ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

FACULTY OF INFORMATICS

DEPARTMENT OF INFORMATION SCIENCE

THE APPLICATION OF WEBSOM FOR AMHARIC TEXT RETRIEVAL

By

BIZUNEH MAMUYE BIRHAN

A thesis submitted to the school of Graduate Studies of Addis Ababa University in Partial fulfillment for the Degree of Master of Science in Information Science

JULY 2003

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

FACULTY OF INFORMATICS

DEPARTMENT OF INFORMATION SCIENCE

THE APPLICATION OF WEBSOM FOR AMHARIC TEXT RETRIEVAL

By

BIZUNEH MAMUYE

Name and Signature of Members of the Examining Board

W/rt Saba Amsalu, Advisor _____

Ato Tesfaye Biru, Advisor _____

W/ro Wonishet Abdela, Advisor _____

DECLARATION

This thesis is my original work, and has not been presented for a degree in any other university and all sources of material used for the thesis have been duly acknowledged.

Bizuneh Mamuye Birhan

The thesis has been submitted for examination with our approval as university advisors

Ato Tesfaye Biru

W/rt Saba Amsalu

W/ro Woineshet Abdela

ACKNOWLEGNEMNT

Warm thanks are due to my advisors W/rt Saba Amsalu, W/ro Woinshet Abdela and Ato Tesfaye Biru for their constructive reviews of my work, and scientific and practical advice. I also wish to thank Ato Mulugeta Bayeh for providing me the Nenet tool and technical support till the end of the study. I also wish to extend my sincerest gratitude to Henock for his technical support while coding.

I would like to thank the Ethiopian News Agency for the collection of the news articles. I am indebted also for the opportunity Ato Tesfaye Jemaneh gave me to join SISA and for his interactive participation throughout my educational career.

Finally, I would like to thank all my friends, especially Nebiyou Asfaw, for their moral and material support and many enjoyable moments we had.

TABLE OF CONTENTS

ACKNOWLEGNEMNT II

TABLE OF CONTENTS III

LIST OF FIGURES V

LIST OF TABLES VI

LIST OF ABBREVIATIONS VII

LIST OF APPENDICES VIII

CHAPTER ONE

INTRODUCTION 1

1.1 BACKGROUND 1
1.2 STATEMENT OF THE PROBLEM AND JUSTIFICATION 14
1.3 OBJECTIVE OF THE STUDY 18
1.3.1 General Objective *19*
1.3.2 Specific Objectives *19*
1.4 METHODOLOGY 20
1.4.1 Review of Related Literature *20*
1.4.2 Data Sources and Data Preparation for the Experiment *20*
1.4.3 Experimentation Method *21*
1.4.4 Tools Used *21*
1.4.5 Evaluation Method *21*
1.5 APPLICATION OF RESULTS AND BENEFICIARIES 22
1.6 SCOPE AND LIMITATION OF THE STUDY 22
1.7 ORGANIZATION OF THE THESIS 22

CHAPTER TWO

LITERATURE REVIEW 24

2.1 INTRODUCTION 24
2.2 VISUAL DISPLAYS AND BROWSING 24
2.2.1 Types of Visual Displays *25*
2.3 WEBSOM 28
2.3.1 Document Encoding *32*
2.3.2 SOM *37*
2.3.3 User Interface for document maps *44*
2.3.4 Choosing a good map *48*
2.3.5 Interpretation, evaluation and use of the maps *48*
2.3.6. Adding new documents *50*
2.4 APPLICATION OF SOM 50
2.5 RELATED RESEARCH WORKS 51
2.6. AMHARIC WRITING SYSTEM AND ITS CHARACTERISTICS 54
2.6.1. History of the Amharic language *54*
2.6.2. The Amharic Writing System *55*
2.6.3. Characteristics (features) of the Amharic Writing System *58*

CHAPTER THREE

DESIGN AND DEVELOPMENT OF THE PROTOTYPE SYSTEM 62

3.1 INTRODUCTION 62
3.2 DATA PREPARATION 63
3.3. PREPROCESSING 67

3.4. TEXT-TERM MATRIX CONSTRUCTION 75
3.5 CONSTRUCTING THE SELF-ORGANIZING MAP USING NENET 79
3.5.1 Overview of Nenet *80*
3.5.2 Initialization of the Map *80*
3.5.3 Training the Map *82*
3.5.4 Testing the Map *84*
3.6. DEVELOPMENT OF THE PROTOTYPE BROWSING INTERFACE 90
3.6.1. Labeling the Map *91*
3.6.2 The Image Map and HTML Page *92*
3.6.3. The news articles database *92*
3.6.4. The Active Server pages *92*
3.6.5. Evaluation of the Prototype Browsing Interface *95*
3.7 DISCUSSION ON THE PROTOTYPE MAP 96

CHAPTER FOUR

CONCLUSIONS AND RECOMMENDATIONS 98

4.1 CONCLUSION 98
4.2 RECOMMENDATION 101

REFERENCES 103

APPENDICES 109

LIST OF FIGURES

FIGURE 1.1 CLASSIFICATION OF INFORMATION RETRIEVAL TECHNIQUES 4

FIGURE 2.1 THE BASIC ARCHITECTURE OF THE WEBSOM METHOD (HONKELA ET AL., 1996; KOHONEN, 2000). 31

FIGURE 2.2. TYPES OF TOPLPGY: NEIGHBOURHOODS (0, 1 AND 2) OF THE CENTRE MOST UNIT: HEXAGONAL GIRD ON THE LEFT, RECTANGULAR ON THE RIGHT. THE INNERMOST POLYGON CORRESPONDS TO 0-, NEXT TO THE 1- AND THE OUTERMOST TO THE 2-NEIGHBOURHOOD: SOURCE VESANTO (2000). 40

FIGURE 2.3. STRUCTURE OF A SELF-ORGANIZING MAP 41

FIGURE 3.1 SELF-ORGANIZED MAP DEVELOPMENTS. 62

FIGURE 3.2 FORMATS OF THE DATASET IN A TEXT EDITOR 68

FIGURE 3.3. THE WEIGHTED TEXT-TERM MATRIX 79

FIGURE 3.5. THE NENET INTERFACE (THE MAP IN THE INITIALIZED STATUS) 81

FIGURE 3.6. THE PARAMETERS SET WHILE TRAINING THE MAP 83

FIGURE 3.7. THE PARAMETERS SET DURING TESTING THE MAP. 85

FIGURE 3.8. A MAP SHOWING THE DISTRIBUTION OF NEWS ARTICLES FROM THE CLASS SPORT " ስፖር. 86

FIGURE 3.9. A MAP SHOWING THE DISTRIBUTION OF NEWS ARTICLES FROM THE CLASS ACCIDENT " ጋድአ" 87

FIGURE 3.10. A MAP SHOWING THE DISTRIBUTION OF NEWS ARTICLES FROM THE CLASS AGRICULTURE " ግብጉ" 89

FIGURE 3.11. THE MAP FOR THE 330(ENTIRE NEWS ARTICLES) AUTOMATICALLY LABELED WITH NEWS ID 91

FIGURE 3.12. THE THREE DIFFERENT VIEW LEVELS: THE WHOLE MAP, THE NEWS TITLES AND THE FULL STORY FOR THE SECOND NEWS TITLE. 94

LIST OF TABLES

TABLE 2.1 DIFFERENT FORMS OF THE BASE ALPHABET WITH THE SAME SOUND 59

TABLE 2.2. DIFFERENT ALPHABETS HAVING THE SAME SOUND 59

TABLE 2..3 SOME EXAMPLES OF WRITING COMPOUND NOUNS IN DIFFERENT WAYS ADAPTED FROM BETHLEHEM (2002) 60

TABLE 3.1. CLASSES CONSIDERED WITH THE CORRESPONDING NUMBER OF NEWS ARTICLES 66

TABLE 3.2. NUMBER OF TRAINING SET AND TEST SET CONSIDERED FROM EACH CLASS 67

TABLE 3.3 EXAMPLES OF THE POSSIBLE COMBINATION OF SUFFIXES FROM "ች" "ን","ም", AND "ና" ADAPTED FROM ZELALEM (2001) 72

TABLE 3.4. NEWS ARTICLES WITH THE CORRESPONDING TITLES LOCATED AT THE RIGHT TOP OF THE MAP SHOWN IN FIGURE 3.8 86

TABLE 3.5. NEWS ARTICLES WITH THE CORRESPONDING TITLES LOCATED AROUND THE TOP OF THE MAP SHOWN IN FIGURE 3.8. 87

TABLE 3.6 NEWS ARTICLES WITH THE CORRESPONDING TITLES LOCATED AROUND THE BOTTOM LEFT OF THE MAP SHOWN IN FIGURE 3.9. 88

TABLE 3. 7. NEWS ARTICLES WITH THE CORRESPONDING TITLES LOCATED AROUND THE CENTER OF THE MAP SHOWN IN FIGURE 3.10 89

TABLE 3.8. NEWS ARTICLES WITH THE CORRESPONDING TITLES LOCATED AROUND THE BOTTOM OF THE MAP SHOWN IN FIGURE 3. 90

LIST OF ABBREVIATIONS

ASP	Active Server pages
BMU	Best Matching Unit
CGI	Common Gateway Interface
HTML	Hypertext Markup Language
IDF	Inverse document frequency
IR	Information retrieval
LSI	Latent Semantic indexing
Nenet	Neural Network Tool
SOM	Self Organizing Map
SVD	Singular Value Decomposition
TF	Term frequency
VSM	Vector Space Model
WEBSOM	Web Based Self Organizing Map

LIST OF APPENDICES

APPENDIX 1: LIST OF AMHARIC CHARACTERS AND NUMBER SYSTEMS 109

APPENDIX 2: CLASSES OF NEWS ARTICLES 111

APPENDIX 3: LISTS OF SUFFIX 112

APPENDIX 4. THE NEWS ARTICLES TABLE 113

ABSTRACT

This research explored the applicability of WEBSOM (Web Based Self Organizing map) for retrieving texts written in Amharic language. The method applies a neural network's self organizing algorithm for generating the map display. The map display detects complex relationships among given documents, and reveals the relationships based on the arrangements of terms abstracted from the documents.

To conduct the experiment, 330 Amharic news articles of three classes were collected from the Ethiopian News Agency. 248 of the news articles were taken as a training set and the remaining as a test set. For the purpose of document representation, the Vector Space Model was used. Non-content bearing terms were removed from the lists of terms identified from the headline and slug parts of the news articles and suffix/prefix-stripping technique was applied on the remaining list. After changing terms having different writing forms in to one common form, terms with a total frequency of above 70 and below 3 were discarded from the list. Then, a matrix both for the training and test set were constructed on the remaining 142 terms. A normalized weight was assigned to each term in a given news article based on TF-IDF (Term Frequency-Inverse Document Frequency) weighting technique and the vector matrix were prepared in appropriate format for the tool to be used.

Using Nenet (Neural Network Tool), the SOM map was trained with the 248 articles in the training set and tested with three test sets selected from the three classes of news articles. From the distribution of these articles on the map, it was observed that the map placed similar articles near to each other. The results obtained from the three tests made, indicated that the clustering capability of the SOM for Amharic documents is promising.

Lastly, a map was constructed for the entire (330) news articles and an HTML based prototype browsing interface map was developed and labled with descriptive terms that convey properties of the area. A link was also made with the actual database through the Active Server Pages created so that users can browse on the map for relevant articles.

Chapter One
Introduction

1.1 Background

Advances in information and communication technology (ICT) have significantly revolutionalized our possibilities to collect, generate, distribute and store text data beyond any system can afford to provide a solution to determine its value (Lagus, 2000). Text databases are thus rapidly growing with the increasing amount of information available in electronic format, such as electronic publications, email, CD-ROM and World Wide Web (WWW) (which can also be viewed as a huge, interconnected, dynamic text databases). The need to process these large, growing collections of electronic documents, which consists of documents from various sources such as journal articles, research papers, books, digital libraries has become everyday challenge (Mulegeta, 2002).

Different information processing systems have been designed as solutions to such problems. Some examples of information processing systems are database management systems, management information systems, decision support systems, and information retrieval systems (Salton and McGill, 1983). Information retrieval (IR) systems are designed to facilitate access to stored information. Moreover, according to Salton and McGill (1983), the functions of information retrieval (IR) systems extend to include the representation, storage, and organization of information.

In nutshell, the goal of any retrieval system is to aid the user in fulfilling his/her information need. The user is the focal point of all information retrieval systems, whose sole objective is to provide information that will satisfy user's information need. Finding text that satisfies a user's information

need is not simple since user's information need is a vague concept, which doesn't remains constant (Croft, 1995). Different writers have said a number of things regarding people's needs of information. According to Lin (1997) and Lagus (2000a), for instance, people approach large information spaces with different kinds of motives. Their prime motive might be searching for a specific piece of information or locating something that might be of interest, without a clear prior notion of what "interesting" should look like. It may also arise when an individual recognizes that his/her current state of knowledge is insufficient to cope up with the task at hand, or to fill a void in some area of knowledge.

One major issue that is related with user's information need is user's task - a task that is required from the user while interacting with the retrieval system. Baeza-Yates and Ribeiro-Neto (1998) discussed that users of an information system will approach the system having different tasks, which can be categorized, in general, in to two: retrieval and browsing. If the task is retrieval, for instance, the user of the system is expected to translate his information need into a query in the language provided by the system. This normally implies specifying a set of words that convey the semantics of the information need.

Similarly, Chen et al. (1998) and Lagus (2000b) summarized three major tasks related to various information needs: searching, browsing and visualisation.

Searching: - The user specifies his/her need by a query and expects the system to locate individual documents that correspond to it. Internet search engines are a familiar example of such tools. This approach presupposes a clear knowledge of the user as what is to be found, and a potential to explicitly express them. However, the need may be often vague, the domain unknown,

and the user is constrained by the appropriate, specialized vocabulary (often called the vocabulary problem).

Browsing: - The user navigates via links between individual documents (WWW for instance) or via some hierarchical structure such as the contents section of a book or the directory structure of an explorer window. The browsing approach reduces the information need to be more vague or unconscious, since no explicit description of the need is required. Instead, the need is implicitly communicated via the choices made in browsing, such as the links followed.

Visualization: - Something familiar is used as a means for illustrating something unfamiliar. There exist information needs that require assessing and conveying similarities, differences, overlaps, and other relationships between collections of documents. One might, for example, wish to find out the relationship between familiar set of documents, e.g. personal files or familiar mailing list, to yet unfamiliar collection. Using suitable visualizations, intricate relationships between large collections of items can be communicated fast and intuitively.

In summary, understanding user's information requirements and user's task will help in determining the nature of information to be collected by the retrieval system. It also help in determining the nature and level of analysis to be made in order to store the information and the nature of user interface to be designed so that users can interact with the retrieval system in order to search and retrieve the required information (Chowdhury, 1997).

Different techniques have been devised in the field of information retrieval (IR) with the aim of representing, storing, organizing and accessing information items, which in turn will help in

addressing the different information needs discussed above (Salton, 1983; Belkin and Croft, 1987; Lagus, 2000a).

Scholars made distinctions among those techniques based on different factors. The following discussion is based on the classification made by Belkin and Croft (1987) and (Hildreth, 1995). At the broader level, information retrieval systems can be classified in two categories: query-oriented systems and non-query-oriented systems (see Figure 1.1 below).

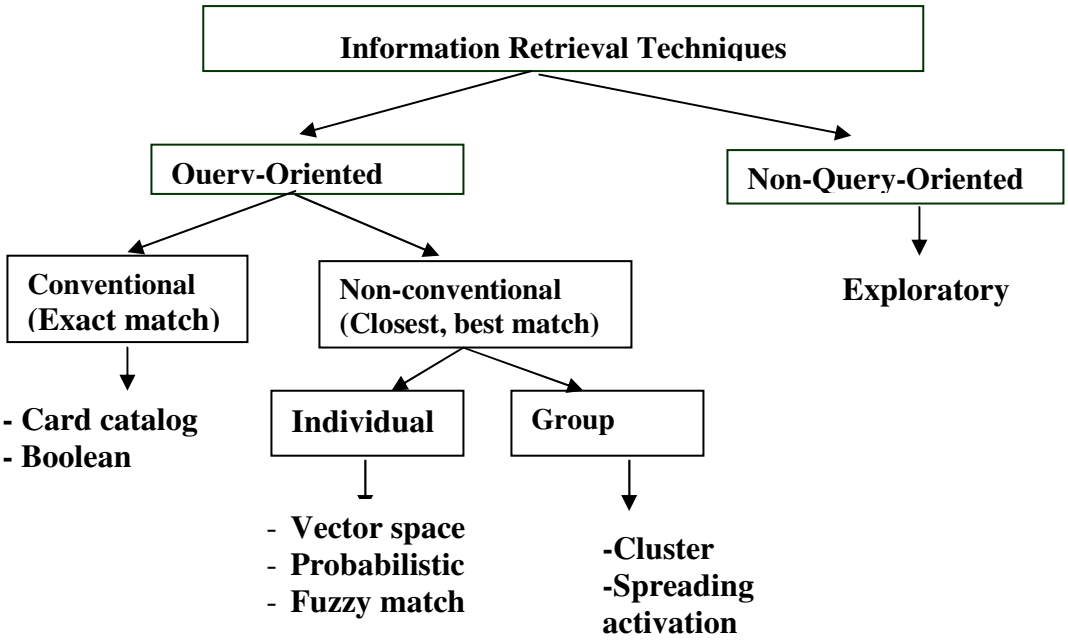


Figure 1.1 Classification of Information retrieval Techniques

Query-Oriented Retrieval Techniques

In Query- Oriented retrieval techniques, first the documents that need to be stored by the system will be represented by assigning appropriate terms and identifiers capable of representing the content of the collection items, and sufficient enough for the retrieval of the document in response to subsequent queries (Salton and McGill, 1983; Tesfaye, 1987;Croft,1997). In the retrieval

process, the user expresses his/her information need using query terms and the strategy implemented in the system will analyze the query, compares it with the collection items and those items that match with the query will be retrieved.

Distinction can be made under query-oriented systems depending whether the set of retrieved documents contains only documents whose representations are an exact match (conventional systems) with the query or a partial match (non-conventional systems) with the query. For a partial match, the set of retrieved documents will include also those that are an exact match with the query.

A. Conventional (Exact match) Techniques

Exact match techniques require that the specifications of the query (e.g., the search terms and their specified logical or textual relationships) be satisfied precisely by all document representations that would make up the retrieval set. Among the most common classical retrieval models that applies this technique is the Boolean model, which is a simple retrieval model established based on set theory and Boolean algebra. This model requires the document to be represented by a set of index terms that appear in the document. It involves binary weighting[1] while indexing documents. Then, to search for a document, a query consisting of a set of index terms combined with Boole's operators[2] need to be formulated. Moreover, in this model, the semantics of the query is also well-defined, i.e. each document either fulfils the Boolean expression or does not (Salton, 1987; Belkin and Croft 1987; Baeza-Yates and Riberio-Neto, 1999; Lagus, 2000).

[1] Binary weighting assigns 1 if that index term is present or 0 if it is absent in a given document.

[2] Boole's operators are the logical operators AND, OR, and NOT.

Implemented as Boolean, full-text, or string searching, exact match retrieval technique is widely used by most of today's operational information retrieval systems and online catalog systems (Hildreth, 1995). However, this type of technique suffers from major drawbacks, which are well known and well documented. First, formulating a suitable query, i.e., the selection of appropriate query terms is difficult, especially if the domain is not well known. Second, its retrieval strategy is based on a binary criterion (i.e., a document is predicted to be either relevant or non relevant) without any notion of ranking scale, which prevents good retrieval performance (Baeza-Yates and Riberio-Neto, 1999). Third, the size of the output cannot be controlled: the result may as easily contain zero or thousands of hits. Furthermore, without a concept of 'partial match', one cannot know what was left out of the query definition (Lagus, 2000).

Despite these mentioned drawbacks, exact match technique in general and Boolean model in particular is the dominant model with commercial document database systems and remains the paradigm for operational systems.

B. Non-conventional Retrieval Techniques

Where conventional information retrieval (IR) systems employ an exact match retrieval strategy, non-conventional query-oriented systems employ a "closest" or "best match" strategy where degree of closeness or similarity of a candidate item's content description to the textual query is taken into account. Various measures of "closeness" or "similarity" between query and document representation or document have been in use (Hildreth, 1995).

Non-conventional techniques can also be further classified as retrieval techniques that compare the query with individual document representatives and techniques that use a representation of

documents that emphasizes connections to other documents in a network. The former category has employed different retrieval strategies such as vector space processing, fuzzy match, and probabilistic techniques and the later category has employed networking strategies like clustering etc.

Vector Space model: In the vector space model, both documents and queries are represented as t-dimensional vectors, where each dimension corresponds to an index term. In general, the procedures used in constructing the vector space model can be divided in to three stages. The first stage is concerned about document indexing where content bearing terms are extracted from the document text. The second stage deals about weighting of the indexed terms to enhance retrieval of document relevant to the user and the last stage ranks the document with respect to the query according to a similarity measure.

The model considered the drawback of binary weighting as well as proposes a frame work of partial matching, which is possible by applying some sort of distance or angle calculation between a given query and set of documents so as to improve the retrieval performance. The notion of degree of similarity is possible because of the assignment of non-binary weights to index terms in queries and documents. While weighting terms, mostly, it applies the combination of within-document frequency (tf)[3] and inverse document frequency (idf)[4] weighting technique, where the weights can be calculated for document terms either as part of the retrieval process or when the documents are indexed (Salton and McGill, 1983; Salton and Buckley cited in Lagus, 2000; Baeza-Yates and Riberio-Neto, 1999).

[3] tf refers for the number of times a term occurs in individual document.

[4] idf refers for the inverse of the number of documents in which a term occurs

As far as ranking is concerned, different ways of similarity measures have been proposed and used by the model such as cosine coefficient, Jaccard, Dice coefficients etc., so as to compare the query against each document in the collection (Salton and McGill, 1983; Rijesbergen, 1996). However, the cosine coefficient, which measures the angle between the document vector and the query vector in the Euclidian space, is the dominant one (ibid).

Proposed by Salton, the model has formed the basis of a large part of IR research (Honkela, 1997; Lagus, 2000b). Furthermore, its term-weighting scheme, its partial matching strategy and its ranking technique base on degree of similarity can be considered as the major advantage of the model. On the other hand, theoretically, its high dimensionality and term-independence assumption made by the model can be regarded as a disadvantage since the occurrence of one term in a given document may affect the occurrence of another term. However, in practice, the model remains as a popular retrieval model nowadays (Salton and McGill, 1983; Baeza-Yates and Riberio-Neto, 1999).

Probabilistic Model: Robertson and Spark Jones (1976), cited in Baeza-Yates(1999) originally introduced another classic retrieval method called the *probabilistic retrieval model*. It makes explicit the probability ranking principle that can be seen underlying most of the current IR research. In brief, the principle states that optimal performance is achieved by that retrieval system which ranks retrieved documents in decreasing order of their probability of relevance to the query that has been submitted to the system. The model has assumed that terms are distributed differently in relevant and non-relevant documents (Rijsbergen, 1996; Baeza-Yates and Riberio-Neto, 1999; Lagus, 2000). A binary weighting technique is applied while document representation and terms are assumed independent. Given this assumption, the model ranks documents in decreasing order of their probability of being relevant.

Relevance feedback from the searcher is considered to be essential to the effective performance of probabilistic retrieval systems. While querying, a user initiates interaction with the system by providing (formulating) a query. The system estimates all the relevant documents for the query and returns it in order of their decreasing probability of relevance. The searcher then may explicitly change the values (system calculated weights) assigned to search terms or may respond to the first-listed, top-ranked documents. In this way, relevance feedback may lead to a refinement or expansion of the user's query and "fuel" the system for even better performance (Robertson and Spark Jones, 1976; Rijsbergen, 1996; Baeza-Yates and Riberio-Neto, 1999).

Even though the model has a sound theoretical ground, there has been many drawbacks associated with it. The first challenge is its requirement to initially guess relevant and non-relevant documents for a given query, i.e., the model doesn't state explicitly how to obtain the estimates regarding which documents are relevant and which are non-relevant. The second problem is the model doesn't take into account non-binary weighting technique, which are presumed to increase retrieval effectiveness and lastly its term independence assumption can also be considered as a problem since in reality terms might exist as dependent to each other (ibid).

Fuzzy Set: - Representing documents and queries through sets of keywords yields descriptions, which are only partially related to the real semantic contents of the respective documents and queries. As a result, the matching of a document to the query terms is approximate (or vague). An alternative technique is to use the fuzzy set, where documents and queries can be modelled by considering that each query term defines a fuzzy set[4] and that each document has a degree of membership in this set, i.e, in stead of inclusion or exclusion of an element to a set. A membership

[4] Classes whose boundaries are not well defined and it also defines a membership function that takes values in the interval [0, 1] with 0 corresponding to no membership in the class and 1 corresponding to full membership.

function is defined by the model to express the degree to which the element is a member of a set. (Salton and McGill,1983; (Baeza-Yates and Riberio-Neto, 1999).

The main contribution of this work in terms of retrieval techniques has been the integration of Boolean queries with ranking techniques (Belkin and Croft, 1987). This integration is limited, however, when compared with extended Boolean retrieval based on the vector space model or the use of term dependencies in probabilistic models. Moreover, fuzzy set models are not popular among the information retrieval community. The vast majority of the experiments with fuzzy set models have considered only small collections which make comparisons to the models (Baeza-Yates and Riberio-Neto, 1999).

Clustering: - In a retrieval environment, a cluster technique can serve as a means for representing knowledge as well as a means for providing efficient search strategy, where its search effectiveness stems from the *cluster hypothesis*, which asserts that closely associated documents tend to be relevant to the same query (Salton and McGill, 1983; Rijesbergen, 1996).

The final result of a clustering process is a cluster hierarchy, where large clusters are analyzed and broken down into a number of smaller clusters, which are themselves broken down into still smaller clusters, until finally the lowest level clusters are broken down into individual documents. Then, a top-down or bottom-up search strategy is implemented in the system.

Clustering technique can be applied for organizing information in circumstances where there is a large mass of collection that needs to be manipulated by the retrieval system, i.e. clustering will provide order among a collection and simplifies searching manipulation (Rijesbergen, 1979; Salton and McGill, 1983). The emphasis on small, well-defined clusters has also led to the development of

retrieval techniques based on the generation of the documents nearest neighbours on which WEBSOM (Web Based Self -Organizing Map), the current focus of this research, is primarily based (see detail in later section).

Spreading activation*:* A query is used to 'activate' parts of a network that describes the contents of documents and how they are related to each other. In the simplest case, the query would activate index term nodes that are connected to document nodes and other terms. The links and nodes represent concepts from the subject domain and how they relate to each other as well as the documents that contain those concepts. From the 'start nodes' provided by the query, other nodes connected to those nodes are in turn 'activated' (hence, the term 'spreading activation'). Criteria, such as threshold values that decrease as the activation propagates through the network or rules about the reasonableness of the inference implied by using a particular link, are used to control the spread of activation. Activation can converge on particular document nodes from a number of links. These highly activated nodes are retrieved (Salton and Buckley, no year; Belkin and Croft, 1987).

In a simple network of words, for instance, selected words in the network get an activation value, and that this activation follows the links to associated words. The activation reaching a word is the sum of the activations coming through all its input links, each proportional to the strength of the link. From an activated word, activation diffuses further to the words linked to it, and so on. This makes it possible to find words which are related to all the words that are initially activated. For example, the activation diffusing from the initial selections "paper" and "education" will concentrate most strongly in the word "book". Other examples are "control" and "society", which activate "government", "car" and "town", which activate "road", and "work", "room" and "building" which together activate "office". This is similar to the way thoughts diffuse in the brain, moving along

intuitive, fuzzy pathways, rather than retrieving exact matches like the Boolean retrieval techniques.

Non-Query Oriented Retrieval Techniques

Though most commercial systems still function in accord with general principles of traditional query-based framework, the research output of so many years suggested that this retrieval technique do not always satisfy the needs of prospective users. For instance, since all models under this category do not take into account information seeking that is not query-based or centered, for example, many kinds of browsing and exploratory searching are not supported (Lin, 1997).

Browsing is an explorative and interactive process in which one will scan large amounts of information, perceive or discover information structures or relationships, and select information items through focusing one's visual attention. If the documents, terms, and other bibliographic information are presented in the system as a network of nodes and connections, the user can browse through this network with system assistance. Browsing is an interesting retrieval technique in that it places less emphasis on query formulation than do other techniques and relies heavily on immediate feedback provided by user browsing decisions (Lin, 1997).

Through dialogue with the user, the system uses the network to build a model of the user's information need that includes relevant documents found during the process. Other research in browsing, for instance WEBSOM, focuses on the use of Visual representations of the document database to acquire information from the user interactively (Mulugeta, 2002). In fact, the WEBSOM

method can be described as a combination of the cluster, spread activation and browsing retrieval techniques described above.

In WEBSOM method, documents are organized using a self-organizing map (SOM) algorithm on to a document map (Kohonen, 2002). A graphical display of the map provides a general overview of the information contained in the document collection (Kaski et al, 1998). The map displayed can then be readily used to explore the document collection. They can also be labeled with automatically identified descriptive terms that convey properties of each area and act as a landmark during exploration. With the help of an HTML-based interactive tool the ordered landscape is then used in browsing the collection and in performing searches on the map.

The self-organizing map algorithm is a special architecture of neural networks that cluster high dimensional data vectors according to a similarity measure (Kclose et al, 2002). The clusters are arranged in a low-dimensional topology that preserves the neighborhood relations in the high dimensional data. Thus, not only objects that are assigned to one cluster are similar to each other (as in every cluster analysis), but also objects of nearby clusters are expected to be more similar than objects in more distant clusters.

The method also addresses the vocabulary problem of all query-based techniques by organizing the words into categories on a word category map (also called self-Organizing semantic map). The word category maps are SOMs, which have organized words according to similarities in their roles in different contexts. They can then be used to represent documents in a manner that explicitly express the similarity of the word meanings to take into account the fundamental problem of vector space model. Each unit on the SOM corresponds to a word category that consists of a set of similar words.

It is the aim of this research to see the potential applicability of this method (WEBSOM) for Amharic text retrieval since no one considered it before for the language under consideration. In fact, there was one research that has been done at SISA on English research documents of ILRI (International Live Stock Research Institute) by Mulegeta (2002).

1.2 Statement of the Problem and Justification

Although more than 80 languages are spoken in Ethiopia (Bender, 1976), Amharic is the most widely used and it is the working language of the Federal Government of Ethiopia. As stated in Bethlehem's (2002) Masters Thesis, according to a census report by ECSA (Ethiopian Central Statistics Authority) (1998), it is the first language for more than 17 million and second language for over 5 million people. The language is used in offices, Schools and other public services. As a result of the language's use by many users, a large number of documents are increasingly published and distributed in both hard and soft copies. As the amount of information to be processed gets larger and larger, systematic representation of information is of paramount importance. Such representation might be difficult unless automatic information storage and retrieval is used.

On the other hand, research work in the area of automatic information retrieval for Amharic is still in its infancy. Researches that have been done in the area of information retrieval for the language under consideration is Nega's (1999) work, conducted to develop a stemmer for the Amharic language, Automatic Classification of Amharic News by Zelalem (2001), The Application of Information Retrieval Techniques to Amharic Documents on the Web by Saba (2001) and n-grams Approach to Automatic Indexing for Amharic Text by Bethlehem (2002).

Recent technological advances continue to lead to more drastic changes in the way information is organized and retrieved, even though traditional query-based retrieval systems have since established themselves fairly well (Kemp, 1988). Startling advances in computer networks also afford instant global access (the internet for example) to tremendous diversity and volumes of information making information overload an increasingly pressing research problem (Chen et al, 1998). Although these advances might be said without sound trait, the implications of them all continue to call for a shift from a simple query-based (Key-word based) search to content-based search (based on the content similarity or relationship between documents) paradigm (Chen et al, 1998).

This shift in the human-computer interaction paradigm is stimulated by the realization that the same idea or concept can be presented in many different ways (Kaski et al, 1998). Natural language (any language that human beings learn from their environment and make use of it to communicate with each other in their day–to-day activities) gives freedom for enormous variation in expression, from choosing between synonymous to using different styles, emphasis, different levels of abstraction, and metaphoric expressions. Furthermore, authors use their unique style that depends on their background, knowledge and personal style of communication. These variations in natural language potentially render information retrieval systems, based solely on the keywords as they appear in the text, misleading because these methods have limited possibilities to tackle these phenomenon (Kaski et al 1998; Lagus et al., 1998; Kohonen et al 2000). According to Lagus (200b), these methods of organizing and managing text have also become both inadequate and too expensive to perform and to maintain for the majority of the available collection.

Moreover, large quantities of textual data available for example on Internet pose a continuing challenge to applications (like search engines) that help users in making sense of the data. Hence,

this may currently require considerable efforts in devising suitable search expressions for information retrieval from a full-text online database (Kohonen et al, 2000), which in turn heavily rely on the users competence to generate adequate queries. Re-formulation of the search expression also required when the scope of the search needs to be changed. At best, when interaction is involved, users can modify queries based on the retrieved results of a previous query. But again, this is analogous to searching books in a library without light by walking around from stack to stack in the dark, without knowing what stacks were walked through (Lin, 1997; Lagus et al, 1996a). In this situation, success in finding a book greatly depends on whether one can walk to the right place in the dark (to generate a good query), and whether one knows how to adjust one's locations until one gets to the right place (to modify queries interactively(Mulugeta, 2002).

The problem is further complicated by the degree of similarity between documents. Most users have difficulty in specifying their needs by a specific query formulation; even if users are successful in doing so, systems have difficulty in retrieving all relevant documents without overwhelming the users with irrelevant documents (Lagus et al., 1996b; Lin 1997; Chen et al., 1998). A problem also arises when the exposure to the system and domain knowledge of the user are limited. Fulfilling a vague information need regarding an unknown domain, or obtaining an overview of a topic or a domain still remains more illusion. Even these significant advances in information organization and retrieval are left much to be desired at when the answers sought relate to a set of documents instead of a single document, or when unexpected patterns or trends should be identified (Lagus et al., 1999; Lagus, 2000b).

One problem left unmentioned with these query-based retrieval techniques is their 'linear display'. That is, documents matched to the query are displayed as a list of document titles. Even when relationships between those matching documents exist, what the displays do is simplifying all the

relationships by reducing it only to a one-dimensional, linear order, chronologically, alphabetically, or according to some kind of relevance (weighted) order between the matching documents and the query (ibid).

Many information retrieval systems including online catalogs are appearing supporting some form of browsing. In traditional, query-oriented systems, browsing plays a subordinate, supporting role in assisting the formulation or modification of a query that is to be matched exactly or partially with document representations. This probably explains why some people view browsing as a secondary activity, and not as real searching. Some forms of browsing are quite different than this and may serve as the primary information seeking method used by most people in real-life searching situations. In light of this, some researchers have suggested that a browsing paradigm for searching should replace the query-matching paradigm in the design of information retrieval systems (Hildreth, 1995). Hence, designers of information retrieval systems and online catalogs must expand their knowledge of the browsing requirements of searchers, and should provide capabilities and search options in their systems that will support these requirements (Hildreth, 1995).

The method used in this research, which is applying WEBSOM method for Amharic text retrieval, can contribute to the simplification and promotion of automatic information retrieval efforts for Amharic text documents. The researcher is motivated on this particular retrieval method because it provides interesting features such as visualizing and browsing, where other traditional query-based retrieval techniques do not taken into consideration. Mulugeta (2002) has also recommended for WEBSOM to be applied for Amharic text retrieval.

WEBSOM method to text retrieval, which applies explorative or browsing technique, has begun to be successfully applied to address a variety of these traditional query-based retrieval problems discussed above on complex document sets being implemented as integral part to those systems (Lagus and Kaski, 1999). With this method a textual document collection may be organized onto a graphical map display that provides an overview of the collection and facilitates interactive browsing. The organized map offers an overview of an unknown document collection helping the user in familiarizing herself/himself with the domain. Map displays that are already familiar can be used as visual frames of reference for conveying properties of unknown text items. Interesting documents also can be located on the map using a content-directed search since nearby locations contain similar documents.

The method has been well-studied and developed especially for small-scale problems, although with increasing computer power it has also become possible to tackle much larger problems of document retrieval (Kohonen et al, 2002). Preliminarily results in a text retrieval experiment also indicated that the WEBSOM method and the resulting document maps perform at least comparably with more conventional retrieval methods (Lagus, 2000a). Besides, it has been claimed that the method is scaleable and generally applicable on document collections of various sizes, text types, and languages (Lagus, 2000b). Hence, this research attempts to look into the potential application of this popular artificial neural network algorithm-based method to solve the problem of document organization and to ease exploration of Amharic text.

1.3 Objective of the Study

To conduct the research the following general and specific objectives were established.

1.3.1 General Objective

The general objective of this research is to explore the possibility of applying WEBSOM method for Amharic text retrieval. By using WEBSOM, an Amharic interface (Map) that is user friendly for inexperienced information seekers as well as for experienced users with little or no knowledge of the subject domain of available documents was developed.

1.3.2 Specific Objectives

In line with the general objective, the researcher attempted to deal with the following specific objectives.

- Review the research context of the WEBSOM method, namely text information retrieval, with emphasis on the visual and exploratory aspects in the context of document collection.
- Collect Amharic text documents (News) from Ethiopian News Agency.
- Preprocessing the selected documents to avoid some of non-textual information like numbers, symbols, punctuation marks etc.
- Identifying a list of terms from the document titles and body of the news articles.
- Comparing the list to stop words to delete common terms that do not contribute much to document representation
- Removing some of the most and least frequently occurring terms from the list by fixing an appropriate threshold value
- Weighting and normalizing the remaining terms by applying weighting techniques and representing the collection as a vector of these lists of terms.
- Generating visual display of the document collection using the self-organizing map (SOM) algorithm using Nenet software.

- Evaluate the document map constructed
- Constructing a user interface
- Drawing useful conclusions and forward recommendations for further study

1.4 Methodology

In order to achieve these objectives, the following methods were employed:

1.4.1 Review of Related Literature

Relevant literatures (books, journals, magazines and web documents) pertaining to the research under consideration were reviewed. In addition, since developing a document map for the language under consideration requires one to investigate and identify the property of Amharic words and phrases that are useful for representing contents of documents, related literatures were also reviewed on the language in general. Also, researches done so far in the area of automatic information retrieval for Amharic in particular and researches done in IR using WEBSOM method for other languages in general were reviewed.

1.4.2 Data Sources and Data Preparation for the Experiment

The researcher collected data sources (Amharic documents) that helped in applying the selected retrieval method, i.e, WEBSOM from Ethiopian News Agency. The agency was chosen as a data source since it has large collection of Amharic News articles in its electronic database. After the raw Amharic documents were collected, different processes were undertaken like text preprocessing (removing non-textual information), term identification that can possibly represent the documents and weighting them accordingly. Finally, the News articles selected were represented and formatted in such a way that the Nenet can handle it.

1.4.3 Experimentation Method

After the documents have been collected and represented, the data were modeled (document map was constructed) using the neural network machine learning algorithm called SOM to answer the objective of the study under consideration. Here documents were categorized as a training set and testing set to train the model and test respectively.

1.4.4 Tools Used

This study combined and used various programming techniques to implement the complex preprocessing (feature extraction for the purpose of document encoding) and user interface development phases of the work in addition to a SOM implementation software tool called Nenet. Deliphi programming was applied for encoding the required functions and procedures to extract the key terms from both the headlines and slug part of the news articles to represent the news articles. Also, Java and Visual Basic was used for the development of the browsing interface which requires the development of appropriate HTML pages.

1.4.5 Evaluation Method

Good evaluation methods for measuring the quality of visualization, exploration and navigation are very difficult to define (Lagus, 2000). However, an attempt was made to test whether the SOM adapted itself to the characteristics of the data pattern using the test set. Then, the final user interface (web based map) was presented to subject experts (Journalists) and an attempt was made to analyze their comments and reactions.

1.5 Application of Results and Beneficiaries

The results of this research could be applied in organizing web based Amharic texts of different institutions such as for organizing Amharic-based official letters, personal files, research papers, library collections, newspaper articles and corporate-full text databases. In general, the results obtained out of this research will support any organization or institutions that have Amharic text databases that need to be organized for the purpose of retrieval.

1.6 Scope and Limitation of the study

The study mainly considered Amharic News articles. News articles of other languages were not considered. Moreover, it aimed in developing a prototype web based user interface since it is impossible to develop a full-fledged system for the entire database. Developing a full-fledged system demands one to construct maps of different levels for easy exploration, which in turn requires a long period of time. Hence in this research only a one level map was constructed with out having any zooming and searching facilities. The user can click from the map to get the head line (titles) of the news items and with further clicking the full story can be obtained. Moreover, the research was conducted only on texts. Organizing images, picture and numerals were not considered.

Since the current researcher couldn't get a tool that supports Amharic font types, labeling of the map was done using the Latin characters.

1.7 Organization of the thesis

The overall content of the thesis is organized into four chapters. The first chapter provides an overview of the classical retrieval techniques and trends of retrieval systems, statement of the problem with proper justification and objectives of the current work. In chapter two, review of

related literature regarding WEBSOM method was presented in the context of text exploration. It particularly discusses the different techniques for data exploration and the basic architecture of WEBSOM, and the details of the self-organizing map as well. At the last part of the chapter, a general overview of the Amharic language was given.

Under chapter three a discussion was made about how the SOM was used for organizing the Amharic news articles, the design and implementation issues of the user interface prototype for Amharic news articles for the purpose of ease exploration. At last, a summary was made by giving some useful conclusion and recommendations for further study.

Chapter Two

Literature Review

2.1 Introduction

This chapter is organized in to two parts. The first part discusses different issues regarding WEBSOM. First, overviews of the concept of visual displays and browsing have been discussed and then the theoretical and technical issues related with WEBSOM have been presented in detail. Finally, researches done so far in relation with this new method of text retrieval have been revised.

Under the second part, a general overview of the Amharic language will be given. Features of the language which are believed to be pertinent to the current research will be reviewed.

2.2 Visual Displays and Browsing

Browsing is an explorative and interactive process in which one will scan large amounts of information, perceive or discover information structures or relationships, and select information items through focusing one's visual attention Lin (1997). He further says, in relation to information retrieval, browsing is particularly useful when:

- There is a good organizational structure and related information items are often located near each other
- Users are not familiar with the content of the collection and they need to explore the collection
- Users have less understanding of how the information is organized in the system and they prefer to take a low cognitive load approach to explore the system

- Users have difficulties in articulating their information needs and
- Users look for information that is easier to recognize than to describe

On the other hand, browsing is considered much difficult to be supported by the computer because of its interactivity and its dependence on human perception. Even though most retrieval system designers believe that information systems need to support browsing, it is less clear to them what functions of browsing may be supported and what techniques should be used to support these functions (Lin, 1997). However, through time researchers have explored some techniques to provide browsing for the purpose of information retrieval.

One of the major issues that are addressed by these techniques is the way information is organized and fitted to the computer so that people can still browse and find the information they want. In so doing, these techniques attempt to define a structure that will reduce complexity of information structures and fit a large amount of information to a limited display space. They also create and add some link mechanisms to the information in order to facilitate association and easy navigation. Moreover, these techniques seek a way to present information only at the most appropriate level of details to the user in order to avoid confusion and cognitive overload.

2.2.1 Types of Visual Displays

One of the central issues in organizing information for visualization is the need to decide on what formats and features of visual displays will help to organize large amounts of information (Marchionini, 1995; Lin, 1997). Four types of display formats have been devised with the aim of organizing information and supporting effective use of human visual capabilities. These formats are hierarchical, network, scatter and map displays.

Hierarchical Display: - It is a graphical display that shows data in a hierarchical manner. In so doing, the hierarchy simplifies complex data structures by separating the data into different levels, branches, or clusters (Lin, 1997; Beza-Yates and Ribeiro-Neto, 1999). For instance, while browsing an electronic book, the user will browse the content of the book in a hierarchical manner. That means, a first level of content in the hierarchy could be the chapters, the second level all sections, and so on, and the last level would be the text itself. Also, on the World Wide Web, Yahoo search engines provide a hierarchical directory which can be used for browsing (ibid).

There are certain problems which are inherent to hierarchical displays. First, it is difficult to generate and display large information spaces using hierarchical displays. Second, there is an increased cognitive over load for users who are forced to make selections among the hierarchical branches, especially when the whole hierarchy is not displayed on the screen.

Network Display: - It refers to the graphical display of information. Network displays show both information contents and structures in the form of links and nodes (Lin, 1997). In this type of display, an associative structure to the links and nodes are displayed on the screen and the viewer is allowed to follow the links to browse items represented by the nodes. They are often based on network representational models such as semantic nets.

Network displays can show complex data structures, and they facilitate and encourage visual inference through explicit links on the display. However, complex structures on network displays may confuse and distract the viewer, particularly when the organization of the display is not clear. Various techniques are needed to simplify the display and to create useful structures of the data. Because of the limited display space, network displays often show only a small portion of a network thus making it difficult to show the overall structure and to use spatial information on the displays. Like hierarchical displays, automatic construction of network displays remains a difficult problem. It is even more difficult to automatically identify different types of link and nodes in a network display.

Scatter Display: - It refers to the graphical (dotted) image resulting from mapping high-dimensional data to a two-dimensional visual space (Lin, 1997). The basic elements of scatter displays are dots (or other small individual icons) that represent the mapping data. The displays often show meaningful structures of underlying data. Among the three display formats reviewed, scatter displays most faithfully reflect underlying data structures.

In scatter display, the viewer is not constrained to follow predetermined links as in the network display or to follow rigid hierarchical structure as in the hierarchical display. However, the lack of regularity in the scatter display also poses problems for the viewer trying to discover the underlying structure. To overcome the problem, the scatter display needs the help of other context or interactive probes such as verbal labeling or mouse sensitive areas (when the mouse moves in, some verbal descriptions will pop up).

Map Display: - Come from the idea of applying the geographical map metaphor to the information space (Lin, 1997). The geographical map is clearly the best example of using graphical displays to show large amount of information and their relationships. The geographical map creates a spatial

analog for the physical space based on the survey data of the space. It represents any size of the physical space in a limited display framework and provides various levels of details to support different needs of the user. It keeps dedicated balance among scale, content, and depiction for an effective representation of the physical space. It also uses a set of conventional and elaborate signs and symbols to reveal properties of geographic data such as shapes, locations, and distances.

Adapting such metaphor to information retrieval is much desired by researchers. As quoted by Lin (1997), Doyle (1961) was one of the earliest researchers who articulated the need of semantic maps for retrieval systems. He envisioned a map that could provide a view of the entire library at a distance and help the searcher to narrow his focus by recognition. He believed that such semantic road maps could "increase the mental contact between the searcher and the information store".

2.3 WEBSOM

Different Query-based information retrieval techniques have been developed in the field of Information Retrieval (IR). This techniques help search for a specific piece of information or for information of a specific topic (Lagus, 2000).

However, there have been some basic problems with such traditional search methods. One major problem is the difficulty to devise suitable search expressions, which would neither leave out relevant documents, nor produce long listings of irrelevant hits. Even with a rather clear idea of the desired information it may be difficult to come up with all the suitable key terms and search expression. An even harder problem, for which such classical search methods are usually not even expected to offer much support, is encountered when the idea concerning the object of interest is

vague. The same holds true if the area of interest resides at the outer edges of one's current knowledge (Honkela et al., 1996).Thus, a method of encoding the information based on, for example, semantically homogeneous word categories rather than individual words would be helpful (Salton and McGill, 1983; Rijesbergen, 1996; Honkela et al., 1996).

Since the motives and information needs of users in actual sessions alternate, a combination of visualization, exploration, and search tools and facilities are likely to be more useful in real systems than any single approach alone (Hildreth , 1995;Honkela et al., 1996). To this end, some mechanisms need to be devised which take into consideration the exploration and browsing aspects of information needs. WEBSOM is a tool for organizing documents, which has made an attempt to provide a visual display of documents for easier browsing and exploration.

WEBSOM is an explorative full-text information retrieval and browsing tool that is used to organize collections of text documents onto a visual map display where similar documents are found near each other, like the books on the shelves of a well-organized library (Kohonen, 2002). This organization is achieved automatically using the self-organizing map algorithm, which is used for projecting documents from an initially very high-dimensional space onto a two-dimensional map grid, so that nearby locations on the map contain similar documents. Subsequently the map can be used for visually conveying a general overview of the information about the document collection and for exploring the collection as well (Kohonen, 1995; Honkela, 1996; Kaski et al., 1996; Kohonen et al., 1996b; Lagus et al., 1998; Kaski et al., 1998).

The problem addressed by the WEBSOM method is thus to automatically order, or organize, arbitrary free-form textual document collections to enable their easier browsing and exploration

(Honkela et al., 1996). The method aims at reducing the information overload associated with managing large document collections (Lagus et al., 1998).

The WEBSOM method is readily applicable for organizing any kind of collection of textual documents (Lagus et al., 1998; Kaski et al., 1998). Especially, it is suitable for exploration tasks in which the user may either do not know the domain very well, or have only a limited idea of the contents of the full-text database being examined.

The visual text exploration paradigm, especially the map metaphor is very recent. As a result, neither the possibilities nor the difficulties in navigation of vast text collections of different languages using visual landscapes have been fully explored (Lagus, 2000). Hence, it is the aim of this research to apply the method on Amharic text document retrieval. In the next subsections, the components that form the basic WEBSOM architecture (Figure 2.1) will be discussed in detail.

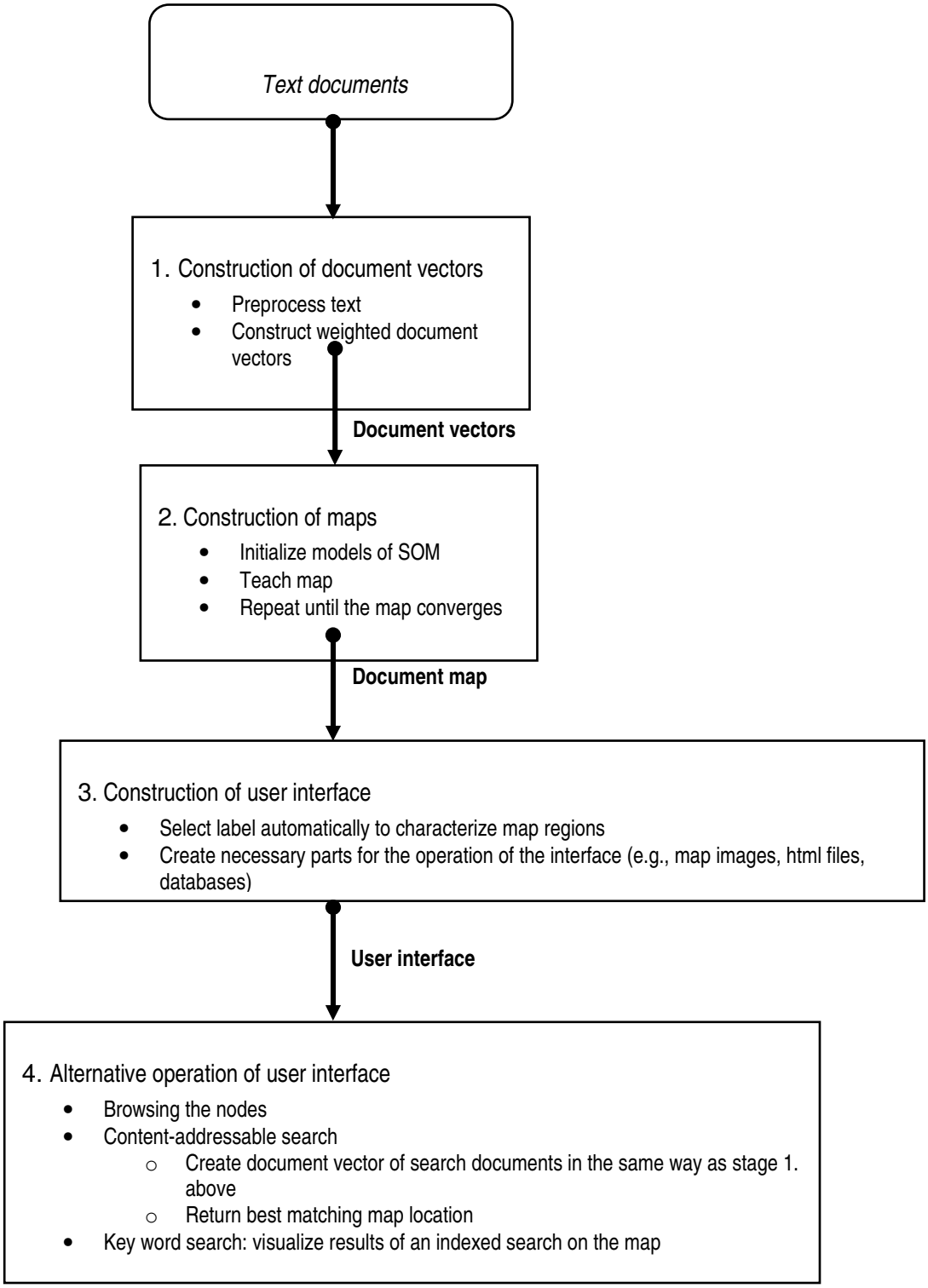


Figure 2.1 The basic architecture of the WEBSOM method (Honkela et al., 1996; Kohonen, 2000).

2.3.1 Document Encoding

Of all the operations required in information retrieval, the most crucial and probably the most difficult task is document encoding (representation). Encoding deals with assigning appropriate terms or identifiers capable of representing the content of the collection items, and sufficient for the retrieval of the document in response to subsequent queries (Salton and McGill, 1983; Croft, 1997).

The representations should be as compact as possible in order to allow efficient processing of large document collections at the same time containing all the information needed in identifying the relevant content.To this end, before applying the SOM algorithm to generate the map, documents need to be represented in a way that the algorithm handles it i.e., to give better results.

Preprocessing

Preprocessing methods are used in representing the data before giving it to the SOM algorithm. This involves the application of sophisticated feature extraction procedure, for which no evident automatic semantic feature exists (Kaski, 1997). A key step in the analysis, feature extraction, involves the choice of suitable representation for the data items (Lin 1997; Lagus 1998). In this section, some document encoding methods, particularly those that are sufficiently similar to the one used in WEBSOM, will be reviewed.

Vector Space Model (VSM):- Both documents and queries are represented as t-dimensional vectors, where each dimension corresponds to an index term. In general, the procedures used in constructing the vector space model can be divided in to three stages. The first stage deals with document indexing where content bearing terms are extracted from the document text. The second

stage deals with weighting of the indexed terms to enhance retrieval of document relevant to the user. The last stage ranks the document with respect to the query according to a similarity measure. Since this procedure is the one that was applied in the current research, each stage will be discussed below in detail.

Before encoding the documents into vector of terms in VSM, the first step is to be done is to clean them by removing the parts that are not considered relevant for the organization. This involves removing some non-textual information like numbers, symbols, punctuation marks etc (Honkela et al, 1996, 1997).

The following steps are required to convert documents to vectors (Salton and McGill, 1983; Baeza-Yates and Riberio-Neto, 1999; Salton and Buckley cited in Lagus, 2000).

- Identifying a list of words from a document collection; it may include every word from the document titles, from the titles and abstracts, or from the full text of the documents.
- Comparing the list to a stop list to delete common words such as "and", "of", "or' etc.
- Using a word-stem procedure to reduce the list to a stem form and remove duplicates
- Removing some of the most and least frequently occurring terms from the list.
- Indexing the collection based on the remaining list. A vector is then created for each document where each component of the vector can be:
    - a binary digit:- each component will be either 1 or 0 depending on whether the corresponding term appeared in the document or not;

- A weight based on the term frequency (the term frequency is the number of times a term appears in a document);
- A weight based on both the normalized[5] term frequency (tf) and the inverse document frequency (idf) (the inverse document frequency is the inverse of the number of documents in which a term occurs).
- Entropy-based weighting of the words: - If the documents can be classified into a set of groups having different topics, the word counts can be further weighted by the information-theoretic entropies (Shannon entropies) of the words.

Choosing a different type of indexing may represent a document space differently. The map displays, based on different types of indexing, may also show different characteristics of the document space (Lin, 1997).

As it was also discussed in chapter one, different ways of similarity measures have been proposed and used by the model such as cosine coefficient, Jaccard, Dice coefficients etc., to compare the query against each document in the collection (Salton and McGill, 1983; Rijesbergen, 1996). However, the cosine coefficient, which measures the angle between the document vector and the query vector in the Euclidian space, is the dominant one (Salton and McGill, 1983; Baeza-Yates and Riberio-Neto, 1999). For instance, if the query $\vec{q}$ is defined as $\vec{q}$ =($w_{1,q}, w_{2,q},..., w_{t,q}$) and a vector of document =($w_{1,j}, w_{2,j},..., w_{t,j}$), where t is the total number of index terms in the system and w is the weight associated with each term, then the similarity between the query and the document can be calculated using the cosine method as:

[5] Normalization is required to reduce the effect of differing document lengths and to moderate the effect of high frequency terms.

$$\text{Sim}(d_j, q) = \frac{\sum_{i=1}^{t} w_{ij} \; x \; w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2 \; x \; \sum_{i=1}^{t} w_{iq}^2}}$$

Proposed by Salton (1982), the model has formed the basis of a large part of IR research (Honkela, 1997). One explanation for this is that vector operations can be performed very fast, and efficient standard algorithms exist to manipulate vectors (Lagus, 2000b). Furthermore, its term-weighting scheme, its partial matching strategy and its ranking technique based on degree of similarity can be considered as the major advantage of the model.

On the other hand, there are two major problems identified with the vector space model. First, the dimensionality of the vectors that represent the documents equals the size of the vocabulary which is immense in large document collections. Therefore, it is practically impossible to use a small proportion of the vocabulary, and some more or less heuristic methods are needed for selecting the set of the most important words. Another problem is due to the orthogonality of the vectors that correspond to the words, semantic relationships of the words are not taken into account. However, in practice, the model remains a popular retrieval model nowadays (Salton and McGill, 1983; Baeza-Yates and Riberio-Neto, 1999; Kaski et al.1998; Salton and Buckley cited in Lagus, 2000).

Latent Semantic Indexing:- Although the vector space model was one of the first models proposed, its variants like latent semantic indexing (LSI) has been introduced with the aim of addressing the high dimensionality problem of the model (Belkin and Croft 1987; Honkela, 1997). The main idea in LSI is to map each document and query vector into a lower dimensional space which is associated with concepts instead of index terms (Baeza-Yates and Riberio-Neto, 1999).

The model tries to take into account the co-occurrence of terms in the documents when encoding the documents.

Random Projection: - In the vector space model discussed above, the documents are encoded as vectors in a very high-dimensional space. Unfortunately, it is not viable to encode the documents in a large document collection using the vector space model as such. The resulting codes would have a very high dimensionality. Hence, one attempt that has been made to obtain a suitable lower-dimensional representation is through random projection or mapping.

For many applications and methods, the central aspect in document representation is the distance between documents. It has turned out that an initially high-dimensional but sparse data space can be projected onto a randomly selected, much lower-dimensional space so that the original distances are nearly preserved. In effect, the exactly orthogonal basis vectors of the original space are replaced by vectors that are with high probability nearly orthogonal, even with randomly chosen directions if the final dimensionality is sufficiently high. An intuitive reason for this perhaps surprising finding is that in very high-dimensional space the number of nearly orthogonal vectors is much larger than the dimensionality of the space.

The advantage of random projection is that it is extremely fast. Furthermore, it can be applied to any high-dimensional vector representation, and any algorithm that relies merely on vector distances, can in principle be applied after the random projection and therefore in a much lower dimensional space (Lagus, 2000).

Random projection has been used for representing words before averaging in, for document encoding in text exploration prior to the application of SOM, as a preprocessing for latent semantic indexing in document representation, and as a preprocessing for SOM in retrieval of Spoken documents (ibid).

2.3.2 SOM

As it was introduced in chapter one, the self -organizing map (SOM)[6] (Kohonen, 1982; Kohonen, 1995; Kohonnen et al., 1996a) is a means for automatically arranging high-dimensional statistical data so that alike inputs are in general mapped close to each other. The algorithm was originally introduced by a Finnish researcher Teuvo Kohonnen and is one of the best known artificial neural network algorithm.

Kohonnen (1995) bases his neural network on the neural properties of the brain, i.e., it is patterned based on the biological ganglia and synapses of the neuron systems. The essential element of the neural network is the neuron. A typical neuron j receives a set of input signals from other connected neurons, x_i , each of which is multiplied by a synaptic weight factor of w_{ij} . All activation weights are then summed to produce the activation level for neuron j. The adjustment of the weights of the nodes of the neural network enable the total network to learn in that a neural network's performance can be adjusted to fit a known set of data characteristics.

In neural networks, in particular, and in machine learning, in general, there are two ways of learning the characteristics of the data: Supervised and Unsupervised.

[6] Also called Kohonen's map

In supervised learning, the class to which a particular data vector belongs must be known before they are given to the training algorithm. Then, a set of training examples is presented, one by one, to the network and the network then calculates outputs based on its current input. The resulting output is next compared with the desired output for that particular input example. The network weights are then adjusted to reduce the error.

In unsupervised learning models, on the other hand, the class of the data vector is not known during the training phase (i.e. the training data does not contain the class information). Instead, the neural algorithm itself tries to find a suitable representation for the data according to certain criteria. That means, based on the network learning rule, weights are adjusted so that input examples are grouped into classes based on their statistical properties. SOM algorithm belongs to the category of unsupervised learning since only after the training procedure has been finished labeling the map units is possible (Lagus et al., 1998).

How the SOM Algorithm works

The SOM maps are trained in an unsupervised manner (i.e. no class information is provided) from a set of high-dimensional sample vectors. The following line of thought may help in gaining an intuitive understanding of how the self-organizing map algorithm works (Lagus, 1998).

Consider an information processing system that must learn to carry out various tasks. Assume that the system may assign different tasks to different sub-units that learn from experience. Each new task is assigned to the unit that can best complete the task. Since they learn from them they become even more competent in those tasks. This is a model of specialization by competitive learning. Furthermore, if the units are interconnected in such a way that also the predefined

neighbors of the unit carrying out a task are allowed to learn some of the task, slowly the system becomes ordered: nearby units have similar abilities and the abilities change slowly and smoothly over the whole system. This is the general principle of the self-organizing map (SOM). The system is called a map and the task is to imitate, i.e., represent the input. The representations become ordered according to their similarity in an unsupervised learning process.

Specifically before discussing how the SOM algorithm constructs the map, in this section some basic and common terminologies in relation with the two-dimensional map, which is the output of the algorithm, are defined.

Neuron (Map Unit):- Neuron (or Map unit) is the basic building block in a SOM. Each neuron is actually a reference vector which contains a certain number of components. The number of components is the dimension of the map and it must also match the dimension of the data that the map processes. All neurons in a given map have exactly the same dimensionality. Neuron is also the general name for the basic calculation unit in many other neural network algorithms. However, the term map unit is only applicable to Self-Organizing Maps.

Reference Vector (Weight Vector):- Reference vector (also known as weight vector) of a neuron is a data vector that encodes the task the neural network is used for. The result of learning is thus encoded in the reference vectors. Each neuron has exactly one reference vector and they all are of the same dimension.

BMU: - It is the short for Best Matching Unit and it is the neuron that is found to be the closest to a given data vector. In other words, a given data vector is given to the map. Then its distance from

each of the map neurons is calculated and the one that is the closest is the best matching unit. BMU plays an important role in the SOM algorithm.

Learning Rate: - Initial learning rate also called ('alpha'). It starts from the value specified in the dialog and decreases to zero during learning.

Function Type: - Determines how the learning rate parameter decreases.

Neighborhood radius: - It determines how the neighborhood structure shrinks during the learning process. There is a start radius, which is the size of the neighborhood in the beginning of training and end radius, which is the size of the neighborhood reached when the training sequence ends.

Topology:-The topological structure of the map determines how many neighboring neurons each single neuron has. Though the most commonly used one is hexagonal, the topology can also be rectangular or irregular (see figure 2.2).

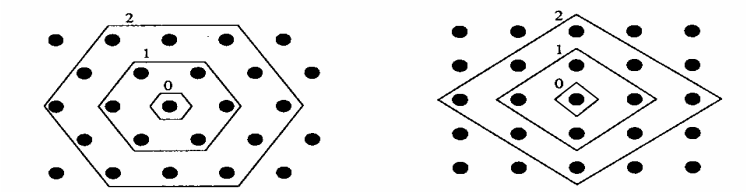


Figure 2.2. Types of toplpgy: Neighbourhoods (0, 1 and 2) of the centre most unit: hexagonal gird on the left, rectangular on the right. The innermost polygon corresponds to 0-, next to the 1- and the outermost to the 2-neighbourhood: Source Vesanto (2000).

The network structure of SOM has two layers: input and output layer. The algorithm takes a set of input objects, which are represented by an N-dimensional vector, and maps them onto nodes of a two-dimensional grid. Each node of the mapping layer also has the same number of features as there are input nodes. Thus, the input layer and each node of the mapping layer can be represented as a vector which contains the number of features of the input.

The network is fully connected in such a way that every mapping node is connected to every input node as shown in figure 2.3. Assume some sample dataset of input variables $\{\varepsilon_j\}$ have to be processed by the SOM algorithm. The set of input samples is defined as a real vector $x = [\varepsilon_1, \varepsilon_2, \varepsilon_d]^T \in \Re^n$ where ε_j is the index sample. Each mapping node on the map contains a model vector (also known as reference vectors) $m_i = [m_{i1}, m_{i2}, \ldots, m_{id}]^T \in \Re^n$, which has the same number of elements as the input vector *x*.

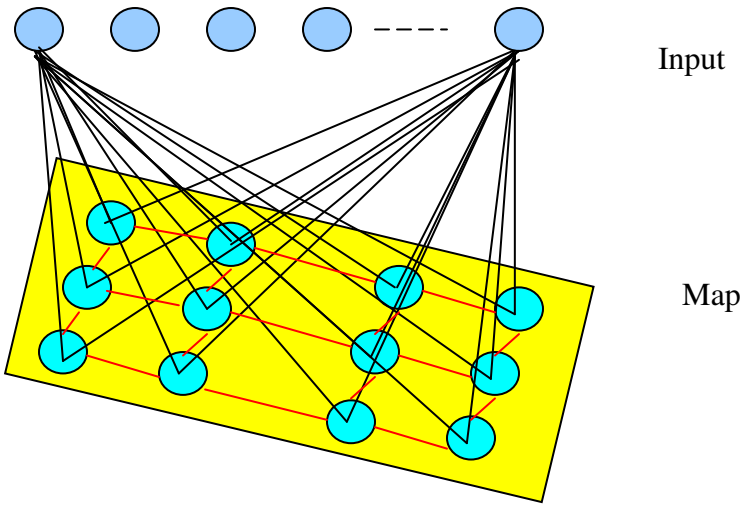


Figure 2.3. Structure of a Self-Organizing Map

Assuming a general distance measure between *x* and m_i, the SOM algorithm then takes the input vector *x* and compares it with the entire model vectors m_i in any metric. In many practical applications, the smallest of the Euclidean distances $\|x - m_i\|$ can be made to define the best-matching node (BMU) (the winner node) on the map, i.e., the node where the model vector is most similar to the input vector.

After the BMU for the input vector has been found, the BMU as well as the neighboring nodes are drawn to the input data vector hit as shown in Figure 2.4. The magnitude of the attraction is governed by the learning rate. As the learning (training) proceeds and new input vectors are given to the map, the learning rate gradually decreases to zero according to the specified learning rate function type. The neighborhood radius decreases along with the learning rate.

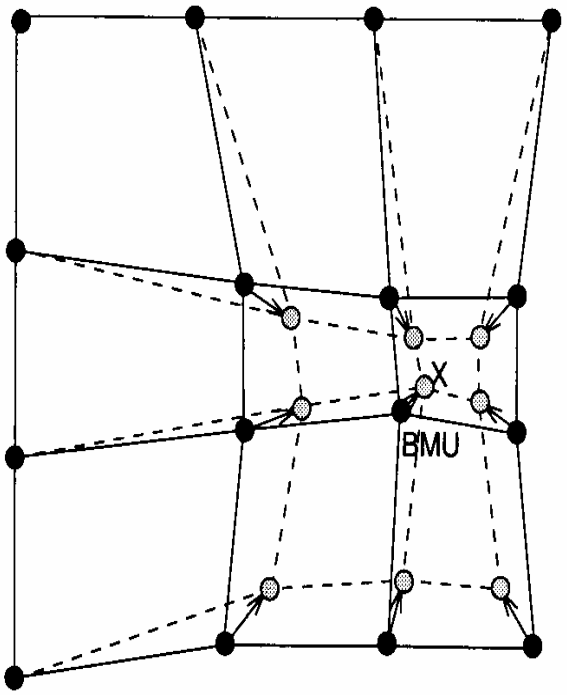


Figure 2.4. Updating the SOM and its neighbours towards the input sample marked with x.. The solid and broken lines correspond to situation before and after updating (Vesanto et al. 2000

In this way, the multi-dimensional (in terms of features) input nodes get mapped to a two-dimensional output grid. After all of the input is processed, usually after hundreds or thousands of repeated presentations, the result should be a spatial organization of the input data organized into clusters of similar (neighboring) regions.

In summary, the mapping procedure is a recursive learning process of the following steps:

- select an input vector randomly from the set of all input vectors
- find the node (which is also represented by an N-dimensional vector called weights) closest to the input vector in the N-dimensional space,
- adjust weights of the node (called the winning node) so that it will more likely be selected again if this input is presented later.
- adjust the weights of those nodes within a neighborhood of the winning node, so that nodes within this neighborhood will have similar weight patterns.

SOM imposes two control mechanisms to ensure map convergence (Lin, 1997). The first is the updating parameter (learning rate), which approaches to zero as the number of iterations increases. The second is the neighborhood structure that shrinks gradually during the process. A large neighborhood will achieve ordering and a small neighbourhood will help to achieve a stable convergence of the map (Kaski, 1997; Kohonen, 1998). By beginning with a large neighborhood and then gradually reducing it to a very small neighbourhood, the feature map achieves both ordering and convergence properties (Lin, 1997).

To finalize, the Self-Organizing Maps are a special architecture of neural networks that combine projections and clustering operations. The algorithm places a set of reference vectors into the input data spaces so that the dataset is approximated by the model vectors. The model vectors are constrained to a two-dimensional regular grid that, by virtue of the learning algorithm, follows the distribution of the data in a non-linear fashion.

Furthermore, the algorithm obtains a clustering of the data onto the reference vectors. The clusters are arranged in a low-dimensional topology that preserves the neighbourhood relations in the high dimensional data. Thus, not only objects that are assigned to one cluster are similar to each other (as in every cluster analysis), but also objects of nearby clusters are expected to be more similar than objects in more distant clusters (Kohonen, 1995, 1996, 2000).

2.3.3 User Interface for document maps

The practical purpose of developing the WEBSOM document map is to provide an interface to a collection by reorganizing documents on their associative relationships so that any one could conveniently explore them. The visualisation is aimed to offer a view of the collection that would help in forming a general view of the domain, as well as to guide exploration towards potentially interesting particular areas (Honkela et al., 1996; Lagus, 2000b). With good interface, the user can interact with the visual display to control or manipulate various views for effective searching and browsing (Lin, 1997).

In addition, a convenient and efficient strategy of moving from general view to specific details and back have been designed, including methods of interaction and a strategy for providing the user

with a sense of location and context (Lagus, 2000b). Suitable visual and textual means for conveying information about the content of the map is also developed (Kohonen, 2000).

Navigation interface

The WEBSOM navigation interface consists of an image of a document map, a hierarchy of zoomed pieces of the map at various zooming levels, and a set of HTML pages, image map files and CGI (Common Gate away) or ASP (Active Server Page) scripts. Zooming-in is achieved by pointing and clicking with the mouse at the desired location on the map image. Horizontal movement to nearby areas as well as panning out is carried out by clicking on a compass image. On the more detailed zoom levels white dots mark units of the regular, map gird. By clicking near a dot the list of documents associated with the map unit is accessed. From the list, individual documents can be selected for reading by clicking on the title, for instance.

Visualising the document map

The WEBSOM document maps are visualised using two methods, a smoothed version of the unified distance matrix or U-matrix and a smoothed document density diagram (Lagus, 2000b). In the U-matrix visualisation, the average distances between neighboring model vectors[7] are represented by shades in a gray scale (or eventually pseudocolor scales might be used). If the average distances of neighbouring m_i is small, a light shade is used; and vice versa, dark shades represent large distances (Kohonen, 1995; Lagus, 2000b).

In contrast, in the density diagram dark color denotes a large number of similar documents and light colour an emptier area. Due to the relationship in SOM between density of model vectors and

[7] also known as reference vectors or codebook vectors

density of documents, both methods visualise the cluster structure to some degree, although U-matrix does this more faithfully (Lagus, 2000).

The smoothed document map landscape carries out two main functions: 1) it describes the document density at each area and 2) it provides texture that can help maintaining a sense of location and context across movements and across dynamic visualizations (Lagus, 2000).

Labeling the map display

Interpretation of a document map display can be aided by labeling the display with a selection of descriptive words that characterise regions of the map. The labels can be utilized for multiple functions: 1) to describe the underlying area, contrasting it against the rest of the map, 2) collectively to summarize aspects of the collection, 3) and in navigation, to act as landmarks or anchor points that help orientation by providing reference points during transitions across views that have different resolutions.

Labeling of the unit map can be carried out in different ways. For instance, if the SOM is used for the classification of entries, for each of which there are several samples available, the problem is to divide the SOM area into non overlapping regions, each of each corresponds to the class. In other words, each map unit is assigned the most probable class label which is determined by first labelling the map units using all the available input samples. Then, the class label is carried out by a majority voting over the labels at each unit (Kohonen, 1995).

Another way of labeling is to label the map after the models have converged to their stationary states, by inputting the entries again and assigning the labels to the winner units. This can be done automatically, since the labels are usually recorded in the data files with the entries.

In case of WEBSOM, a different kind of labeling is needed since each entry contains a number of words. Among the words found in each entry, the most descriptive ones should be selected to describe a cluster area on the SOM. This kind of labeling can be done automatically. According to Kohonen (1995) a good landmark should be a word that occurs often in the documents of that area and rarely elsewhere.

Once the map is labeled, the labels can be used as "landmarks" or "signposts" for starting browsing. In case where several zooming levels are implemented in SOM, fewer landmarks can be made on the top level of WEBSOM and more landmarks can be made to appear in zooming, i.e., to magnified portion of the map (Kohonen, 1995; Lagus, 2000b).

The search facility

In addition to exploration tasks, the WEBSOM may also be used for content-directed document search especially when browsing large document maps since it may be difficult to decide where to start browsing the map. In such cases, a search facility can be implemented to provide suitable starting points for exploration (Kohonen, 2000). The description of interest written by the user—either a whole document or a few words—is encoded (see section 2.3.1) as a document vector and a number of best-matching map units are marked on the display with circles the radius of which convey the goodness of the match. That is, the map nodes close to the position of the new document most likely contain related information and its position on the map also provides a starting point for exploring related documents. It should be noted that this facility does not perform document retrieval, i.e. return the best-matching documents, but only returns the best-matching map units (Lagus, 2000b).

This facility is implemented using client-server architecture: A search server holds the map reference vectors in memory and when a search is initiated, a client program encodes the query as a document vector, passes it to the server and requests for a number of the best-matching units. Upon receiving the results the client draws them on existing static map images, constructs an appropriate HTML page and returns it to the WWW browser.

2.3.4 Choosing a good map

The stochastic-based SOM learning process allows some variations in the learning results (Kaski, 1997). Thus, to ensure good quality map several maps may need to be computed and the best map chosen according to some cost function specific to the size and topology of the map. Note that this limits the possibilities to use a cost function to compare maps of different sizes or neighborhood function (Kaski, 1997; Lagus, 2000b). The value of the cost function is inversely proportional to the size of the map and increases as the width of the neighbouring function increases.

2.3.5 Interpretation, evaluation and use of the maps

Interpretation*:* Some general methodology may aid in the interpretation of maps (Kaski, 1997; Lagus, 2000b), although the interpretation of the map should be predominately local (because SOM attempts above all to preserve local structures) based on the local relations of the data items on the map. Different properties of the reference vectors and the data items can be visualised on the map display to help in the interpretation.

Another method that helps in the interpretation provided that some external information like class labels are available, is to plot the labels on the organised map. Lagus (2000b) utilized a measure

based on an external topical classification of documents, best described as the purity of map nodes. He defines purity as the proportion of documents that fall into a map unit where their own class form a majority for evaluating document maps. The classification accuracy then indicates how well the classes are separated on the map, and the classification accuracy of new samples measure the generalizability of the results. If the distributions of the known classes are overlapping, such displays can even be used to explore the degree of overlap in different types of samples. It may then be possible to gain insight to whether the classes actually co-exist or whether new kinds of features should be added to the data items to make the classes easily separable.

Evaluation*:* Although good evaluation methods for measuring the quality of visualisation, exploration and navigation are very difficult to define (Lagus, 2000b), the quality of the map display may generally be evaluated by an expert in the application area (Kaski, 1997). User studies may also be required until more direct, automatically applicable measures are determined (Lagus, 2000b). If samples having known classes are available, it is potentially useful to try to classify the samples using the map. Each map unit is labelled according to a majority voting of the samples, after which all samples that are projected into a unit (node) are classified according to its label. The generalizability of the results could also give some indication of the quality of the mapping. Generalizability could be measured as the sensitivity of the map to small variations in the input data, caused for instance by adding artificial noise.

Use of the organized map*:* The illustrations formed by the SOM can be used as tools for gaining insight into a dataset. They can also be used to summarise data sets, together with explorative research, or even as a decision support system (Kaski, 1997).

2.3.6. Adding new documents

New documents can be inserted onto an existing map simply by locating the best-matching map unit for each document (Lagus, 2000b). However, the map may not be such a good representation of the collection in a non-stationary (dynamic) document collection where new topic areas and terms are introduced after a while. An intuitive reason for this is that in a very high-dimensional and very sparse space an unseen document will not be near any area of the two-dimensional map unless its own topical domain is discussed by documents that contributed to the map construction. In a non-stationary collection the map should therefore either be incrementally adapted or fully re-calculated after a time (Lagus, 2000b).

2.4 Application of SOM

The SOM is one of the most widely used neural network algorithms. Studies in which SOM has been used or analyzed have been reported in over 4000 scientific articles (Lagus et al., 1998; Lagus, 2000). Most of the early applications were in the fields of engineering (e.g. image recognition, signal processing), but nowadays a very diverse range of applications is covered, from medicine and biology to economics. Several recent studies also adopted the SOM approach to textual analysis and classification. Overviews of the applications are given in (Kohonen, 1995; Kohonen et al., 1996c).

The usefulness of the SOM stems from its two properties: first it creates models or abstractions of different types of data in the dataset, and second it organizes the abstractions onto a usually two-

dimensional lattice which can be used to generate an illustrative graphical display of the dataset (Kohonen, 1995; Lagus etal, 1998; Lagus, 2000).

2.5 Related Research Works

In support of using Kohonnen map for textual document classification, Lin, Soergi and Marchionini (1997) used the Kohonnen SOM for classifying documents for information retrieval. Documents are represented as vectors of binary values. Each coordinates of the vector represents a specific term or phrase with the value set to "1" if the term or phrase is found within the document and "0" otherwise. After several passes through the input file (a collection of similar documents), the Kohonen layer is trained. The resulting map provides an intuitively appealing organization of the input data. The documents are classified according to their content and conceptual regions are formed and named on a two-dimensional grid.

Lin (1997) also made an effort towards applying map displays for information retrieval. He pointed out some of the advantages that can be obtained by applying visual displays for information retrieval through a thorough examination of relationships of among visual displays, information retrieval, and browsing. The ability to convey a large amount of information in a limited space, the potential to reveal semantic relationships of terms and documents as well as the facilitation of browsing and perceptual inferences on retrieval interfaces are the merits of visual displays.

To strengthen those merits, he further demonstrated three map displays that have been generated by a neural network's self organizing algorithm. The first map was constructed using 311 set of multilingual information documents. The documents were indexed based on the general vector space model, where potential index terms were identified based on only the titles. A vector of 85

dimensions was created for each document and the relative importance of terms in each document was considered by applying binary weighting technique. Finally, the document vectors were used as an input to train the feature map of 85 inputs and a 10 by 14 output nodes arranged in a grid. After several iteration of training, the map was constructed that showed the content of the collection roughly in three parts, the languages on the left, the technologies in the middle, and the tools on the right.

The second map was constructed based on 660 personal document collections, which were accumulated over many years as a by-product of a researcher's research activities. Like the first map, the vector space model was applied for indexing the documents where the indexing for this collection was based on full text, i.e. every word in the titles, keywords, and abstracts were used. After following each steps of the VSM, 1472 unique terms were identified and weights of each term were computed based on both the term frequency and the inverse document frequency. The 660 document vectors of 1472 dimensions were then used as input to train a 10 by 14 Kohonen's feature map. The final map display successfully clustered the researcher's major research areas and the relationships of these areas.

The third example is about 143 documents from 1990-1993 SIGIR conference proceedings. The indexing terms for this collection were collected from titles only, but the weights of terms were computed based on the term frequency in titles, keywords, and abstracts. Such indexing uses the same low dimensions as the title indexing, but the indexing vectors reflect how the indexing terms are distributed in titles, keywords, and abstracts. After the same stopword removing, stemming procedures, and elimination of the most-and-the least-frequently occurring terms, 154 terms were used to index the collection, resulting in 143 vectors of 154 dimensions. These vectors were then used to train a 14 by 14 feature map of 154 input features.

As it is pointed out in the abovementioned paragraphs, all applied the Vector Space Model for document encoding. Many studies revealed that the vector space model of Salton is very efficient for small document collections (Kohonen, 1995; Lagus et al, 1998; Lagus, 2000). In large document collections, however, the vocabulary is very large. Since each word in the vocabulary requires one dimension in the document vector, the resulting vectors may have hundreds of thousands of dimensions. Computing with large amounts of such vectors is, of course, not computationally feasible, and therefore the dimensionality of the document vectors needs to be reduced before further processing.

To this end, Lagus et al (1998) encoded the documents based on the traditional vector space model. The importance of the potential identifiers (index terms) was considered by weighting using Shannon's entropy-based weighting technique. Then, they reduced the large dimension into smaller by applying a random mapping technique. In this way, 10000 documents were encoded and given to the SOM for constructing the map. Furthermore, they introduced Latent semantic Indexing technique as an alternative for reducing the dimension.

The WEBSOM method has been developed by a team of several people since the onset of the project in 1995. In particular, two doctoral theses (Lagus (2000) and (Kaski (1997)) have been published that partially consists of research on the method. One effort that has been made to apply the method in practical environment here in Ethiopia was by Mulegeta (2002). Furthermore, experiments on WEBSOM have been done on documents of various languages. For instance in addition to English documents and articles, the method was also applied on Finnish news articles (Lagus, 1998) and News article of the Chinese (Lee et al, 2000).

To summarize, experiments with the various datasets showed that the method can be successfully applied to organizing both very small and very large collections ranging, for instance, from few hundreds to about seven million (Lagus, 2000).

2.6. Amharic Writing system and Its Characteristics

Amharic, or Amarënya, was the national language of Ethiopia until 1983 E.C. Currently it is the official language of the Federal Government of Ethiopia. Moreover, it is the working language of different governmental and non-governmental organizations through out the country. Mass Medias like radio, television broadcasts and the press are also using it for disseminating information to the public.

As a result of its wide application, large Amharic documents are compiled both in hard copy and electronic forms. Like any documents of other language, the contents or meanings of these documents are represented using important features[8] of the language.

For the purpose of this research since Amharic documents are considered, it is important to investigate these potential features or terms that are the capability of representing the contents of the documents, which in turn demands one to understand the characteristics of the language in particular. Hence, under this section, important features of the Amharic language that are believed to be pertinent to the current research will be reviewed.

2.6.1. History of the Amharic language

[8] Features also called terms are words and /or phrases and are obtained by analyzing the document.

Being a Semitic language of the afro-Asiatic language group, this language is related to Hebrew, Arabic, and Syrian. The current Amharic writing system was adopted from the Ge'ez writing system, which was the classical language of the Axum Empire of Northern Ethiopia (Bender et al., 1976). It existed between the 1st Century A.D. and the 6th Century A.D. The ancient Sabaean script is in turn attributed as the source of the Ge'ez script. As the Sabean script descended into Ge'ez and later into Amharic, the numbers of symbols in its original Sabean script and their shapes have been changed.

When the power base of Ethiopia shifted from Axum to Amhara between the 10th and 12th Century A.D., the use of the Amharic language spread its influence, hence became the national language of the country until 1983 E.C.

2.6.2. The Amharic Writing System

Amharic uses its own alphabets, numbers, punctuation marks etc., for its writing system.

The Amharic Alphabets: Amharic is a syllabic-alphabetic language which has 33 basic characters each having 7 forms or orders for each consonant-vowel combination (Bender et al., 1976; Hudson, 2001). Each character occurs in one basic form and in six other forms which is made in accordance with the sound that goes with the symbol. The non-basic forms are derived from the basic forms by following regular modifications For instance, the seven orders of the consonants of ሀ, ሰ and ሠ can be written as:

ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ
hä	hu	hi	hä	he	h	ho

ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ
lä	lu	li	lä	le	l	lo
ሠ	ሡ	ሢ	ሣ	ሤ	ሥ	ሦ
sä	su	si	sä	se	s	so

In the above examples, each symbol represents a consonant together with its vowel. The vowels are fused to the consonant form in the form of diacritic (accent) markings (Bethlehem, 2002). The diacritic markings are strokes attached to the base characters to change their order. For instance, the above ሀ (ha) is changed into the second order ሁ (hu) by attaching " ◌ُ " and into the third order ሂ (hi) by attaching the marking " ◌ِ " and so on. In this way, the 33 core characters yield 231 distinct symbols.

The above transformation of the base form into the non-basic forms indicates that the Amharic writing system does not use independent symbols for vowels in representing a syllable. As Bender (1976) explains, this is a characterization known as syllabic. However, currently there is a debate whether the language is actually syllabic or alphabetic (Baye, 1997; Hundson, 2001). Alphabetic writing systems are systems that present the consonants and the vowels separately such as the English and Greek language. On the other hand, syllabic writing systems are systems that combine both the consonant and the vowel together (e.g. Amharic writing system). However, Baye (1997) argues that Amharic is alphabetic on the grounds that each symbol can be broken down into consonant and vowel phonemes which can be independently represented by separate symbols. In fact, he describes the Amharic script in terms of 27 consonant and 7 vowel phonemes.

In addition to the 231 basic characters, there are also four labio-vellars each having five orders and eighteen additional labialized consonants. See appendix 1 for the complete list of the Amharic characters.

Amharic Number Systems: Except for zero the Amharic language has its own ways of representing numbers, which are derived from Greek letters as described by Bender (1976). The numbers consists of a single character for one to ten, for multiples of ten (twenty to ninety), hundred and thousand symbols having a horizontal stroke above and below each character (appendix 1**).** These numerals are largely used in writing dates and page numbers in text. Also it is widely applied in the environment of Ethiopian Orthodox Church. Since there is no way of representing numbers greater than 10,000 and for 0 as well, Arabic numerals are used in many Amharic texts however.

Punctuation Marks: In Amharic, there are around 17 punctuation marks (Beletu, 1982). In this section only some of the most commonly used punctuation marks that are used both in handwritten and computer written text will be reviewed.

Two dots (**:** Hulet netib) similar to a colon are used to separate words, though in languages such as Amharic blank spaces are generally used instead. A full stop or period is four dots (**::** Arat netib) and a comma is two dots with horizontal lines over them (፣). An equivalent for semi-colon is two dots with horizontal bar above and below them (፤ dirib serez)**.** Others include borrowed symbols like **?**, **!**, ", ", ', **/**, ****, etc.

As far as the application of these punctuation marks is concerned, the word delimiter (two dots) is mostly used in handwritten text but it is becoming a common practice to exclude it from computer

written text. Rather space is being used as word separator. In case of sentence delimiter, the four dots continue to be used. The remaining punctuation marks are used where appropriate.

2.6.3. Characteristics (features) of the Amharic Writing System

As it was discussed in many literatures, the Amharic writing system has many features, which may cause some problem from the perspective of computation (Getachew, 1967; Bender, 1976).The next section deals some of them.

Consonants with the same sound: At the time of borrowing its script from Geez, Amharic did not select consonants which are only important to its writing system. As a result, in Amharic writing system, there has been found different symbols with the same pronunciation and meaning (i.e., in Geez those symbols are different in meaning as well as in spelling, which is not the case for Amharic) and they have been used interchangeably (Getachew, 1967; Bender et al., 1976). As Getachew (1967) noted, however, for the case of Amharic there is no defined rule that differentiates their proper usage.

In Amharic, these consonants with the same sound falls into two categories: (1) the first and the fourth order alphabets of the same base form having the same sound and (2) different alphabets with the same sound.

For the first case, for instance, it is not clear whether one should write "ሀይማኖት " (`religion) and "ሃይማኖት" since both "ሀ" and "ሃ" have the same sound. Those alphabets that exhibit such characteristics are listed in table 2.1.

1st order	4th order

ሀ (hä)	ሃ (hä)
ሐ (hä)	ሓ (hä)
ኀ (hä)	ኃ (hä)
አ (ä)	ኣ (ä)
ዐ (ä)	ዓ (ä)

Table 2.1 different forms of the base alphabet with the same sound

Similarly, table 2.2 shows lists of different alphabets that have the same meaning and sound. Here, not only the base forms listed have the same sound but also all the corresponding orders (6 orders) of them have the same sound too. For example, writing "ሰማይ" and "ሠማይ" to mean "the sky" does not make difference in meaning even though "ሰ" and "ሠ " are used interchangeably. The same holds true for "አይን" (eye) and " ዐይን" although "አ" and "ዐ" are two different alphabets with the same sound.

Alphabet	Other alphabet with the same sound
ሀ (hä)	ሐ,ኀ
ሰ (sä)	ሠ
አ (ä)	ዐ
ጸ (tsä)	ፀ

Table 2.2. Different alphabets having the same sound

Moreover, a complex case comes when the same word appears to be in many forms (more than two forms) by using interchangeably these alphabets having the same sound. We can take " ገብረስላሤ ", " ገብረሥላሴ", " ገብረሥላሤ " and " ገብረስላሴ " as a good example, which refers to the name of a person (Gebresilase). As all the above discussion indicates, there arises some confusion and inconsistencies in Amharic alphabet and as a result these redundant consonants add their contribution to make the vocabulary to be large.

Different forms of writing compound Nouns: In Amharic writing system, there also exist different ways of writing compound words with out affecting their meaning (Bender and Ferguson, 1964) as cited by Zelalem (2001) and Bethlehem (2002). That means, at one time the compound noun can be written as two separate words and at another time as single word. For instance, it makes no difference in meaning at all while writing the compound word "ወጥ ቤት" as one word "ወጥቤት" which is to mean that "Kitchen". Additional examples of such Nouns are mentioned in table 2.3.

Compound Noun as two separate words	Compound Noun as single word	Its meaning in English
ማዕድ ቤት	ማዕድቤት	Dining room
ብርድ ልብስ	ብርድልብስ	Blanket
ብረት ድስት	ብረትድስት	Cooking pot (metallic)
ቤተ መቅደስ	ቤተ መቅደስ	Temple
ቤተ መዘክር	ቤተ መዘክር	Museum

Table 2..3 some examples of writing compound nouns in different ways adapted from Bethlehem (2002)

Different ways of writing the same word: This is a problem that exists when the language has some words having different forms of writing system. In Amharic the word "äytoäl" (he has seen) may be spelled as "**አይታ᎓ል**", "**አይቶአል**" or " **አይትዋል**", which are different variants of the same word.

This problem is also a common issue while translating some foreign words into Amharic, which varies as the number of possible pronunciations (Getachew, 1967). "**ዬሬክተር**", "**ዳይሬክተር**" and "**ዲሬክተር**" which is to mean that "director", "**ኤሌክትሪክ**" and "**ኤሌትሪክ**" which is to mean that "electricity" are some of the examples.

Different forms of writing Abbreviation: In Amharic, it is also found that there is no consistency while spelling abbreviations. For instance, the phrase "**ዓመተ ምህረት**" can be abbreviated as "**ዓ.ም**", "**ዓም**" and "**ዓ/ም**". Similarly, the use of hyphen is also not consistent. The same word "**ዓመተ ምህረት**" can also be written as "**ዓመተ-ምህረት**". Hence there should be a mechanism to handle these problems while representing Amharic documents.

Chapter Three

Design and Development of the Prototype System

3.1 Introduction

The development of SOM map using Nenet along with a design of prototype interface is presented in this chapter. Discussion on the results of the experiment on three classes of Amharic news articles is also made. The general procedures used in the development of the map are shown in the figure below.

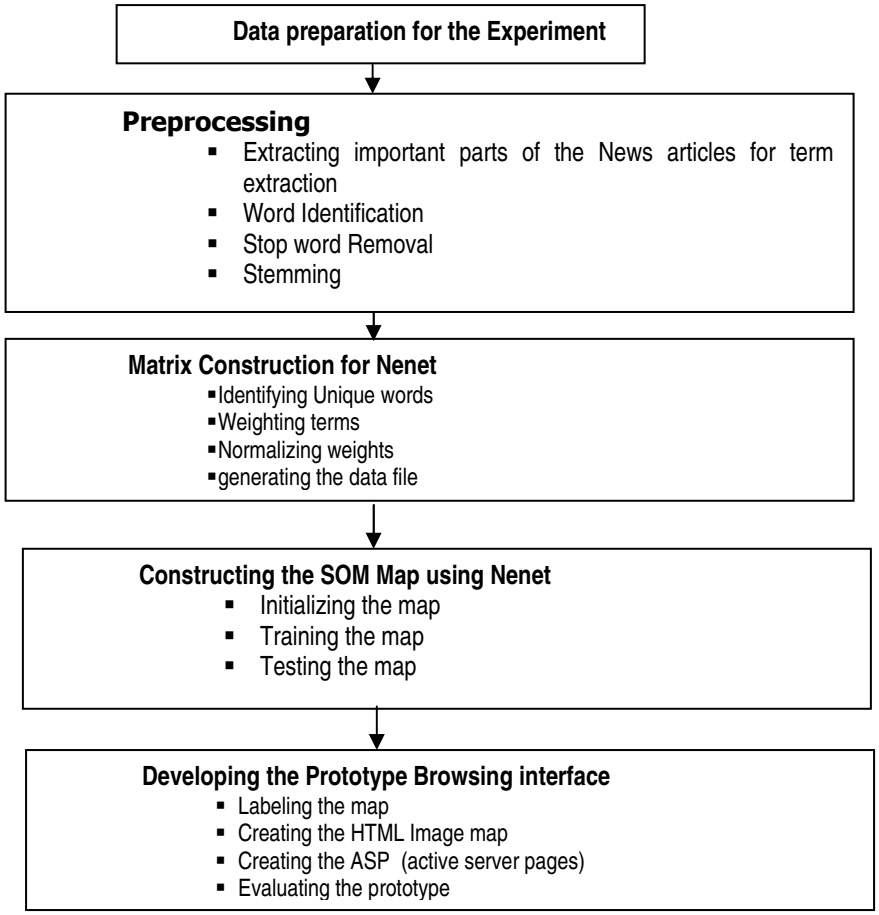


Figure 3.1 Self-organized Map developments.

3.2 Data preparation

To address the current research problem under consideration, Amharic News documents were collected from Ethiopian News Agency (ENA), which is the first news agency that has been remained as a news source for half a century for mass media like Radio and *Television* of the country.

ENA news articles report on several agendas that are classified hierarchically. This hierarchical classification scheme is adapted from that used by Reute's and has sixteen parent classes.

To undertake the current research, samples of news articles from three classes were selected randomly: Agriculture, Sport and Accident. The details of the processes involved in selecting the actual number of the articles are discussed later in this chapter.

The structure of the ENA news articles is that a news item has four main parts. These parts are

Header: the header part of an article consists of a classification code, slug, Author's name and a deadline. The classification code indicates the class topic the news belongs to. The slug on the other hand is a general identification of the subject in the form of a generic master slug, which may be followed by a slug specific to that story. Dateline gives information about date and place of story's origin and Agency's acronym.

Headline: the headline part of a news article gives the main content of story in a few crisp words to catch the reader's interest. This part seldom exceeds a single line. The lead is the opening paragraph and would contain no more than 30 words. It captures the essence of a situation event clearly, and if possible, dramatically. The last part of the news articles is the body that elaborates

on the lead and provides any necessary details.

The Headline and the Slug part of the news articles were considered very important for term extraction. This was done so because these fields retain the main contents of the story. The full story only narrates the content of the headline by giving description about the where, when and why of the story.

The data preparation process started by separating the Amharic news articles from English articles since the agency stores both in the same database (Sql Server). For this study, these news articles have been imported to MS Access. The articles in the SQL server were stored in a table consisting of the fields: ID, NewsID, HeadLine, Slug, Keyword,NewsDate, ClassificationCode, FullStory, Date and others too. The fields imported to MS Access are, however, ID, NewsID, HeadLine, Slug, ClassificationCode, FullStory and Date only as they are the parts necessary for the research. Using the Access query analyzer, the Amharic articles were separated from the English by writing filtering criteria in the NewsID field. Then, another query was made on the Amharic database to separate the three classes considered for this study using their classification code as a criteria generating three tables.

The Amharic news articles that were considered are stored in a file with rich text format (rtf) where many formatting characters are stored along with the text of the news inside the files. Extracting terms from these articles without removing these formatting characters thus would result in the construction of noisy data since the introduction of these formatting characters may cause some symbols to be changed. To alleviate such problems, the news articles from the Access database were exported and loaded to an rtf control that extracted each term without the formatting characters (exported to NotePad and saved as a plain text file, which removes automatically all the

formatting characters from the text).

The analysis made on the database indicated that there are news articles that have missing values. That is, some of them do not have headline, slug and keyword or a combination of them. Others have meaningless values in their fields (for instance, the news article with Identification number ኢዜአ10350 has a value "ቀ1ሀ3ቷ4ደ6ነ7 ሀ1ፀ4መ5 ጸ6ደ7 ሀ1ነ5ተ8ጸ1 ቈ4ደ3ሀ9" under its full story). Hence such articles were dropped out. Spelling errors were also checked and edited manually to avoid misinterpretation and negative effect on the dimension of the document space.

As it was discussed in section 2.6.3 of chapter two, there are characters with the same sound and meaning but having different writing forms. For the purpose of document representation, these characters should be treated similarly. An algorithm was developed that reduces all different symbols of the same sound and meaning to achieve the equivalency wanted. Thus, the algorithm developed converts ሐ, ኀ, ኃ, ሐ, ሃ and ኻ (having the sound h) into ሀ. In similar manner, all the seven forms of ሰ and ሠ (having the sound s) were changed to ሰ, all the seven forms of አ and ዐ (having the sound a) were changed to አ and finally all forms of ፀ and ጸ (having the sound tse) were changed to ፀ.

The algorithm works as follows:

- Read the character
- if the character identified is any of ሐ, ኀ, ኃ, ሐ, ሃ ኻ or any other order of them then
  - Change in to ሀ
- Else if it is ሠ or any other order of it

- Change it to ሰ

- Else if it is ፀ or any other order of it
  - Change it to ጸ
- If the character that follows is a diacritic marking, attach it to the changed base character

After all the necessary cleaning was made on the three classes, a total of 330 samples of articles were placed in a new table in the database. Table 3.1.summarizes the type of classes and the total number of articles considered in this study. The total size of the news articles were decided based on other related researches done so far in the area of the application of SOM in clustering documents.

No	Type of class	Description	Total Number of articles
1	ግብጉ Agr	ግብርና ጉዳዮች Agriculture	80
2	ስፖር Spo	ስፖርት Sports	146
5	ጋደአ Acc	አደጋዎች Accidents	104
		Total News articles	330

Table 3.1. Classes considered with the corresponding number of news articles

Among the 330 articles, 75% (248) were assigned as a training set and the remaining 25% (82) as a test set, giving due attention to the proportion of each class. Table 3.2 shows the details of the number of news articles considered for each class as training and test sets.

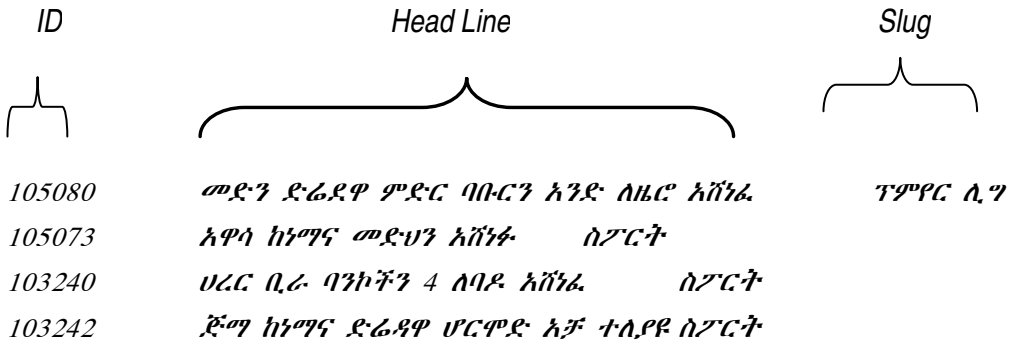
Type of Class	Training set	Test set	Total
ግብጉ Agr	60	20	80
ስፖር Spo	110	36	146
ጋደአ Acc	78	26	104

Table 3.2. Number of training set and test set considered from each class

3.3. Preprocessing

The preprocessing phase involves selection of relevant words (representative words) from the contents of each article. To do this, words were first identified; those words that exist in stopword list were removed and the words were reduced to their corresponding stems.

Word identification: To select terms from the headline and slug part of the news articles, the Identification number (ID), Headline and slug of each news articles were placed in one line in a document editor (Notepad) separated by tab delimiter (see figure 3.1).



.

Figure 3.2 Formats of the dataset in a text editor

Identification number (ID), Headline and slug of each news articles were placed in NotePad file were again placed each in a separate file named after the ID of the articles. The code for creating these files was written using Delphi programming language. The algorithm is as follows:

- Initialize a variable
- read the character and hold it in the variable until it gets the tab delimiter
- if it gets the tab delimiter
- Assign all the characters (ID) in the variable as a file name
- then until it reaches the first line
    - read all the remaining strings
    - write on the file
- If end of file is reached then
    - Exit
- Else go to the first step

In this way, 330 files were created for each news articles. After writing each news articles in its own file, terms that potentially represent the news articles have been extracted. To extract words or terms, space and punctuation marks that end words in the language are taken as delimiters.

In Amharic, there are around 17 punctuation marks (Beletu, 1982). In this section only those punctuation marks supported by VG2 are considered. Words may be separated using the space or two dots (**:** Hulet netib). However in computer written texts it has been a common practice to use space. Both options were taken into account to handle possible occurrences of anyone of them**.** A full stop or period is four dots (**::** Arat netib) and a comma is two dots with horizontal lines over them (፣). An equivalent for semi-colon is two dots with horizontal bar above and below them (፤ dirib serez)**.** Others delimiters in VG2 are borrowed symbols like **?**, **!**, ", ", ', **/**, ****, etc.

Using the delimiters discussed above as word separator, a Delphi programming code was written using an algorithm that was developed by Zelalem (2001) for term extraction from each file. The general procedure of the algorithm that was applied can be summarized as follows:

- Initialize the variable to hold the word from each file
- Read a character from a sentence (document)
- Check the character against Amharic word delimiters list
- If the character is any one of the delimiters (the characters you read so far make up a word (i.e. the word variable now contains a complete word))
- and if the character length is above one character
    - check it whether it consists numeric character
- If no numeric character is found in the word, consider the word as a valid Amharic word

- If end of file (no more characters to read) then
    - Exit
- Else go to the first step

Stop Word Removal: After a valid Amharic word was identified, it was compared against a stop word dictionary (the stop word dictionary contains the lists of words that do not contribute in content representation of documents and are common to all documents) to avoid non--content bearing terms. Since there is no standard list of stopwords for Amharic documents, 750 stop words used by Zelalem (2001), and 80 additional words added by the current researcher were applied for the current study.

The stop word lists used in the study were identified in two ways. First, stop words that are common to the Amharic language were identified and later stop words that are domain (News) specific were considered. The news specific stop words were identified from the very fact that reporters and Journalists use vocabularies that are peculiar to their profession at the time of reporting. For instance, **አስታወቀ**, **አስታወቁ**, **አስታውቀዋል፡ አካሄዱ,ተካሄደ፡ ተባለ**: **ዘገበ**, **ዘግቧል**, **ዘገባ** etc., are words that are commonly used by most of the journalists.

Word Stemming: One problem faced in the use of free text for indexing and retrieval is word variation which occurs in languages. Morphological variants of words have similar semantic interpretations and for the purpose of IR applications they are considered to be equivalent. One way to alleviate such variations of words is to use a stemming technique, a computational procedure that is designed to bring together words that are semantically related, and to reduce them to a single form for retrieval purpose.

Stemming is not only a means that different variants of a term can be conflated to a single representative form, but also reduces the dictionary size. That is, the number of distinct terms needed for representing a set of documents. A smaller dictionary size in turn results in saving of storage space and processing time.

In Amharic, there are many inflectional and derivational affixes to a noun (Bender, 1976). To address the word variations, in this research depularaization and prefix/suffix stripping technique was used.

As it was noted by Rijesbergen (1996), one way of reducing morphological variants of a word to its root is via suffix stripping. One standard approach that was mentioned by Rijesbergen is to identify a complete list of suffixes and to remove the longest possible one.

The plural form of a noun in Amharic is expressed by attaching the suffix "ዎች" (woch) to singular nouns. Other possible suffixes that can be attached to nouns include "ን","ም", and "ና", where their attachment to a noun indicates possession, emphasis, and object marker respectively (Abiyot, 2000).

A given noun can also have a chance to take one or more of these suffixes in a number of ways. One possible example can be shown by combing the suffix "ች" (wöch) and the above three suffix in table 3.4. For the purpose of this study, the suffix lists made by Abiyot (2000) and Zelalem (2001) were used (see appendix 3 for the complete lists of suffix used in the study).

Suffixes Combined	**Suffix used**
ች + ን	ችን

ች + ን + ና	ችንና
ች + ን + ም	ችንም
ች + ና	ችና
ች + ና + ም	ችናም
ች + ም	ችም
ች + ም+ ና	ችምና

Table 3.3 Examples of the possible combination of suffixes from "ች" "ን","ም", and "ና" adapted from Zelalem (2001).

For removing suffixes from a given Amharic word, first the possible suffix lists were put in a suffix dictionary. Then, the suffixes that were attached with a given word were stripped off by the algorithm and compared against the lists in the dictionary. The algorithm does this task iteratively as follows:

- It checks if the maximum length suffix exists in the word by stripping the right most five characters. That is, if the length of the word is greater than six, then strip the right most five characters and check if it exists in the suffix dictionary (database). If it hits, then report the word without the suffix.

- Else, if the length of the word is greater than five, then strip the right most four characters and check if it exists in the suffix dictionary. If it exists then report the word without the suffix

- Else, if the length of the word is greater than four, then strip the right most three characters and check if it exists in the suffix dictionary. If it exists then report the

word without the suffix

- Else, if the length of the word is greater than three, then strip the right most two characters and check if it exists in the suffix dictionary. if it exists then report the word without the suffix

- Else, if the length of the word is greater than two, then strip the right most one character and check if it exists in the suffix dictionary. if it exists then report the word without the suffix

The algorithm in general starts stripping the longest suffix from the word and searching for it in the suffix database. If it is not found in the database, then it strips the next longest suffix etc., iteratively until the suffix is found in the list. Finally, when the suffix that was stripped is found in the database, it reports the remaining word without the suffix.

The most commonly used Amharic prefixes were also stripped off from the words to bring different forms of the same word to its common form. The single character prefixes that occur mostly in Amharic text "በ", "ለ","ከ", "የ" and four additional prefixes "እንደ", "ስለ", “እየ” and “እስከ" were removed.

To remove one character prefixes the following algorithm was used.

- Checks whether the length of the word is greater than two. It also checks whether the second character is not the diacritic form of the prefix. That means even though the first character found is "በ", "ለ","ከ", and "የ" it doesn't necessarily mean it is considered as a prefix. The character may appear in one of the six forms of these characters.

Hence, it is the second character that determines whether the first character is a prefix or not.

- If it is not, then it checks also if there is a diacritic in the remaining string. That is, if there is a diacritic in the remaining string, the length should be greater than two
- Else the length should be greater than one.
- If the length satisfies the requirement then strip the prefix and report the word.

After the word was stemmed and reduced to its common form, a text file with a file name "wordList.txt" was created for holding words identified in a given file with its frequency. The algorithm:

- writes the file name of individual articles as a heading
- writes the word under the file name (news article identifier) with a frequency count of 1
- If similar word is found in the list

  appends its frequency
- If end of file reached

  Exit
- Else if more files to process
- go to the first step.

In this way, for every file (news article), the words identified in it with the corresponding frequency were written in the file "wordList.txt ". The ID number which is the file name for every article was assigned as a heading and lists of words identified were written under it. This was done iteratively for the training and test set.

3.4. Text-Term Matrix Construction

Nenet, the tool that was used in this study, demands the data to be prepared in a specific matrix format (The detail of the tool and the data format is presented in section of 3.5.1). After the words in the stop word and prefix-suffix dictionary were removed, the next step performed was constructing the news articles - term matrix. This by itself involves identifying the unique words in the entire dataset, term weighting, normalizing the weights and finally generating the data file in the required format.

Identifying unique words in the collection for matrix construction: So far, words in given news article were identified with their corresponding frequencies (frequencies in each document) and stored in a file called "wordList.txt". However, it is necessary to identify unique words in the entire collection. A new text file called "Uniqueword.txt" was created to hold the unique words with their frequencies in the entire collection. The algorithm developed to handle this task:

- opens the file "WordList.txt"
- until it reaches end of the file
- reads a word from the file
- counts and writes its total frequency
- while end of file reached it exits.

High frequency terms were also removed in addition to the stop words removed previously to avoid non-discriminating terms from the collection. Different threshold values for the minimum and maximum frequency were set and the news articles were represented with the remaining terms. An experiment was conducted (the map was trained and tested) on the articles for each pair of threshold values. Thresholds of 80 and 5, 75 and 3, 70 and 5, 70 and 3 were set for the maximum and minimum threshold respectively. Though the map organized the articles based on some pattern for each threshold value, the map created for the 70, 3 values seem better. That is, as compared to other maps, most of the clusters consists news articles from the same class when using 70, 3 threshold values. Hence for the purpose of this research terms having maximum frequency value above 70 and minimum frequency value below 3 were discarded out of the lists of unique words identified. The final result was a total of 142 unique words.

Term Weighting: Weights were assigned to terms to indicate the degree of importance (representativeness) of individual terms to each document. Before assigning weights to terms, using the 142 terms column wise and the news articles row wise, a matrix was initialized, i.e., the values of the matrix were initially set to 0. Then, the frequency of each term in a given article were brought from the file "WordList.txt" and filled in its appropriate position in the matrix. In this way, all the corresponding values were filled by the algorithm.

Among the different approaches for term weighting, the current study applied the commonly used TF- IDF term weighting technique. The formula used to calculate and assign weight is as follows:

$$w_{ij} = tf * (\log \frac{N}{n} + 1) \quad \text{\_\_\_\_\_} (3.1)$$

where **tf** is the frequency of term **i** in document **j,**

N is the total number of documents in the collection

n is the total number of documents containing the word i

Using the above formula, relative importance of the term (weight) was calculated for all words in each news article and the result was filled in the matrix constructed.

The weights obtained had to be normalized before being input to Nenet. This step is crucial for SOM since it uses Euclidean metric to measure distances between vectors. If Euclidean distance formula is used directly on attributes that are measured on different scales, the effect of some attributes might be completely dwarfed by others that have larger scales of measurement. For instance, if one variable has the values in the range of (0,...,1000) and the other in the range of (0,...,1) the former will almost completely dominates the map organization because of its greater impact on the distances measured. Hence, it was necessary to normalize weights to have values between 0 and 1.

Different formulas have been established for normalization (Witten, 2000). One approach is to divide all values by the maximum value encountered. The other alternative is to subtract the minimum value from the actual value and then to divide the result by the range. There is also another normalization technique which calculates the statistical mean and standard deviation of the attribute values, subtracts the mean from each value, and divides the result by the standard deviation.

The second alternative was chosen for this study. All the weighted values were normalized in between 0 and 1 using the following formula:

Normalized weight (w_i) = $\frac{V_i - MinV_i}{MaxV_i - MinV_i}$

Where V_i is the actual value of attribute *i*, and the maximum and minimum are taken over all instances in the dataset.

The minimum and the maximum weight value for each term in the entire dataset were calculated, and then the differences (range) of the two were determined. Finally, after subtracting the minimum value from the actual weight value of a term, the result was divided by the range. In this way, all the weights were scaled to have value in between 0 and 1.

The final text-term matrix with normalized weights look like figure 3.3.

142

0.0000	0.0000	0.0000	0.0000	0.3333	0.0000	0.0000	0.0000	0.0000	0.6667	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.3333	0.3333	0.3333
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	**102620**

0.0000	0.3333	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.6667	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	**102659**

Figure 3.3. The weighted text-term matrix

In the above figure, the first line indicates the total dimension of the space used and the columns indicate the normalized weight given to each word. The row corresponds to a news article. At the end of each row the News ID (see the bolded numbers in the matrix in the above figure) was attached so as to understand later how well the final map adapts itself to the data. Then both the training and the test files were saved as a plain data file format with the file names "training.dat", "Adegatest.dat", "Sporttest.dat" and "Gibernatest.dat".

3.5 Constructing the Self-Organizing Map using Nenet

After the plain ASCII data file was constructed, the news articles were fed into Nenet to create a two dimensional map. The Kohonen's feature map algorithm (SOM), takes a set of input matrices, each represented by the 142-dimensional vector and maps them onto nodes of a two-dimensional grid. The mapping procedure is recursive learning process that:

- Selects an input vector randomly from the set of all input vectors
- Finds the node closest to the input vector in the 142-dimensional space
- Adjusts weights of the winning node, so that it will more likely be selected again if this input is presented later
- Adjusts the weights of those nodes within a neighbourhood of the winning node, so that nodes within this neighbourhood will have similar weight patterns.

There are three essential steps that have been performed until the final map was constructed by Nenet. These are **Initialization, training** and **testing** the map. In the subsequent sections, a general overview of the tool used is given and then every task in the three steps undertaken in constructing the map is discussed.

3.5.1 Overview of Nenet

Nenet is a 32-bit Windows 95 and Windows NT 4.0 application designed to illustrate the use of a Self-Organizing Map (SOM). Nenet originally resulted from a project programmed for the course 'Software Project' at Helsinki University of Technology, Finland 1996-1997. Since then, the development of Nenet has continued and several new features have been added.

With Nenet, all the basic steps in map control can be performed. The map can be visualized in several ways and it can be labelled to make understanding of results easier. In addition, Nenet also includes some more exotic features in the area of visualization. For instance, different color scheme for visualization and labelling facilities are provided by Nenet (see figure 3.5).

3.5.2 Initialization of the Map

Using the New command button on the toolbar, the map is initialized for every experiment. Initialization sets reference vectors of each neuron to tentative values. The map initialization process requires several parameters to be set before it can be initiated. These include the X and Y dimensions, the map topology to be used and the neighborhood function. The topological structure of the map determines the total number of neighboring neurons each single neuron will have. The neighborhood function, on the other hand, determines how the neighborhood structure shrinks during the learning process. There is a start radius, which is the size of the neighborhood in the beginning of training and end radius, which is the size of the neighborhood reached when the training sequence ends.

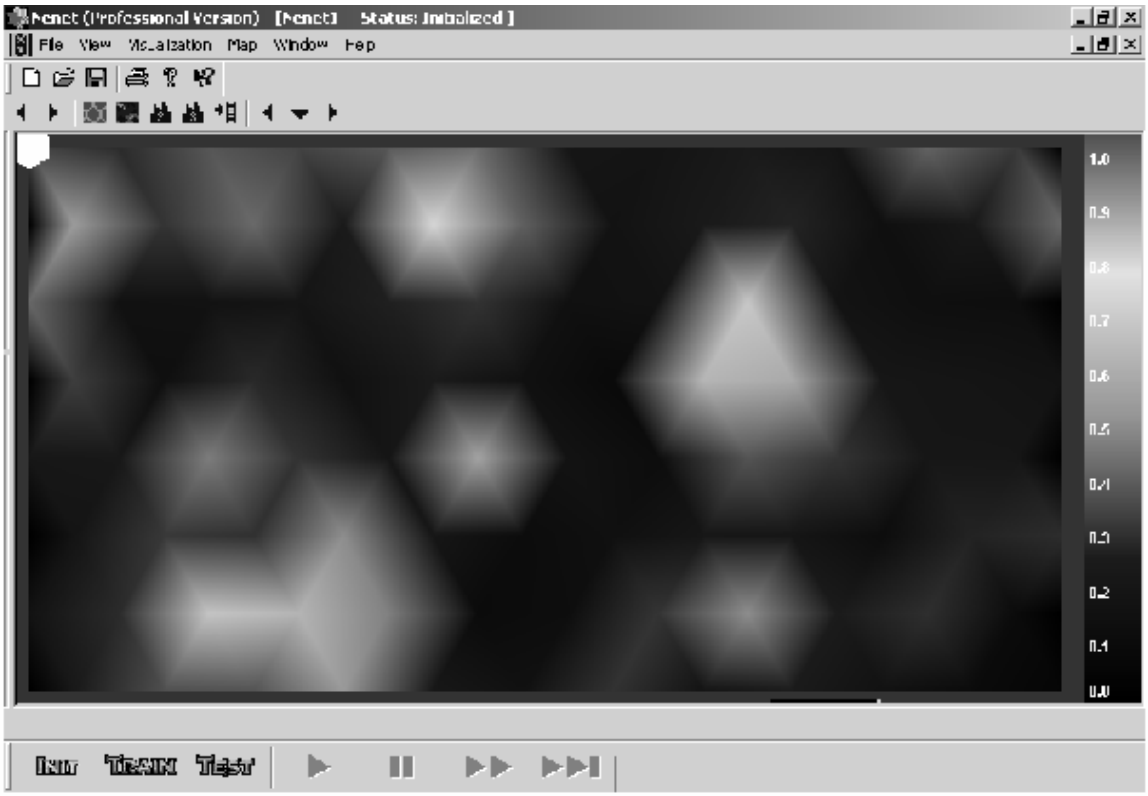


Figure 3.5. The Nenet Interface (The map in the initialized status)

While initializing the map for this study, the X and Y dimensions were set to **20** by **20** (enough for

the 248 news articles). The structure used was hexagonal map topology and a bubble neighbourhood function (the function that describes how the neighbourhood is taken into account in the SOM algorithm) and a linear initialisation type with random seed value of 0. No pre-processing method was specified for the initialisation. Thus, 248 News articles of 142 dimensions in the training dataset were fed as an input to the algorithm and the map was then initialised and the codebooks (reference vectors) were generated.

3.5.3 Training the Map

Training can be carried out after the map initialization. The map training is fundamentally used to execute the SOM algorithm for the map (to learn the characteristics of the data) and requires several parameters to be set before it can be initiated. The training process activates only after the OK button in training dialog has been pressed. The process flow control buttons under the bottom of figure 3.5 on the map control bar can be used to control the training sequence (i.e. to start, pause etc.).

During training, a linear initialisation type was used and the training proceeded in one phase because this initialisation type does the ordering of the reference vectors of the map neurons. Learning rate (´alpha´) was set initially to 0.05 and the value of learning rate was decreased after one million training cycles (length of the training measured in steps, each corresponding to one data vector) from 0.05 to 0.02 and remained 0.02 for a further 10 thousand training cycles (see figure 3.6).

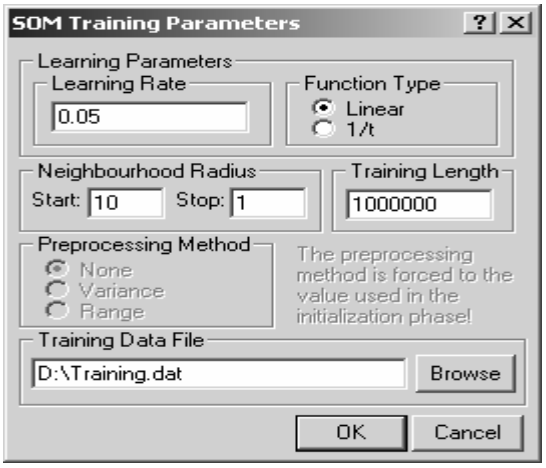


Figure 3.6. The parameters set while training the map

The size of the neighbourhood radius in the beginning of training was set to 10 and it stopped at 1 when the training sequence finished. The map training session then started with a 142-dimension data vectors representing the 248 news articles in the training data set. This process went through many iterations and is left to proceed until the mapping is ordered and is descriptive of the distribution of the data vector or when the adjustments all approach to zero.

In addition, to check the efficiency of the parameters used in this study, map training was initially conducted using different training parameters: learning rates and training regimes. Several training sessions were run using different parameter values and without a specific preprocessing method. In one case, for instance, learning rate (α) was set to 0.02 throughout and in another α was set to 0.05 initially (gave the map high opportunity to move initial clusters) but was then set to 0.02 for the final few training cycles. The ordering of the data vectors differs with each learning rate, training regime and map dimension giving minimal chance of comparison of the various training sessions although the SOM could still efficiently classify the data.

3.5.4 Testing the Map

The final step in constructing the map is the testing phase which is used to verify the correct adaptation of the map. Before a testing process can be initiated In Nenet, some testing parameters need to be determined. The testing process activates after the OK button in testing dialog has been pressed. The process flow control buttons on the map control bar can be used to control the testing sequence.

Primarily testing the efficiency of the ordering of similar pattern data vectors is required to make sure how the map has successfully adapted to the input data. This is a crucial step for constructing the Kohonen feature map for the entire data collections. As it was discussed in section 2.3.5 in chapter two, there are two alternatives for testing the map in clustering documents. One way is to ask subject experts for their judgment. The other option is to compare documents that are found on the map against their class information if there is class information that is already made by subject experts. In this study, the later option was used.

The adaptation of the map was tested initially using three test sets from three classes which were not included in the training set.

i. Test conducted using test set from Class Sport "ስፖርት"

The test was conducted first by testing the trained map with the test set from the class sport by automatically labeling the news articles with their identification number (ID). As shown in Figure 3.7, two parameters were set during testing the adaptation of the map. The first one is the initialize BMU (Best Matching Unit) hits, which sets all BMU hits to zero during the training phase. However, in this step clear BMU hits box was ticked to add new BMU hits to the existing ones. The second

parameter selected was the automatic labeling facility to display the ID of the news for the winning node on the map. After setting these parameters the map was tested.

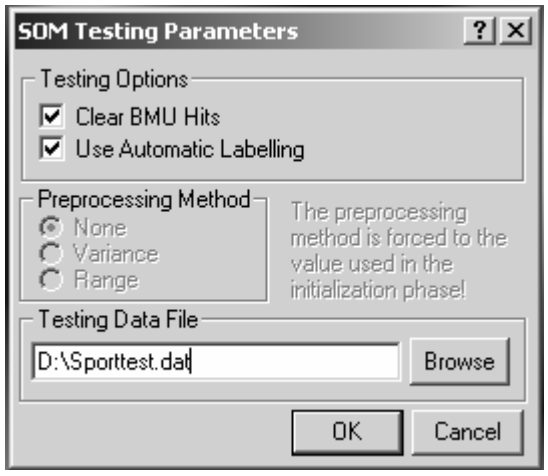


Figure 3.7. The parameters set during testing the map.

The resulting distribution of the articles was observed visually to determine how well the map adapts itself to the data. The result obtained is promising.

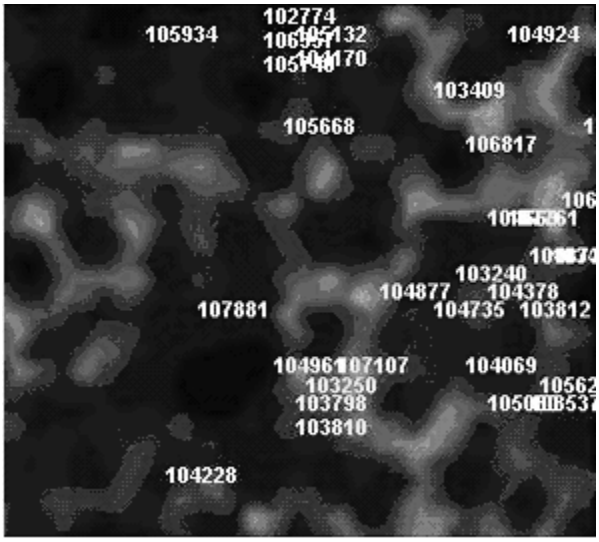


Figure 3.8. A map showing the distribution of News articles from the class sport " ስፖር

As it can be observed from the map (Figure 3.8), the test news articles were distributed close to one another in the right half of the map region almost forming three sets of clusters. For ease of visualization and observation the location of the winning nodes for each news articles is indicated by their IDs on the two-dimensional map. The news articles distributed at the right top of the map, for instance, region exhibit similar characteristics, i.e., their corresponding news titles indicated that they were news about the primer league sport of Ethiopian clubs and running competition (see table 3.4).

ID	News Title (head Line)
102774	በብሔራዊ ሊግ ውድድር ጉና ንግድና ደብረ ብርሀን ብርድ ልብስ አቻ ተለያዩ
105132	ክልል አቀፍ የስርት ውድድር በዱብቲ ተጀመረ
106357	የሩጫ ውድድር ተካሄደ
105740	የኢትዮጵያ አትሌቲክስ ሻምፒዮና ውድድር ለማካሔድ የሚያስችል ዝግጅት ተጠናቀቀ

Table 3.4. News articles with the corresponding titles located at the right top of the map shown in figure 3.8

As one can observe around the bottom of the map, there are also news articles forming a cluster. Their title describes that they were all about running competition of the famous Ethiopian runners (Table 3.5).

ID	News Title (Head Line)
103798	ቀነኒሳ የብራስልስን አገር አቋራጭ አሸነፈ
103810	ቀነኒሳ በቀለ በብራስልስ አገር አቋራጭ ውድድር አሸነፈ

103250	አትሌት ኃይሉ ንጉሴ ጃን በተደረገ የማራቶን ውድድር አሸነፈ
107107	አትሌት ብርሃኔ አደሬ በሳምንት ሁለት ታላላቅ ውድድሮችን አሸነፈች። አዲስ የዓለም ክብረወሰን ለማስመዝገብ ሳይሳካላት ቀረ።

Table 3.5. News articles with the corresponding titles located around the top of the map shown in figure 3.8.

ii. Test conducted using test set from Class Accident " ጋደአ "

The second test was made on the test set from the class "accident (**ጋደአ**)". The result showed that the news articles were clustered around the bottom left corner of the map (Figure 3.9). The analysis made on the news ID and the corresponding title (head line) revealed that the map has adapted itself to the features of the data.

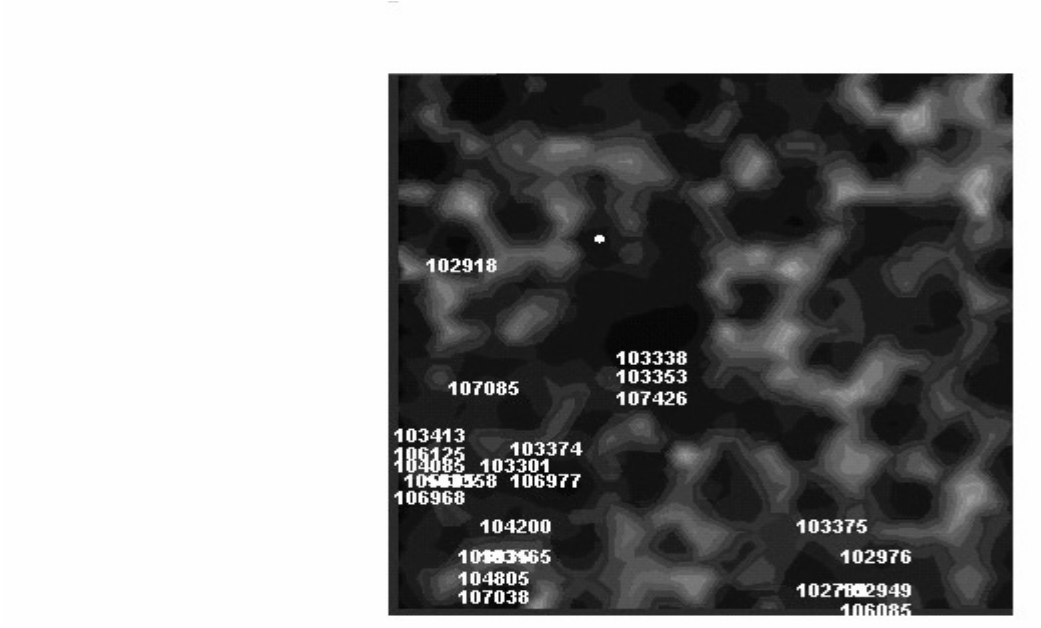


Figure 3.9. A map showing the distribution of news articles from the class Accident " ጋደአ"

For instance, the corresponding titles of the News ID which are located at the left bottom of the

map indicated that they were news about car accidents (Table 3.6).

ID	News of title (Head line)
103413	ሁለት ተሽከርካሪዎች ተገልብጠው ሁለት ሰዎች ሞቱ
106125	ተሽከርካሪ ተገልብጦ ሶስት ሰዎች ሞቱ
104085	የጭነት መኪና ተገልብጦ ሶስት ሰዎች ሞቱ ሰባቱ ከባድና ቀላል ጉዳት ደረሰባቸው
106968	በመኪና አደጋ ጋዜጠኛ መታፈሪያ አበበን ጨምሮ ሶሰት ሰዎች ሞቱ

Table 3.6 News articles with the corresponding titles located around the bottom left of the map shown in figure 3.9.

iii. Test conducted using test set from Class Agriculture *"ግብጉ"*

Finally, the third test was made using the test from Agriculture *"ግብጉ"*. As the map shows almost all of the dataset were clustered around the center of the map except some outliers (Figure 3.10). To determine how far the map ordered the dataset, the ID numbers of the following articles were taken and the corresponding news titles were retrieved from the database. The result is shown in table 3.8

Figure 3.10. A map showing the distribution of News articles from the class Agriculture " ግብጉ"

ID	Titles
104403	የዱር እንስሳት መመናመንን ለመግታት የሚረዳ ፖሊሲና ስትራቴጂ ተረቀቀ
107149	ወደ አፍሪካ የሚገቡ ባዕዳን የእፅዋትና የእንስሳት ዝርያዎች ከፍተኛ ጫና እንደሚያሳድሩ ተገለጸ
102683	የእንስሳት ገበያ ባለስልጣን 3 ረቂቅ አዋጆችን ማዘጋጀቱ አስታወቀ ።
104770	በህገወጥ አደን ሳቢያ በርካታ ቁጥር ያላቸው የዱር እንስሳት ቀንሰዋል።

Table 3. 7. News articles with the corresponding titles located around the center of the map shown in figure 3.10

As it can be observed from the titles of the articles in table 3.7, all deals with some issues about wild animals. Similarly, if we look at the three articles at the bottom of the map, except one the two

are concerned about the water retaining project that has been undertaken all over the country (table 3.8). Although the third article does not exactly mean what the others describe, it expresses about the rain condition. Hence from this it can be deduced that all the three are conceptually the same.

ID	Title
104075	ሰዎች ለሰዎች ውኃ በማቆር ድርቅ ለመከለካል የሚረዳ ስልጠና ሰጠ
103336	በኢትዮጵያ የሚካሄድ ማንኛውም የገጠር ልማት ውሃን ማዕከል ያደረገ ይሆናል
10388	በአሁኑ ወቅት በአንዳንድ የኢትዮጵያ ክፍሎች እየጣለ ያለው ዝናብ ለጥቂት ቀናት ይቀጥላል

Table 3.8. News articles with the corresponding titles located around the bottom of the map shown in figure 3.

3.6. Development of the Prototype Browsing Interface

The development of the prototype browsing interface is the main concern of this research. The practical purpose of developing a WEBSOM map is to provide an interface to a document collection so that any one could conveniently explore the map. This part of the project involved labeling the map, creating the map image- the map that was constructed using Nenet, creating the active server pages and database connections to these pages.

3.6.1. Labeling the Map

After the test was made and the adaptation of the map to the data has been confirmed, the whole map for the entire news articles (both the training and test set) was constructed and automatically labeled with the news ID of each articles (see figure 3.11).

Figure 3.11. The map for the 330(entire news articles) automatically labeled with News ID

SOM can be labeled to give each neuron a meaning. The labeling of the map is carried out by assigning one or several labels to selected neurons. In this study, labeling was performed by double clicking on each neuron and assigning a key term that mostly represents the contents of news articles held by that neuron or a group of neurons within a certain radius. The appropriate term for the area was selected on the basis of the analysis and observation that was made on the

nature of the news articles organized at a particular region on the map. Then, subject experts (Journalists[9] in this case) were consulted for their judgment of the terms used for labeling the area on the map**.** The labeled map of the prototype was presented in Figure 3.12. Since Nenet does not incorporate Amharic font types for labeling the map, Latin characters were used as an alternative.

3.6.2 The Image Map and HTML Page

The labelled map that was constructed by the SOM using the Nenet tool was converted into a Jpg image file. Then, an HTML (hypertext mark-up language) page was created that displays the image map. A Java script that reads the X and Y coordinates of the area of the image map was embedded into the HTML tags. The script not only reads the X-Y coordinate but it passes these values as a parameter to the Active Server pages when the user clicks on the map.

3.6.3. The news articles database

The news articles are stored using MS Acess. The table contains the news article Identification number (ID), the headline (titles of the news), slug, , the classification code, and the full story the X coordinate and Y coordinate as a column name and every news articles are recorded using these column headings. A sample of the database table structure was presented in appendix 4.

3.6.4. The Active Server pages

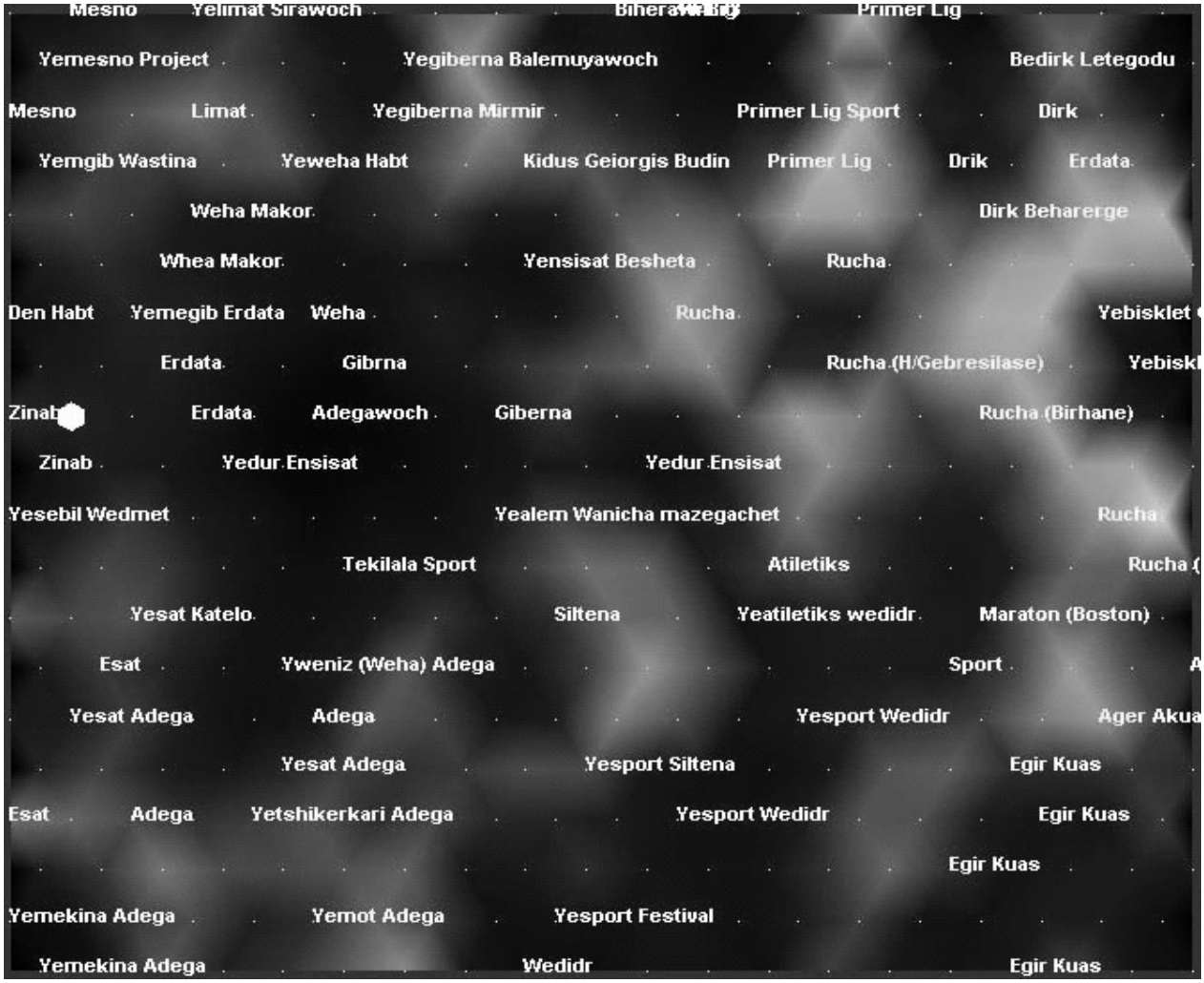
There are two active server pages that were created for facilitating browsing of the News articles from the Access database. The first page accepts the X-Y coordinate values passed to it when the user clicks on any location on the map. Based on the values of the parameters, it calculates the

[9] Zenebe Desta, and Abebaw Zewde

values of the neighbouring coordinates in all directions (left, right, up and bottom) and then searches the database for matching coordinates and retrieves the Head line or the titles of the associated News articles with these coordinates. The lists of titles of the News articles that are found are displayed on a new HTML page. The titles are hyperlinked to the corresponding full story. This active server page also holds the full story of the corresponding title and passes it as a parameter to the second active server page.

The user then can click on any one of the titles and get the full story. When the user clicks on the title, the full story held by the first active server page will be passed as a parameter to the second active server page. Then, this page displays the full story to the user. Thus getting the full story from the database can be considered as a two step actions: First the user clicks on the image map which gives a list of titles that met the criteria. Second further clicking on the title will display the full story.

A sample example was shown below. When the user clicks around "premier lig sport" label term on the map (at the right top position), a coordinate matching was found for five news articles in the database and their titles were displayed along with the date information. Further clicking on the first news title retrieved the corresponding full story (Figure 3.12).



አርእስተ ዜና

News Titles

አርባምንጭ ጨርቀጨርቅና መ·ገር ሲ·ሚንቶ አቻ ተለያዩ 11.30.2002
አርባምንጭ ጨርቃ ጨርቅ አዋሳ ከነማን 2 ለ 0 አሸናፈ 11.25.2002
አዋሳ ከነማ መ·ርን ሁለት ለአንድ አሸነፈ 12.14.2002
ሐረር ቢራ ቡናን 2 ለ 1 አሸነፈ 11.18.2002
የፕሪሚየር ሊግ ኒያላ አዋሳ ከነማን 2 ለ1 አሸነፈ 12.18.2002

አርባምን ጨርቀጨርቅና ሙገር ሲሚንቶ አቻ ተለያዩ ፕሪሚየር ሊግ ፕሪሚየር ሊግ በአርባምን ከተማ በተካሄደው የኢትዮጵያ ፕሪሚየር ሊግ ውድድር አርባምን ጨርቃጨርቅና ሙገር ሲሚንቶ በአቻ ውጤት ተለያዩ በሜዳው ላይ ባካሄዳቸው አምስት ተከታታይ ጨዋታዎች ያሸነፈው አርባምን ጨርቃጨርቅ ትናንት ባደረገው ጨዋታ ግን አልተሳካለትም

Figure 3.12. The three different view levels: the whole map, the news titles and the full story for the second news title.

3.6.5. Evaluation of the Prototype Browsing Interface

As it was discussed in section 2.3.5 of chapter two, it is difficult to define a standard evaluation method for measuring the quality of visualization, exploration and navigation systems. However, the quality of the map display may generally be evaluated by an expert in the application area (Kaski, 1997). User studies may also be required until more direct, automatically applicable measures are determined (Lagus, 2000b).

Although a complete user study has not been undertaken and is outside the scope of this study, an attempt was made to evaluate the prototype browsing interface developed. This was done by making available the interface to users in the domain area for a few days. Then an observation was made about their feeling to the system. Initially, it took some hours until they adapt to this new system. After a while, they became fascinated as the labels on the map provided them a general overview of the entire database. As they navigate on the map, they found articles which deals with the same thing are very near than others. Except some overlaps found at the left middle of the map, they appreciated the clustering power of the tool.

A general comment was also obtained from these users. That is, to apply this tool for developing a ful-fledged browsing interface that takes into consideration users need so that the overall burden lies on users in finding similar items (news articles in this case) will be reduced.

3.7 Discussion on the Prototype Map

The labelled map (Figure 3.11) was found to reflect relations among the three news articles. That means similar news articles occur near each other on the map and are thus retrieved together when the user clicks on a particular area based on his/her interest. As one can observe from the test results of the test sets (Figures 3.8, 3.9 and 3.10 and Tables 3.4, 3.5., 3.6, 3.7, and 3.8 indicated that the SOM classified and clustered the news articles effectively. That means similar news articles were found very near each other as well as clusters having some common points also fall near each other. For instance, in Figure 3.12 as the labels indicated news articles from class "Agriculture" and class "Accident" was located near each other than the other class (sport). The reason was that some of the articles in class "Accident" deals about providing support for farmers who were unable to cultivate due to different factors. Hence, there was an overlap in some terminologies between these two classes. This indicates that the presence of news articles which may belong to different news classification. The visualised clustering tendency or density of the articles in different areas of the collection, presented with the colour of the document map, can also be used to aid in finding related articles.

Also, the distribution of the news articles on the map can be easily identified based on the different colour visualization that the map provided. The dark colour in Figure 3.11, for instance, shows that most of the articles were concentrated on the area where as the light green colour at the right side of the map indicates articles were distributed widely.

From the above figures and discussion made, it is possible to say that the trained map thus makes successful distinction of the various types of news articles in the samples used. The map can be categorized into four major areas based on the distribution of the News articles. On the right side, bottom to up, of the map the articles from class “Sport” were distributed. On the right corner of the map, there is a cluster of news that deals with drought. On the left bottom half of the map, the articles were from class ”Accident” and at the left top of the map they were from class “Agriculture”. This reveals that the classification and clustering obtained by the SOM is consistent with the manual classification.

To wind up, it is therefore quite evident that the SOM has effectively adapted to the data and is thus suitable, it enables interactive browsing and exploration of the document database. Map regions are appropriately characterised with keywords, to be regarded as some kind of landmarks on the map display, to provide guidance to the exploration. These keywords serve as navigation hotspots during the exploration of the map, as well as provide information on the topics discussed in the news articles on the respective map area. When clicking a point on the map display with a mouse, links to the news article database enable reading the full story of the news article.

Chapter Four
Conclusions and Recommendations

4.1 Conclusion

Searching for relevant documents from very large collections has traditionally been based on keywords and their Boolean expressions. However, such systems are not suitable for exploration tasks in cases where the user either do not know the domain very well, or they have only a limited idea of the collection being examined.

Recently an alternative method to such query based retrieval systems called WEBSOM has been developed. It is a full-text information retrieval and exploration method for large document collections, which is based on the self-organizing map (SOM) algorithm. The self-organizing map is one of the major unsupervised artificial neural network models. It basically provides a way for cluster analysis by producing a mapping of high dimensional input vectors onto a two dimensional output space while preserving topological relations as faithfully as possible. After appropriate training iterations, the similar input items are grouped spatially close to one another.

In this study, 330 Amharic news articles of three classes were collected from the Ethiopian News Agency. To conduct the experiment, 248 of the news articles were assigned as a training set and the remaining as a test set. For the purpose of document representation, the Vector Space Model was used. Non-content bearing terms were removed from the lists of terms identified from the headline and slug parts of the news articles and suffix/prefix-stripping technique was applied on the remaining list. After changing terms having different writing forms in to one common form, terms with a total frequency of above 70 and below 3 were discarded from the list. Then, a vector matrix

both for the training and test set were constructed on the remaining 142 terms. At last, a normalized weight was assigned to each term in a given news article based on TF-IDF weighting technique and the vector matrix were prepared in appropriate format for the tool to be used.

A 20 by 20 unit map was initialised and trained with 248 news articles selected from three classes: Agriculture, Accident, and sport. Then, to get an idea of the quality of the ordering of the final map, a test was made by running three different sets of input data that were not included originally in the training dataset.

While testing, the winning map nodes for the test data were automatically labelled by the news identification (ID) number. The accuracy (the purity of the nodes) were evaluated by comparing the news articles that fall in to a particular area on the map against the manual class determined by subject experts using the ID of the articles as a key. The result obtained was promising and the map has successfully adapted to the data since similar vectors were distributed closer to one another forming various clusters. The clusters formed are: one in the right half of the map (Sport), the other on the lower left corner of the map (Accident) and the last at the central part of the map (Agriculture).

After testing the purity of the map in clustering data, the Kohonen feature map was constructed for the entire (330) News articles considered in this study. The distribution of the articles on the map has shown the powerfulness of the tool in clustering similar articles near each other. In general, the map area was categorized in to four major regions based on the distribution of the dataset. The clusters in the right half of the map were articles from class “Sport ”. At the upper right corner a cluster consisting news about drought , which are all from class Agriculture are found. On the lower left side of the map seems to have specialized particularly on articles which are all about accidents

made by different factors and the left upper shows articles from class “Agriculture”. Of course, there has been found some overlaps between articles from class “Accident ” and class “Agriculture” at the middle left of the map. The analysis made on the contents of the documents in the overlapping indicated that the overlap is due to vocabularies that articles from both classes have in common. An attempt has also been made to compare the local clusters formed by the map with the actual classes of the articles that formed the cluster. The result showed that the accuracy of the map in clustering is compatible with the human classification that was made except for few outliers.

Using the labels on the map that characterize the clusters on the region, users can navigate the whole map. Once an interesting area on the map is identified using those descriptive terms as a landmark, with a simple mouse click lists of news titles that are related to that area will be displayed along with the date information and then the links made with the titles will lead to the corresponding full story.

User's reaction to the prototype browsing interface developed seems encouraging: the system enabled them to visualize and explore unfamiliar articles in the collection. Moreover, they have got a general overview of the collection in the database.

The output of the current research has proven that the tool can be used for organizing Amharic texts. The analysis made on the map indicated that similar news articles became mapped near each other. Moreover, similar clusters are also organized to be near each other on the map than from those less similar to each other.

4.2 Recommendation

The experiment was conducted on small-scale texts and the test result is promising. However, much has to be done to improve the clustering performance of the tool.

One immediate possible future work that can be commenced from this research is to apply and test the tool for very large Amharic texts. Since the map created was a prototype and considered data only from three classes, it has to be tested and validated for the other classes that are not considered in the development process of this study. It is also possible to apply the tool for organizing documents containing images and pictures.

Another effort that has to be made is in the preprocessing aspect. In this study, the Vector Space Model was used for representing documents and generating the matrix. By applying the latent semantic indexing done for Amharic by Tewodros(2003) and that of the n-gram indexing approach of Bethlehem (2002) for document rpresentation and random mapping for the reduction of the dimension space, it is possible to test and validate the results obtained in this study.

Furthermore, a simple depularization and suffix/prefix stripping technique was applied while preprocessing to reduce variants of a word into a common form. One improvement that can be made is to make use of a complete Amharic Stemmer. Besides there was no standard stop word list established for Amharic. Most of the stop word lists used in this study are terms of journalists (commonly used words for reporting news). With use of standard stop word list better results may be attained. An effort should also be made to develop an interface that can support Amharic font types. At last, interested researcher can also apply and test this novel method on texts of other local languages.

References

Abiyot Bayu. (2000). Design and developemnt of word parser for Amharic language (Masters Thesis). School of Information Studies for Africa. Addis Ababa University, Addis Ababa. (Unpublished)

ባዬ ይማምና ተ.ም (1997) . "ፊደል እንደገና" የኢትዮጵያ የቋንቋዎችና የሥነ ፅሁፍ መፅሄት ቁጥር 7 (1-32)

Beletu Reda. (1982). Graphemes Analysis of the writing system of Amharic paper for the requirement of the Degree of Bachelor of ART in Linguistics. Addis Ababa University.

Belkin N.J. and Croft W.B. (1987). Retrieval Techniques. In: Williams M.E. (ed). *Annual Review of Information Science and Technology,*22,109-145. Amsterdam:Elsevier

Bender, M.L., et al.,(1976). *The Ethiopian Writing System*. In Bender et al (Eds). Language in Ethiopia.London: Oxford University Press.

Bethlehem Mengistu. (2002). *N-gram based automatic indexing for Amharic text.(*Masters Thesis). School of Information Studies for Africa. Addis Ababa University, Addis Ababa. (Unpublished)

Beza_Yates R., and Ribeiro-Nets. (1999). Modern Information Retrieval. Harlow; Addison Wesely Longman Limited.

Chen H., Houston A.L., Sewell R.R., Schatz B.R. (1998). Internet browsing and searching: User evaluations of category map and concept space techniques. *Journal of the American Society for Information Science* 49(7):582-603.

Chowdhury, G. (1997). Introduction to Modern Information Retrieval. A teaching Material for the course INST 534, Addis Ababa University. (Unpublished).

Croft, B. (1997). Retrieval Effectiveness of Various Indexing Techniques on Indonesian News Articles. Univesrsity of Massachsett,Amherst.

ENA. (1993a). Handbook for editorial Staff. Addis Ababa.

Getachew Haile. (1967).The problems of Amharic Writing System. Addis Ababa University.(Unpublished).

Han J. and Kamber M. (2001). *Data mining: Concepts and techniques.* Morgan Kaufamm Publisher, Academic Press, San Diago, USA.

Harter S.P. and Hert C.A.(1997). Evaluation of information retrieval systems: Approaches, issues and methods. In: Williams M.E. (ed). *Annual review of information science and technology.* 32:3-94.

Hildreth, R. (1995). Information Retrieval Models. IR Research Online Catalog Models

phoenix.liunet.edu/~hildreth/clr-opac.html - 3k

Honkela T.(1997). *Self-organising maps in natural language processing.* PhD Thesis, Helsinki University of Technology, Neural Networks Research Centre, Finland.

www.cis.hut.fi/~tho/thesis/honkela.ps.Z

Honkela T. (1997). WEBSOM- Self-Organizing Maps Of Document Collections. Neural Networks Research Centre, Espoo, Finland.

websom.hut.fi/websom/doc/ps/honkela97wsom.ps.gz

Honkela T., Kaski S., Lagus K. and Kohonen T.(1996). Newsgroup exploration with WEBSOM method and browsing interface. Report A32. Helsinki University of Technology, Faculty of Information Technology, Laboratory of Computer and Information Science. Helsinki, Finland.

www.websom.hut.fi/websom/doc/websom.ps.gz

Hudson, G.(2001). Aspects of the History of Ethiopic Writing, IES Bulettin, 25, 1-10.

Kaski S.(1997). *Data exploration using self-organising maps.* PhD Thesis, Helsinki University of Technology, Neural Networks Research Centre, Finish Academies Technology, Acta Polyte chnica Scandinavia.

www.websom.nucleus.hut.fi/~sami/thesis.ps.gz

Kaski S., Honkela T., Lagus K., and Kohonen T.(1998). WEBSOM—Self-organising maps of document collections. *Neurocomputing.* 21:101-117.

Kemp D.A. (1988). *Computer-based knowledge retrieval.* Aslib, London, UK.

Kohonen T. (1998). Exploration of very large databases by self-organising maps. In: *Proceedings of ICNN'97, International Conference on Neural Networks*. pp. PL1-PL6. IEEE Service Center, Piscataway, New Jersey.

websom.hut.fi/websom/doc/ps/kohonen97icnn.ps.gz

Kohonen T.(1999). Self-organisation of very large document collections: State of the art. In: Niklasson L., Doden M., and Ziemke T. (eds). *Proceedings of ICANN98, the 8th International Conference on Artificaial Neural Networks.* Vol. 1, Springer, London, UK. pp. 65-74.

websom.hut.fi/websom/doc/ps/kohonen98.ps.gz

Kohonen T. (2001). *Self-Organising Maps.* 3rd edition. Springer-Verlag, Berlin, Germany.

Kohonen T., Kaski S. Lagus K. and Honkela T.(1996). Very-large two-level SOM for the browsing of newsgroups. In: von der Malsburg C., von Seelen W. Vorbruggen J.C. and Sendhoff B. (eds). *Proceedings of ICANN96, International Conference on Artificial Neural Networks, Bochum, Germany, July 16-19, 1996.* Lecture Notes in Computer Science, Vol. 1112, pp. 269-274. Springer, Berlin.

Kohonen T., Kaski S., Lagus K., Salojarvi J. Paatero V. and Saarela A.(2000). Self organisation of a massive document collection. *IEEE Transactions on Neural Networks.* Special issue on

Neural Networks for Data Mining and Knowledge Discovery, Vol 11, Number 3, Pages 574-585.

Kohonen T., Kaski S., Lagus K., Salojarvi J., Honkela J., Paatero V. and Saarela A.(2000). Self-organisation of a massive text document collection. *IEEE Transactions on Neural Networks, Special Issue on Neural Networks for Data Mining and Knowledge Discovery.* 11: 574-585.

Lagus K.(1997). Map of WEBSOM'97 abstracts—alternative index. In: *Proceedings of WSOM'97, Workshop on Self-organising Maps, Espoo, Finland, June 4-6.* pp. 368-372. Helsinki University of Technology, Neural Networks Research Centre, Espoo, Finland.

Lagus K. (1998). Generalisability of the WEBSOM method to document collections of various types. In: *Proceedings of 6th European Congress on Intelligent Techniques and Soft Computing (EUFIT'98)*. Verlag Mainz, Aachen, Germany, Vol. 1, pp. 210-214.
websom.hut.fi/websom/doc/ps/lagus98eufit.ps.gz

Lagus K.(2000a). *Text retrieval using self-organised document maps.* Technical Report A61, Helsinki University of Technology, Laboratory of Computer and Information Science.
websom.hut.fi/websom/doc/ps/lagus00tr.ps.gz

Lagus K. (2000b). *Text mining with WEBSOM*. PhD Thesis, Helsinki University of Technology, Neural Networks Research Centre, Finish Academies Technology, Acta Polytechnica Scandinavia, Mathematics and Computing Series No. 110.

Lagus K.(2002). *Text retrieval using self-organised document maps.* Technical Report A61, Helsinki University of Technology, Laboratory of Computer and Information Science.
websom.hut.fi/websom/doc/ps/lagus00tr.ps.gz

Lagus K. and Kaski S.(1999). Keyword selection method for characterising text document maps. In: *Proceedings of the Ninth International Conference on Artificial Neural Networks (ICANN'99)*. IEEE Press, London, pp. 371-376.

Lagus K., Honkela T., Kaski S. and Kohonen T.(2000a). WEBSOM for textual data mining. *Artificial Intelligence Review*. 13(5/6):345-364.

Lagus K., Honkela T., Kaski S., and Kohonen T.(1996a). *Self-organising maps of document collections*: A new approach to interactive exploration. In: Simoudis E., Han J., and Fayyad U., (ed). *Proceedings of the second International Conference on Knowledge Discovery and Data Mining (KDD'96)*. AAAI Press, Menlo Park, CA, pp. 238-243.

Lagus K., Kaski S., Honkela T., and Kohonen T.(1996b). Browsing digital libraries with the aid of self-organising maps. *Proceedings of the Fifth International World Wide Web Conference WWW5, May 6-10*. Paris, France, pp. 71-79.

Lin X. (1997). Map displays for information retrieval. *Journal of the American Society for Information Science* 48(1): 40-54.

Marchionini, G. (1995). Information seeking in electronic environments. New York: Cambridge University Press.

Mulegeta Bayeh. (2002). *Text retrieval using self-organizing document map*: The case of ILRI digital Libarary. *(*Masters Thesis). School of Information Studies for Africa. Addis Ababa University, Addis Ababa. (Unpublished)

Nega Alemayehu.(1999). *Development of stemming algorithm for Amharic language text retrieval.* PhD Thesis, University of Sheffield, Sheffield, UK.

NNI (Neural Networks Information).(2001). http://koti.mbnet.fi/~phodju/nenet/NeuralNetworks/NeuralNetworks.html

Pris.(2001). Extracting Meaningful Labels for WEBSOM Text Archives www.comp.nus.edu.sg/~arnulfo/cikm2001.ps

Saba (2001). *The Application of Information Retrieval Techniques to Amharic Documents on the Web. (*Masters Thesis). School of Information Studies for Africa. Addis Ababa University, Addis Ababa. (Unpublished).

Salton G. and McGill M.J. (1983). *Introduction to Modern Information Retrieval.* McGraw-Hill, New York, USA.

Swanson D.R. (1988). Historical note: Information retrieval and the future of an illusion. *Journal of American Society for Information Science* 39(2):92-98.

Van Reijsbergen X. (1996). *Information Retrieval*. London:Butterworths

Wise J.A. (1999). The ecological approach to text visualisation. *Journal of American Society for Information Science* 50(13):1224-1233.

Witten, I. and Frank, E. (2000). Data Mining: Practical machine learning tools and techniques with JAVA implemewntations. San Fransico: Morgan Kaufman Publishers.

Zelalaem Sintayehu. (2001). *Automatic Classification of Amharic News Items*: the case of Ethiopian News Agency. *(*Masters Thesis). School of Information Studies for Africa. Addis Ababa University, Addis Ababa. (Unpublished)

Appendices

Appendix 1: List of Amharic Characters and Number Systems

		u	i	a	y	e	o
h	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ
l	ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ
H	ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ
m	መ	ሙ	ሚ	ማ	ሜ	ም	ሞ
S	ሠ	ሡ	ሢ	ሣ	ሤ	ሥ	ሦ
r	ረ	ሩ	ሪ	ራ	ሬ	ር	ሮ
s	ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ
q	ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ
b	በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ
t	ተ	ቱ	ቲ	ታ	ቴ	ት	ቶ
c	ቸ	ቹ	ቺ	ቻ	ቼ	ች	ቾ
n	ነ	ኑ	ኒ	ና	ኔ	ን	ኖ
N	ኘ	ኙ	ኚ	ኛ	ኜ	ኝ	ኞ
x	አ	ኡ	ኢ	ኣ	ኤ	እ	ኦ
k	ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ
w	ወ	ዉ	ዊ	ዋ	ዌ	ው	ዎ
X	ዐ	ዑ	ዒ	ዓ	ዔ	ዕ	ዖ
z	ዘ	ዙ	ዚ	ዛ	ዜ	ዝ	ዞ
Z	ዠ	ዡ	ዢ	ዣ	ዤ	ዥ	ዦ
Y	የ	ዩ	ዪ	ያ	ዬ	ይ	ዮ
d	ደ	ዱ	ዲ	ዳ	ዴ	ድ	ዶ
j	ጀ	ጁ	ጂ	ጃ	ጄ	ጅ	ጆ
g	ገ	ጉ	ጊ	ጋ	ጌ	ግ	ጎ
T	ጠ	ጡ	ጢ	ጣ	ጤ	ጥ	ጦ
C	ጨ	ጩ	ጪ	ጫ	ጬ	ጭ	ጮ
P	ጰ	ጱ	ጲ	ጳ	ጴ	ጵ	ጶ

		u	i	a	y	e	o
f	ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ
p	ፐ	ፑ	ፒ	ፓ	ፔ	ፕ	ፖ
v	ቨ	ቩ	ቪ	ቫ	ቬ	ቭ	ቮ
Q	ቐ	ቑ	ቒ	ቓ	ቔ	ቕ	ቖ

		u	i	a	y	e	o
h	ኀ	ኁ	ኂ	ኃ	ኄ	ኅ	ኆ
H	ሸ	ሹ	ሺ	ሻ	ሼ	ሽ	ሾ
s	ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ
t	ጸ	ጹ	ጺ	ጻ	ጼ	ጽ	ጾ
T	ፀ	ፁ	ፂ	ፃ	ፄ	ፅ	ፆ

IIBILIAZED

lwa	ሏ	Nwa	ኗ	bwa	ቧ	Cwa	ቋ
mwa	ሟ	zwa	ዟ	twa	ቷ	tswa	ፏ
rwa	ሯ	Zwa	ዧ	cwa	ቿ	fwa	ፉ
swa	ሷ	dwa	ዷ	nwa	ኗ	vwa	ፗ
shwa	ሧ	jwa	ዯ	Twa	ጧ	sft+2	ሻ

	o	i	u	a	e
kw	ኰ	ኲ	ኰ	ኳ	ኴ
gw	ጐ	ጒ	ጐ	ጓ	ጔ
qw	ቈ	ቊ	ቈ	ቋ	ቌ
hw	ኈ	ኊ	ኈ	ኋ	ኌ

The Amharic Numerals System

፩	፪	፫	፬	፭	፮	፯	፰	፱	፲	፳	፴	፵	፶	፷	፸	፹	፺	፻	፼
1	2	3	4	5	6	7	8	9	10	20	30	40	50	60	70	80	90	100	1000

Appendix 2: Classes of News Articles

Classification Code	
Code	**Description**
ብለ Npo	ብሔራዊ ፖለቲካ National Politics
ዓአጉ Iaf	ዓለም አቀፍ ጉዳዮች International Affairs
ኢኮኖ Eco	ኢኮኖሚ Economy
መናጸ D&S	መከላከያና ጸጥታ Defense and Security
ትምህ Edu	ትምህርት Education
ጤናጥ Hel	ጤና ጥበቃቀ Health
ሣናተ S&T	ሣይንስና ቴክኖሎጂ Science and Technology
ኃደአ Acc	አደጋዎች Accidents
የአፀ Wea	የአየር ጸባይ Weather
ስፖር Spo	ስፖርት Sports
ማቆጥ Loe	ጥቆማ List of Events
ማኅበ Soc	ማኅበራዊ Social
ባሕጉ Cul	ባሕል ጉዳዮች Culture
ግብጉ Agr	ግብርና ጉዳዮች Agriculture
ሕናፍ L&J	ሕግና ፍትሕ Law and Justice
ሌፈዓ Ocs	ሌሎች የፈርጅ ዓይነቶች Other Classifications

Appendix 3: Lists of Suffix

List Of Suffix (adapted from Zelalem (2001))		
ናም	ን	ና
ናምና	ንና	ናና
ናምን	ንናም	ች
ናን	ንም	ቹ
ናንና	ንምና	ቹና
ናንም	ንምን	ቹናም
ናንን	ንምው	ቹናን
ናዉ	ንን	ቹም
ናዉና	ነት	ቹምና
ናዉም	ዉ	ቹምን
ናውን	ዉና	ቹን
ም	ዉናም	ቹንና
ምና	ዉናን	ቹንም
ምናም	ዉናዉ	ችና
ምናን	ውም	ችናም
ምቹ	ዉምና	ችናን
ምን	ዉምን	ችም
ምንና	ዉን	ችምና
ምንም	ዉንና	ችምን
ምንን	ዉንም	ችን
ምንዉ	ዉንን	ችንና
ምዉ		ችንም
ምው		
ምው·ና		
ምዉና		
ምዉም		
ምው·ም		
ምው·ን		
ምዉን		

Appendix 4. The news articles Table

ID	HeadLine	Slug	Classification	FullStory	xcoordinate	ycoordinate
10386363	እርዳታን ከዛላቂ ልማት ጋር በማቀናጀት ፕሮጀክት ቀረጻ ላይ ለሰለጠኑ ምስክር ወረቀት ተሰጠ	ስልጠና	ጋደአ	እርዳታን ከዛላቂ ልማት ጋር በማቀናጀት ፕሮጀክት ቀረጻ ላይ ለሰለጠኑ ምስክር ወረቀት ተሰጠ ስልጠና ስልጠናአደጋ መከላከልና ዝግጁነት ኮሚሽን ከሁለት ግብረሰናይ ድርጅቶች ጋር በመተባበር ዕለታዊ እርዳታን ከዘላቂ ልማት ጋር ለማቀናጀት በፕሮጀክት ቀረጻና አተገባበር ላይ ያሰለጠናቸው 24 ሰዎች ምስክር ወረቀት ተቀበሉ የኦሮሚያ አደጋ መከላከልና ዝግጁነት ኮሚሽን ኮሚሽነር አቶ ደምሴ ለገ	0	0
10391919	በአቃቂ አካባቢ በደረሰ የመኪና አደጋ የአንድ ሰው ህይወት አለፈ	የመኪና አደጋ	ጋደአ	በአቃቂ አካባቢ በደረሰ የመኪና አደጋ የአንድ ሰው ህይወት አለፈ የመኪና አደጋ የመኪና አደጋአዲስ አበባ ውስጥ አቃቂ አካባቢ ዛሬ ማለዳ በደረሰ የመኪና አደጋ የአንድ ሰው ህይወት ሲያልፍ በሶስት ሰዎች ላይ ደግሞ የመቁሰል ጉዳት ደረሰ በወረዳ 27 ቀበሌ 08 እና 09 ክልል ውስጥ ከጧቱ 1 ሰዓት ከ45 ላይ ይህ አደጋ የደረሰው ከደብረ ዘይት ወደአዲስ አበባ ይገባ የነበረው ተ	0	475

103930	ዝናብ በ114 ሄክታር ላይ የነበሩና የደረሱ ሰብሎችን አወደመ	ዝናብ	ጋደአ	ዝናብ በ114 ሄክታር ላይ የነበሩና የደረሱ ሰብሎችን አወደመ ዝናብ ዝናብአዜአ- 33በደቡብ ወሎ ዞን በጀማ ወረዳ በረዶ ቀላቅሎ ሰሞኑን የጣለው ዝናብ በ114 ሄክታር ማሳ ላይ የነበሩና የደረሱ ሰብሎችን ማውደሙን የወረዳው ማስታወቂያ ጽህፈት ቤት ገለፀ የጽህፈት ቤቱ ኃላፊ አቶ ረሽድ ሐሰን ዛሬ እንዳመለከቱት በወረዳው ዜሮ ስምንት ቀበሌ ውስጥ በረዶ ቀላቅሎ ያለማቋረጥ ለሁለ	0	225
103935	በመኪና አደጋ ሁለት ሰዎች መሞታቸውና በአንድ ሰው ላይ የአካል ጉዳት መድረሱ ተገለጠ ።	አደጋ	ጋደአ	በመኪና አደጋ ሁለት ሰዎች መሞታቸውና በአንድ ሰው ላይ የአካል ጉዳት መድረሱ ተገለጠ አደጋ አደጋ በአዳማ ከተማ ከትናንት በስቲያ በደረሰው የመኪና አደጋ ሁለት ሰዎች መሞታቸውንና በአንድ ሰው ላይ ከባድ የአካል ጉዳት መድረሱን የልዩ ዞኑ ፖሊስ አስታወቀ በልዩ ዞኑ ፖሊስ የትራፊክ ክፍል ሃላፊ የመቶ አለቃ አህመድ ኩሩ ዛሬ እንደገለጡት ይሄው ቶዮታ አነስተኛ የበ	32	475