논문 2010-3-12

유전자 알고리즘을 이용한 웹 검색 랭킹방법

Ranking Methods of Web Search using Genetic Algorithm

정용규*, 한송이**

Yong-Gyu Jung, Song-Yi Han

요 약 검색엔진을 사용하는 이용자의 정보 즉 선호도에 따른 지속적인 피드백으로 검색 결과의 랭킹을 향상시켜 유연한 검색이 가능하게 하는 방법에는 학습된 인공 신경망을 이용한다. 인공 신경망 학습은 신경망이 여러 다른 검색어로 학습된 후 다른 사용자들이 과거에 실제 검색했던 결과를 좀 더 반영하기 위한 것이다. 가중치의 지속적인 변경을 위해서는 네트워크에서 역방향으로 움직이면서 가중치를 변경하는 역전파 알고리즘을 이용하여 학습한다. 그러나 이러한 학습은 초기에는 훈련데이터에 적합한 성능을 보이나 학습의 횟수가 증가할수록 점점 과대적합되는 것을 알 수 있다. 따라서 본 논문에서는 최적화해야 할 개체가 많을 때 강한 장점을 가지고 있는 유전자 알고리즘을 적용하여 검색어에 관련성이 높은 페이지들 유연하게 랭킹하기 위해 URL리스트를 개체로 랜덤으로 선택하여 학습하는 기법을 제안한다.

Abstract Using artificial neural network to use a search preference based on the user's information, the ranking of search results that will enable flexible searches can be improved . After trained in several different queries by other users in the past, the actual search results in order to better reflect the use of artificial neural networks to neural network learning. In order to change the weights constantly moving backward in the network to change weights of backpropagation algorithm. In this study, however, the initial training, performance data, look for increasing the number of lessons that can be overfitted. In this paper, we have optimized a lot of objects that have a strong advantage to apply genetic algorithms to the relevant page of the search rankings flexible as an object to the URL list on a random selection method is proposed for the study.

Key Words : Artificial Neural Network, Search Ranking, Backpropagation, Genetic Algorithms

I. Introduction

The rapid increase in Web information for millions of pages of the web site to determine the most appropriate is an important issue has been decided that the way in terms of results the search ranking has become a key issue. Extract pages that match the returned page, simply crawl is ranked in the order that was. If you found this page contains the word, but unrelated to the choice problem of many pages are facing. Therefore, in order to solve this problem for a given problem by applying the score accordingly higher scores corresponding to the first page as a way to return the user can narrow the choices. Traditionally, the highest score by calculating a score with a priority ranking to the results based on the contents of the article that way. As a way to search the frequency of words appearing on the page, document the location of my key words, words of the proximity of the measured distance to determine the degree of fit of the article. However,

*종신회원, 을지대학교 의료IT마케팅학과
**정회원, 을지대학교 의료산업학부 의료전산학전공
접수일자 2010.06.05 수정일자 2010.06.14

these rankings are relatively simple to operate using many search engines, but the rapidly changing web environment, how to calculate the fixed points of the information that applies to the page smoothly with the problem of feedback when you can.

Bbackpropagation learning of neural network trained with several other queries by other users and the actual search results in the past is to reflect a little more. But early in the training data performance, look for increasing the number of learning for overfitting can see that. In this paper, there are many objects you need to optimize the benefits that have a strong genetic algorithms applied to the relevant pages in the search rankings in order to flexibly URL (Uniform Resource Locator) list of objects selected at random Learning the technique is proposed.

II. Related research

2.1 Neural networks for Searching and Tracking

Neural network node consists of the connection. Relevant to a given word is a node that is to track the web neuron is composed of several layers MLP (Multi Layer Perceptron) is using the network.

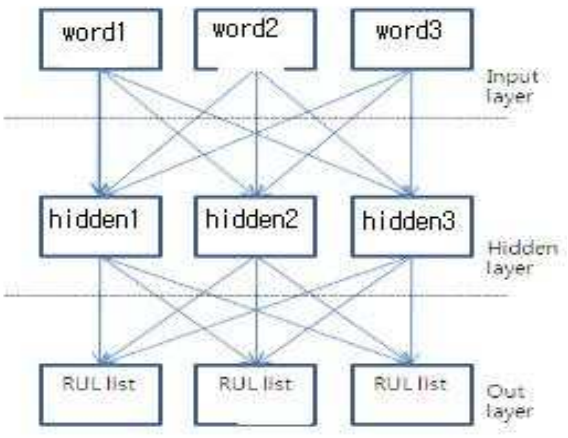

그림 1. 추적 신경망 설계
Fig. 1. Neural network design for tracking

Input layer provided by the search word is a word, a number of hidden layers are independent of the hierarchy may exist (hidden layer) of the query that takes place in a hierarchy of direct interaction with the outside world enter the union only to respond by cutting be. Finally, the output layer by applying a weighting in query words with the URL to determine the relationship between the output list.

Construction Site tracking neural network for continuous learning and back propagation algorithm is applied.

2.2 Backpropagation

Forward multilayer neural network training in the normal way inclined gradient descent algorithm is using the back propagation algorithm. Learning of the squares of the errors that appear to minimize the cost function. Squared error cost function as follows.

$$E=\frac{1}{2}\sum_{i=1}^{p}\| y_i-d_i \|^2 \quad (1)$$

The expression of the actual neural network output is y_i, and the target value d_i. Of the backpropagation algorithm and then pseudocode is.

표 1. 역전파 알고리즘
Table 1. Backpropagation algorithm

```
v,wm<--radomvalue
p<--number of traing pattern pairs
k<--1
E<--0
set learning rate and Emax

repeat: :
 ForwardCalculate()
 ErrorCalculate()
 updateWeight()
until k=P

if E<Emax
 stop
else
 E<--0 ,k<--1
 goto repeate
```

Learning how to adjust the error to sell the station to learn about the sensitivity training data reflects the excellent performance under the optimal, but local minimum for excessive fall in that concern, and it does not work in the saturated zone problems can.

III. Genetic Algorithm Learning

As a fixed percentage of the weight saved by the manner of deviation Backpropagation was applied instead of the genetic algorithm. Genetic algorithms to populations of the initial mutation and crossover, while the process of evolution through several generations and the best selection of the object is. Applying this value of current generated by the random initial populations evolve. Each population size of 50, the maximum evolution of Households 100 Elite, select the ratio of 0.2, mutation occurs the probability 0.2 has, for each object is 0-8 between the number of elements go into a length of an array object used as should. Received via the existing value of the weights can be calculated by measuring the minimum value set in the direction chosen after subtracting the value of the final object is used as a percentage. When you apply this approach in the search engines are subtracting the proportion of genetic evolution is not fixed because the variable is determined by the object. Components of an object and the medium weight of the product were obtained, then the values calculated using the average and the only positive to the selected URL, and the rest is treated as a negative, the weight is less for.

IV. Experimental methods and results

The experiments in this paper is available from the website of Toby Segaran Parent Directory - search index data set were used, the language Python were used.

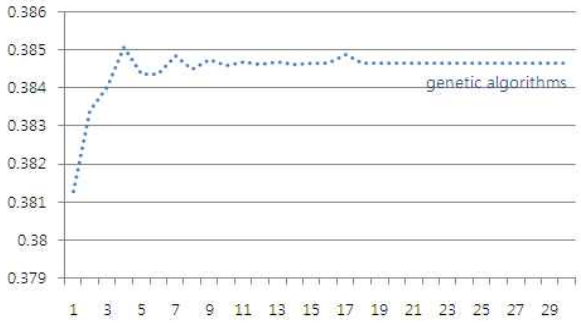

그림 2 카테고리별 리스트 실험 (Genetic)
Fig. 2 Categorical list experiments (Genetic)

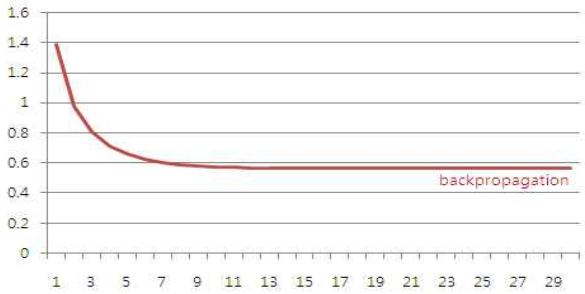

그림 3 카테고리별 리스트 실험 (Backpropagation)
Fig. 3 Categorical list experiments (Backpropagation)

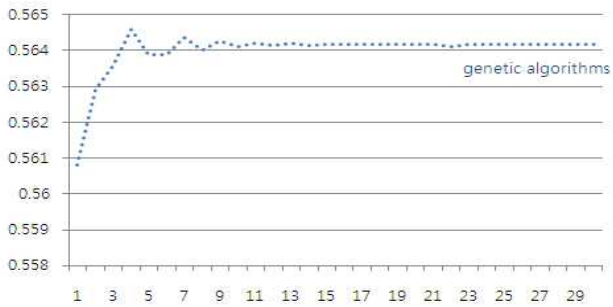

그림 4 멀티 패러다임 실험 (Genetic)
Fig. 4 Multi paradigm experiments (Genetic)

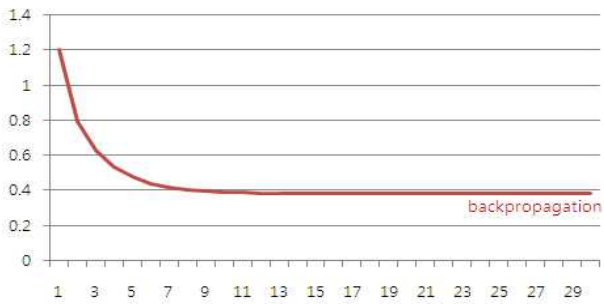

그림 5 멀티 패러다임 실험 (Backpropagation)
Fig. 5 Multi paradigm experiments (Backpropagation)

표 2. 학습방법에 따른 실험결과
Table 2. Experiments for each methods

Algorithm / Iteration / URL	content-based	Back Propagation			Genetic Algorithm		
	1	1	10	30	1	10	30
Functional_programming.html	3.12114	8.121124	8.121124	8.121124	8.121124	8.121124	8.121124
Programming_language.html	2.074506	2.895387	2.082266	2.074511	2.071146	2.074437	2.074511
Object-oriented-programming.html	0.712191	1.533073	0.719952	0.712197	0.708832	0.712123	0.712197
Programming_paradigm.html	0.622044	1.442925	0.629804	0.622049	0.618685	0.621975	0.622049
Categorical_list_of_programming.html	0.564171	1.385052	0.571932	0.564176	0.560812	0.564102	0.564176
Procedural_programminng.html	0.469566	1.290447	0.477326	0.469571	0.466207	0.469497	0.469571
Lisp_programmuing_language.html	0.46369	1.284572	0.471451	0.463696	0.460331	0.463622	0.463696

In table 2, The content-based, backpropagation learning, genetic algorithms were applied relevance of Web pages for the query table that shows the weight value will be. Content-based learning does not proceed, Backpropagation is the minimum area tend to focus on the value. However, using genetic algorithms for learning the initial population random values, except to accept a position was found to be variable. This definitely shows more volatility through the graph.

V. Conclusion

Results of using genetic algorithms, ranking methods compared to the propagation of a flexible weighting function for a couple of positions in many objects can be shown fluctuations. In addition to the selection of the initial population and calculated weights to improve the function, click rates, unlike the backpropagation method, as well as to consider the impact of the subjective and can be found that can be studied. Simply click on reversing the current ratio of wave propagation depending on the choice so that the user inadvertently or similar query can be connected include the search results do not reflect the impact. In contrast, the initial population of genetic algorithm selection and the cost function depending on the configuration and utilization reflect a variety of factors can make it, so future research is needed.

References

[1] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. “Indexing By Latent Semantic Analysis”, Journal of the American Society For Information Science, 41, 391-407. (1990)

[2] D. Mertz, “Spam Filtering Techniques. Six approaches to eliminating unwanted e-mail.”, Gnosis Software Inc., September, 2002. Ciencias Físicas, Universidad de Valencia, 1992.

[3] Ian H. Witten, Frank Eibe, "Data Mining: Practical Machine Learning Tools and Techniques" Morgan Kaufmann, 2000

[4] Jiawei Han, Micheline Kamber, “Data mining - Concepts and Techniques”, Morgan Kaufmann Publishers, 2001.

[5] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A Bayesian Approach to Filtering Junk E-Mail," AAAI Technical Report WS-98-05, 1998

[6] M. Vinther, “Junk Detection using neural networks”, MeeSoft Technical Report, June 2002. Available: http://logicnet.dk/reports/JunkDetection/JunkDetection.htm.

[7] Nils J. Nilsson, 1998. Artificial Intelligence. Morgan Kaufmann, Inc.

[8] Pang-Ning Tan & Michael Steinbach & Vipin Kumar, "Introduction to Data Mining", ELSEVIER, 2006

[9] Toby Segaran. Parent Directory - searchindex [cited 2009.3.28] <http://kiwitobes.com/db>

[10] Toby Segaran, "Programming collective intelligence", O'REILLY, 2007

저자 소개

정 용 규(종신회원)

- 1981년 서울대학교 (이학사)
- 1994년 연세대학교 (공학석사)
- 2003년 경기대학교 (이학박사)
- 1999년~현재 을지대학교 교수
- 2001년~현재 ISO/TC154K위원장

<주관심분야: 임상데이터마이닝, 의료정보시스템, 전자거래표준>

한 송 이(정회원)

- 2007년~현재 을지대학교 의료산업학부 의료전산학전공

<주관심분야: 데이터마이닝, 비즈니스 인텔리젼스>