The Hybrid of PSO and SOM for Blog Success Prediction

2C4-IOS-3c-7

Chi-I Hsu[*1] Shelly P.J. Wu[*2] Chaochang Chiu[*3]

[*1]Dept. of Information Management, Kainan University, Taiwan
[*2]Dept. of Information Management, National Sun Yat-Sen University, Taiwan
[*3]Dept. of Information Management, Yuan Ze University, Taiwan

Particle Swarm Optimization (PSO) is a population-based optimization method simulating the social behavior of populations. Self-Organizing Map (SOM) is a neural network technique with good clustering performance. This research proposes the hybrid PSO and SOM for predicting Blog success level. Basically, PSO algorithm is adopted to find the optimal weights of the input variables of SOM in order to reinforce the data in a single cluster (or certain clusters) with the same outcome level. A research model of Blog Success was proposed and 210 valid samples were collected from Internet users with Blog using or building experience. Compared with other prediction methods, the PSO-SOM approach of yielded better results than those of C5.0, Classification and Regression Trees (CART), Support Vector Machine (SVM) for 10-fold cross-validation.

1. Introduction

Blog applications have become more diversified with the rapid development of blogs in the era of the Internet. From plain narratives of feelings and experiences in the early days, to the publication of one's works and sharing of audios and videos today, the blog has made continuous progresses, and its free features have attracted a huge number of users. There are successful and popular blogs, such as that of Wanwan, a famous Taiwanese network writer. Total visits to her blog have exceeded one billion to date. Many creative blogs that have rich content and allow frequent communication in the network have received high ratings and created remarkable values. However, there are more blogs that are less interesting and have fewer visitors. Why is there a marked difference in the number of visitors attracted by different blogs? What factors influence a blog's success, and how can the success of a blog be predicted more accurately?

The objective of this research is to build a Blog Success model and create a PSO- and SOM-based, supervised machine learning method to predict blog success. The research design can be divided into two parts. The first part involves building a Blog Success model, investigating and measuring the influential factors of Blog Success, and collecting necessary data through the questionnaire investigation method. The second part discusses a PSO and SOM-based, supervised classification and prediction method. Particle Swarm Optimization (PSO) is a population-based optimization method simulating the social behavior of populations. Self-Organizing Map (SOM) is a neural network technique with good clustering performance. This research proposes the hybrid PSO and SOM for predicting Blog success level.

2. The Literature Review

A number of studies in the past have focused on the combination of optimized searching calculation and clustering technology, whereas others discussed the combination of GA and SOM, such as the clustering analysis of Gorza_lczany & Rudzi´nski (2004) and application of intrusion detection of Xiao, Shao, & Liu (2006). Al-Harbi & Rayward-Smith (2006) presented a research on supervised clustering approach developed from the combination of GA and k-means. To solve and address the oversensitivity problem of k-means to the original value, Laszloy & Mukherjee (2007) adjusted the original parameters using GA. Similarly, Li & Zhang (2010) used a combination of PSO and SOM for his network intrusion warning system; Hung & Huang (2010) used PSO to extract the classification rules from the clusters produced from SOM. Generally, in these studies, PSO and SOM were combined by intensifying PSO and extending the clustering effect of SOM, but not through changing the unsupervised learning clustering mechanism of SOM into a supervised classification mechanism.

3. The Research Model of Blog Success

Blog success can be discussed using three important value dimensions, namely, content value, technology value, and social value (Du & Wagner, 2006). Among these, content value is the factor that directly affects the success of blogs the most, because it is blog content that first attracts majority of users when browsing; a blog can have rich content and include text, pictures, or multi-media. The format of blogs is free and casual. Content can be presented as a simple diary, writings to express one's emotions, opinions on issues, and creative presentations. Blogs attract various readers through their rich content. A blog's technology value serves as the foundation for both content and social value, because the creative technology of a blog supports the development of its content and social value (Wagner & Bolloju, 2005). The technology value manages and presents the content value, and enriches the social value in blogs. Finally, the social value in blogs refers to the community resources and communication features; a blog's community support functions encourage users to regularly visit certain blogs and react with other users; its rich internal and external links help form a close virtual community.

Contact: Dr. C. Chiu, Dept. of Inf. Magt., Yuan Ze Univ., 135 Yuan-Tung Rd., Chungli, Taiwan, 886-34638800 ext 2601, 886-34352077, imchiu@saturn.yzu.edu.tw

The Blog Success model introduced in this research contains three dimensions, as shown in Figure 1. The content dimension refers to currentness and design; the technology dimension pertains to reliability and security; and the social dimension involves interaction and connectivity. These definitions are presented in Table 1.

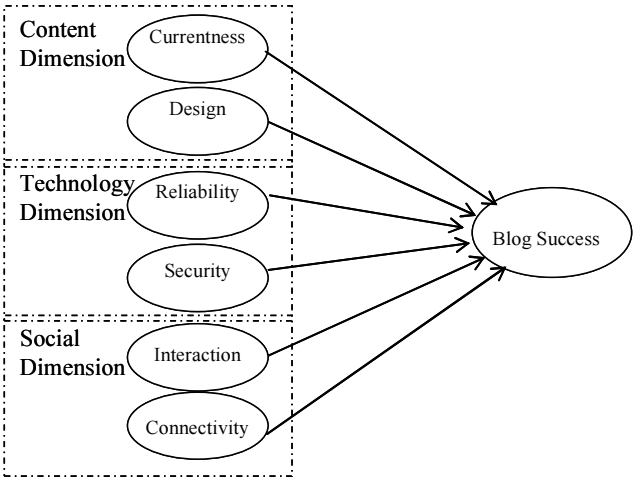

Figure 1: Blog success model

Table 1: Construct Definition

Construct	Definition
Blog Success	An index for evaluating the success of a blog, including number of visitors, popularity, loyalty of readers, and recommendation
Currentness	Refers to the rapid and frequent renewal of blog posts. A blog should provide complete and real-time information to the visitors.
Design	Pertains to the free and unique design features of the webpage in the blog. This dimension should present one's personality adequately.
Reliability	Connotes the browsing stability and speed provided by the blog, and the effective use of its files, pictures and links
Security	Refers to security problems that may arise in using the blog
Interaction	Denotes the degree of interaction between the webmaster and the readers, and the virtual community produced by the blog
Connectivity	Pertains to the super links included in the blog

4. The Hybrid of PSO and SOM

This study applies the PSO algorithm in converting the clustering mechanism of SOM into a supervised classification mechanism. The PSO algorithm was used to adjust the weight of SOM input variables and affect the SOM clustering result by changing the weighted vector represented by each particle. The evolution goal of the PSO adaptive value is to assemble the same outcome class into one cluster. We regard the class that accounts for the highest proportion in the cluster as the representative of that cluster. If the occupation ratio of the classification result of any cluster generated by SOM is higher than, or equal to, the preset threshold value, then this is termed an effective cluster. The PSO adjustment attribute weight is used to maximize the effective cluster amount, thus making each cluster a collection of one classification result data, and reaching the goal of optimized classification.

Figure 2 is the system frame for the predicated mode of supervised PSO-SOM proposed by this study. The system frame mainly consists of three parts: one is where PSO generates weight to influence the calculation of SOM clusters; another is the threshold value for measuring adaptive value, and this adaptive value for calculating SOM clustering results will be applied in classification evaluation; and the third is the internal processes in which SOM optimizes the concentration ratio of a classification result in each cluster, by continuously adjusting the weight and changing the speed and location of the particle.

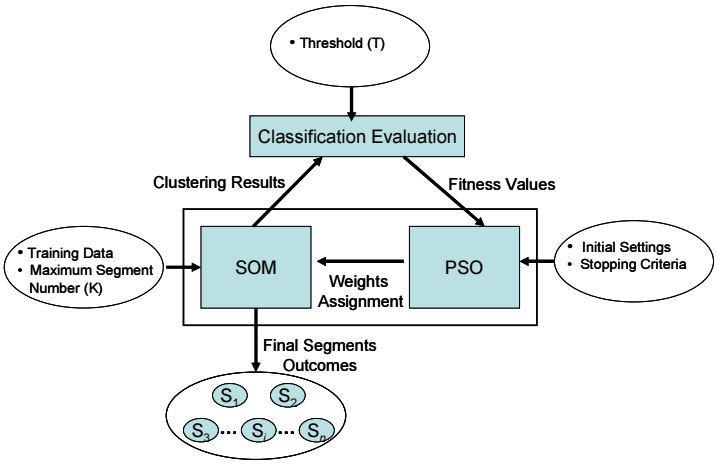

Figure 2: Supervised PSO-SOM framework

The detailed execution of the PSO-SOM steps is shown below: (1) Original values are set in measuring the threshold value T of the adaptive value and the biggest evolution algebra Gmax for the stop condition, such as the cluster scale size m of PSO, or the maximum speed limit Vmax for the particle (one can also preset the execution time or the acceptable adaptive value). (2) Execution of PSO-SOM begins. A cluster is initialized with a speed scale m in a random location. (3) PSO is used to generate a group of weights. (4) SOM calculations that include weight are made. (5) The adaptive value is calculated to measure the classification result. The adaptive value function f is shown in formula la. (6) If the stop condition is not reached, the speed and location of the particle is changed according to formulas la and lb. Next, the adaptive value of each particle is recalculated until the cycle achieves the stop condition (e.g., once a preset maximum algebra Gmax is reached). (7) Calculations of PSO-SOM are completed.

$$f = \sum_{i=1}^{n} T_j * C_i \qquad (1a)$$

$$If\ T_j < \mathrm{T}\ then\ T_j = 0\ else\ T_j = \max_{i=1}^{n} \left\{ \frac{o_j}{c_i} \right\}$$

In the above, n refers to the total cluster amount; i=1 ton; m refers to the total outcome class amount; j=1 tom; Tj refers to the result threshold value in j classification; Ci is the total material stroke count of the i cluster; T is the set threshold value; and Oj refers to the total data stroke count of the j classification result.

5. The Experiment

The questionnaire was employed in collecting the experimental data for this study. The questionnaire was designed based on the Blog Success model. This design conforms to the definition of each construct, and the questions were developed according to related references. Likert's Five-Point scale was used in the study. Thirty items were used for related input variables, as shown in Table 2, and six items for related output variable, shown in Table 3.

Table 2: Questionnaire Items for Input Variables

Input Variables		Questionnaire Items
Currentness (V1)	var1	The content of this blog is updated frequently.
	var2	There is always new information in this blog.
	var3	I can learn new knowledge from this blog.
	var4	I can discover the latest trends by browsing this blog.
	var5	This blog discusses burning issues.
Design (V2)	var6	The blog's interface is easy to use.
	var7	This blog has a vivid, versus rigid, design.
	var8	The color scheme of this blog is pleasing to the eyes.
	var9	I find that this blog has its own unique features.
	var10	I believe that the design of this blog is original.
Reliability (V3)	var11	Pictures in this blog appear normally.
	var12	The display or download speed of this blog page is steady.
	var13	Documents or content of this blog can be downloaded normally.
	var14	The webpage functions of this blog can operate normally.
	var15	I believe that this blog uses a stable server.
Security (V4)	var16	I am not worried about being infected by viruses while browsing this blog.
	var17	I think this blog addresses network safety issues.
	var18	I find the network safety of this blog to be reliable.
	var19	I do not need to worry about my personal information being revealed while browsing this blog.
	var20	In general, I find this blog to be safe to use.
Interaction (V5)	var21	This blog promotes friendly interaction.
	var22	The posts in this blog always solicit replies from new friends.
	var23	I will post my own opinions to this blog.
	var24	I think this blog provides good communication space.
	var25	I am satisfied with the interaction between the moderator and blog visitors.
Connectivity (V6)	var26	This blog is connected with other related blogs.
	var27	This blog is frequently invited to join the association lists of other blogs.
	var28	I always click this blog's associated links.
	var29	This blog has numerous associated links.
	var30	I find that most of the associated links of this blog have a reference value.

Table 3: Questionnaire Items for Output Variable

Outcome Variable	Questionnaire Items
Blog Seccess	I believe that many people will browse this blog.
	I think this blog has high popularity ratings.
	This blog is able to attract visitors for a second visit.
	I think this blog has loyal reader groups.
	I will recommend this blog to relatives and friends.
	I find the operation of this blog to be successful.

210 valid samples were collected from Internet users with Blog using or building experience. To compare the prediction results, the accuracy of the 10-fold prediction of the hybrid PSO–SOM method was cross validated and compared with that of other commonly used methods.

The parameter settings are as follows. In PSO-SOM, the particles size is 50 and cycle is 100. In C5.0, the expected noise (%) is 0. In CART, the maximum surrogates is 5, minimum change in impurity is 0.0001, and impurity measure for categorical targets is Gini. In SVM, the stopping criteria is 1.0E-3, regularization parameter is 10, regression precision (epsilon) is 0.1, kernel type is RBF, RBF gamma is 0.1, bias is 0, Gamma is 1, and degree is 3.

The experiment results are shown in Figure 3. Values of the six input variables are considered the average of the related items in Table 2. Compared with other prediction methods, the PSO-SOM approach of yielded better results than those of C5.0、CART and SVM in the testing results with 10-fold average basis.

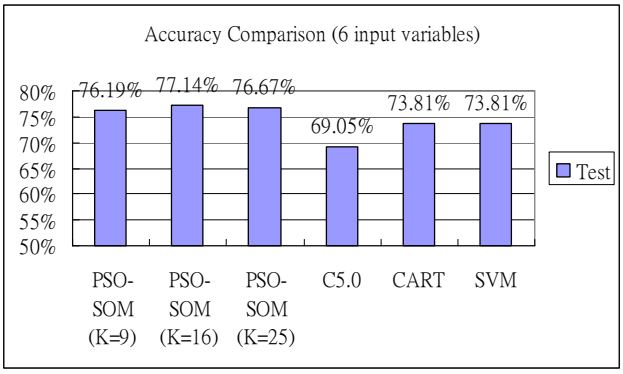

Figure 3: The Experiment Results

6. The Conclusion

In our blog success model, this research proposed six influential factors of blog success, including currentness, design, reliability, security, interaction and connectivity. In assessing blog content, currentness refers to the frequency in updating and providing new information and knowledge, the latest trends and burning issues; design, on the other hand, considers the overall design of the blog and whether it is original, pleasing to the eyes, and easy to use. In assessing technology, a blog's reliability is measured by the stability of its browsing speed, webpage display, and connection. In evaluating security, attention is paid to webpage safety issues such as the presence of viruses and loss of data privacy. In measuring the social value of a blog, interaction refers to whether there is good communication space and whether the net users are willing to publish their opinions and whether or not they receive replies; connectivity, on the other hand, pertains to a blog's hyperlinks and the associated reference value that the blog offers.

This study proposes a supervised machine learning method combining PSO and SOM for predicting Blog Success. Although k-means is a common clustering technology, this study found that this technology is unsuitable for clustering when processing non-continuous input variable values, and the K value can only

be decided after continuous trying. In comparison, SOM is more flexible as it not only processes continuous and non-continuous data separately, but also processes mixed type data. Although there are several other technologies that can perform optimized search such as SA and GA, PSO contains relatively better searching ability for continuous weight values. Although PSO and SOM have been widely applied individually in many research studies, few of these studies have investigated combining them for predictive purposes. This research has proposed a supervised classification mechanism to predict Blog success. The prediction performance of this approach has also been compared with the predictions of C5.0, CART and SVM. In the future, various clustering and optimization methods may be combined and compared their prediction performance.

References

Al-Harbi, S. H., Rayward-Smith,V. J. 2006. "Adapting K-means for Supervised Clustering." Applied Intelligence 24: 219-226.

Du, H. S., and Wagner, C. 2006. "Weblog Success: Exploring the Role of Technology." International Journal of Human-Computer Studies 64: 789-798.

Gorza_lczany, M., Rudzi´nski, F. 2004. "Application of Genetic Algorithms and Kohonen Networks to Cluster Analysis." Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science) 3070: 556-561.

Hung, C. and Huang, L. 2010. "Extracting Rules from Optimal Clusters of Self-Organizing Maps." Second International Conference on Computer Modeling and Simulation 382-386.

Laszlo, M. and Mukherjee, S. 2007. "A Genetic Algorithm that Exchanges Neighboring Centers for k-means Clustering," Pattern Recognition Letters 28(16): 2359-2366.

Li, L. and Zhang, C. 2010. "Alert Clustering Using Integrated SOM/PSO. " 2010 International Conference On Computer Design And Applications (ICCDA 2010) 2: 571-574.

Wagner, C., and Bolloju, N. 2005. "Supporting Knowledge Management in Organizations With Conversational Technologies: Discussion Forums, Weblogs, and Wikis." Journal of Database Management 16 (2): i–viii.

Xiao,L., Shao,Z., Liu,G. 2006. "K-means Algorithm Based on Particle Swarm Optimization Algorithm for Anomaly Intrusion Detection." Proceedings of the World Congress on Intelligent Control and Automation (WCICA) 2: 5854-5858.