ABSTRACT
In this paper, we present an HIV/AIDS FAQ retrieval using Artificial Neural Network (ANN). One of the challenges in FAQ retrieval systems is the mapping of user queries to appropriate questions in the FAQ repository so that the corresponding answer can be retrieved as an answer to the user query. To address this issue, a number of approaches have been proposed, most of which are based on traditional information retrieval techniques. In this paper we discuss our approach based on Neural Network which maps user queries to existing answer questions in the domain on which it is trained. The experimental results demonstrate that the Neural Network approach can effectively improve the performance of the HIV/AIDS FAQ retrieval system.

KEY WORDS
FAQ Retrieval system, HIV/AIDS FAQ, Neural Networks, Health Informatics.

1. Introduction

HIV/AIDS is an epidemic disease that has been widely spread especially in the African continent. African governments in conjunction with other stakeholders have implemented a number of counter measures to address the spread of HIV and AIDS. For instance, the government of Botswana has embarked on educational and awareness programmes using publications such as MASA HIV/AIDS FAQs (Frequently Asked Questions) booklet, conducting awareness promotions through radio dramas like Makgabaneng, establishment of call centres such as IPOLETSE [1-4] and training through the Knowledge, Innovation & Training Shall Overcome AIDS (KITSO) [5]. Many of the educational and awareness strategies implemented use HIV/AIDS FAQs as an effective approach to share information about the disease’s infection and many other issues. However, access to these FAQs has been a major challenge as this info is available either on printed materials or through call centers. Masizan-Katongo et al [6] have already noted that there are no appropriate, conducive and contemporary IT systems for accessing information on HIV and AIDS in Botswana.

A more convenient solution would be to get the question answer service through mobile phones just by sending SMS (Short Message Service) questions. The ultimate goal of our research project is to develop a question answering (QA) system that can answer any question people may have about HIV/AIDS through standard mobile phones. With such a system, people can send SMS questions using mobile phones and get the answer as an SMS on their cell phone. In this paper, we focus on the development of an automated FAQ retrieval system (a special type of question answer service) on HIV/AIDS. One of the major tasks in an FAQ retrieval service is to find questions in the FAQ repository that are semantically similar to a user’s question.

An automated FAQ retrieval system will automatically search the FAQ repository to see if the same or similar question exists in the repository. If the same or similar question is found, then the corresponding answer can be provided. However, determining the semantic similarity between a user question and questions in the FAQ repository is a difficult task. The difficulty is due to the fact that the same question can be expressed using different words which have similar meanings. To address this issue, a number of approaches have been proposed to improve the accuracy of measures of similarity between user question and FAQ questions.

One area that has a potential to address this issue is the use of machine learning techniques such as neural networks. In a neural network based system, the FAQ system is trained with sufficient number of question-answer pairs through which the system can learn how to answer questions in the domain on which it is trained.

In this paper we discuss our approach based on Neural Networks which maps user queries to existing questions in the FAQ repository using Multi-layered Feed forward Neural Network architecture (with back-propagation training). One of the advantages of the Neural Network approach is its ability to learn based on a given
training data set and then map user queries to FAQ questions based on its training.

The remainder of the paper is organized as follows: Section 2 presents related works on automated FAQ retrieval systems by emphasizing on the use of machine learning techniques to map user questions to existing questions in the FAQ repository. Sections 3 and 4 provide a discussion of the neural network approach to FAQ retrieval. Section 5 presents a discussion of the experimental evaluation and analysis of the results. Finally, Section 6 summarizes the main points of the paper and highlights of our future work.

2. Related Work

FAQ retrieval systems can be implemented using various models such as statistical, template-based, relevance feedback, probabilistic, Boolean, Fuzzy Set, Language model and machine learning [7-11].

Machine learning refers to the construction and study of systems that can learn from data. Consequently, a computer system is trained from examples using learning rules and is able to pick knowledge patterns, trends, relations from the training examples. The learning rules are stochastic or symbolic and use mathematical models which exhibit cognitive simulation, parallel processing, probabilistic computations, neuroscience techniques to extract knowledge or discover data patterns in huge databases or user queries[10]. After the training the system is able to predict, classify, or cluster. The most common machine learning models include Bayesian networks, Hidden Markov, Decision tree, Nearest-Neighbour, Artificial Neural Network (ANN), Genetic Algorithms and Support Vector Machines [8-11].

In this section, we discuss question answering systems that have employed machine learning techniques such as neural networks.

Desjardins et al [12] designed a self-organizing map neural network to extract the concepts embedded in a collection of documents. The system uses unsupervised learning and the competitive (Kohnen) learning rule. The system was trained using document vectors to enhance its sensitivity. However, the unsupervised training has been found to be time consuming.

Pandey et al [13] implemented a sentence recognition system using Hopfield Neural Networks that has an ability of identifying sentence similarity which have lexicon – grammatical diversity. Their system has a single layered binary recurrent Hopfield ANN. The user queries and stored questions key words are represented in binary and the system uses a nearest neighbour classification rule to map to respective clusters. The system managed to match input sentences to stored sentences with a precision of 92.2%. However, the system has large number of neurons and hence computationally expensive.

In [14, 15], Genetic Algorithm based Neural Network was used for information retrieval and compared with traditional information retrieval systems. The results indicate that it outperformed the traditional information retrieval systems. In [15], another comparison of the Genetic Algorithm approach with the Back propagation Neural Network Information retrieval system was carried out. The results show that the back-propagation neural network approach performed better than the genetic algorithm based approach. Another interesting work in [16], which combined traditional information retrieval approach, relevance feedback technique and Genetic Algorithms, demonstrated a better performing system.

Chen [17] made a comparison of Neural Network information retrieval system based on Genetic Algorithms and Back Propagation training rule and discovered that the Genetic Algorithm based model performed at best equal to variants of a Back propagation algorithm in very small scale networks but far less efficient in larger networks.

In summary, most neural network based information retrieval systems use Hopfield networks or Self Organizing Maps without hidden layers. In [18, 19], recommendation has been made that a Neural Network with one or more hidden layer is capable of solving nonlinear and complex problems. Mandal [20] indicated that the hidden layer acts as the intuitive processor of a Neural Network and can properly manage the non-linearity of problems during mapping text terms to appropriate text documents. In our research, the task of mapping an arbitrary HIV/AIDS user query to similar question in the HIV/AIDS FAQ collection is a nonlinear process and therefore requires a Neural Network with a hidden layer like the MultiLayered Feed Forward (MFN) Neural Network. Back propagation is a Multi-layer Least Squares (MSE) neural network training rule which adopts the generalization of the least squares procedure that works for networks which have layer (s) of hidden units between the input and output units.

3. The Artificial Neural Network Approach

An Artificial Neural Network (ANN) is an information processing system composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems by learning from examples. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process. Learning involves adjustments to the synaptic connections that exist between the neurons. A trained neural network can be thought of as an "expert" in the category of information it has been given to analyze. This expert can then be used to provide projections given new situations of interest.

The network is composed of a number of highly interconnected processing elements (neurons) working in parallel to solve a specific problem such as prediction or classification.

Neural Networks are typically organized in layers. Layers are made up of a number of interconnected nodes which contain an activation function. Patterns are presented to the network via the input layer, which
communicates to one or more hidden layers where the actual processing is done via a system of weighted connections. The hidden layers then link to an output layer where the answer is output. Most ANNs contain some form of learning rule which modifies the weights of the connections according to the input patterns that it is presented with.

There are four main steps in developing ANN-based solution as indicated in [21].

The first step is concerned with data transformation, scaling or normalization. The main purpose is to create input data on a uniform scale as variation in the input data can slow down or even prevent the training of the network. Common scaling techniques used include min-max, Sigmoidal or principal component transformations.

The second step focuses on the selection of appropriate neural network architecture. This involves the selection of architecture type, number of layers (input, hidden, and output) and the number of neurons in each layer. The selection depends on the particular problem to be solved. For example, in some applications, one hidden layer is enough to approximate a function but we may need more in other applications. The hidden layers provide the network with its ability to generalize. Increasing the number of hidden layers may increase the risk of over-fitting and increases computation time.

The third step focuses on the selection of a learning algorithm. The purpose of the learning algorithm is to train the network so that it can respond correctly to a given set of inputs. During learning, the weights are dynamically adjusted and when we reach the best performance of the network (i.e., difference between the ANN output and the actual output is small), we get parameters that result in the best performance. At the beginning of training, the initial value of weights can be assigned randomly or based on experience. In the process, the learning algorithm changes the weights to correctly perform a desired input-output association.

The fourth step is the validation step in which the performance of the trained ANN model is tested through some selected validation set data. Validation can be employed to monitor the network error during training to determine the optimal number of training iterations. It can also be used to find out the optimal number of hidden neurons. Usually, we select the NN which performs the best over the testing set. The testing set’s size usually ranges from 10% to 30% of the training set. For a three-layer network, it has been suggested that the hidden layer should have approximately the square root of the number of input neurons times the number of output neurons.

4. The ANN Model Development

The ANN architecture selected is a multi-layer feed-forward neural network. The multi-layer feed-forward neural network has three layers defined as input layer, hidden layer and output layer. Each layer consists of neurons which process input depending on the layer it belongs to.

![Multilayered Feed Forward Neural Network Structure](image)

\[ V_{ij} \] represents connection weights between input layer neurons and hidden layer neurons. For our system, \( V_{ij} \) represents connection weights between user question terms and FAQ terms. The entire input neurons represent the input feature vector as term weights for a user question or questions from the FAQ that are used during training of the system. The \( W_{jk} \) represents connection weights between hidden layer neurons and output layer neurons. In our system, \( W_{jk} \) represents connection weights between the FAQ term neurons to FAQ document neurons. The FAQ document neurons represent the set of Question-Answer pairs of the FAQ. Each neuron in the hidden layer represents a question in the HIV and AIDS FAQ collection. Each selected output layer neuron shall have a value computed by a mapping function which maps an input query to FAQ question.

The number of input neurons shall be defined by the Principal Component Analysis (PCA) transformed Vector Space Model (VSM) matrix of the HIV and AIDS FAQ questions. The number of output neurons shall be the total number of Question–Answer pairs in the HIV and AIDS FAQ collection which is 467 questions. The numbers of hidden neurons shall be determined experimentally since they affect the neural network’s ability to generalize its answering ability.

For our system, back propagation learning algorithm has been used. Back propagation works best with the supervised training method and Multi-layer feed-forward neural network. The learning rule autonomously extracts the functional relationship between input data and the expected output data embedded in a set of historical data set (i.e. training data set) and encodes it into connection weights. The training process defines a question to question similarity mapping through training the neural
network with an arbitrary HIV and AIDS questions feature vector and mapping it into existing HIV and AIDS questions of the FAQ.

For the implementation of our neural network based HIV and AIDS FAQ Retrieval system, we have experimentally determined the appropriate neural network parameters using MATLAB. An HIV/AIDS FAQ question term-FAQ questions (712 x 467) matrix training was used from the MASA booklet, IPOLESTE HIV/AIDS FAQ manual and others sources such as UN AIDS.

The neural network system uses 467 HIV/AIDS FAQs stored in hidden/output layer connection weights with 467 neural nodes at the output layer. Input and output layer nodes play a role in determining the hidden layer nodes as observed from heuristic knowledge [18, 19, 22, 23]. We, thus, used heuristic knowledge as a guide to define the least and highest hidden bounds to consider when experimenting for an optimal set of hidden neurons. Constructive technique was used to filter the optimal hidden neurons as deduced from the bounds set by heuristic knowledge. We used early stopping to prevent a neural network being over trained by using a validation data set from within the HIV/AIDS FAQ question term-FAQ questions.

The best performing neural network can be determined based on a measure of similarity [24] between the input HIV/AIDS query and HIV/AIDS FAQs. This measure would be any value ranging from 0 to 1. If \( \text{sim} \) has a value 1 it means that the two questions match and if it falls between 0 and 1, there is a degree of similarity. Five cut off points for the degree of similarity were determined as 0.5, 0.4, 0.3, 0.2 and 0.1.

Results based on neural network simulations using MATLAB were evaluated by students at Limkokwing University who are members of an HIV/AIDS awareness campaign coordinated by the Human Resources Development Council, student counsellors and lecturers as well. Participants were presented with 120 HIV/AIDS questions and the corresponding answers as determined by our system from the FAQ collection. Participants were asked to rate the correctness of the query results of the system. For each question, participants rated correctness of Top1, Top5, Top10 and Top 15 answers as shown in Table 1. For example, Top 1 means that only the top one answer is provided for a given HIV/AIDS question. Top 5 means that the 5 top answers are provided for given HIV/AIDS questions.

Table 1: Recall and Rejection Rate Values

<table>
<thead>
<tr>
<th>FAQ Retrieval</th>
<th>Recall Rate</th>
<th></th>
<th></th>
<th></th>
<th>Rejection Rate</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top 1</td>
<td>Top 5</td>
<td>Top 10</td>
<td>Top 15</td>
<td>Top 1</td>
<td>Top 5</td>
<td>Top 10</td>
<td>Top 15</td>
</tr>
<tr>
<td>VSM Based</td>
<td>18.33%</td>
<td>38.33%</td>
<td>48.33%</td>
<td>55.83%</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Neural Network Cut Off Point 0.1</td>
<td>38.33%</td>
<td>62.50%</td>
<td>76.67%</td>
<td>77.50%</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Neural Network Cut Off Point 0.2</td>
<td>38.33%</td>
<td>64.17%</td>
<td>76.67%</td>
<td>79.17%</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Neural Network Cut Off Point 0.3</td>
<td>37.50%</td>
<td>61.67%</td>
<td>75.00%</td>
<td>77.50%</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Neural Network Cut Off Point 0.4</td>
<td>38.33%</td>
<td>63.33%</td>
<td>75.83%</td>
<td>78.33%</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Neural Network Cut Off Point 0.5</td>
<td>39.17%</td>
<td>56.67%</td>
<td>72.50%</td>
<td>73.33%</td>
<td>0.83%</td>
<td>0.83%</td>
<td>0.83%</td>
<td>0.83%</td>
</tr>
</tbody>
</table>

For user’s questions, there are three scenarios: 1: the same question exists in FAQ collection; 2 a similar question exists in FAQ collection and finally 3: No same or similar question exists in the FAQ collection.

5. Experimental Evaluation

The evaluation of the system was done using recall and rejection rate. Rejection indicates the ability of the system to report the unavailability of similar questions in the FAQ collection. The results show that the neural network outperforms the traditional retrieval system according to Table 1. The two types of FAQ retrieval methods used and their performance based on recall rates for top 1, top5, top10 and top15 FAQs is shown in Table 1. The traditional information retrieval system based on the vector space model performs poorly against all the variants of the neural network based system which have been sensitised at various ranking cut off points. The best performing system is the neural network with a sensitised ranking cut off point of 0.2 which shows high recall rates as shown in Figure 2. The rejection rate shows a poor result. This could be caused by the fact that the new questions used for participants to rate had HIV/AIDS related FAQs and therefore the system was able to pick question-answer pairs which have a degree of relevance. We need to test the system with none HIV/AIDS questions and see how it performs. Secondly there is need to increase the FAQ collection so that the depth and coverage of FAQ questions becomes sufficient.
6. Conclusion

In this paper, we presented a neural network based approach for HIV/AIDS FAQ retrieval. The main purpose of the approach is to match arbitrary users’ questions to existing questions in an HIV/AIDS FAQ collection.

A comparison with keyword-based FAQ retrieval system shows a better recall and rejection of the Neural Network based approach. The experimental results demonstrate that the Neural Network approach can effectively improve the performance of the HIV/AIDS FAQ retrieval system.

Even though the neural network based FAQ retrieval system is a promising approach, there is a need to further explore this approach to improve its performance (e.g., recall rate) by increasing the size of HIV/AIDS FAQ corpus and the diversity of FAQs in the collection. It may also be necessary to explore incorporating other complementary information retrieval models.

References

[20] T. Mandl, Vague Transformations in Information Retrieval, Social Science Information Centre,


