieve relevant (useful?!) documents
  - (efficiency is also studied, but focus is on effectiveness)

The importance of evaluation in IR

- Everybody agrees that evaluation is of paramount importance in IR
- One of the most evaluation-oriented disciplines in computer/information sciences
- We're busy doing a lot of evaluation since the 60s
  - So this talk is relevant. I do not know if it is useful :-)

A short history of nearly everything about IR evaluation



The importance of evaluation in IR

- Everybody agrees that evaluation is of paramount importance in IR
- One of the most evaluation-oriented disciplines in computer/information sciences
- We're busy doing a lot of evaluation since the 60s
- And we don't know (agree on) how to evaluate

(Other) Issues in IR evaluation

- Relevance
  - "Topicality"?
  - "Utility"?
  - …
- Methodology
  - Test collection, benchmark ,TREC-like
  - User study (--> Diane)
  - Large log analysis
  - …

Outline

- Evaluation [5']
- Measures / metrics [15']
- Measurement theory [15']
- Metrics analysis [5']
- Axiometrics framework [15']

We go down the dark evaluation metrics rabbit hole

IR effectiveness metric

- (or, measure)
- ""A number telling us how effective an IR system is""
- Simple, isn't it?

Simple, isn't it?

1. Precision
2. Recall
3. Swets's E measure
4. Fallout
5. Normalized Recall
6. Normalized Precision
7. P/R curve
8. Fallout/R curve
9. Brookes's S measure
10. Expected Search Length
11. ESL Reduction Factor
12. Precision Averaged over Recall Points
13. Sliding Ratio
14. Coverage Ratio
15. Novelty Ratio
16. Relative Recall
17. E-measure
18. F-measure
19. Utility
20. Average Precision
21. Mean Average Precision
22. Interpolated MAP
23. Precision at rank r
24. R-Precision
25. Generality
26. Miss
27. Shaw's D
28. PRecall
29. Satisfaction
30. Frustration
31. Total
32. Usefulness measure
33. Hull's Averaged Precision
34. Average Search Length
35. Quality based on ASL
36. Normalized Distance-Based Performance Measure
37. Recall Effort
38. Relative Relevance
39. Ranked Half Life
40. Reciprocal Rank
41. Mean Reciprocal Rank
42. Instance Precision
43. Instance Recall
44. Classification Accuracy
45. Discounted Cumulative Gain
46. Normalized DCG
47. NDCG at rank r
48. Average Weighted Precision
49. Weighted R-Precision
50. Average Distance Measure
51. PRecall adopted to XML
52. Precision and Recall with overlap
53. \no
54. area
55. Success at rank r
56. Average Gain Ratio
57. R-Gain Ratio
58. Binary Preference
59. Bpref-10
60. Q-measure
61. R-measure
62. Tolerance to Irrelevance
63. Estimated Ratio of Relevant
64. Egghe's measure
65. Passage Recall
66. Passage Precision
67. Passage bpref
68. Kendall, Spearman
69. Normalized xCG
70. Mean Average nxCG
71. Effort-Precision/Gain-Recall
72. Mean Average Effort-Precision
73. Precision-Recall with User Modelling
74. Geometric MAP
75. Generalized AP
76. Inferred AP
77. Expected Precision-Recall with User Modelling
78. Rpref
79. Generalized Success@10
80. k-call
81. Highlighting XML Evaluation
82. Ranking Efficiency
83. Rank-Biased Precision
84. Average Search Length Transformed
85. Logarithm of Found Position
86. Normalized Search Length
87. Polynomial Natural Order Imposition
88. Constant Penalty Natural Order Imposition

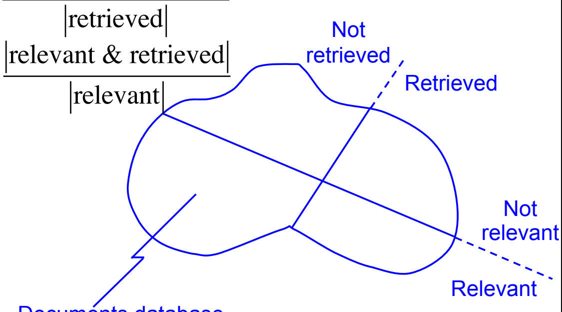
A shorter list

- Precision, Recall
- Precision-Recall curve
- MAP (Mean Average Precision)
- P@n (Precision at n)
- NDCG (Normalized Discounted Cumulative Gain)
- MRR (Mean Reciprocal Rank)
- RBP (Rank Biased Precision)
- TBG (Time Based Gain)

Precision & Recall

$$P = \frac{|\text{relevant \& retrieved}|}{|\text{retrieved}|}$$

$$R = \frac{|\text{relevant \& retrieved}|}{|\text{relevant}|}$$



[Salton & McGill, 84]

P & R: probabilistic definition

- $P = \text{p(relevant | retrieved)}$
- $R = \text{p(retrieved | relevant)}$

Ranking!

- But today all IR systems **rank** the documents
- Limitations of P&R
  - 2 numbers, not just one
  - Not affected by the rank of retrieved docs.
- Solutions: (too?) many.
  - Precision/Recall curve
  - MAP (Mean Average Precision)
  - …
- Let us see some examples

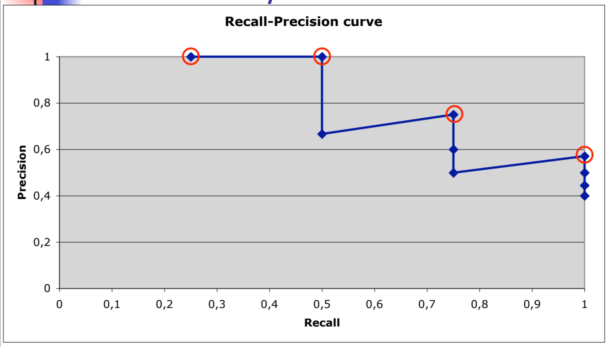
Rank

Rank	Rel?	R	P
1	1	0,25	1
2	1	0,5	1
3	0	0,5	0,67
4	1	0,75	0,75
5	0	0,75	0,6
6	0	0,75	0,5
7	1	1	0,57
8	0	1	0,5
9	0	1	0,44
10	0	1	0,4

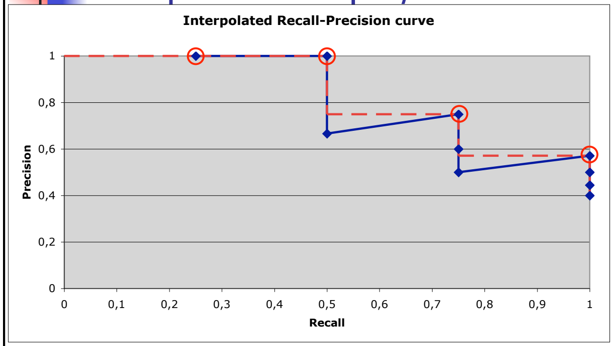
Rank

Rank	Rel?	R	P
1	1	**0,25**	**1**
2	1	**0,5**	**1**
3	0	0,5	0,67
4	1	**0,75**	**0,75**
5	0	0,75	0,6
6	0	0,75	0,5
7	1	**1**	**0,57**
8	0	1	0,5
9	0	1	0,44
10	0	1	0,4

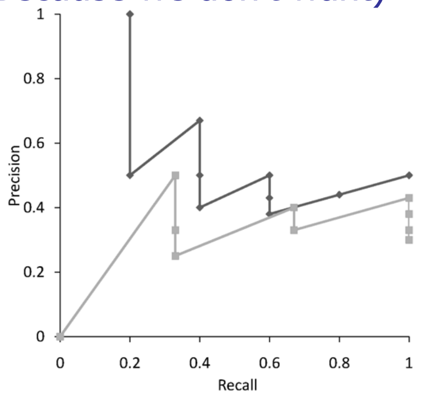
Sawtooth P/R curve



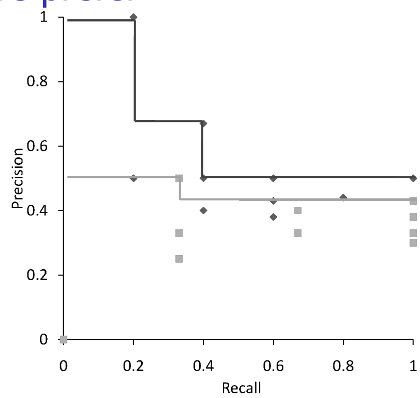
Interpolated step P/R curve



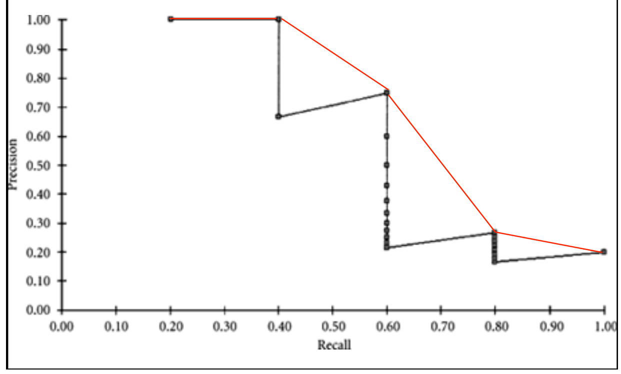
(Because we don't want)



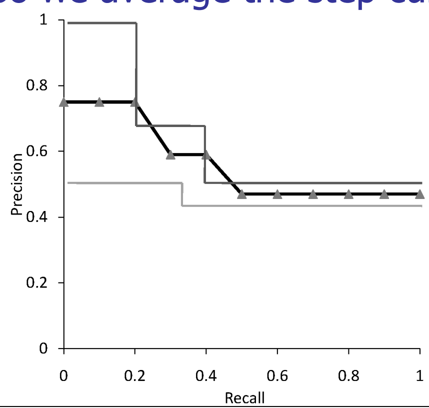
We prefer



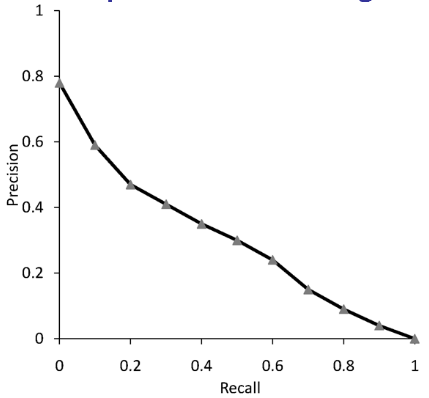
(We could do but we don't do)



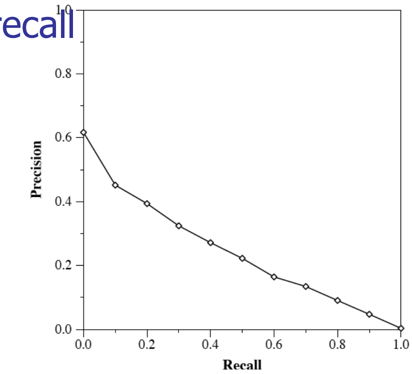
So we average the step curves



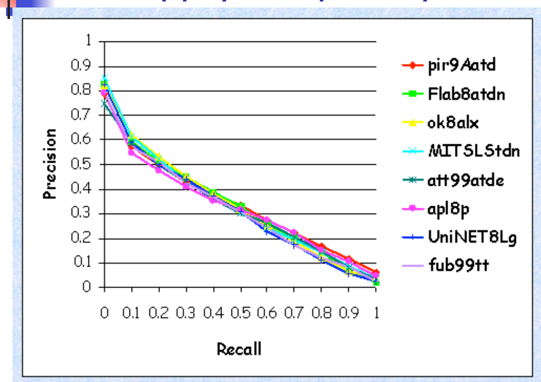
Over N queries and we get



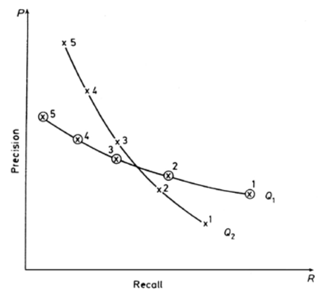
And of course on 11 levels of recall



We happily compare systems ?



Although often…



P/R curve --> MAP

- P/R curve
  - It is not a number
  - It can be transformed into a number by measuring **the area below the curve**
- --> **AP (Average Precision)**
- --> **MAP (Mean Average Precision)**
- Good property: top-heavyness

Rank	1	2	3	4	5	6	7	8	…
Rel	1	0	1	0	0	0	1	0	…

- User model?

P@n

- Simply count how many relevant documents are retrieved in the first n positions of the rank
- P@10 useful for classical Web search engines
- P@1 for "Feeling lucky"

Non binary relevance

- Some documents are "more relevant" than others
- Discounted Cumulative Gain (DCG, NDCG)
  - Different relevance --> different gain for the user
    - E.g., H --> 3, R --> 2, P --> 1, N --> 0
  - Sum of the gains while walking down the rank
  - Discounting more and more: late rank positions give less gain even if of equal relevance ("top-heaviness")

Example

Rank	Rel
1	H
2	P
3	N
4	R
5	H
6	R
7	N
8	P
9	N
10	P

Example

Rank	Rel	Gain
1	**H**	**3**
2	**P**	**1**
3	**N**	**0**
4	**R**	**2**
5	**H**	**3**
6	**R**	**2**
7	**N**	**0**
8	**P**	**1**
9	**N**	**0**
10	**P**	**1**

Example

Rank	Rel	Gain	CG
1	H	**3**	**3**
2	P	**1**	**4**
3	N	**0**	**4**
4	R	**2**	**6**
5	H	**3**	**9**
6	R	**2**	**11**
7	N	**0**	**11**
8	P	**1**	**12**
9	N	**0**	**12**
10	P	**1**	**13**

Example

Rank	Rel	Gain	CG	Discount
1	H	3	3	~~*log(1)*~~ **1**
2	P	1	4	**log(2)**
3	N	0	4	**log(3)**
4	R	2	6	**log(4)**
5	H	3	9	**log(5)**
6	R	2	11	**log(6)**
7	N	0	11	**log(7)**
8	P	1	12	**log(8)**
9	N	0	12	**log(9)**
10	P	1	13	**log(10)**

Example

Rank	Rel	Gain	CG	Discount	DG
1	H	**3**	3	~~*log(1)*~~ **1**	**3/1=3**
2	P	**1**	4	**log(2)**	**1/log(2)=1**
3	N	**0**	4	**log(3)**	**0/log(3)=0**
4	R	**2**	6	**log(4)**	**2/log(4)=1**
5	H	**3**	9	**log(5)**	**3/log(5)=1.3**
6	R	**2**	11	**log(6)**	**2/log(6)=.8**
7	N	**0**	11	**log(7)**	**0/log(7)=0**
8	P	**1**	12	**log(8)**	**1/log(8)=.3**
9	N	**0**	12	**log(9)**	**0/log(9)=0**
10	P	**1**	13	**log(10)**	**1/log(10)=.3**

Example

Rank	Rel	Gain	CG	Discount	DG	DCG
1	H	3	3	~~log(1)~~ 1	3/1=3	3.0
2	P	1	4	log(2)	1/log(2)=1	4.0
3	N	0	4	log(3)	0/log(3)=0	4.0
4	R	2	6	log(4)	2/log(4)=1	5.0
5	H	3	9	log(5)	3/log(5)=1.3	6.3
6	R	2	11	log(6)	2/log(6)=.8	7.1
7	N	0	11	log(7)	0/log(7)=0	7.1
8	P	1	12	log(8)	1/log(8)=.3	7.4
9	N	0	12	log(9)	0/log(9)=0	7.4
10	P	1	13	log(10)	1/log(10)=.3	7.7

Example

Rank	Rel	Gain	CG	Discount	DG	DCG	DCG Ideal
1	H	3	3	~~log(1)~~ 1	3/1=3	3.0	(H) 3.0
2	P	1	4	log(2)	1/log(2)=1	4.0	(H) 6.0
3	N	0	4	log(3)	0/log(3)=0	4.0	(R) 7.3
4	R	2	6	log(4)	2/log(4)=1	5.0	(R) 8.3
5	H	3	9	log(5)	3/log(5)=1.3	6.3	(P) 8.7
6	R	2	11	log(6)	2/log(6)=.8	7.1	(P) 9.1
7	N	0	11	log(7)	0/log(7)=0	7.1	(P) 9.4
8	P	1	12	log(8)	1/log(8)=.3	7.4	(N) 9.4
9	N	0	12	log(9)	0/log(9)=0	7.4	(N) 9.4
10	P	1	13	log(10)	1/log(10)=.3	7.7	(N) 9.4

Example

Rank	Rel	Gain	CG	Discount	DG	DCG	DCG Ideal	NDCG
1	H	3	3	~~log(1)~~ 1	3/1=3	3.0	(H) 3.0	1.00
2	P	1	4	log(2)	1/log(2)=1	4.0	(H) 6.0	0.67
3	N	0	4	log(3)	0/log(3)=0	4.0	(R) 7.3	0.55
4	R	2	6	log(4)	2/log(4)=1	5.0	(R) 8.3	0.61
5	H	3	9	log(5)	3/log(5)=1.3	6.3	(P) 8.7	0.72
6	R	2	11	log(6)	2/log(6)=.8	7.1	(P) 9.1	0.78
7	N	0	11	log(7)	0/log(7)=0	7.1	(P) 9.4	0.75
8	P	1	12	log(8)	1/log(8)=.3	7.4	(N) 9.4	0.78
9	N	0	12	log(9)	0/log(9)=0	7.4	(N) 9.4	0.78
10	P	1	13	log(10)	1/log(10)=.3	7.7	(N) 9.4	0.82

Example

Rank	Rel	Gain	CG	Discount	DG	DCG	DCG Ideal	NDCG
1	H	3	3	~~log(1)~~ 1	3/1=3	3.0	(H) 3.0	1.00
2	P	1	4	log(2)	1/log(2)=1	4.0	(H) 6.0	0.67
3	N	0	4	log(3)	0/log(3)=0	4.0	(R) 7.3	0.55
4	R	2	6	log(4)	2/log(4)=1	5.0	(R) 8.3	0.61
5	H	3	9	log(5)	3/log(5)=1.3	6.3	(P) 8.7	0.72
6	R	2	11	log(6)	2/log(6)=.8	7.1	(P) 9.1	0.78
7	N	0	11	log(7)	0/log(7)=0	7.1	(P) 9.4	0.75
8	P	1	12	log(8)	1/log(8)=.3	7.4	(N) 9.4	0.78
9	N	0	12	log(9)	0/log(9)=0	7.4	(N) 9.4	0.78
10	P	1	13	log(10)	1/log(10)=.3	7.7	(N) 9.4	0.82

A shorter list

- Precision, Recall
- Precision-Recall curve
- MAP
- P@n
- NDCG
- MRR (Mean Reciprocal Rank)
- RBP (Rank Biased Precision)
- TBG (Time Based Gain)

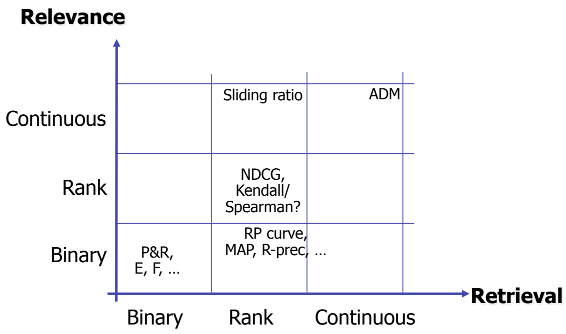
… even continuous relevance

- Dynamometer, Hand force grip, even physiological data
- IR System showing a "relevance bar" close to each document
  - "estimation of the amount of relevance"
- Magnitude estimation (paper @ last SIGIR)
- …

Metrics classification

- Metrics could be classified on the basis of underlying notions of:
  - relevance (binary, ranking, continuous)
  - retrieval (binary, ranking, continuous)
- 3 x 3 (or N x N) grid, …

Classification (incomplete!)



So, if you understand it, I'm not being clear enough :-)

--> Julio, Enrique, Evan…

Take home messages so far

- In IR, Evaluation is important.
- There are **many (100+)** metrics
  - System-oriented metrics only
    - (R, P, MAP, NDCG, …)
  - Let alone human/user-oriented metrics
    - (user satisfaction, fatigue, …)
  - And still IR only, no clustering, filtering…
- Just a few are indeed used
  - (which ones? Why those? --> Julio)
- Definitions of some of them

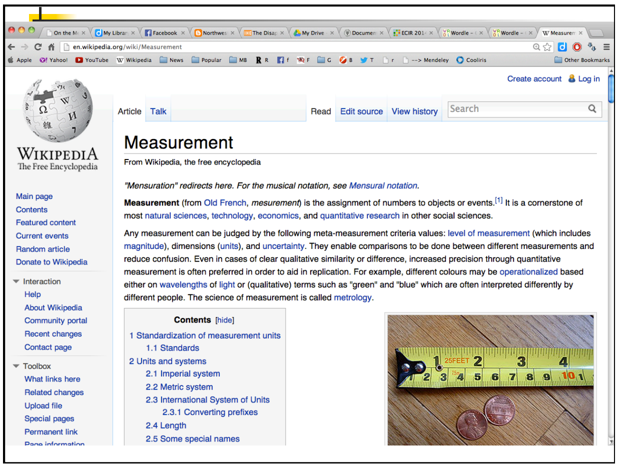
And now for something completely different…

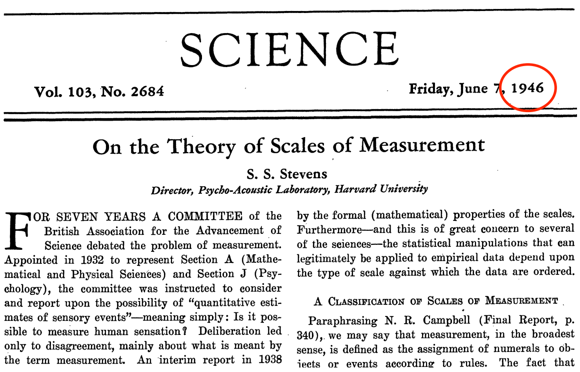
Outline

- Evaluation [5']
- Measures / metrics [15']
- Measurement theory [15']
- Metrics analysis [5']
- Axiometrics framework [15']

Measurement

- Definition: A process aimed at determining a relationship between a physical quantity and a unit of measurement
- Typically, one assigns numbers to objects/events
- Studied in Measurement Theory
  - Reasonably settled





SCIENCE

Vol. 103, No. 2684 — Friday, June 7, 1946

On the Theory of Scales of Measurement

S. S. Stevens

Director, Psycho-Acoustic Laboratory, Harvard University

FOR SEVEN YEARS A COMMITTEE of the British Association for the Advancement of Science debated the problem of measurement. Appointed in 1932 to represent Section A (Mathematical and Physical Sciences) and Section J (Psychology), the committee was instructed to consider and report upon the possibility of "quantitative estimates of sensory events"—meaning simply: Is it possible to measure human sensation? Deliberation led only to disagreement, mainly about what is meant by the term measurement. An interim report in 1938 by the formal (mathematical) properties of the scales. Furthermore—and this is of great concern to several of the sciences—the statistical manipulations that can legitimately be applied to empirical data depend upon the type of scale against which the data are ordered.

A Classification of Scales of Measurement

Paraphrasing N. R. Campbell (Final Report, p. 340), we may say that measurement, in the broadest sense, is defined as the assignment of numerals to objects or events according to rules. The fact that

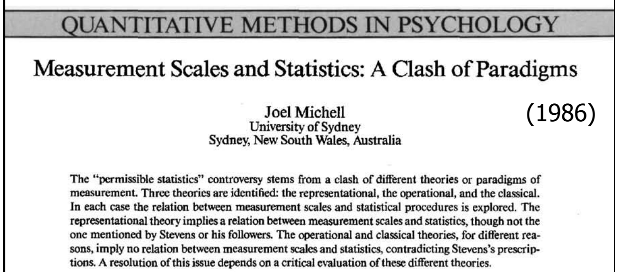
Measurement Theory

- One important concept: which numbers can I select when measuring? What properties do they have?
- Which **measurement scale**?
  - (or "**level**")
  - E.g., to measure length:
    - Meters
    - Inches
    - "Longer than" (?)

Measurement scales

- Standard set of scales:
  1. Nominal
  2. Ordinal
  3. Interval
  4. Ratio
- (other proposals exist)

Perhaps not so well settled…



QUANTITATIVE METHODS IN PSYCHOLOGY

Measurement Scales and Statistics: A Clash of Paradigms

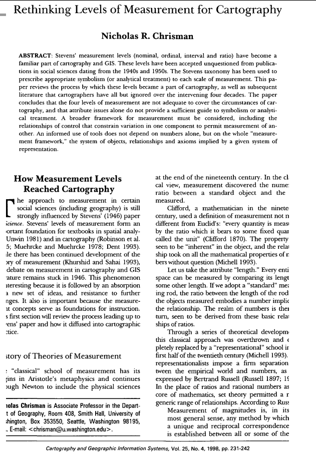
Joel Michell
University of Sydney
Sydney, New South Wales, Australia

(1986)

The "permissible statistics" controversy stems from a clash of different theories or paradigms of measurement. Three theories are identified: the representational, the operational, and the classical. In each case the relation between measurement scales and statistical procedures is explored. The representational theory implies a relation between measurement scales and statistics, though not the one mentioned by Stevens or his followers. The operational and classical theories, for different reasons, imply no relation between measurement scales and statistics, contradicting Stevens's prescriptions. A resolution of this issue depends on a critical evaluation of these different theories.

And indeed

- Nicholas Chrisman
- 1998
- Proposes 10 scales, not just 4



Rethinking Levels of Measurement for Cartography

Nicholas R. Chrisman

ABSTRACT: Stevens' measurement levels (nominal, ordinal, interval and ratio) have become a familiar part of cartography and GIS. These levels have been accepted unquestioned from publications in social sciences dating from the 1940s and 1950s. The Stevens taxonomy has been used to prescribe appropriate symbolism (or analytical treatment) to each scale of measurement. This paper reviews the process by which these levels became a part of cartography, as well as subsequent literature that cartographers have all but ignored over the intervening four decades. The paper concludes that the four levels of measurement are not adequate to cover the circumstances of cartography, and that attribute issues alone do not provide a sufficient guide to symbolism or analytical treatment. A broader framework for measurement must be considered, including the relationships of control that constrain variation in one component to permit measurement of another. An informed use of tools does not depend on numbers alone, but on the whole "measurement framework," the system of objects, relationships and axioms implied by a given system of representation.

How Measurement Levels Reached Cartography

Cartography and Geographic Information Systems, Vol. 25, No. 4, 1998, pp. 231-242

Anyway "Good Old Fashioned" Measurement scales

1. Nominal
2. Ordinal
3. Interval
4. Ratio

4. Ratio scale

- I'm twice taller than him
- He is twice richer than me
  - (Both "how much" & "how many")
- I'm twice older than you
  - Years, months (*12), days, …
- Zero
  - Age starts from zero!
- But not
  - Today is twice as hot as yesterday?

3. Interval scale

- Today is twice as hot as yesterday?
- In the last two days we had the same increase of 5ºC in temperature
  - The difference between today and yesterday temperature is the same as …
- Dates are another example
  - Difference: OK (2014 – 2012 = 2006 – 2004)
  - Ratio: KO (2000 is not 2 * 1000)
  - (ratios of differences: OK)

2. Ordinal scale

- A measure is not an amount but a **rank**
- It is a form of measurement!
- Ex: Lines ranked according to their length
- It does **not** mean that:
  - the first is twice as long than the second
  - the length difference between the 1st and the 2nd is the same as the 2nd and the 3rd

1. Nominal scale

- Qualitative
- Categories
- Names, gender, nationality, …
- Can be Dichotomous or Non-dichotomous
- No assumptions on ratios, distances, ranks.

Legit operations

- Given a scale, only some operations make sense
  - Arithmetic: +, -, *, /
  - Statistic: Mean, median, mode, …
- Ex:
  - Average height, weight, …: OK
  - Average gender: KO

Legit relational/math operations

	=, ≠	>, <	+, -	×, ÷
Nominal	✓	✗	✗	✗
Ordinal	✓	✓	✗	✗
Interval	✓	✓	✓	✗
Ratio	✓	✓	✓	✓

Permissible transformations

	a · x	a · x + b	Monotonic	1-to-1
Nominal	✓	✓	✓	✓
Ordinal	✓	✓	✓	✗
Interval	✓	✓	✗	✗
Ratio	✓	✗	✗	✗

Meaning

- If you transform the measure, are you still measuring the same thing?
  - Nationality
  - Rank
  - Temperature
  - Money

Examples

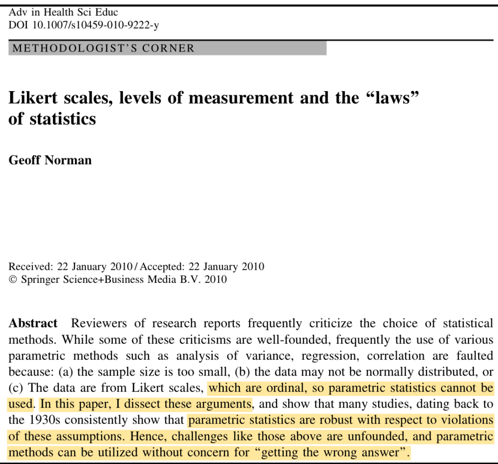
- Nominal scale for nationality
  - Greek= 1
  - Italian = 2
  - Spanish = 3
  - Japanese = 4
  - …
  - 3 – 1 = 4 – 2 Uh?!?
- Whereas interval scale for temperature ºC
  - 30º – 10º = 40º – 20º: ok

Legit statistics

	Mode	Median	Mean: Arithmetic	Mean: Geometric, Harmonic
Nominal	✓	✗	✗	✗
Ordinal	✓	✓	✗	✗
Interval	✓	✓	✓	✗
Ratio	✓	✓	✓	✓

Examples

- Nationality, mode, ok
- Mean rank? Uh?!
  - (think of ranking ten students…)
- Mean temperature, ok


Adv in Health Sci Educ
DOI 10.1007/s10459-010-9222-y

METHODOLOGIST'S CORNER

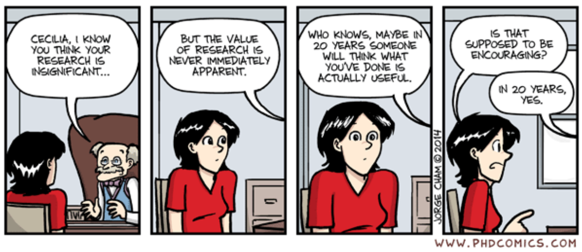
Likert scales, levels of measurement and the "laws" of statistics

Geoff Norman

Received: 22 January 2010 / Accepted: 22 January 2010
© Springer Science+Business Media B.V. 2010

Abstract Reviewers of research reports frequently criticize the choice of statistical methods. While some of these criticisms are well-founded, frequently the use of various parametric methods such as analysis of variance, regression, correlation are faulted because: (a) the sample size is too small, (b) the data may not be normally distributed, or (c) The data are from Likert scales, which are ordinal, so parametric statistics cannot be used. In this paper, I dissect these arguments, and show that many studies, dating back to the 1930s consistently show that parametric statistics are robust with respect to violations of these assumptions. Hence, challenges like those above are unfounded, and parametric methods can be utilized without concern for "getting the wrong answer".

Why are you telling me this?



Why are you telling me this?

- Two reasons
  - 1. Measurement theory and scales can be used to directly analyze IR metrics
  - 2. Because IR can be seen as a measurement. Of relevance --> Language to define axioms on metrics
- Let's see 1. first

Outline

- Evaluation [5′]
- Measures / metrics [15′]
- Measurement theory [15′]
- Metrics analysis [5′]
- Axiometrics framework [15′]

MRR (Mean Reciprocal Rank)

- Take the rank of the 1st relevant doc.
- Take the reciprocal (...)

$$RR = \frac{1}{rank(i)}$$

- Then take the mean (...) over some topics

$$MRR = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{rank(i_q)}$$

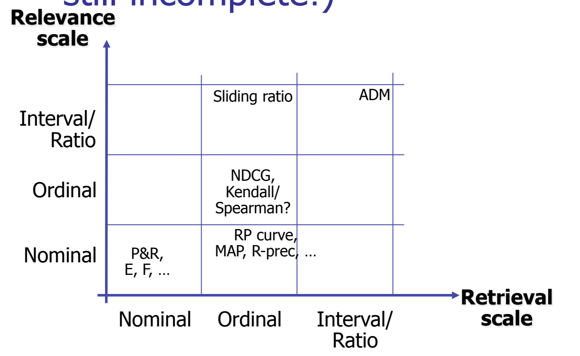
MRR?!

- It is used
- In many papers
- Even in some TREC tracks
- When analyzed with "measurement theory glasses", is is not "measurement theory proof"
- A **reciprocal** is taken on an **ordinal** scale...
- ... then it is **averaged**...

Even worse than that... NDCG

- H,R,P,N --> 3, 2, 1, 0 ?
  - Linear, most common
- H,R,P,N --> 100, 10, 1, 0 ?! (or 4, 2, 1, 0)
  - Exponential, sometimes used, actually
- H,R,P,N --> 100, 99, 90, 0 ?!?
  - "Crazy", never heard of... why?
- Arbitrary choice!
- By transforming relevance levels into gains, we transform an 2. Ordinal scale --> 4. Ratio scale!
- And we also discount
  - Dividing by log(rank)...
  - All metrics that can be modeled as gain/discount...

Classification (more principled, still incomplete!)



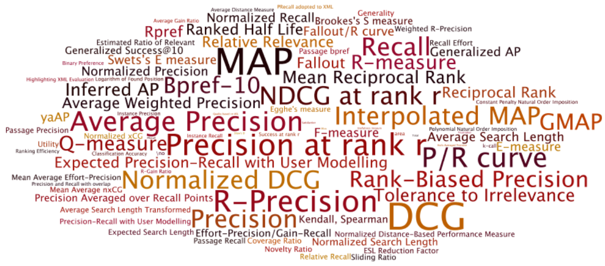
To summarize, but not to conclude

- 100+ metrics
- Measurement theory seems a useful tool
  - Metric classification
  - Some arbitrary choices
  - Some metrics are not "Measurement theory proof"
- Metric **engineering** seems more an **art** (artisan) than a **science**...
- A more principled approach?

Outline

- Evaluation [5′]
- Measures / metrics [15′]
- Measurement theory [15′]
- Metrics analysis [5′]
- Axiometrics framework [15′]

So, back to metrics…



Hm. One hundred.

- What do they have in common?
- Research question:
  Are there some axioms that any effectiveness metric should satisfy?
- "Axioms"? ("Axiometrics"! [Arjen])
  - Properties, constraints, Laws, …
- Maybe a not so crazy idea:
  - [van Rijsbergen 1979], [Bollman 1984], [Amigó, Gonzalo 2009] (& more), [Sebastiani 2015], SWIRL 2012, …

2. IR as measurement of relevance

- Both an IR system and a human assessor **measure** the **relevance** of a document
- Maybe on different **scales**:
  - Ranked SE output: ordinal scale
  - Human TREC-like relevance assessor: nominal (ordinal) scale (binary, relevant/nonrelevant)
  - … IR related tasks like categorization, filtering, …

So, relevance measurement(s)

- This is general
  - Change the scale and you get the corresponding relevance measurement
  - Allows to take into account binary, graded, ranked,… relevance and retrieval

Notation: σ and α

- 2 relevance measurements
  - by a system (σ)
    - $\sigma(q,d)$, $\sigma(q,D)$, $\sigma(Q,D)$
    - e.g. SE ranked output
  - by a human (user/assessor) (α)
    - $\alpha(q,d)$, $\alpha(q,D)$, $\alpha(Q,D)$
    - e.g., TREC qrels

Similarity

- Given two relevance measurements α and σ, we can define a notion of similarity between them (given a query/topic q and a document d)

$$\text{sim}_{q,d}(\alpha, \sigma)$$

- And an IR system should provide a measurement σ that is similar to the human measurement α

Now, similarity…

Psychological Review

Copyright © 1977 by the American Psychological Association, Inc.

VOLUME 84 NUMBER 4 JULY 1977

Features of Similarity

Amos Tversky
Hebrew University
Jerusalem, Israel

The metric and dimensional assumptions that underlie the geometric representation of similarity are questioned on both theoretical and empirical grounds. A new set-theoretical approach to similarity is developed in which objects are represented as collections of features, and similarity is described as a feature-matching process. Specifically, a set of qualitative assumptions is shown to

Similarity comparison

- And we can compare similarities
- For example,

$$\underset{q,d}{\text{sim}}(\alpha, \sigma) < \underset{q,d}{\text{sim}}(\alpha, \sigma')$$

- means that on the query q and the document d, and given the human relevance judgment α, system σ is worse than σ' (i.e., less similar to α)

Similarity --> Metric

- On the basis of the notion of similarity, we can define an IR effectiveness metric
- The more σ is similar to α, the higher the metric value

So, to summarize (but not to conclude!)

- **Measurement** theory, Measurement scales, IR as relevance measurement
  - $\sigma(q,d)$, $\sigma(q,D)$, $\sigma(Q,D)$, $\alpha(q,d)$, $\alpha(q,D)$, $\alpha(Q,D)$
- **Similarity**

$$\underset{q,d}{\text{sim}}(\alpha, \sigma) < \underset{q,d}{\text{sim}}(\alpha, \sigma')$$

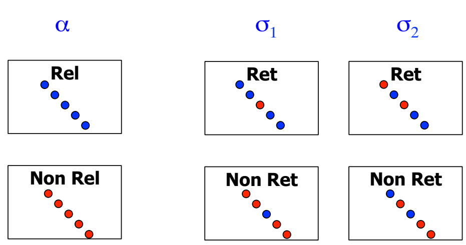
- **Metric**

$$\underset{Q,D}{\text{metric}}(\alpha, \sigma)$$

Generality: Across different measurement scales

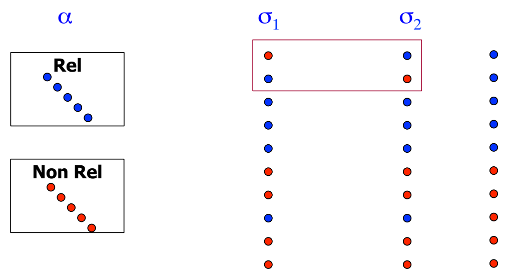
- We can compute the similarity of two relevance judgments when they are on the same scale
- …but more than that…
- … also when they are on different scales, in some cases
  - E.g., the classical ad-hoc retrieval
    - scale(σ) = [[ordinal]]
    - scale(α) = [[nominal]] (binary relevance. Ordinal)

Same scales: binary IR



sim(α, σ_1) > sim(α, σ_2)

Different scales: ad hoc IR



sim(α, σ_1) < sim(α, σ_2)

Ok, but we want Axioms!

Some details…

- I do not trust our axioms too much yet…
- … preliminary work…
- Actually:
  - I think axioms are correct and consistent
  - I don't know if they are complete
- Stating axioms is also useful to "test the framework"
  - Measurement theory is an effective language to state them!

A First Axiom

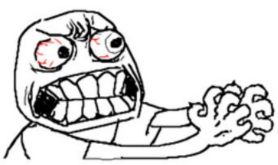
Axiom 3 (Similarity of two systems). *Let q be a query, d a document, α a human relevance measurement and σ and σ' two system relevance measurements such that*

$$\sigma(q,d) = \sigma'(q,d). \quad (1)$$

Then

$$\underset{q,d}{\text{sim}}(\alpha,\sigma) = \underset{q,d}{\text{sim}}(\alpha,\sigma'). \quad (2)$$

Come on you're kidding!



- All this and then such a stupid axiom????

Ok, ok. Second Axiom

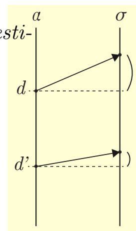
Axiom 6 (Overestimated documents). *Let q be a query, d and d' two document, α a human relevance measurement and σ a system relevance measurements such that*

$$\alpha(d) > \alpha(d'),$$

$$\underset{q,d}{\text{sim}}(\alpha,\sigma) < \underset{q,d'}{\text{sim}}(\alpha,\sigma)$$

and (6) and (8) hold (i.e., both d and d' are overestimated), then

$$\underset{q,d}{\text{metric}}(\alpha,\sigma) < \underset{q,d'}{\text{metric}}(\alpha,\sigma).$$



- (d is more relevant, sim on d is lower, then metric value on d has to be lower)
- d is both "more wrong" and "more visible" to the user

Ok, ok. Third Axiom

Axiom 8 (System relevance). *Let q be a query, d and d' two documents, α a human relevance measurement and σ a system relevance measurement such that* $\mathrm{sim}_{q,d}(\alpha,\sigma) = \mathrm{sim}_{q,d'}(\alpha,\sigma)$, $\sigma(d) > \sigma(d')$, *and*

$$\alpha(d) \geq \alpha(d'). \qquad (10)$$

Then

$$d \sqsupset_{metric(\alpha,\sigma)} d'.$$

(document d affects the metric value more than d')

Meaning?

- Corollary:
- By taking scale(σ) = [[Rank]] we derive that:
  - Early rank positions affect a metric value more than later rank positions
  - IR metrics should be "**top-heavy**"
- Previous Axiom 8 states a more abstract/ general principle, independent of the scales

Now, a last Axiom

Axiom 9 (User relevance). *Let q be a query, d and d' two documents, α a human relevance measurement and σ a system relevance measurement such that:* $\mathrm{sim}_{q,d}(\alpha,\sigma) = \mathrm{sim}_{q,d'}(\alpha,\sigma)$, $\alpha(d) > \alpha(d')$, *and*

$$\sigma(d) \geq \sigma(d'). \qquad (11)$$

Then

$$d \sqsupset_{metric(\alpha,\sigma)} d'.$$

(document d affects the metric value more than d')

Meaning?

- A metric should weigh more, and be more affected, by more relevant documents
- "α top heavyness", "human top heavyness"
  - Perhaps less intuitive than previous axiom,
  - but it does indeed seem natural in the framework
  - by symmetry (treat α as σ)
  - To evaluate a nonrelevant document as nonrelevant is an easy job (the vast majority of documents in a collection are nonrelevant)

Meaning?

- Consequence: linear gain values of 3, 2, 1, 0 in NDCG (for H, R, P, N) can be questioned
  - Exponential 100, 10, 1, 0 (or 4, 2, 1, 0) might be better
  - (already proposed in the original paper, but not much used…)
  - And "crazy" 100, 99, 90, 0 is wrong!

A theorem

Theorem 2 (Unbalanced query). *Let Q be a query set, $q \notin Q$ a query, D a document set, α a human relevance measurement and σ and σ' two system relevance measurements such that*

$$\underset{Q,D}{\mathrm{metric}}(\alpha,\sigma) > \underset{Q,D}{\mathrm{metric}}(\alpha,\sigma')$$

and

$$\underset{Q\cup\{q\},D}{\mathrm{metric}}(\alpha,\sigma) \leq \underset{Q\cup\{q\},D}{\mathrm{metric}}(\alpha,\sigma').$$

Then

$$\underset{q,D}{\mathrm{metric}}(\alpha,\sigma) < \underset{q,D}{\mathrm{metric}}(\alpha,\sigma').$$

A theorem

Theorem 2 (Unbalanced query). *Let Q be a query set, $q \notin Q$ a query, D a document set, α a human relevance measurement and σ and σ' two system relevance measurements such that*

$$\underset{Q,D}{\text{metric}}(\alpha,\sigma) > \underset{Q,D}{\text{metric}}(\alpha,\sigma')$$

on the query set Q and the document collection D, and given the human relevance judgment α, σ is more effective than σ'

and

$$\underset{Q\cup\{q\},D}{\text{metric}}(\alpha,\sigma) \le \underset{Q\cup\{q\},D}{\text{metric}}(\alpha,\sigma').$$

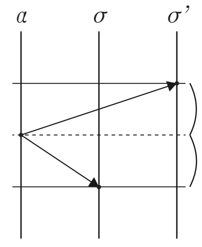
we add a new query q and then σ becomes less effective than σ'

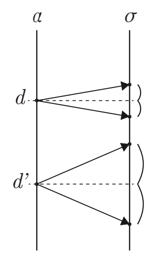
Then

$$\underset{q,D}{\text{metric}}(\alpha,\sigma) < \underset{q,D}{\text{metric}}(\alpha,\sigma').$$

σ is less effective than σ' also on the new query q

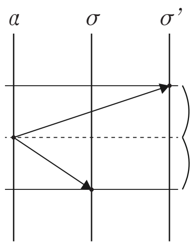
When we can't say anything (i.e., we can't state an axiom)



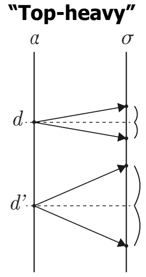


Why we can't say anything

"P-oriented"
"R-oriented"



"Top-heavy"



Hopefully you got the point

- By relying on **measurement theory**,
- one can define **relevance measurement,**
- that in turn allows to define **similarity** between (human and system) measurements,
- that in turn allows to define **Axioms** and **Theorems** on **metrics**
- that seem somehow interesting
  - not (always) trivial, more general, even inspiring …

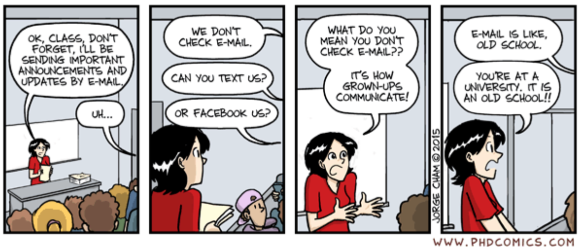
Outline + wrap-up

- Evaluation [5']
  - Yes, it's complex!
- Measures / metrics [15']
  - Oh dear, so many metrics?
- Measurement theory [15']
  - Maybe a useful tool…
- Metrics analysis [5']
  - Are we doing it wrong?!
- Axiometrics framework [15']
  - Attempt to shift focus from metric A vs. B to metric properties

Biblio

- [Amigó et al. 2009] A comparison of extrinsic clustering evaluation metrics based on formal constraints. IRJ 2009.
- [Bollman 1984] Two axioms for evaluation measures in information retrieval, SIGIR 1984.
- [Busin & Mizzaro 2013] Axiometrics: An Axiomatic Approach to Information Retrieval Effectiveness Metrics, ICTIR 2013.
- [Maddalena & Mizzaro 2014] Axiometrics: Axioms of Information Retrieval Effectiveness Metrics, EVIA@NTCIR 2014.
- [Sebastiani 2015] An Axiomatically Derived Measure for the Evaluation of Classification Algorithms, ICTIR 2015, to appear.
- [van Rijsbergen 1979] Information Retrieval, 1979.

mizzaro@uniud.it



(I’m here until Friday, and I’m also on facebook and Twitter :-))

…

(Ahem… Instagram? Pinterest?! Periscope?!? Snapchat?!?!)