FILTERED BOOK FEATURES WITH ARTIFICIAL NEURAL NETWORK: A CASE STUDY OF THE BAN NONG SAENG SAMAKKEE BOOK OF THE AMSS ++ SYSTEM OFFICE OF KHON KAEN PRIMARY EDUCATION AREA 5

Sangkaphat Khampong[1] and Urachart Kokaew[2]
[1]Department of Computer Science, Faculty of Science, Khon Kaen University,
123 Moo 16 Mittapap Rd., Muang District, Khon Kaen 40002,Thailand.
[2]Ubiquitous Computing Laboratory, Department of Computer Science, Faculty of Science,
Khon Kaen University, 123 Moo 16 Mittapap Rd., Muang District,
Khon Kaen 40002, Thailand.

ABSTRACT

Receipt letters were documents submitted to education institutes in primary educational service area to inform assigned educational institutes or relevant educational institutes to perform any transactions without specifying specific educational instituted. Thus, both relevant and irrelevant receipt letters were delivered to the educational institutes and it considerably affected small schools where there were no administrative teachers. Then, other teachers had to check if there were any receipt letters delivered each day and it wasted their teaching time. For this reason, an artificial neural network was developed to sort out the receipt letters using the data of receipt letters from AMSS++ for 4 January 2016-10 January 2018. When tested with 1,348 subjects of the receipt letters, the accuracy of the back-propagation neural network (BPNN) was at 97.60 percent which represented that the analytical efficiency of this model was at good level.

KEYWORDS: Back-propagation Neural Network, Receive documents

1. Introduction

AMSS ++ System is education Area Management Support System for the Office of Educational Service area to use the information in four management areas and is the general management. The system is used by the Office of Educational Service Area for all educational institutions. The revised and supplemented information will be provided as a supplementary information booklet that is constantly updated and used by school personnel of each school.

The content of the information is obtained from the book. Sometimes it is not relevant to what the institution is doing. For example the researcher noted that the problem was solved and developed an analytical system that will help to get the information that meets the needs of the institution. The neural network model was studied to find the most appropriate data model for the study series in order to be able to select information and respond quickly.

2. Model Methodology

2.1 AMSS++ System

AMSS++ System is a support system for the operation of the Education Service Area Office which is the education for district office to store data directly. The Office of Educational Service Area Management (AMSS ++) has been developed as a tool to support the Office of Educational Service Area Office to efficiently and effectively manage the Office of Educational Service Area Office. AMSS ++ is an open source software that gives everyone the opportunity to develop programs that can be tailored to meet the needs of the office. The study area is complete.

2.2 Neural Network

MLP is a form of multi-layer neural network used for complex tasks that work well. The training process is supervised and using the reverse procedure. The backpropagation for training the return process consists of two subheadings, forward pass and backward pass for forward passage. The data passes through the neural network at the input layer and is passed from one layer to another until the data layer is out. The return value of the connection weight will be adjusted in accordance with the error correction rule. Error-Correction is the difference between the actual response (response) and the target response (Error Signal). This error signal is sent back to the artificial neural network in the opposite direction to the connection and the weight of the connection is adjusted until the actual response approaches the target response.

MLP has two types of signals: function signal and error signal. Function Signal is the input signal from the node in the previous layer. Signal Error is a reverse signal that occurs at the node in the data layer of the artificial neural network. It is transmitted from one layer to another [1].

The MLP principle is that each layer of the hidden layer has a function for calculating the output from the node in the previous layer, called the activation function. Each layer does not need to be the same function hiding, is an important function. It tries to convert the data into a layer so that it can distinguish it using a single linear (Linearly Separable) and before the data is sent to the output layer. More than one layer is required to convert data into a linear form [2].

To calculate the output for a classification problem, input the info into the neural network that we have already found. Then compare the values of the output in the output layer and select the higher value of the output (higher neuron) and take the predicted value corresponding to the selected neuron. If the value is in the receiving range (error is less than the error we set), then the next set of data is received. If the value is greater than the acceptable value, adjust the weight and biased according to the procedures mentioned above. When weight adjustment is done, continue the next set of data and repeat the process until the last data set. When finished, it is counted as 1 cycle of calculation (1 Epoch). Of the average, I have kept it, and use to check that the value on average, in the classification. For the value less than the acceptable error value, the artificial neural network generated can produce the correct result of every data. If you do not, go back to the first step. Get started with the new data set [3].

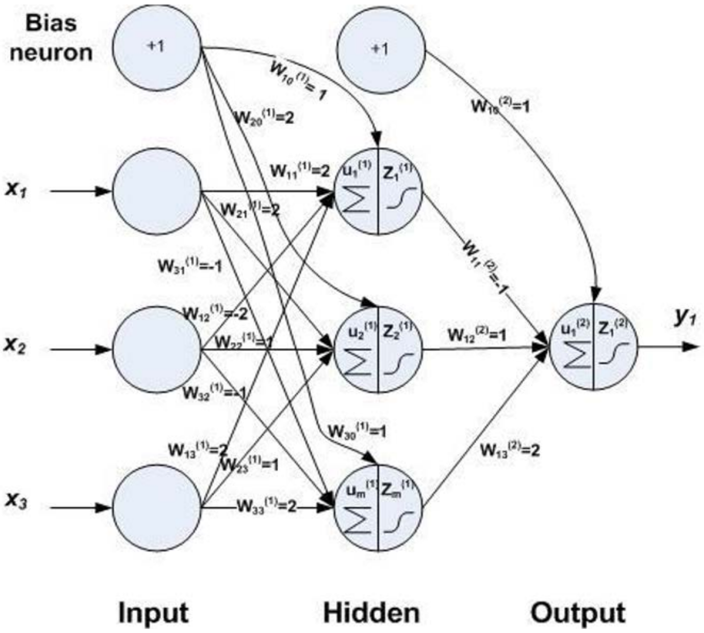

Figure 1 Connected feed-forward network with one hidden layer and one output layer

2.3 Data mining

Data Mining is a process that involves a lot of data searching for the patterns and relationships that are hidden in that data set. Web-based data mining techniques involve processing web-based text and searching for the same knowledge within unstructured data. Data mining involves the use of data stored in text in a form suitable for automatic processing. The purpose of data mining is to convert unstructured source data to retrieve meaningful numbers from text and make the information available in the message accessible through a variety of data mining techniques. This includes retrieving summary information, displaying text, grouping messages, and grouping documents [4].

2.4 Word Segmentation

It separates each word in a sentence from a Thai document that has a contiguous word. The purpose is to index the document (Document indexing) by the main method used. There are three main methods: Rule-based, dictionary use and corpus use, used in word wrapping.

2.4.1 Longest word pattern matching

This method examines the input (left-to-right) string and then compares it to the existing word in the dictionary. If one syllable is found in the dictionary, the longest syllable is chosen.

2.4.2 Shortest word pattern matching

This method is similar to the longest word matching method, but only the shortest word is found first, but this method has the highest number of words, but the accuracy of the word is less than the word used in the longest word matching method.

2.4.3 Word usage frequency

This method is one way of solving the problem of Thai sentences. Analyze the frequency of daily use of words by arranging the words in the dictionary at the frequencies found and using the word wrapping method in the same way as in clauses 2.4.1 and 2.4.2.

2.4.4 Back tracking

When comparing the words that are cut to the word contained in the dictionary, if can be that more than one word is found, and the longest word is selected, so the strings that follow that word cannot be cut because they are not found by dictionary. In this case, the word is not selected and the next word is cut [5].

3. Related work

3.1 Data pre-processing

The data for this analysis is based on the book received in AMSS ++ from 4 January 2017 until 18 January 2018. It was sent to all schools in Khon Kaen Area 5 which content within the book. Then take the whole topic and make a list of words and symbols. The list of words and marks will be used as a feature, with a value of 0 if it is not related to the school, and a value of 1 if it is related to the school.

3.2 Feature identification

In the text to be considered, there will be compiled words that do not correspond or words to be used to separate whether the content in this topic needs to be read or not compile in the XML document (data dictionary). Word wrapping from the import document split into sub-clauses using spaces as separators and determine which clauses can be immediate words. Follow by the sentence to cut the first word for example, numbers, abbreviations in parentheses, etc., for easy navigation. After that, the words from all available topics will be compared to the XML documents provided. The term used to identify the characteristics of the classification, such as the sentence that the words red cross, aliens, etc. will consider that the book does not need to be read.

3.3 Data processing

Use the breakIterator class, which is IBM's library for word wrapping. The Longest Machine is a dictionary that divides the document into two parts, the title and the content. This research has defined the qualities of word or phrase by statistical analysis. Consider up to four words of contiguous word groupings as they appear in the document can be considered as a group of words used to indicate. Evaluation of the importance of words or phrases calculates the priority values that words or phrases have on a document. Use keyword-specific configuration techniques to replace documents as vector models. The vector-space model is the product of TF (Term Frequency) and Inverse Document Frequency (IDF), which compares the frequency of word segments in a document with the frequency of words in another document. The presence of the word in different parts of the document is calculated from the formula.

$$TF{*}IDF \quad = \quad freq(P,D)\ {*}\ \log_2 \frac{N}{n_p} \tag{1}$$

By D: document, P: word or phrase, freq(P,D): Frequency found in word P in D, n_p: Number of documents found in the database , N: Total number of documents in the database.

Each group of words in the document will be calculated TF * IDF, 2 words, separated by title and content in summary. This value is the input of the neural network for word extraction.

3.4 Neural Network design and training data

Artificial neural network modelling uses a multi-layer perceptron network consisting of 56 input nodes, 6 hidden nodes, and 2 output nodes.

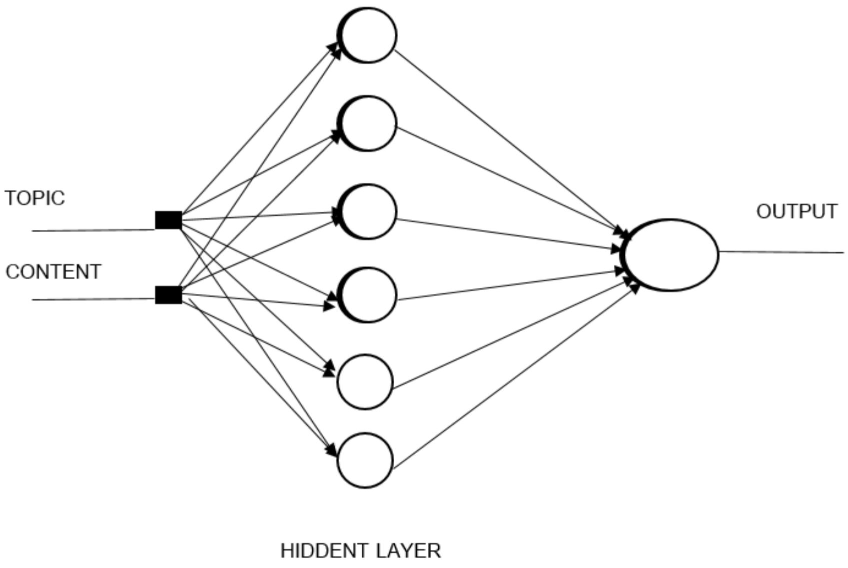

Figure 2 An example of a feed-forward neural network uses model

4. Experimental evaluation

The performance test of the 56-6-2 model (input layer - hidden layer - output layer), which was adopted as the format for extracting the text from the dataset, was chosen by selecting a series of 1,348 data sets. The key messages with no importance are 3 sets.

1. Learning Resources 1000 topics divided into 800 topics and 200 topics.
2. Data set for examination 300 topics divided into 200 topics, 100 topics, not important 100 topics.
3. Data set used in the test 48 topics divided into 40 important topics, not important 8 topics.

5. Conclusion

This research aims to introduce the theory of multi-layer forward neural networks using reverse lookup to adapt to the filtering of sensitive topics. Then the stabilizer and the value of the variables can be used to calculate. These procedures are used to separate the topics from the data. By comparing the performance, the 56-6-2 neural network has the smallest error value. It is suitable to be used for sorting important topics.

Acknowledgement

The authors would like to thank Department of Computer Science, Faculty of Science, Khon Kaen University for financial support.

References

[1] Trappey AJC, Hsu F-C, Trappey CV, Lin C-I. Development of a patent document classification and search platform using a back-propagation network. Expert Systems with Applications 2006;31(4):755–65.

[2] Monirul Kabir M, Monirul Islam M, Murase K. A new wrapper feature selection approach using neural network. Neurocomputing 2010;73(16):3273–83.

[3] Alghamdi HM, Selamat A. Arabic web page clustering: a review. Journal of King Saud University - Computer and Information Sciences [Internet]. 2017 [cited 27 April 2018]. Available from: http://www.sciencedirect.com/science/article/pii/S1319157817300290

[4] Wang R, Ji W, Liu M, Wang X, Weng J, Deng S. Review on mining data from multiple data sources. Pattern Recognit Lett [Internet]. 2018 [cited. 27 April 2018]. Available from: http://www.sciencedirect.com/science/article/pii/S0167865518300199

[5] Kafetzopoulou L, Boocock D, Dhondalay GKR, G Powe D, Ball G. Biomarker Identification in Breast Cancer: Beta-Adrenergic Receptor Signalling and Pathways to Therapeutic Response. Comput Struct Biotechnol J 2013;6:e201303003. doi: 10.5936/csbj.201303003.

Author's Profile

Sangkaphat Khampong, is a graduate student Department of Computer Science, the Faculty of Science, Khon Kaen University (KKU), Khon Kaen, Thailand. E-Mail: sangkaphat@kkumail.com

Urachart Kokaew, Ph.D. is an Assistance Professor Department of Computer Science, Faculty of Science, Khon Kaen University (KKU) in Khon Kaen, Thailand. She has a BSc in Computer Science (Hons), and MSc in Computer Science from school of advanced technology at Asian Institute of Technology (AIT), and PhD in eLearning Methodology at Assumption University in Bangkok, Thailand. E-Mail: urachart@kku.ac.th.