The Application of Artificial Neural Networks with the Multilayer Perceptron Classification for Discovering Webpages

Arpaporn Angsachon, Chidchanok Lursinsap, and Somsak Sriborisutsakul

***Abstract*—**This experimental study aims to apply the method of artificial neural networks with the multilayer perceptron for machine learning in pattern recognition and producing better retrieval results of unstructured data sets. The researchers used the Waikato Environment for Knowledge Analysis (WEKA) software as a toolbox to generate machine learning algorithms for data mining tasks. The main datasets were divided into five sub-datasets. The multilayer perceptron classification was applied to every group, but there were differences in word frequencies and sequences among them. After the experiment, it is found that the first sub-dataset provided the best overall results because of its learning performance of recognition. It implies that the application of artificial neural networks with the multilayer perceptron classification significantly improves the pattern of recognizing word meanings. This potential technique also reinforces an effort to develop a novel system of web information retrieval.

***Keywords*—**digital collection, information retrieval, multilayer perceptron

I. Introduction

INFORMATION technology plays on important role to our everyday life and the huge amounts of data increase rapidly. Information retrieval is an important tool for new researchers to find information, new ideas, and new topics for their works. It has become a challenge to a developer to create the effective tool for information retrieval in order to serve individual's needs. Collections in the digital libraries are major resources for researchers, but the current system of information retrieval does not support semantic search patterns. It simply combines textual search with an important rank, relevant output and keyword appearance [1]. For this reason, the existing retrieval system usually generates many irrelevant results which information searchers do not want them. Time constraint is another hindrance.

Arpaporn Angsachon is with the Technopreneurship and Innovation Management Program, Graduate School, Chulalongkorn University, Bangkok 10330 Thailand (telephone: 66-86-775-7109; e-mail: arpapora@yahoo.com).

Chidchanok Lursinsap is with the Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University, Bangkok 10330 Thailand (e-mail: lchidcha@chula.ac.th).

Somsak Sriborisutsakul is with the Department of Library Science, Faculty of Arts, Chulalongkorn University, Bangkok 10330 Thailand (e-mail: somsak.sr@chula.ac.th).

General users cannot get appropriate results quickly because of hundred items found in each display. They decide to choose some results from only the first two webpages. Thus, the filtering process of desired information among a large number of results challenges digital library experts to the improvement of the information retrieval system.

The digital collections or web pages can be undoubtedly discovered by search engines with specific keywords. The users, however, expect to get results showing anticipated meanings. In fact, many words have varied meanings and can be used in different parts of speech. The search engines cannot distinguish one from other homonyms. For example, "fly" (noun or verb) can be used in different parts of speech and various meanings. The users have to filter the results again by themselves and spend more time to extract the results precisely and suitably from the first few pages. This affects laypersons who have less experience to express their needs in exact keywords in any queries. The rationale behind this study is to apply the method of artificial neural networks with the multilayer perceptron classification for improving the retrieval system that provides semantic keywords found on web sites or in digital documents.

II. Related works

Whenever web authors want to compose their digital documents; it is unavoidable to meet the homonyms appeared in them. Retrieving these collections, it is necessary to have a useful aid to differentiate between search terms spelled the same on the basis of their true meanings to enable us to obtain information we need. In practice the searchers look at the context that comes before and after a word, phrase or statement to help find out its meaning [2]. In other words, the users select documents to match to their needs by considering the context of the documents relative to the search terms. Previous attempts at building the information retrieval systems for filtering relevant keywords in titles, abstracts, and paragraphs of full papers have been developed gradually. Later on, the concept of the machine learning process is introduced to simulate information extraction that resembles human recognition by displaying the precise search results with less time constraint.

The evolution of retrieval system development in general is similar to the development of digital library systems in particular. Digital collections are the digitized format of

http://dx.doi.org/10.15242/IIE.E0414059

information resources delivered by information service units or organizations. These collections allow users to have a remote, online access according to topics in which they are interested. Examples of the digital collections include journal articles, electronic books, digital images, bibliographies, gazetteers, chronologies, directories and so on [3].

Researchers in the study area of digital library systems apply some techniques in relation to machine learning and information retrieval approaches of webpages to the development of digital collection access. They are as follows: [4]

1) Topic and event tracking

2) New adaptive information retrieval methods based on training-language models for the collections and queries.

3) Novelty detection

4) Hierarchical text categorization

5) Unsupervised clustering, including hierarchical clustering for discovering potentially useful ontologies;

6) Cross-language information retrieval, tracking and clustering

Recently, data mining has been employed to advance the digital library system. This technique is the process of using language-learning patterns to discover unknown, hidden, and meaningful information in databases. Its tasks consist of description, estimation, prediction, classification, clustering, and association of information to be searched [5]. The data mining field of research arises from the explosive growth in the number of digital information collections, information repositories, corporate memories, data warehouses, business intelligence, etc. There are many software packages developed to mine such data in the information technology sector [1], [6].

The data mining applied to this experiment relies on three components – artificial neural network, multilayer perceptron, and back propagation. Each component has its own details.

Artificial neural network (ANN) often called a neural network. ANN is a mathematical model or computational model based on biological neural networks. ANN represents an attempt at a very basic level to imitate the type of nonlinear learning that occurs in the networks of neurons found in nature. The most common activation function is the sigmoid function [5], [7]. ANN consists of an interconnected group of artificial neurons. The information processing uses a connectionist approach to computation. ANNs are usually used to model complex relationships between inputs and outputs or to find patterns in the given data [8], [9]. The output value is compared to the actual value of the target variable for the training observation. The error calculation is the difference between actual value and output value. Neural network model uses the sum of squared error in prediction measurement. The fitting of output predictions becomes the actual target values [6], [10]-[12].

Multilayer perceptron or multilayer feed forward neural network comprises multiple layers of computational units (input layer, hidden layer, and output layer). Each neuron in a layer directs connections to the neurons of a subsequent layer. It is usually interconnected in a feed-forward way [5], [7]. There are no cycles or loops in the network. The combination function of the structure of input layer nodes, hidden layer nodes, and output layer nodes produces a linear combination of the input node. The connection weighs into a single scalar value in term of "net" [1].

Back propagation is a popular method of training multilayer feed forward ANNs [7]. The machine learning process has two stages which are a training stage and a testing stage [6], [8], [9]. The training stage is used for supervising and training the designed multilayer perceptron in accordance with its training data. It is a pattern for the supervised test set. In the testing stage, the test set of calculated weights of neural network from the first stage. Its result is used to estimate the overall ranking of documents for new examples in this second stage. The test output used for estimating the values of the attributes for information corresponding to the desired output and its ranking in the normalized rank table.

The back propagation algorithm uses the supervised learning to provide the network computation algorithm with the examples of the inputs, outputs, and errors, i.e. differences between the actual and the expected results [9]. The artificial neurons send their signals "forward" and then the errors are propagated backwards. The ANN contains many nodes with weights assigned to each connection. It can learn to work around noisy data or uninformative examples in the data set. Each connection between nodes has a weight associated with it. At initialization, these weights are randomly assigned to the values from zero to one. The training set and the test set include columns of input nodes and interconnections in a feed-forward way.

III. Proposed Concept

Enabling a computer to understand human languages remains to be done. An alternative information retrieval system for digital collections, i.e. webpages, is proposed in this experimental study. It presents the system that can train and test the relative keywords derived from the context of digital documents or web sites. The interesting function is multilayer perceptron classification. It simulates the learning process for the retrieval model. It is also adopted in many previous works to calculate the vector of keywords [6], [13]-[15]. This study reviews and compares the existing retrieval techniques and machine learning techniques currently used in the present. The investigators chose the simple word with several meanings to be the datasets for testing the learning processes. The word "house" was selected with the specific meaning as a building people living in, usually for one family [2]. There are two main dataset schemes: the training set and the test set prepared for data mining. Classification function in the data mining is one of the suitable choices in this research setting. Next, the application of the machine learning is employed in this study. It seems similar to humans' decision making tasks encountered in everyday life. The process for grouping classes of data belongs to its feature by finding common traits or characters. The main objective of classification is to predict categorical class attributes for new samples [10]. Six steps conducted in this study are summarized below:

1) Get initial input data from webpages and databases through search engines

2) Extract the data from each web or document containing the assigned term "house" by focusing on its frequency, order of appearances, and statistical data in each selected sample. Then consider the words to understand the structure of data representation in each input data.

3) Group the data into the contextual types of the studied theme

4) Create the sub-datasets with the input file format for learning classification and complete all missing values.

5) Run the machine learning process with the multilayer perceptron function in both the training stage and testing stage

6) Compare the output derived from the multilayer perceptron with other classification techniques.

IV. METHODS

From the above-mentioned concept, the study began with data preparation from the assigned keyword "house". The data were gathered from webpages and databases. The 38 selected attributes relevant to "house" in terms of the learning pattern and testing were divided into four groups. The first group consisted of attributes with the same meanings or synonyms for "house". The second group had thirteen types of related terms which were frequently found and rarely found. The third group was comprised of the compound words with "house". The last group had five attributes which were nouns possibly relevant to the meanings of "house". All scheme attributes are shown in table I and II.

The second process was to build up the dataset from raw data. Each dataset had 250 training set in instances and 50 test set in instances. The actual value of the data was set to "1" if the relevant records were in the condition of correct meanings. Meanwhile, "0" would be set for the irrelevant records. There were five sub-datasets in the learning system. The selected data followed the sequential records and were categorized into each sub-dataset. The training set had 250 instances that could separate class "1" from class "0" in an equal number of 125 instances per class. In the meantime, there were 25 instances of the test set per class.

The datasets were stored in CSV format which were readily executed with the WEKA (Waikato Environment for Knowledge Analysis), an open source data mining program [7], [16]-[17]. The back propagation algorithm used the supervised learning pattern and its testing was calculated. All the attributes of the first training and testing round were executed one by one until every dataset was completed and sorted by the least error found in the evaluation results. The feed forward processing was also done through all the attributes to accomplish the learning pattern in the context of keywords regarding to the extraction of irrelevance. The learning and testing continued to show least error-ranking attribute+1. Remaining attributes would be new input nodes, and then lead to the next learning and testing loops.

The training and testing sets were processed by the back propagation algorithm until the input attributes of both had the evaluation results that reached the least error and would be stopped the learning and testing [8]-[9]. The performance measures for numeric prediction were, therefore, used to assess results of the multilayer perceptron [6]. The predicted values on the test instances were $p_1, p_2, \ldots, p_n$. The actual value sets were $a_1, a_2, \ldots, a_n$. The measurements were

TABLE I
THE SCHEME ATTRIBUTES GROUP A AND GROUP B

TABLE II
THE SCHEME ATTRIBUTES GROUP C AND GROUP D

Group C	Group D
C1 house/home builder	D1 family
C2 house/home buyer/purchaser	D2 housing
C3 house cleaning/housekeeping	D3 live/living
C4 household	D4 rent
C5 house/home income	D5 staying
C6 homeless	
C7 house/homeownership	
C8 house/home price	
C9 housewife/homemaker	
C10 house/homework	

$$\text{Correlation coefficient} = \frac{S_{PA}}{\sqrt{S_P S_A}} \quad (1)$$

Where,

$$S_{PA} = \frac{\sum_i (p_i - \bar{p})(a_i - \bar{a})}{n-1}$$

$$S_P = \frac{\sum_i (p_i - \bar{p})^2}{n-1}$$

$$S_A = \frac{\sum_i (a_i - \bar{a})^2}{n-1}$$

Which p_i referred to the value of the prediction for the i^{th} test instance, and $\bar{a}$ was the mean value over the test data.

$$\text{Mean-absolute error} = \frac{|p_1-a_1|+\cdots+|p_n+a_n|}{n} \quad (2)$$

$$\text{Root mean-squared error} = \sqrt{\frac{(p_1-a_1)^2+\cdots+(p_n-a_n)^2}{n}} \quad (3)$$

$$\text{Relative-absolute error} = \frac{|p_1-a_1|+\cdots+|p_n-a_n|}{|a_1-\bar{a}|+\cdots+|a_n-\bar{a}|} \quad (4)$$

$$\text{Root relative-squared error} = \sqrt{\frac{(p_1-a_1)^2+\cdots+(p_n-a_n)^2}{(a_1-\bar{a})^2+\cdots+(a_n-\bar{a})^2}} \quad (5)$$

referred to "(4)" and "(5)", $\bar{a}$ was the mean value over the training data

The final process was to make a comparison of the multilayer perceptron results with the other retrieval techniques. This study compared the results of F-measure [1], [7], [13], [18] with cosine similarity, euclidean distance, and extended jaccard coefficient [19]. F-measure became the harmonic mean of precision and recall.

$$\text{F-measure} = \frac{2 \times recall \times precision}{recall + precision} \quad (6)$$

$$= \frac{2 \times true\ positive}{2 \times true\ positive + false\ positive + false\ negative}$$

The new information retrieval model would be developed toward a novel system suitable for non-expert users who want to search for webpages and collections of digital libraries.

V. Experimental Results

The multilayer perceptron classifier of the learning model for every sub-dataset is accomplished. The contextual relationships of each sub-dataset's learning output are as follow:

1) Sub-dataset 1 consists of attribute A2, D2, A8, C7, C9, and C2.

2) Sub-dataset 2 consists of attribute A2, D4, D2, C3, C7, A10, C4, C6, A8, C8, and B13.

3) Sub-dataset 3 consists of attribute A2, D4, A10, A9, D3, and A4.

4) Sub-dataset 4 consists of attribute A2, B2, D4, A7, C2, and C3.

5) Sub-dataset 5 consists of attribute A2, A9, D1, C3, A4, D5, and C2.

The expression of each sub-dataset output is relevant to the first attribute of every sub-dataset that is A2 home. There are combinations of keywords from three groups (A, C, and D) in sub-dataset 1, sub-dataset 3, and sub-dataset 5, which recognize the words from every group. These results indicate that there are relationships among the context of the words in these three groups. None of independent group can be found.

The evaluation results of multilayer perceptron learning are summarized in table III. The sub-dataset 1 retrieves the best evaluation testing results in correlation coefficient 0.9978, root mean squared error 0.1004, and root relative squared error 20.09%, whereas sub-dataset 5 has the worst overall measurement results.

Table IV presents the percentage of the correctness prediction results in the multilayer perceptron learning. The test set in sub-dataset 1 completes the correct prediction in both classes. The sub-dataset 3 has the least correct prediction in class "1" 87.5%, whereas the sub-dataset 5 has the least correct prediction in class "0" 71.4%.

TABLE III
THE PERFORMANCE MEASUREMENT OF MULTILAYER PERCEPTRON

	Correlation coefficient	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
Sub-dataset 1					
Train	0.6411	0.2495	0.3923	49.90%	78.47%
Test	0.9978	0.0964	0.1004	19.29%	20.09%
Sub-dataset 2					
Train	0.7436	0.1541	0.3491	30.83%	69.82%
Test	0.9395	0.0656	0.1713	13.12%	34.26%
Sub-dataset 3					
Train	0.8239	0.1729	0.2892	34.58%	57.84%
Test	0.8424	0.1708	0.2697	34.16%	53.95%
Sub-dataset 4					
Train	0.8036	0.2262	0.3184	45.25%	63.68%
Test	0.9137	0.1547	0.2132	30.93%	42.64%
Sub-dataset 5					
Train	0.7816	0.1544	0.3166	30.89%	63.33%
Test	0.7422	0.1775	0.3461	35.49%	69.22%

TABLE IV
THE PERCENTAGE PREDICTION RESULT OF MULTILAYER PERCEPTRON

	Predict=1 (%)		Predict=0 (%)	
	✓	×	✓	×
Sub-dataset 1				
Train	82.3	17.7	25.5	74.5
Test	100	0	100	0
Sub-dataset 2				
Train	90.1	9.9	69.5	30.5
Test	96	4	100	0
Sub-dataset 3				
Train	95.1	4.9	63.0	37.0
Test	87.5	12.5	92.3	7.69
Sub-dataset 4				
Train	96.7	3.3	48.9	51.1
Test	96	4	85.7	14.3
Sub-dataset 5				
Train	86.0	14.0	86.9	13.1
Test	95.8	4.2	71.4	28.6

Table V shows the results of the F-measure comparison between the multilayer perceptron and the other methods by using cosine similarity, euclidean distance, and extended jaccard coefficient. The F-measure rate comparison of test data in sub-dataset 1 for multilayer perceptron is equal to the extended jaccard coefficient at 1. The outcome seems to provide the perfect correct prediction. The cosine similarity has F-measure poorest rate, especially in the test data in sub-dataset 3 at 0.63. The highest rate of euclidean distance is 0.89 in the test sub-dataset 2. In conclusion, the average of the F-measures of all five sub-datasets test rates range from the multilayer perceptron and extended jaccard coefficient to the euclidean distance and cosine similarity respectively.

TABLE V
F-MEASURE COMPARISON

	Multilayer Perceptron	Cosine Similarity	Euclidean Distance	Extended Jaccard Coefficient
Sub-dataset 1				
Train	0.85	0.76	0.72	0.83
Test	1	0.70	0.81	1
Sub-dataset 2				
Train	0.88	0.77	0.76	0.88
Test	0.98	0.83	0.89	0.92
Sub-dataset 3				
Train	0.90	0.53	0.77	0.84
Test	0.91	0.63	0.74	0.84
Sub-dataset 4				
Train	0.90	0.63	0.78	0.88
Test	0.96	0.61	0.65	0.94
Sub-dataset 5				
Train	0.89	0.81	0.73	0.88
Test	0.90	0.75	0.67	0.78
Average Train	0.88	0.70	0.75	0.86
Average Test	0.95	0.70	0.75	0.89

VI. DISCUSSION

This experiment is undertaken with few cautions regarding to the provision of the dataset. The existing dataset aims to test in the basic search environment of general Internet users. The raw data are brought from the search results of three search engines — Google, Yahoo, and Bing. The selected items are not limited to the first page result because the higher ranks occasionally depend on Search Engine Optimization for competitive marketing reasons. Therefore, it causes the actual results ranking to sometimes disconnect the relevant or matched meanings from the search term. Several results have wrong linked or limited accessibility. The sample keyword "house" is chosen because it is a simple, generic term which can retrieve numerous search results. This word also has several appropriate for applying the learning condition based on it. To really cover digital collections, the dataset for the future study should come from open access journal articles or other databases subscribed to various digital libraries.

VII. CONCLUSION

The system learning applied in this study yields the satisfactory results. They illustrate the output attributes from five sub-datasets by emphasizing the attribute A2 home rated as the least error test. The unique result of the raw data A2 home in every dataset has the highest degree of frequency differentiation of the actual value between correct class and incorrect class. This outcome implies that the proposed concept works for the function of keyword search and retrieval. This system generates filtered results with the contextual meaning like an information searcher does in practice. The differences in attributes between these datasets can be described in detail relating to the meanings of words at the context. The multilayer perceptron classification provides the best overall test F-measure results in comparison with other techniques. The outcome of the study suggests that the multilayer perceptron can be an alternative to digital library experts developing the information retrieval system for discovering webpages or digital documents.

REFERENCES

[1] M.L. Khodra, D.H. Widyantoro, E.A. Aziz, and R.T. Bambang, "Information extraction from scientific paper using rhetorical classifier," *2011 International Conference on Electrical Engineering and Informatics*, July 2011. http://dx.doi.org/10.1109/ICEEI.2011.6021634

[2] A.S. Hornby, *Oxford advanced learner's dictionary of current English*, 6th ed., Oxford University Press, 2003, pp. 267, 630.

[3] M.K. Buckland, "The difference in reference collections", *Journal of Library Administration*, vol. 46(2), 2007, pp. 87-100. http://dx.doi.org/10.1300/J111v46n02_07

[4] J. Carbonell, Y. Yang, and W. Cohen, "Special issue of machine learning on information retrieval introduction", *Machine learning*, vol. 39, 2000, pp. 99-101. http://dx.doi.org/10.1023/A:1007676028106

[5] D.T. Larose, *Discovery knowledge in data: an introduction to data mining*, New Jersey: John Wiley & Sons Inc., 2005. pp. 5, 128-146.

[6] I. Colak, S. Sagiroglu, and M. Yesilbudak, "Data mining and wind power prediction: an literature review," *Renewable Energy*, vol. 46, 2012, pp. 241-247. http://dx.doi.org/10.1016/j.renene.2012.02.015

[7] I.H. Witten, F. Eibe, and M.A. Hall, *Data mining: practical machine learning tools and techniques*, 3rd ed., Morgan Kaufmann, 2011. pp. 180, 232-241, 469-472.

[8] R. Ali and I. Naim, "Neural network based supervised rank aggregation," *2011 International Conference on Multimedia, Signal Processing and Communication Technologies*, 2011, pp. 72-75. http://dx.doi.org/10.1109/MSPCT.2011.6150439

[9] R. Ali and I. Naim, "User feedback based metasearching using neural network," *International Journal of Machine Learning and Cybernetics*, November, 2013, pp. 1-11.

[10] E.M. El-Alfy and R.E. Abdel-Aal, "Using GMDH-based networks for improved spam detection and email feature analysis," *Applied Soft Computing*, vol. 11, 2011, pp. 477-488. http://dx.doi.org/10.1016/j.asoc.2009.12.007

[11] V. Golovko, H. Vaitsekhovich, E. Apanel, and A. Mastykin, "Neural network model for transient ischemic attacks diagnostics," *Optical Memory and Neural Networks (Information Optics)*, vol. 21, Allerton Press, Inc., 2012, pp. 166-176.

[12] TK.R.K. Rao and R.U. Rani, "Content based image retrieval through feed forward neural network," *2012 IEEE 8th International Colloquium on Signal Processing and its Applications*, 2012, pp. 228-231.

[13] F. Siraj, M.A. Salahuddin, and S.A.M. Yusof, "Digital Image Classification for Malysian blooming flower," *Second International Conference on Computational Intelligence, Modelling and Simulation*, 2010, pp. 33-38. http://dx.doi.org/10.1109/CIMSiM.2010.92

[14] M. Sasikumar and Y.S. Kumaraswam, "An improved support vector machine kernal for medical image retrieval system," *Proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering*, March 2012, pp. 257-260.

[15] L. Sylvain, C. Valerie, and G. Pierre, "Principles and properties of a MAS learning algorithm: a comparison with standard learning algorithms applied to implicit feedback assessment," *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, 2011, pp. 228-235.

[16] R.R. Bouckaert., et al., *WEKA Manual for Version 3-6-10*, Hamilton: University of Waikato, 2013. pp. 35-51.

[17] M. Selvanayaki, M.S. Vijaya, K.S. Jamuna, and S. Karpagavalli, "Supervised learning approach for predicting the quality of cotton using WEKA," *Information Processing and Management Communications in Computer and Information Science*, vol. 70, 2010, pp. 382-384. http://dx.doi.org/10.1007/978-3-642-12214-9_61

[18] B. Cheng, R.J. Stanley, S. Antani, G.R. Thoma, "Graphical figure classification using data fusion for integrating text and image features," *2012 12th International Conference on Document Analysis and Recognition*, 2012, pp. 693-697.

[19] Y.N Tu and J.L. Seng, "Research intelligence involving information retrieval – an example of conferences and journals," *Expert Systems with Applications*, vol. 36, 2009, pp. 12151-12166. http://dx.doi.org/10.1016/j.eswa.2009.03.015