Local Histograms of Character N-grams for Authorship Attribution

Escalante, Solorio, Montes-y-Gómez

Michael Träger

Studienstiftung des deutschen Volkes
Sommerakademie Nizza — Who wrote the Web?

September 25, 2015

Content I

1 Approach
- BOW
- LOWBOW
- Authorship Attribution with LOWBOW

2 Experiments
- Datasets
- Methodology

3 Results
- Conclusion

4 Reproduction

Local Histograms (LH)

- **enriched** histogram representations
- separate LH for each document-part
- combine more LHs:
  word/char usage (**frequency**) + **sequential** information

- more helpful than global histograms
- also challenging situations:
  - imbalanced training sets
  - small training sets

Histograms

Evolution

- word histograms

  ⇓

- n-grams at *word* level

  ⇓

- n-grams at *character* level

Bag of words

Representation (BOW)

- one document: histogram over vocabulary
- weighting: binary (or other)

Locally-weighted bag-of-words

Representation (LOWBOW)

- several local histograms per document
- terms of documents weighted:
  - smoothed by kernel function $K_{\mu,\sigma}(x)$
  - term position weighting
  - term frequency weighting
- over terms in vocabulary

Locally-weighted bag-of-words

Representation (LOWBOW)

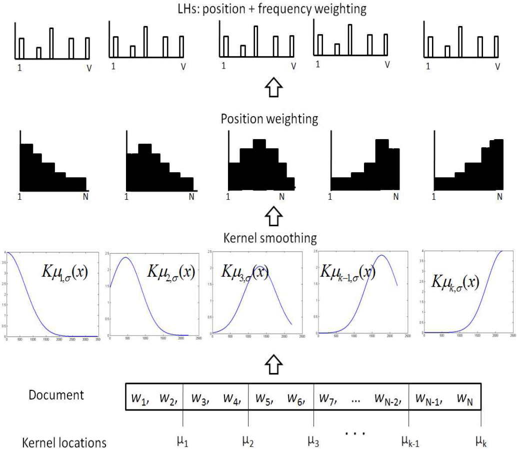


Figure 1: Process for obtaining local histograms. [291]

Approach

LOWBOW & BOLH

LOWBOW histogram

- unweighted sum of LHs
- term usage + sequential information

BOLH (Bag of local histograms)

- term occurrence frequencies across different locations on document

Approach

SVM

- multiclass SVM
- associate patterns-outputs (results of LOWBOW / set of LHs) to documents authors

LOWBOW

- linear kernel

BOLH

- no standard kernel
- Diffusion
- Eucidean
- χ^2

Experiments

Data-Set

- Plakias and Stamatatos, 2008a+b
- subset of RCV1 collection
- docs authored by 10 authors
- same topic
- 50 docs per author for training and also 50 for testing

Experiments

Settings

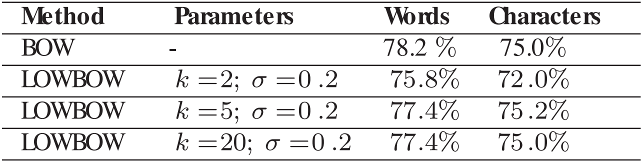
- 3-grams

- balanced corpus (BC)
- balanced reduced data sets (RBC)
- imbalanced reduced data sets (IRBC)

Results

Balanced Data

- LOWBOW histogram vs BOW



Method	Parameters	Words	Characters
BOW	-	78.2 %	75.0%
LOWBOW	$k=2; \sigma=0.2$	75.8%	72.0%
LOWBOW	$k=5; \sigma=0.2$	77.4%	75.2%
LOWBOW	$k=20; \sigma=0.2$	77.4%	75.0%

Figure 2: Accuracy for BOW and LOWBOW, with char/word n-grams

- with char and word n-grams

- BOW very effective
- LOWBOW worse when $k = 2$ LHs

Results

Balanced Data

- BOLH (superior to LOWBOW, BOW)

Kernel	Euc.	Diffusion	EMD	χ^2
		Words		
Setting-1	78.6%	81.0%	75.0%	75.4%
Setting-2	77.6%	82.0%	76.8%	77.2%
Setting-3	79.2%	80.8%	77.0%	79.0%
		Characters		
Setting-1	83.4%	82.8%	84.4%	83.8%
Setting-2	83.4%	84.2%	82.2%	84.6%
Setting-3	83.6%	**86.4%**	81.0%	85.2%

Figure 3: Accuracy for BOLH, with char/word n-grams

- setting 1, 2, 3 correspond to $k = 2, 5, 20$
- diffusion kernel outperforms best results
- characters better than words

Results

RBC - Reduced

- more realistic setting
- BOW, LOWBOW histogram, BOLH (diffusion kernel, $k = 20$)

Results

Balanced Data

			WORDS		
Data set	**Balanced**				
Setting	*1-doc*	*3-docs*	*5-docs*	*10-docs*	*50-docs*
BOW	36.8%	57.1%	62.4%	69.9%	78.2%
LOWBOW	37.9%	55.6%	60.5%	69.3%	77.4%
Diffusion kernel	52.4%	63.3%	69.2%	72.8%	82.0%
Reference	-	-	53.4%	67.8%	80.8%

			CHARACTER N-GRAMS		
Data set	**Balanced**				
Setting	*1-doc*	*3-docs*	*5-docs*	*10-docs*	*50-docs*
BOW	65.3%	71.9%	74.2%	76.2%	75.0%
LOWBOW	61.9%	71.6%	74.5%	73.8%	75.0%
Diffusion kernel	70.7%	78.3%	80.6%	82.2%	86.4%
Reference	-	-	50.4%	67.8%	76.6%

Figure 4: Accuracy for RBC, with char/word n-grams

Results

RBC - Reduced

- best performance: BOLH (diffusion kernel)
- LHs more beneficial with less documents
- character-level significantly better than word-level

Results

Imbalanced Data

WORDS								
Data set	**Balanced**					**Imbalanced**		
Setting	*1-doc*	*3-docs*	*5-docs*	*10-docs*	*50-docs*	*2-10*	*5-10*	*10-20*
BOW	36.8%	57.1%	62.4%	69.9%	78.2%	62.3%	67.2%	71.2%
LOWBOW	37.9%	55.6%	60.5%	69.3%	77.4%	61.1%	67.4%	71.5%
Diffusion kernel	52.4%	63.3%	69.2%	72.8%	82.0%	66.6%	70.7%	74.1%
Reference	-	-	53.4%	67.8%	80.8%	49.2%	59.8%	63.0%

CHARACTER N-GRAMS								
Data set	**Balanced**					**Imbalanced**		
Setting	*1-doc*	*3-docs*	*5-docs*	*10-docs*	*50-docs*	*2-10*	*5-10*	*10-20*
BOW	65.3%	71.9%	74.2%	76.2%	75.0%	70.1%	73.4%	73.1%
LOWBOW	61.9%	71.6%	74.5%	73.8%	75.0%	70.8%	72.8%	72.1%
Diffusion kernel	70.7%	78.3%	80.6%	82.2%	86.4%	77.8%	80.5%	82.2%
Reference	-	-	50.4%	67.8%	76.6%	49.2%	59.8%	63.0%

Figure 5: Accuracy for RBC and IRBC, with char/word n-grams

Results

IRBC - Imbalanced

- BOW + LOWBOW OK
- BOLH performed best
- BOLH robust to reduction and imbalanced data

Conclusion

- local histograms are advantageous
- paper-conclusion:
  LHs can uncover writing preferences of author
- improvements larger in reduced + imbalanced data sets

Reproduction

Implementation

//TODO implement me.

References I

[1] Escalante, H.J., Solorio, T., Montes-y-Gómez, M.: Local Histograms of Character M-grams for Authorship Attribution. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, 288-298, (2011)

Questions?

Michael Träger

michael.traeger@stud.uni-bamberg.de

Comic

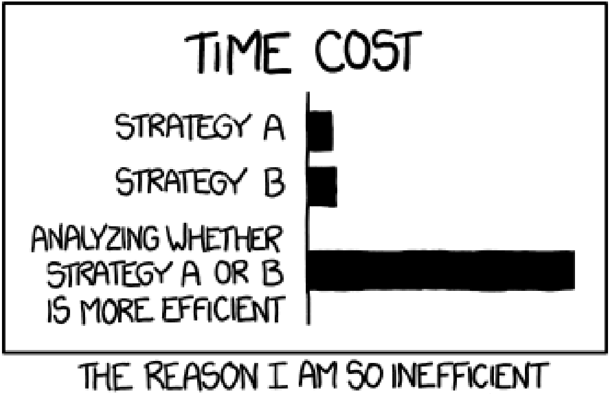

TIME COST
STRATEGY A
STRATEGY B
ANALYZING WHETHER STRATEGY A OR B IS MORE EFFICIENT
THE REASON I AM SO INEFFICIENT

Figure 6: Randall Munroe - xkcd.com/1445