Want to Normalize Scores? Ask Me How!

Unsupervised Linear Score Normalization: Intuition, Assumptions and Performance

Ilya Markov[1], Avi Arampatzis[2], Fabio Crestani[1]

[1]University of Lugano,
Lugano, Switzerland

[2]Democritus University of Thrace,
Xanthi, Greece

August 08, 2012

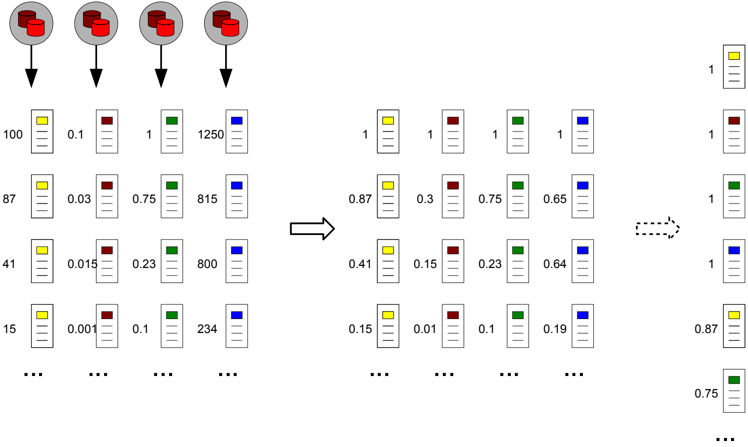
$$s_{norm} = \frac{s}{max}$$

$$s_{norm} = w \frac{s}{max}$$

Outline

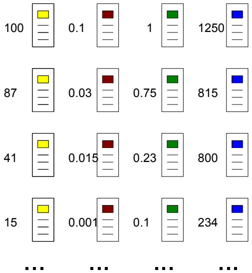
1. Introduction
2. Linear Score Normalization
3. Improving Linear Score Normalization

Score Normalization

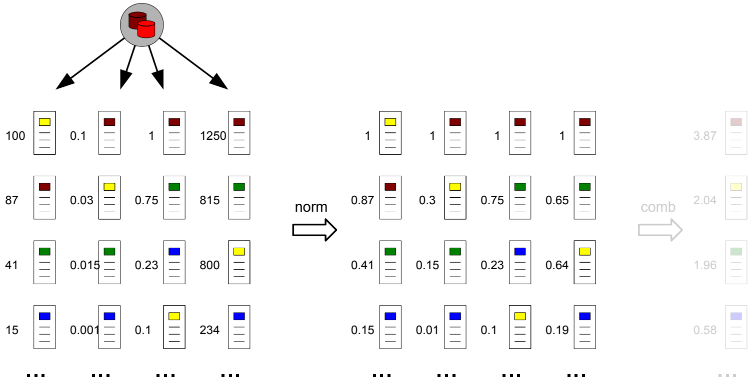


Assumptions

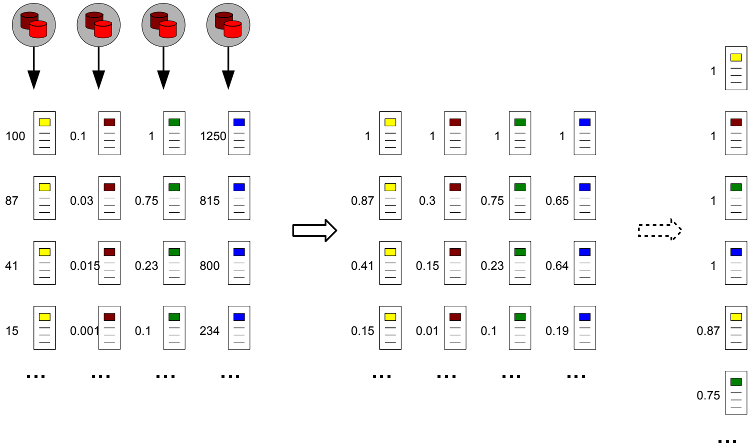
1. Document scores are provided
2. Document lists are disjoint



Data Fusion


100
0.1
1
1250
87
0.03
0.75
815
41
0.015
0.23
800
15
0.001
0.1
234
...
norm
1
1
1
1
0.87
0.3
0.75
0.65
0.41
0.15
0.23
0.64
0.15
0.01
0.1
0.19
comb
3.87
2.04
1.96
0.58

Score Normalization



Outline

1. Introduction
2. Linear Score Normalization
3. Improving Linear Score Normalization

Linear Score Normalization

- MinMax

$$s_{norm} = \frac{s - min}{max - min}$$

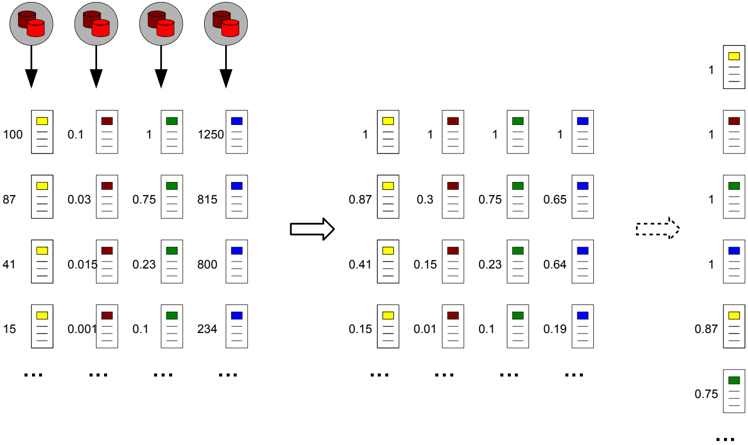
- Z-Score

$$s_{norm} = \frac{s - \mu}{\sigma}$$

- Sum

$$s' = s - min, \;\; s_{norm} = \frac{s'}{\sum_i s'_i}$$

MinMax



$$s_{norm} = \frac{s - min}{max - min}$$

MinMax

- Formula

$$s_{norm} = \frac{s - min}{max - min}$$

- Assumptions
  - Each collection contains at least 1 relevant document.
  - This document is most likely to be ranked 1st.
- Discussion
  - 1st documents are ranked before any other in the merged list.
  - High early precision is achieved when assumptions are hold.
- Results

	gov2.30		gov2.1000	
	p@10	p@20	p@10	p@20
top-10	0.1966	0.1936	0.0161	0.0148
top-1000	0.1953	0.1919	0.0161	0.0148

Max

- Observations
  - The lowest theoretical score of many scoring functions is 0.
- Formula

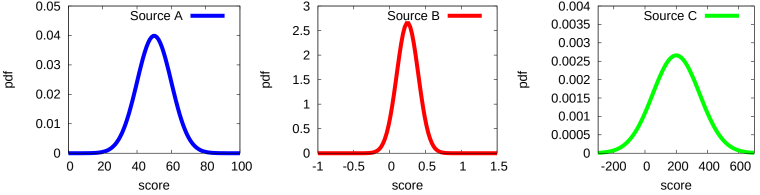
$$s_{norm} = \frac{s}{max}$$

- Results

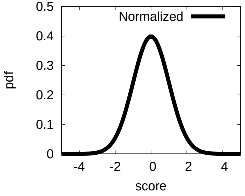
		gov2.30		gov2.1000	
		p@10	p@20	p@10	p@20
top-10	MinMax	0.1966	0.1936	0.0161	0.0148
	Max	0.1953	0.1926	0.0161	0.0148
top-1000	MinMax	0.1953	0.1919	0.0161	0.0148
	Max	0.1953	0.1919	0.0161	0.0148

- Discussion
  - Performance similar to MinMax.
  - Not affected by minimum score.

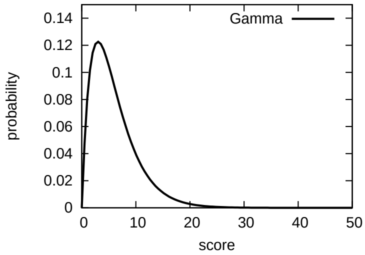
Z-Score

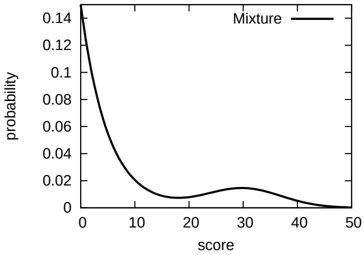

Source A
pdf
0.05
0.04
0.03
0.02
0.01
0
0
20
40
60
80
100
score
Source B
pdf
3
2.5
2
1.5
1
0.5
0
-1
-0.5
0
0.5
1
1.5
score
Source C
pdf
0.004
0.0035
0.003
0.0025
0.002
0.0015
0.001
0.0005
0
-200
0
200
400
600
score

⇓


Normalized
pdf
0.5
0.4
0.3
0.2
0.1
0
-4
-2
0
2
4
score

Z-Score


Gamma
probability
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
10
20
30
40
50
score


Mixture
probability
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
10
20
30
40
50
score

Z-Score

- Formula

$$s_{norm} = \frac{s - \mu}{\sigma}$$

- Assumptions
  - Score distribution is a bell-shape curve.
- Discussion
  - May be true for top documents.
  - Low ranked documents affect normalized scores of top ranked documents.
- Results

	gov2.30		gov2.1000	
	p@10	p@20	p@10	p@20
top-10	0.1839	0.1852	0.0007	0.0037
top-1000	0.0819	0.0953	0.0013	0.0023

Unit Variance (UV)

- Formula

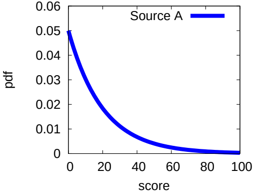
$$s_{norm} = \frac{s}{\sigma}$$

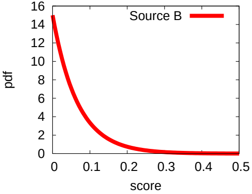
- Results

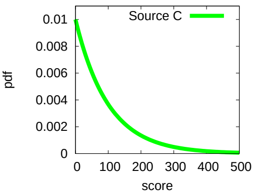
		gov2.30		gov2.1000	
		p@10	p@20	p@10	p@20
top-10	Z-Score	0.1839	0.1852	0.0007	0.0037
	UV	*0.1980*	*0.1936*	*0.0228*	*0.0188*
top-1000	Z-Score	0.0819	0.0953	0.0013	0.0023
	UV	*0.1248*	*0.1144*	0.0013	0.0007

- Discussion
  - More robust and less affected by low scores.

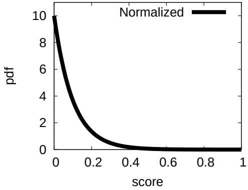
Sum







$\Downarrow$



Sum

$$s' = s - min, \;\; s_{norm} = \frac{s'}{\sum_i s'_i}$$

$$p(s) \sim \mathcal{E}(s; \lambda), \;\; \lambda = \frac{n}{\sum_i s_i}$$

$$p(s_{norm}) = p\left(\frac{s'}{\sum_i s'_i}\right) \sim \mathcal{E}(s; n)$$

Sum

- Formula

$$s' = s - min, \;\; s_{norm} = \frac{s'}{\sum_i s'_i}$$

scores become probabilities

- Assumptions
  - Document scores are distributed exponentially.
  - The mean of different exponentials is made the same.
- Discussion
  - Should not hold especially for top documents.
- Results

	gov2.30		gov2.1000	
	p@10	p@20	p@10	p@20
top-10	0.1785	0.1758	0.0013	0.0023
top-1000	0.0134	0.0114	0.0013	0.0017

Overall Results

	gov2.30		gov2.1000	
	p@10	p@20	p@10	p@20
MinMax	0.1966	0.1936	0.0161	0.0148
Max	0.1953	0.1926	0.0161	0.0148
Z-Score	0.1839	0.1852	0.0007	0.0037
UV	0.1980	0.1936	*0.0228*	*0.0188*
Sum	0.1785	0.1758	0.0013	0.0023

Table: Top 10 documents are retrieved.

	gov2.30		gov2.1000	
	p@10	p@20	p@10	p@20
MinMax	*0.1953*	*0.1919*	*0.0161*	*0.0148*
Max	*0.1953*	*0.1919*	*0.0161*	*0.0148*
Z-Score	0.0819	0.0953	0.0013	0.0023
UV	0.1248	0.1144	0.0013	0.0007
Sum	0.0134	0.0114	0.0013	0.0017

Table: Top 1000 documents are retrieved.

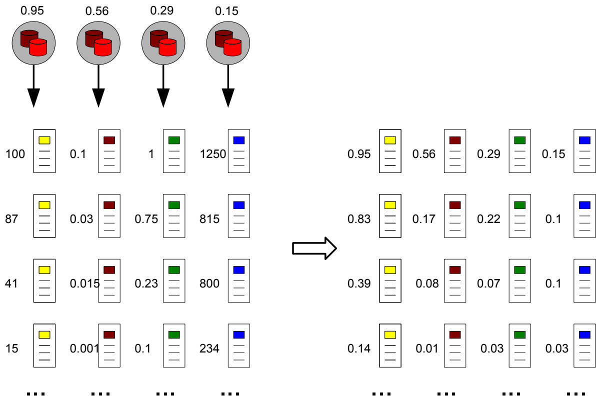
$$s_{norm} = \frac{s}{max}$$

Outline

1. Introduction
2. Linear Score Normalization
3. Improving Linear Score Normalization

$$s_{norm} = w \frac{s}{max}$$

Weighted MinMax

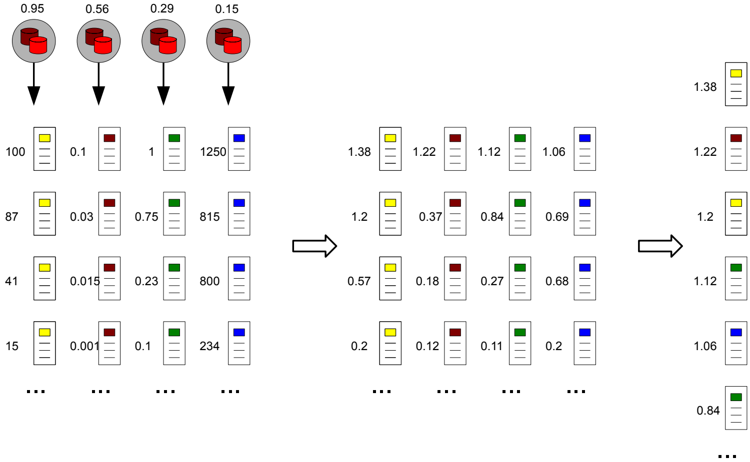


$$s_{norm} = w_{coll} \frac{s - min}{max - min}$$

Weighted MinMax

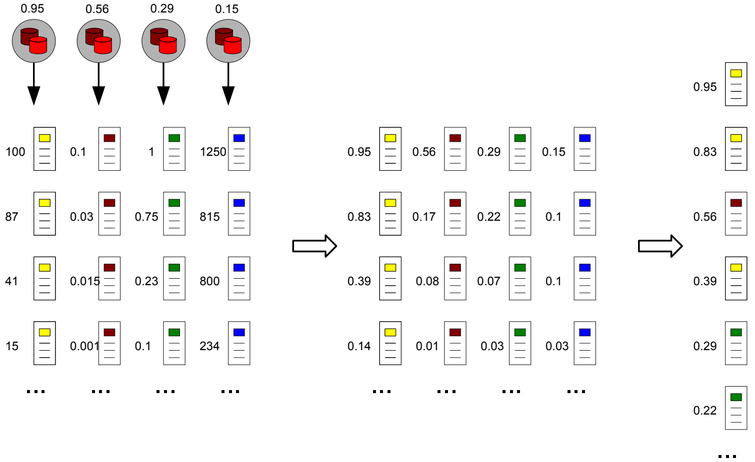
	gov2.30		gov2.1000	
	p@10	p@20	p@10	p@20
MinMax	0.1966	0.1936	0.0161	0.0148
WeightedMM	*0.3658*	*0.2802*	*0.2705*	*0.2205*

Collection Retrieval Inference Network (CORI)



$$s_{norm} = (1 + 0.4 \cdot w_{coll}) \frac{s - min}{max - min}$$

Weighted MinMax



$$s_{norm} = w_{coll} \frac{s - min}{max - min}$$

Collection Retrieval Inference Network (CORI)

	gov2.30		gov2.1000	
	p@10	p@20	p@10	p@20
WeightedMM	*0.3658*	*0.2802*	*0.2705*	*0.2205*
CORI	0.2906	0.2295	0.1879	0.1131

Open Question

Is it possible to calculate collection weights simpler and with minimum additional information?

$$s_{norm} = \frac{s}{max}$$

$$s_{norm} = w \frac{s}{max}$$

Alternative ways to calculate collection weights?