Recursion identify algorithm for Gender Prediction with Chinese names

Hua Zhao
School of Mathematical and Computer Sciences
Heriot-Watt University
Edinburgh, UK
hz103@hw.ac.uk

Fairouz Kamareddine
School of Mathematical and Computer Sciences
Heriot-Watt University
Edinburgh, UK
f.d.kamareddine@hw.ac.uk

***Abstract*—Biological gender is a significant feature in representing a person. The literature is rich in work on gender prediction using people's names. Researchers have investigated the gender prediction based on names in a number of different languages. However, we found that there is low accuracy with the existing gender prediction systems on predicting gender using Chinese names. Since Chinese names have different types, the gender prediction of names needs a method for identifying each type of the Chinese names. This paper proposes a recursion identify algorithm for identifying Chinese names in Hanyu Pinyin. We will also display a method for identifying different types of Chinese names. Those methods can get the highest accuracy in predicting gender using Chinese names among all the gender prediction systems of names.**

***Index Terms*—Data Science, Machine Learning.**

I. Introduction

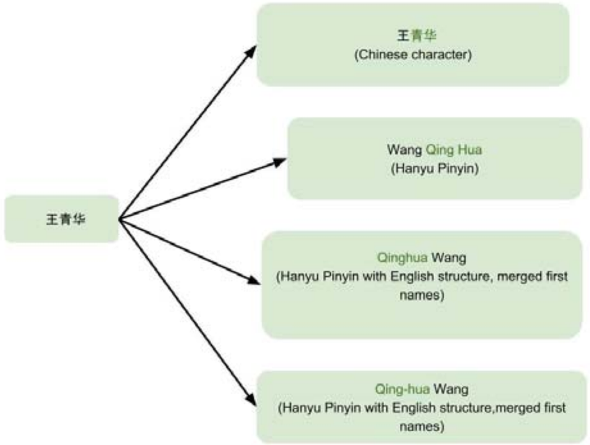
A name is a crucial information of a person. Gender is an essential data of people. This necessary personal information is valuable in many areas' research, such as in biology, psychology, sociology.

Gender prediction using people's names has been popularly used. There are many existing gender prediction systems with people's names in different languages [1], [6], [9], [17]. However, we found that these exsiting systems have low accuracy on predicting gender with Chinese names.

As we know, many Chinese names have been translated into Hanyu Pinyin Romanization as they can be understood globally. Hanyu Pinyin Romanization uses Roman letters, and it depends on the pronunciation of Mandarin Chinese.

For Chinese names, Han people's names normally contain first names and surnames. Han people's first names could be classified into two first names or one first name. For example, ”ChenNing Yang ” is a Chinese name where the first name is Chen Ning which is officially displayed as Chenning. It is more easy for people to understand this full name's first name and last name. Whereas, in Chen Ning Yang, it is difficult for people to know which is the surname, since Chen and Yang can both be surnames for Han people's names. Fig 1 displays the different representation of one example of a Chinese name.

Fig. 1. The different representation of an example Chinese name


王青华
王青华
(Chinese character)
Wang Qing Hua
(Hanyu Pinyin)
Qinghua Wang
(Hanyu Pinyin with English structure, merged first names)
Qing-hua Wang
(Hanyu Pinyin with English structure,merged first names)

In this paper, our purpose is using a recursion identify algorithm for Gender Prediction of names in order to identify and classify Chinese names in Hanyu Pinyin. We will incorporate this algorithm in a method for predicting gender using different types of Chinese names.

Our contributions are:

1) A new recursion identify algorithm for identifying Chinese names in Hanyu Pinyin.
2) A new method of processing different types of Chinese names on gender prediction.
3) Increasing the accuracy on predicting gender with Chinese names.

In section II, we describe the related work and the reason for building a Recursion identify algorithm. In section III, we display our new algorithm and a new method, apply it to a gender predicting system in detail. In section IV, we describe the data for training and testing of our new algorithm in detail. In section V, we outline the experiments' results of testing the system. We show some results on testing the accuracy of our algorithm. In section VI, we conclude and give some future work.

II. Related Work and need for new algorithm

A. Existing gender prediction systems of names

Generize [17], Gender API [1], Namsor [6], Name API [9] and Ngender [2] are popular gender prediction systems.

Genderize is an API that can predict gender using people's first names [17]. It can predict in 89 languages of people's first names [17]. This API can be used for predicting names at a limit of 1000 per day [17].

Gender API is a gender prediction API which uses people's full names to predict their gender [1]. It supports 178 countries' languages of names [1]. This API can do names prediction with 500 names per month for free [1].

Namsor is a classification software that can predict gender in 1000 names for free every month [6]. This software can classify all languages names [6].

Name API is a functional name classification web API [9]. It supports a large number of languages to do name gendering on predicting gender using people's full names [9]. This API can predict 1000 names on gender prediction for free each day and 10000 names prediction for free each month.

Ngender is a gender prediction that only works with Chinese names [2]. Although it can support unlimited number of names for free gender prediction, it does not have Hanyu Pinyin name prediction.

Table I shows the results of example Chinese names given by these existing gender prediction systems. The last row shows the real gender of each name.

TABLE I
The results of example names from exisiting gender prediction systems of names

Names Systems	Xu Zhang	Qinghua Wang	王青	赵金标
Generize	Female	Unknown	Male	Unknown
Gender API	Male	Male	Male	Male
Namsor	Unknown	Unknown	Unknown	Male
Name API	Unknown	Male	Male	Unknown
Ngender	Unknown	Unknown	Male	Male
Real Gender	Female	Female	Female	Male

Table II shows the output labels of gender prediction of names in each API.

TABLE II
Output labels of gender prediction APIs with names

Genderize	Male	Female	Null	
Gender API	male	female	unknown	
Namsor	male	female	unknown	
Name API	Male	Female	Neutral	Unknown
Ngender	male	female	unknown	

Generize API only can input people's firstname to predict gender. It predicts a limited number of people's first names with two first names in Hanyu Pinyin and Chinese language. Gender API can predict gender using people's full name. However, the results are low accuracy on predicting Chinese names. Namsor can predict gender with all languages' names. However, it cannot identify two first names in Chinese names, and users have to classify all the input names into first name and surname. Name API can identify different types of Chinese names. But the accuracy of the prediction results with Chinese names is deficient. Ngender can predict gender using Chinese names. However, it cannot predict Hanyu Pinyin in Chinese names.

In an earlier paper [20], we implemented a gender prediction tool of names, that can predict Chinese names and English names simultaneously for an unlimited number of names. However [20] could not deal with the Hanyu Pinyin romanization. Due to the nature and development of first, second and last names, we noted that a recursive method is needed.

In this paper, we propose a Recursion identify algorithm for identifying and classifying two first names in Chinese names. We give such an algorithm for processing different types of Chinese names in the next section.

B. Recursion

Recursion is a method for solving problems. It uses a breakdown method to process a problem into a similar subproblem and continues the process of breakdown method until the problem would be solved [7], [21].

Many articles show that robots use recursion for identifying the holes/hulls within the given distribution to prevent the obstacle [24], [25]. Guillaume et al. [26] use different versions of Kleene's second recursion theorem to classify Viruses. This method can gain the solutions of fixed point equations.

We propose to use recursion method for processing Chinese names in Hanyu Pinyin for gender prediction. Therefore, the gender prediction system can understand and identify different types of Chinese names.

III. The Recursion identify algorithm for gender prediction

In this section, we will describe our recursion identify algorithm. It works for gender prediction to identify when relevant the two first names of Chinese. This algorithm is used in systems that could identify different types of Chinese names.

Here we define $n + 1$ as the number of letters in a cycle input string of names. A normal input name can be defined as

$$N = [a_0, a_1, ..., a_n]$$

N is an input string of name and a_i where for $0 \leqslant i \leqslant n$ is the letter of each character in a string name. We have available, a list L of first names. Until a successful match against list L is found, we split N into N_1, N_2 where N_1 consists of the first i characters and N_2 consists of the second character, for $1 \leqslant i \leqslant n$. We start with $i = 1$ because a first name can never be empty but second first name can.

For example, a string can be like $N = [a_0, a_1, a_2]$, so it would work in the recursion identify algorithm as follow,

$$N = ([a_0], [a_1, a_2]), i = 1$$

$$N = ([a_0, a_1], [a_2]), i = 2$$

$$N = ([a_0, a_1, a_2], []), i = 3$$

Algorithm 1: Recursion Identify algorithm

```
List of names L
Input  : first names $N = [a_0, ..., a_n], n \geqslant 0$
Begin i =1
while  $i \leqslant n$ do
    firstname = $[a_0, ..., a_{i-1}]$;
    secondname = $[a_i, ..., a_n]$;
    if firstname and secondname $\in L$ then
        return firstname;
        return secondname;
    else
        $i = i + 1$;
    end
    if  sucess then  return firstname;
    return secondname;
    else
        print Unknow
    end
end
```

Algorithm 1 shows the process of the Recursion Identify algorithm. If the processed strings match with the training datasets which contained in list L , then it would process to output the result of gender in the system.

Fig 2 displays an example of how the Recursion Identify algorithm works with the Hanyu Pinyin name 'Qinghua Wang'. We can see that this name has been classified within the two first names. The algorithm can identify the correct first names and then send these first names to our system to predict gender.

The second algorithm (Algorithm 2) displays the process of how the recursion identify algorithm works with the gender classifier in our system to classify gender within Chinese names. Here, G is the result of the gender prediction, gender. We set the output as male, female, unisex and unknown. Note that N is the input name. We use the gender prediction algorithm of names in [20]. The authors use the Naive Bayes conditional probebility model which assigns probabilities to name instance using the following formula [2], [20]:

$$P(G \mid N) = P(G) * P(N \mid G)/P(N).$$

Fig 3 shows the process of the gender prediction classifier with different types of Chinese names.

Fig. 2. Demo of the Recursion Identify algorithm

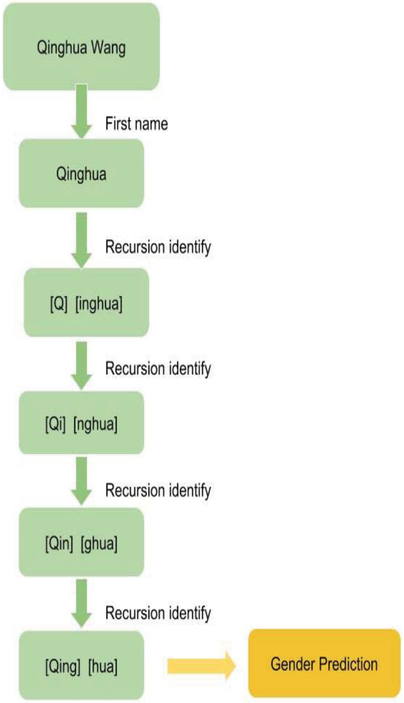
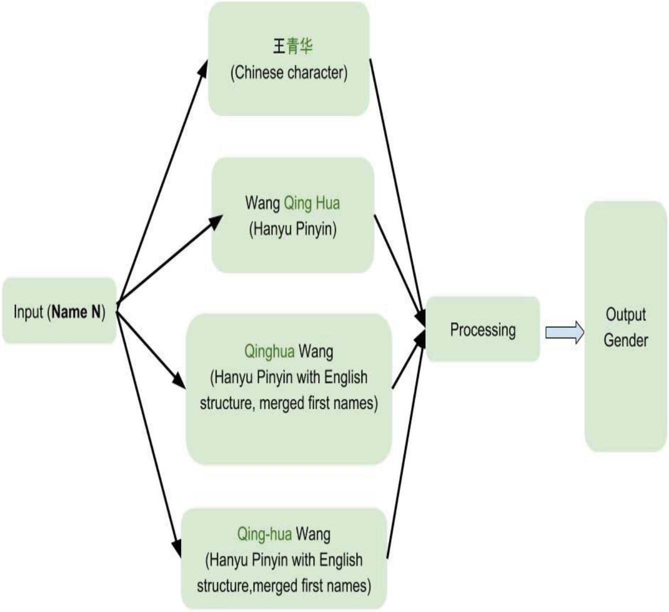


Fig. 3. Demo of the Gender Prediction with different types of Chinese names



Algorithm 2: Recursion Gender Prediction within different types of Chinese names

```
Input  : Different types of Chinese names N
Output: Gender G
 List of names L
while $G = None$ do
    if $N$ in Chinese character then
        continue
    else if $N \in L$ then
        continue
    else
        $n \geqslant 0$
        Recursion Identify Algorithm(N)
    end
    $P(G \mid N) = P(G) * P(N \mid G)/P(N)$
    if sucess then then print Male or Female or
     Unisex;

    else
        print Unknow
    end
end
```

IV. Data

A. Training Dataset

For testing our Recursion Gender Prediction algorithm and processing the gender prediction of names, we need a training dataset.

We collected a database of Chinese characters from Ngender [2]. This database has the names of Chinese characters and their frequencies. We used this training database for predicting Chinese characters.

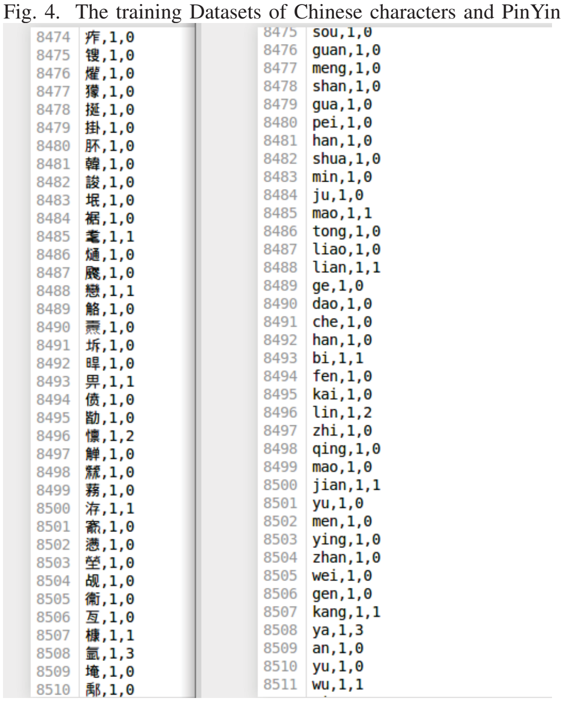
We processed 9441 Chinese names into Hanyu Pinyin to extend the training dataset for our tool. We used PinYin API [15] and Google translator [5] to translate those Chinese names into PinYin. Fig 4 shows two training datasets in our gender prediction system. They are Chinese characters training dataset and Hanyu PinYin training dataset.

On processing the Hanyu PinYin training dataset, we found that lots of Chinese characters have same characters in Hanyu Pinyin. Fig 5 and 6 show an example of this problem. In fig 5, we can see that there are two different Chinese names, but they have the same character in Hanyu Pinyin as fig 6 shows. Therefore, we used a method to solve this problem, see fig 7. On processing the Hanyu Pinyin training dataset, we combined all the same Hanyu Pinyin characters. Fig 8 displays the final dataset on processing the Hanyu Pinyin character 'Yong'.

B. Testing Data

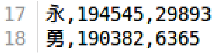
In our research, we collected data from Wikidata [22], Baidu RenWu [3], and CNKI [23] to test our algorithm and existing gender prediction APIs.

Fig. 4. The training Datasets of Chinese characters and PinYin



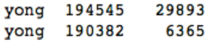
```
8474  痄,1,0          8475  sou,1,0
8475  锼,1,0          8476  guan,1,0
8476  燿,1,0          8477  meng,1,0
8477  獴,1,0          8478  shan,1,0
8478  挻,1,0          8479  gua,1,0
8479  掛,1,0          8480  pei,1,0
8480  肧,1,0          8481  han,1,0
8481  韓,1,0          8482  shua,1,0
8482  誜,1,0          8483  min,1,0
8483  垊,1,0          8484  ju,1,0
8484  裾,1,0          8485  mao,1,1
8485  耄,1,1          8486  tong,1,0
8486  熥,1,0          8487  liao,1,0
8487  飂,1,0          8488  lian,1,1
8488  戀,1,1          8489  ge,1,0
8489  骼,1,0          8490  dao,1,0
8490  燾,1,0          8491  che,1,0
8491  坼,1,0          8492  han,1,0
8492  晘,1,0          8493  bi,1,1
8493  畀,1,1          8494  fen,1,0
8494  偾,1,0          8495  kai,1,0
8495  勓,1,0          8496  lin,1,2
8496  懍,1,2          8497  zhi,1,0
8497  觶,1,0          8498  qing,1,0
8498  檾,1,0          8499  mao,1,0
8499  蓩,1,0          8500  jian,1,1
8500  洊,1,1          8501  yu,1,0
8501  斎,1,0          8502  men,1,0
8502  濷,1,0          8503  ying,1,0
8503  塋,1,0          8504  zhan,1,0
8504  觇,1,0          8505  wei,1,0
8505  衛,1,0          8506  gen,1,0
8506  亘,1,0          8507  kang,1,1
8507  槺,1,1          8508  ya,1,3
8508  氬,1,3          8509  an,1,0
8509  埯,1,0          8510  yu,1,0
8510  鄔,1,0          8511  wu,1,1
```

Fig. 5. Same characters in PinYin in different Chinese words (Chinses words)



```
17  永,194545,29893
18  勇,190382,6365
```

Fig. 6. Same characters in PinYin in different Chinese words (Hanyu Pinyin)



```
yong  194545   29893
yong  190382    6365
```

Fig. 7. A method on processing Hanyu Pinyin for training datasets

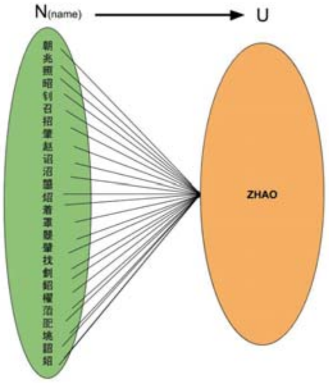
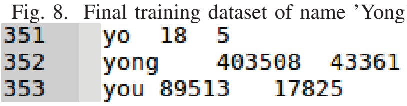


Fig. 8. Final training dataset of name 'Yong'



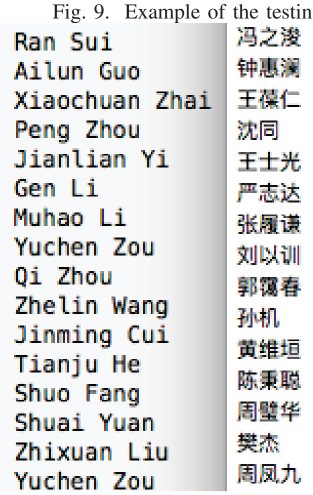
```
351    yo   18   5
352    yong     403508   43361
353    you 89513    17825
```

The labelled data is the lists of Chinese writers, Chinese Olympic participants, Chinese scientists. The data is labelled with genders.

These labelled data are in different types of Chinese names.

Fig 9 shows an example of the testing data in different types of Chinese names.

Fig. 9. Example of the testing data



Ran Sui
Ailun Guo
Xiaochuan Zhai
Peng Zhou
Jianlian Yi
Gen Li
Muhao Li
Yuchen Zou
Qi Zhou
Zhelin Wang
Jinming Cui
Tianju He
Shuo Fang
Shuai Yuan
Zhixuan Liu
Yuchen Zou

冯之浚
钟惠澜
王葆仁
沈同
王士光
严志达
张履谦
刘以训
郭霭春
孙机
黄维垣
陈秉聪
周璧华
樊杰
周凤九

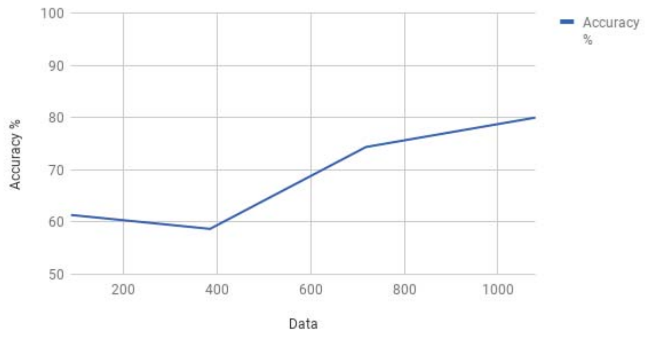
We used 1080 labelled data on testing our algorithm and existing gender prediction APIs. In this next section, we will show some results on testing our algorithm and compare with existing gender prediction APIs.

V. Experiments

A. Testing the algorithm on Predicting different types of Chinese names

We tested our system with real collected data. We used 1080 labelled data to test the gender prediction using different types of Chinese names. The accuracy is 80 %. Fig 10 shows the accuracy of our algorithm on predicting Chinese names.

Fig. 10. Accuracy of our algorithm on predicting different types of Chinese names



B. Accuracy on existing systems and our algorithm on Predicting different types of Chinese names

We used 503 data test existing gender prediction APIs and also test with our algorithm. Of this 503 dataset, 251 data is the name labelled with gender in Hanyu Pinyin, whereas 252 data is the name labelled with gender in Chinese characters.

We used different types of Chinese names for testing these existing systems as well as our algorithms. Table III shows the accuracy of each system on testing with Chinese names in Hanyu Pinyin.

TABLE III
The accuracy of the existing systems and our algorithm on gender prediction of names within Hanyu Pinyin

Generize	58 %
Gender API	72 %
Namsor	6 %
Name API	45 %
Ngender	0 %
Our algorithm	76 %

Fig 11 displays the accuracy of the existing systems and our algorithm on gender prediction of names within Hanyu Pinyin detail.

We can see that our algorithm gets the highest accuracy on gender prediction of names in Hanyu Pinyin.

Fig. 11. Accuracy of the existing systems and our algorithm on gender prediction of names within Hanyu Pinyin

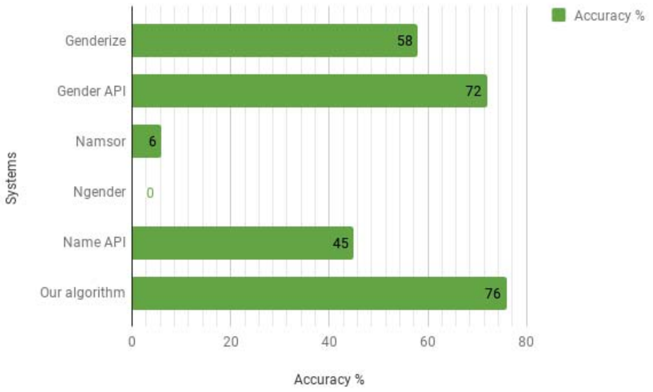


Fig 12 shows the accuracy on predicting gender using Chinese names with the existing systems and our method.

Fig 13 displays the accuracy on predicting gender using all types of Chinese names with the existing systems and our system. We can see that our system has the highest accuracy on gender prediction of names.

In our research, we found that our Recursion identify algorithm can highly increase the accuracy on gender prediction of Chinese names in Hanyu Pinyin. Our new method on predicting gender also gets the highest accuracy on using different types of Chinese names.

Fig. 12. Accuracy of the existing systems and our system on gender prediction of names within Chinese characters

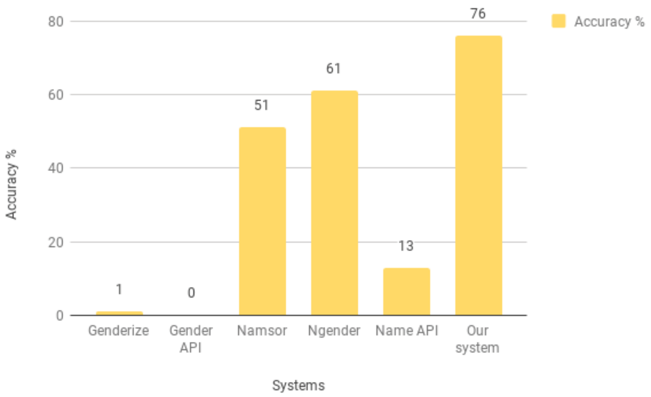
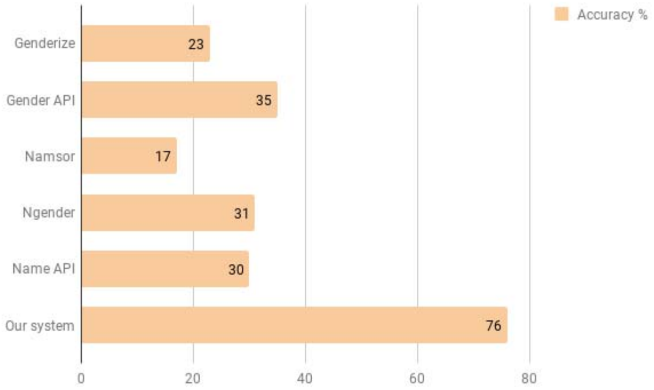


Fig. 13. Accuracy of the existing systems and our system on gender prediction of names within all types of Chinese names



VI. Conclusion and Future Work

In this paper, we have presented an algorithm for identifying and classifying Chinese names in Hanyu Pinyin. We have demonstrated the new method of predicting different types of Chinese names. We did increase the accuracy on gender prediction of names in Chinese names. Our method can identify different types of Chinese names and process them for gender prediction. This method is useful for many research areas on analysing their data of Chinese names. We did the experiments with our algorithm in analysing the collected, labelled data. In the next step, we want to develop a new method that can output high accuracy results for predicting gender, data's subjects and their culture origin simultaneously in high speed.

References

[1] gender-api.com. Gender API. Available at: https://gender-api.com/, Last accessed: 16 April 2018.

[2] Jingchao Hu. ngender 0.1.1: Guess gender for Chinese names. Available at: https://pypi.python.org/pypi/ngender/ 0.1.1, Last accessed: February 2017.

[3] baidu.com.com. BAIDU RENWU. http://baike.baidu.com/renwu, Last accessed: 30 April 2018.

[4] Jacob Perkins. Python Text Processing with NLTK 2.0 Cook- book. Packt Publishing, 9 Nov. 2010. isbn: 1849513600.

[5] google.com. Google translator. Available at: https://translate.google.co.uk, Last accessed: 16 April 2018.

[6] Namsor. NamSor Gender API. Available at: http://www. namsor.com, Last accessed: April 2018.

[7] Ronald L. Graham, Donald E. Knuth, Oren Patashnik: Concrete mathematics - a foundation for computer science (2. ed.). Addison-Wesley 1994, ISBN 978-0-201-55802-9, pp. I-XIII, 1-657

[8] webofknowledge.com. Web of Science. Available at: https: //apps.webofknowledge.com, Last accessed: June 2017.

[9] nameapi.org. NameAPI. Available at: https://www.nameapi.org/en/home/, Last accessed: 23 April 2018.

[10] Andrew Flowers. "The Most Common Unisex Names In America: Is Yours One Of Them?" In:FiveThirtyEight (2015).

[11] saedsayad.com.Naive Bayesian. Available at: http://www.saedsayad.com/naivebayesian.html, Last accessed: May2017.

[12] wikipedia.org.List of British scientists. Available at: https://en.wikipedia.org/wiki/ListofBritishscientists, Last accessed: June 2017.

[13] python.org. langdetect,Python object serialization. Available at: https://pypi.python.org/pypi/langdetect, Last accessed: January 2018.

[14] Suriyan Laohaprapanon and Gaurav Sood. Ethnicolr. Available at: https://pypi.python.org/pypi/ethnicolr/0.1.2, Last accessed: 16 April 2018.

[15] Lx Yu, pinyin 0.4.0. Available at: https://pypi.python.org/pypi/pinyin, Last accessed: January 2018.

[16] Steven Loria, textblob 0.15.1. Available at: https://pypi.python.org/pypi/textblob, Last accessed: March 2018.

[17] Kamil Wais. Gender Prediction Methods Based on First Names with genderizeR. In: The R Journal 8.1 (2016), pp. 17,37.

[18] wikipedia.org, Category: given names. Available at: https://en.wikipedia.org/wiki/Category, Last accessed: March 2018.

[19] fileformat.info, Unicode Character. Available at: http://www.fileformat.info/info/unicode/char/4e00/index.htm, Last accessed: March 2018.

[20] Hua Zhao, Fairouz Kamareddine. Advance gender prediction tool of first names and its use in analysing gender disparity in Computer Science in the UK, Malaysia and China. CSCI 2017: 222-227

[21] Susanna S. Epp. Discrete Mathematics with Applications 2nd.p. 427. 1995

[22] WIKIDATA.org. WIKIDATA. Available at: https://www.wikidata.org/wiki/Wikidata, Last accessed: 30 April 2018.

[23] CNKI.NET. Journal of China Academic Database. Available at: https://www.ssa.gov/oact/babynames.html, Last accessed: June 2017.

[24] Barber, C.B., Dobkin, D.P., and Huhdanpaa, H. (1996) The quickhull algorithm for convex hulls, ACM Transactions on Mathematical Software, Vol. 22, No.4.

[25] Teizer, J.; Mantripragada, U.; Venugopal, M. 2008. Analyzing the travel patterns of construction workers, in Proceedings of the 25th International Symposium on Automation and Robotics in Construction, ISARC 2008, June 26 - 29, 2008, Vilnius, Lithuania.

[26] Guillaume Bonfante, Matthieu Kaczmarek, Jean-Yves Marion: A Classification of Viruses Through Recursion Theorems. CiE 2007: 73-82.