Scientific Research and Essays Vol. 5(22), pp. 3405-3410, 18 November, 2010
Available online at http://www.academicjournals.org/SRE
ISSN 1992-2248 ©2010 Academic Journals

Full Length Research Paper

Feasibility research of text information filtering based on genetic algorithm

Zhenfang Zhu[1*] and Peiyu Liu[2]

[1]School of Information Science and Engineering, Shandong Normal University Ji'Nan, 250014, China.
[2]Shandong Provincial Key Laboratory for Distributed Computer Software Novel Technology, Ji'Nan 250014, China.

Accepted 11 August, 2010

For the problem of content-based information filtering, this paper introduces genetic algorithm to solve this problem, because it could find optimal solutions within global context. In order to illustrate the effectiveness of this approach, it gives a new approach based on set theory, and it also gives an experiment. From the theoretical proof and the experimental results, we could see that the method is feasible and could obtain better information filtering results. With the development of information technology, the Information world provides several network information to computer users, however, people are also inevitably exposed to a lot of spam, while they enjoy the convenience of information. Hence, the network information filtering arises at this historic moment. The text information filtering (Belkin and Croft, 1992; Liddy et al., 1994; Xuan-Jing, 2003) is a process that extracts the texts from a large amount of text stream in order to meet specific user requirements.

Key words: Information filtering, content-based, feasibility analysis, convergence.

INTRODUCTION

In the current research, the main research of information filtering are collaborative filtering and content filtering, and the content-based method is the focus of information filtering research, which is divided into statistical-based method and learning-based method. At the same time, construction and updating of the filter template are the cores of the machine learning-based content filtering method.

Genetic algorithm (Holland, 1975), arising from the 70s of last century, in which many institutions and researchers have done a lot of in-depth research and made a number of important results, has made its applications extend quickly to the optimization, search, machine learning and other fields. Consequently, it has gradually developed into a computing model, which could solve optimization problem through a simulation of the natural evolutionary process.

*Corresponding author. E-mail: zhuzhfyt@163.com. Tel: 86-13793100702. Fax: 86-0531-86180514.

Text information filtering based on content is an important part of machine learning. At the beginning of the research, genetic algorithm is applied to solve some simple learning problems, for example, Holland and Reitman proposed CS-1 system (Holland, 1992), which is the first time to apply genetic algorithm into solving maze problem, and Goldberg (Goldberg, 1989) also introduced genetic algorithm into engineering control. These researches produced a real sense of machine learning based on genetic (Genetic-based machine learning, GBML).

During the research, we found that the studies, which introduced genetic algorithm into the text information process are very few, particularly fewer with text information filtering. And most of these studies are focused on feature selection and generating filter template in the practical application by using genetic algorithm. Genetic algorithm was introduced into feature selection for the first time by Burns and Danyluk (2000) and Pan et al. (2004) for the first time introduced feature selection based on genetic algorithm into the field of text classification. Since then, researchers have proposed

many improvement projects, our research group (Zhao et al., 2009) also proposed an adaptive genetic algorithm and applied it into feature selection. In recent years, besides Lv (2007), our group gave a method to construct and improve the filter template and filter model (Zhu et al., 2009) using genetic algorithm in text information filtering. In Zhi-Long's research, the genetic algorithm was used to optimize the template, and there is no direct practical application of generating profile with genetic algorithm, while in the other research group (in this study), there is no corresponding theory proof, rather than its application.

PROBLEM DESCRIPTION

Text filtering is an important part of information filtering, TREC-9 gives the definition of text filtering (Belkin and Croft, 1992): According to the given user needs, text filtering system establishes a filter template (profile), which could select relevant texts automatically from text stream. Using this profile, text filtering system automatically accepts or rejects the texts and correct filter template adaptively according to feedback information.

To a certain extent, text information filtering can be regarded as a type of binary text classification; the text to be filtered will be mapped to a legitimate or illegal text collection. This process can be expressed with mathematical language as follows:

For each $< d_i, c_i > \in D \times C$

Where D is the document set to be filtered, d_i is one document of D, C is the set of category, and it has two value called filtering text collection and normal text collection, c_i is a value of C. To determine the value of document C, if the value is true (T), the text d_i is a number c_i, otherwise, it does not belongs to c_i. So the process of information filtering is described by constructed function $\alpha : D \times C \Rightarrow \{T, F\}$

Information expression

Training texts and the texts extracted from Internet are mapped into a vector of n-dimensions in the vector space model, each dimension is made up of a feature and its weight, and therefore, a text D could be formally expressed as follows:

$$D =< w_1, w_2, \ldots, w_n > \quad (1)$$

From Equation 1, w_1 is the first dimension of the vector, it symbolizes the weight of the first feature.

Encoding and initialization

During the optimization process of genetic algorithm, it is necessary to encode the problem space first. In the text filtering, an improved binary encoding method is as follows:

(1)Generate a random binary sequence using the random generator;
(2)Multiply the aforementioned binary sequence by word segmentation results;
(3)Generate the initial population consisting of individuals, which would be calculated in Equation (2).

Fitness

Fitness function indicates the individual's ability to adapt to the environment; we often select different fitness function for different problems. In text information filtering, template must have much similarity with the texts to be filtered in the same category and less similarity in different categories. Therefore, it is a desirable program (Zhu et al., 2009) to select the similarity between the individuals as the fitness function in application. However, this approach will inevitably arise the local optimal solution problem in the optimization process, and it is also found that the evolution curve is too flat when we select the average fitness function. At last, we choose the absolute value of the subtraction similarity between two individuals as the criteria for evaluating the individual merits.

Definition 1: Similarity*:*

$$\begin{aligned} &\text{Similar}(individual[i], individual[j]) = \\ &\cos(individual[i], individual[i]) = \\ &\cos < weight[i], weight[j] >, j \neq i \end{aligned} \quad (2)$$

Definition 2: Average similarity:

$$\text{Similar}[insividual[i]] = \frac{\sum_{j=1}^{group\_size} \cos < weight[i], weight[j] >}{(group\_size - 1)} \quad (3)$$

From Equation 3, $weight[i]$ and $weight[j]$ are the weight of $individual[i]$ and $individual[j]$.

Definition 3: Fitness function:

$$Fitness[i] = |\text{Similar}[individual[i]] - \text{Similar}[individual[i+1]]| \quad (4)$$

$\text{Similar}[insividual[i]]$ and $\text{Similar}[insividual[i+1]]$ are the average similarity as shown in definition 2.

CONVERGENCE

In the genetic algorithm convergence analysis, the main methods are schema theorem (Holland, 1992), stochastic theory (Kai, 1989), dynamics (Guo et al., 2002), and so on. Wang et al. (1996) proposed a method based on superior class and similarity, and this article proved the convergence of information filtering problem learning from this approach.

Problem reduction

Text information filtering is a natural language that understands problem to some extent, and the understanding of natural language belongs to multi-layer neural network architecture, so it is difficult to discuss the convergence in the multidimensional space.

Meanwhile, from the problem space description, it could be observed that network information filtering, based on genetic algorithm, is a mapping from a high-dimensional space of data-processing to a binary space. Therefore, we can transform the text information filtering problem to two-dimensional space and discuss its convergence or divergence.

Relevant definitions

In the convergence analysis, several definitions involved are as follows:

Definition 1: Problem solution:

Suppose I is the problem space of text information filtering, and $C = \{1,2.....n\}^k$ is used to code the problem for each possible solution of C, it has a corresponding point in problem space I. Otherwise, it does not established necessarily.

Definition 2: Space change function

Suppose f is the space change function, which is called intensity function:

$$f(I) = \begin{cases} \text{Objective function} & \forall i, has \quad a \quad value \\ \text{Minimum of objective function} & \forall i, doesn't \quad has \quad a \quad value \end{cases}$$

In this function, the domain is I, and the value field is as the same as the objective function. Like doing this, we could map the problem into a two-dimension space. In the two-dimensional space collection, we can define the relevant class to discuss the complex issues of convergence and divergence.

Definition 3: Concept of class

The set S is called a class , if and only if $S \subseteq I$, the intensity of class S in the population category POP is the average intensity of all the individuals in the population; for S, if $f(S, POP) \geq f(POP, POP)$, then S has become the dominant class in the population; if the class S in any of the populations are dominant, it is called similarity class.

From the aforementioned concept, it is observe that the intersection of the similarity class is the set of optimal solutions, but the stability can be easily destroyed during the evolutionary process. Following the aforementioned reasons, if the genetic process can find the optimal solution, it is necessary to ensure that the similarity class solution set is not replaced or disappears, so, some assumptions are explained.

Convergence hypothesis

If S is a similarity class, POP is the population, for any competitive class $S^{'}$,if $f(S^{'}, POP) \geq f(S, POP)$, in the following two conditions, there must be a condition established:

(1) If a individual contains the set $S^{'}$, and it also contains the set S, it means that $S^{'} \cap POP \subseteq S$;

(2) If the intersection of $S^{'}$ and S is $S^{''}$, then $f(S^{'}, POP) \geq f(S, POP)$.

The similarity class of the above conditions can not be replaced to ensure the stability of the optimal solution set, which is conducive to a final solution. So in this case, population must be convergent.

Convergence analysis of the problem

According to the aforementioned definition and the description of genetic algorithm to solve the text information filtering, we have transformed the text information filtering into the hypothetical question with

Table 1. Training set distribution.

Category	Violence	Pornography	Computer	Environment
Number	276	192	1358	1218
Category	Agriculture	Economic	Politics	Sports
Number	1022	1601	1026	1254

Table 2. Contingency table.

	Text number classified correctly	Text number classified falsely
Judge to the category	a	b
Judge to other categories	c	d

convergence, that is, the text information filtering problem is convergent.

APPLICATION

Training sets

The training set comes from the corpus sorted by Li et al. (2004) (Natural Language Processing Group, the International Database Centers, Department of Computer Information and Technology, Fudan University), in this corpus, there are 20 categories and 9804 texts. In these texts, 11 categories have less than 100 texts, and 6 categories have more than 1000 texts, such as Computer, Environment, and Economy and so on. Because this classifier would be used to filter spam, the project team collected two categories called Pornography and Violence. The distribution of training set is shown in Table 1

Test sets

In order to test the performance of the algorithm, we set two test sets as thus explained.

Closed test

The first test sets(set), which comprised a total of 902 texts, consists of two parts, one is the categories, which does not have more than 100 texts of corpus sorted by Li et al., (2004), and the other comes from the training sets, and each category randomly selects 50 texts.

Open test

The second test set, which comprised a total of 600 texts, comes from Chinese text corpus -TanCorpV1.0 Corpora. This corpus sorted by TAN Song-bo has 2 levels of about 14150 texts, and the first level has 12 categories, from which we select 3 categories that are relevant to training set such as economics, computer and sports.

Performance validation

Test for single category

The most common evolution methods of classification and filtering are precision (p) and recall(r), for each category o, researchers use contingency table to calculate precision (p) and recall(r). Table 2 gives an example of contingency table (Miao and Wei, 2007).

On the basis of contingency table, we could define precision (p) and recall(r) as follows:

$$p=\frac{a}{a+b}\ r=\frac{a}{a+c} \tag{5}$$

Test for overall categories

The contingency table as shown in Table 2 could only evaluate single category, and if the classifier is to be evaluate, another parameter named Micro-averaging should be used:

$$\bar{r}=\frac{\sum_{1}^{|c|} r_c}{|c|}\ \bar{p}=\frac{\sum_{1}^{|c|} p_c}{|c|} \tag{6}$$

Table 3. Accuracy on closed sets.

Agriculture	Political	Sports	Violence
79.969	74.364	75.211	96.053
Environment	Economic	Computer	Pornography
83.345	91.585	87.468	98.446

Table 4. Accuracy on open set.

	Sports	Economic	Computer
Accuracy	46.154	90.697	84.314
Recall	79.969	75.000	82.690

Table 5. Statistical data of filtering test.

	Text number	Be filtered	Accuracy(%)
Illegal	300	293	97.67
Legal	300	257	85.67

RESULTS

Results on closed test

From Table 3 it is shown that in those two categories, which have lower effect to classification, the texts are similar to each other. For example, the political category has a lot of texts with information regarding other categories, as a result of this, these categories (e.g. political) have a lower accuracy.

In order to research the effect of the classifier, Micro-averaging of the data in Table 3 was calculated, and the Micro-averaging is 85.810. From the calculation, the average accuracy of data is $\overline{p}$ =85.810, our result was compared with several basic methods, which also has been applied to the data set Reuters-21578 in recently years. In these methods (Su et al., 2006), the method is better than Naive Bayes method, Tree method and nearest neighbour classification method.

Results on open test

From the described experiment, our method could have a good effect, but we can not exclude the possibility of excessive fitting because a closed set comes from training sets. For these reasons, it is necessary to carry out the test in open set.

From the data in Table 4, we can find that the categories such as computer and economic decline a little, meanwhile, the category sports has a big decline. During the analysis discovered that there were great differences between the training sets and test sets, the text of training sets were theory of physical education, but the text of test sets were about sports from practice and sports news.

Results for filtering

At last, the method designed would be used to filter spam, so it is necessary to test the filtering effect of the classifier.

During the test, we divided the closed set into two parts, one of which is legal texts set and the other, illegal texts set. The illegal texts set involves all the texts, which belong to the categories of pornography and violence, and the legal texts set involves the rest of the training sets. The results are shown in Table 5.

From Table 5, it was discovered that the improved algorithm has better performance, at the same time, the category of illegal texts has a better effect because of its distinctive characteristics, and finally we must filter such spam. In order words, the tactics of classification have greater capability than some of the classification methods, and the improvement in this paper is effective.

From the theoretical and experimental analysis, it is seen that the method can solve the problem of text information filtering effectively.

DISCUSSION

The next step is to improve network information filtering model based on genetic algorithm, including giving an improved genetic algorithm based on ant colony algorithm to solve the local optimization problem, using artificial neural network to solve the dimensionality problem of item space by construct semantic network, and so on.

ACKNOWLEDGMENTS

This work is supported by National Nature Science Foundation of China (60873247), Nature Science Foundation of Shandong Province (ZR2009GZ007), Independent Innovation of High and New technology Special Project of Shandong Province (2008ZZ28). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have

improved the presentation. Zhenfang ZHU, Doctoral students, Main research: Network Information Security, Genetic Algorithm, Management of Engineering and Industrial Engineering; Peiyu LIU, professor, doctoral supervisor, Main research: Network Information Security, Network System Planning, Development of the Network Information Resources.

REFERENCES

Belkin NJ, CroftW B (1992). Information filtering and information retrieval: two sides of the same coin? Communications of the ACM 35(12): 29-38.

Burns BD, Danyluk AP (2000). Feature Selection vs Theory Reformulation: A Study of Genetic Refinement of Knowledge-based Neural Networks[J]. Machine Learning,38: 89-107.

Goldberg DE (1989). Genetic Algorithms is Search, Optimization,Machine Learning. Reading MA: Addison Wesley.

Guo D, Liu D, Zhou Chun G, et al (2002). Dynamic analysis of GA's convergence and its application [J]. J. Comput. Res, Dev., 39(2): 225-230.

Holland JH (1975). Adaptation in natural and artificial systems. Ann Arbor, MI: The University of Michigan Press.

Holland JH (1992). Adaptation in natural and artificial system: An introductory analysis with application to biology, control, and artificial intelligence. 1st edition, Ann Arbor, MI:The University of Michigan Press, 1975;2nd edition, Cambridge,MA: MIT Press.

Holland JH (1992). Adaption in Nature and Artificial Systems. MIT Press.

Kai LC (1989). Probability Tutorial [M]. Translation by Liu Wen,WU Rangquan. Shanghai: Shanghai Science and Technology Press.

Li RL,Tao XP,Tang Hu YF (2004). Using mxaimum entropy model for Chinese text categorization. In proceedings of the 6ht Asia-Pacific Web Conference. Hnagzhou,China, pp. 578-587.

Liddy ED, Paik W, Yu ES (1994). Text categorization for multiple users based on semantic features from a machine-readable dictiony.ACM Transactions on Information Systems 12(3): 278-295.

LV Z (2007). Research of Adaptive Text Filtering based on Genetic Algorithm [D]. Harbin Engineering University Master's degree thesis.

Miao D-G, Wei Z-H (2007). Principle and application Chinese text information processing [M]. Beijing: tsinghua university press.

Pan L, Zheng H, Zhang Z, Zhang J (2004). Genetic Feature Selection for Texture Classification.Geo-spatial Information Science (Quarterly). 7(3):163-173.

Su JS, Zhang BF, Xu X (2006). Advances in machine learning based text categorization. J. Software, 17(9): 1848-1859.

Wang L, Hung Y, Hong J (1996). On the convergence of genetic algorithms [J]. Journal of Computers, 19 (10): 794-797.

Zhao L-N, Liu P-Y, Zhu Z-F (2009). improvement and application of adaptive genetic algorithm in feature selection[J].Computer Engineering and Applications, 45 (7): 39-41.

Zhu Z-F, Liu P-Y, Zhao L-N, Lv T-X (2009). Research of feature weights adjustment based on Semantic paragraphs matching. ICIC Express Letters, 4(2): 559-564.