Negation Naive Bayes for Categorization of Product Pages on the Web

Kanako Komiya [1] **Naoto Sato** [1] **Koji Fujimoto** [1,2] **Yoshiyuki Kotani** [1]
Tokyo University of Agriculture and Technology [1]
Tensor Consulting Co.Ltd. [2]
{kkomiya, kotani}@cc.tuat.ac.jp
50009646113@st.tuat.ac.jp
koji.fujimoto@tensor.co.jp

Abstract

We propose the negation naive Bayes (NNB): a new method to categorize product pages on the Web depending on their information. It is a modified version of the naive Bayes (NB) and we got the idea from the complement naive Bayes (CNB). We compared the NNB with the NB and the CNB. Our experiments show that the NNB outperformed the other methods significantly when the product pages were distributed non-uniformly through categories.

1 Introduction

In late years, e-commerce, the services by which users can easily purchase products on the Web without visiting a store, is introduced in many companies. When products are purchased via Internet, the user narrows down the candidate categories of each product in incremental steps. We categorized the products on the Web automatically depending on their information using the method of text classification (Sato et al., 2011).

Many researchers have investigated text classification and the naive Bayes (NB) is one of the most famous methods for it. However when we use the NB classifier to categorize the products, the accuracies were not very high, especially when the data distribution is very skewed.

Hence this paper proposes the negation naive Bayes (NNB): a new method of text classification especially for the product pages on the Web depending on their information. It is a modified version of the NB and we got the idea from the complement naive Bayes (CNB). Our experiments showed that the NNB outperformed the NB and the CNB when the product pages were distributed non-uniformly.

This paper is organized as follows. Section 2 reviews related works on text classifications and the NB classifiers. Section 3 describes the classification methods including our proposal method: the NNB. Section 4 describes the system to categorize the product pages and explains the experimental setting. We describe results in Section 5 and discuss them in Section 6. Finally, we conclude the paper in Section 7.

2 Related Work

Many works on text classification have been accomplished so far. Approaches of Bayes are often used within the area of text classification (Mochihashi, 2006). Izutsu et al. (2005) categorized the html documents and compared the NB classifier with discriminant analysis and the rule-based method. They suggested the simple implementation and the high scalability of the NB classifier. McCallum and Nigam (1998) suggested the difference between multinomial model and multivariate Bernoulli model of the NB classifier in text classification. Lewis (1992) compared the difference of the effect between the types of features used for text classification: words, phrases, clustered words, clustered phrases and indexing terms. W.Church (2000) used a concept called “Adaptation” as the weighting method to the words in substitution for IDF value, and defined the words related to contents but not included a document as “Neighbor”. The feature terms were extracted depending on them.

In addition, the method called “Complement Naive Bayes” attracts attention. It estimates parameters of a category using data from all categories except the category which is focused on (J.D.M.Rennie et al., 2003).

On the other hand, there have been the works that used the product information of Internet auctions (Nishimura et al., 2008).These works suggest a method to extract the attribute information from the description of the product pages.

This paper proposes the NNB. Its equation is

Proceedings of Recent Advances in Natural Language Processing, pages 586–591,
Hissar, Bulgaria, 12-14 September 2011.

derivable from the equation of the NB unlike the CNB but it has the same advantages; it tackles the ununiformity of the texts of each category. We got the classification accuracies that exceed the NB and the CNB significantly when the data distribution is non-uniformly.

3 Classification Method

In this section, we describe the classification methods to categorize the product pages on the Web including our proposal method: the NNB. The distribution of the product pages of each category is very skewed in Internet auctions. Therefore, the classification model which tackles the ununiformity of the data distribution is necessary.

3.1 Naive Bayes Classifier

We used the NB classifier to classify the product pages as a baseline. Let $d = w_1, w_2, \ldots, w_n$ denote the text containing the words and let c denote a category. Here, let $\hat{c}$ denote the category that d belongs to, and $\hat{c}$ is as follows:

$$\hat{c} = \operatorname*{argmax}_c P(c|d) \qquad (1)$$

where $P(c)$ and $P(d)$ each represent the prior probability of c and d.

By substituting theorem of conditional probability into the equation, we obtain the following:

$$\begin{aligned}\hat{c} &= \operatorname*{argmax}_c P(c|w_1, w_2, \ldots, w_n) \\ &= \operatorname*{argmax}_c P(w_1, w_2, \ldots, w_n|c)P(c) \qquad (2)\end{aligned}$$

We assume that w_i is conditionally independent of every other word. This means that under the above independence assumptions, $P(w_1, w_2, \ldots, w_n|c)$ is approximated by the following:

$$P(w_1, w_2, \ldots, w_n|c) \approx \prod_i P(w_i|c) \qquad (3)$$

Finally, the category $\hat{c}$ that d belongs to is determined by following:

$$\hat{c} = \operatorname*{argmax}_c P(c) \prod_i P(w_i|c) \qquad (4)$$

When there is the pair of w_i and c where $P(w_i|c) = 0$, the left-hand value of eq. (4) equals 0. Therefore, let N denote the total number of training data, and substitute following eq. (5) for eq. (4) in order to avoid this case.

$$P(w_i|c) = \frac{0.1}{N} \qquad (5)$$

3.2 Complement Naive Bayes Classifier

The NB classifier tends to classify documents into the category that contains large number of documents. The CNB classifier is a modification of the NB classifier. This classifier improves classification accuracy by using data from all categories except the category which is focused on. This classifier is also used as a baseline.

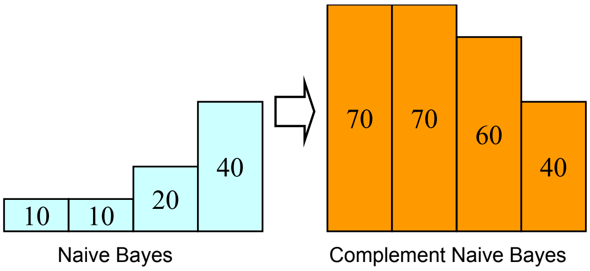

Figure 1: The difference of training data between two methods

Figure 1 shows the difference of the training data between the NB classifier and the CNB classifier. The NB classifier estimates parameters of a category using the data from the category which is focused on. When there are four categories that each contain 10, 10, 20, 40 training data, and the category with the most data has four times data as many as the category with the least data.

On the other hand, the CNB classifier estimates parameters of a category using the data from all categories except the category which is focused on. Therefore, the category with the least data is 40 and the category with the most data is 70. The gap of the number of the training data is less than the NB classifier.

The CNB classifier estimates the likelihood from probability of occurrence of words and decides the category which the product pages are classified into. The CNB estimates $P(w_i|c)$ using data from all categories except c ($\bar{c}$ denote those categories):

$$P(w_i|c) = \prod_{\bar{c}} \frac{1}{P(w_i|\bar{c})} \qquad (6)$$

When there is the pair of w_i and $\bar{c}$ where $P(w_i|\bar{c}) = 0$, we used the same the smoothing method as eq. (5).

Finally, the category $\hat{c}$ that d belongs to is determined by following:

$$\hat{c} = \operatorname*{argmax}_c P(c) \prod_{\bar{c}} \frac{1}{P(w_i|\bar{c})} \qquad (7)$$

3.3 Negation Naive Bayes Classifier

The CNB is a method that tackles the ununiformity of the data distribution. However we think eq. (7) is not derivable from eq. (1). J.D.M.Rennie et al. (2003) also ignored $P(c)$ assuming it is enough small comparing with $P(w_i|\bar{c})$ but we think $P(c)$ cannot be always ignored and should be calculated especially when the data distributionis very skewed.

Therefore, we propose the NNB, which is derivable from eq. (1) but also have the advantage like the CNB. The derivation of the equation of the NBB is as follows.

First, equ. (8) is obtained from equ. (1) because we would like to find $\hat{c}$: the category which maximaizes the posterior probability $P(c|d)$ here again. Here, we focus on $\bar{c}$: the categories which d is not supposed to belong to, like the CNB.

$$\begin{aligned}\hat{c} &= \underset{c}{\operatorname{argmax}}(1 - P(\bar{c}|d)) \\ &= \underset{c}{\operatorname{argmin}}\, P(\bar{c}|d) \qquad (8)\end{aligned}$$

Next, equ. (9) follows from equ. (8) and Bayes' theorem as follows:

$$\begin{aligned}\hat{c} &= \underset{c}{\operatorname{argmin}} \frac{P(\bar{c})P(d|\bar{c})}{P(d)} \\ &= \underset{c}{\operatorname{argmin}}(\bar{c})P(d|\bar{c}) \qquad (9)\end{aligned}$$

Finally, by substituting theorem of conditional probability like and assuming independence of every other word like , the category $\hat{c}$ that d belongs to is determined by following:

$$\hat{c} = \underset{c}{\operatorname{argmin}}(\bar{c}) \prod_{i} P(w_i|\bar{c}) \qquad (10)$$

This is an equation of the NNB that we propose. We used the same smoothing method as the CNB.

4 Classification Experiments

In this section, we describe the system to categorize the product pages and explain the setting of the classification experiments.

4.1 Data Set for Experiments

We used the product pages assigned to subordinate category of "Windows desktop PC", "baby products", and "memorial stamps" on Yahoo! auctions[1] as the training and test data. These categories can be narrowed down as follows from top category of Yahoo! auctions.

- All products > Computers > Personal computers > Windows > desktop PC
- All products > Baby products
- All products > Antiques or Collections > Stamps or cards > Japanese > Memorial stamps

The left-hand of the mark ">" is the parent category and right-hand is the child category.

We regard the categories assigned by the sellers as the correct labels. In addition, each product belongs to only one category in Yahoo! auctions. Categories are hierarchical and each product is assigned to terminal categories.

We used only one product page by one seller for each category to get rid of bias of notation habits of each seller like (Nishimura et al., 2008). The number of the categories and the product pages before and after removing the product pages of the same sellers is shown in Table 1. In addition, the number of the product pages of Windows desktop PC, baby products, and memorial stamps that we used for classification is each shown in Figure 2, Figure 3, and Figure 4. The categories are sorted by the number of the product pages in these figures. They show the numbers of the product pages are distributed non-uniformly through the categories.

Genre	Before	After	categories
PC	19,849	4,403	21
Baby product	29,477	10,389	62
Stamp	16,543	3,980	53

Table 1: The number of the categories and the product pages before and after removing the product pages of the same sellers

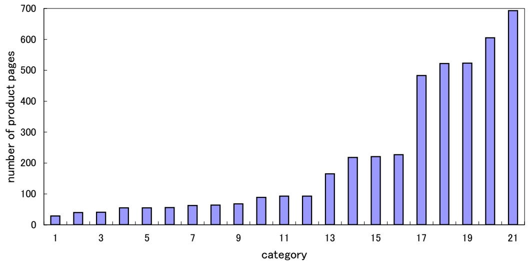

Figure 2: The number of the product pages of Windows desktop PC for each category

The product pages are described in HTML but we removed the HTML tags assuming that they were unnecessary for classification.

[1] http://auctions.yahoo.co.jp/

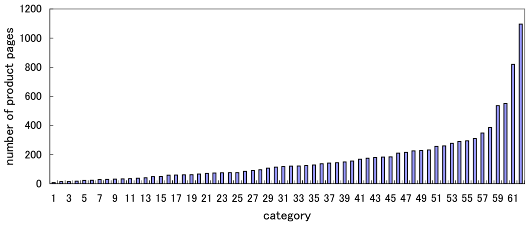

Figure 3: The number of the product pages of baby products for each category

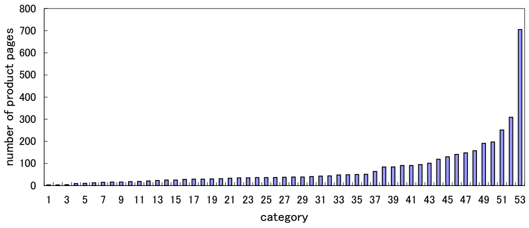

Figure 4: The number of the product pages of memorial stamps for each category

4.2 Features for Classification

The product pages of Yahoo! auctions contain many technical terms and many words which have a very small effect about the classification (e.g. symbols, shipping address, and so on). They also contain itemization and their sentences are short and colloquial. From these properties, we thought that it is not important for classification to see the whole product pages, but to extract words which represent the category of the product. We performed the classification experiments depending on the following four kinds of information.

- All the words in the titles
- The nouns extracted from the titles
- All the words in the titles and the descriptions
- The nouns extracted from the titles and the descriptions

4.3 Procedure of Classification

The procedure of the classification is following.

1. Obtain product pages with the category label which they are classified into.
2. Extract the titles and the description if necessary.
3. Perform morphological analysis on each product pages using Chasen [2].
4. Extract the features for classification.
5. Classify the product pages using the methods shown in Section 3.

We used the default settings of Chasen. We used the 5-fold cross validation for the test. The chi-square test was performed to see if the difference is significant or not and its level of significance was 0.05.

5 Results

In this section, we describe the results of the classification experiments. First, we compare the accuracies of classifiers in four settings according to the features to categorize the product pages that we discribed in 4.2 about Windows desktop PC.

Table 2 shows the classification accuracy of the NB, the CNB, and the NNB. The "-" mark in "Descriptions" column means the descriptions of the product pages were not used for classification and the "+" mark means the classification was performed depending of the words from both the title and the description of the product pages. In addition, "Nouns" in "POS" column means only the nouns were used for the features of classification and "all" means all the words were used. Table 2 shows that whatever classifier was used, the accuracies when the titles were used were higher than when the titles and the descriptions were used. The difference was statically significant. Table 2 also shows that product pages can be classified little more correctly depending on only the nouns than all the words, but the difference was not significant.

Descriptions	POS	NB	CNB	NNB
-	all	**0.613**	**0.698**	**0.711**
-	nouns	**0.629**	**0.701**	**0.713**
+	all	0.456	0.623	0.623
+	nouns	0.481	0.641	0.642

Table 2: The classification accuracy using the product pages of Windows desktop PC

Next, we compare the NNB classifier with the NB classifier and the CNB classifier using the data of the following three genres: Windows desktop PC, baby products, and memorial stamps. In view of Table 2, we performed the classification experiments depending on two kinds of features, all the words of the titles and the nouns extracted from them.

[2] http://sourceforge.net/projects/masayu-a/

Table 3 summarizes classification accuracy of the NB, the CNB, and the NNB using the data of these three grnres. The MFC is an abbreviation for the most frequent category. The "Total" in "Genre" column means the total average of three genres.

Genre	POS	NB	CNB	NNB
PC	all	0.613	0.698	**0.711**
PC	nouns	0.629	0.701	**0.713**
PC	the MFC	0.158		
Baby product	all	0.479	0.445	**0.508**
Baby product	nouns	0.484	0.436	**0.507**
Baby product	the MFC	0.105		
Stamp	all	0.451	0.452	**0.489**
Stamp	nouns	0.436	0.447	**0.490**
Stamp	the MFC	0.177		
Total	all	0.505	0.506	**0.552**
Total	nouns	0.508	0.501	**0.552**
Total	the MFC	0.133		

Table 3: The classification accuracy using the product pages of the three grnres

Table 3 shows that whatever features were used, and the data of whatever genre were used, the accuracies of the NBB classifier were higher than other classifiers. The second best classifier varies depending on the genre of product pages. The difference between the CNB and the NNB of Windows desktop PC, the NB and the CNB of memorial stamps, and the NB and the CNB of the total product pages were not significant. All the other differences were statically significant. Table 3 also shows that sometimes the product pages were classified little more correctly depending on only the nouns than all the words and sometimes not. In addition, all these differences were not significant. Table 2 also shows the accuracies of the three classifiers of the product pages about Windows desktop PC, when the titles and the descriptions were used. The tendency of the results is almost the same as when the titles were used for classification.

Finally, we compare the three methods to classify by three-class classification using the data of the three genres. Here, we classify all the product pages of three genres into three classes: Windows desktop PC, baby products and memorial stamps. Table 4 shows the accuracy of this experiment.

POS	NB	CNB	NNB
all	**0.982**	0.978	0.978
nouns	**0.977**	**0.977**	0.976
the MFC	0.553		

Table 4: The classification accuracy of three class classification

It shows the NB classifier outperformed the other classifiers significantly when all the words in the title were used for the features of classification, and the NB and the CNB slightly outperformed the NNB when only the nouns on the title were used. When the nouns were used for the features, the difference among the NB, the CNB, and the NNB were not significant. In addition, when the features were compared, the classifier the accuracy when all the nouns were used was higher than when the nouns were used. The difference between the nouns and all the words was significant when the NB classifier was used. All the other differences were not statically significant.

6 Discussion

Table 2 shows that the product pages can be classified more correctly depending on only the titles of the product pages than both the titles and the descriptions of them. It means that the titles of the product pages were better features for classification than the titles and descriptions of them, at least for the product pages about Windows desktop PC. We think that this is because there are lots of words which are unnecessary for classification in the description of the product pages and they obstruct effective classification.

Next, Table 3 shows that whatever features were used, and the data of whatever genre were used, the accuracies of the NNB classifier were higher than the other classifiers. The second best classifier varies depending on the genre of product pages; the CNB was for the texts of Windows desktop PC or memorial stamps and the NB for the texts of baby products. Therefore when the NB classifier and the CNB classifier were compared, it is still unansuwered question that which is the better method to classify these product pages. However, the experiments show that our proposal method, the NNB is always the best method for the classification of the product pages of the three genres. When the total averages were compared, the NNB classifier also outperformed the other NB classifiers significantly.

In the experiments of Table 3, the products that belong to the categories with a few product pages tended to be classifed into the categories with many product pages when the CNB was used. We think we can see the reason form the equation of the CNB. Here, equ.(10), the equation of the NNB, can be rewritten as the following equ. (11)

$$\hat{c} = \underset{c}{\text{argmax}} \frac{1}{1 - P(c)} \prod_i \frac{1}{P(w_i|\bar{c})} \qquad (11)$$

From the equation of the CNB equ. (7) and equ. (11) , we can see that the difference of the equations between the NNB and the CNB is the usage of the prior probability $P(c)$. We think that the usage of the prior probability $P(c)$ in the equation of the CNB caused this problem.

In addition, Table 4 shows that the NB classifier outperformed the NNB classifier. It means that the NNB is not always the best method to classify the product pages of any genres. We think this is because the uniformity of the data. In this three-class classification, the product pages of each category are 4403, 10389, and 3980 and the distribution is not so non-uniform. The NNB tackles the ununiformity of the text but the advantage does not help in this situation. We think that that is why our proposal method could not classify more correctly than the other classifiers in the three-class classification. The measure of the uniformity of the distribution of texts such as deviation can be considered in the future in order to decide the best classification method for each category.

Finally, the differences between the nouns and all the words are almost always not significant. Only one exception is the difference of the three-class classification when the NB was used. We think this condition is not important comparing with the other conditions.

7 Conclusion

In this paper, we proposed the NNB to categorize product pages on the Web. It is a modified version of the NB and we got the idea from the CNB. Its equation is derivable from the equation of the NB unlike the CNB and it has the same advantage as the CNB: it tackles the ununiformity of the data distribution through categories.

We performed classification experiments using four kinds of features and product pages of three genres to compare three kinds of classification methods: the NB, the CNB, and the NNB. The features are all words in titles of the products pages, nouns extracted from the titles, all words in titles and descriptions of the product pages, and nouns extracted from them. The genres are Windows desktop PC, baby products, and memorial stamps.

The experiments gave us following three observations: (1) The titles of the product pages were better features for classification than the titles and the descriptions of them, at least for the product pages about Windows desktop PC, (2) When the classifiers were developed based on the titles of the product pages, our proposal method the NNB is always the best classification method in the three genres. (3) The NNB is not the best classification of three-class classification of the three genres. Therefore we think that the NNB is good for non-uniformly distributed data but is not so good for uniformly distributed data.

References

Kiyoshi Izutsu, Makoto Yokozawa, and Takeshi Shinohara. 2005. Comparative evaluation and applications of automatic web-based document classification methods. In *IPSJ SIG Notes 2005(32) (In Japanese)*, pages 25–32.

J.D.M.Rennie, L.Shih, J.Teevan, and D.R.Karger. 2003. Tackling the poor assumptions of naive bayes text classification. In *ICML2003*, pages 616–623.

David D. Lewis. 1992. An evaluation of phrasal and clustered representations on a text categorization task. In *15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 37–50.

Andrew McCallum and Kamal Nigam. 1998. A comparison of event models for naive bayes text classification. In *AAAI/ICML-98 Workshop on Learning for Text Categorization*, pages 41–48.

Daichi Mochihashi. 2006. Bayesian approaches in natural language processing. In *IEICE Technical Report. NC, Neurocomputing (In Japanese)*, pages 25–30.

Jun Nishimura, Rintaro Miyazaki, Naoto Maeda, Tatsunori Mori, Shorei O, Yusuke Ishikawa, Hiroyuki Kobayashi, Yuya Tanaka, and Fuyuko Kido. 2008. Attribute-value extraction from description of exhibits for facetted search in net auction system. In *ANLP2008 (In Japanese)*, pages 392–395.

Naoto Sato, Kanako Komiya, Koji Fujimoto, and Yoshiyuki Kotani. 2011. Categorization of product pages depending on information on the web. In *UCSIP2011*, pages 393–398.

Kenneth W.Church. 2000. Empirical estimates of adaptation: The chance of two noriegas is closer to $p/2$ than p^2. In *COLING ' 00*, pages 173–179.