A repetition based measure for verification of text collections and for text categorization

Dmitry V. Khmelev and William J. Teahan

Lucas Rettenmeier

September 25, 2015

Overview

Repetition measure R

- Main goal: Verification of text collections.
- Can be used for Multi-class categorization.
- Very simplistic approach to authorship attribution.
- Works in every language, without any preprocessing.

Theory

Notation

$$R(T|T_1,...T_m) \in [0,1]$$

- Each document is a string $T_i[1,...|T_i|]$, thereby works like a mapping:

$$T_i : \{1,2,..|T_i|\} \rightarrow \{\texttt{ASCII}\}$$

- Length of the document:

$$|T| =: l$$

- Suffix of a string T:

$$T[i,...l]$$

- Lenght of the longest prefix of S, repeated in one of the documents $T_1,...T_m$:

$$Q(S|T_1,...T_m)$$

Theory

Definition

- Squared R - measure:

$$R^2(T|T_1,\dots T_m) = \frac{2}{l(l+1)} \sum_{i=1}^{l} Q(T[i,\dots l] \mid T_1,\dots T_m)$$

Properties:

- Well-behaved in many situations.
- Possible values:

$$R \geq 0: \quad R^2(\text{"}abc\text{"}|\text{"}def\text{"}) = \frac{2}{l(l+1)} \times 0 = 0$$

$$R \leq 1: \quad R^2(T|T) = \frac{2}{l(l+1)} \sum_{i=1}^{l} (l-i) + 1 = 1$$

Theory

Example

R-measure of:

$$\texttt{"cat sat on"}$$

in respect to the collection:

$$T_1 = \texttt{"the cat on a mat"} \text{ and } T_2 = \texttt{"the cat sat"}$$

Result

$$R^2(T|T_1, T_2) = \frac{2}{10 \times (10+1)} [(7+5+4+3)) + (5+4+3+2+1)] \approx 0.727272$$

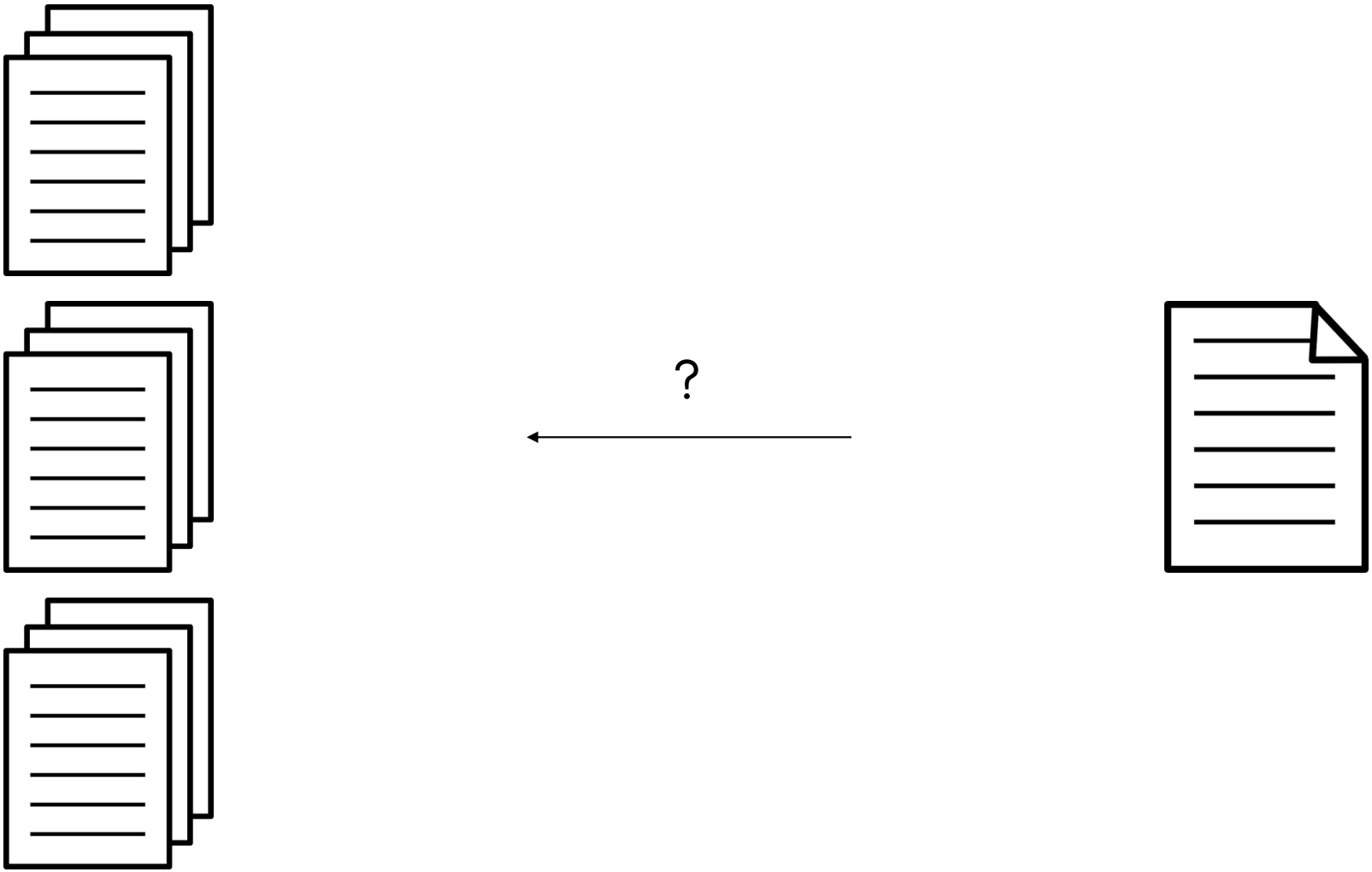
$$R(T|T_1, T_2) = \sqrt{R^2(T|T_1, T_2)} \approx 0.852802$$

Demo

Locating duplicates/plagiarism

- R-measure itself delivers no information about which document is similar to our test document.
- Heuristic pruning algorithm $\Rightarrow$ list of greatest contributing documents to R.

Multi-Class categorization



Multi-Class categorization

Authorship attribution

- Set of training documents $\{T_i\}$ for different authors.
- Document U with unknown authorship.

$$\textbf{Real} \text{ author: } r = \arg\max_i R(U|T_i)$$

Identify foreign and/or non-typical documents

- Training documents for different languages $\Rightarrow$ very fast language identification.
- Very small R-values (≤ 0.01) indicate special properties of the document (almost only names, numbers,...).

Results

Method	$R < 0.25$	$R < 0.5$	$R < 0.75$	$R < 1.00$	$R \leq 1.0$
R-measure	82.1	86.4	87.1	87.8	89.0
Multi-SVM	80.6	83.4	83.5	84.6	85.0
Bzip2	56.9	55.2	45.9	51.9	48.2
Gzip	55.7	53.5	53.9	50.1	59.4
Markov Chains, order 1	62.3	64.6	63.2	64.3	66.1
Markov Chains, order 2	60.9	64.4	61.8	64.7	64.5
Markov Chains, order 3	48.6	60.3	59.3	61.7	63.3
RAR	**84.3**	**86.9**	**87.3**	88.5	**89.4**
PPMD, order 2	77.8	79.1	79.4	80.5	81.3
PPMD, order 3	80.6	82.3	84.0	85.0	86.4
PPMD, order 4	82.5	85.4	86.0	87.7	88.4
PPMD, order 5	82.2	86.1	86.3	**88.8**	89.2

Table: Results of the authorship attribution

Reproducing the Results

- Programm written in C++.
- Corpus (C50) consists of newspaper articles of 50 different authors (each about 5000 characters long).
- Used 10 authors, 50 training + 50 test articles per author.
- High computing time: $\Theta(l, a, n) = l^2 \times a^2 \times n^2$.

Own Results

RL/AT	AP	BH	EF	JM	JB	LZ	NL	SD	TF	WK
AP	**47**	0	0	1	0	0	0	2	0	0
BH	0	**43**	0	0	0	0	0	7	0	0
EF	0	0	**34**	2	4	0	0	6	4	0
JM	0	0	0	**42**	1	0	0	0	2	5
JB	0	0	0	0	**49**	0	0	1	0	0
LZ	0	0	0	0	0	**43**	4	3	0	0
NL	0	0	0	1	0	0	**50**	0	0	0
SD	0	0	0	1	0	0	0	**49**	0	0
TF	0	0	0	0	0	0	0	1	**49**	0
WK	0	0	0	14	0	0	0	2	0	**34**

Table: Results of the multi-class categorization

$\Rightarrow$ **87,8 %** correctly identified documents

Improved Results

RL/AT	AP	BH	EF	JM	JB	LZ	NL	SD	TF	WK
AP	**49**	0	0	0	0	0	0	0	0	1
BH	0	**48**	1	0	0	0	0	0	0	1
EF	0	0	**40**	0	4	0	0	0	6	0
JM	0	0	0	**38**	0	0	0	0	0	12
JB	0	0	0	0	**48**	0	0	0	1	1
LZ	0	0	0	0	0	**45**	4	1	0	0
NL	0	0	0	0	0	0	**50**	0	0	0
SD	0	0	1	1	0	1	0	**43**	0	4
TF	0	0	1	0	1	0	0	0	**48**	0
WK	0	0	0	8	0	0	0	0	0	**42**

Table: Improved Results of the multi-class categorization

$\Rightarrow$ **90,2 %** correctly identified documents

Reproducibility

- Simple approach $\Rightarrow$ No real problems occured.
- Very detailed description.
- References to other papers where quite clear.

Conclusion

Positive

- Precision of over 90 % with a very simplistic approach.
- Very effective in finding duplicates and plagiates in large text collections.

Negative

- Only tested with large training sets (about 100.000 characters).
- Newspaper articles may not be representative.

What to do next?