International Journal of Database Theory and Application
Vol.6, No.6 (2013), pp.85-94
http://dx.doi.org/10.14257/ijdta.2013.6.6.08

Predicting Age Range of Users over Microblog Dataset

Lizhou Zheng, Kaifan Yang, Yongbo Yu and Peiquan Jin*

School of Computer Science and Technology
University of Science and Technology of China, 230027, Hefei, China

jpq@ustc.edu.cn

Abstract

In this paper, we present the idea and methodologies on predicting the age span of users over microblog dataset. Given a user's personal information such as user tags, job, education, self-description, and gender, as well as the content of his/her microblogs, we automatically classify the user's age into one of four predefined ranges. Particularly, we extract a set of features from the given information about the user, and employ a statistic-based framework to solve this problem. The measurement shows that our proposed method incorporating selected features has an accuracy of around 71% on average over the training dataset.

Keywords: *Microblog; User classification; Age prediction*

1. Introduction

Nowadays, microblogging services, such as Twitter and Sina Weibo, play an important role in our daily life. The rapid development of microblogging platform has introduced lots of challenging issues in many research areas including event detection, sentiment analysis, user classification, and so on, because of the rich information generated by the large number of microblogging users.

The study of this paper is based on a dataset collected from Sina Weibo [1]. We aim at predicting the age ranges of given users, which is an important attribute of a microblog user. The dataset contains over one million users and their basic information including user ID, tags, job, personal description, age, gender, and education level, as well as the content of the microblogs (in which the words were mapped into unique numbers) posted by them. Our work is to build a system for predicting users' age range which is predefined as [0-18], [18-24], [24-35] and above 35.

This paper describes the method we present to solve the problem described above. We leverage the observable information and extract a set of features that are helpful in predicting the age range of user. Those features include: linguistic user base features (content of tags, job, education, and self-description), content of user microblogs, and numeric features (the posted count of user microblogs, count of words in users' tag, count of words in users' job, etc.). We finally employ a statistic-based method to accomplish this task.

Our experimental results show that our method has a classification precision of around 71% on average over the given dataset.

The rest of the paper is organized as follows. In Section 2 we briefly summarize the related work. Section 3 describes the general framework of our method. Section 4 gives the experimental result, and conclusions as well as future work are in Section 5.

ISSN: 2005-4270 IJDTA

2. Related Work

Since microblog platforms attract large amount of register users and contains rich information, researchers are now paying more and more attention on mining useful knowledge from microblog platforms such as Twitter and Sina Weibo. Those research mainly focus on two typical problems, one is knowledge extraction on microblogs. Since microblog data is multi-domain and short in length, it needs new methods to deal with those noisy but informative data, the real-time attribute of microblog data is another challenge for knowledge extraction researchers. The other typical problem of researches on microblog is how to mine user-based information such as the analysis of social network and user information extraction. For the two typical problems, there are hot research areas such as event detection [15, 16, 8, *etc.*,], sentiment analysis [17, 18, *etc.*,], user classification and user-based attributes mining. The latter two research areas are most relevant to our work.

Classify users in microblog platform such as Twitter and Sina Weibo is a meaningful work and thus attracts more and more researchers to participate in. In those works researchers always define several user-based categories in different views then classify a user into one of them. In the literature [9], the researchers classified Twitter users into many fields such as political orientation or ethnicity. In [10], Twitters users were classified into three kinds of users (*i.e.*, information seeking users, information sharing users and friend making users) by analyzing users' social network. In [11], three classes for a Twitter user were defined, which are human, bot and cyborg then employed mainly two kinds of features (*i.e.*, frequency of posting tweets and content of tweets) to perform classification. Twitter user accounts can also be divided into open accounts and closed accounts [13], in which the open accounts refer to users who publish information to general public and their intentions is to promotion products, services or themselves, while the closed accounts are those who mainly post for their daily lives. In [13], the classification was conducted based on user profiles and their followers' distribution.

Another research area which is related to our work is user-based attribute detection. Those user-based attributes include users' gender, location, interests and any other items which related to a user. Mining user-based attributes is first done on normal text, blogs, e-mails, *etc.*, For those traditional well written text sources, there are previous works which aim at detecting user attributes based on user communication streams. For example [3, 4, 5 and 6] detect user gender from text, blogs, reviews and e-mails respectively. As to microblog platform, [12] extracts the location of a Twitter user based purely on the content of a user's tweets, the work estimates k possible locations for each user in descending order of confidence. [7] uses some simple features such as n-gram model to classify latent user attributes (*e.g.*, location of origin, age, *etc.*,) in Twitter. [14] utilizes user-centric features and graph-based features for automatic user classification and profiling. [15] produces a method to predict the personality of Twitter users. They utilize the user profile information and employ machine learning techniques to finish the work.

User-based attributes play an important role in describing a microblog user. In this work we focus on extracting an important attribute of a user, i.e. the age of a user. Here the age of a user is divided into four spans, *i.e.*, [0-18], [18-24], [24-35] and above 35. Different from the works we mentioned above, in our dataset words are mapped into unique numbers, so we could not obtain any domain information in our work. In our work we extract features based on user base information (*e.g.*, user tags, user jobs, *etc.*,) as well as user microblog content and employed a statistic-based model to finish the task.

3. The General Framework for User Age Prediction

The framework of our method is shown in Figure 1.

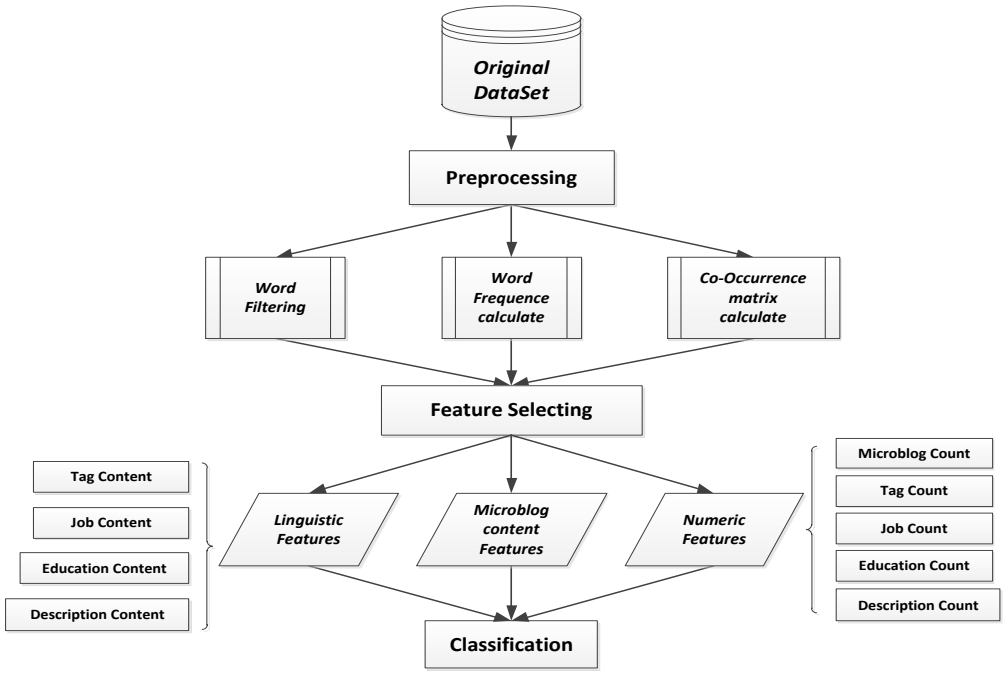

Figure 1. The Framework of our Method

We first perform preprocessing on the original dataset to obtain statistic information about the dataset. Then, we extract three types of features on those preprocessed data. Finally, we make classification on the dataset.

3.1. Preprocessing

As the original dataset is quite large and contains some noise information, we first make preprocessing on the dataset. The preprocessing includes word filtering, term frequency calculation and co-occurrence matrix calculation.

Word Filtering. As we mentioned above, the Chinese words in microblogs are preprocessed and mapped into unique numbers in the dataset, so we could not manually construct a stop word list to filter the stop words. Therefore, we simply abandon the words which have a significant high term frequency (*e.g.*,"77" in the dataset) or those rarely appearance (*e.g.*,"63385"). The word filtering procedure is performed on each linguistic item (user base information and microblog content) respectively.

Term Frequency Calculation. In order to extract linguistic features, we calculate frequency for both the unigrams and the bigrams of the whole training dataset. For each word w_i, we calculate the term-frequency TF_i (shown in Formula 3.1), as well as its term-frequency under each age range r_c, which is represented as $TF_{i,c}$ (shown in Formula 3.2). Here, we introduce the total term-frequency of age range r_c for smooth.

$$TF_i = count(w_i) / \sum count(w) \tag{3.1}$$

$$TF_{i,c} = \frac{count(w_i, r_c)}{count(w_i) * \sqrt{\frac{\sum count(w, r_c)}{\sum count(w)}}} \tag{3.2}$$

For the bigram case, we calculate the bigram frequency under the whole corpus and each range for each consecutive two-word-pair in users' microblog content and self-description content, as to tag, job and education, we calculate the frequency for each pair of words no matter they are consecutive or not.

Co-Occurrence Matrix Calculation. Since two words that simultaneously appear in many microblogs or user item fields may provide similar information, we pre-calculate the term co-occurrence matrix which is to be used in the next process of classification. The co-occurrence matrixes include the user microblog term matrix *MM*, the user tag term matrix *TM*, the user job term matrix *JM*, the user education term matrix *EM*, and the user self-description term matrix *DM*. Note that elements in those matrixes are in probabilistic format that range from zero to one.

3.2. Feature Selection

In this section we describe in detail how we derive a set of features from the dataset after preprocessing. The extracted features are finally classified into three types:

(1) Linguistic user base information features, such as content of tags, jobs, education, and self-description.

(2) User microblog features. This feature is based on the content of user microblogs.

(3) Numeric features, such as user microblog count, count of words in user tag, count of words in user job, and so on.

3.2.1. Linguistic User-Based Features & Microblog Content Features

The user base information provided by the dataset can help characterize a user and predicting a user's age range. People with different ages may use different lexical items in their personal information fields. For example those users with "university" in their education field may not be under 18 years old, and those with "youngster" in their tag field may not be above 35 years old. Note that since in this dataset words were mapped into unique numbers, introducing domain knowledge about words is not practicable in our work, so we employed statistic method to extract the linguistic user base features.

Typical Words. Since the count of words in each item of user base information as well as user microblog content is quite large (with over 50K different words in tag item, over 30K different words in job item, *etc.*,), and some words have strong ability to distinguish different classes but many are not. Here we select typical words which could be used to well distinguish different age ranges for each item by employing the entropy measurement.

Given n categories described as four age ranges r_c (c = 1, 2, 3, 4), the term-frequency TF_i for each word w_i and the term-frequency under each categories $TF_{i,c}$ we obtain from preprocessing step, we employed the entropy statistic for word w_i as Formula 3.3.

$$entropy(w_i) = -\sum_{c=1}^{4} TF_{i,c} * \log TF_{i,c} \tag{3.3}$$

The entropy measurement could help to decide whether a given word have strong ability in distinguishing different categories. We calculate the entropy for each words in different user base information item as well as user microblog content, then we select typical words respectively by both entropy and word frequency that satisfied Formula 3.4.

$$entropy(w_i) < \sigma \ \&\& \ TF_i \ in \ top\text{-}K \tag{3.4}$$

Here we tuned the parameter σ and K by experiment.

Other Words. Since the typical words we obtained above are just part of the whole word set, many users' user base information does not contain any of the typical words. In this case, we need to employ some process on those non-typical words.

Suggest we have got M typical words for an item t, M is far less than the size of the whole word set, and a given user has a word list $wlist_t = \{w_1, w_2, \ldots, w_n\}$ for item t, where none of w_i is a typical word, we employ two methods to process $wlist_t$, namely the word mapping method and the individual probabilistic method.

- **Word Mapping.** As we describe in Section 3.1, we calculate the term co-occurrence matrix for all the user base information items. Here we just map each word in $wlist_t$ into each typical word by their co-occurrence value. After that we could obtain an M-element vector of which each element represents the co-occurrence frequency of the word with a typical word and values from zero to one.
- **Individual Probabilistic.** Here we do not map those non-typical words into a typical words co-occurrence vector, instead, we calculate the combination probability of all non-typical words under each class c respectively by their term frequency and make use of the result in further classification step, as shown in Formula 3.5.

$$prob^{c}_{non\_typical\_word} = \prod_{w_i \in non\_typical\_words} TF_{i,c} \tag{3.5}$$

3.2.2. Numeric Features

People of different ages may have added different counts of words in their user base information, for example the ratio of users whose word count in job item is zero in age range [0-18] is larger than users in other age ranges. This is because those users may be too young to get a job. So we extract the numeric features that could be listed as word count in microblog count, tag, job, education and self-description. We manually partition the total count value set into spans (*e.g.*, count 0 as span 0, count 1,2,3 as span 1, *etc.*,) and calculate the count span frequency $CSF_{i,c}$ of span i under each class c as features, as shown in Formula 3.6.

$$CSF_{i,c} = \frac{count_i(u, r_c)}{count_i(u) * \sqrt{\frac{count_c(u)}{\sum count_c(u)}}} \tag{3.6}$$

Here $count_i(u)$ is the count of users who have an item count in count span i, and $count_i(u, r_c)$ is the count of users who have an item count in count span i and their age range is c. The frequency of age span c is used for smooth.

3.3 Classification

Here we introduce our method for classification. For each user to be classified, we assign a score for each of the four age spans and choose the range with a max score as the result. The score contains two parts (as shown in Formula 3.7): the linguistic feature score *ling_score* and the numeric feature score *num_score*.

$$\max_{c}(ling\_score_{c}^{\alpha} * num\_score_{c}^{1-\alpha}) \tag{3.7}$$

The parameter α is tuned as 0.8 in the experiment.

Ling_Score. As we introduced in Section 4, we got *M* typical words for each item and tried two methods to process the non-typical words. For the word mapping method, we got an *M*-element vector with its *i*-th element represents the weight of the *i*-th typical word. So the linguistic feature score for non-typical words *nt_ling_score* under age range *c* is defined in Formula 3.8.

$$nt\_ling\_score_{(word\ mapping)\ c} = \prod_{i} TF_{i,c} * weight(i) \tag{3.8}$$

For the individual probabilistic method, we got the combination probability of non-typical words under each age range *c*, so the *nt_ling_score* could be represented as Formula 3.9.

$$nt\_ling\_score_{(individual\ probability)\ c} = prob^{c}_{non\_typical\_word} = \prod_{w_i \in non\_typical\_words} TF_{i,c} \tag{3.9}$$

We test the two methods while experiment and find the individual probability method performs better, so we employ the individual probability method to finish the task, we will show the details of our experiment in Section 4.

In both cases, we get typical word score as well, for each user to be classified, we obtain the typical words appeared in each item of user base information, and the score of typical words *t_ling_score* is defined as Formula 3.10.

$$t\_ling\_score_{c} = \prod_{w_i \in typical\_words} TF_{i,c} \tag{3.10}$$

At last we combine the score of typical words and non-typical words as the *ling_score* (shown in Formula 3.11), the parameter β is tuned as 0.6 by experiment.

$$ling\_score_{c} = t\_ling\_score_{c}^{\beta} * nt\_ling\_score_{c}^{1-\beta} \tag{3.11}$$

Num_Score. The score of numeric feature in our method is also in probability format. We first get the count span of different items and calculate the combination probability of different items under each age range *c*, as shown in Formula 3.12.

$$num\_score_{c} = \prod_{items} CSF_{i,c} \tag{3.12}$$

Here $CSF_{i,c}$ is the count span frequency of span *i* over age range *c*.

Apart from the feature scores we describe above, we employ some strong *n*-gram rules to enhance the performance of the result. Those rules include the unigrams and bigrams which only occur in one of the four age range. We give higher scores for those strong *n*-gram rules.

4. Performance Evaluation

In this section we show the performance of our method. We first describe the dataset we use then show our experiment results when using different features and under different values of parameters.

4.1. Dataset

The dataset we used is collected from Sina Weibo, the most popular microblog service in China. There are 1.12 million users in the dataset, along with their user profile information and microblog post information. The user profile information includes user ID (which is anonymized), tags, jobs, personal description, gender, education and age. Microblog post information contains microblogs which posted by a specific user (retweet content and name mentioned is removed, content with Chinese words are mapped into unique numbers).

4.2. Results

Here we describe our experiment results. As we shown in Section 3.2 and Section 3.3, we extract linguistic features which contain scores for typical words and scores for non-typical words. For typical words, the count of typical words K is crucial to the result. For non-typical words, we test two methods to deal with this feature, i.e. the word mapping method and the individual probabilistic method. And for both set of words, we use unigram and bigram to represent the features. So in our experiments, we test mainly three groups of methods: 1) the results under different values of K; 2) the results for word mapping method and individual probabilistic method when dealing with non-typical words; 3) the result when utilize bigram feature and only utilize unigram feature. The results of those experiments are shown in Table 1 to Table 3, and Figure 2 to Figure 3. Due to the fact that there are many users without typical words given in their user base information and microblog content, and some users even do not publish any microblogs, we divide our dataset into three categories, *i.e.*, users with typical words, users without typical words and users without microblogs. We list the results under different categories. Figure 2 shows the percentage of users with typical words, without typical words and without microblogs when we choose different values of K, we could discover that with the decrease of K, the percentage of users with typical words is decrease while users without typical words is increase, that could lead to the change of precision under each category as well as the overall precision. Note that the result we list for each category of the users does not include the strong rules, when employing strong rules, the overall precision would have some increase.

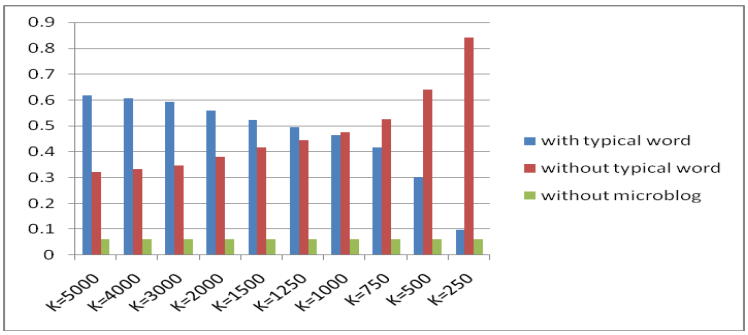

Figure 2. Percentage of Users With/ Without Typical Words and Without Microblog

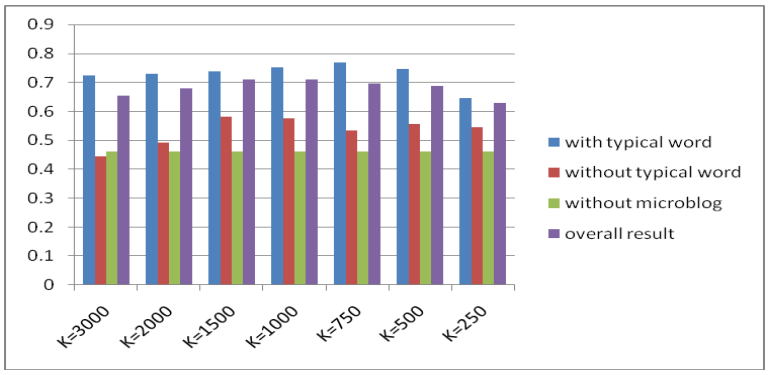

Figure 3. Precision Under Different Values of *K*

Table 1. Results When Utilize Word Mapping and Individual Probabilistic Method (*K*=1500)

Metric \ Case	Without microblogs	Without typical words	With typical words	Overall
Percentage	6.04%	38.3%	61.7%	100%
Precision When Utilize Word Individual Probabilistic Method	46.2%	58.2%	73.5%	71.2%
Precision When Utilize Word Mapping Method	46.2%	55.7%	69.6%	67.8%

Table 2. Results When With/ Without Bigram (*K*=1500)

Metric \ Case	Without microblogs	Without typical words	With typical words	Overall
Percentage	6.04%	38.3%	61.7%	100%
Precision With Bigram	46.2%	58.2%	73.5%	71.2%
Precision Without Bigram	46.2%	56.8%	71.4%	69.4%

Table 3. Best Performance of our Method

Metric \ Case	Without microblogs	Without typical words	With typical words	Overall
Percentage	6.04%	38.3%	61.7%	100%
Precision	46.2%	58.2%	73.5%	71.2%

As shown in Figure 3, the best result of users with typical words is when *K*=750, but since the percentage of users with typical words is relatively small, it's not the best choice of *K* for the overall result. Here we choose *K*=1500 for our number of typical word under each item. Table 1 and Table 2 show the result of employing different methods for feature representation. Here we utilize individual probabilistic method and employing bigram feature since they show higher precision on the dataset.

Table 3 shows our best performance, the overall precision of our method is around 71%, and the performance in the category that the typical words are given is even higher, *i.e.*, 73.5%.

5. Conclusions

In this work we employed a statistic-based framework to solve the problem of user age span classification. For a given user with his/ her user base information and microblog content, we predict the age span of the user which could be in [0-18], [18-24], [24-35] and above 35 years old. For our best performance, we could obtain an overall precision of around 71%. By testing for different values of parameters and different methods of feature representation, we have some conclusions as follows:

(1) The number of typical words could affect the classification result, so choosing a proper number is important.

(2) When dealing with non-typical words, an individual probabilistic method could obtain better result compared with word mapping method. That is because the performance of mapping into typical words depends on the quality of typical words.

(3) By utilize bigram feature, we could obtain better result since it contains more semantic information. The weakness of our method lies in that for users whose user base information contains no typical words and users who do not post microblogs, the precision of age prediction is relatively low, so in our future work, we may focus on extracting more features to deal with users of these two cases.

Acknowledgements

This work is supported by the National Science Foundation of China (no. 71273010 and 61379037), and the National Science Foundation of Anhui Province (no. 1208085MG117).

References

[1] http://twitter.com.
[2] http://t.sina.com.cn.
[3] S. Herring and J. Paolillo, "Gender and genre variation in weblogs", Journal of Sociolinguistics, vol. 10, no. 4, **(2010)**, pp. 439-459.
[4] J. Burger and J. Henderson, "An exploration of observable features related to blogger age", Proc. of AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, **(2006)**, pp. 15-20.
[5] J. Otterbacher, "Inferring Gender of Movie Reviewers: Exploiting Writing Style, Content and Metadata", Proceedings of CIKM, **(2010)**.
[6] N. Garera and D. Yarovsky, "Modeling latent biographic attributes in conversational genres", Proceedings of CIKM, **(2007)**.
[7] D. Rao, Y. D., A. Shreevats and M. Gupta, "Classifying Latent User Attributes in Twitter", Proceedings of SMUC-10, **(2010)**, pp. 710-718.
[8] T. Hua, F. Chen, L. Zhao, C.-T. Lu and N. Ramakrishnan, "STED: semi-supervised targeted-interest event detection in in twitter", Proc. Of KDD, **(2013)**, pp. 1466-1469.
[9] M. Pennacchiotti and A. Popescu, "A Machine Learning Approach to Twitter User Classification", Proc. Of ICWSM, **(2011)**, pp. 281-288.
[10] A. Java, X. Song, T. Finin and B. Tseng, "Why We Twitter: Understanding Microblogging Usage and Communities", Proc. Of SNA-KDD, **(2007)**, pp. 56-65.
[11] Z. Chu, S. Gianvecchio, H. Wang and S. Jajodia, "Who is Tweeting on Twitter: Human, Bot, or Cyborg?", Proc. Of ACSAC, **(2010)**, pp. 21-30.
[12] Z. Cheng, J. Caverlee and K. Lee, "You are where you tweet: A Content-based Approach to Geo-locating Twitter Users", Proceedings of CIKM, **(2010)**.

[13] L. Yan, Q. Ma and M. Yoshikawa, "Classifying Twitter Users Based on User Profile and Followers Distribution", Proceedings of DEXA, **(2013)**, pp. 396-403.
[14] M. Pennacchiotti and A.-M. Popescu, "Democrats, republicans and starbucks afficionados: user classification in twitter", Proceedings of KDD, **(2011)**, pp. 430-438.
[15] T. Sakaki, M. Okazaki and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors", Proceedings of WWW, **(2010)**.
[16] J. Weng and B. Lee, "Event Detection in Twitter", Proc. Of ICWSM, **(2011)**.
[17] Y. Hu, F. Wang and S. Kambhampati, "Listening to the Crowd: Automated Analysis of Events via Aggregated Twitter Sentiment", Proc. Of IJCAI, **(2013)**.
[18] K.-L. Liu, W.-J. Li and M. Guo, "Emoticon Smoothed Language Models for Twitter Sentiment Analysis", Proc. Of AAAI, **(2012)**.

