Journal of Theoretical and Applied Information Technology
30th September 2012. Vol. 43 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN IMPROVED TEXT CLASSIFICATION METHOD BASED ON GINI INDEX

[1]XIAOQIANG JIA, [2]JIANGYAN SUN

[1]School of Mathematics and Information Science, Weinan Normal University, Weinan 714000, Shaanxi, China

[2]Modern Education Technology Center, Xi'an International University, Xi'an, Shaanxi, 710077, China

E-mail: [1]394892738@qq.com , [2]sunnyjy@126.com

ABSTRACT

In text classification, the purity of the Gini index can be used. When purity value is greater, the characteristic of the information contained in the attribute is higher, and the feature distinguishing capability is stronger. But using the Gini purity formula on feature weight, the classification result is not very good, one of the main reasons is those rare words only appearing in one category and not appearing in other categories can not be strictly differentiated. In order to solve this problem, On the basis of Gini index, an improved feature weight method based on Gini index has proposed. By introducing the approximation quality of features term in the categories, according to the category distinguishing ability adjust term weight, using the purity formula feature weight comparison, the above problem is well solved, which can effectively improve the performance of text classification. The experiments have verified the feasibility of the proposed method.

Keywords: *Gini Index, Approximation Quality, Term Weigh, Text Classification*

1. INTRODUCTION

As a result of natural language is a kind of semi structured information in text classification [2], so prior to text classification, the structured handling is needed. The current text categorization is often used to use vector space model representing the text, but during the process of constructing the formation of the high dimensionality of the feature, for the majority of the classification algorithms are difficult to handle. Therefore, in order to reduce the burden of machine learning, it is necessary to carry out the dimension reduction. Feature dimension reduction methods advantages and disadvantages impact the final classifier performance. In the current text mining classification system [3] of performance has not achieved satisfactory situation, text feature dimension reduction technology is still the field of text classification a fundamental and decisive job. The common feature selection and weight method with TFIDF, mutual information, information gain and x^2 Statistics etc. But these methods have some defects, detailed analysis is as following:

(1) TFIDF

TFIDF is very simple, but more obvious shortcomings: First, a typical situation is that the total number is not change, if a feature term appears in a category, according to the IDF formula, when in the same category containing the features of a text is increased, the value of IDF is instead reduced

Obviously, it is not compatible with the actual situation; if a feature term in only one category, with the emergence of the text number increasing, the feature of the category distinguishing ability should be gradually become strong. Second, TFIDF method only consider the feature in the whole sample space distribution, and did not introduce categories information (attributes) into the formula of weight, so it is difficult to improve the precision of classification.

(2) Mutual information

Intuitively, score in the mutual information is higher, the degree of correlation between features and categories is greater. But the biggest drawback is not considering the characteristics of the term frequency.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

(3) Information gain

The shortage of the information gain is considered the feature appearing and not appearing, when feature term in most of text is not appear, weight characteristic feature is decided by the term not appearing, therefore, the effect of the information gain will be significantly reduced.

(4) "x^2"statistics

The "x^2" statistics is a good feature selection method in the classification, there are not enough when feature term in the whole training set is of very high frequencies, and in the designated class rarely appeared, the "x^2" statistics score will be higher , in fact this kind of word is needed to be filtrated.

Text classification is a key technology of the Web text mining, including data mining, machine Learning, neural networks, statistical and natural language processing and many other fields of study, and has a wide range of applications in information retrieval, information extraction, information filtering, automatic indexing, document organization etc. In Past 40 years, domestic and foreign scholars on the text classification technology conducted in-depth research, and obtained many achievements. Research of text classification has entered a practical stage. Chinese text classification technology research began relatively late, to 90 time, as a result of artificial intelligence technology, expert system was applied and developed in the field of text classification. Although many domestic and foreign scholars in the field of text classification has been gotten very fruitful research achievements, but Some key technology to some extent still affected the text classification system performance and practicality, especially at home, because of the late start, so the research level is relatively backward in many.

(1) The text representation

Text representation of text classification is an important research direction. At present, there are mainly 3 kinds of text representation model: Boolean model (Boolean Model), vector space model (Vector Space Model, VSM) and almost Rate model (Probabilistic Model). VSM in the knowledge representation has a huge advantage, not only is simple in concept, and the operation is convenient, which is currently the most popular text representation model and in many systems have very good application. However, the biggest deficiency is the assumption of text characters independent of each other, so the loss of text semantic context and underlying conceptual structure information. And the classification effect based on the purity of the Gini index probability model is not very good.

(2) Dimension reduction technology

Dimensionality reduction techniques impact text classification performance, when text is expressed in VSM, the dimension of the feature vector often up to thousands and even higher, most classification methods are unable to bear such a large set of features. If has no dimensionality reduction, not only the calculation can be costly, and a large number of not containing text classified information terminology is existed, which will cause the classification performance being greatly reduced. Dimension reduction method in Gini can be divided into feature selection (weight) and feature reconstruction. Feature selection is mainly TFIDF, mutual information, information gain and x^2 statistics.

In view of the several common feature selection and weight method is insufficient, many scholars put forward different modification, and the construction of the new feature selection and weight method, and these methods to some extent improve the text classification performance. Gini index was developed by Breiman etc. Putting forward the earliest, it is a kind of multi valued attributes splitting method, many scholars based on it carried out research on feature selecting and weight method, for example: Shankar adjusted feature weight application based on the basic principle of the Gini index, but due to used iterative method, which caused more time consuming; Charu based on the Gini formulas impurity degrees studied text feature selection problem; Shang Wenqian on the basis of Gini purity formula launched a study of feature selection, a variant of formula for feature selection, while these variant of formula and the other feature selection methods in different classifier performance were comparied. The classification effect is good, but the results are not very satisfied, and classification effect on several variants formula of feature selection are not significant, among which, classification effect of the purity formula of Gini index(Purity Gini Index, the PG for short) effect is slightly better, one of the main reasons is PG for the rare words in only one category was calculated as 1, which is not strictly distinguished from those feature term appearing in large amounts in a category, and not appear in other categories. In view of the above problems, an improved Gini index text classification method is put forward.

Journal of Theoretical and Applied Information Technology
30th September 2012. Vol. 43 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2. THE TEXT CLASSIFICATION BASED ON THE IMPROVED GINI INDEX

2.1 The Principle Of The Gini Index

Gini index [4, 7] is a kind of purity split method, it is suitable for the category, binary, continuous numerical and other types of fields, it was the Breiman that in 1984 in Classification and Regression Trees [6] proposed in the paper, the main algorithms of thought is [1]:

Hypothesis U is object not empty finite set, according to the category of attribute value, which can be divided into N different categories, and then Gini index number is:

$$Gini(u) = 1 - \sum_{i=1}^{n} [p(C_i \mid u)]^2 \quad (1)$$

The $P(c_i \mid u)$ expressed in node u conditional probabilities that object set U belongs to the C_i class, when the minimum value of $Gini(u)$ is 0, namely the nodes all objects belonging to the same category, which would be the most useful information; when the all objects in the node have the uniform distribution to category field, $Gini(u)$ gets maximum, as means can get minimum useful information. If the set according to a subset of attributes is divided into the number of K set ($u, j = 1, 2...k$), after splitting $Gini(u)$ is the

$$Gini_{split}(u) = \sum_{j=1}^{k} \frac{n_j}{n} Gini(u_j) \quad (2)$$

Among this, the n is objects number of u , n_j is sub-node object number. The basic idea is that the Gini index for each attribute to traverse all possible segmentation method, if which can provide the minimum Gini coefficient, as in this node can be made splitting criteria, whether for the root node or a child node.

2.2 The Improved Gini Index

The Gini index measure contained information in the attribute by complex degree forms, complexity is bigger, and attribute information contained in the volume is smaller. However, many scholars in the application of Gini index for text feature selection and weight tend to use pure metric form PG , i.e.

$$PG(u) = \sum_{i=1}^{n} [p(C_i \mid u)]^2 \quad (3)$$

The purity measure form conforms to the thinking habit, when purity value is greater, the information contained in the term is higher, it is said that the feature distinguishing capability is stronger, this article is also use the purity measurement method. Good feature weight methods are required to complete two tasks:

(1) Must be able to sort according to the characteristics of the categories of ability;

(2) Must be able to adjust feature weight and increase those with a high category distinguishing ability of feature weight

According to the Gini purity formula, although to a certain extent, which reflect the features distinguishing ability of a category, if do not adjust purity formula in the feature weight assignment, mainly include the following two questions:

First, if a feature term in only the Ci class appeared in small numbers, according to the formula (3), the result is equal to 1. However, when the feature terms [4] appeared in large numbers in the C_i class, and a few in other categories, according to the formula (3), the results should be less than 1. In fact, the latter feature weights should be bigger.

Second, if the feature term appeared only in the c_i class, no matter how many times, according to the equation (3), the results are equal to 1, it is not consistent with the actual situation, because the feature term only appeared in a category text in the case, only a small number of appearing is compared to appearing in large numbers, important degree is much smaller, which can even think of those few appearing words without any effect on decision classification making. Article [26] in feature selection, Gini purity formula was changed into the following form:

$$GiniTxt(u) = \sum_{i=1}^{N} P(c_i \mid u) P(u \mid c_i) \quad (4)$$

In the formula one factor $P(c_i \mid u)$ was replaced by $P(u \mid c_i)$, to a certain extent overcome the defect increasing rare word weight of Gini purity formula, but which will appear the new problems in Figure 1.

Figure 1 in the experiment may occur a kind of special situation, in which , $|c_i| = N \neq 0$,that is, in training set, the text total number of each category is N ; Case1 representation term t_1 only appearing number in c_1 text were x ; Case2 showed the term

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

t_1 only appearing number in c_1 and c_2 text were x and y ; Case3 was t_1 appearing number in c_1 , c_2 and c_3 text were x, y and z .When $x = y = z \neq 0$, the three kinds of circumstances by the formula (4) obtained the result is the same, but this is clearly not realistic, especially when $x = y = z$ and the value is relatively large, then tend to N , terminology in Case1 should have a strong category distinguishing ability, while Case2 and Case3 in terms of categories is weak.

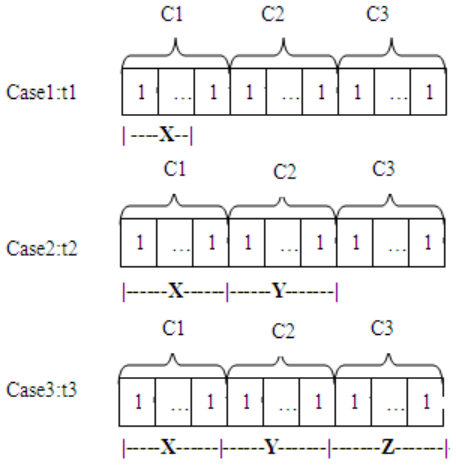

Figure 1: The question of improved the formula (4)

Analysis can be found, when the x, y, z and $|c_i|$ is not zero, the Case1 and Case2 in formula (4) results in a sufficient condition is equal to $c_1 / c_2 = |x| / |y|$; which makes the Case1 and Case3 in formula (4) results in an equal a sufficient condition is $x / y = |c_1| / |c_2|$, and $x / z = |c_1| / |c_3|$. Therefore, formula (4) to Gini purity formula adjustment method is not very effective.

In the purity measure formula, $P(c_i | u)$ means conditional probability object set U of the node u belong to the c_i class. In the text classification, according to one feature, it is c_i degree the text was accurately classified to the decision classes.

Which can be found, if $i \in \{1, 2, L, N\}$, and $P(c_i | u) \in [0.5, 1]$, then $P(c_i | u)$ is bigger, feature category distinguishing ability in Gini will be more strong; conversely, if $i \notin \{1, 2, L, N\}$, such that $P(c_i | u) \in (0.5, 1]$, this feature categories ability is poor. In the process of the feature weights adjustment, the former type of feature based on specific situation may appropriately increase the weight; the latter type of feature term should be given less weight. Therefore, introducing features term of a category of approximation quality of $\alpha_i(u)$ is defined as:

$$\alpha_i(u) = P(C_i | u) = \frac{|U \cap C_i|}{|U|} \qquad (5)$$

Definition of feature weight adjustment operator μ is for:

$$\mu_t(\alpha_i(u)) = \begin{cases} 1 & \alpha_i > t \\ \alpha_{i(U)} & \alpha_i \leq t \end{cases} \qquad (6)$$

Among which, the threshold $T \in (0.5, 1]$, used to control the quality of approximation feature category, when the category of the approximation quality of $\alpha_i(u)$ is superior to t, in most cases which has good category distinguishing ability, and was able to make as c_i class feature. At the same time, the introduction of t could increase the fault tolerance, especially in the results for the selected corpus inventory in the case of multi-subject text. Analysis shows that, the appropriate selection of T, which can improve the classification performance. Thus, by the μ operator, under the control of the T, in the c_i class appearance in great quantities, whereas in the other relatively few appearance to feature, are similar to think that these features only appear in c_i class, which can alleviate on certain level the first class of problems (Experiments t = 0.8).

In order to solve first and two prevalent in increasing rare words weight problems, based on the formula (4.6) the feature weights were adjusted further, and increase the high category distinguishing ability of feature weight, the lower ability characteristics maintain constant value, so, defining feature weight adjustment operator λ is for:

$$\lambda(\mu(i)) = \begin{cases} \alpha_i(U) \times |U| & \mu = 1 \\ 1 & 0 \leq \mu < 1 \end{cases} \qquad (7)$$

Through the above definition, μ operator and λ operator feature was adjusted, improved feature weight formula based on Gini purity is as follows:

$$NG(u) = \sum_{i=1}^{N} [\mu(p(c_i | u))]^2 \, \mathrm{X} \, \lambda(\mu(p(c_i | u))) \qquad (8)$$

3. EXPERIMENTS AND RESULTS ANALYSIS

At Windows XP environment, using the development platform Eclipse 3.2 and JDK 1.5, the algorithm was prepared and realized. Experimental machine configuration for 1.6GHz Processor, 512MB memory, 80GB Hard Disk. Experimental

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

data derived from Fudan University, Li Ronglu with the Chinese text classification corpus.

3.1 Experimental Content

Feature selection efficiency can be tested through the classification results, in order to verify the effectiveness of algorithms based on Gini index of feature weight method, and use the most Gini performance evaluation parameters: recall, precision and F1 value.

Experimental data came from a source of selected physical, economic and artistic three texts, the training set is 3 ×700 texts, and test set is 1254 sports texts, 1601 economy texts, 850 art texts. Design of two kinds of classifiers: KNN classifier (K=40) and NB classifier. And in the KNN classifier using TFIDF, Gini purity formula *PG* (3) and the improved weight formula NG(8), as well as in the NB classifier based on word frequency mutual information of MIDF and NG, respectively, for each category of recall, precision and F1 value were censused, by macro average overall assessed the effectiveness of feature weight method. Based on the frequency of mutual information MIDF is a kind of mutual information method, which mainly added the various frequency information to the mutual information in the formula, the specific formula is as follows:

$$MI_{DF}(f,c) = \log\frac{p(f,c)}{p(f)p(c)} = \log\frac{TF(f,c)\mathrm{X}|db|}{TF(f,db)\mathrm{X}|c|} = \log\frac{TF(f,c)}{TF(f,db)} + \log\frac{|db|}{|c|} \quad (9)$$

Among them: $TF(f,c)$ is the times in C text feature; $TF(f,db)$ is the number of occurrences of f feature in DB text sets; $|db|$ is text feature number in the entire set DB; $|c|$ is the feature number in c text of terms.

3.2 Results And Nanlysis

Figure 2-4 are respectively TFIDF, PG and NG three kinds of feature weight method, which choose a different feature number in KNN classifier on three kinds of classification performance evaluation index and macro average performance.

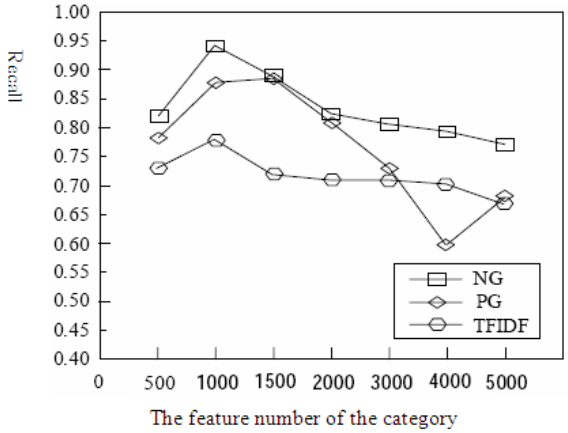

Figure 2: The TFIDF, PG and NG in KNN classifier recall performance

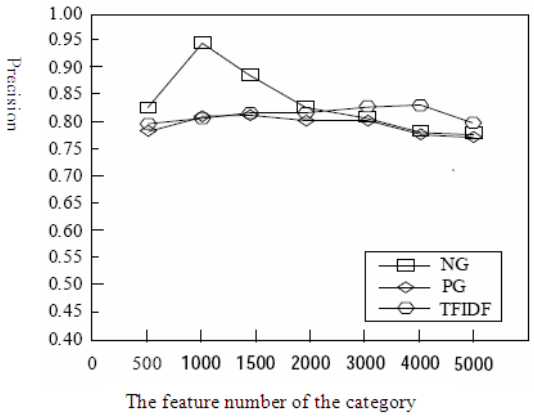

Figure 3: The TFIDF, PG and NG in KNN classifier precision performance

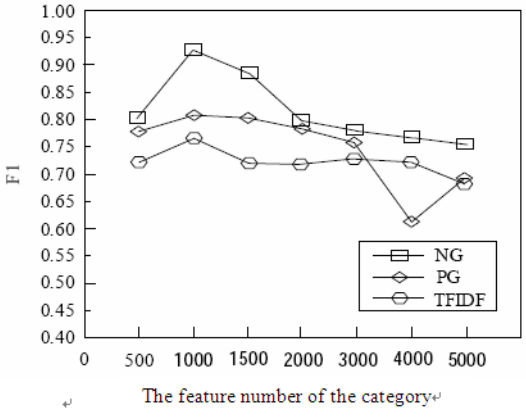

Figure 4: The TFIDF, PG and NG in KNN classifier F1 performance

Journal of Theoretical and Applied Information Technology
30th September 2012. Vol. 43 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

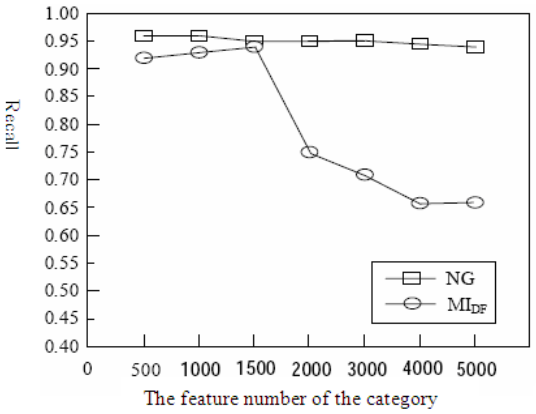

Figure 5: NG and MIDF in NB classifier recall performance

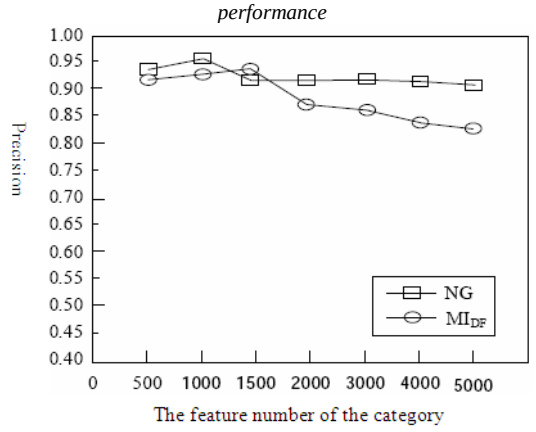

Figure 6: The NG and MIDF in NB classifier precision performance

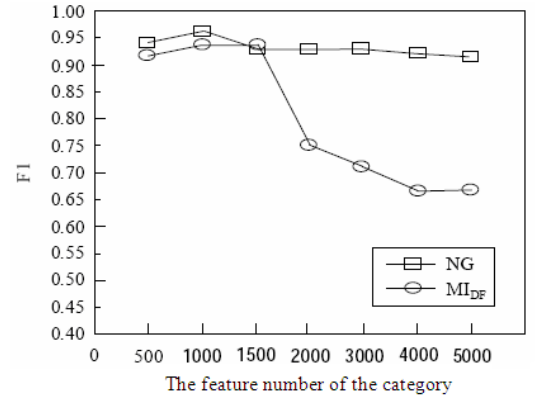

Figure 7: The NG and MIDF in the NB classifier F1value of the performance

Figure 5 to Figure 7 are respectively based on word frequency mutual information MIDF and NG two kinds of feature weight method, which choose a different feature number in NB classifier three classification performance evaluation parameters and make macro average performance.

From the experimental curve in Figure 2 to Figure 7:

(1) No matter how to adopt which classifier, based on Gini index feature weight method shows good classification effect. In the KNN classifier, NG classification results is better than TFIDF and Gini purity formula PG; And in the NB classifier, NG overall classification result is better than that of MIDF.

(2) In the experiment, the Gini purity formula PG weighted characteristics, the classification effect is not so good, one of the main reasons is PG for those in only one category in the rare words is calculated as "1", which is not strictly distinguished from those in large amounts in a category, in other categories do not appear in the feature.

For example: in this paper the experimental data, economy class has nearly more than 2000 feature by the purity of formula PG results for 1, from more than 2000 features selected in 1000 and there is no certain standard, so the resulting classification effect is not very good. In addition, the [7] in experiments using Gini purity formula for feature selection, after the selection of 1000 characteristics using TFIDF weighted, but according to the IDF formula, assuming the same total number of text is unchanged, if a feature term all appear in a category, when in the same category contains the features text to increase, the value of IDF will decrease. Therefore, in paper [7] classification result is not a particularly good reason may be appear in the example above.

(3) From the selection of features number of different perspectives, feature number in the interval [1000, 1500], or in the vicinity of interval, the classifier performance will reach a maximum, before the characteristics dimensionality reduction, three types of text features number is more than 30000, so that means after the deletion approximately 80%-90% feature, not only won't reduce the performance of the classifier, but improve the classification effect.

4. CONCLUSION

On the analysis of the Gini index and feature weight based on the basic principle, through the introduction of category approximation quality as a evaluation feature weight importance standard, the feature weights are further adjusted, a new feature weight formula based on Gini index has defined, finally, in the KNN classifier and NB classifier, respectively has compared with TFIDF, PG and MIDF, which validate this new feature weight method on the classifier performance has improved to some extent.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ACKNOWLEDGMENT

This work was supported by the project of education department of Shaanxi Province research foundation, the project number is "2011JM1010". By the project of Weinan Normal University special research foundation, the project number is "12YKZ052". By the project of science and technology department of Shaanxi Province research foundation, the project number is "2012JM8048".

REFERENCES

[1] Wenqian Shang, Houkuan Huang, Yulin Liu, "Reasearch on feature selection algorithm in text classification based on Gini coefficient", *Journal of Computer Research and Development,* Vol.43, No.10, 2006, pp.1688-169.

[2] Luka Cehovin, Zoran Bosnic1, "Empirical evaluation of feature selection methods in classification", *Intelligent Data Analysis*, Vol.14, No.3, 2010, pp.265-281.

[3] Levent OEzguer,Tunga Guengoer, "Text classification with the support of pruned dependency patterns", *Pattern Recognition Letters,* Vol.31, No.12, 2010, pp.1598-1607.

[4] Heum Park, Hyuk-Chul Kwon, "Improved Gini-Index Algorithm to Correct Feature-Selection Bias in Text Classification", *IEICE transactions on information and systems,* Vol.31, No.12, 2011, pp.855-865.

[5] Levent OEzgueir, Tunga Guengoer, "Optimization of dependency and pruning usage in text classification", *Pattern analysis and applications*, Vol.15, No.1, 2012, pp. 45-58.

[6] Liu RL, "Interactive high-quality text classification", *Information Processing& Management*, Vol.44, No.3, 2008, pp.1062-1075.

[7] Pu Wang, Jian Hu , Hua-Jun Zeng ,Zheng Chen, Using Wikipedia knowledge to improve text classification, *Knowledge and information systems,*Vol.19, No.3, 2009, pp.265-281.