IDENTIFYING AUTHORS OF TEXT BASED ON ASSOCIATION RULES

Hiroshi Sugimura[1], Ryosuke Saga[2] and Kazunori Matsumoto[2]
[1]*Course of Information and Computer Sciences, Graduate School of Engineering*
[2]*Department of Information and Computer Sciences, Faculty of Information Technology,*
[1 e 2]*Kanagawa Institute of Technology, 1030 Shimo-ogino, Atsugi-shi, Kanagawa 243-0292, JAPAN*

ABSTRACT

This paper proposes a system that identifies the author of a text based on association rules. In the case of European languages, distribution of words, length of words and sentences, patterns of punctuations, and etc. are effective to identify authors. We cannot apply this approach to many Asian languages because they do not have explicit word boundaries in text. Preliminary processing of text data such as morphological analysis may influence the final results. Therefore, we propose a language independent method for identifying authors based on association rules of N-grams. The distribution of N-grams greatly depends on genre, domain, and writing period. Therefore, a user may not obtain interesting knowledge as author's feature from the distributions. We focus on combinations of N-grams, and extract association rules from these N-grams. The feature vector of an author is created by a set of probabilities of occurrence of association rules. For author identification, the system measures dissimilarity between two feature vectors.

KEYWORDS

Identifying authors, Authorship detection, Classification, Association rule, N-gram.

1. INTRODUCTION

Identifying authors is the problem of detecting the author of an anonymous text, or text whose authorship is in doubt [6]. Recently, huge databases of electronic texts have become available on the Internet, making the problem of managing large text collections increasingly important. Plagiarism detection is another application area for identifying authors. The techniques for this problem are proposed [4,6,7]. However, there are several problems of the standard approaches. Many Asian languages do not have explicit word segmentations. Thus finding the word segmentation is a difficult problem in Asian languages. This problem creates an extra problem with this process introduces. Our approach is based on N-gram. Hence, we do not use any language dependent information.

In [5], they propose a method for identifying authors based on the distribution of N-grams of characters in sentences. However, this method extracts N-grams of genre, domain, and writing period. And a user cannot obtain interesting knowledge as author's feature from these N-grams. We focus on co-occurrence of N-grams. Even if both two authors write a work of same genre, we expect that the co-occurrences of N-grams between two authors are different. In the author's feature discovery process, the system extracts association rules from obtained N-grams. An association rule is defined as an expression $X \Rightarrow Y$ with a support s and confidence c. The rule means that a transaction T contains X then T contains Y also with a probability of c. We use the Apriori algorithm [3] to discover all association rules. The Apriori algorithm discovers all rules whose support and confidence are greater than give minimal thresholds. We expect the obtained association rules as author's feature.

In the process for prediction of authorship, the system computes the dissimilarity between author's feature and an anonymous text. For computation of the dissimilarity, we compare four methods that are Tankard's method [2], dissim [5], Kullback-Leibler divergence, and cross entropy. As a result, it is shown that the calculated dissimilarities certainly look like the each author's feature. But, according to balance between total numbers of each author's feature, the dissimilarities between a work and each author's feature

becomes unbalanced. We solve this problem by normalizing the dissimilarities. Finally, the system is able to classify based on normalized dissimilarity.

2. OUTLINE OF THE SYSTEM

We show an outline of the system in Fig. 1. A user inputs pairs of author and work into the system. *Feature miner* extracts author's features by using N-gram and the Apriori algorithm. Extracted features are stored into *AFDB*. In process for prediction of authorship, a user inputs a work of unknown author into the system. *Identifier* computes dissimilarities between inputted work and stored author's features. The system outputs author's name with smallest dissimilarity as prediction.

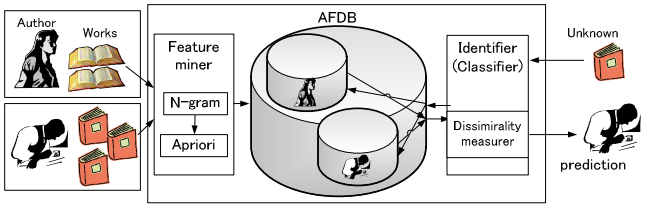

Figure 1. An outline of the system

3. EXTRACT ASSOCIATION RULES WITH N-GRAM

In Fig. 2, we illustrate the flow of the method of making rules and database. An association rule set A_i is made from each author database D_i.

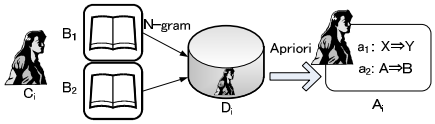

Figure 2. A simple example of training

The system converts each sentence of work B_j of known authors C_i to each transaction of database.N-gram set is generated by N-gram from a sentence. The N-gram set is labeled C_i as author i. A transaction is a pair of a N-gram set and a class label. Transaction database D_{ij} is created by applying to all sentence of a work B_j. Database D_i contains all transactions of an author C_i. Database D contains all authors.

Table 1. Example of database D

D_1	D_{11}	AUT, UTH, THO, HOR, ORS	C_1
		IDE, DEN, ENT, NTI, TIF, IFY, FYI, YIN, ING	C_1
	D_{12}	ASS, SSO, SOC, OCI, CIA, IAT, ATI, TIO, ION	C_1
D_2	D_{21}	A s, si, sim, imp, mpl, ple, e c, ca, cas, ase,	C_2
		Hol, olm, lme, mes, es, s r, re, rem, ema, mar, ark, rke, ked	C_2

Association rules extracted from database D. Let $I = \{i_1, i_2, \ldots, i_m\}$ be the set of database items. Each transaction T in the database D has a unique identifier, and contains a set of items which is called an itemset. An association rule is a conditional implication among itemsets, $X \Rightarrow Y$, where itemsets X, $Y \in I$, and $X \cap Y = \phi$. The support s of an itemset is the percentage of transactions in D which contain the itemset. The confidence c of the association rule a is the conditional probability that a transaction contains Y, given that it contains X. The support s and the confidence c are defined as:

$$s = \frac{|X \cup Y|}{|D|}, c = \frac{|X \cup Y|}{|X|}$$

The system finds all the association rules which satisfy both the two thresholds: minimum support S_{min} and minimum confidence C_{min}.

4. IDENTIFICATION BY ASSOCIATION RULES

Identification by individual association rule is carried out by using its support. The fraction of instances which contains an association rule may be a very small fraction of all instances. Hence, it cannot yield very accurate predictions if it is used by itself on all instances.

A set of the supports of all rules for a class make a vector. We call this feature vector. Classifier predicts the class membership by using feature vectors. In order to avoid zero frequency, we apply additional smoothing. When a rule set A_p for class C_p which has a total number of i rules (a_1,… , a_i), and s is a very small value for additional smoothing. Each x_i in the feature vector $X(A_p) = \{x_1, x_2,\dots , x_i\}$ is defined as:

$$x_i = \frac{\mathrm{p}(a_i) + s}{\sum_{r \in A_p} (\mathrm{p}(r) + s)}$$

In order to compute the dissimilarity between two feature vectors, we test four types of similarity calculation methods. This paper reports the experimental results on using Tankard [2], dissim [5], Kullback-Leibler divergence, and cross entropy. For example, let two probability distributions of two feature vector P,Q to be $P(x)$,$Q(x)$. In [2], Tankard's method is proposed as follows:

$$\mathrm{Tankard}(P, Q) = \sum_x |P(x) - Q(x)|$$

Table 2 shows example of the dissimilarity by using Tankard's method. A_i is a feature vector for class C_i. Work D_{ij} is known to be class C_i.

Table 2. Example of dissimilarity

	Number of rules	$D_{11}(C_1)$	$D_{21}(C_2)$	$D_x(C_2)$
Association Rules for $C_1(A_1)$	20	80	200	200
Association Rules for $C_2(A_2)$	100	300	150	300

In Table 2, D_{11} and D_{21} are classified well. However, the numbers of association rules for different classes may not be balanced. As a result, the classifier assigns class label which has the largest number of rules. In table 2, work D_x which is known to be class C_2 is assigned to class label C_1. For this problem, we carry out to normalize the dissimilarity.

To normalize by dividing them using a dissimilarity at a fixed percentile base(C_i) for the training instances of each class. We can let base(C_i) be the median of the dissimilarities of the training instances class C_i. Given database D_{ij} and feature vector A_i for class C_i, a method to compute dissimilarity is S(D_{ij} ,A_i), base(C_i) is defined as follows:

$$\mathrm{base}(C_i) = \mathrm{Median}(S(D_{i1}, A_i), \dots, S(D_{ij}, A_i))$$

Normalized dissimilarity, norm_dissim(D_x,C_i), between database D_x and feature vector A_i for class C_i is defined as:

$$\mathrm{norm\_dissim}(D_x, C_i) = \frac{S(D_x, A_i)}{\mathrm{base}(C_i)}$$

Result of requesting standardized difference level with each database. Table 3 shows normalized dissimilarity between each database when we assume the value in Table 2 to be the median. We predict the class membership of work D_x which is known to be class C_2. The dissimilarity between the class C_1 is *200/80 = 2.5*, and C_2 is *300/150 = 2.0*. D_x is assigned to class label C_2.

Table 3. Examples of the normalized dissimilarity

	$D_{11}(C_1)$	$D_{21}(C_2)$	$D_x(C_2)$
Association Rules for $C_1(A_1)$	80/80=1.0	200/80=2.5	200/80=2.5
Association Rules for $C_2(A_2)$	300/150=2.0	150/150=1.0	300/150=2.0

5. EXPERIMENT

We obtain text data of literary works from Aozora Bunko [1]. The total number of the text data is thirty (by ten authors, with three literary works for each author). In [6] concludes that the author of sentences with

22,336 characters in Japanese modern novels at least can be identified. We thus use 20,000 characters for each text. We prepare three kinds of transaction database (2-gram, 3-gram, and 4-gram) from these texts. Table 4 shows the number of extracted N-grams.

Table 4. Number of extracted N-grams

	2-gram	3-gram	4-gram
N-grams	69526	232513	374894

We set the confidence to a fixed value of 0.90 for the extraction of association rules, and carry out with three kinds of the support (0.01, 0.05, and 0.10). Table 5 shows the average number of population of association rules with minimum support. We examine and compare four kinds of measure described above. KL-D is KL divergence, and CE is cross entropy. Table 5 shows identification accuracy on each measure. Dashes indicate that results are unavailable which in the case of too many rules (or too few) to compute dissimilarity.

Table 5. Result of experiment

	2gram			3gram			4gram		
Support	0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10
Rules	134685.3	41.5	2.4	14465.8	7.2	0.5	1829.8	1.2	0.0
Tankard	-	40.0	75.0	77.8	44.4	-	44.4	-	-
Dissim	-	40.0	44.4	29.6	50.0	-	14.8	-	-
KLD	-	33.3	66.7	81.5	44.4	-	59.3	-	-
CE	-	20.0	58.3	77.8	25.9	-	70.4	-	-

6. CONCLUSION

We propose a system which identifies authorship based on association rules, and also show experimental results. In addition, our model of identification is easy to understand the underlying meaning of an identification pattern. Even if the prediction accuracy is not so high, the system that outputs understandable model is important for knowledge management. For example, in Table 5, we seem the rule extracted by both 2gram and Support 0.10 becomes interesting knowledge. Because, extracted rules by this settings has the higher accuracy than other settings of the parameters even though the total number of rules is very small. The number of extracted rules depends on the values of N and the support. Thus it is necessary to set appropriate values.

REFERENCES

[1] Aozora Bunko Project. http://www.aozora.gr.jp/

[2] J. Tankard, 1986. *The Literary Detective*, BYTE, pp. 231-238.

[3] R. Agrawal, T. Imielinski, and A. Swami, 1993. *Mining association rules between sets of items in large databases*. In Proc. of the ACM SIGMOD Conference on Management of Data, pp. 207-216.

[4] S. Kim, H. Kim, T. Weninger, and J. Han, 2010. *Authorship classification: a syntactic tree mining approach*, in Proceedings of the ACM SIGKDD Workshop on Useful Patterns. ACM, 2010, pp. 65-73.

[5] T. MATSUURA and Y. KANADA, 2000. *Extraction of authors'characteristics from japanese modern sentences via n-gram distribution*, in Discovery Science, ser. Lecture Notes in Computer Science, Springer, Berlin / Heidelberg, vol. 1967, pp. 315-319.

[6] V. Ke.selj, F. Peng, N. Cercone, and C. Thomas, 2003. *N-gram-based author profiles for authorship attribution*, in Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING '03. Citeseer.

[7] W. Zhang, T. Yoshida, and X. Tang, 2008. *Text classification based on multi-word with support vector machine*, Knowledge Based Systems, vol. 21, no. 8, pp. 879-886.