International Journal of Civil Engineering and Technology (IJCIET)
Volume 8, Issue 12, December 2017, pp. 339-345, Article ID: IJCIET_08_12_040
Available online at http://www.iaeme.com/ijciet/issues.asp?JType=IJCIET&VType=8&IType=12
ISSN Print: 0976-6308 and ISSN Online: 0976-6316

LOCATION PREDICTION OF ANONYMOUS TEXT USING AUTHOR PROFILING TECHNIQUE

Raghunadha reddy T
Associate Professor, Dept of IT, Vardhaman College of Engineering, Hyderabad

Gopi Chand M
Professor, Dept of IT, Vardhaman College of Engineering, Hyderabad

Hemanath K
Assistant Professor, Dept of IT, MLR Institute of Technology, Dundigal, Hyderabad

ABSTRACT

Author Profiling is used to predict the demographic characteristics of the authors like age, gender and location by analyzing their written text. Several researchers proposed different types of approaches to predict the gender and age of the authors from different types of corpuses. Few researchers concentrated on location prediction of the texts. In this work, we concentrated on prediction of the location of the authors. Most of the existing work in Author Profiling concentrates on extraction of various types of stylistic features to differentiate the writing style of the text. In this work a profile specific document weighted model is used to predict the location of the authors. In this model content based features were used to compute the weights of the documents specific to each profile group. These document weights were used to represent the vectors of documents for generating classification model. The obtained results were good when compared with exiting approaches for location prediction.

Key words: Location Prediction, Author Profiling, Profile specific Document Weighted model, Term Weight Measure, Document Weight Measure.

Cite this Article: Raghunadha reddy T, Gopi Chand and Hemanath K, Location Prediction of Anonymous Text Using Author Profiling Technique, International Journal of Civil Engineering and Technology, 8(12), 2017, pp. 339-345.
http://www.iaeme.com/ijciet/issues.asp?JType=IJCIET&VType=8&IType=12

1. INTRODUCTION

The text in the World Wide Web is increasing exponentially day by day mainly through blogs, reviews, twitter and other social media content. It was very difficult for the information analyst to analyze the text when the text was uploaded without any details. In this context, the researchers were looking for a solution to predict the details of the anonymous text. The Author

Profiling is one such technique, which is used to predict the demographic characteristics of the authors like gender, location, age, nativity language, educational background and personality traits of the texts.

The author Profiling techniques were used in different applications such as security, social media and business. In social media, the users created profiles with false information and they posted harassing messages and threatening messages to other users. The Author Profiling techniques were used to predict the correct details of the users by analyzing these harassing and threatening texts. The terrorist organizations send threatening mails to government agencies to convey information without specifying their correct details. In this context, Author Profiling is used to predict the details of the mail like from which country the mail came, the gender and age group of the person. In business point of view, the business analysts take decisions about their products by analyzing the reviews of the product given by the users. It was difficult for analysts when the reviews do not contain any known details. Here, the Author Profiling techniques were used to predict the gender, age and location of authors of reviews.

This paper is structured in 6 sections. The existing work in Author Profiling for location, age and gender prediction was described in section 2. The dataset characteristics and evaluation measures were explained in section 3. The Profile specific Document Weighted model was explained in section 4. The experimental results of PDW model presented in section 5. This paper concluded in section 6 with conclusions and future work.

2. LITERATURE REVIEW

In Author Profiling, most of the researchers proposed several types of stylistic features like word, character, syntactic, structural, readability, content and information retrieval features for predicting location, age and gender of the authors [1]. The major problem in Author Profiling approaches were the collection of corpus for experimentation. In Koppel, M. et al. [2], 566 documents were collected from the British National Corpus (BNC). They achieved an accuracy of 77.3% for gender prediction using 1081 features. Argamon, S. et al. [3] collected a blog posts of 19320 blog authors. The result of 76.1% accuracy is achieved for gender dimension by using both content based and stylistic features. It was observed that the style based features were most useful to discriminate the gender. In another work [4], they achieved better accuracy of 80.1% using 1502 features of content based and stylistic features on 37478 blogs.

Dominique Estival used [5] a corpus of 9836 e-mails of 1033 authors which contains English, Spanish, Arabic language e-mails. While predicting native language of authors they extracted 689 features of word based, character-level and structural features. They used many classifiers for this analysis but Random forest algorithm gave a overall accuracy of 0.8422. While predicting the education dimension, the bagging algorithm obtained an accuracy of 0.7998 by using the function words as features. The SMO machine learning algorithm gave an accuracy of 0.8113 for country dimension by using all features.

Wee-Yong Lim et al., used [6] TFIDF scores of words to find the rare or common words in the entire corpus. Seifeddine Mechti et al., computed [7] the ranked list of words, then group these words into classes according to their similarity. The TFIDF measure was used to calculate the scores of each class for each document to find the stylistic differences between male and female. Suraj Maharjan et al., used [8] word n-grams as features and TFIDF as the weighting measure. TFIDF scores of the word n-grams were used to filter the n-grams that were not been used by at least two authors.

Alonso Palomino-Garibay et al., experimented [9] on tweets corpus and represented each tweet with a bag of words in a vector space. TFIDF measure was used to assign a value to each word in a vector. Octavia-Maria S et al., used [10] the combination of type/token ratio and TFIDF scores of character n-grams. The TFIDF scores were extracted from scikit-learn's

TfidfVectorizer(). It was observed that this combination of features obtained good accuracies for Dutch and Spanish language and also observed that the best TFIDF scores were obtained for character level n-grams in the range 'n' value between 2 to 6.

Dang Duc, P. et al. experimented [11] with 73 Vietnamese bloggers posts of 3524. They extracted 298 features such as character based and word based features from the blogs corpus. It was observed that the word based features influence is more for gender prediction when compared with character based features and the classifier IBK obtained a good accuracy of 83.34% for gender prediction using combination of word and character based features. In another work [12], the researchers collected 1000 blog posts of Greek language and extracted 300 most frequent n-grams such as word n-grams and character n-grams and standard stylometric features. In their observation, the accuracy was increased to 82.6% for gender prediction using Support Vector Machines when longer sequences of character n-grams and word n-grams were used.

Shlomo Argamon used [13] the corpus of International Corpus of Learner English (ICLE) which is a culmination of non-native English speakers from various countries and the corpus was tested to predict the age, gender and native language. He also used essays of 251 psychology undergraduates at the University of Texas at Austin for neurotism prediction. They considered five sub-corpora namely Russian, Czech Republic, Bulgaria, French and Spanish from ICLE. They used 258 authors writings from each sub corpus to avoid class imbalance problems. While predicting the age, gender, nativity language and neurotism they observed that style based features gave an accuracy of 65.1%, content based features gave an accuracy of 0.823 and both style based and content based features together gave an accuracy of 0.793.

Edson R. D. Weren treated [14] test document as a query and training document set as search engine space for information retrieval. They used information retrieval features for gender and age prediction from social media texts. In [15, 16], Edson R. D. Weren experimented with more number of features on various corpus sets for profile identification including personality traits of the authors.

3 CORPUS CHARACTERISTICS AND EVALUATION MEASURES

3.1. Corpus Characteristics

The corpus for location prediction was collected from a web site www.tripadvisor.com. The researchers in Author Profiling faced many problems in collection of reviews. In this work, we take some special precautions in the collection of corpus. Some of them are, the reviews were considered which are written in English language only, the reviews were considered who gave the details of location in their profile and the reviews were considered which contains at least five sentences. The location corpus contains 4000 hotel reviews of different countries authors. We collected 400 reviews from each country to balance the corpus countrywide.

3.2. Evaluation Measures

The approaches for Author Profiling used recall, precision, F1-score and accuracy measure to evaluate the efficiency of their model for predicting demographic characteristics of the authors. In this work, accuracy measure is used to test the efficiency of our approach for location prediction. The accuracy is the ratio of the number of reviews correctly predicted their location to total number of reviews in the test corpus.

4. PROFILE SPECIFIC DOCUMENT WEIGHTED MODEL

The PDW model for location prediction is depicted in Fig 1.

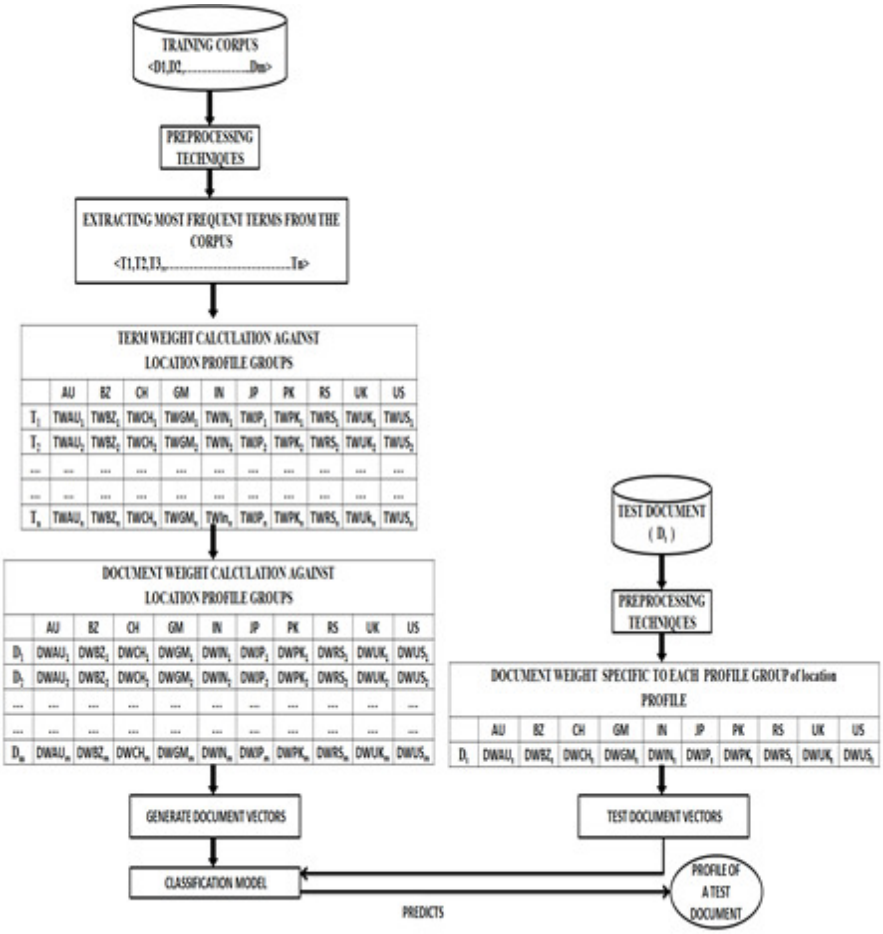

Figure 1 The Profile specific Document Weighted model for location prediction

The Profile specific Document Weighted (PDW) model was proposed in [17] for document representation. In this work, The PDW model is used for location prediction. In this model, {D1, D2 … Dm} is a set of documents in the corpus and {T1, T2… Tn} denotes the set of vocabulary terms. The hotel reviews corpus for location prediction contains documents of 10 location groups such as AUSTRALIA (AU), BRAZIL (BZ), CHINA (CH), GERMANY (GM), INDIA (IN), JAPAN (JP), PAKISTAN (PK), RUSSIA (RS), UNITED KINGDOM (UK), UNITED STATES OF AMERICA (US). $TWAU_n$ is the weight of the term T_n in the profile group of AU. $DWAU_m$ is the weight of document D_m in the profile group of AU.

In this model, first collect the corpus carefully. Then, apply preprocessing techniques like stop word removal stemming to extract interested features from the corpus. Identify most frequent terms which are occurred at least five times in the corpus. Compute the weights of these terms specific to all location groups of documents using term weight measure. Document weights were calculated specific to each location profile group of documents using the terms in that documents. Document weights were used to generate the vector representations of documents. Finally, these document vectors were used to generate the classification model using different machine learning algorithms.

In this model, the choice of term weight measure and document weight measure influence the accuracy of location prediction.

4.1. Term Weight Measure

The term weight measures assign suitable weights to the terms by considering different types of information of terms such as inner-document, intra-class and inter-class distribution of terms. Inner-document distribution means how many times the term occurred in a document. Intra-class distributions mean how the term is distributed within a specific class of documents. Inter-class distribution talks about the distribution of terms across the classes of documents. In this work, 10 countries corpus was used that means 10 classes of documents were used. In this work, SUTW measure [18] is used to compute the weights of terms specific to each profile group. The SUTW measure is represented in equation (1).

$$W(t_i, p_j) = \sum_{k=1, d_k \in p_j}^{m} \left(\frac{tf(t_i, d_k)}{tf(t_i, p_j)} \left[\frac{\log(d_{tk})}{0.8 * AVGUT_k + 0.2 * UT_k} \right] \right) \times \frac{a_{ij}}{(a_{ij} + b_{ij})} \times \frac{c_{ij}}{(c_{ij} + d_{ij})} \quad (1)$$

Where, $W(t_i, p_j)$ is the weight of the term t_i in profile group p_j. $tf(t_i, d_k)$, $tf(t_i, p_j)$ are the number of times the term t_i occurred in document d_k and the documents of profile group p_j respectively. UT_k is the number of unique terms in document d_k. $AVGUT_k$ is the ratio of the number of unique terms in document d_k to total number of terms in document d_k. a_{ij}, b_{ij} are the number of documents in profile group p_j which contain the term and do not contain the term t_i respectively. c_{ij} and d_{ij} are the number of documents of other than profile group p_j which contain the term and do not contain the term t_i respectively.

4.2. Document Weight Measure

The document weight measure is used to compute the weight of the document based on the terms in a document. In this work, a document weight measure is used proposed in [19]. This measure considers two factors to calculate the weight of a document. First factor is the weight of terms within a document and the second factor is the weight of the terms that were computed by using term weight measure. The document weight measure is represented in equation (2).

$$W(d_k, p_j) = \sum_{t_i \in d_k, d_k \in p_j} TFIDF(t_i, d_k) * W(t_i, p_j) \quad (2)$$

Where, $w(d_k, p_j)$ is the weight of document d_k in profile group p_j. TFIDF (Term Frequency and Inverse Document Frequency measure) is used to compute the term weight within a document.

The next section presents the results of PDW model for location prediction using various classifiers.

5 EXPERIMENTAL RESULTS

Table 2 The Accuracy of PDW model for location prediction using different Classifiers

Classifiers/ Number of Terms	SL	IBK	LOG	BAG	NBM	RF
1000	60.39	64.70	67.07	70.70	72.07	69.82
2000	62.41	65.54	68.34	71.54	73.34	71.33
3000	63.26	68.36	70.24	72.36	75.24	73.56
4000	63.83	71.79	73.12	73.79	75.12	76.21
5000	65.51	71.85	73.87	73.85	76.87	77.22
6000	65.80	73.20	75.64	75.20	77.64	79.76
7000	67.39	74.14	76.25	76.14	79.25	82.48
8000	68.47	75.91	78.18	78.27	80.36	83.49

By analyzing the corpus of location profile, it was observed that the content based features like the words they used to write a review differentiates the different countries authors. It was assumed that, in general the choice of words selected by the users of one country was same to write a review. With this assumption most frequent 8000 terms as features were extracted from the corpus. The experiment start with 1000 terms then increased to 8000 terms with an increasing of 1000 terms in each iteration. We obtained good results for location prediction when compared with existing approaches of Author Profiling for location prediction. Various classification algorithms such as Logistic (LOG), Simple Logistic (SL), IBK, Bagging (BAG), Random Forest (RF) and Naïve Bayes Multinomial (NBM) were used to generate the classification model.

The accuracies of location prediction in PDW model using term weight measures are shown in Table 2. The Random Forest classifier obtained highest accuracy of 83.49 for location prediction when compared with all other classifiers. The Naïve Bayes Multinomial classifier obtained an accuracy of 80.36% for location prediction. It was observed that in all classifiers the accuracies of location prediction was increased when the number of terms increases.

6 CONCLUSIONS

In this work, a Profile specific Document Weighted model is used to predict the location profile of the authors from hotel reviews corpus. The Random Forest classifier obtained highest accuracy of 83.49% for location prediction when most frequent 8000 terms were used as features. It was planned to increase the accuracy of location prediction by considering the semantic relationship between terms in a document.

REFERENCES

[1] T. Raghunadha Reddy, B.VishnuVardhan, and p.Vijaypal Reddy, "A Survey on Authorship Profiling Techniques", International Journal of Applied Engineering Research, Volume 11, Number 5 (2016), pp 3092-3102.

[2] Koppel, M., Argamon, S., Shimoni, A.R.: Automatically categorizing written texts by author gender. Literary and Linguistic Computing 17(4), pp. 401-412 (2002).

[3] Argamon, S., Koppel, M., Pennebaker, J. W., and Schler, J. (2009). Automatically profiling the author of an anonymous text. Communications of the ACM, 52(2):119.(2009).

[4] Schler, J., Koppel, M., Argamon, S., and Penebaker, J. "Effects of Age and Gender on Blogging". AAAI Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW), AAAI Technical report SS-06-03, (2006).

[5] Estival, D., Gaustad, T., Pham, S.B., Radford, W. and Hutchinson, B., "Author profiling for english emails", in Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING'07)., (2007), 263-272.

[6] Wee-Yong Lim, Jonathan Goh and Vrizlynn L. L. Thing, "Content-centric age and gender profiling", Proceedings of CLEF 2013 Evaluation Labs, 2013.

[7] SeifeddineMechti, Maher Jaoua,Lamia Hadrich Belguith, "Author Profiling Using Style-based Features", Proceedings of CLEF 2013 Evaluation Labs, 2013.

[8] Suraj Maharjan and Thamar Solorio, "UsingWide Range of Features for Author Profiling", Proceedings of CLEF 2015 Evaluation Labs, 2015.

[9] Alonso Palomino-Garibay, Adolfo T. Camacho-Gonzalez, Ricardo A. Fierro-Villaneda, Irazu Hernandez-Farias, Davide Buscaldi, and Ivan V. Meza-Ruiz, "A Random Forest Approach for Authorship Profiling", Proceedings of CLEF 2015 Evaluation Labs, 2015.

[10] Octavia-Maria S, ulea1;2 and Daniel Dichiu, Bitdefender Romania, "Automatic Profiling of Twitter Users Based on Their Tweets.", Proceedings of CLEF 2015 Evaluation Labs, 2015.

[11] Dang Duc, P., Giang Binh, T., Son Bao, P.: Author Profiling for vietnamese blogs. Asian Language Processing, 2009 (IALP '09), pp. 190-194. (2009).

[12] Dang Duc, P., Giang Binh, T., Son Bao, P.: Authorship Attribution and Gender Identification in Greek Blogs. 8th International Conference on Quantitative Linguistics (QUALICO), April 26-29, 2012, (2012).

[13] Argamon, S., Koppel, M., Pennebaker, J. W., and Schler, J. (2009). Automatically profiling the author of an anonymous text. Communications of the ACM, 52(2):119.(2009).

[14] Edson R. D. Weren, Viviane P. Moreira, and José P. M. de Oliveira, "Exploring Information Retrieval features for Author Profiling", Proceedings of CLEF 2014 Evaluation Labs, 2014.

[15] Edson R. D. Weren, Viviane P. Moreira, and José P. M. de Oliveira, "Using Simple Content Features for the Author Profiling Task", Proceedings of CLEF 2013 Evaluation Labs, 2013.

[16] Edson Roberto Duarte Weren, "Information Retrieval Featuresfor Personality Traits", Proceedings of CLEF 2015 Evaluation Labs, 2015.

[17] Raghunadha Reddy T, Vishnu Vardhan B, Vijayapal Reddy P, "Profile specific Document Weighted approach using a New Term Weighting Measure for Author Profiling ", International Journal of Intelligent Engineering and Systems, Nov 2016, 9 (4), pp. 136-146. DOI: 10.22266/ijies2016.1231.15

[18] Raghunadha Reddy T, Vishnu Vardhan B, Vijayapal Reddy P, "A Document weighted Approach for Gender and Age Prediction", International Journal of Engineering, Volume 30, No. 5, 2017, PP. 647-653.

[19] Dheyaa Jasim Kadhim, Tagreed Mohammed Ali and Faris A. Mustafa, Location Prediction In Cellular Network Using Neural Network, Volume 4, Issue 4, July-August (2013), pp. 321-332, International Journal of Computer Engineering and Technology (IJCET)

[20] Raghunadha Reddy T, Vishnu Vardhan B, Vijayapal Reddy P, "Author profile prediction using pivoted unique term normalization", Indian Journal of Science and Technology, Vol 9, Issue 46, Dec 2016.