

---

# ID Identification in Online Communities

---

**Yufei Pan**

Rutgers University

YUFEIPAN@CS.RUTGERS.EDU

## Abstract

We are trying to identify IDs of users in an online community from the text they've produced. In particular, we focus on two questions about ID identification: 1. Given a piece of text, could we identify the ID for it from known IDs? 2. Given an ID with all its text evidences, could we identify it with any other known ID, i.e. are they belonging to the same entity? For the first question, we discuss the difference between text identification and text categorization. Then we give a generic approach based on stylometric features and machine learning techniques. For the second question, based on the observation that IDs may demonstrate various text behaviors in different online environments, we try to find out the constant variation pattern for an entity, which may be independent of the ID it uses. Besides, because IDs are interacting with each other in an online community, we also propose the idea to exploit this relational information to achieve a better identification as our future work.

## 1. Introduction and Motivation

Nowadays online communities are a large part of lives of people. In those communities, the beings interact with each other via different IDs. Naturally, an interesting yet fundamental question arises: who is that ID? People become curious about the ID they are interacting with, directly or indirectly, actively or passively, frequently or infrequently. They want to get certain information about the IDs they are interacting with, provided the information they've already had or could get easily.

In this paper we focus on two questions about ID identification: 1. Given a piece of text, could we identify the ID for it from known IDs? 2. Given an ID with all its text evidence, could we identify it with any other known ID, i.e. do they come from the same entity?

We call them as Text Identification and ID Identification.

### 1.1 Text Identification VS Text Categorization

Text categorization is a well-known area in the AI research and attracts a lot of intellectuals. Its major task is to categorize the type of text based on the content. Hence, it often treats the text as a bag of words then uses many classification techniques in machine learning to classify, such as decision trees, Bayesian probabilistic approaches, or neural networks.

However, for the text identification, we are trying to identify the ID who produces the text. The similarity of content couldn't give us any clue of the producer. Instead, we need find out the constant features for an ID to identify the text, independent of text content.

Motivated by the work in the author attribution research and progress in the machine learning, we summarize a generic approach for text identification in the section 2.

### 1.2 Entities VS IDs

In an online community, it is a well-known fact that an entity may have many different IDs. Here we use entity to represent the certain being such as human being, and use ID to represent the identification within an online community, which could be observed by other entities.

For example, people may use more than one ID in a forum to post all kinds of stuff, good or bad, legal or illegal. Or during online chatting, people will use multiple IDs even while talking to the same ID. The purposes could vary from attacking others to self-protection, from making fun to make others sad. So it might be interesting to infer which IDs correspond to the same entity.

This also motivates us to do some preliminary work on ID identification: identify if two IDs belong to same entity.

### 1.3 Text Identification VS ID Identification

At the first sight, it seems we could easily solve the ID Identification if we could solve Text Identification. We could test the text evidences of one ID with the other's stylometric model. If they are matched well, then they belong to the same entity.

However, this depends on the consistency of the stylometric features over different IDs. What if the entity controls the text styles for each ID intentionally,

or he/she unconsciously changes the text behavior to match the expected behavior of ID?

Based on the observation that IDs may demonstrate various text behaviors in different online environments, we try to find out the constant variation pattern for an entity, independent of the ID it uses. The pattern could help to achieve a better identification.

Furthermore, since IDs are interacting with other IDs in an online community, this also motivates us to explore the possibility of exploiting the ID interactions to solve the ID identification problem.

The rest of paper is organized as follows: Section 2 discusses the generic approach to solve text identification problems. We also give two original ideas to improve ID identification accuracy. Section 3 presents the experiment, results and the analysis. Section 4 gives the conclusions of the paper. Section 5 points out the limitations of the work and discusses the possible future work.

## 2. ID Identification

### 2.1 Stylometric Features

Roughly speaking, text identification is very similar to author attribution, which enjoys a long history in the literature research. The problem is to identify who is the author of an article or book. The major technique they use is to discover the style features of an author with his/her known text. These are so-called stylometric features.

Rudman(1997) and other researchers reviewed the history of author attribution. They discussed the problems they've encountered and the important achievement they've accomplished. Among those achievements, stylometric features are of most importance. Also according to Rudman, over 1,000 stylometric features have been proposed. Tony (2000) also gives many interesting stylometric features.

Diederich et al(2000) extracts stylometric features then uses support vector machines to find the author of a unknown text.

Summarizing the main points of above research, we can articulate the following generic for text identification:

Firstly, we would extract some kind of stylometric features. Secondly, we would choose some kind of machine learning algorithms. Finally, we conduct experiments to get the good results. In this approach, the selection of stylometric features, the selection of algorithms and the parameters of algorithms are crucial to the final performance. And there is no best rule for all author attribution application due to the difference of language, topics and nature of people.

### 2.2 Style Variation Pattern

As we pointed out in Section 1.2, an entity may have multiple IDs. And the entity may assign different text behavior style to different IDs intentionally or unconsciously. Thus, we could not simply rely on the generic approach for ID identification.

Yet based on our observation, we realized that for the same ID, an entity would demonstrate a certain style variation over changed environment. For example, the text behavior varies when the topic categories are changed to some degree, which is reflected in our experiment result.

Inspired by this observation, we try to find the constant variation pattern for an entity, which is independent of the ID it uses. And we argue that if two IDs have the similar variation pattern, it is very likely they belong to the same entity. Roughly speaking, we try to examine the variations of IDs over the environment dimension to give the possible answer to the ID identification.

It is not easy to calculate and represent the variation directly. Here we continue to use stylometric features and machine learning algorithms to achieve this indirectly.

#### Definition:

$IDS = \{ID_1, ID_2, \dots, ID_n\}$ , define the ID set.

$ES = \{E_1, E_2, \dots, E_m\}$ , define the Environment Dimension Set. For any  $E \in ES$ , we denote  $D(E)$  as the domain of that environment dimension.

$TRS = \{TR_i \mid TR_i \text{ is the training data set for } ID_i\}$

$TSS = \{TS_i \mid TS_i \text{ is the testing data for } ID_i\}$

#### Learning the variation pattern:

Given an  $ID_i$  and  $E_j$ , we first drill down the  $TR_i$  along  $E_j$ . For any  $e_k \in E_j$ , we have a subset of  $TR_i$ :  $TR_{ijk}$ . Similarly, we get  $TS_{ijk}$  for testing data subset for  $ID_i$ ,  $e_k \in E_j$ .

Then we apply the generic text ID identification approach on the  $TR_{i[j,e]}$  to get the trained model  $M_{ij}[k]$ .

For each  $M_j[e]$ , we use testing data set to get the variation matrix:  $VM_{ij}$

$VM_{ij}[k,l] =$  the correctness of  $M_{ij}[k]$  tested with  $TS_{ijl}$ , range from 0 to 100.

#### Inferring:

Once we have the Variation Matrix Cluster for each ID, then we could compare the similarity of these Matrixes over a specific Environment dimension. For example, we could simply convert the variation matrix to a binary matrix according some rule. Or we could apply

various clustering techniques over the row vectors or column vectors. Or we could use linear algebra techniques such as eigenvalue, inner product of row vectors or column vectors. Matrix similarity itself is a quite interesting problem.

To achieve a good result and reduce the calculation complexity, we also need to choose reasonable environment dimension that may require the involvement of human knowledge.

In the proof of concept experiment, we give two simple variation matrices and provide a simple comparison.

### 3. Experiments

#### 3.1 Experiment setup

##### 3.1.1.1 INPUT DATA

We collect the IDs and their text from **2nd light forum** (<http://2ndlight.com/forum42ndlight/index.cfm>). The feature of this forum is that many IDs will post across different topic categories, which is good for our Style Variation Experiment.

Due to the lack of existing tools to collect text data from a specific forum automatically, we develop our own data collector and a data processor to parse and generate the final data file we could use. However, it is still not an easy task to get a large amount data for it need a lot of human involvement.

##### 3.1.2 STYLOMETRIC FEATURES

We select the features, which are fit for the short text, according to De vel, 2001. The number of features is 56.

Number of blank lines/total number of lines
Average sentence length(number of words)
Average word length (number of characters)
Total number of function words/W
Function word frequency distribution (44 features)
Total number of short words/W
Total number of alphabetic characters in words/C
Total number of upper-case characters in words/C
Total number of characters in words/C
Total number of digit characters in words/C
Total number of white-space characters/C
Total number of tab spaces/number white-space characters
Vocabulary richness ( V/W)

Note: W = total number of words, C = total number of characters, T = total number of types (i.e. distinct words).

#### 3.1.3 MACHINE LEARNING ALGORITHM

We use support vector machine to train our identification model because we have 56 features for each data point. The advantage of SVM is that the number of free parameters used in training only depends on the margin and does not depends on the number of input features (Diederich 2000). Also, the implementation we used is weka.classifiers.SMO, included in WEKA package.

### 3.2 Experiment Result

#### 3.2.1 TEXT IDENTIFICATION

	Training	Testing
Correctly Classified	63.6364%	56.6038%
Incorrectly Classified	36.3636%	43.3962%
Kappa statistic	0.2542	0.1147
Mean absolute error	0.4646	0.4731
Root mean squared error	0.4742	0.4845
Relative absolute error	93.1145 %	95.5107 %
Root relative squared error	94.9291 %	97.7028 %
Total Number of Instances	88	53

The ID we use is Floridave. The number of training instances is 88, and the number of testing instances is 53. Both are not overlapped with each other and are across multiple topic categories. For the test confusion, we identify 13 over 32 Floridave's texts not of Floridave. We also identify 10 over 11 non-Floridave's text of Floridave.

False Positive	True Positive	False Negative	True Negative
0.41	0.59	0.48	0.52

#### 3.2.2 STYLE VARIATION PATTERN

We use topic category as our environment dimension. It means we will drill down the training data along the topic category. Then we learn the stylometric models with each subset of training data. Then we also drill down the testing data along topic category and perform the cross-testing on each learned stylometric model with subsets of testing data.

For each ID, then we could see the topic category affect the test result for each stylometric model with specific category, such as Surfing, Gardening and so on.

IDs	Topic Categories			
		Surfing	Gardening	NSR
Floridave	Surfing	83.33%	22.22%	18.18%
	Gardening	25%	44.44 %	9.09%
	NSR	16.66%	11.11%	81.82%
paddleout	Surfing	85.71%	50%	40%
	Gardening	42.86%	75%	0%
	NSR	14.29%	50%	60%

The above table gives 2 Variation Matrixes for 2IDs along the topic category environment dimension.

VM[Floridave] =	83.33	22.22	18.18
	25	44.44	9.09
	16.66	11.11	81.82
VM[paddleout] =	85.71	50	40
	42.86	75	0
	14.29	50	60

The eigenvalue for VM[Floridave] is: 109.2, 67.1, 33.3

The eigenvalue for VM[paddleout] is: 1.4, 0.4 + 0.2i, 0.4 - 0.2i

The obvious difference between VM[Floridave] and VM[paddleout] helps us that Floridave and paddleout are not corresponding to the same entity, even a simple aggregated test indicates that they may belong to the same one.

## 4. Conclusions

Combined with stylometric features extraction and machine learning algorithm, we could identify the id better than simple guess, i.e., the probability is above 50%.

Along the environment dimension, the styles will vary for different dimension value. And the patterns of variation are different for different ID.

## 5. Limitations and Future Work

### 5.1 Style Variation Pattern

Since the lack of available data from which we could actually know which IDs belong to the same entity, we could not prove that ids of the same entity will have similar variation pattern.

Meanwhile, how to represent the style variation pattern and how to examine the similarity of pattern is not very clear to us. We think it deserve more work to get the good result.

### 5.2 Exploit IDs' Interaction

IDs are not isolated, because of the basic nature of entities: desire for involvement. IDs communicate with each other a lot in online communities. We give two observations, which we learned from our own experience.

1 For two IDs belonging to same entity, the interaction groups overlap a lot, if not identical.

2 For two IDs belonging to same entity, the interaction degrees with the overlapped IDs are proportional.

As a future work, we could use data mining techniques to discover the interaction information to improve our identification with that graph on both false positive and false negative.

## Acknowledgements

Thanks a lot to Professor Michael Littman, who gives me nice suggestion on this project and useful reference lists. We also thank Professor Matthew Stone to lend us his nice linguistic book. Besides, we really appreciate the encouraging words we received from Yihua Wu.

## References

- J. Rudman. (1997) "The state of authorship attribution studies: Some problems and solutions". Computers and the Humanities, 31(4):351--365, 1997.
- De Vel. (2001) "Mining E-mail Content for Author Identification Forensics" SIGMOD Record, 2001.
- Joachim Diederich, Jörg Kindermann, Edda Leopold, Gerhard Paass (2000) "Authorship Attribution with Support Vector Machines", 2000.
- Tony McEnery, Michael Oakes. (2000) "Authorship Identification and Computational Stylometry", in Handbook of Natural Language Processing, chapter 23, pages 545-562. Marcel Dekker Inc.
- B. Taskar, M. F. Wong, P. Abbeel and D. Koller. (2003) "Label and Link Prediction in Relational Data" To appear in Neural Information Processing Systems Conference(NIPS03), Canada, 2003.