Web page rank estimation in search engine based on SEO parameters using machine learning techniques

Hengameh Banaei[1]† and Ali Reza Honarvar[2]††

[1] Department of Computer Engineering and Information Technology, Safashahr Branch, Islamic Azad University, Iran

Summary

E-Commerce is growing day by day so that most of the requirements of customers can easily be provided through websites. In such conditions, the Web markets to stay on competition needs one of two methods of advertising and increasing site rank in search engine results.This research wants to analysis website’s internal parameters that affect determining site rank in Google's search engine. Forasmuch as how to choose and number of HTML tags for a particular website has direct impact on page rank, important concepts in this area should be examined. These concepts includs Number, Proximity and Density. For extract required data for this research we design a crawler and parser. Then dataset has been learning by ANN. Then, this ANN is used to predict new websites rank that has good performance on test data.

Key words:

Search Engine Optimization, E-Commerce, Site Rank, Artificial Neural Network

1. Introduction

There are different methods to have viewer on a site, or in other words, marketing for a site: the first method is the advertising on the outside of the web. This type of advertising even if attracts the attention of viewer, the possibility that he searches the considered site is low because the environmental interval in which he accesses the internet is much and the effectiveness of the ad is removed. The second method is the advertising on the web, for example, embedding an advertising flash file on a popular site. The second method is better than the first method in most cases, but also it has weaknesses such as high cost and that it cannot be certain that users of the full viewer site are interested in the ads or not. There is also another new method that is in contrast with traditional methods completely. In this way, in fact, instead of presenting site to viewer, it will be responded to its need and question and rather than following the viewer, the viewer will find it and go toward them.

When the user refers to Internet for search something, actually search engines act to find information in millions of different sites. Because the results are so high, the user never examines all the results and may after investigating results in the first page, he refers to the second page but he never refers to the results of the tenth page. So the position of placing site in result pages is very important. That means that the more a site to be placed in the initial pages of results, the possibility of being seen and using the site will be increased that it causes the more recognition of site. SEO or search engine optimization based is actually response to the main need of web sites that targeted traffic or high number of visitors from the site.

Choosing a proper keyword for Websites' contents is one of the main concerns of website's authors because these words are one of the most important parameters in increasing rank of site.

Search engines have specified certain factors as search criteria and based on these factors and keyword searched do the search operation.

These factors are divided into five groups of factors related to the key words, factors related to the link, factors related to site, factors related to the page and factors related to laws of specific algorithms that the factors related to keywords are the most important. So many website optimization techniques have focused on the extraction of appropriate keywords from a website to enhance the site's rank in search engines.

Therefore authors should follow the principles that would enhance the credibility of the text and the Web sites when inserting new content on their website. Based on this, estimate rank of site will be valuable for writers of websites according to the selected keywords from text. This is despite the fact that many factors that search engines are considered when searching for a word or phrase, cannot be managed by website's authors, such as the number of previous visits, the credits the type of range. The basic problem is how can choose the key words based on the content that caused to increase the site's rank. Since the website rank is affected in the priority of displaying search engines, the behavior of a search engine considering keywords can be analyzed and learned. Machine learning techniques are used for this purpose. After learning the behavior of search engine based on keywords, we can choose the best possible keywords according to considered text

Manuscript received May 5, 2017
Manuscript revised May 20, 2017

2. Related works

SEO science was begun by webmasters of major sites in the mid-1990s. The first one who has written concepts about the search engine optimization is John Audette and he recorded his company as Multimedia Marketing Group in one of the pages of MMG site in August 1997. First, by Kan and his colleagues in an article entitled "fast classification of Web site using the features of Web Site", referred to the problem of predicting rank of websites in 2005 [1]. In this way, they took advantage of linear regression that its complexity was increased by increasing the number of features extracted exponentially. Also the results of this study were much lower than expected, but it opened new ways for the researchers to predict rank of website. After this unsuccessful experience, Vazirgiannis tried to improve the prediction of web rank using generating Markov models on the ranked lists of sites [2]. Broder improved the method of estimate rank without parsing the total graph using the method of producing web graph that was introduced by Chien and colleagues [3] in 2003 [4]. Another study that was done in this way, they used web structure to estimate rank that provided better results both in terms of computing time and yield [5] In 2009, a study was conducted entitled "estimate rank of website using the method of reducing PCA dimension and clustering method of EM" where they had collected their features used from lists ranked of Wikipedia websites [6] In 2012, Voudigari and his colleagues used previous results of search engines to improve the method presented in [6] and they could improve well the performance of the previous method [7]. A study that Chen and his colleagues had done in 2005 [8] was improved in 2014 [9]. Instead of parsing the entire graph that was obtained from parsing and obtaining the required information from previous websites that caused slow speed, they only parsed a small part of the graph. This method that was first introduced by Chen improved by other researchers several times. Two examples of these tasks that had higher speed than other methods were conducted by Lafgern [10] and Mytlyakas [11]. Cheng Zhi Lu estimated web site rank in the other search engines using the Google search methods [12]. For this purpose, he used the factors of Google and extracted this information including title, links contained on the website, etc. by crawling method and on the basis of these factors, he assigned a weight to each web site using genetic algorithm. Then, he estimates the rank of considered website based on the weights and statistical models.

3. Proposed method

3.1 Factors based on keywords

Many of the factors affecting the rank of the website are out of the site's author access (not the designer of site) and it is required to focus only on other components to increase the rank of site. One of the most important components that can be derived from the previous mentioned issues is key words. So factors should be determined based on keywords and examined them on websites. Three main factors of number, density and proximity of keywords have been expressed in this field. [13]

Number of keywords: Of course, the higher number of key words with regard to all its circumstances will increase rank. For example, using keywords that is not considered by Internet users or keywords that are not compatible with the subject of web site or text will have no effect.

Density of keyword: It is a measuring criterion that shows the percentage or actual number that a keyword is appeared compared to the total number of words of text. The more the total number of words used in the text is greater; the number of keywords can be increased.

Prominence of keywords: It is a measuring criterion which measures the position of the words in HTML tags. Keywords that used earlier in tags are more prominent. A definition that we will use to prominence is as follows: [13]

$$Prominence = \left(Words\ in\ string - \frac{Sum\ of\ positions\ of\ keywors - 1}{repetitions\ of\ keywords\ in\ string} \right) * \frac{100}{Words\ in\ string} \quad (1)$$

In this definition, the key words that are at the beginning of the tag sentence are more emphasized Now that factors based on keywords and affecting rank of Web site were determined, HTML tags should be determined that they should be evaluated.

3.2 Tags influencing rank in Google

Necessarily, all tags of the web site in terms of search engines, especially Google, are not considered. According to what stated by Google and many experts use to optimize the websites; it can be considered multiple tags more important than others. These tags include Meta tags, title, body, paragraph, picture, ALT, div, head of the page and link. So, only expressed tags should be investigated in any website analyzed and values of factors based on keywords to be extracted in them.

3.2 The keywords in several areas

The last step before the stage of crawling in web is to determine the appropriate and varied keywords in several specific areas. These words can be determined without any rules and restrictions but previous studies have shown that the choice from a few specific areas led to centralized data collection that has significant impact on the learning process.

3.4 Crawling in web

In the process, dataset is provided to process and extract the required information using the keywords extracted in the previous step that consists of two steps.

Google Search: In the step, any of the specified keywords from the previous step is search in Google and the results returned by the Google search engine are stored in the next steps for processing. To do this, the first 5 pages of Google results should be maintained.

Extract Html codes and determine rank of each website:

After returning the search results of Google for any keyword, HTML codes of every result are maintained to provide data for parse. So for the crawling in the first 5 page of Google results (and given that in per page, Google results of 10 websites is returned as a result), 50 HTML pages are stored. It should be considered this number of result is only for a keyword. Determining the rank of each website will be derived from number of Google results page so that rank one has been allocated to all of the web sites that are placed on the first page of Google search, rank two has been allocated to all Web sites that placed on the second page of Google search and so on ... So ranks are from one to five, and 10 HTML code will be existed from every rank for each keyword.

3.5 Parser

The final and basic step to gather dataset collection is the design of parser that actually extracts all the information needed to process. The parser parses all HTML codes stored (50 pages of code per keyword) and it extracts values of three factors of number, density and proximity of the considered keyword through 15 introduced tags in the previous chapter. For example, in the tag of title, the numerical value of the number of keyword x is equal one, its density is 0.25, proximity is 100 and in the tag of descripting meta, the number of keyword is equal one, its density is 0.043 and proximity is 100. All values extracted for all keywords are placed in a text file to be entered the neural network for training. For example, if we have 200 keywords, 10,000 records are stored in a text file as a dataset that each record has 45 values. (15 tags, that from each, three values of number, density and proximity are extracted). Any of the records is called data or sample or pattern and each of the 45 values within each data is called property.

3.6 Multilayer Perceptron Neural Network

Perceptron is a non-recursive network that utilizes a supervised learning algorithm. Therefore, its training categories include a set of input vectors with their preferred target vectors. In this network, input vectors include continuous limits of values but target vectors include binary numbers of one and zero that are produced after training. [15]

Layers, respectively, are connected in the multi-layered networks, in such a way that the outputs of the first layer, the second layer inputs and so on that the last layer outputs form the main outputs and real answer of network. In other words, the flow of network signal is carried out in an appropriate direction that starts from the input layer and leads to output layer.

Two types of signals are used that are different with each other. Category of (MLP), generally in multilayer networks of first perceptron of function signals that according to the inputs of each neuron and weight parameters and motion function are calculated and the second category, error signals that are calculated by the reversal of the output layer and branching to other hidden layers.

The number of hidden layer neurons depends on the view of network design and it is obtained by attempt and error. In case of insufficient number of neurons, the network will not be able to create accurate mapping between input and output vectors. There is a linear function and learning process in all neurons and layers MLP takes place in the output of each neuron. All weights and biases that are on the network can be changed during the learning process. [15]

4. Experiments and results

As mentioned, search engines determine 5 categories of factor and as consider as the rank of site that many of which are not managed by the author of website. It can be identified those factors that can be manipulated and improved by the author of website and have a significant impact on rank of websites among all the factors used by search engine of Google. The most important of these factors are related to keywords. So in this study, in the first step, a limited number of parameters based on keywords affecting rank of site in search engine of Google was determined, such as the number of repetition of keywords in text, the number of keyword in the title of text.

The concepts that is loaded on any Web site can be included a large number of key words according to its scope that search any of these words in search engine will return different ranks for websites. So in the second step, a number of words related to the scope of website concepts are searched in search engine of Google. It can be extracted training data for machine learning from these results according to searched keywords and parameters set in the first step.

In the third step, each of the results is parsed in order of preference of display in Google searcher page using crawling method and the amount of each of the parameters that obtained in the first step is extracted in returned website. The values of these parameters and rank of the website page form the training data. The data is given to machine learning algorithm and the learning will be taken place.

So far, by the processes that were conducted, the data that have been extracted from the web sites and rank of websites is available. Any of these data should be mapped to its considered rank using a proper way. A category of machine learning techniques called "categories" in these cases have shown an extraordinary performance. Categories learn how to map the input data to the desired result by taking training data (information extracted from web sites) and result (rank of each site). After learning, any data that is given to them can determine the considered output with a mapping.

Categories that apply the learning process on the data of this study are MLP neural network that their results will be compared with previous methods.

Because many of the factors that Google uses to rank sites and search are not manageable by author (improving) or have not been revealed yet, the result of numerical rank assigned to the website will not have the required accuracy. So we divide ranks into 5 categories that category one has the highest rank and the category 5 has the lowest rank.

There are many machine learning techniques that are improving every day and some have shown an extraordinary performance for specific applications. A method that we have used for this study is the artificial neural networks that their parameters should be set for this problem.

4.1 Keywords

Keywords were selected in three fields of medical, students and sport. It is tried to examine the most common words in selecting these words.Totally are 148 keywords. For example in table 1:

Table 1: Keywords

Dataset Name	**Model**	***Total***
Student	[Lesson], [School], [Graduation]	50
Sport	[Player], [Federation], [coach]	50
Medical	[Health], [Doctor], [Hospital]	48

4.1 Dataset

Each keyword is searched separately in the search engine of Google and is parsed by crawling and parser. The final dataset has 7400 data and 45 characteristics. For example in table 2, the results from the evaluation of the first result of first page of Google for the search of keyword "knowledge" are as follows (the first result of first page of search is for the word of site of Wikipedia):

Table 2: The results from the evaluation of the search keyword "knowledge"

Number	0.01,0,0.01,0.002,0,0,0,0,0,0,0.001,0,0.001,0.003,0
Density	0.048,0,0.5,0.011,0,0,0,0,0,0,0.006,0.008,0.009,0.014,0
Prominence	0.95,0,1,0.143,0,0,0,0,0,0,0.037,0.021,0.025,0.204,0
Rank of website	1

Of 46 number, 15 first number represents the number of (as normal) keyword of knowledge in each of the 15 tags considered, 15 second number represents the keyword density of knowledge in each of the 15 tags considered and 15 third number represents the prominence of keyword of knowledge in each of the 15 tags considered. The final number represents the number of Google result page for the Web site that is in fact the rank of website.

Of course, all the data have been normalized that they are divided to the values related to number of keyword on 1000 and values related to proximity on 100. The value related to density, which is a number between 0 and 100 does not need to be normalized.

4.2 Design Neural Network

As simple neural networks do not have the ability to categorize the data, so the MLP neural network was used. In this network, learning rate is equal to 0.25 that as needed, neural network with rate of 0.1 reduces and with rate of 1.2 increases during the learning. Its minimum value is selected 0.0001. The network has two hidden layers with 14 and 10 neurons. Its activated function is Sigmoid and data have been applied as binary to network.

The number of iterations of every time of the running neural network is 6000 times and it has been done 100 times. Training and test data have been cluttered up every time to be avoided of data of a class successively. Also testing data in every one hundred times running randomly are selected among all data extracted from previous steps but the number of it has been fixed.

The number of training data is 6000 and the number of test data is 1400. Least squares error method was used to measure the performance of the neural network in training and testing stages. Initial value of delta parameter is 0.5 and its reduced rate is 0.1. The minimum delta value is considered equal to 0.000005.

Trial and error method is used to determine the adjustable parameters of network as learning, the rate of increase and decrease learning, the Delta value, the amount of increase and decrease delta, the number of iterations of neural network in each time running, the number of training data and test and speed of reducing and increasing delta and learning rate.

It is considered the rate of learning network is initially intended high to steps of learning to be done faster. If learning error after T consecutive time decreases, the learning rate decreases to be efficiently optimized and vice versa if the learning error increases T consecutive time, the rate of learning will be added. In Figure 1, the results of one of the best functions of 100 running of neural network with mentioned parameters are shown above.

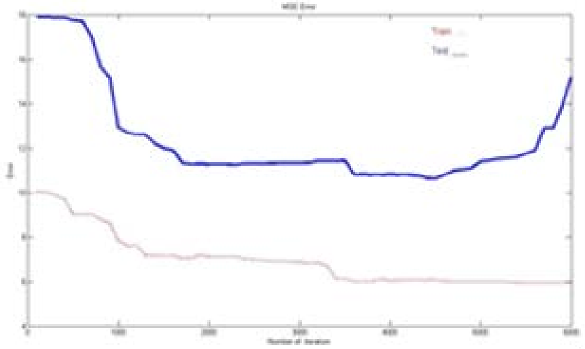

Figure 1: Shows the results of neural network training and testing performance in 6000 replications

As seen in the figure, error of learning process is reduced to 1200 iterations with the same slope and in contrast, the test error is reduced from 600 iterations with steep and almost constant to 1700 iterations. Traing error has not had high volatility from 1200 to 3300 iterations and relatively has been fix. Test error also has behaved similarly.

The error value of training is reduced in 3300 iterations and then has remained quite stable. Error of test is declined from 3400 iteration and has had the best performance with an average error of 10.8 (Note 89.2) but as can be seen, test error is increasing after 4500 iterations, that indicates the fact that neural network is "remembering" training data instead of learning. So the learning process not only is unnecessary after 4500 iterations but has negative impacts on network performance. The best number of iteration of learning is 4000 iterations.

As mentioned, 100 times neural network were trained and tested with the same parameters listed, which the only difference has been in the arrangement of training and testing data. The result shown in Figure 1 is related to the most appropriate distribution of data, where the data of each of the five classes (5 ranks of web site) was placed within the data of testing and training with the same number.

In the following, a few other performances with the distribution of data in each class are shown:

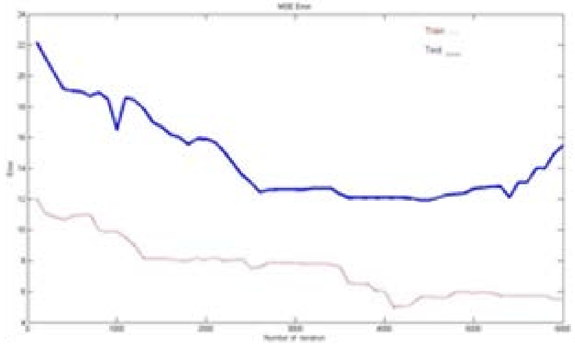

Figure 2: Shows the results of neural network training and testing performance in 6000 replications

As you can see, the behavior of this network is almost the same as the previous network, the difference is that the slope of decline error was slower and a bit is added to its volatility. Of course, this point should not be simply ignored that despite the decrease in the mean error of learning, the test error rate is higher than the previous state that was predictable according to the distribution of data in each class.

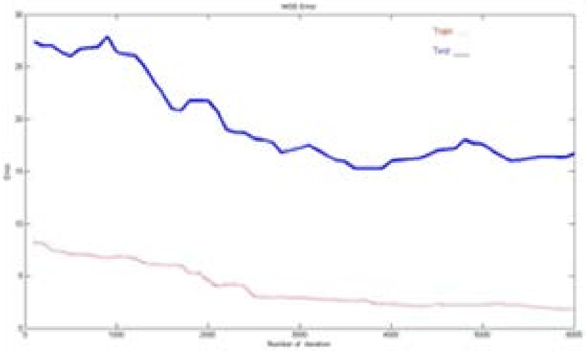

Figure 3: The results of the training and testing the neural network repeated in 6000 with very uneven data

As can be seen, test error of the network is highly volatile. Average test error of the network is much higher than the two other samples (15.27) while the mean of learning error of the network is lower than others. This behavior is due to imbalanced use of data of different classes that caused data of a few classes to be learned well, but in the face of other classes data has practiced badly.

The results of previous studies have been shown in Figure 4. The best performance among the results is the research of Zachurly and colleagues that is the use of regression classifier. In this study, EM classification method is used that has had acceptable results.

Testing error of our proposed method and regression method is the same. But certainly, set the parameters of the neural network is much more difficult than regression method that is considered the disadvantage of this proposed method.

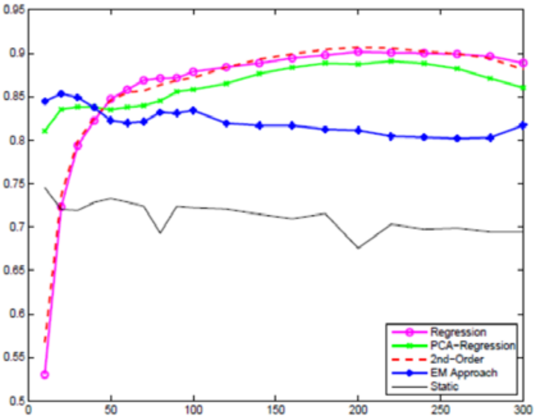

Figure 4: Shows the results of previous research on data sets extracted from Google

Conclusion

In this research, first, it was investigated the most important factors managed affecting rank of site which can increase the rank and validity of website in search engines, especially Google. These factors can manage all keywords. Then criteria based on key words were determined and values of each of them for each of the results returned were calculated in results of Google search engine. Values obtained were taught using MLP Neural Network.

In the last step, the results were evaluated. The evaluation of this neural network is obtained by dividing dataset obtained from previous stage in two parts of training and testing data. Training data are used for learning neural network and test data are given to the network after completing the learning process and output of network is obtained per input. Obtained output is compared with the class of each data, if they are equal, the new data is properly classified (its ranked is correctly predicted) and otherwise the output is faced with error. The accuracy of neural network designed for the system (Predicting rank of site based on keywords) is 89.36% at the best mode that is a very good performance.

Also it can be focused on a site in the work ahead instead of investigating results on different sites, in such a way that for a valid site, it can be searched different key words in Google and extracted values of key word parameters searched then, for that keyword to be applied dataset to the neural network based on the website rank in Google.

References

[1] Kan, M.-Y., Thi, H.O.N.”Fast webpage classification using URL features”. In: Proc. CIKM, 2005. Bremen, Germany

[2] Vazirgiannis, M., Drosos, D., Senellart, P., Vlachou, A.”Web Page Rank Prediction with Markov Models”. WWW poster, 2008. Beijing, China.

[3] Chien, S., Dwork, C., Kumar, R., Simon, D.R., Sivakumar, D.”Link evolution: Analysis and algorithms”. Internet Mathematics, 2003.1(3), 277–304.

[4] Broder, A.Z., Lempel, R., Maghoul, F., Pedersen, J.”Efficient PageRank approximation via graph aggregation”. Information Retrieval, 2006.9(2), 123–138.

[5] Yang, H., King, I., Lyu, M.R.”Predictive ranking: a novel page ranking approach by estimating the Web structure”. In: Proc,2005.

[6] Zacharouli, P., Titsias, M., Vazirgiannis, M.”Web page rank prediction with PCA and EM clustering”. Proc. of the 6th Intern. Workshop on Algorithms and Models for the Web-Graph: Springer-Verlag Berlin,2009.104-115.

[7] Voudigari,E .Pavlopoulos, J.Vazirgiannis,M.”A Framework for Web Page Rank Prediction”.IFIP Advances in Information and Communication technology, 2012.p 240-249.

[8] Chen, Y.-Y., Gan, Q., Suel, T.: “Local methods for estimating pagerank values”. In: International Conference on Information and Knowledge Management, 2005,pp. 381–389.

[9] Sakakura Y, Yamaguchi Y, Amagasa T, Kitagawa H. “An Improved Method for Efficient pageRank Estimation”. InDatabase and Expert Systems Applications 2014 Sep 1 (pp. 208-222). Springer International Publishing.

[10] Comandur S, Lofgren P, Banerjee S, Goel A. "FAST-PPR: Scaling Personalized PageRank Estimation for Large Graphs". Sandia National Laboratories (SNL-CA), Livermore, CA (United States); 2014 Feb 1.

[11] Mitliagkas I, Borokhovich M, Dimakis AG, Caramanis C. “FrogWild!: fast PageRank approximations on graph engines”. Proceedings of the VLDB Endowment. 2015 Apr 1;8(8):874-85.

[12] Luh CJ, Yang SA, Huang TL. “Estimating Google’s Search Engine Ranking Function from a Search Engine Optimization Perspective”. Online Information Review. 2016 Apr 11;40(2).

[13] Michael Wong. "How Keyword Density, Frequency, Prominence And Proximity Affects Search Engine Rankings". from Mike's Marketing Tools : http://www.mikes-marketing-tools.com/marketing-tips/keyword-densities.html, 2009.

[14] Rosenblatt, Frank. “Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms”. Spartan Books, Washington DC, 1961.

[15] [15] Rumelhart, David E., Geoffrey E. Hinton, and R. J. Williams. "Learning Internal Representations by Error Propagation". Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1: Foundations. MIT Press, 1986.

[16] McCulloch, Warren; Walter Pitts ."A Logical Calculus of Ideas Immanent in Nervous Activity". Bulletin of Mathematical Biophysics. 5 (4): 115–133.1943.