The Application of Naive Bayes Classifier In Name Disambiguation

Na Li* and Jin Han

School of Computer and Software, Nanjing University of Information Science and Technology,Nanjing 210044,China
757165407@qq.com,284615715@qq.com

Abstract. Name repetition exists in the academic resource management system, which brings difficulties to academic evaluation, information retrieval, citation analysis and so on. According as different authors use function words in different habits, the Naive Bayes classifier was used to study in this paper. Based on the assumption of feature independence, this paper selects 26 common function words with high frequency as statistical frequency standard, use Naive Bayes classifier to classify texts. Experiments show that the method has a high accuracy rate.

Keywords: Naive Bayes classifier, feature independence, function words analysis, name disambiguation.

1 Introduction

The authors of many scientific papers have same name, some authors' names change with time or living environment, these problems bring difficulties to academic evaluation, information retrieval, citation analysis and so on. Joint Conferences Digital Libraries is held for this issue, it began in 2001 in the United States, has been successfully held 16 sessions.

In order to solve name repetition in web search, some systems have been developed. In 2007, Chen and Martin proposed a robust unsupervised name disambiguation method, developed the Poly UHK system[1]. Masaki Ikeda and Shingo Ono developed an ITC_UT system[2], using two-step clustering, the first step using hierarchical clustering, the second step based on hybrid keyword clustering algorithm. Lorenza Romano and Krisztian Buza developed a XMedia system[3], the system used the quality threshold clustering algorithm and used machine learning methods on the similarity comparison.

In recent years, some new name disambiguation algorithm is proposed. The first kind is a similairty calculation-based clustering disambiguation mehtod, such as Huang proposed an algorithm for the same name differentiation based on multi-view nonnegative matrix decomposition[4]. The second is a hierarchical-based clustering disambiguation approach, such as Zhang used hierarchical clustering algoirithm to solve the multi-document ambiguity issue of Chinese names[5]. Huang put forward person name disambiguation based on hierarchical clustering and web page relationship[6]. The third is a clustering disambiguation method based on the specific

relationships. For example, Li presented a name disambiguation approach based on the relationship of document collaborators[7].

Most of above are clustering algorithms, which use the metadata of the papers as the basis of clustering, calculate the similarity between the papers, and then use the appropriate clustering algorithm to cluster the papers according to the similarity degree. But these algorithms depend on a lot of conditions. When the conditions are less or often changes, the test accuracy rate will be greatly reduced. In addition, the application of this kind of algorithm is narrow. But the algorithm of name disambiguation in this paper is based on the frequency of using function words, so that it eliminates the dependence on a lot of conditions, has a wider scope of application and performs good in name disambiguation.

2 The Principle of Naive Bayes Classifier

In the construction method and theory of many classifiers, the Naive Bayes classifier has been widely used because of its computational efficiency, high accuracy and solid theoretical basis. The basis of the idea is as follows: For the given items to be classified, solve the probability of occurrence of each category under the conditions in which this item occurs, and sort this item as the category whose probability is the largest. Each training sample data is decomposed into one-dimensional eigenvector X and decision category variable C , and it is assumed that the components of the eigenvector are independent of each other.

The specific definition is as follows:

$x = \{a_1, a_2, ..., a_m\}$ is a feature to be classified, and each "a" is a characteristic attribute of x, $C = \{y_1, y_2, ..., y_n\}$ is a decision category variable, calculate $P(y_1|x), P(y_2|x), ..., P(y_n|x)$.

If $P(y_k|x) = max\{P(y_1|x), P(y_2|x), ..., P(y_n|x)\}$, then $x \in y_k$.

According to Bayes theorem:

$$P(y_i|x) = \frac{P(x|y_i)P(y_i)}{P(x)} \quad . \tag{1}$$

Because the denominator is constant for all categories, we only need to maximize the molecules can be. And because the characteristic attributes are conditional independent, so there are:

$$P(x|y_i)P(y_i) = P(a_1|y_i)P(a_2|y_i)...P(a_m|y_i)P(y_i) = P(y_i)\prod_{j=1}^{m} P(a_j|y_i) \tag{2}$$

Therefore, it is only necessary to solve the conditional probability of each characteristic attribute under each category, put the result into the above formula, and then we can get the probability of occurrence of each category under the condition of the item to be classified, compare out the maximum value, and classify the item to be classified as the category that the maximum probability represents.

3 Key Technology and Algorithm Process

3.1 Feature Extraction

At present, there are seven kinds of methods of feature extraction: mutual information, expected cross entropy, information gain, text evidence, probability ratio, word frequency method and CHI. In this paper, word frequency method is used to calculate the frequency of the function words. According to statistical results, there is a great difference in the frequency of the function words used by different authors (**Fig. 1.)**. The object of this study is Chinese scientific papers. For a paper, calculate the number of the words that need to be counted in this paper, and then divide the total number of words in the paper to get the frequency of the statistical function words needed in the paper. In this paper, 26 commonly used Chinese function words with highest frequency are selected as the standard of statistical frequency.

Since the value of the frequency is continuous, the value of each feature attribute is infinite, so Naive Bayes classifier can not be applied to this problem. In this paper, according to the distribution of the calculated frequency, the accuracy of the frequency will be reduced to 2, 3 or 4 bits after the decimal point, and take reduced precision frequency as the value of the characteristic attribute.

The extracted feature vector format is as follows:

[a1, a2, a3,..... , a24, a25, a26, type]

"a1-a26" is the requency of 26 words , "type" is the category that the feature vector represents (yes or no).

This paper selects some now commonly used Chinese function words, including prepositions, conjunctions, adverbs, auxiliary words, through the statistics of 452 papers in the frequency, select the 26 function words of highest frequency. The 26 selected function words are as follows:{"进而", "但", "且", "而且", "按", "及", "以及", "和", "同", "因此", "于", "与", "假如", "如果", "当", "并且", "或","根据", "然而", "跟", "但是", "或者", "因而", "在", "按照", "从而", "一样"}.

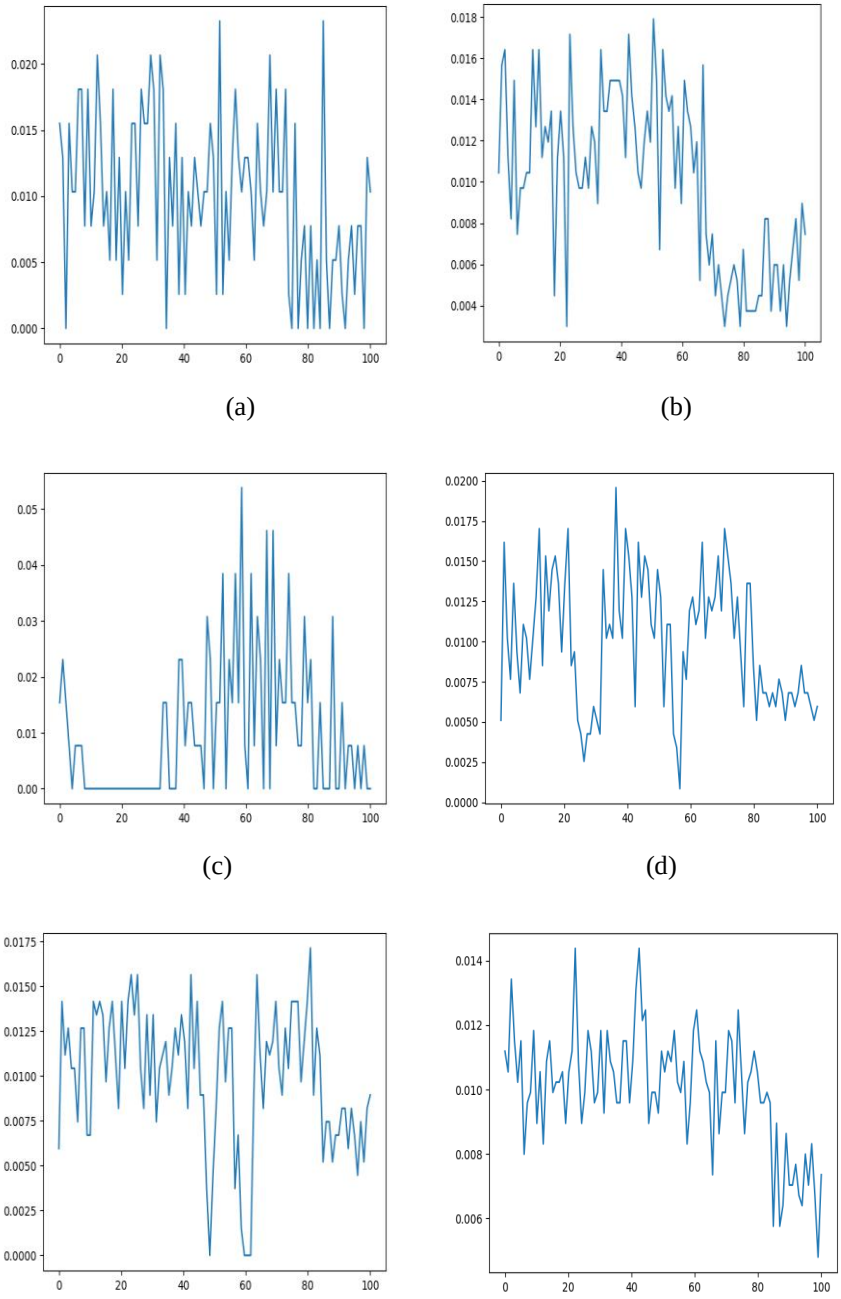

Fig. 1. Distribution of function words in papers from six different authors

3.2 Algorithm Process

3.2.1 Data Input

The program enters a path that contains all the papers of a given scholar, and uses the tika plugin to extract the pdf format of the paper into a string. And use ansj to divide the text into word segmentation, extract the function words, calculate the frequency of each function word. The final output of these papers as a two-dimensional array of data. Each line represents a paper,marked with "yes", take a small number of papers as a test sample. In the same way, take the equivalent of another author's paper, marked the category with "no", constitute a complete sample in this way.

3.2.2 Process

For each test sample, calculate the probability that each feature attribute of the sample appears in the "yes" category, multiply the product of these probabilities with the probability of "yes" in the training sample to obtain the result P1. Calculate the probability P2 which represents "no" in the same way, compare P1 and P2, and classify the test sample as the category represented by the larger value of the two.

Due to some feature of the training sample value does not appear in the experimental process, resulting P(a|y)=0, so that the quality of the classifier greatly reduced. In order to solve this problem, we introduce Laplace calibration, the idea is very simple, that is, all the categories under the division of the count plus 1, to avoid the embarrassing situation caused by the frequency of 0.

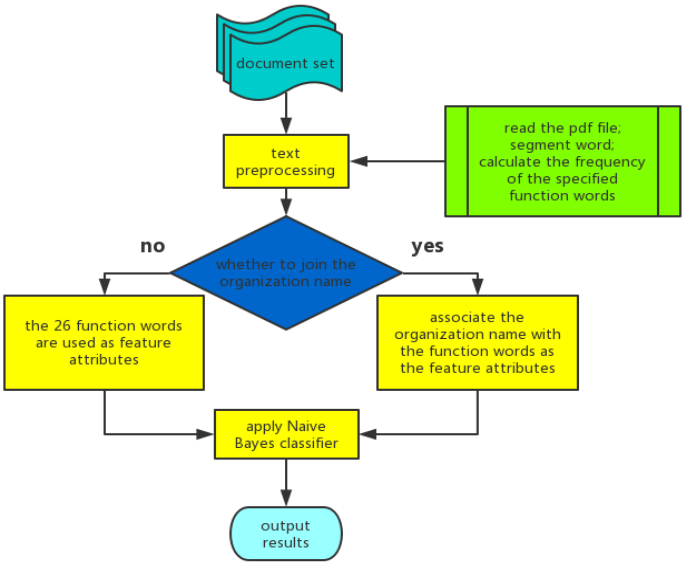

Fig. 2. Algorithm flow chart

3.2.3 Core Pseudo-code of the Algorithm

```
Algorithm 1 Realization of Naive Bayes Classifier
BEGIN
The training samples were divided into two groups according to the results of yes or no，
  put groups into resultMap;
    Traversal resultMap{
            If the key is “yes”{
             Assign the proportion of the sample with the result of "yes" in the total
             training sample to yesCurrent;
                Traversal testList{
             yesCurrent=yesCurrent* The probability of the current feature attribute
             value in the testList that appears in the “yes” sample;
           }
        }
          If the key is “no”{
              Assign the proportion of the sample with the result of "no" in the total
            training sample to noCurrent;
                Traversal testList{
             noCurrent=noCurrent* The probability of the current feature attribute value
             in the testList that appears in the “no” sample;
          }
        }
        If yesCurrent>noCurrent
            return “yes”;
        Else return“no”;
    }
END
```

4 Experiment

4.1 Data Sources

The data studied in this paper are the Chinese periodicals downloaded from CNKI. In order to ensure that the sample will not be doped with the scholar who has the same name, thoses papers were all published by scholars on the home page. A total of 52 scholars and 1282 papers were studied.

4.2 Experiment Result

This paper has studied 52 samples, the average accuracy rate of classification without orgnization name is 85.273%, the average accuracy rate of classification with orgnization name is 93.074%.

The selected 6 representative data are as follows:

Table 1. Partial experiment data

Subject (Author)	Training Samples (articles)	Test Sample (articles)	Correct Rate %（without organization）	Correct Rate %（with organization）
Zhuge Jianwei	17	5	100.00	100.00
Wu Libing	19	5	80.00	100.00
Li Wei	28	7	85.71	85.71
Deng Zhengchun	54	11	72.73	90.91
Ni Ming	105	13	84.62	92.31
Xu Zeshui	167	21	95.24	95.24

As is shown in the table, when use the organization name as a characteristic attribute, the correct rate of classification will be significantly improved, which is consistent with the real situation in life that the probability that the author has the same name will be very low in the same organization. Thus it can be seen that organization name can be used as an important indicator of the name disambiguation.

At the same time, by observing the experimental results and training samples, this paper found that the higher the correct rate of several groups of data, the emergence of high frequency words is more. Therefore, it is concluded that the classifier realized in this paper has a better classification effect on those papers in which the frequency of the function words that this paper select is relatively high.

5 Conclusion

Many algorithms regard name disambiguation as a clustering problem. The field, the title, the author, the journal, and the abstract of the paper are used as the basis of clustering. By applying a similarity function to the document attributes to measure the similarity between the paper and the paper or the collection of papers and the collection of papers, and then use the appropriate clustering algorithm to cluster the paper according to the calculated similarity. The well-performed algorithm of the name disambiguation needs to rely on a large number of document metadata attributes, but some attributes in the actual academic resource platform is missing or difficult to obtain, which led to the poor practical application of these algorithms and poor scalability.

This paper mainly studies the common same name problem in academic resource platform, and proposes a new algorithm of name disambiguation based on function words analysis and organization name. The algorithm eliminates the dependence of the algorithm on the field, the title, the collaborators, etc, and focuses on the author's habit of using the function words. The experiment has fully demonstrated that the

algorithm based on the function words has a high accuracy, can be applied to combat writing paper for others. For some ancient books whose author is abscent, we can take some authors who study the same direction with the ancient books in the same age as the categories, extract some representative function words, modify the classifier into a multi-classifier, then we can find possible author of these ancient books. When the organization name is added to the algorithm of the function words analysis, the average accuracy of the algorithm is improved by 7.8%, which can be used to track the experts of different disciplines, identify and study the hot topics in different fields. In summary, this algorithm saves a lot of human resources, achieve a certain accuracy, and have better flexibility and scalability.

References

1. Chen, Y., Martin, J.: Towards Robust Unsupervised Personal Name Disambiguation. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-Co NLL) (2007).
2. Ikeda, M., Ono, S.: Issei Sato. Person Name Disambiguation on the Web by Two Stage Clustering. In: Second Web People Search Evaluation Workshop, WWW2009 (2009).
3. Romano, L., Buza, K., Giuliano, C.: XMedia: Web People Search by Clustering with Machinely Learned Similarity Measures. In: Second Web People Search Evaluation Workshop, WWW2009 (2009).
4. Huang, Z.: Research on name disambiguation algorithm based on multi - view nonnegative matrix factorization. Dalian University of Technology. Master Thesis (2015).
5. Zhang, S., You, L.: Chinese people name disambiguation by hieracrhical clustering. New Technology of Library & Information Service **2010**(11), 64-68 (2010).
6. Li, Q.: Person name disambiguation based on hierarchical clustering and web page relationship. Shan Dong University. Master Thesis (2012).
7. Li, W.J.: The research and application of name disambiguation algorithm based on multi-level clustering. Dalian University of Technology. Master Thesis (2013).
8. Yang, Y.L., Zhou, J., Li, B.C.: Name disambiguation algorithm based on ensemble. Application Research of Computers **33**(9), 2716-2720 (2016).
9. Chen, C., Wang, H.F.: Social network based cross-document personal name disambiguation. Journal of Chinese Information Processing **25**(5), 75-82 (2011).
10. Guo, S.: Research on author name disambiguation algorithm in the literature database. New Technology of Library and Information Service **29**(Z1), 69-74 (2013).
11. Gu, B., Sun, X.M., Sheng, V.S.: Structural Minimax Probability Machine. IEEE Transactions on Neural Networks and Learning Systems (2016). DOI : 10.1109/TNNLS.2016.2544779
12. Zhou, Z.L., Wang, Y.L., Wu, Q.M.J., Yang, C.N., Sun, X.M.: Effective and Efficient Global Context Verification for Image Copy Detection. IEEE Transactions on Information Forensics and Security, vol.12 no.1. pp 48-63, 2017. DOI: 10.1109/TIFS.2016.2601065, 2016.
13. Gu, B., Sheng, V.S., Tay, K.Y., Romano, W., Li, S.: Incremental Support Vector Learning for Ordinal Regression. IEEE Transactions on Neural Networks and Learning Systems **26** (7), 1403-1416 (2015).
14. Tian, Q., Chen, S.C.: Cross-Heterogeneous-Database Age Estimation Through Correlation Representation Learning. Neurocomputing **238**, 286–295 (2017).

15. Li, X., Xie, H., Chen, L., Wang, J., Deng, X.: News impact on stock price return via sentiment analysis. Knowledge-Based Systems **69**, 14-23 (2014).
16. Xie, H., Li, X., Wang, T., Chen, L., Li, K.: Personalized search for social media via dominating verbal context. Neurocomputing **172**, 27-37 (2016).
17. Rao, Y.H., Li, Q., Wu, Q., Wang, T.: A multi-relational term scheme for first story detection. Neurocomputing (2017). DOI:10.1016/j.neucom.2016.06.089