Writeprints: A Stylometric Approach to Identity-level Identification and Similarity Detection in Cyberspace

Author: Ahmed Abbasi

University of Arizona

Presenter: Faizan Ahmad

https://qdata.github.io/deep2Read

Outline

1. Introduction
2. System Design
3. Extended Feature Set
4. Testbed
5. Results

Introduction

- Authorship attribution and similarity detection
- Feature based approach
- Contributions
  - Better Feature Sets (Extended Feature Set)
  - Better Feature Selection (Karhunen Loeve Transform)

System Design

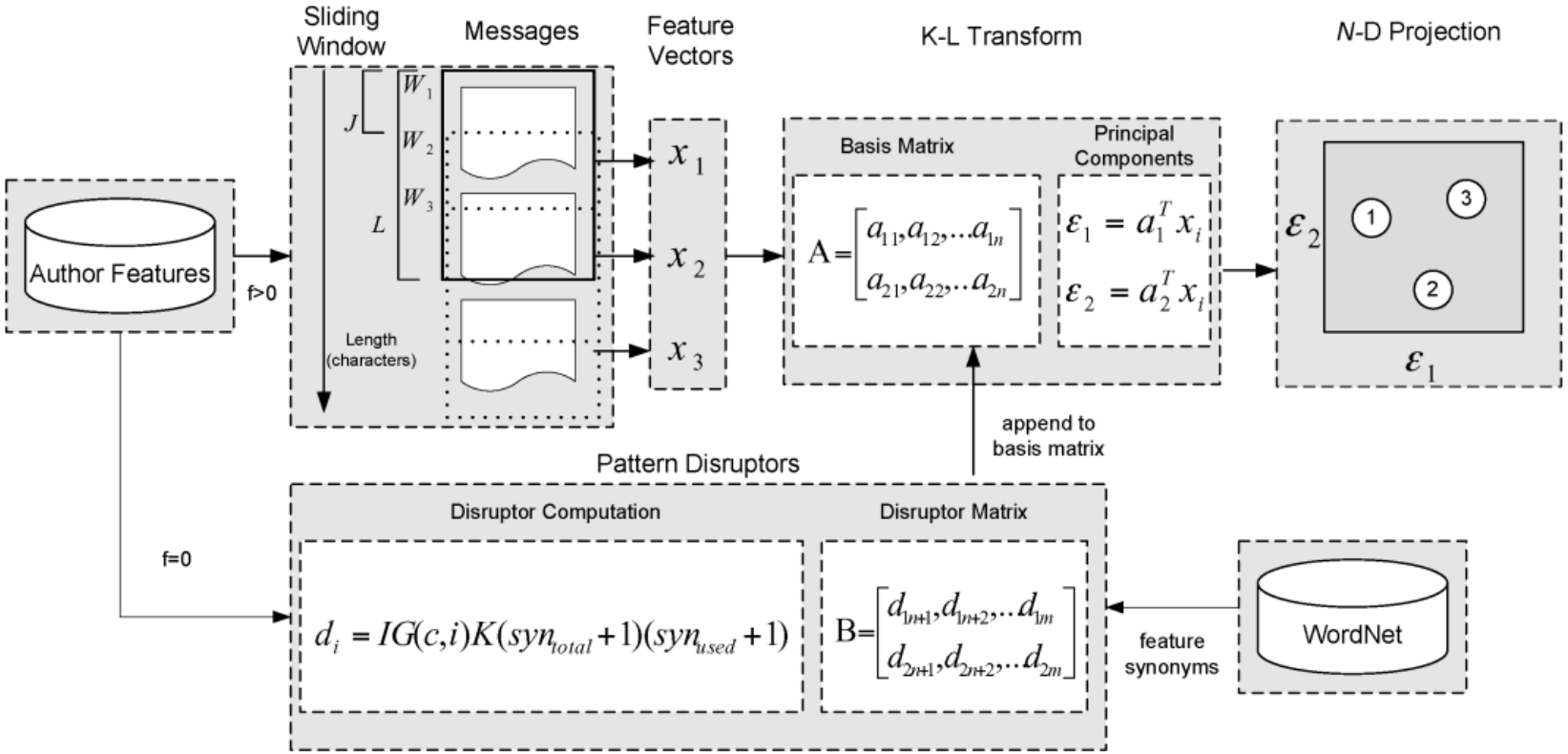

Fig. 3. Writeprints creation illustration.

Figure: Writeprints system design overview

Extended Feature Set

Group	Category	Quantity		Description
		Baseline (BF)	Extended (EF)	
Lexical	Word-Level	5	5	total words, % char. per word
	Character-Level	5	5	total char., % char. per message
	Letters	26	26	count of letters (e.g., a, b, c)
	Character Bigrams	—	<676	letter bigrams (e.g., aa, ab, ac)
	Character Trigrams	—	<17,576	letter trigrams (e.g., aaa, aab, aac)
	Digits	—	10	digits (e.g., 1, 2, 3)
	Digit Bigrams	—	<100	2 digit number frequencies (e.g., 10, 11)
	Digit Trigrams	—	<1,000	frequency of 3 digit numbers (e.g., 100)
	Word Length Dist.	20	20	frequency of 1–20 letter words
	Vocab. Richness	8	8	richness (e.g., hapax legomena, Yule's K)
	Special Characters	21	21	occurrence of special char. (e.g., @#$%^)

Figure: Extended Feature Set

Extended Feature Set

Syntactic	Function Words	150	300	frequency of function words (e.g., of, for)
	Punctuation	8	8	occurrence of punctuation (e.g., !;:,.?)
	POS Tags	—	<2,300	frequency of POS tags (e.g., NP, JJ)
	POS Tag Bigrams	—	varies	POS tag bigrams (e.g., NP VB)
	POS Tag Trigrams	—	varies	POS tag trigrams (e.g., VB JJ)
Structural	Message-Level	6	6	e.g., has greeting, has url, quoted content
	Paragraph-Level	8	8	e.g., no. of paragraphs, paragraph lengths
	Technical Structure	50	50	e.g., file extensions, fonts, use of images
Content	Words	20	varies	bag-of-words (e.g., “senior”, “editor”)
	Word Bigrams	—	varies	word bigrams (e.g. “senior editor”)
	Word Trigrams	—	varies	word trigrams (e.g., “editor in chief”)
Idiosyncratic	Misspelled Words	—	<5,513	misspellings (e.g., “beleive”, “thougth”)

Figure: Extended Feature Set

Testbed

Data Set	Domain	No. Authors	Words (per Author)	Time Duration	Noise
Enron Email	Asynchronous (D1)	100	27,774	10/98–09/02	Yes
EBay Comments	Asynchronous (D1)	100	23,423	02/03–04/06	No
Java Forum	Program Code (D4)	100	43,562	04/03–05/06	Yes
CyberWatch Chat	Synchronous (D2)	100	1,422	05/04–08/06	No

Figure: Details for Datasets in Testbed

Results

Test Bed	Techniques/Features	No. Authors		
		25	**50**	**100**
Enron Email	Writeprint	**92.0**	**90.4**	**83.1**
	Ensemble	88.0	88.2	76.7
	SVM/EF	87.2	86.6	69.7
	Baseline	64.8	54.4	39.7
eBay Comments	Writeprint	**96.0**	**95.2**	**91.3**
	Ensemble	**96.0**	94.0	90.9
	SVM/EF	95.6	93.8	90.4
	Baseline	90.6	86.4	83.9
Java Forum	Writeprint	88.8	66.4	52.7
	Ensemble	92.4	85.2	**53.5**
	SVM/EF	**94.0**	**86.6**	41.1
	Baseline	84.8	60.2	23.4
CyberWatch Chat	Writeprint	**50.4**	**42.6**	**31.7**
	Ensemble	46.0	36.6	22.6
	SVM/EF	40.0	33.3	19.8
	Baseline	37.6	30.8	17.5

Figure: Classification Results