Is it possible to guess the question on a web search?

Cheng Chen

Research School of Computer Science,

The Australian National University Canberra, Australia

Abstract. This paper will look at the relationship between the type of searched question and user' net searching habits. The data chosen comes from a research report "What Snippet Size is Needed in Mobile Web Search." [1] The research report focuses on the snippet size for mobile web search. The paper conducts an experiment to examine the relationship between snippet size and indicators. In this paper, we adapt the data and change the research direction slightly. Our goal is to find a relationship between searched question and net searching habit. The result has shown that the use of MLP does not improve the chances of getting the right prediction

1 Introduction

Personal privacy has been a popular topic in the recent years. Search engine provider such as Google and Yahoo has detailed and solid privacy policy. However, is it possible to guess a question based on user's web searching habits and time spent on different types of question? Guessing a detailed searched question seems challenging because they are relatively unrelated to user's web searching habits. However, it seems reasonable to guess a type of question based on few indicators. The overall findings suggest that, it is also difficult to guess the type of the questions based on few indicators with the use of multilayer perceptron network.

2 Data Preparation

In the research paper, the type of question is divided into two categories, informational and Navigational. [1]. One example of informational question is "You are interested in some facts about the Golden Gate bridge in the U.S. In what years was the bridge construction completed". One example of navigational question is "You are interested in shoes from Adidas. Find the

official Adidas homepage".

In this paper, we are going to use features such as satisfaction, accuracy, time to complete to find out the correlations between these features and type of question (informational or navigational) A research paper suggests that users take more time to complete informational tasks. The feature time to complete might be a good indicator for us to guess the type of question. [2]. In contrast, another paper suggests that users tend to spend less search time with the long snippets for informational tasks, whereas the long snippets for navigational tasks required more time. [3]

In the previous paper, the input features are snippet length, time to first click, Log(TTF), Accuracy, Satisfaction, Scroll, Task_Num and show_taskNum. [4] However, the result does not show any strong correlation between the features above and type of question. This time, features such as Task_Num and show_taskNum are removed. It is aimed to increase the accuracy of the model.

Furthermore, values in satisfaction are normalized. Before normalization, the value is in the range of 1-7. As compare to values in other column, the range is relatively large. Hence, normalization of the satisfaction column might help to increase the accuracy of the model.

3 Method

The model we are using is called Multi-layer perceptron (MLP). There are many other deep learning techniques such as Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). Although CNN and RNN are more popular in deep learning nowadays. However, they are commonly used in certain context.
Traditionally, the CNN input is two-dimensional. The network is often used to classify images. Therefore, CNN works well with data that has a spatial relationship. In this case, the data clearly has no spatial relationship. RNN is designed to work with sequence prediction problems such as text prediction. It is therefore not suitable for this case.

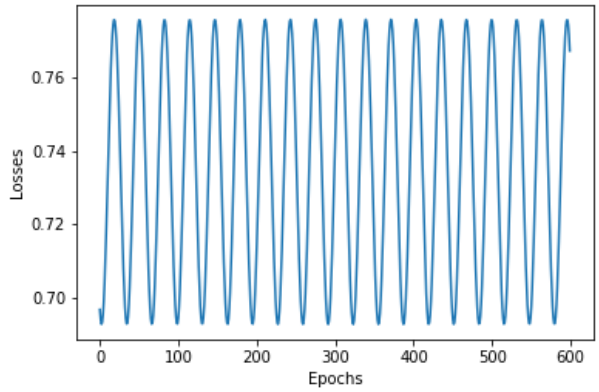
The number of hidden layers is an important parameter for neural network architecture. [5] In the single-layer neural network, there are 6 input neurons which represent the input features, five hidden neurons and 2 output neurons which represent the two classes. The activation function is sigmoid function. We use sigmoid function because the output exists between 0 and 1. It is often used for model where we want predicts the probability as an output. In this case, we want to predict the whether a question is informational or navigational. Therefore, the probability of informational and navigational will be either 0 or 1.

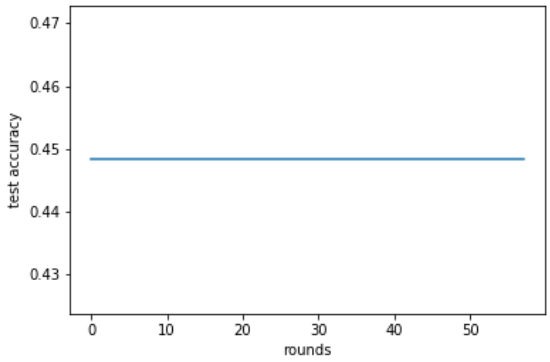
The second neural network is a multilayer neural network. It consists more than one hidden layer. Multilayer network is more powerful when the activation function is non-Linear. The sigmoid functions act almost linear for small absolute values of the argument and are saturated. [6]. This validate the property of the multilayer perceptron to act as a universal approximator. [6]. We aim to increase the hidden layers to improve the accuracy of the model. It is not sure whether the accuracy will increase or not because it depends on the complexity of the problem. It is also important to note that if there are more than sufficient number of layers, there might be a problem of overfitting and the accuracy will decrease. [7]

4 Results and Discussion

When using MLP, it is important to tune the hypermeter. Hypermeters play an important role in training MLP network. In the following part of the paper, we test the MLP network with different parameters to decide the best combination of the parameters.

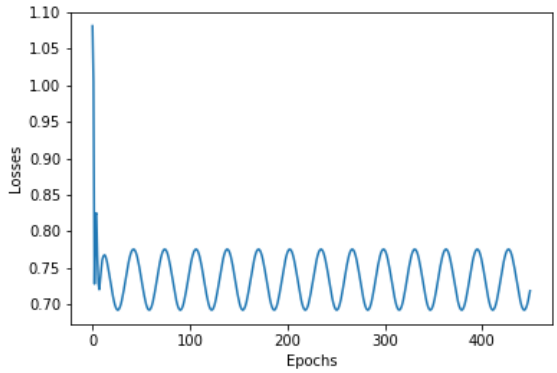
Combination 1(learning rate =0.01, Epochs=600):

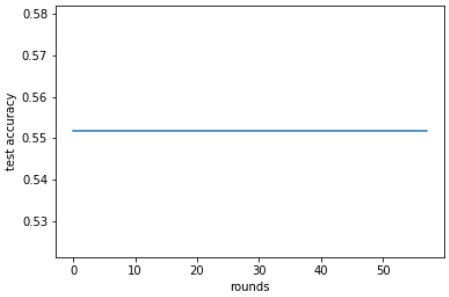




The training loss is between 0.7 and 0.77. and the test loss is under 0.45

Combination 2: (learning rate=0.01, Epochs =450)





The training loss is between 0.7 and 1.1 and the test loss is consistent at 0.55.

From the experiment above, we can see that the combination of learning rate=0.01 and epochs = 450 gives a better model.

5 Conclusion and Future Work

The overall finding suggests that the proposed model has a slightly better accuracy than normal prediction. From the model above, we can see that the test accuracy is 55.1% which is slightly better than normal guessing 50%. Even though the model is improved by tuning the parameters, the model does not seem improved a lot.
For the future works, we may be can include more kinds of data in building up a model. The more relevant input features can improve the accuracy of the model. Moreover, an evolutionary algorithm can be used to determine the hypermeters. The future work can focus on how evolutionary algorithm can improve the model

by choosing parameters.

References

[1] P. T. R. S. T. G. H.-. J. Y. Jaewon Kim, "What Snippet Size is Needed in Mobile Web Search," Reseach School of Computer Science, Statistical Consulting Unit. The Australian National University, Canberra, Australia, 2017.

[2] B. H. T. J. L. a. G. L.Lorigo, "The influence of task and gender on search and evaluation behavior using google," in *Information Proceeding & Management* , 2006.

[3] E. a. Z.Guan, "What are you looking for ? An eye-tracking study of information usage in web search," in *Proceedings of the SIGCHI conference* , 2007.

[4] C. Chen, "Is it possible to guess the question on a web search," in *ANU 3rd bio-computing conference*, Canberra, 2020.

[5] C. M. H. P.-M. Graham R. Brightwell, "Multilayer Neural Networks: One or Two Hidden Layers," in *NIPS*, 1996.

[6] V. E. B. L. P.-P. N. E. M. Popescu Marius, "Multilayer perceptron and neural networks," in *WSEAS Transactions on* , July 2009.

[7] I. G. a. Y. B. a. A. Courville, Deep Learning, MIT Press, 2016.