ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
DEPARTMENT OF INFORMATION SCIENCE

# Collaborative News Filtering for Amharic: An Experiment Using Neural Networks

BY

LEMMA NIGUSSIE HABTE



*A thesis submitted to*
*the School of Graduate Studies of Addis Ababa University*
*in partial fulfillment of the requirements for the Degree of Master of Science in Information Science*

July 2005

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
Faculty of Informatics
Department of Information Science

# COLLABORATIVE NEWS FILTERING FOR AMHARIC: AN EXPERIMENT USING NEURAL NETWORKS

BY

LEMMA NIGUSSIE

**Name and Signature of Members of the Examining Board**

| | |
|---|---|
| Ato Getachew Jemaneh, Chairman, Examining Board | Getachew Jemaneh |
| Dr. Björn Gambäck, Advisor | for |
| Dr. Nega Alemayehu , Examiner | |

| | | |
|---|---|---|
| Chairman, Faculty | Getachew Jemaneh | 12/07/05 |
| | Signature | Date |
| Chairman, Graduate Council | Signature | Date |

# Dedication

*This thesis is dedicated to all my family, primarily to my late brother, Mulugeta Nigussie, who was eager to see my future!*

# *Table of Contents*

*Page*

*Page*

## ACKNOWLEDGMENTS

First of all, I would like to thank my advisor, Dr. Björn Gambäck, for the critical comments he has been giving me through all the work. I would like to thank Ato Kibur Lisanu, Department of Information Science at AAU, and Ato Paulos Gemechu for their continuous support during my stay at the department. I would also like to thank Ato Ermias Abebe, Ato Mesfin Getachew and Ato Workshet for their supportive comments.

My special thanks go to my beloved family for their support and encouragement and especially to my brothers Tewodros Nigussie and Abebe Nigussie, who helped me in typing my paper. My mother Etalemahu Tessema deserves special thanks for her love and encouragement from the start till the end.

My deep thanks go to my beloved friends, Meshesha Legesse and Samuel Eyasu, for the careful reading of the thesis and helpful comments. I would also like to thank all the staff members of the Department of Information Science, and all my friends at the school for their love.

## List of Tables



## List of Figures

## List of Abbreviations

| | |
|---|---|
| Artificial Neural Network | ANN |
| Automatic Collaborative Filtering | ACF |
| Backpropagation Network | BPN |
| Backpropagation | BP |
| Collaborative Filtering | CF |
| Content-Based Filtering | CBF |
| Information Filtering | IF |
| Information Retrieval | IR |
| Neural Network | NN |
| Processing Element | PE |
| Selective Dissemination of Information | SDI |
| Self Organizing Map | SOM |

# Abstract

*Information Filtering (IF) is an area of research where only a few documents are selected from a large collection in a dynamic flow of information. Particularly, filtering out news items from a collection has paramount importance in order that a news reader can easily find what he/she likes to read. Several research projects are underway to implement such a system.*

*Collaborative filtering aims at learning predictive models of user preferences, interests or behavior from community data, e.g., a database of available user preferences. It is complementary to content-based filtering and retrieval that is mainly built on the fundamental assumption that users are able to formulate queries that express their interests or information needs in terms of intrinsic features of the items sought.*

*Many newspapers are being published in Amharic. Almost none of them use automated systems for filtering news. With the increasing number of such news, it is evident that a lot of textual information is accumulated which makes it difficult for the reader to find few desired news from a collection. It was felt that an experiment should be underway to extract such news on the basis of collaborative interest.*

*The purpose of this research was, therefore, to explore the potential application of Artificial Neural Networks (ANN) for filtering Amharic news based on preferences of readers. The Backpropagation (BPN) and Self Organizing Map (SOM) algorithms were used to develop a model for Amharic news filtering where news items were selected from two popular Amharic newspapers. The preferences of reading these items, collected from active readers of the newspapers, were used to develop the first model whereas the weighted term-by-document matrix of the news items in the sample was used to classify the news items.*

*The experiment was undergone in twofold; developing a model for predicting user preferences of reading news items and classifying the news items in the sample to predefined categories. The results showed that ANNs can be used to model user preferences of news items written in Amharic. The Matlab neural network toolbox was used to develop both models.*

*The result indicated that with Model 1, containing the preference list, it could predict 83.3% of the preferences in the training set and 79.8% of the preferences in the test set. That is, a news item is likely to satisfy the readers in the test set 79.8% of the time.*

*Model 2, the Self-Organizing Map (SOM) model, was also trained so as to classify news items into each category of the news. The best model could classify the items 76.5% of the time. 72.9% of the news items in the test set were correctly classified into the respective category. However, as neural networks learn from large examples, extended research is recommended.*

# CHAPTER ONE

# INTRODUCTION

## 1.1 BACKGROUND

The availability and production of large amounts of information in the form of newspapers, magazines and other forms of human communication entails the question of sorting or filtering out the right one as human choices are limited.

The public needs to be updated on important public events such as actions of governments, social or economic trends, education and international relationships, which are often referred to as hard news (Ethiopian News Agency, 1993). The production of such news in large amounts makes it difficult to choose the right document from a given collection. So, in the area of information storage and retrieval, a lot of research has been carried out to help readers retrieving and filtering out only desired documents.

Information filtering (IF) is studied in wider domains of readers of these documents. Previously, Selective Dissemination of Information (SDI) was the interest of research. SDI, as described by Foltz & Dumais (1992) and Oard (1997), was designed as an automatic way of keeping scientists informed of new documents published in their areas of specialization. As research on SDI progressed, IF appeared to be the area of focus. SDI used user profiles to match the keywords with respect to new articles for predicting new articles which were most relevant to the scientist's interest.

An IF-based technique, called Collaborative Filtering (CF) has become another area of research. To this effect, Jonatan, et al. (2000) pointed out that Automatic Collaborative Filtering (ACF) systems predict a user's affinity for items or information. These systems were basically designed to match the interest of one reader based on the interests of other groups of readers with similar interests.

CF helps people make choices based on the opinions of other people (Resnik et al., 1994). As an instance, a system for IF of Netnews, called GroupLens, helps people find articles they will like in the huge stream of available articles.

Filtering documents helps people looking for documents in many respects. Palme (1998) explained that filtering is designed to help people find the most valuable information, so that the limited time spent on reading/listening/viewing can be spent on the most interesting and valuable documents. Filters are also used to organize and structure information.

Amharic is spoken in most parts of Ethiopia. There are several newspapers, magazines and other written documents in the language. The variations among the language's usage such as lexical variations are common (Bender et al., 1976). Amharic dictionaries written by Kesate-Birhane Tessema (1951) and by Desta Teklewolde (1970) show these variations of the Amharic words with their different meanings in different parts of the country.

The number of newspapers published in Amharic is increasing over time. From such a large amount of news items, readers need automatic ways to filter out those items that are relevant to them. However, there is no such effective automatic way of filtering news items written in Amharic. Therefore, research needs to be underway concerning filtering out these items from a collection, based on a given profile. This study is one such attempt.

## 1.2 STATEMENT OF THE PROBLEM

Newspapers play a crucial role in informing people about current events. Many of the newspapers published in Amharic are on weekly basis and owned by private agencies. As a number of such newspapers are published, it is evident that a lot of textual information is accumulated, which makes it difficult to handle them manually. A news reader who is interested only in 'Sport' category may want to read only such items integrated from more than one source published with similar topics. Hence, filtering systems must be established so as to accomplish this task.

Amharic newspapers are dispatched to the public manually through distributing agents. Every reader needs to look into all the pages in order to find out what he/she would like to read. However, as many of such newspapers are published, the reader wants to be more selective and specific in the topics that are of interest.

In this regard, the issue of filtering news items from a large collection of news becomes paramount. Before posted, the news must be edited for errors and inconsistencies. Whatmore (1978) stated that every news is processed (shaped and sharpened) by editors and assistant editors before it is dispatched to reach the public. It implies that not all news collected is fit for publication. There are redundancies, errors and unwanted stories that are not interesting to the readers. With the increasing number of news published a system for retrieving and/or filtering out only the items that are of interest to the reader should be devised.

Newspapers written in Amharic should then be filtered in such a way that they can somehow satisfy the reader who is in need of particular news items. There are, of course, many parameters that should be considered before the filtering process is on track. Foltz & Dumais (1992) argued in this respect, while automatic filtering of information sounds like a wonderful vision, there are many difficulties in determining what information a person would actually want to see.

Experimental efforts reveal that it is possible to predict the preference of a reader based on the preferences of other readers. One such technique called CF, has become a dominant area of research. ACF systems predict a user's affinity for items or

## 1.3 JUSTIFICATION OF THE STUDY

Research on filtering out relevant documents from large databases has undergone notable progress. There is a wide range of topics included in these collections of databases. Among such topics, a retrieval system must be able to handle the information stored so as to satisfy its users.

Finding and filtering information are the challenges in the area of digital libraries (Adam & Yesha, 1997). Users, however, are often overloaded with large amounts of information that are irrelevant. Preventing users receiving an overload of irrelevant information is a major problem in IF (Belkin & Croft, 1992). Hence, readers should be able to find precise and useful information.

Some reasons that justify the importance of news filtering for Amharic are listed below:

- News currently published in Amharic do not use automated systems for news filtering purpose,
- The choice of readers, such as selection among categories of newspapers, makes the reader's choice specific and helps to find out what the user is in need of timely. Consequently, the search time is reduced and the item of interest can be found easily,
- Large number of news have been published and documented manually at each publishing center of the agencies. Hence, searching for a news item for reading is difficult,
- A user may easily find the news item of interest without going through all of a collection,
- The potential of ANN has been explored by many researchers in filtering items, it was felt that its application to Amharic had to be seen,
- Part of this research is concerned with classifying news items to predefined categories such as 'Sport'. News classification and machine learning techniques could be applied to make the news items ready for information filtering purpose.
- The techniques used in this research could be expanded to recommending online products on the basis of collaborative filtering.

## 1.4 OBJECTIVES OF THE STUDY

### 1.4.1 General Objective

The general objective of the study was to explore the potential application of ANNs to filtering Amharic news using CF techniques.

### 1.4.2 Specific Objectives

This research tried to meet the following specific objectives:

- Review literatures related to Information filtering and Artificial Neural Networks,
- Select sample news items,
- Select potential readers of Amharic news,
- Develop a neural network model for predicting users' preferences of reading Amharic news,
- Develop a neural network model for classifying Amharic news in predefined categories,
- Test the model developed for accuracy,
- Draw conclusions and recommend future research areas.

## 1.5 METHODS

The following methods were employed in order to achieve the above stated objectives:

### 1.5.1 Literature Review

A literature review was carried out to get further understanding of information filtering, artificial neural networks, and the Amharic writing system. A neural network modeling technique was used for modeling users' preferences of news items and therefore enhances the filtering techniques for the sample selected. For this purpose, reference was made particularly to books, journal articles and available materials on the Internet.

The news items[1] in the preference set were indexed using an application developed by Theodros Hailemichael (2003). Matlab Neural Network Toolbox[2] was used to produce the term-by-document matrix of the indexed news items so as to train the network to predict the category to which an item belongs. The training and test sets of the news items were selected from the indexed matrix of term-by-document.

### 1.5.5 Testing Techniques

The testing technique was implemented in two phases. First, the news items likely to be chosen by readers were used as target groups for the network model. The model suggested the extent to which the actual output and the network output agree. Next, the term-by-document matrix of the news items in the sample was tested using 50 of the matrix fields out of the 150 news items by 493 unique terms identified.

The network's performance was measured to what extent it could correctly classify the categories under consideration. Using the SOM model, the percentage of correct assignments was used to decide its effectiveness.

[1] The sample news items are included in Annex 1.
[2] Matlab 7.0 is a software developed by Mathwork Inc.

## 1.6 SCOPE AND LIMITATIONS OF THE STUDY

The news filtering process in this study is limited to two newspapers that are published weekly for the very reason that they comprise common categories. However, there are daily Amharic newspapers like Addis Zemen[3] that are worth considering. The study entertains few readers of the newspaper. A larger number of readers need to be included so as to attain a better result.

Most Amharic newspapers are not dispatched on the Internet. As a result, it was not possible to collect fast rating information of news items from news readers. Hence, manual methods were used to collect users' preferences.

Another major problem is that the researcher couldn't find a full-fledged stemmer for the language. In addition to these, due to time constraint, it was not possible to integrate the readers' preference results and the classified items of the news in each category of the newspapers. The news titles selected for the study were based on only the unique keywords identified from the news titles. However, to achieve a better result, the contents of all the news could be considered, which hopefully gives a better result.

[3] *Addis Zemen* is a government owned daily newspaper.

## 1.7 ORGANIZATION OF THE THESIS

This thesis is organized into five chapters. The first chapter describes the area of the research. It also lists the statement of the problem, the general and specific objectives, and the methods employed in the study.

Chapter two is concerned with the discussion of the available IF techniques used by other researchers. Since the focus of this study is on filtering Amharic news, the language's properties and its writing systems were discussed in line with the discussions in chapter two.

The third chapter looks into the notion of ANNs with an emphasis on the Backpropagation (BP) and the SOM algorithms as they were used in the study.

In chapter four, the actual experiment phase is seen in detail. The experimentation and results of the study were discussed.

The conclusions drawn and the recommendations for future research areas were pointed out in chapter five.

# CHAPTER TWO

# INFORMATION FILTERING

## 2.1 INTRODUCTION

The amount of information produced, primarily in electronic form, basically exceeds the amount of information a person is in need of. Hence, getting the right information at the right time is relatively difficult. IR research on the basis of keyword matching and other techniques have come to be used extensively to extract relevant items from large collection.

Together with the notion of IR, IF plays a major role in recommending documents that were read by others.

Folth, et al. (1992) and Oard (1997) pointed out that there have been several studies and systems developed to test the abilities of information filtering. The experiments carried out by Allen (1990), cited by them, showed that through the analysis of what texts are read, and what is in the text, a user model is developed. Models developed in this process create a structure to represent the information based on user preferences.

From looking at several models, predictions were successful for user preferences of general categories of articles, but not very good for predicting preferences of specific articles (Jason, 2000). Even though it is possible to develop different models, there are many parameters that should be taken into account in order to come up with a good model.

## 2.2 INFORMATION FILTERING

IF is concerned with extracting only wanted documents from large databases. Filtering documents like news items and filtering commercial products like computers in electronic form is becoming a common day to day activity based on a profile of choice. It is customary to define what filtering is.

Oard (1997) defined Automatic IF systems as

> "...systems with the goal of automatically sorting through large volumes of dynamically generated information to satisfy a relatively stable and specific information need, as opposed to IR systems dealing with relatively static databases and short-term information needs."

Retrieving documents, as described by Baudisch (1997) from document streams, based on the information needs of users long term information interests are represented by the so-called profiles. The filtering activity involves the removal of those documents that are not worth to the user in a dynamic flow of information. The goal of IF systems is to keep users from being swamped with information. Filtering systems are devised to remove those items from a large information store that are considered to be non-relevant to users. Hence items that are relevant to the user are returned.

Comparisons between IF (or Selective Dissemination of Information, SDI) and information retrieval (IR) were given by Belkin et al., (1992) and Parker et al., (1979). Among others, IF systems have been applied to personal mail and Usenet news. The comparison between IR and IF is discussed in section 2.3.

Belkin et al., (1992) described filtering in many contexts though the above definition seems general.

- delivering information to users, most likely unstructured data not from a controlled imaged or optimized database, primarily textual information in large amounts,
- it is based on descriptions (types) and profiles (of users) as algorithms or tables,
- It consists of constantly incoming information,
- It compares a history of queries.

In other words, IF often involves extracting information in text form, though other forms can also be extracted. It works based on an existing users' profile. The information in an IF system is dynamic, in that there is a continuous flow of information. Moreover, as

filtering is carried out based on previous profile, it uses the history of queries supplied or used by the users.

Malone (1997) stated three forms of information filtering: cognitive (or content), economic, and social. Content-based filtering (CBF) is dominant in IR (Folth and Dumais, 1992), typified by profiles based on keywords, and economic filtering will become increasingly important as digital cash, micro-payments, and secure payment technologies emerge from research laboratories onto the Internet.

Kai Yu et al. (2003) classified IF as recommender systems which assist users to find their favorite products like movies, CDs or books; image retrieval, where the goal is to locate images that match a given query concept and automatic news filtering based on news reading habits of a specific user.

Major IF research undergone in the TREC (Text Retrieval Conference) and the ACM (Association of Computing Machine) reveal the development of IF researches.The filtering experiment undergone by Hongbo Xu et al (2004) indicates that each profile vector represents a user's interest; After retrieving more and more relevant or non-relevant documents, one can get more and more useful information about the user's interest, which can help us adapt the profile. The adaptation, according to them, includes positive adaptation, negative adaptation and adaptation based on undetermined documents.

## 2.3 INFORMATION RETRIEVAL vs INFORMATION FILTERING

IF is a special type of IR. Their similar feature as stated by Salton et al. (1983) and van Rijsbergen (1979), lies in that traditional IR and IF are related in that they both have the goal of retrieving information relevant to what a user wants, while minimizing the amount of irrelevant information retrieved.

There are three primary differences between IR and IF: First, user preferences (profiles) in IF typically represent long term interest, while queries in IR tend to represent a short term interest that can be satisfied by performing the retrieval (Belkin et al., 1992).

Second, IF is typically applied to streams of incoming data, while in IR, changes to the database do not occur often and retrieval is not limited to only the new items in the database. Finally, a distinction can be made between the two, in that filtering involves the process of 'removing' (i.e. removing documents that were given little attention by users even if they are relevant) information from a stream while IR involves the process of 'finding' information in that stream (Ibid).

In both IR and IF, a textual database can be represented by a word-by-document matrix whose entries represent the frequency of occurrence of a word in a document (Folth et al., 1992). The matrix produced represents the occurrence of each unique term along with the document labels that are to be indexed. Thus, documents can be thought as vectors in a multidimensional space, the dimensions of which are the words used to represent the documents.

In a standard 'key word' matching vector system (Salton et al., 1983), the similarity between two documents is computed as the inner product or cosine of the corresponding two columns of the word-by-document matrix. Queries can also be represented as vectors of words and thus compared against all document columns with the best matches being returned (Ibid).

## 2.4 LATENT SEMANTIC INDEXING

An important assumption in this vector space model is that the words (i.e. dimensions of the space) are orthogonal or independent (Folth, 1990). The assumption that words posses some 'latent' structure with respect to the items sought is worth seen, in that there are term associations and domain semantics. This method is called Latent Semantic Indexing (LSI).

A technique known as Singular Value Decomposition (SVD) is used to create this concept space. First, a term by document matrix $A$ is generated. (Jason, 2000). In such representation, every term is represented by a row in matrix $A$, and every document is represented by a column. An individual entry in $A$, $a_{ij}$, represents the frequency of the term $i$ in document $j$ (Ibid).

Researchers were looking for a novel approach that had representations whose power could be adjusted to the specific collection, had explicit representations of terms and documents in order to simplify retrieval, and was computationally tractable (Jason, 2000). LSI was then used to address these problems.

SVD is a technique closely related to eigen vector decomposition and factor analysis. (Dearwester et al., 1990). LSI is a statistical IR method designed to overcome two common problems in information retrieval, synonymy and polysemy (Ibid).

## 2.5 RECOMMENDER SYSTEMS

Personalized recommender systems, which recommend specific products (e.g., books, movies) to individuals, have become very prevalent. Recommender systems help overcome information overload by providing personalized suggestions based on a history of a user's likes or dislikes (Melville et al. 2001). On-line databases such as Amazon.com provide recommending services in that a product is recommended based on other choices of previous customers.

Melville et al. (2001) identified common approaches to building recommender systems: CF and Content Based (CB) recommending. CF systems, as described by them, work by collecting user feedback in the form of ratings for items in a given domain and exploit similarities and differences among profiles of several users in determining how to recommend an item. On the other hand, CB methods provide recommendations by comparing representations of content that interests the user.

## 2.6 FILTERING SYSTEMS

Research on intelligent information agents in general and recommendation systems in particular, has attracted much attention (Shardanand et al., 1995; Resnik et al., 1994). The alarming rate of increase in the amount of information produced has contributed to the emergence of researches on intelligent information agents, making it difficult to surf through today's information streams.

Personalized filtering as a research area could be implemented in different ways, where different machine learning algorithms can be used to learn a mapping from the features of an item to a number indicating the item to the user based on previous ratings that the user has made on other items. Most of the literature shows that these ratings are usually scaled from 1 to 5.

The following section discusses the different typed of filtering systems.

### 2.6.1 USER BASED FILTERING

User Based CF has been the most widely used technology for building recommendation systems these days, and is used in many commercial recommendation systems (Sonja, Kangas, 2002). The computational complexity of these methods increases linearly with the number of customers that sometimes in commercial applications can grow to large amounts as the systems are on the net and possibly the customers can be around the globe (Ibid).

#### 2.6.1.1 COLLABORATIVE FILTERING

CF is a method applied to predict the preference of a user based on the preferences of other readers. It compares one's likes or dislikes to those of other people to predict user preference. Based on the subjective evaluations of other readers, CF is a promising form of social filtering. Problems arising from information overload have given considerable attention to overcome the problems of search through a database. Users usually tend to choose some items of a collection to read.

Recommender systems based on ACF predict new items of interest for a user based on predictive relationship discovered between that user and other participants of a community (O'Connor, M. et al., 1999). Most of the successful research and most commercial systems in CF use a nearest-neighbor model for generating predictions. ACF systems based on the nearest-neighbor method work in three simple phases, as discussed by (ibid):

- Users of an ACF system rate items that they have previously experienced.
- The ACF system matches the user with other participants of the system who have similar rating patterns (i.e., they have similar opinions on experienced

items). This is usually done through statistical correlation. The closest matches are selected, becoming known as neighbors of the user, or collectively as the neighborhood.

- Items that neighbors have experienced and rated highly, but which the user has not yet experienced, will be recommended to the user and the consistency of opinion within the neighborhood.

CF aims at learning predictive models of user preferences, interests or behavior from community data, i.e. a database of available user preferences (Hofmann, 2003). ACF systems have been used in several research areas. Jonathan I. (2000) pointed out that ACF systems have been successful in research, with projects such as GroupLens (Konstan, J.A., et al. 1997; Resnick, P., et al. 1994) and Ringo (Shardanand, U. et al., 1995), to site few, are gaining large followings on the Internet.

CF is a complementary to content-based filtering and retrieval that is mainly built on the fundamental assumption that users are able to formulate queries that express their interests or information needs in terms of intrinsic features of the items sought.

CF is preferred to CB filtering in that

- It can perform in domains where there is not much content associated with items, or where the content is difficult for the computer to analyze ideas, opinions, etc.
- A CF system has the ability to provide serendipitous recommendations, i.e. it can recommend items that are relevant to the user, but do not contain content from the user's profile.

Because of these reasons, CF systems have been used fairly successfully to build recommender systems in various domains (Goldberg et al., 1992).

#### 2.6.1.2 SOCIAL FILTERING

Another form or often only a synonym for CF is social filtering. In the basic level social filtering often means the same as CF. Although the terminology differs between various

research groups, social filtering is often compared to item-based CF systems, it overcomes some of the limitations of CB filtering. (Alspector et al., 1997).

Natatalie Glance, (1997) described social filtering as matching items by first matching people to each other and that social filtering via ACF is based on the premise that information concerning personal relationships is not necessary.

In social filtering, documents are assigned to ratings, similar to CF systems. It can be compared to the stars which newspapers often assign to films, books and other consumer products (Alspector, et al. 1997). Shardanand, et al, (1995) described it as the process of 'word-of-mouth' recommendations because items are recommended to a user based upon values assigned by other people with similar taste.

Social filtering is similar to that of the filtering done by editors, journalists and publishers in that in both situations it is the task of humans to filter the information needed; humans are more capable of really deciding the value of a document (Alspector, J. A. et al. 1997).

Ratings for use in social filtering can be provided by:

- Editors, special people with the task of doing such rating. An example is the people selecting which messages to put into services like Yahoo,
- *Readers,* ordinary readers might input ratings on what they read, and these ratings might be collected and put into databases to help other people. Firefly and Grouplens (Resnick, et al 1994) are systems based on this method.
- *Authors* can provide certain kinds of ratings themselves. The advantage is that authors may be more willing to produce ratings; a disadvantage may be that an author might give too high ratings to his/her own documents. Because of this, author ratings are mostly useful if objective scales are used. (Ibid)

#### 2.6.1.3 CLIQUE-BASED FILTERING

Clique-based filtering another term for CF and often used as a synonym for CF. In the clique-based filtering approach, the interests of the user is determined by similar minded people. The assumption behind this approach is that users who feel similarly about

previous items will feel similarly about new items. (Alspector, et al. 1997). Clique-based filtering may work based on a clustering algorithm, a correlation-based approach, vector-based similarity technique, or according to a Bayesian network (Fink, et al., 2000).

#### 2.6.1.4 ADAPTIVE FILTERING

Adaptive filtering is a kind of a combination of user-based and item-based filtering. The idea is that the system adapts itself through learning, by asking the user to rate things and by monitoring the click stream to watch what the user does (Alspector, et al. 1997).

Promising approaches to this problem are those where the user community itself provides the needed structure through their preferences and actions (Åsa Rudström et al., 1997). This structure can be supplied directly by the user (filling in keywords, setting rules in e.g. email filters etc) (Ibid).

Among the many researches, the works of Kevyn et al (2004) used the Roccio algorithm for profile updating.

### 2.6.2 ITEM-BASED FILTERING

The other perspective to filtering an item is item-based CF recommendation techniques which have been developed to analyze the user-item matrix to identify relations between the different items, and use these relations to compute the list of recommendations. (Sonja, 2002). Unlike the user-based CF, discussed in section 2.6.1., item-based approach looks into the set of items the target user has rated and computes how similar they are to the target item. (Badrul, et al., 2001). Search between the similarities or differences between items are done and then the evaluations are connected to users or a group of users (Sonja, 2002).

#### 2.6.2.1 FEATURE-BASED FILTERING

Feature-based approach is based on the idea that it is possible to capture the features a user likes and does not like about an item and thus provide feedback about various items to the user (Sonja, 2002). The filtering system would understand the features of an item. The limitation is that often the features can only be retrieved from text data (Ibid).

### 2.6.2.2 CONTENT-BASED FILTERING

In CB filtering (also called cognitive filtering), document representations can exploit only the piece of information that can be derived from document contents (Sonja, 2002). But as described above, the content of the document is taken into consideration unlike other user based filtering techniques.

CB approaches, according to Pazzani (2000), involve describing items that are rated for the purpose of learning relationships between the ratings and the rated items to learn a relationship between the ratings of a single user and the description of the items rated by others (Sonja , 2002).

### 2.6.2.3 KEYWORD-BASED FILTERING

Keyword-based filtering is one type of CB filtering. It is based on the fact that keywords express the concepts to filtered (Sonja, 2002). However, there can be cases where it is difficult to retrieve using keywords. User based CF systems are preferred for this particular case.

### 2.6.2.4 PROFILE FILTERING

Profile filtering is the most straightforward approach. Users either describe their interests by picking words from list of profile or entering keywords, and a software developed for such filtering purpose rejects anything that doesn't match with the user's choice (Sonja, 2002).

## 2.7 AMHARIC WRITING SYSTEM

In a system where there is large number of documents, retrieval of a given document can be possible if the collection is organized systematically (Zelalem Sintayehu, 1998). One aspect of retrieving documents is that among the collection of documents, the issue of filtering out the right document that can satisfy the user in search of few documents.

A number of news is published on daily and weekly basis in Amharic. These newspapers are distributed to the public manually. There is no automatic way of reading them though attempts are progressing to view some of them on the Internet. There are

also few newspapers published in English. However, few Amharic newspapers are considered in this study.

The Amharic language faces some problems regarding its script that must be considered in the development of automatic text processing systems (Zelalem Sintayehu, 1998).

The Amharic writing system consists of a core of 33 characters each of which occurs in a basic form and in seven other forms known as orders. Each graphic symbol represents a consonant together with its vowel. The vocalic symbol cannot be detached from the consonant element. That is, Amharic does not use independent symbols for vowels (Zelalem Sintayehu, 1998). Lesalu (1965) argued that the Amharic script is a syllabic rather than an alphabet.

The seven orders represent the different forms of a consonant. Each form is made in accordance with the sound that goes with the symbol. The non-basic forms are derived from the basic forms by more or less regular modifications. The orders for the two first alphabets are given below.

| ሀ<br>hä | ሁ<br>hu | ሂ<br>hi | ሃ<br>ha | ሄ<br>he | ህ<br>h | ሆ<br>ho |
|---|---|---|---|---|---|---|
| ለ<br>lä | ሉ<br>lu | ሊ<br>li | ላ<br>la | ሌ<br>le | ል<br>l | ሎ<br>lo |

Table 2.1 *Orders of the first two Amharic alphabets*

There are in total 33*7 basic alphabets. However, the alphabets are not restricted to this list only. There are the so-called labiovelars with five orders and eighteen labialized consonants. There are a total of 290 letters. Amharic also has its own numbers though not widely used. A list of Amharic character set called Fidel (ፊደል) is given in Annex III.

Amharic borrowed most of its scripts from Ge'ez. However it did not select from Ge'ez alphabet those symbols that are only necessary for its consonants. As a result there are

certain phonemes with different symbols, where they have meaning in Ge'ez, but their meaning is not known in Amharic.

There are problems related to orders that are used interchangeably such as using "ሀ" or "ኀ" in the name "ሀይሉ". Amharic word abbreviations, the problem of prefixes and suffixes do also contribute to the problem of automating the language. Details of these issues are discussed in Zelalem Sintayehu (1998) and Theodros Hailemichael (2003).

# CHAPTER THREE

# ARTIFICIAL NEURAL NETWORKS

## 3.1 INTRODUCTION

The human brain consists of billions of interconnected neurons. These are cells which have specialized membranes which allow the transmission of signals to neighbouring neurons (Fausett, 1994). ANNs are representations of these biological neurons for solving complex real world problems. As the brain learns through experience, ANNs learn from many examples supplied to them.

There are different types of ANNs that are found to be good in solving problems in medicine, speech and pattern recognition, astronomy, and the like. (Siganos, et al., 1999) stated: An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process and that in the process of learning, the synaptic connections in biological systems involves adjustments between the neurons.

When trained with large number of examples, an ANN can perform the task of humans just like we reason and respond to the environment. For this purpose, data are supplied for training it; and another data for testing its performance. During the process of learning from examples, the network uses the inputs given to it, with the help of a learning rule and activation function, results in an output.

## 3.2 THE BIOLOGICAL NEURON

The human brain consists of billions of neurons with complex interconnections. These neurons have the capability to process the information accepted from other similar neurons. A biological neuron accepts inputs through its dendrites. Signal passes through the neucles and axon, and then to another neuron.

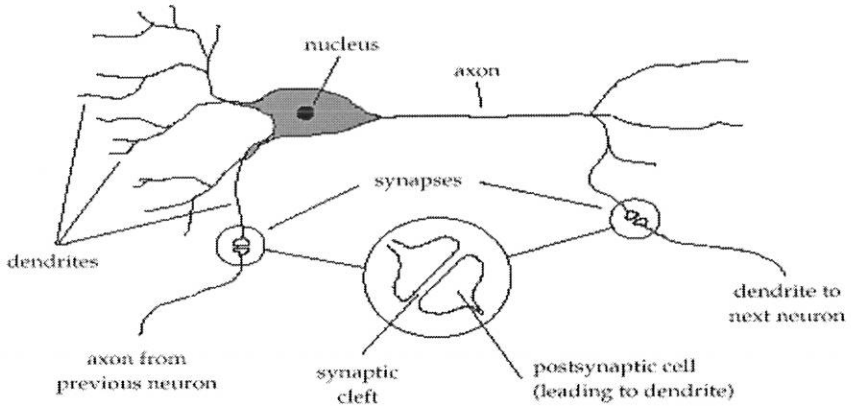



Figure 3.1.The structure of a biological neuron (*Adapted from* Coryn A.L. et al., 2001)

The process of information processing within the neuron, as described by Fausett (1994), proceedes as follows: Dendrites accept the information sent from other neurons (left), The input stimuli is accepted and then the cell sends an output signal along the axon to the synapses. Finally the signal is sent to other neurons (right).

Each of these neurons accumulates the input it receives, producing an output according to an internal activation function (discussed in section 3.4) (McCelland, 1986).

Unlike earlier programmed computing where a series of instructions are defined, ANNs learn from large examples (Simon, 1993).

## 3.3 AN ARTIFICIAL NEURON

An artificial neuron, as many of the literature show, resembles a biological neuron. It accepts many different signals, $x_i$, from many neighboring neurons and processes them in a pre-defined simple way (Simon, 1994).

Fausett, (1994) gave the following assumptions between natural neurons and ANNs:

- Information processing occurs at many simple elements called neurons.
- Signals are passed between neurons over connection links.
- Each connection link has an associated weight, which, in a typical neural net, multiplies the signal transmitted.

- Each neuron applies an activation function (usually nonlinear) to its net input (sum of weighted input signals) to determine its output signal.

Depending on the outcome of the processing, a neuron decides either to fire an output signal or not (Simon, 1994). The output signal, when triggered can be either 0 or 1, or may assume any value between 0 and 1. A number of such neurons can then be connected together for fast processing. ANNs combine artificial neurons in order to process information (Gershenson, 1999).

According to Simon (1994), there are three basic elements of the neuron model. These are:

- A set of *synapses* of connecting links, each of which is characterized by a weight of strength of its own,
- An *adder* for summing the input signals weighted by the respective synapse of the neuron, the operations described here constitute a linear combiner.
- An *activation function* for limiting the amplitude of the output of the neuron. Typically the normalized amplitude range of the output of a neuron is written as a closed unit interval [0, 1] or alternatively [-1, 1].

A threshold function can be externally applied in order to lower the effect of the net input of the activation function.

The simple model of an ANN is shown below.

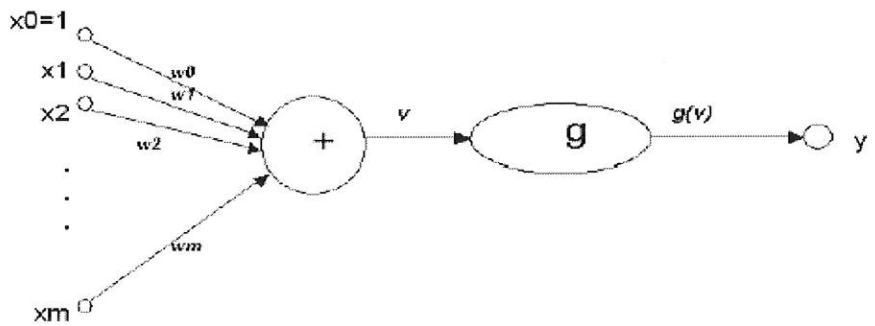



Figure 3.2 *An artificial neuron model (Adapted from Coryn A.L. et al., 2001)*

where $x_1, x_2 \ldots\ldots x_n$ is the input vector with corresponding $w_1, w_2, \ldots . w_n$ weight vector.

The summing function, shown by the + sign, adds the dot product of the inputs $x_i$ and the corresponding weights $w_i$, The result of the sum is shown as v in figure 3.2 and passed to an activation function indicated by g(v). Eventually, it outputs the result of the processing, which in this case is y. Essentially, v is the dot product of the input and the associated weights given by $\Sigma x_i w_i$.

## 3.4 ACTIVATION FUNCTIONS

An ANN has an internal state, called its activation or activity level. An activation function accepts the inputs as an argument. Typically, a neuron can send, through its dendrites signals to many other neurons simultaneously.

The activation rule, a local action that each node simultaneously carries out in updating its activation level following input from neighboring nodes. It is important to note that massive parallelism is involved as activation spreads through the network (Doszkocs, 1990).

To determine the output of neuron y in figure 3.2 above, we sum all the inputs combined with their respective weights:

$$y = b + w_1x_1 + w_2x_2 + \ldots\ldots + w_nx_n$$

The bias *b* serves to allow one to change bias of the output independently of the inputs. We then apply an activation function to this sum (Simon, 1994).

Common activation functions are the threshold and the logistic sigmoid function (an S-shaped curve) shown in figure 3.3. The most popular transfer function among neural network users, the sigmoid acts as a "squasher", compressing the input function when it gets takes on large positive or large negative values. Large positive values asymptotically approach 1, while large negative values are squashed to 0. The sigmoid function is given by:

$$y = f(y_{in}) = \frac{1}{1 + \exp(-y_{in})}$$

where $y_{in}$ is the input argument.

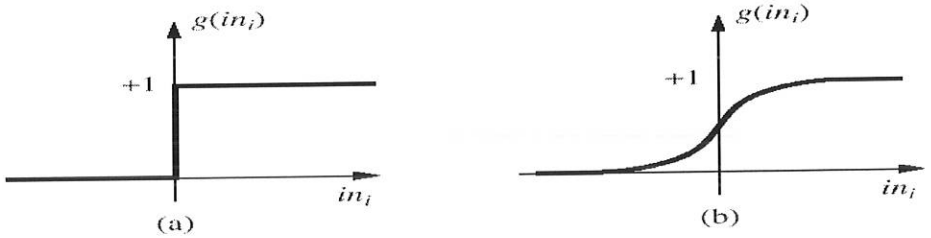



Figure 3.3 *(a) A threshold function (b) A sigmoid function; $in_i$ is the input argument to the activation function and $g(in_i)$ is the activation function. (Adapted from Zupan, 1994)*

According to kröse, et al. (1996), this activation function is a non-decreasing function of the total input; it is not restricted to non-decreasing function, generally a hard limiting threshold function (a sgn function), of a linear or semi-linear function, or a smoothly limiting threshold is used.

An ANN has two modes of operation; the training mode and the using mode (Siganos, D. et al., 1999). In the training mode, the neuron is given training examples in order to fire (or not) for the given example, whereas in the using mode, when other input pattern is supplied to the input, its associated output becomes the current output.

## 3.5 WHY USE ARTIFICIAL NEURAL NETWORKS?

Neural networks, with their remarkable ability to derive meaning from complicated or imprecise data, can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. A trained neural network can be thought of as an 'expert' in the category of information it has been given to analyze. It can respond to the input patterns given to it after training. This expert can then be used to provide projections given new situations of interest and answer 'what if' questions.

Other advantages include:

*Adaptive learning*: An ability to learn how to do tasks based on the data given for training or initial experience.

*Self-Organization*: An ANN can create its own organization or representation of the information it receives during learning time.

*Real Time Operation*: ANN computations may be carried out in parallel, and special hardware devices are being designed and manufactured which take advantage of this capability.

*Fault Tolerance via Redundant Information Coding*: Partial destruction of a network leads to the corresponding degradation of performance. However, some network capabilities may be retained even with major network damage.

The neuron presented in figure 3.2 doesn't do anything that conventional computers don't do already. A more sophisticated neuron (figure 3.2) is the McCulloch and Pitts model (MCP). The difference from the previous model is that the inputs are 'weighted'; the effect that each input has at decision making is dependent on the weight of the particular input. The weight of an input is a number which when multiplied with the input gives the weighted input. These weighted inputs are then added together and if they exceed a pre-set threshold value, the neuron fires. In any other case the neuron does not fire.

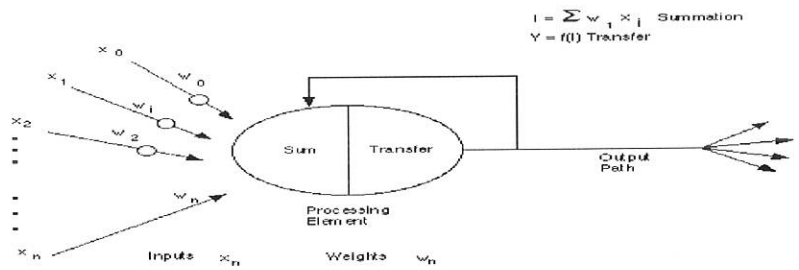



Figure 3.4 *An MCP neuron (Adapted from Zupan, 1994)*

## 3.6 PREPROCESSING DATA

### 3.6.1 Mean/Std Deviation Preprocessing

Mean standard deviation preprocessing is the most popular of the preprocessing methods. Each input is handled independently. The two calculated parameters are then stored as weights for the input layer of the network.

Then, for every iteration of the network the input is modified by subtracting the mean for that input and then dividing by the standard deviation for that input. The new input is then presented to the network in place of the original input. The formula is as follows.

$x_i' = (x_i - mean_i) / stddev_i$ for each input i.

where $x_i$ is the value to be input and $x_i'$ is the transformed value.

### 3.6.2 Max Min Preprocessing

This form of preprocessing calculates the maximum and minimum values for each of the inputs over the training set. The maximum and minimum for each input are stored as weights for the input layer of the network.

For every iteration of the network, the input is modified by the following formula.

$x_i' = (x_i - (max_i + min_i) / 2)) / (max_i - min_i)$.

### 3.6.3 Sum to 1 Normalization Preprocessing

This preprocessing function creates no weights. It is calculated directly from the inputs themselves. The result of the calculation is that the sum of all the modified inputs within a given input vector equals 1.0. This is accomplished by calculating the sum of the inputs and then dividing each input by the sum.

This kind of preprocessing is not effective if there can be large positive and negative inputs to the network. The negative inputs cancel the effect and the modified inputs still have a large dynamic range. This method does work well if the inputs are all positive.

### 3.6.4 Sum of Squares to 1 Preprocessing

This preprocessing function is similar to Sum to 1 Normalization in that it does not create weights and acts immediately upon input data. It is sometimes referred to as unit vector normalization.

Each of the inputs is divided by the square root of the sum of squares of the inputs. When the sum of squares of the modified inputs is taken, the sum will equal 1.0 exactly. Thus, the input vector becomes a unit vector (length of 1.0).

This form of preprocessing loses the magnitude information related to the input vector while preserving its direction or relation between inputs.

## 3.7 DESIGNING NEURAL NETWORK MODELS

The developer must go through a period of trial and error in the design decisions before coming up with a satisfactory design. The design issues in neural networks are complex and are the major concerns of system developers.

Designing a neural network consists of:

- Arranging neurons in various layers.
- Deciding the type of connections among neurons for different layers, as well as among the neurons within a layer.
- Deciding the way a neuron receives input and produces output.
- Determining the strength of connection within the network by allowing the network learns the appropriate values of connection weights by using a training data set.

The process of designing a neural network is an iterative process. Biologically, neural networks are constructed in a three dimensional way from microscopic components. These neurons seem capable of nearly unrestricted interconnections. This is not true in any man-made network. Artificial neural networks are the simple clustering of the primitive artificial neurons. This clustering occurs by creating layers, which are then connected to one another. How these layers connect may also vary. Basically, all artificial neural networks have a similar structure of topology. Some of the neurons

interface the real world to receive its inputs and other neurons provide the real world with the network's outputs. All the rest of the neurons are hidden form view.

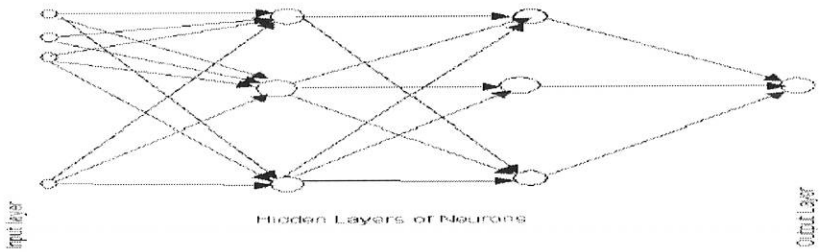



Figure 3.5 *A neural network with two hidden layers (Adapted from Zupan, 1994)*

As figure 3.5.shows, the neurons are grouped into layers The input layer consist of neurons that receive input form the external environment. The output layer consists of neurons that communicate the output of the system to the user or external environment. There are usually a number of hidden layers between these two layers; the figure above shows a simple structure with only one hidden layer.

When the input layer receives the input its neurons produce output, which becomes input to the other layers of the system. The process continues until a certain condition is satisfied or until the output layer is invoked and fires their output to the external environment.

To determine the number of hidden neurons the network should have to perform its best, one are often left out to the method trial and error. If you increase the hidden number of neurons too much you will get an over fit, that is the net will have problem to generalize. The training set of data will be memorized, making the network useless on new data sets.

## 3.8 LEARNING

Learning is a fundamental ability of neural networks. Supervised learning ('teaching by doing') as Doszkocs, Tamas E. (1990) explained, implies that the network is trained by presenting to it examples of input and desired output from a training set. The iteration process results in a neural network model where it can process new data given to it

when a test set is supplied to it. The desired or expected output is then obtained with the accuracy of the model. In the other way, an unsupervised learning ('self teaching') learning, the neural network can automatically detect regularities in input patterns and is able to group patterns with similar structure into the same category (Doszkocs, Tamas E., 1990). In an unsupervised learning, there is no expected or target output. The neurons in the hidden layers are supposed to learn the internal processing with no prior knowledge of the output. In most literature, unsupervised learning is also called 'learning by doing'.

Since this study uses the BPN and SOM models, it is discussed below. Most of the review was adopted from the user manuals of the neural network software.

## 3.9 BACK PROPAGATION (BPN) LEARNING RULE

Back propagation is probably the most widely recognized and most commonly used supervised-learning algorithm. As one of the first algorithms to be used effectively with neural networks, it is relatively unsophisticated by today's standards and considered outmoded and slow by some. Nevertheless, its success is due to a robust ability to achieve generalization, and it remains a useful standard by which others are compared.

*Learning Rate Parameter (Alpha):* This is the most important parameter. It scales the magnitude of weight adjustments and thus can dramatically affect the rate of learning.

Settings are typically between 0 and 1 and tend to be very small. Too large of a value can stall training. The default value is 0.01.

*Momentum Parameter (Beta):* This parameter can improve performance by adding inertia to the trajectory of the weights during learning. This has an averaging and smoothing effect. Too much momentum, however, can cause training to overshoot a goal. When in doubt, use of little or no momentum is recommended. Settings are constrained to 0 to 1 and tend to be close to 0 which is equivalent to no momentum.

*Weights Decay (WtsDecay)*: BPN is a gradient descent algorithm. Gradient descent operates much as the name implies: the algorithm adjusts network weights so that the overall network continually progresses downhill. A drawback of gradient descent technique algorithms is that the network may settle into a local minima.

Steepest descent proceeds by looking at the gradient of the error in the immediate vicinity of the current set of weights. This gives the learning algorithm information about how to perturb the weights to reduce the error. There are situations when the steepest descent approach becomes quite inefficient. This is because the direction of steepest descent is not always the best direction.

## 3.10 BACKPROPAGATION ALGORITHM FOR MULTIPLE HIDDEN LAYERS

The algorithm for training a network with multiple hidden layers is shown. The procedure is as follows:

*BACKPROPAGATION PROCEDURE:*

1. *Initialize the weights in the network to small random values*
2. *Pick a pattern x(l)*
3. *Propagate the input pattern forward to determine the network output O(l)*
4. *Compute deltas at the output layer*
5. *Compute deltas for preceding layers by propagating backwards until a delta has been computed for every unit*
6. *Compute the adjustment to weight*
7. *Go back to step 2 and pick a new input pattern.*

We have yet to specify when to stop the procedure. The procedure is only guaranteed to converge to a local minimum and usually requires many iterations to do so. The stochastic gradient descent procedure converges to a global minimum (for small enough), but there the cost function was a simple quadratic form in the weights (Klerfors, 1998).

## 3.11 SELF-ORGANIZING MAP (SOM) MODELS

Self-Organizing Map (SOM) is an unsupervised learning neural network architecture. A detail discussion of this network type (adapted from the user manual of Neuralware neural network software) is given below.

Self-Organizing Map (SOM) network architecture and general learning methodology was pioneered by Teuvo Kohonen. The SOM network architecture was originally developed to visualize topologies and hierarchical structures of higher dimensional input spaces.

The Kohonen layer, which is the heart of a SOM network, transforms any n-dimensional space into an ordered, z-dimensional map. When an input vector is passed to a trained SOM, the distance from each SOM Processing Element (PE) to the input vector is evaluated. The closest PE is declared the winner. Since the winning PE is the closest to the input vector in the input space, it represents the input on the Kohonen map. In this way, high dimensional input vectors can be visualized on a lower dimensional map. A SOM network can also be used to create area-filling curves.

A key difference between a SOM network and many other networks is that the SOM network learns without supervision, hence the term self-organizing. The SOM network is sometimes combined with other neural layers for categorization and prediction. In this case the SOM network begins training in an unsupervised mode and then requires supervised training for the output layer.

SOM networks are typically used for sorting items into appropriate categories of entities with similar features, a difficult yet fundamental and frequent data mining activity. The SOM neural network identifies categories by creating a (typically) two-dimensional feature map of the input data in such a way that order is preserved. In other words, if two input vectors are close in some multi-dimensional input space, they will be mapped to processing elements that are close together in the two-dimensional layer that represents the clusters (i.e. features) of the input data. The Kohonen layer in the SOM network is thought to act similar to biological systems, in that it preserves order, compacts the representation of sparse data, and disperses dense data throughout a (typically) two-dimensional region.

Neighbors of the winning PE are determined by the neighborhood shape (type) and size. Neighborhood shape depends on the distance measure selected; neighborhood size is the radius of the neighborhood.

# CHAPTER FOUR

# THE EXPERIMENT

## 4.1 INTRODUCTION

This chapter presents the techniques used in the experiment phase and the results thereof obtained in order to filter Amharic news items from predefined categories such as Sport and Art.

The experiment in general comprises the following steps:

- Indexing the news items based on their titles,
- Training the neural network model,
- Predict future user preferences based on previous user profiles.

When the system is fully implemented, after a user of the system retrieves the news items of his/her interest and completes using the search for the news items, the profile is used to predict the preferences of other readers.

## 4.2 THE PROCEDURE

The following procedure was used in the process of filtering Amharic news:

- News items published each week are fed to the proposed system in accordance to their categories,
- The neural network model is initialized so as to determine what news items are of interest to the user who come to the proposed system according to the previous profile,
- The user selects the category of his/her interest and the history is saved for later profile updating.

Figure 4.1 depicts the steps followed.

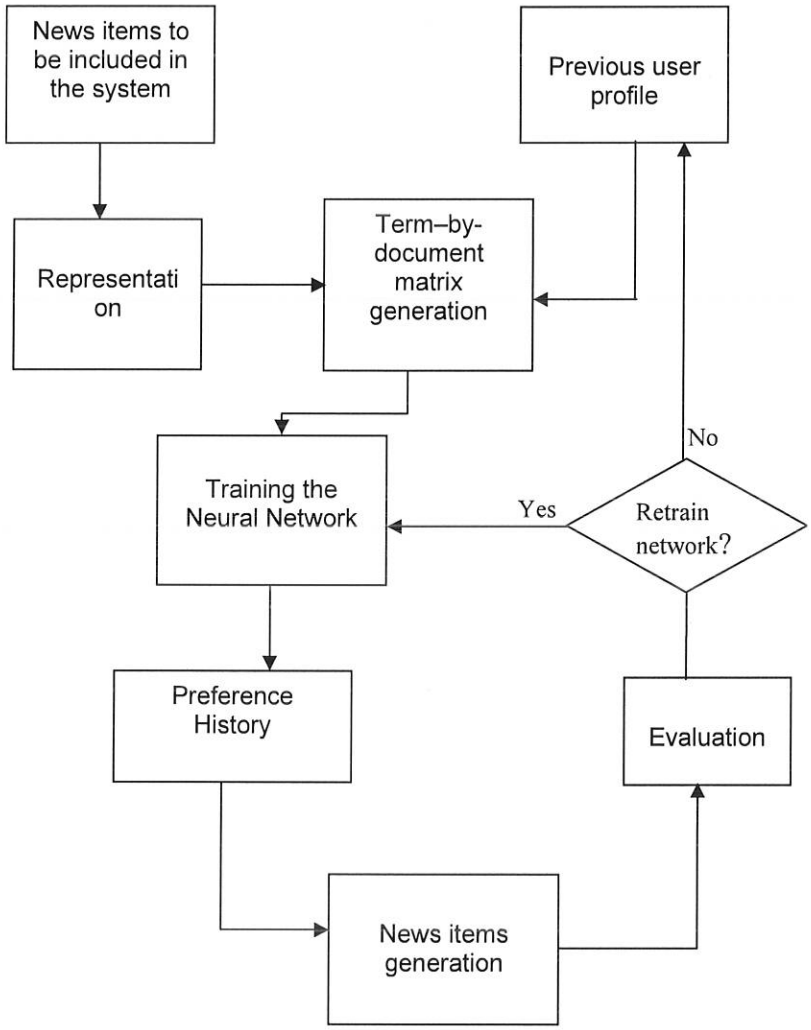



Figure 4.1 *The process of Filtering Amharic News*

## 4.3 IMPLEMENTATION

**Step 1: News items included in the system**

This procedure is mainly concerned with the addition of new news items that are to be included in the already existing database during publication. At the beginning of the experiment, the selection of the news items that are worth reading was done manually by experts since there is no such automatic way. Moreover, it was unlikely that all the items would be chosen by the readers at a given time.

**Step 2: Previous user profile**

User profiling can be approached in one of three ways, as described by Langley (2000) and cited by MacDonald (2001)

- Using stereotypes,
- Using surveys/questionnaires, or
- Using a learned model.

The first two approaches rely on traditional marketing methods using known information or information collected in person or over the phone to build appropriate profiles (ibid). However, the approach followed in this study was based on the last approach, using a neural network Model.

It involves creating a system that has no knowledge of users' choices of the categories under consideration or the contents of the news items at all. Users must interact with the system through a number of steps in order that it develops a profile model on the basis of collaborative interest (MacDonald, 2001).

The profile of the reader in the database refers to all the news items that were read at the moment a user comes to use the system. In other words, it contains those items that were visited before the most recent news items were included.

**Step 3: Representation**

The categories of the new news items were given the identification indexes as shown in table 4.1 below. The binary combination of bits then identifies each of the ten categories of the newspapers in the sample.

Eight newspapers were selected from a collection of 48 newspapers, each of them published within the time range of May, 2003 to April, 2004 (i.e. until the date this sample was considered). Hence, the collection embraced only a sample of one year of data. The common categories of the newspapers were identified as follows:

| *Nº* | *Category* | *Identifier* | *AmharicTranslation* |
|---|---|---|---|
| 1 | Hot News | 0001 | ዜና |
| 2 | Editorial | 0010 | ርዕሰ አንቀፅ |
| 3 | Politics | 0011 | ፖለቲካ |
| 4 | Business and Economy | 0100 | ቢዝነስና ኢኮኖሚ |
| 5 | Society | 0101 | ህብረተሰብ |
| 6 | Culture | 0110 | ባህል |
| 7 | Science and Technology | 0111 | ሳይንስና ቴክኖሎጂ |
| 8 | Health | 1000 | ጤና |
| 9 | Art | 1001 | ጥበብ |
| 10 | Sport | 1010 | ስፖርት |

Table 4.1 *Common categories of the newspapers in the sample*

**Step 4: Term-by-document matrix generation**

The news items in the sample were merged and preprocessed so as to produce the term-by-document matrix. The generation of this matrix was done by a preprocessor developed by Theodros Hailemichael (2003). It uses SVD technique to generate the matrix. A preview of the generated matrix is shown in figure 4.2 below.

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Hot News_1 | Hot News_2 | Hot News_3 | Hot News_4 | Hot News_5 | Hot News_6 | Hot News_7 | Hot News_8 | Hot News_9 |
| 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 11 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 12 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 13 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 |
| 14 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 15 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 16 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 17 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 19 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 20 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 21 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 22 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 23 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 24 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 25 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 26 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 27 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 28 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 29 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 30 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 31 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |

Figure 4.2 *Term-by-document matrix generated*

**Step 5: Training a Neural Network**

The generated term-by-document matrix is fed to the neural network so as to train it to predict the preferences of news readers in the collection as well as classifying the news items in the predefined categories.

**Step 6: Preference history**

The preferences of those news readers who recently chose to read the items are saved for the purpose of training the neural network, which predicts the news item likely to be chosen by the next user.

**Step 7: News items generation**

At this stage, the reader is interacting with the system. After the network is trained, the preferences of news items likely to satisfy the reader are generated using the developed neural network model.

**Step 8: Evaluation**

The next task is to study the extent to which the users are satisfied with the generated news items. Either the network is retrained to train the network again or the profile is included in the list of previous profile for later use.

## 4.4 SELECTION OF TRAINING AND TEST SETS

The experiment in this study was carried out in twofold, basically consisting of training the neural network using the backpropagation algorithm for predicting preferences, and applying the SOM algorithm for classifying news items in a predefined category.

### 4.4.1 Preference List

For the purpose of this experiment, only 15 of the news items were included in the experiment from each of the 10 category, so as to obtain an adequate number of inputs for training the neural net. A full list of the sample news is given in Annex I.

In the first phase, out of 100 preference lists, the training and test sets consisted of 50 and 50 preference lists, respectively. This set comprised the set of preference pairs of

active readers[4] collected through interview. The readers were interviewed to supply the category and item of choice that he/she chose to read and then the category and item of news he/she would likely to choose next was noted. The latter was considered as the target output for the network. This is synonymous to rating the choices of the readers as 1 (first choice) and 2 (second choice). For clarity, the preferences of the 6th and 71st readers are summarized below.

| | *Top choice of category and news item* | | *Next choice of category and news item* | |
|---|---|---|---|---|
| | Category | Item No | Category | Item No |
| User 6 | 0001 | 0101 | 1010 | 0001 |
| User 71 | 0011 | 0011 | 0101 | 0001 |

Table 4.2 *Preferences of user 6 and 71 in the sample*

### 4.4.2 CLASSIFICATION

The second set consisted of the term-by-document matrix generated from the matrix of the collection of news items incorporated in the sample. The news items rated were preprocessed so as to generate this matrix. At the beginning of the experiment, the 15 news items selected from each category were indexed. The matrix then contained 15*10=150 fields (since there were10 categories) with the item labels as NewsItem_1, NewsItem_2, ..., NewsItem_150 and the records (vectors) of the 493 unique words identified from all the items.

## 4.5 THE INDEXING PROCESS

For this experiment, the matrix of term frequencies of the selected news titles from each category was fed to an application developed by Theodros Hailemichael (2003). The application has the capability of accepting Amharic texts (particularly news) and clean it by removing repeated consonants such as "ሀ", "ሐ", "ነ", "ሓ", "ኀ", and "ሃ" with one of the single characters, removing stop words such as "እንደ", and producing a term-by-

[4] Active readers, literally speaking, are those readers who regularly read the newspapers. In this context, the terms "users" and "readers" are used interchangeably.

document matrix. Moreover, it can list the frequency of each unique term that appears in the collection of the news titles.

The news titles after being processed in this preprocessor were analyzed for correctness by use of inspection. Erroneous results were identified with repeated trial, manually tracing the words.

Before explaining how the network was trained, below is given the way the news items were represented.

The category identifier, one of the numerical values[5] given in table 4.1, is attached to the choice of a reader. Each title is given an identifier 1 through 15 according to the order in which the news items appeared during publication. The following table shows the preferences of two active readers included in the sample. The labels "Top choice of category and news item" and "Next choice of category and news item" in table 4.2 are referenced as Preference 1 and Preference 2 respectively.

| | Preference 1 | | Preference 2 | |
|---|---|---|---|---|
| User 3 | News title | Category (Identifier) | News title | Category (Identifier) |
| | 7. በኢ.ሲ.ኤ. አካባቢ ነዋሪዎች ቤታቸውን በአንድ ቀን እንዲያፈርሱ ተነገራቸው | Hot News (0001) | 1. ቤክሀም ትልቁ የእግር ኳስ አጀንዳ ሆኗል | Sport (1010) |
| User 17 | 2. ቢወዱት ንቆ የሞቱ ቢጠሉት ምን ሊሆን ነበር | Editorial (0010) | 11. የመድሀኒት ጥቅም ጉዳት እና አጠቃቀም | Health (1000) |

Table 4.3 *Preferences of two readers in the sample*

Hence, the first preference of user 3 in the sample was identified as

*Category_Identifier + NewsItem_Identifier = 00010111*

And preference 2 is denoted as *10100001*. Similarly, the preferences of user 17 were represented respectively as *00100010* and *10001011*. The data collected for the 100 readers of the newspapers were used for training the neural network to measure the

[5] The reason behind choosing a maximum of 10 is due to the fact that a) it is unlikely that a user may read more than these news items, and b) the network is supposed to be fed with the binary value 0001 to 1010 (the total number of identified categories).

extent to which the next reader's preference can be predicted. A full list is given in Annex IV.

In order to obtain good data for training the neural network, the news items in the experiment were carefully analyzed for errors. These were found to be both typographical and grammatical.

There are also cases where a single word is treated as one word or two separate words. For example, the words 'እንደተነገረው' and 'እንደ ተነገረው' may appear in different news titles. However, the former is treated as a single word while the latter involves two different words where the word "እንደ" is a stop word to be removed from the main text. Such inconsistencies may make the desired result erroneous. In addition to this, the preprocessed words in the matrix were seen carefully as they could affect the prediction result.

## 4.6 MODEL BUILDING

The model building process in this experiment involves creating and training a network that can predict user preferences and classify news items in predefined categories. But before discussing these major parts of the experiment, it is customary to discuss why the software used for this experiment was chosen.

### 4.6.1 MATLAB

MATLAB 7.0 was selected for this research work for many reasons. First, it was easily available for this experiment. Second, it has a number of toolboxes, one of which is the neural network toolbox that embraces many algorithms. MATLAB is a high-performance language for technical computing. It integrates computation, visualization, and programming in an easy-to-use environment where problems and solutions are expressed in familiar mathematical notations.

### 4.6.2 SELECTED ALGORITHM

The Backpropagation and Self Organizing Map (SOM) algorithms were selected for this research. The BP algorithm is the most popular one in use. The BPN was developed for predicting the preferences of news readers in the sample while the SOM network was developed in order to cluster the news items.

***Training function*:** There are many training function in the toolbox, out of which the trainlm (Levenberg Marquardt algorithm) was selected for it showed satisfactory results compared to other functions. However, traingdm, (Gradient descent with momentum) and traingdx (Gradient descent with momentum and adaptive learning rate) were also good during the training process. For the second set in the experiment, the trainr function with a neuron organization function hextop (hexagonal layer topology function) was used.

***Performance function*:** The default performance function mean square error (mse) was applied. The mean square error is the average squared error between the network outputs and the target output. During training the weights and biases of the network are interactively adjusted to minimize the network performance function (mse). The performance function is necessary in that it determines how well the neural network is doing its task.

***Activation Function:*** The log sigmoid activation function, whose range is between 0 and 1, was applied both for the hidden and the output layers.

## 4.7 INTERPRETATION

In the process of building the model, the final step is to analyze the simulated results so far attained through the training and test sets. The predicted numerical values in the first model were analyzed so as to determine the correctness of prediction with respect to the predicted values in the sample.

In the second model, the SOM network was trained multiple times as to choose a good model for simulating the test elements. This model was used to classify the news items in the test set and its accuracy was noted. The following section discusses the test result in the experiment.

## 4.8 TESTING

### 4.8.1 Preference List

As described in section 4.4, the preference set consisted of 100 preferences and trained using the BPN. A summary of this list (table 4.4) shows that out of the 100 readers of the newspapers, more readers were likely to choose categories 1, 2 and 3. The user supplied the category of choice twice, the news title chosen at the start and the news title the user to read next, comprising 200 choices of news items.

| Category of choice | 0001 | 0010 | 0011 | 0100 | 0101 | 0110 | 0111 | 1000 | 1001 | 1010 |
|---|---|---|---|---|---|---|---|---|---|---|
| Nº of choices | 27 | 32 | 25 | 23 | 17 | 15 | 10 | 15 | 16 | 20 |
| Percentage | 13.5% | 16% | 12.5% | 11.5% | 8.5% | 7.5% | 5% | 7.5% | 8% | 10% |

Table 4.4 *The number of preferences in each category for the sample*

The first 50 of the preferences were chosen for training the net while the remaining 50 were reserved for testing purposes. The testing was done using the collected preference lists. The backpropagation training function, trainlm, in the toolbox was used to train the feed-forward neural network in order to build a model for predicting the preferences of the readers. There are generally four steps in the training process:

- Assemble the training data
- Create the network object
- Train the network
- Simulate the network response to new inputs.

The data was first loaded onto the Matlab workspace in such a way that the field labels and the actual values of the matrix were assigned to variables, as [z p] where z was a 50-by-16 matrix containing the values and p was a 1-by-50 matrix containing the labels. Figure 4.3 shows the network model. The net was created using the command:

```
net=newff(minmax(z),[3,1],{'logsig','logsig'},'trainlm');
```

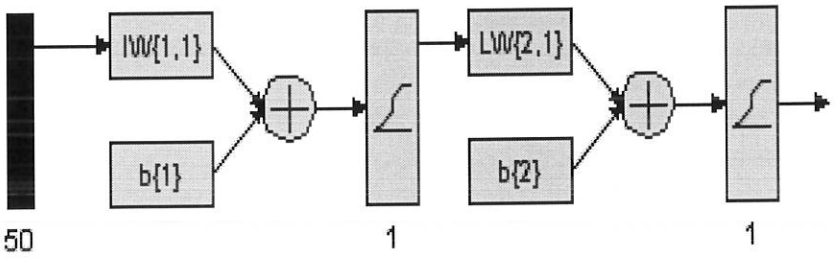



Figure 4.3 *The feed-forward network created*

It was trained for training epochs 100, 200, 300 ... 1000 and the error was observed to decrease during each step. However, the error tended to increase for more than about 950 epochs.

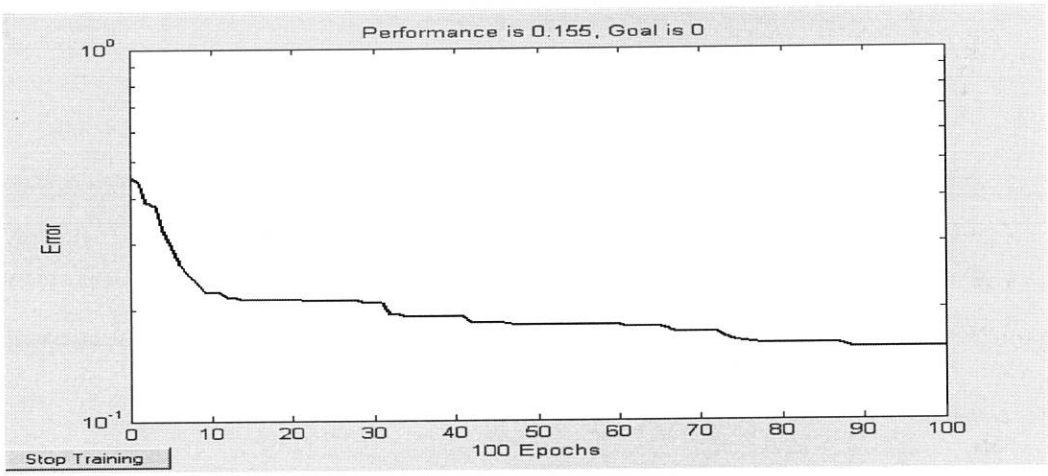



Figure 4.4 *The error graph for 100 epochs*

The simulated output was organized in order to count the correct predicted numbers and the result of correctly predicted preferences of the 100 preference lists was summarized in the following table.

| Class | # of Preferences | Accuracy (%) |
|---|---|---|
| Training set | 100 | 83.3 |
| Test set | 100 | 79.8 |

Table 4.5 *Accuracy of the sample in the preference list*

Table 4.5 shows that the model developed correctly predicted 83.3% of the preferences in the training set and 79.8% of the preferences in the test set. That is, a news item is likely to satisfy the readers in the test set 79.8% of the time.

### 4.8.2 Classification of News Items

In phase two of the experiment, the weighted term-by-document matrix generated from the news items was fed to the neural network for training it to classify the identified news items to a unique category. The original model was developed from the 100 labels of the news items, comprising 10 fields from each category. This model was used to simulate the remaining 50 news items consisting of 5 news items from each category.

The SOM model recognized the labels of the news items at the neuron positions in the diagram, and hence the items in the neighboring node of the model were collected for classification.

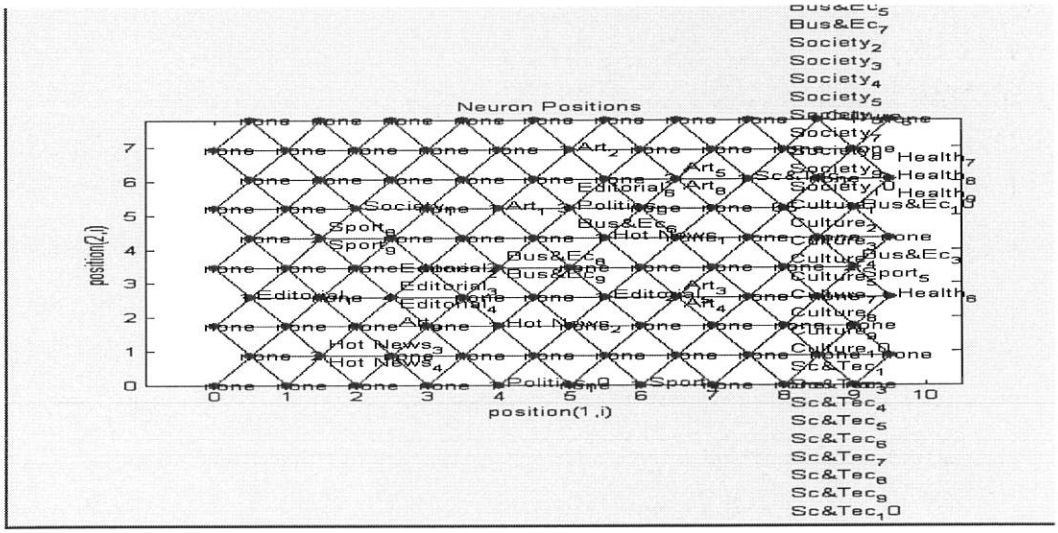



Figure 4.5 *The SOM diagram before training (Most of the News items seem to cluster at one neuron position)*

The labels of the news items seem to concentrate at one neuron position, which shows that the news items are not sparse for good classification. In order to be these labels to cluster at different neuron positions, training is required.

The SOM was trained for a maximum of 100 to 200 iterations at the first experimental trial. Since it clusters the items well as the number of iterations increased, the number of epochs was increased to 500 and proceeded at an interval of 500. Table 4.6 shows the procedure followed in training the SOM model.

***Input: Training Set Field***
***Output: SOM model***

- *Select the first 10 fields of the items from the total collection of term-by-document matrix from each category.*
- *Train the network.*
- *Cluster the categories at each neighboring node of the SOM model.*

Table 4.6 *The procedure for training the SOM model*

Once the model developed, the test set was simulated so as to classify the news items in the neighboring nodes of the model. Table 4.7 shows the simulation procedure. The function *LemTrain,* which was created using the Matlab code

```
Lemtrain=newsom(minmax(z),[10 10],'hextop');
```

trains the network based on the following procedure: The implementation of this classification code in Matlab is given in Annex II. The procedure follows the following steps:

***Input: Test Set Field***
***Output: SOM model***

- *From each category, select the last 5 fields of the news items from the total collection of term-by-document matrix*
- *Cluster the categories at each neighboring node of the SOM model*
- *Generate a SOM model*

Table 4.7 *The procedure for testing the SOM model*

The SOM could able to cluster the test set fields around the neurons at iteration steps, i.e. the model was simulated for unseen data.

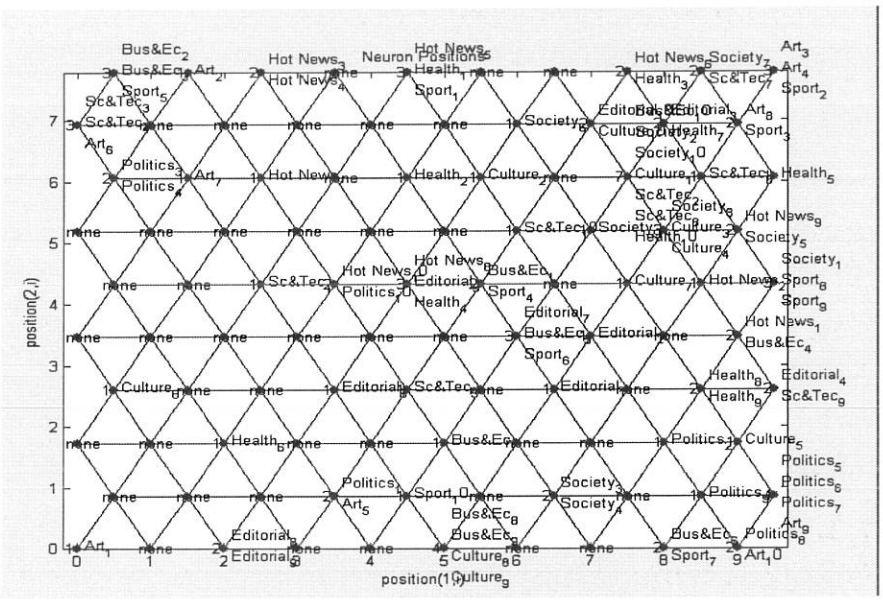



Figure 4.6 *One of the SOM models generated during training*

Each node of the figure was defined to cluster 6 of the neighboring nodes in its class of the surrounding nodes. A single node has a maximum of 6 neighboring nodes. The number of nodes in the class was then counted as to determine the percentage of correct classification.

| | *Epochs* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | *100* | *500* | *1000* | *1500* | *2000* | *2500* | *3000* | *3500* | *4000* |
| Training Set (66.7%) | 31 | 40.2 | 55 | 68 | 74 | 76.5 | 68 | 42 | 37.6 |
| Test Set (33.3%) | 25 | 39 | 43.7 | 62.4 | 73.1 | 72.9 | 41.6 | 29.5 | 31 |

Table 4.8 *Percentage of correctly classified items*

The result of correct classification of each category from the whole collection for the 100 and 500-4000 iterations with the range of 500 iteration steps is summarized in table 4.8

below. It is customary to represent the summary given in table 4.8 using the line graph presented in figure 4.7 below.

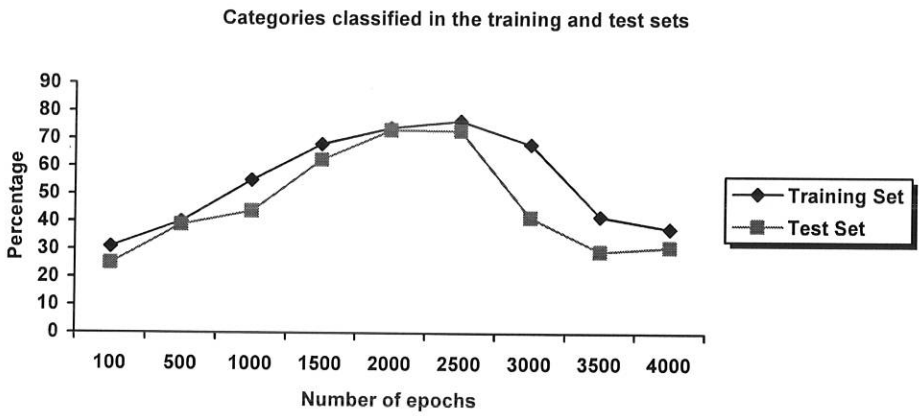



Figure 4.7 *Percentage of categories correctly classified*

It can be seen from Figure 4.7 that the predictive trend in classifying the vector of unique terms (in the news items) to a category achieved a better result at the 2500$^{th}$ epoch. This model was used to classify the news items in the test set. The result at this epoch is shown in table 4.9.

| **Correctly Classified at 2500 epoch** | |
|---|---|
| Training set | Test set |
| 76.5% | 72.9% |

Table 4.9 *Percentage of news items classified for 2500 iterations*

Table 4.9 indicates that the SOM model could correctly classify 76.5% of the news items in the total collection in the training set whereas 72.9% of them were assigned to the correct category in the test set.

## 4.9 SUGGESTED PROTOTYPE

This section proposes the development of an easy to use prototype for implementing Amharic news Filtering. The filtering process, as described before, is restricted to only two newspapers due to the following reasons.

Nevertheless, there are other newspapers published on daily and weekly basis that can be included in the suggested model, which thus will be more representative. Users of the system may approach it using an interface such as the following:

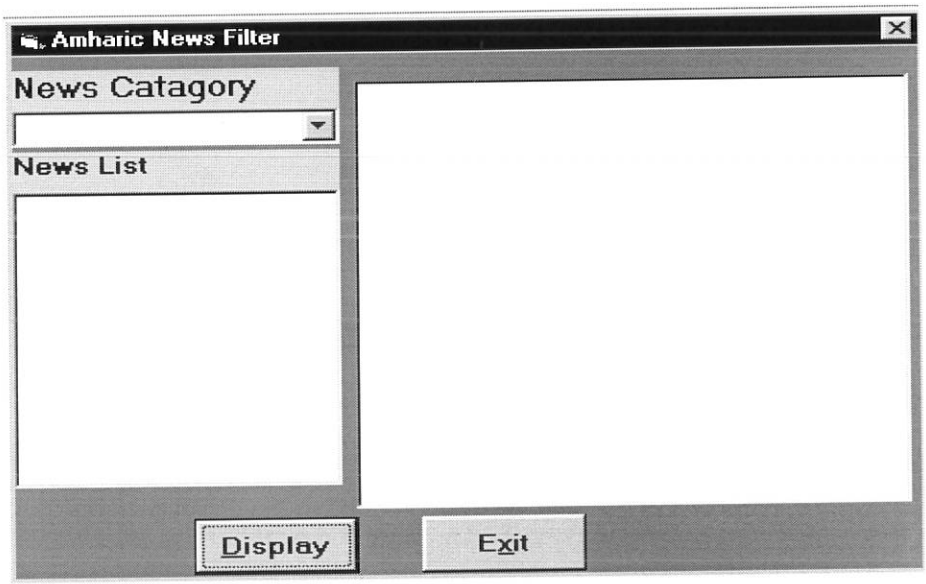



Figure 4.8 *The user interface*

A news reader selects the category of choice from the *News Category* combo box. After selection of a category is made, a list of news items is displayed based on the preferences of previous news readers. The reader then selects the news item of interest. A click on the Display button generates the content of the news item. Upon completing, the *Exit* button closes the dialogue box, saving the choice of the reader.

## 4.10 DISCUSSION

The test result for the preference list ensured that the feed-forward network was capable of producing a good model for predicting user preferences of news items. The result obtained from the prediction of the model shows that the predictive power of the ANN model is promising. The model was able to predict 79.8% of the preferences of the readers in the test set.

Moreover, the model used to classify the news items to each of the categories, the SOM model, could be able to classify the news items in each category with an accuracy of 72.9% in the test set. Hence it could be employed for classification purposes with this accuracy. However, increasing the training epochs may make the model perform better than this result.

In the course of the experiment, as stated before, very few of readers were incorporated. Most previous researches on IF used large datasets containing a number of user ratings. The EachMovie dataset, as an instance, is open for researchers and contains millions of ratings collected from users.

In the presence of such ratings supplied, the result may be smaller than what is attained in this research work as there are many factors, such as missing values (ratings).

# CHAPTER FIVE

# CONCLUSION AND RECOMMENDATION

## 5.1 CONCLUSION

The availability and production of large amounts of information in the form of newspapers, magazines and other forms of human communication entails the question of sorting or filtering out the right one as human choices are limited. As a result of this fact, filtering out the right information from a collection of information is an issue of research.

Filtering documents helps people looking for documents in many respects. It was designed to help people find the most valuable information, so that the limited time spent on reading, listening or viewing can be spent on the most interesting and valuable documents. Filters are also used to organize and structure information.

The number of newspapers published in Amharic is increasing from time to time. From such a large amount of news items, readers need automatic ways to filter out those items. However, there is no such effective automatic way of filtering news items written in Amharic. Therefore, research needs to be underway concerning filtering out them from a collection, based on a given profile.

It was seen in this study that filtering Amharic texts was possible using a machine learning approach. The application of ANNs to predicting the preferences of news readers as well as its capability for classifying news items in a predefined category have been seen. Sample training and test sets were selected from readers and the news items. Two separate sets were formed. The first set was the set of preference pairs of active readers containing the preferences of news items of 100 active readers of Addis Admas and Reporter newspapers whereas the second set contained the term-by-document matrix of the news items indexed and fed to the neural network model so as to classify them into predefined categories.

ANNs can be used to model an information processing system as applied in this study. ANN consists of a large number of processing elements, called neurons. Each neuron

has an internal state, called its activation or activity level, which is a function of the inputs it has received. Typically, a neuron sends its activation as a signal to several other neurons. A neuron can send only one signal at time, although that signal may be broadcast to several other neurons. In this research, the potential of the BPN and SOM architectures were used.

The preference list in the sample was trained using the Matlab neural network toolbox so as to determine its predictive power. The results showed that the model developed could correctly predict 83.3% of the preferences in the training set and 79.8% of the preferences in the test set. That is, a news item is likely to satisfy the readers in the test set 79.8% of the time.

In order to determine the potential of ANNs for classifying the news items in a predefined category, a SOM model was developed. The model was developed using the training set group consisting of 100 fields of the term-by-document matrix generated from the news items in the sample. The indexed 50 news items were simulated using the model. The training and test sets sampled for training this particular network comprised 66.7% and 33.3% respectively.

The SOM model could correctly classify *76.5%* of the news items in the total collection in the training set whereas *72.9%* of them were assigned to the correct category in the test set. This showed that ANNs could be used for classifying Amharic news items.

## 5.2 RECOMMENDATIONS

The results found in this research showed that ANNs could be applied to reasonably predict users' preferences of reading news items and for classification purpose. There are many algorithms of ANN out of which only two, the BP and SOM networks were considered. Apart from this, more research efforts need to be conducted to efficiently explore ANN technology regarding news filtering.

The following areas are worth studying:

1. The method is used only for the purpose of the research. It has to be tested for a larger collection of news so that it may be more effective, as neural networks learn from large examples,
2. The preferences of news readers were collected using manual technique. This is due to the fact that there are no fully operational web sites to collect readers' ratings of choice. In the presence of such sites, making it possible to collect users' ratings, the research can be extended to a wider domain,
3. User preferences were predicted on the basis of a pure CF system. Other research areas such as using content based CF systems and a combination of the two filtering methods (also called Content-Boosted filtering) can be used,
4. The preferences of the newsreaders included in the sample were studied using the BP algorithm. Other machine learning algorithms could be used and compared with the result attained in this study.
5. Integrating users' preferences and the neural network database in dynamic user interaction is also one area of study.

## BIBLIOGRAPHY


Adam, N. and Yesha, Y. 1997. Introduction, International Journal on Digital Libraries, Springer-Verlag. 1 (1-2).

Allen R. 1990. User models: Theory, method and practice. *International Journal of Man-Machine Studies*. 32. 511-543.

Alspector, J. A. Kolez, and N. Karunanithi. 1997. Feature-Based and Clique-Based Models for Movie Selection: A Comparative Study. User Modeling and User Adapted Interaction. 37(4).

Åsa Rudström et al. 1997. Edited Adaptive Hypermedia: Combining Human and Machine Intelligence to Achieve Filtered Information.

Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-Based Collaborative Filtering Recommendation Algorithms. Department of Computer Science and Engineering. University of Minnesota.

Baudisch, B. 1997. The Profile Editor: Designing a direct Manipulative Tool for Assigning Profiles. *Institute for Integrated Information and Publication Systems, IPSI.* German National Institute for Information Technology. Darmstadt.

Belkin, N. and Croft, W. 1992. Information Filtering and Information Retrieval: Two sides of the same coin? Communication of ACM, 35(12), 29-38.

Bender, Marvin L., Sydney W. Head, and Roger Cowley. (1976). The Ethiopian

Chislenko, A. 1997. Collaborative Information Filtering and Semantic Transports. MIT Media Lab. Available at:

http://www.lucifer.com/~sasha/articles/ACF.html

Coryn A.L. Bailer-Jones, Ranjan Gupta and Harinder P. Singh. 2001. An introduction to artificial neural networks. *Max-Planck-Institute of Astronomie.* Konigstuhl. Heidelberg, Germany.

Dawit Yimam. 1998. Applying Interface Agent Technology to Selective Dissemination of Information (SDI) User Profile Management: The Case of ILRIAlerts. Masters Thesis at The School of Information Studies for Africa. Addis Ababa University. Addis Ababa.

Deerwester, S. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science.* 41(6).

Kevyn Collins-Thompson, Paul Ogilvie, Yi Zhang, and Jamie Callan (2004). At the TREC. Information Filtering, Novelty Detection, and Named-Page Finding. Language Technologies Institute. Carnegie Mellon University.

Klerfors, D., 1998. Artificial Neural Networks: *What are they? How do they work? In what areas are they used?* Louis University. School of Business & Administration.

Kröse, Betal. An Introduction to Neural Networks. Eighth Edition. University of Amsterdam.

Langley, P. 2000. User Modeling and Adaptive Interfaces. Seventeenth national conference on Artificial Intelligence. Daimler Chrysler research and Technology Center.

Lesalu, W. 1965. An Amharic Text Book of Everyday Usage. University of California. Los Angeles.

MacDonald, R 2001. Web-based User Profiling Using Artificial Neural Networks. Honours Thesis, Acadia University, Nova Scotia, Canada.

Malone, T.W., Grant, K.R., Turbak, F.A., Brobst, S.A. and Cohen, L.R. and Riedl, J. (1997), Applying collaborative filtering to Usenet news, *Communications of the ACM*. 30(5). 390-402.

McCelland, J. L. et al. 1986. Parallel Distributed Processing: Explanations in the Microstructure Cognition. The MIT Press.

Melville, Prem, Raymond J. Mooney and Ramadas Nagarajan. 2001. Content-Boosted Collaborative Filtering. *Proceedings of the ACM*. Workshop on Recommender Systems. New Orleans.

Natalie Glance, Damian Arregui and Manfred Dardenne. 1997. Knowledge Pump: Community-Centered Collaborative Filtering. Xerox Research Center Europe. Grenoble Laboratory.

Oard D. 1997. Information Filtering Defined. At: Information Filtering Resources, http://www.glue.umd.edu/~oard/

Oxford University Press.

Palme, Jacob. 1998. Information Filtering. Department of Computer and Systems sciences, *Published in the proceedings of the* ITS'98 conference. Stockholm University/KTH

Parker, K. H. et al. 1979. The importance of SDI for Current Awareness in Fields with Severe Scatter of Information. *Journal of the American Society for Information Science*. 30(3). 125-135.

Resnik, Paul, Neophytos Iacovou, Mitech Suchak, Peter Bergstrom and John Riedl. 1994. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. *Proceedings of the ACM. Conference on Computer Supported Cooperative Work*. Chapel Hill, NC. 175-186.

Salton, G., and McGill, M. J. 1983. Introduction to Modern Information Retrieval. McGraw Hill.

Shardanand, U. and Maes, P. 1995. Social Information Filtering: Algorithms for Automating 'Word of Mouth'. *Proceedings of the Conference on Human Factors in Computing Systems*. 210-217.

Siganos, D. et al. 1999, Neural Networks. Available at: http://www.baclace.net/piif_sidebar_cacm.html

Simon, Haykin, 1994. Neural Networks, A Comprehensive Approach. Macmillan College Publishing Company, Inc. USA.

Sonja Kangas. 2002. Collaborative Filtering and Recommendation. VIT Information Technology Systems.

Sung Young Jung, Jeong-Hee Hong and Taek-Soo Kim. 2002. A Formal Model for User Preference. *Machine Intelligence Group*. Electronics Institute of technology. Seoul, Korea.

Theodros Hailemeskel. 2003. Amharic Text Retrieval: An Experiment Using Latent Semantic Indexing (LSI) With Singular Value Decomposition (SVD). Masters Thesis at the School of Information Studies for Africa. Addis Ababa University. Addis Ababa.

Thomas, Hofmann. 2003. Collaborative Filtering via Gaussian Probabilistic Latent Semantic Analysis. Brown University, Providence, RI, USA.

van Rijsbergen, C. J. 1979. Information Retrieval. Butterworths, London.

Whatmore, Geoffrey. 1978. The Modern News Library. *Documentation of Current Affairs in Newspaper and Broadcasting Libraries*. London: The Library Association. Writing System. In Bender et al. (eds.) Languages of Ethiopia. London:

Zelalem Sintayehu. 1998. Automatic Classification of Amharic News Items. The Case of Ethiopian News Agency. Masters Thesis at the School of Information Studies for Africa. Addis Ababa University. Addis Ababa.

Zupan, Jure, 1994. Introduction to Artificial Neural Network (ANN) Methods: What They Are and How to Use Them. Department of Chemistry, University Rovira i Virgili. Tarragona. Spain.

ከሣቴ ብርሃን ተሰማ፤ 1951፤ የአማርኛ መዝገበ ቃላት፤ አርቲስቲክ ማተሚያ ቤት ታተመ፤ አዲስ አበባ

ደስታ ተክለ ወልድ፤ 1970፤ ዐዲስ ያማርኛ መዝገበ ቃላት፤ አርቲስቲክ ማተሚያ ቤት፤ አዲስ አበባ

## Annex I

## News Items in the sample

### Category 1: Hot News: (0001) ዜና

0001 የኢህአዴግ አባል ድርጅቶች ፈርሰዋል መዋቀራቸውን ቀጥለዋል
0010 የአውሮፓ ህብረት የግሉን ዘርፍ ለማጠናከር እንደሚሰራ ተገለፀ
0011 ኢትዮጵያና ኤርትራ ግንኙነት እንደማይፈልጉ ገለፁ
0100 የግሉ ዘርፍና መንግስት አይተማመኑም
0101 ባህላዊ አርሻ ተፈላጊ እድገትን አያመጣም
0110 በተባበሩት መንግስታት ጠቅላላ ጉባኤ የኢትዮ ኤርትራ ድንበር ጉዳይ ተነሳ
0111 በኢ.ሲ.ኤ. አካባቢ ነዋሪዎች ቤታቸውን በአንድ ቀን እነዲያፈርሱ ተነገራቸው
1000 ቡድኑ ከማስታወቂያ ሚኒስቴር ጋር አልተገናኘም
1001 ቴሌ የቀለጠው መንደር
1010 መንግስት ለ3 አመት የሚያገለግል አዲስ የኢትዮጵያ እቅድ አፀደቀ
1011 የኤርትራ አመራር ሌላ ጦርነት ለመጀመር ካሰበ ውጤቱ ግዛት ከማስጠበቅ በላይ ያልፋል ጠ/ሚ መለስ ዜናዊ
1100 ኤርትራ አለማቀፍ ህግጋትን የጣሰ ተግባር ማከናወኗን የካሳ ኮሚሽኑ አረጋገጠ ለኢትዮጵያ በተደረገው የብድር ቅነሳ ውሳኔ የአሜሪካ ገንዘብ ሚኒስቴር/ ትሬዠር ዲፓርትመንት/ ድምጽ ተአቅቦ አደረገ
1101 የገቢዎች ሚኒስቴር ረቂቅ የስነ ምግባር መመሪያ አዘጋጀ
1110 ለኤች አይቪ ኤድስ መከላከል ተግባር 7.5 ሚሊየን ዶላር ስራ ላይ ዋለ
1111 በአማራ ክልል የመንግስት ሰራተኞች ግምገማ እንዲቀመጡ አዘዘ

### Category 2: Editorial: (0010) ርዕሰ አንቀጽ

0001 ቁና አህልና መኮትኮቻ የሚሰጡህ የቀረውን እንድትተውላቸው ብለው ነው
0010 በልጄ አልተኛሁ፣ በልጄ ተጎዳሁ አልተኛሁ አለበት
0011 ቢወዱት ንቆ የሞቱ ቢጠሉት ምን ሊሆን ነበር
0100 ከህገ መንግስቱ ምን ይጠበቃል
0101 ቅን ስራ ሰርተን እንኳ እውሮች እየወለዱን ነው አለች ውሻ
0110 ሹም ለሹም ይጐራረሳል፤ ድህ ለድህ ይላቀሳል
0111 በራሱ ምንም የሌለው ለሌላው በቅሎ አሞሌ ይገዛል
1000 የፖለቲካው ማቅለጥም መበየድም
1001 ኢትዩጵያዊነታችንም አንድነታችንም አደጋ ተጋርጦበታል
1010 መንግስት በድንበሩ ውሳኔ ያለው አቋም ሀቀኝነት የሚፈተሸው በተግባር ብቻ ነው
1011 መንግስት ጥገኛ ሆነሳ
1100 በህዝብ ተቋማት በቢሊዮኖች እየተቀለደ ነው
1101 ስልጣን የማራዘሚያ ፍተን መድህኒት የህዝብን ጥያቄ መመለስ ነው
1110 ለማሰሪያው ገመድ ቢዘጋጅለት
1111 ፍየል ከመድረስዋ ቅጠል መበጠስዋ

## Category 3: Politics: (0011) ፖለቲካ

0001 ገድሏል ገድሏል፤ ዝናሩ ጐድሏል
0010 የመብት ጥያቄዎችና የካሳ ኮሚሽኑ
0011 ስለድህነታችን የመውጫ መንገዶቹ ጥቂት መነሻ ሀሳቦች
0100 ኢህአዴግ እሱው ያመጣው አሱው ያሮጠው
0101 ኢንቨስትመንትን ያደከመው የፖለቲካው ሰርአት ሳይሆን ፖሊሲውና ስትራቴጂው ነው
0110 እርቁ ለፖለቲካ ወይስ ለዘለቂ ሰላም
0111 ሰላማዊ ሰልፍና ስበሳባ ፈቃጅና ከልካይ የሌለበት የዜጎች መብት
1000 የድንበር ኮሚሽኑ ውሳኔ
1001 ፖለቲካ ወይስ የሚዳ ጨዋታ
1010 ኢህአዴግ ከየት ወደ የት
1011 የህዝብ ውሳኔ ይቅደም
1100 ፍቃድ ከመንግስት ምክር ቤት
1101 የብሄር ብሄረሰቦች መብት
1110 የብሄር ብሄረሰቦች መብት
1111 በህገ መንግስቱ ዙሪያ

## Category 4: Business and Economy: (0100) ቢዝነስና ኢኮኖሚ

0001 የኢንፎርሜሽን ቴክኖሎጂን በመጠቀም ከሌሎች አፍሪካውያንም በጣም ወደ ኋላ ቀርተናል
0010 ድርጅቶች የተጨማሪ እሴት ታክስን እንደወጪ መመዝገብ አለባቸው
0011 ኢትዮጵያና ኬኒያ የንግድና ኢንቨስትመንት ግንኙነታቸውን ለማጠናከር ተስማሙ
0100 የቡና አላላክ ዘዴ ከጅንያ ወደ ላስቲክ ከረጢት ለመቀየር
0101 የአፍሪካ ልማት ባንክ 55 ሚሊዮን ዶላር እርዳታ ሰጠ
0110 ወደ አውሮፓ ገበያ ለመግባት የአውሮፓውያንን ፍላጎት ማጤት ያሻል
0111 ተጨማሪ እሴት ታክስ ከአንድ ዓመት በኋላ
1000 የኢትዮጵያ አየር መንገድ ተመላሽ ጭነት የለም በሚል እስከ መቼ የደርሶ መልስ ያስከፍላል
1001 ብዙ ኢንቨስተሮች ወደ ውጭ ገበያ መግባት አማራጭ እንደሌለው እየተረዱ ናቸው
1010 መሬትን በድርድር ጨረታ ነጋዴውንና የአዲስ አበባ አስተዳደርን እያወዛገበ ነው
1011 ከዘጠና በመቶ በላይ የሚሆኑ ጠበቆች የገቢ ግብር አይከፍሉም
1100 ሸማቹና የድምጽ ብክላ
1101 በፌዴራል መ/ቤቶች በ1995 ዓ.ም ሁለት ቢሊዮን ብር የሚጠና የሂሳብ አያያዝና የውስጥ ቁጥጥር ግድፈት ታይቷል
1110 እንደ ሀገር የቅድመ ምርመራ አገልግሎት ጥገኛ መሆን የለብንም
1111 በከተሞች አካባቢ ከፍተኛ ጥቅም ያላቸውን ሰብሎች በማምረት ግብርናን ማሳደግ

## Category 5: Society: (0101) ህብረተሰብ

0001 የቡድህን ሀውልት ስታቆም መንፈሱንም አብረህ ማቆሙን እንዳትዘነጋ
0010 ቦሌ ኤርፖርት ውብ ነበር
0011 በሌሎች ድክመት መመካት ደካማነት ነው
0100 አወይ ግሎባላይዜሽን

0101 የግብርና ኢኮኖሚስቶች አያስፈለጉኝም ያለው ግብርና መር የኢኮኖሚ ፖለሲ
0110 የማጠቢያ አስመጪዎች ሁለት ብቻ ናቸው
0111 ረሀብ የተፈጥሮ ችግር ሳይሆን የተሳሳተ ፖሊሲ ውጤት ነው
1000 እህል ዘርተህ ዝናብ ካልዘነበ ምን ማድረግ ትችላለህ
1001 ማርቲ አህቲስ የተባበሩት መንግስታት ልዩ መልእክተኛ ገ/መድህን
1010 የኢንፎርሜሽን ቴክኖሎጂን ለመጠቀም ከሌሎች አፈሪካውያን በጣም ወደ ኋላ ቀርተናል
1011 ድርጅቶች የተጨማሪ እሴት ታክስን እንደወጪ መመዝገብ አለባቸው
1100 የዱቤ ነገር
1101 በከተሞች አካባቢ ከፍተኛ ጥቅም ያላቸውን ሰብሎች በማምረት ግብርናን ማሳደግ
1110 የጥጥ አምራች ገበሬዎች በውጭ ገቢያ እንዲያመርቱ ማደራጀት ያስፈልጋል
1111 ኢትዮጵያና ኬንያ የንግድና ኢንቨስትመንት ግንኙታቸውን ለማጠናከር ተስማሙ

## ***Category 6: Culture:*** *(0110)* ባህል

0001 ጥቂት ስለ አሳ ነባሪዎች
0010 ጊዜውም ሲባክን እኛም
0011 የአርምሞ ንጋቶች የግጥም ክሰአቶች፣ የጥበብ ምሽቶች
0100 መህል ላይ መቁለጭለጭ
0101 ምልመላና..... ሳይክሊንግ
0110 የጉንዳን አፈጣጠር
0111 የሰው ሆዱ፤ የወፍ ወንዱ
1000 የሚያውቁ የታደሉ ናቸው
1001 አድናቂዎቼን አስቀይሜያለሁ፤ አዲሱ ስራዬ መታረቂያዬ ይሆናል
1010 የጉዲፈቻ እሰጥ ገባ
1011 ከጀርመን አርቲስቶች ለአዲስ አበባ ሊስትሮዎች
1100 ፎቶ ማንሳት ክልክል ነው የሚለው ምልክት መቅረት አለበት ደረጀ እርገጤ
1101 የዋህነት በቀልና ፍቅር በእሳትና ፍቅር
1110 ከጀርመን አርቲስቶች ለአዲስ አበባ ሊስትሮዎች
1111 አምበቶ ተራራ ጥግ ያለ ባህል

## ***Category 7: Science and Technology:*** *(0111)* ሳይንስና ቴክኖሎጂ

0001 አፍሪካዊው ቢል ጌትስና የአፍሪካው ማይክሮሶፍት
0010 ተንሳፋፊ የዘመኑ መንኩራኩሮች
0011 ማርስ በሮቦት ልትፈተሽ ነው
0100 አፍካውያን የኮምፒውተር እውቀት ሊጨብጡ ነው
0101 የቢል ጌትስ ስመ የቫይረስ መደበቂያ ሆኗል
0110 ፀሀይ ተጋረደች
0111 የሙዚቃውን አለም በጭንቃላቱ ያቆመ ፈጠራ
1000 በፀሀይ ሀይል የሚሰሩ መኪናዎች እየተሞከሩ ነው
1001 አንግሊዝ ከነፋስ ተርባይን ኤሌክትክሲቲ ልታመነጭ ነው
1010 የቴክኖሎጂ አደገኛነት የቪዲዮ ጌሞችና የህልም አለም ሩጫ
1011 ሞባይልና የፍቅር ድብብቆሽ
1100 የሰው ልጅ አንድ ቀን ማረስ ላይኖር ይሆናል
1101 ማርሶ በርክውስኮ

1110 ኢትዮጵያ ያልወቁ አመት የሚስጢር ቁልፍ
1111 ሶፍትዌርን እንዳ ባንባ ውህ ማደረስ ይቻላል

## ***Category 8: Health:*** *(1000)* ጤና

0001 ይድረስ ለኤች አይ ቪ
0010 የጤና ጥግ
0011 መሳሳም ለጤንነት
0100 የቻይናዊቷ ድል ከተሰፋ መቁረጥ በኋላ
0101 ሙዚቃ ህመም ማስታገሻ
0110 በቂ የሆነ የብረት ማእድን እያገኙ ነውን
0111 የደም ነገር
1000 መገለልና መድልዎ
1001 ማልቀስም ለጤንነት
1010 የልብ ድካም
1011 ህኪም ታካሚና ህክምና በኢትዮጵያ ዶ/ር ለጃ ሀምዛ
1100 ጥሩ እንቅልፍ ለመተኛት
1101 የመድሀኒት ጥቅም ጉዳት እና አጠቃቀም
1110 በቃጠሎ ሳቢያ የሚደርስ የአካል ጉዳት
1111 በኤድስ በሽታ የሞተ አንድም ሰው አላውቅም ጤናማ ሆኖ ለመኖር የሚረዱ ጠቃሚ ምክሮች

## ***Category 9: Art:*** *(1000)* ጥበብ

0001 አንድ የቀረሽኝ ዘመዴ
0010 ጋዜጠኝነትና ወንዞቻችን
0011 ቢግ ብራዘር ተቀባይነትን እያገኘ ነው
0100 ቸርችልና አስቀያሚው የፓርላማ አባል
0101 ኪነት በመለኪያ
0110 አለም አቀፋዊ ዝና ያተረፈው ሙዚቀኛ ካምፖይ ሴጉአ አረፈ
0111 አሜጋ አጭር ልብ ወለድ
1000 እንጦጦ ዲልዲላ ፊንፊኔ አዲስ አበባ
1001 የኪነ ጥበብ ደረጃ
1010 መነኩሴ አሳማ በላ ከበላው የሰማ ገማ
1011 አንብብ አቅራራ አስነበብ ትነበብ ዘንድ
1100 መፃህፍትና ክቡሩ ጊዜ
1101 ያሰለጠንከው አውሬ ጥቃት ከፈጸመብህ ጥፋተኛ ማን ይሆናል?
1110 ሸዋዚንገር ሂትለርን አሞግሷል ተብሎ ተተቸ
1111 ትንሽ መጽሀፍ ከባህር ዳር

## ***Category 10 : Sport:*** *(1010)* ስፖርት

0001 ቤክሀም ትልቁ የእግር ኳስ እጀንዳ ሆኗል
0010 የዘመኑ የፊፋ ምርጥ ሴቶች ቡድን ይፋ ሆነ
0011 የዝውውሩ ገንዘብ ተጫዎችንም አስገረመው
0100 በ2003 የአለም እግር ኳስ ንጉሱ
0101 ፓብሎ ግራሲያ አለምን በብስክሌት እየዞራት ነው

0110 የአፍሪካ ዋንጫ ሪኮርዶች
0111 ሰውም አውቆኝ ፀሐይም ሞቆኝ
1000 ማነፃፀር ቢያስቸግርም አሁን ያላቸውን ብቃት አስመልክቶ ከዚህ የሚከተለውን ሙያዊ ንፅፅር ሰጥተዋል
1001 የ24ኛው አፍሪካ ዋንጫ ከዋክብት
1010 ዶ/ር ወልደ መስቀል ኮስትሬ አትሌት ሻለቃ ሀይለ ገ/ስላሴና ቀነኒሳ በቀለን አነጻጽረዋል
1011 የ5ሺ ሜትር የፍጻሜ ውድድር የአለም ትኩረት ስቡዋል
1100 ከስፖርት መሪዎቻችን ጋር ወደፊት
1101 የአዲሰ አበባ ስዲየም የመሮጪያ ትራክ አትሌቶች ላይ የጤና ችግር እያስከተለ ነው
1110 በፌዴሬሽኑ መግለጫ ዶ/ር ወ/መስቀል ኮስትሬ የተሰማቸውን ቅሬ ገለጹ
1111 ኪስና ዋንጫ

### *(c) Matlab Code for generating the SOM diagram*

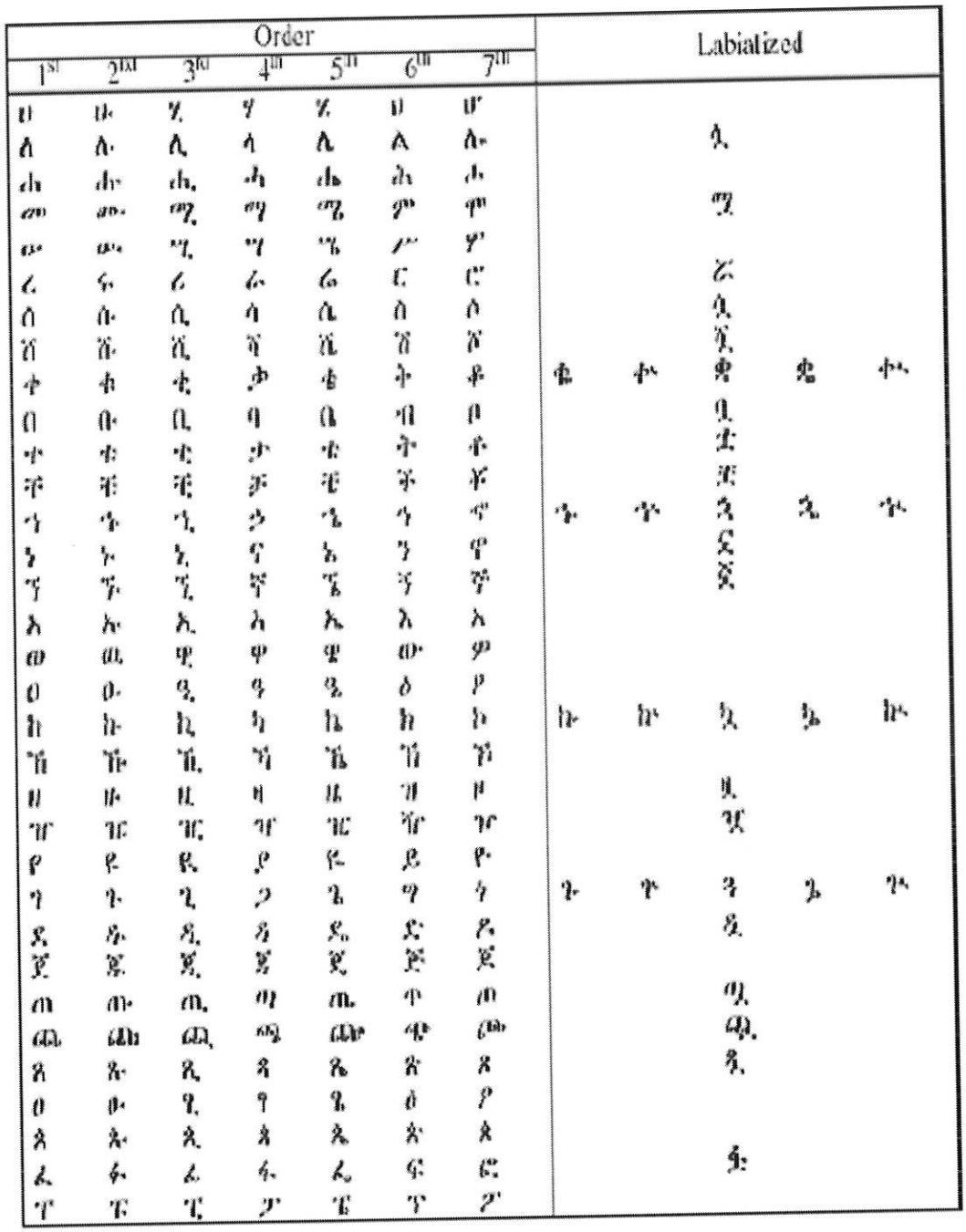
```
function ClasFun(net, data, labels)
% Simulates and plots a SOM
%   plotsimsom(net, data, labels) takes
%   net    - The SOM network.
%   data   - The data that shall be simulated and plotted.
%   labels - A cell-array or double matrix with the labels (optional).

disp('Plotting SOM');

% if double matrix then convert
if ~iscell(labels)
  labels_tmp=cell(length(labels),1);
  for i=1:length(labels)
    labels_tmp{i}=num2str(labels(i));
  end;
labels=labels_tmp;
end

% assign labels to each unit
[output dummy]=sim(net,data);
if nargin==3
  [dummy len]=size(output);
  neuron_label=cell(1,net.layers{1}.size);
  for i=1:len
    winner=find(output(:,i));
    neuron_label{winner}=strvcat(neuron_label{winner}, labels{i});
  end;
end;

% plot the labels at appropriate units
plotsom(net.layers{1}.positions)
for i=1:net.layers{1}.size
  A(i)=length(find(output(i,:)~=0));
  if A(i)~=0
    text(net.layers{1}.positions(1,i)-0.15,net.layers{1}.positions(2,i),int2str(A(i)));
    if nargin==3
      text(net.layers{1}.positions(1,i)+0.1,net.layers{1}.positions(2,i), neuron_label{i});
    end;
  else
    text(net.layers{1}.positions(1,i)-0.15,net.layers{1}.positions(2,i),'none');
  end
end
whitebg(gcf,'white');
axis fill;
drawnow;
```

### *(d) The weighting Scheme*

```
        //Previously used by Theodros Hailemichael (2003)
function [Wmatrix, Gi]=weight(A)
            % signal noise weighting scheme
            % A function used to weight a term by document matrix A and return
            % the weighted matrix W
            % Gi is a vector which holds the global weight of each term in the
            %vocabulary list

           [M K]=size(A) ;      % M-terms and N-documents
           for i=1:M
                 gfi=0;     % collection frequency of term i
             for j=1:K
                 gfi=gfi + A(i,j);   % adds the frequency of term i in each document
             end

          Noise=0;
          Denom=log2(K);
          if(gfi~=0)
            for j=1:K
              pij=A(i,j)/gfi;
                if(pij~=0)
                     Noise=Noise + pij *log2(pij)/Denom;
                end
            end
        end
        Gi(i)=1-Noise;          % global weight of term i.

        for j=1:K
             Wmatrix(i,j)=(log2(A(i,j)+1))*Gi(i);
        end
      end
```

## Annex III

The Amharic character set (Adapted from Bender *et al.*, 1976).



## Annex IV

### Preference List

| | *Preference 1 (Category-NewsItem)* | | | | | | | | *Preference 2 (Category-NewsItem)* | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *User 1* | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| *User 2* | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| *User 3* | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| *User 4* | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 |
| *User 5* | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| *User 6* | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| *User 7* | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| *User 8* | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 |
| *User 9* | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 |
| *User 10* | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |
| *User 11* | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| *User 12* | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| *User 13* | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 |
| *User 14* | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 |
| *User 15* | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| *User 16* | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| *User 17* | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 |
| *User 18* | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 |
| *User 19* | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| *User 20* | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| *User 21* | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 |
| *User 22* | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| *User 23* | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 |
| *User 24* | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| *User 25* | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| *User 26* | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 |
| *User 27* | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 |
| *User 28* | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| *User 29* | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| *User 30* | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| *User 31* | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| *User 32* | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| *User 33* | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 |
| *User 34* | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| *User 35* | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| *User 36* | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| User 37 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| User 38 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| User 39 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |
| User 40 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 |
| User 41 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| User 42 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 |
| User 43 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 |
| User 44 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| User 45 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 |
| User 46 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| User 47 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| User 48 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 |
| User 49 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| User 50 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 |
| User 51 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| User 52 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| User 53 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 |
| User 54 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| User 55 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 |
| User 56 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 |
| User 57 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| User 58 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 |
| User 59 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 |
| User 60 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| User 61 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| User 62 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 |
| User 63 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| User 64 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| User 65 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 |
| User 66 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| User 67 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| User 68 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 |
| User 69 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 |
| User 70 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 |
| User 71 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |
| User 72 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| User 73 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 |
| User 74 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 |
| User 75 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 |
| User 76 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| User 77 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 1 |

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 |
| | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |
| | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 |
| | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 |
| | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 |
| | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |
| | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 |
| | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 |
| | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |
| | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 |
| | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |
| | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 |
| | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |