Web Text Feature Extraction with Particle Swarm Optimization

Song Liangtu†,††, ***Zhang Xiaoming***†

† Institute of Intelligent Machines, Chinese Academy of Sciences,Hefei,230031 China
††Automation Department, University of Science and Technology of China,Hefei,230036 China

Summary

The Internet continues to grow at a phenomenal rate and the amount of information on the web is overwhelming. It provides us a great deal of information resource. Due to its wide distribution, its openness and high dynamics, the resources on the web are greatly scattered and they have no unified management and structure. This greatly reduces the efficiency in using web information.Web text feature extraction is considered as the main problem in text mining. We use Vector Space Model (VSM) as the description of web text and present a novel feature extraction algorithm which is based on the improved particle swarm optimization with reverse thinking particles (PSORTP). This algorithm will greatly improve the efficiency of web texts processing.

Key words:
Web mining, VSM, Particle swarm optimization, Text feature extraction

1. Introduction

The web is growing at a tremendous rate. It contains a huge amount of unstructured, distributed data. This content provides a great potential source for information extraction that needs to be filtered, organized, and maintained in order to permit an efficient use[1]. However, Due to its wide distribution, its openness and high dynamics, the resources on the web are greatly scattered and they have no unified management and structure. This greatly reduces the efficiency in using web information. How to search for the information from the Massive web information speedily and accurately has become a major problem[2]. There was an urgent need of tools that can quickly and effectively find resources and knowledge from the Web. The main form of information on the web is web text. So how to process these Web Texts becomes the key question. Data mining can help discovering potential knowledge and information from mass raw data and effectively solve the problem that information is abundant but knowledge is lacking. However, most classical Data Mining tools require structured information. Therefore, Web-Based Text Mining has become a new theme of Data Mining.

In this paper, a review of Web mining is presented outlining the techniques of Web mining and data processing. A new Web text feature Extraction method is proposed that a modified particle swarm optimization algorithm is used in it.

2. Web Mining and Description of Web Text

2.1 Definition of Web Mining

Web mining can be broadly defined as the discovery and analysis of useful information from the World Wide Web[3]. This broad definition on the one hand describes the automatic search and retrieval of information and resources available from millions of sites and on-line databases, i.e., Web content mining, and on the other hand, the discovery and analysis of user access patterns from one or more Web servers or on-line services, i.e., Web usage mining. Web Mining targeted large, heterogeneous, distributed data of the Web. The Web is a semi-structured or unstructured. Therefore semantic understandable absence of machinery, natural language understanding and text processing are the main methods used in the fields of Web mining.

2.2 Description of Web Text

Most of the information in the Internet is in the form of Web texts. How to express this semi-structured and unstructured information of Web texts is the basic preparatory work of Web Mining. Vector space model is one of good method which is applied more in recent years. It is used to represent each document as a vector of certain weighted word frequencies. In a document vector, the value on each term axis represents the importance of the term in the document. Therefore the document vector is merely a set of all terms with a value representing the importance of each term in the document, like $V(d) = (t_1, w_1(d);...;t_i, w_i(d);...;t_n, w_n(d))$. t_i is the term i and $w_i(d)$ is the value representing the importance of term i in document d. $w_i(d) = \psi(tf_i(d))$ is always defined as a function of ti's frequency- $tf_i(d)$ in document d. TF-IDF is a common method of determining the weights of the term in VSM.The weight of the term is proportional to Term Frequency(TF) and with the Document Frequency(DF) in inverse proportion[4]. The following is a commonly used formula for calculating Term Weight:

Manuscript received June 5, 2007.
Manuscript revised June 20, 2007.

$$W_{ik} = \frac{tf_{ik} \log(\frac{N}{n_k} + 0.01)}{\sqrt{\sum_{k=1}^{n} (tf_{ik})^2 \times \log^2(\frac{N}{n_k} + 0.01)}} \qquad (1)$$

tf_{ik} is the frequency of T_k in document D_i . N is the total number of sample documents . n_k is the number of docments which have the term T_k. The terms which appear in the Heading, subheading and keywords always should be emphasized.

2.3 Feature Extraction

Words and phrases is the basic elements of a normal document and there is certain regularity with the frequency of each term in different documents. So we can use it to distinguish documents which have different contents.

Text feature vectors are obtained through algorithms of words segmentation and statistical approach of term frequency. The dimension of the vector is immense by using this approach .If no disposal is done to the original text vector , the vector calculating expenses will be tremendous and the efficiency of the whole process will be extraordinary inefficient. Therefore, we need to process the text vector for further refinement on the basis of that the original meaning is ensured and the feature vector is rather compact. Extracting text features is a NP-complete problem. Currently, many novel approaches on the study of feature extraction is proposed. In the study of Liu li, a novel term selection and weighting approach based on key words is presented[5]. ZOU Juan proposed a method of Chinese text eigenvalue extraction using multiple heuristic rules in their paper[6]. We believe that the Web texts can be seen as a multi-dimensional information space which is made up of the feature terms. Then the text features selection is just the optimization process in the multi-dimensional information space. Therefore efficient optimization algorithm is needed in finding the text features. As a general intelligent search algorithm , Particle Swarm Optimization(PSO) is discovered through simulation of a simplified social model and it can search the multidimensional complex space efficiently. In this paper, a novel approach of Web text feature Extraction with particle swarm optimization algorithm is presented.

3. Web Text Feature Extraction with Improved PSO Algorithm

Particle Swarm Optimization (PSO) was first introduced by James Kennedy and Russel C.Eberrhart in 1995 and it was discovered through simulation of a simplified social model. It can search the multidimensional complex space efficiently through cooperation and competition among the individuals in a population of particles[7]. We think that the text features selection can be transformed into the optimization process in the multi-dimensional information space. Firstly , the Web text should be coded and the text vector is viewed as the position vector of a particle. Because the number of the text features is unknown, we proposed a novel particle swarm optimization with alterable dimension. Reverse thinking particles are also added in the algorithm in order to improve the global search ability of the PSO. The main process of the approach is presented in the followings:

3.1 The Selection of Text Features

Before extracting text feature, we need to pretreat the text to get the feature terms. That extracting good features is a very important means and a basic require for text processing. From the point of view of natural language understanding, the nouns and verbs form the main content of a text. Currently, the main methods used in text pretreatment include stem-ming of English documents and segmentation of Chinese documents. In this paper, our approach is the positive maximal match segmentation based on dictionary.

3.2 Code The Feature

The weight of a feature term of the web documents is one dimension of a particle's position vector. Because the number of the document features is unknown, we proposed a dynamic PSO. The dimension of particles' position in the algorithm is dynamic in order to adapt to the uncertain number of the document features. The particles in the swarm is initialized by the random selected feature terms from the term list after word segmenting. The particles' position vector is mainly formed from the standardized web document vector model.

3.3 Optimization Procedure

When the PSO starts, a population of particles is generated first with random positions and a velocity is random assigned to each particle[8]. The fitness of each particle is then evaluated according to the objective function. Then the particles begin to search for the best solution. Each particle's trajectory is adjusted by dynamically altering the velocity of each particle, according to its own search experience and other particles' experience. The position vector and the velocity vector of particle j in the d-dimensional search space can be represented as X_j=

(x_{j1},x_{j2},x_{j3},…,x_{jd}) and V_j= (v_{j1},v_{j2},v_{j3},…,v_{jd}) respectively. According to the objective function, let the best position of each particle be P_j = (p_{j1},p_{j2},p_{j3},…,p_{jd}) and the best position of all the particles be G = (g_1,g_2,g_3,…g_d). The formulas which the particles use to adjust its position and velocity is:

$$V_j = V_j + c1*rand1()*(P_j - X_j) + c2*rand2()*(G - X_j) \qquad (2)$$

$$X_{j+1} = X_j + V_{j+1} \qquad (3)$$

where c_1 and c_2 are acceleration coefficients and are always made 2, rand1() and rand2() are random numbers in [0,1][9].

The first part of (2) represents the previous velocity. The second part represents the personal experience. The third part represents the collaborative effect of the particles and it always pulls the particles toward the global best solution which particles found so far.

At each iteration of PSO, the velocity of each particle is calculated according to (2) and the position is updated according to (3). Generally, a maximum velocity vector is defined in order to control the V_j . Wherever a v_{jd} exceeds the defined limit, its velocity will be set to be $v_{\max d}$. If a particle finds a better position than the previously found best position, it will be stored in memory. The algorithm goes on until the satisfactory solution is found or the predefined number of iterations is met.

3.4 Evaluation Function

Evaluation function is an important criterion for judging the merits of particle location. The location of each particle is composed of the feature weight of the same documents when the terms of the web document are acquired. So the best particle should be able to reflect the web document well and it also should contain other particles distribution information.

Therefore, the individual's fitness should be measured by all particles which present the same document. The more similar they are, the bigger their fitness should be.

The fitness function should be defined as following:

$$fitness(P^d{}_i) = \sum_{j=1}^{swarm\_size} similar(P^d{}_i, P^d{}_j) / swarm\_size \qquad (4)$$

The similarity of vectors can be measured by the angle between two vectors or by the distance between two particles. It is easy to use the angle between two vectors and the formula is like following:

$$similar(P,Q) = \frac{\sum_{i=1}^{n} p_i \times q_i}{\sqrt{\sum_{i=1}^{n} p_i{}^2} \times \sqrt{\sum_{i=1}^{n} q_i{}^2}} \qquad (5)$$

p, q is the elements of vector *P* ,*Q*. In addition, the dimension of a vector is also an important criterion. If several particles have the same fitness, smaller the dimension is , better the particle is.

Therefore, we need to increase the penalty function p(R) in the evaluation function in order to lead the particles to the best region. In this paper, we set penalty function p(R) as following:

$$p(d) = \frac{2^{|d|+1}}{1+2^{|d|}} \qquad (6)$$

Where d is the dimension of vectors.

The final evaluation function is formed by adding the penalty function to the fitness formula:

$$value(d) = p(d) \times fitness(d) \qquad (7)$$

4. Feasibility Analysis and Design

Theoretically, the dimension of particles' location should be dynamic. But in fact ,it isn't. We use empty elements to fill the diminished dimension. When the weight of one feature is less than 0.05 , we also replace it with empty element. The structure of the particles' location is like following:

W_1	W_2	……	W_m	B_1	B_2	……	B_n

Fig.1. the structure of a particle

After the particle design is finished ,we use the algorithm of particle swarm optimization with reverse thinking particles[10]. The trait of the reverse thinking particles is that they do not update their position according to (3). They move to an opposite direction. You can look at the Fig. 2. They are selected randomly in the particle swarm and their new formula which will update their position is :

$$X'_j(t+1) = X_j(t) - V_j(t) \qquad (8)$$

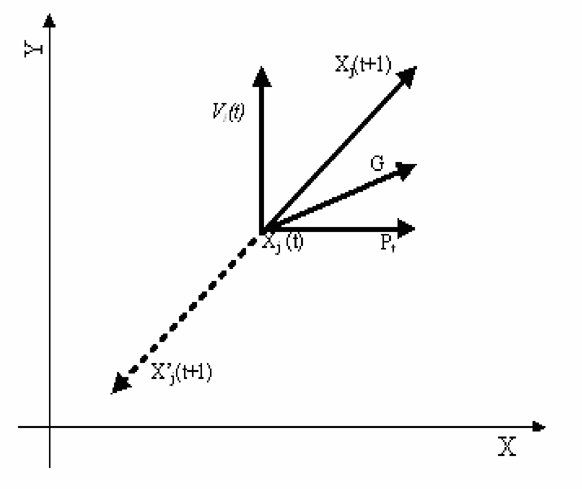

Fig.2. trait of reverse thinking particles

When the algorithm starts, normal particles run following the formula (2) and (3) and the reverse thinking particles run following the formula (2) and (8). If one of the reverse thinking particles finds a better solution than G(the best position of all the particles), the reverse thinking particle will change to be a normal particle and one particle will be selected randomly from the normal particles to become an reverse thinking particle. If the G is acquired by the normal particles, the reverse thinking particle will still be opposite. When the reverse thinking particles' position reach the border of the problem space, they will be set to be the normal particles which has the best position G. The flow chart of PSORTP is depicted graphically in Fig 3

5. Experiments

The main purpose of the experiments is to verify the effectiveness of the novel web text feature extraction algorithm. The main experimental materials is the web pages(HTML) which is Parsed by HTMLParser.
We use two criterions: accuracy rate(AR) and reduction rate(RR).

$$AR(t) = \frac{n_{hl}(n_{ao}(t))}{n_{ao}(t)} \times 100\% \tag{9}$$

$$RR(t) = \frac{n_{fv}(t)}{n_{ov}(t)} \times 100\% \tag{10}$$

$n_{ao}(t)$: the number of feature terms of text t after optimization;

$n_{hl}(n_{ao}(t))$: the number of high weight feature terms of HLSegment[11];

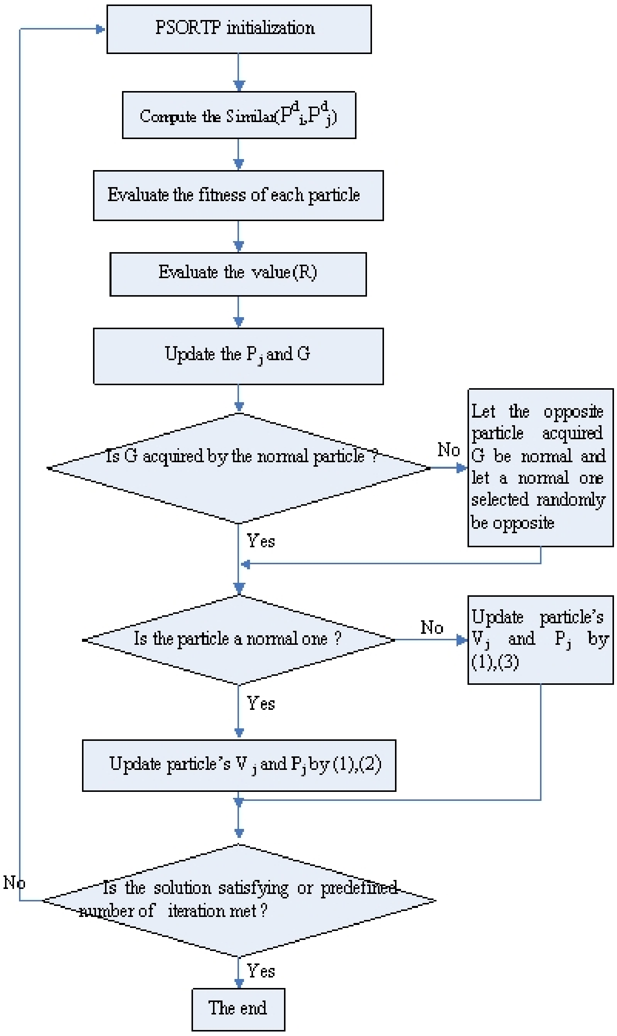
PSORTP initialization
Compute the Similar(P^d_i,P^d_j)
Evaluate the fitness of each particle
Evaluate the value(R)
Update the P_j and G
Is G acquired by the normal particle ?
No
Let the opposite particle acquired G be normal and let a normal one selected randomly be opposite
Yes
Is the particle a normal one ?
No
Update particle's V_j and P_j by (1),(3)
Yes
Update particle's V_j and P_j by (1),(2)
No
Is the solution satisfying or predefined number of iteration met ?
Yes
The end

Fig.3. The flow chart of PSORTP

$n_{fv}(t)$: the number of feature terms of text t;

$n_{ov}(t)$: the number of original terms of text t;

We tested four web pages. The result is the following:

Table 1: The test results of the effectiveness of the novel web text feature extraction algorithm.

No.	Text size	$n_{ov}(t)$	$n_{ao}(t)$	$AR(t)$	$RR(t)$
1	996	396	187	88.9%	47.2%
2	1117	331	156	93.2	47.1%
3	1156	293	98	95.8	33.4%
4	178	88	41	96.1	46.6%

From the results, we can see that by using this method, the text features are optimized. Based on the text feature extraction algorithm, we can greatly improve the efficiency of web documents processing.

6. Conclusion

In this paper , a novel Web text feature Extraction algorithm is proposed. The algorithm is based on the particle swarm optimization with reverse thinking particles and the structure of the particles is also improved.

The algorithm can search the multidimensional complex space efficiently. This method present a good idea in feature dimension reduction and improving the efficiency of web documents processing. Now we are continuing our experiments about the algorithm and we will release more of our results of the experiments in times to come. In the future we will concentrate on applying the method to the fields like Web documents filtering, classification and clustering .

References

[1]. Maya Rupert，etc. The Web and Complex Adaptive Systems Proceedings of the 20th International Conference on Advanced Information Networking and Applications (AINA’06) , pp200 - 204 ,2006

[2]. Wang Shi, etc.. Web mining〔J〕. Computer Science, 27 (4) : 28～ 31, 2000

[3]. http://maya.cs.depaul.edu/~mobasher/webminer/survey/survey.html

[4]. SHI Zhongzhi Knowledge discovery , Tsinghua University Press, 2002

[5]. LIU Li etc. Term selection and weighting approach based on key words in text categorization Computer Engineering and Design 2006.6

[6]. ZOU Juan etc. An Eigenvalue Extraction Met hod for Chinese Texts Using Multiple Heuristic Rules COMPUTER ENGINEERING & SCIENCE 2006.8

[7]. Kennedy J., Eberhart R.C.: Particle swarm optimisation. Proceedings of the IEEE International Conference on Neural Networks. (1995) 1942-1948.

[8]. Ratnaweera, etc.:Self-Organizing Hierarchical Particle Swarm Optimizer With Time-Varying Acceleration Coefficients. IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION. (2004) 240-255

[9]. Shi Y.H., Eberhart R.C.: Parameter Selection in Particle Swarm Optimization. Evolutionary Programming VII: Proc.EP 98. (1998) 591-600

[10]. ZHANG Xiaoming, Wang Rujing. Particle Swarm Optimization Algorithm with Reverse Thinking Particles. Computer Science, 2006, 33 (10) :156～159

[11]. www.hylanda.com

Song Liangtu received the B.E. and M.E. degrees, from Anhui Agri. Univ. in 1987 and 1990, respectively. he is working toward the Ph.D, in Automation department,University of Science and Technology of China.
he has been an associate professor in the Institute of Intelligent Machines, Chinese Academy of Sciences,since 2000. His research interest includes data acquisition,pattern recognition and intelligent systems.

Zhang Xiaoming received the B.E. from Shandong University of technology in 2002. He received M.E. from Graduate University of Chinese Academy of Sciences in 2006. Now he is a Research Assistant of Institute of Intelligent Machines, Chinese Academy of Sciences. His research interest includes computational intelligence, web mining, machine learning.