Then:

$$P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN}, \quad F_1 = \frac{2PR}{P + R} \tag{7}$$

The results shown in table 1.Firstly,the application of classification achieve accuracy of 89.4% and recall of 88.2% after training random sample of 40(Random 10 for each area).Secondly, the classifier performance is stable fundamentally, and the accuracy、recall and F_1 value wave in the small range with the increase of training sample.

Table 1 Accuracy rate、Recalling rate and F1

Number of Training samples	Accuracy rate	Recall rate	F1
40	89.4	88.2	88.1
60	89.5	88.3	88.5
80	88.5	86.8	87.3
100	88.9	87.2	87.8
120	89.6	87.4	87.9
140	88.8	87.6	88
160	88.6	87.8	87.9

4 Conclusions

Web data includes query interface and result. The classification of web data is the key to integrate classification and search. The paper discusses feature extraction from HTML and classification with naïve Bayes. The result shows that the method used in the paper can work well with small samples. But the paper does not consider more content of the words, so it is our further study.

References

[1] GAO Ling, ZHAO Ping-Pong, CUI Zhi-ming. Automatic Judgment of Deep Web Query Interfaces. Computer Technology and Development Vol.17-5(2007), p.150-151

[2] YU Wei, LI Shi-jun, WEN Li-juan, TIAN Jian-wei.Ranking of Deep Web Sources Dased on Data Quality. Journal of Chinese Computer Systems Vol.31-4(2010), p.641-646

[3] JIN Ling-zhi, WANG Xiao-ling, ZHU Shou-zhong.Automatic classification of Deep Web sources. Microcomputer Information Vol.25-4(2009), p.227-230

[4] Tang Huanling, Sun Jiantao, Lu Yuch. A Weight Adjustment Technique with Feature Weight Function Named TEF-WA in Text Categorization. Journal of Computer Research and Development Vol.42-1(2005), p.47-51

[5] Wei Huang，Zhongliang Jing.Multi-focus image fusion using pulse coupled neural network.Pattern Recognition Letters Vol.28-9(2007), p.1123-1132.