HARAMAYA UNIVERSITY

POST GRADUATE PROGRAM DIRECTORATE

SEMANTIC DOCUMENT CLUSTERING BASED INDEXING FOR AFAAN OROMO LANGUAGE INFORMATION RETRIEVAL SYSTEM

MSc. Thesis Proposal

BY:

BELETE BOGALE

College: Computing and Informatics

Department: Information Science

Program: Information Science

Principal Advisor: Dr.Million Meshesha (PHD)

Co-Advisor: Prof. Saravanan Madderi Sivalingam (PHD)

December 2017

Haramaya University, Haramaya

Continued…

LIST OF ACRONYMS AND ABBREVIATIONS

CLIR	Cross Language Information Retrieval
IR	Information Retrieval
IRS	Information Retrieval System
LSI	Latent Semantic Indexing
LC	Lexical Chain
MR	Relevance Model
NLP	Natural Language Processing
SVD	Single Value Decomposition
VSM	Vector Space Model

LIST OF TABLES

1. INTRODUCTION

1.1. Background of the Study

Information Retrieval is defined as finding relevant documents that satisfies information need of users from unstructured large collection corpus (Christopher et al, 2008). Information Retrieval is one of the major branches of Information Science discipline (Moukdad, 2003). The trend in information storage and retrieval can be traced back to 2000 BC when people of Sumerians chose special place to store clay tablets with cuneiform inscription (Christopher et al, 2008). After they understand that their work is efficient on use of information, they developed special categorization system that identifies every tablets and its content.

One of the major evolutions in Information Retrieval is invention of print machine in 1450 A.D (Judit and Gutman, 2003). A German goldsmith Johannes Gutenberg invented the first movable printer, thousand years later after Chinese invented paper which provides means for disseminating and storing knowledge. Gutenberg‘s aim was allowing direct access to mass information that was contained in the Bible and other scholarly works.

At the beginning of the 1990s, a single fact changed once and for all the perceptions towards Information Retrieval was the introduction of the World Wide Web (Baeza-Yates and Ribeiro-Neto, 2000). The Web is becoming a universal repository of human knowledge and culture which has allowed unprecedented sharing of ideas and information in a scale never seen before (Baeza-Yates and Ribeiro-Neto, 1999).

An Information Retrieval system is system that stores and manages information on documents and also enables users to find the information they need (Rijsbergen, 1979). It returns documents that contain answer to users query rather than explicit answer to their information need. Most of the time retrieved relevant documents satisfy users’ information needs, whereas retrieved irrelevant documents are not satisfying users’ information needs (Christopher et al, 2008).

Information Retrieval has two main subsystems (Deerwester et al, 1989), Indexing and Searching. Indexing is an offline process of representing and organizing large document

collection using indexing structure such as Inverted file, sequential files and signature file to save storage memory space and speed up searching time. Searching is on the other hand an online process of relating index terms to query terms and return relevant hits to users query.

Now a days with the advent of digital databases and communication networks, huge repositories of textual data have become available to a large public (Thomas, 1999). Although the use of elaborate ergonomic elements like computer graphics and visualization has proven to be extremely fruitful to facilitate and enhance information access (Thomas, 1999). The need to store and retrieve written information became increasingly important over centuries, especially after the inventions of scientific paper.

(Christopher et al, 2008). The purpose of an Information Retrieval (IR) system is to process a collection of documents as per requests for information, identify and retrieve from the corpus certain documents in response to the information requests (Marco and Kathleen, 2009).

General search engines on the Web such as Google, Yahoo, and MSN are among the popular tools to search for, locate, and retrieve information, and their use has been growing fast (Moukdad, 2003). These engines handle English queries more or less in the same way, but their handling of non-English queries is different from how these queries are handled by non-English search engines that were designed for specific languages. Most general search engines enable people to search using English language. But people usually want to get the information they need in the language that they can understand.

In fact there is no perfect Afaan Oromo Information Retrieval system which retrieves all relevant documents and no irrelevant document (Kula, 2007). This is due to the semantically related words are not taken into account; which can cause problems (Neepa and Sunita, 2012). There are actually two sides to the issue; they are broadly classified into synonymy and polysemy (Thomas, 1999). Synonymy describes the fact that there are many ways to refer to the same object. Users in different contexts or with different needs, knowledge, or linguistic habits will describe the same information using different terms (Thomas, 1999). Polysemy refer to the general fact that most words have more than one distinct meaning (Thomas, 1999). This can be when same term is used in different contexts or by different people.

There are various methodologies and techniques tried so as to make effective and efficient way of retrieving documents from very large and unstructured corpus (Recardo, 1999). Some of the scientific approaches to resolve the problems are ontology based Information Retrieval

System, latent semantic indexing, query expansion and reformulation, statistical method for semantic indexing.

Ontology based retrieval is an approach that uses the knowledge of the language for document retrieval purpose (Recardo, 1999). It works by associating of concepts from knowledge base to documents and queries, for example synonymous of the terms can be found using knowledge base. For a given query first concepts are extracted from the query itself. Then the set of concepts associated with each document is extracted from corpus. Next, these two sets are compared using simple metric, which expresses the similarity between a document ***Di*** and given query ***Q***. Based on the similarity measure relevant documents are retrieved.

The ontology based retrieval system needs intensive work in linguistic, in the domains, and Natural Language Processing (Thomas, 1999). Therefore it is much more language and domain dependent (Recardo, 1999). Ontologies based indexing and document clustering control, synonym and polysemy of words which can basically improve the precision and recall of document in Information Retrieval system (Thomas, 1999).

Latent Semantic Indexing (LSI) is also a method that helps to overcome the problems of lexical matching (Christopher, 2009). It assumes that there is latent structure in word usage that words are partially obscured by variability in words choice. On the other hand, most of the current document clustering algorithms does not consider the semantic relationships which produce unsatisfactory clustering results (Neepa and Sunita, 2012). While semantic document clustering plays an important role in providing intuitive navigation and browsing mechanisms by organizing such large amounts of information into a small number of meaningful clusters (Andreas et al, 2003).

1.2. Statement of the Problem

Today, in Ethiopia Afaan Oromo is among the major languages that are widely spoken and it is considered to be one of the five most widely spoken languages from thousand languages of Africa (Abera, 1988). Afaan Oromo, although relatively distributed within Ethiopia and some neighboring countries like Kenya and Somalia (Ibrahim, 2015). Afaan Oromo is part of the Lowland East Cushitic group within the Cushitic family of the Afro-Asiatic phylum (Etnologue, 2009), unlike Amharic (an official language of Ethiopia) which belongs to Semitic family languages.

Afaan Oromo has a very rich morphology like other African and Ethiopian languages (Ibrahim, 2015). With regard to the writing system, Qubee (Latin-based alphabet) has been adopted and became the official script of Afaan Oromo since 1991 (Abera, 1988). Currently Afaan Oromo is an official language of Oromia Regional State (which is the largest region in Ethiopia) and used as an instructional media for primary and junior secondary schools of the region (Ibrahim, 2015). Furthermore, literature, newspapers, magazines, educational resources, news, online education, books, entertainment Medias, videos, pictures, official documents and religious writings are available on the Internet. Not only this but also humans to interact more naturally with computers, one has to deal with the potential ambivalence, impreciseness, or even vagueness of user requests, and has to recognize the difference between what a user's might say or do and what she or he actually meant or intended (Thomas, 1999). Therefore these huge amounts of information available in electronic format both on the Internet and on offline are need potentially powerful Information Retrieval system with good performance.

There were several research works done in Afaan Oromo Information Retrieval to enhance the effectiveness of the system, especially on text retrieval system;

Tesfaye, (2010) was conducted research on IR, designed and developed search engine for Afaan Oromo language. The search engine mainly consists of three components – crawler, indexer, and query engine that were optimized for Afaan Oromo. In this study query was posed to the search engine, the search was take place using the default OR operator of Lucene search on the terms in the query. This causes more number of documents to be retrieved. In effect, the precision was being negatively affected.

Gezehagn, (2012) further intended to make possible retrieval of Afaan Oromo text documents by applying techniques of modern Information Retrieval system. This study was basically applied Vector Space Model of Information Retrieval system for indexing and searching relevant document from Afaan Oromo text corpus. The performance of proposed system in this study was on the average (57.5%) precision and (62.64%) recall. While the challenging tasks in the study was absence of standard corpus, handling synonymy and polysemy, inability of the stemmer algorithm to all word variants, and ambiguity of words in the language.

Recently Wegari, (2017) conducted a rule based root generation system for Afaan Oromo. This study was show that rule-based method can be used to develop root generation system for

Afaan Oromo. The proposed system of this study was evaluated with testing wordlist and has been experimentally shown that it improves the performance of state-of-the-art methods for the language. The experimental results show that the methodology proposed was effective in identifying root boundaries. A root generator system for concatenate languages could be a starting point for natural language processing related works particularly for morphological analysis and information retrieval systems.

However, those research works are more or less not consider semantic nature of the language. Having words in the information by itself does not meet the users information needs, semantic and contextual understanding of the user information need and information in the collection have to take into consideration to return effective result to the user. It happened because of different reasons but the major one is not properly address semantic natures of the language. The systems simply represent documents and queries as a "bag-of-word" with its weight (Bo-Yeong, 2013). When users give some query to the system the system tried to match those documents which have terms in the query and the documents will be automatically retrieved but not otherwise. There are several problems tied to the simple "bag-of-words" representation of documents (Bo-Yeong, 2013). Two of them are word-sense ambiguity (polysemy and synonymy) and semantic document clustering (some documents are not retrieved because they do not share terms with the query).

Hence the aim of this study is to design Semantic Document Clustering Based Indexing techniques for organizing Afaan Oromo document corpus based on their similarity for effective searching to retrieve relevant documents.

To this end, this study will explore and answer the following research questions:

- What are the unique features of word formations in Afaan Oromo language writing system?
- What are the suitable text preprocessing applied to documents corpus for semantic document clustering based indexing?
- Could a statistical approach combined with a linguistic features have better result in semantic index construction?
- Could semantic document clustering based indexing improve performance of Afaan Oromo Information Retrieval system?

1.3. Objectives of the Study

1.3.1. General Objective

The general objective of this study will be to design semantic document clustering based indexing techniques so as to improve the performance of Afaan Oromo information retrieval system in retrieving relevant documents as per users query.

1.3.2. Specific Objectives

In order to accomplish the above general objective the following specific objectives are formulated:

- To understand basics of Afaan Oromo linguistic features, techniques and approaches in IR
- To perform text preprocessing and extract Afaan Oromo semantic texts.
- To identify and experiment better semantic indexing and semantic document clustering algorithm.
- To design the framework and develop proposed information retrieval system for Afaan Oromo language.
- To evaluate the performance of the framework using Information Retrieval effectiveness measures (precision, recall, and F-measure).

1.4. Scope of the Study

This study will specifically focus on designing Afaan Oromo Information Retrieval System that effectively improves the performance of the system and it only used by individuals who can read and write Afaan Oromo languages. It mainly implements the semantic document clustering based indexing which involves semantic indexing, semantic document clustering and searching Afaan Oromo textual document corpus. To do these the study will able to carry out semantic document clustering using Lexical Chain (LC) which basically include (dataset preprocessing, mapping dictionary (WordNet), generation of candidate words, creations of lexical chains using candidate words, selecting the best chains based on heuristics, using those chains for clustering, and formation of clusters) and Latent Semantic Indexing (LSI) approach for semantic indexing. LSI is a variant of the vector retrieval method or SVD that exploits dependencies or “semantic similarity” between terms. LSI became famous as one of the first

IR techniques exhibiting effectiveness in dealing with the problems of synonymy, polysemy and dimension reduction (Frakes and Baeza-Yates, 1992). It helps with multiple meanings because the meaning of a word can be conditioned not only by other words in the document but by other appropriate words in the query not used by the author of a particular relevant document.

Lexical chains will help in forming topic based clusters as the chains will reflect the themes of the clusters based on the candidate words present in the chains. The document that will relate to the lexical chain of a particular cluster will be placed under that cluster. Thus, based on the relatedness among the documents and the initial clusters of lexical chains, the clusters will be formed.

This study will able develop and test using Afaan Oromo news article documents that will be collected from Oromia Broadcast Network (OBN). As a result of time factor (it takes more time), to prepare all relevant Afaan Oromo documents, limited corpus will be used for developing and evaluating performance of the this Afaan Oromo IR system. On the other hand data types, such as image, video, audio and graphics are out of focus of this study.

1.5. Significance of the Study

The study is concerned with Afaan Oromo Information Retrieval system so as to improve its performance. When the semantic document clustering based indexing will be implemented in Afaan Oromo retrieval system, it helps information seekers to retrieve information needs. The system retrieves as per the information needs of the user by retrieving documents which have similar meaning to the query formulated by the user rather than exact terms match. The primary and target beneficiaries from this study will Afaan Oromo language native speakers who can read, understand and fluent enough to produce queries to search information they need. Thus, these users will able to search and retrieve documents in Afaan Oromo.

In addition to that the computational the cost of the system will be reduced. During the system looking for documents, the system not goes through the whole documents of the corpus instead it will searches only from specific cluster which depends on specific relationship. Also, the system will contributes to future researchers in the area of Information Retrieval especially ontology based Information Retrieval system. Generally the research outcome gave benefit to individuals, groups, and future researchers.

2. LITERATURE REVIEW

2.1. Overview

This section is presents an overview of Information Retrieval in general, giving more emphasis to semantic indexing, semantic document clustering and Afaan Oromo linguistic feature. This is broadly divided into three sections. The first section discusses about concepts related to IR such as text preprocessing, indexing, components of IR system and semantic document clustering. Concepts like extraction semantics of the term and term weighting is also included. In the second section the features of Afaan Oromo writing system related to Information Retrieval is briefly discussed. The Afaan Oromo alphabets, numbers, punctuation marks, morpheme and parts of speech of Afaan Oromo language are also introduced. The third section focuses research works related to Afaan Oromo Information Retrieval system.

2.2. Information Retrieval

The need to store and retrieve written information became increasingly important over centuries, especially inventions like scientific paper (Christopher et al, 2008). Soon after computers were invented, people realized that they could use the machine for storing and mechanically retrieving large amounts of information. In 1945 Vinegar Bush published a ground breaking article titled “As We May Think” that gave birth to the idea of automatic access to large amounts of stored knowledge. In the 1950s, this idea materialized into more concrete descriptions of how to archives the text could be automatically searched (Christopher et al, 2009). Several key developments in the field happened in the 1960s. Most notable was the works of Gerard Salton and his students; they were laid the foundation for retrieval system. They developed Retrieval system and formulated a technique to evaluate the IR system, which is using still today (Singhal, 2008)

Information Retrieval is fast becoming the dominant form of information access, overtaking traditional database style searching (Christopher et al, 2008). It is defined as process of looking for material (usually documents) of an unstructured nature that satisfies an information need of the user from within large collections (usually stored on computers) (Christopher et al, 2009). IR can also cover other kinds of data and information beyond textual document; it could be image, multimedia or other data.

An Information Retrieval system does not inform (i.e. change the knowledge of) the user on the subject of his inquiry (Rijsbergen, 1979). It merely informs on the existence (or non-existence) and whereabouts of documents relating to his request. However, the computer is not likely to have stored the complete text of each document in the natural language in which it was written (Rijsbergen, 1979). It will have, instead, a document representative which may have been produced from the documents either manually or automatically.

Information Retrieval applications concerning textual documents use automatically generated free text index terms (post-coordinated), which are weighted by the statistical frequency of terms in documents and collections (Marco and Kathleen, 2009; Baeza-Yates et al, 2004). For many authors the purpose of an Information Retrieval (IR) system is to process files of records and requests for information, and identify and retrieve from the files certain records in response to the information requests (Marco and Kathleen, 2009).

2.2.1. Information Retrieval Model

A retrieval model specifies representations used for documents and queries, and how they are compared (Turtle and Croft, 1992) as it cited in (Potsdam, 2007). An Information Retrieval model is a formalization of the way of thinking about Information Retrieval (Potsdam, 2007). Such formalism can be defined in form of algorithms, mathematical formulas, etc.

A document model is a formal document representation used by an Information Retrieval model to obtain document similarities (Potsdam, 2007).The three most commonly known models are the vector space model, the probabilistic models, and the inference network model (Recardo, 1999).

The Vector Space Model (VSM) has been a standard model of representing documents in Information Retrieval for almost three decades (Salton and McGill, 1983; BaezaYates and Ribeiro-Neto, 1999).

In this model text is represented by a vector of terms from the documents and query (Recardo, 1999). The term is either word or phrase. Words in the vocabulary and their respective document wills construct two dimensional matrixes, called a vector space. Any text can then be represented by a vector in this two dimensional space. To assign a numeric score to a document retrieved by a query, the model measures the similarity between the query and vector created by vocabulary in the index versus documents in the corpus (Singhal, 2008). The

similarity between two vectors is once again not inherent in the model. Typically, the angle between two vectors is used as a measure of divergence between the vectors. The similarity can be measured using cosine similarity or dot product or others (Singhal, 2008).

Probabilistic Models is based on the general principle that documents in a collection should be ranked by decreasing probability of their relevance to a query or user information needs (Bruce, 1995). Probabilistic IR models estimate the probability of relevance of documents for a user queries (Bruce, 1995), this estimation is the essential part of the model, and this is a point where probabilistic models different from others. The probabilistic models works based on basic probability notation, the probability of relevance for document in the log (Fuhr and Buckley, 1991). In probabilistic IR models, parameters of the models relate to elements of the underlying representations (Fuhr and Buckley, 1991). In order to estimate these parameters, relevance feedback data is used.

Inference Network Model attempts to model document retrieval as an inference process in an inference network (Singhal, 2008). Most techniques used by IR systems can be implemented under this model (Recardo, 1999). In the simplest implementation of this model, a document instantiates a term with certain strength, and the credit from multiple terms is accumulated given a query to compute the equivalent of a numeric score for the document (Recardo, 1999). From an operational perspective, the strength of instantiation of a term for a document can be considered as the weight of the term in the document, and document ranking based on term strength of the documents.

2.2.2. Information Retrieval Process

Information Retrieval can be subdivided in many ways; it seems that there are three main areas of research which between them make up a considerable portion of the subject (Rijsbergen, 1979). They are: content analysis, information structures, and evaluation. Briefly the first is concerned with describing the contents of documents in a form suitable for computer processing; the second with exploiting relationships between documents to improve the efficiency and effectiveness of retrieval strategies; the third with the measurement of the effectiveness of retrieval. The term information structure covers specifically a logical organisation of information, such as document representatives, for the purpose of Information Retrieval (Rijsbergen, 1979).

2.2.3. Text Preprocessing

Document preprocessing is a stage before a document model construction, when data is analyzed, transformed and filtered from the document content (Potsdam, 2007). During the preprocessing phase, some important text operations can be performed. It is performed to control the size of the vocabulary. Some of the major operations are (Christopher et al, 2009; Rijsbergen, 1979):

- Lexical analysis
- Elimination of stop words
- Stemming
- Thesaurus construction.

Lexical analysis is a process of collecting tokenized terms which needs some other operation on it and it may be used for indexing (Christopher et al, 2009; Singhal, 2008). The tokens are taken from the document which may be word or phrase.

Elimination of stop words is the process of avoiding a word which has fewer relevancies to determine the content of the document (Potsdam, 2007). Sometimes, some extremely common words which would appear to be of little value in helping select documents matching a user need are excluded from the vocabulary entirely. In other word it is non-content bearing words. In English language, these words are most of the time articles, prepositions, sometime conjunctions and others. Therefore elimination of stopping word is very critical task to control the size of the vocabulary (Christopher et al, 2009; Singhal, 2008). Stemming is the process of finding the root of the words given word variants (Christopher et al, 2009). In any language we observe occurrence of variants of the same word into little modification in its morpheme. For example a word nation can be written as nationalized nationality, national, and others. So this different form of the same word comes to its root term.

Thesaurus is knowledge base which describes the meaning of a word and finds the synonymy or polysemy of the word. In addition to that it can tell us the context to which the words used. Therefore constructing thesaurus has very important impact on the Information Retrieval effectiveness (Singhal, 2008).

2.2.4. Automatic Indexing

An index language is the language used to describe documents and requests (Rijsbergen, 1979), the elements of the index language are *index terms*, which may be derived from the text of the document to be described, or may be arrived at independently. The transformation from a document text into a representation of text is known as indexing the documents (Sajendra and Ram-Kumar, 2012). Indexing can be defined as a process that collect, parse and store data to facilitate fast and accurate Information Retrieval (Clarck and Cormack, 1995).

Automatic indexing is the ability of a computer to scan large volumes of documents against a controlled vocabulary, taxonomy, thesaurus or ontology and use those controlled terms to quickly and effectively index large document collections (Rijsbergen, 1979).

Typically the indexing of a textual document is obtained through the identification of a set of terms or keywords which characterize the document content that is terms which describe the topics dealt with in the document (Lei Shi, 2005). The terms included in this set have not only to be representative of the topics covered in the documents, but they also need to be distinguishing, in that they should make it possible to discriminate one document against the other documents in the collection covering the same or similar topic.

Indexing techniques have been developed in order to make possible the identification of the information content of documents (be they text documents, hypermedia or multimedia ones) (Lei Shi, 2005). It simply means pointing to or indicating the content, meaning, purpose and features of messages, texts and documents (Xiao, 2010).

There are two major approaches for the automatic indexing of text documents (Lei Shi, 2005): the first is approach is statistical approach that rely on various word counting techniques, vector space model used this approach. The aim of statistical indexing is to capture content bearing words which have a good discriminating ability and a good characterizing ability for the content of a document (Linda, 2010). Discrimination ability means that the words are able to distinguish documents from one another. And those terms are going to be counted and will have some weight. The other approach is linguistic approach that involves syntactical analysis (Aurélie et al , 2008). It is based on knowledge base of the language it is working on (Latifur et al, 2013). Therefore, it needs ontology of the language to determine the terms which

describe a document. It can use the extraction of the semantics the document using content bearing words extracted from document that going to be indexed (Latifur et al, 2013).

Ontology

The Use of ontology enables to define concepts and relations representing knowledge about a particular document in domain specific terms (Paralic and Kostial, 1999). In order to express the contents of a document explicitly, it is necessary to create links (associations) between the document and relevant parts of a domain model, i.e. links to those elements of the domain model, which are relevant to the contents of the document.

Ontology is a collection of concepts and their interrelationships which can collectively provide an abstract view of an application domain (Latifur et al, 2013). The ontology (WordNet) lexical database is now quite large and offers broad coverage of general lexical relations in the English language (David et al, 2013). It describes the relationship between the words in different situations. WordNet has been employed as a resource for many applications in natural language processing (NLP) and Information Retrieval (IR). Word relationships in the database are useful for NLP and IR applications, are not necessarily appropriate for a general, sometime the WordNet can be domain-independent lexical database (Marti, 2008; Singhal, 2008).

Semantic Indexing

To improve the performance of a traditional keyword-based search, a documents should be represented with their concept rather than bag-of-words‘(Bo-Yeong, 2013). However; most of the previous works on indexing and Information Retrieval depend on lexical analysis and statistical methods. Using these techniques, it is difficult to abstract the semantics of the documents (Bo-Yeong, 2013). A better approach would allow users to retrieve information on the basis of a conceptual topic or meaning of a document (Barbara, 2000). One way of expressing their commonality is to think of a searcher as having in mind a certain meaning, which he or she expresses in words, and the system as trying to find a text with the same meaning (Deerwester et al, 1989). Success, then, depends on the system representing query and text meaning in a manner that correctly reflects their similarity for the human.

Semantic Indexing assumes that there's some underlying or context structure in word usage that's partially obscured by variability in word selection and widely used indexing technique is

Latent Semantic indexing (LSI) (Ding, 1999). Latent Semantic Indexing tries to beat the problems of lexical matching by using statistically derived conceptual indices instead of individual words for retrieval. In Latent Semantic Indexing (LSI), the original vector space representation of documents is replaced by a representation in the low-dimensional latent space and the similarity is computed based on that representation (Hofmann, 1999).

The goal of semantic indexing is to use semantic information (within the objects being indexed) to improve the quality of information retrieval. Compared to traditional indexing methods, based on keyword matching, the use of semantic indexes means that objects are indexed by the concepts they contain rather than just the terms used to represent them (Marco and Kathleen, 2009):

• A semantic index is inherently multidimensional, since any combination of properties cast into a document concept can serve as an indexing element.

• As a structured concept the indexing elements are not just attribute values, but can be based on complex descriptions of related objects.

• A semantic index as a whole is highly adaptable to patterns of usage. Indexing concepts can be added or removed at will, making it very dense and precise with respect to interesting sets of individuals, or very sparse in other less interesting areas.

• Since the index is actually a set of partial descriptions for the indexed objects, lots of information can be drawn from the index alone without accessing individual descriptions at all.

2.2.5. Term Frequency and Weighting

The term frequency and weighting for the IR model is handled by statistics (Potsdam, 2007), there are three main factors of term weighting: term frequency factor, term collection frequency factor and document vector length normalization factor. The end term weight might be constructed from all or a subset of mentioned factors. E.g., the inverse document frequency assumes that the importance of a term is proportional with the number of documents the term appears in (Salton, 1983), a document that mentions a query term more often has more to do with that query and therefore should receive a higher score. Thereof, assign to each term in a

document a weight for that term that depends on the number of occurrences of the term in the document (Christopher et al, 2009). This concept can be applied through assigning the weight; it is equal to the number of occurrences of term t in document d. This weighting scheme is referred to as term frequency and is denoted by ***tft, d***, and the document in its order (Christopher et al, 2009). The exact ordering of the terms in a document is ignored but the number of occurrences of each term is material (in contrast to Boolean retrieval). We only tried to capture information concerning to the number of occurrences of each term (Christopher et al, 2009).

Inverse Document Frequency

It will create a problem when we make all terms are equally important; the impact would be explicitly viewed when it comes to assessing relevancy against a query (Christopher et al, 2009). In fact certain terms have little or no discriminating power in determining relevance. An immediate idea is to scale down the term weights of terms with high collection frequency, defined to be the total number of occurrences of a term in the collection. The idea would be to reduce the tf weight of a term by factor that grows with its collection frequency (Christopher et al, 2009). document frequency of the term, ***dft***, defined to be the number of documents in the collection that contain a term **t**. its aim is to obtain document level statistics, which deals with the number of documents containing a term (Christopher et al, 2009). Using ***dft*** is difficult to measure the discrimination power of the term among the documents. To come over this problem, it is better to use inverse document frequency, this is logarithmic function of total number of document N over dft (Christopher et al, 2009). It is defined as:

$$\mathbf{idf = \log N/dft}$$

Tf-idf weighting

Combining the definitions of term frequency and inverse document frequency helps to produce a composite weight for each term in each document. The ***tfidf*** weighting scheme assigns to term ***t*** in document ***d*** given by

$$\mathbf{tfidf_{t,d} = tf_{t,d} \times idf_t}$$

tfidf $_{t,d}$ assigns to term ***t*** is a weight in document ***d*** that is

- Highest when **t** occurs many times within a small number of documents (thus lending high discriminating power to those documents);
- Lower when the term occurs fewer times in a document, or occurs in many documents (thus offering a less pronounced relevance signal);
- Lowest when the term occurs in virtually all documents.

2.2.6. Query Matching and Searching

Comparison of query and document representations is very important task of IR system, it is performed in order to retrieve documents that are more similar to the specific query (Recardo, 1999). The 'Search' is a systematic examination of information in a database, aiming in view to identify the items or objects, which satisfy particular preset criteria (Juban and Falguni, 2007). In other way, searching means the operation of locating a specific object in a given sequence of 'n' objects.

All search strategies are based on comparison between the query and the stored documents (Rijsbergen, 1979), sometimes this comparison is only achieved indirectly when the query is compared with clusters (or more precisely with the profiles representing the clusters). The distinctions made between different kinds of search strategies can sometimes be understood by looking at the query language that is the language in which the information need is expressed and the nature of the query language often dictates the nature of the search strategy.

2.3. Document Clustering

Cluster analysis is a sub-field in artificial intelligence and machine learning that refers to a group of algorithms that try to find a natural grouping of objects based on some objective metric (Christoper, 2010). Clustering is the most common form of unsupervised learning (Cambridge, 2009). No super-vision means that there is no human expert who has assigned documents to classes. Clustering is the process of partitioning a set of data objects into subsets (Recardo, 1999). Text clustering process deals with grouping of an unstructured collection of documents into semantically related groups (Maitri et al, 2015). It is commonly used technique in data mining, information retrieval, and knowledge discovery for finding hidden patterns or objects from a data of different category.

In the Information Retrieval (IR) field, cluster analysis has been used to create groups of documents with the goal of improving the efficiency and effectiveness of retrieval, or to determine the structure of the literature of a field (Edie, 2015). Cluster-based retrieval has as its foundation the cluster hypothesis, which states that closely associated documents tend to be relevant to the same requests (Rijsbergen, 1979). In other words, documents within a cluster should be as similar as possible; and documents in one cluster should be as dissimilar as possible from documents in other clusters (Cambridge, 2009).

Clustering is a powerful technique for large-scale topic discovery from text (Bjornar and Chinatsu, 1999). It involves two phases: first, feature extraction maps each document or record to a point in high-dimensional space, then clustering algorithms automatically group the points into a hierarchy of clusters. Clustering algorithms group a set of documents into subsets or clusters.

The two main types of cluster analysis methods are the nonhierarchical, which divide a data set of N items into M clusters, and the hierarchical, which produce a nested data set in which pairs of items or clusters are successively linked (Bader et al, 2009) (Edie, 2015). One of nonhierarchical partitioning clustering is K-means clustering, which is done by minimizing the sum of squares of distances between objects and their corresponding cluster centroid. It groups the objects into K clusters, and keeps iteratively moving the cluster centers and re-assigning objects into clusters, based on minimum distance to the closest cluster's centroid. The process terminates when cluster centers are not moved anymore and all objects have been assigned to their closest cluster center. Unlike hierarchical clustering, which groups data objects with a sequence of partitions, K-means (partitioned) clustering directly divides data objects into K clusters, without any corresponding hierarchical structure (Bader et al, 2009). For K-means clustering, the cosine measure is used to compute which document centroid is closest to a given document (Michael et al, 2009). While a median is sometimes used as the centroid for K-means clustering, we follow the common practice of using the mean. There are two basic approaches to generate a hierarchical clustering (Recardo, 1999): the first one is Agglomerative, which start with the points as individual clusters and, at each step, merge the most similar or closest pair of clusters. This requires a definition of cluster similarity or distance.

The other one is Divisive, which start with one, all-inclusive cluster and, at each step, split a cluster until only singleton clusters of individual points remain. In this case, we need to decide, at each step, which cluster to split and how to perform the split.

2.4. Afaan Oromo Language

2.4.1. Overview

Afaan Oromo is the second most widely spoken indigenous language in Africa next to Hausa in Nigeria (Demie, 1996). It is one of a highly developed language of the Cushitic languages spoken in Ethiopian, Somalia, Sudan, Tanzania, and Kenya (Ibrahim, 2015). From forty different Cushitic linguistic and cultural groups in Africa, the Afaan Oromo speakers are considered as one of the largest in terms of population and their language (Demie, 1996). In Oromia state, in Ethiopia, it is the official language used in courts, schools and administration (Demie, 1996).

Currently there are a growing number of publications in hard copies and vast amount of information in electronic formats for Afaan Oromo (Tilahun, 2008). This was made for the convenience of the Latin script for the writing of Afaan Oromo from the linguistics, pedagogic, and practical reasons. It is believed that many fold more text were written in Afaan Oromo since then than before. Afaan Oromo is a language that is used in a wide area in the country. According to (Gragg, 1976), quoted in (Wakshum, 2000), four major categories can be identified. These are: Western (Wellega, Iluababor, Kaffa and parts of Gojjam), Eastern (Harar, Eastern showa, and parts of Arsi and Bale), Central (Central Showa, Western Showa and possibly Wollo) and Southern (Parts of Arsi, Sidamo and Borena).

2.4.2. Afaan Oromo Writing Style

According to (Ladefoged, 1955), quoted in (Morka, 2001) some of the finer anatomical feature involved in speech production include the vocal cords, velum, tongue, teeth, palates, the alveolar ridge, the mouth, and lips. These anatomical components move to different positions to produce various sounds and are referred in articulators. Most of the characters of a sound are determined by the position of these articulators in the oral tract.

The Oromo Alphabet

The alphabets of Afaan Oromo language are often called “Qubee Afaan Oromoo (Ibrahim, 2015). Qubee has 33 characters representing distinct sounds. It has both capital and small

letters. Afaan Oromo has a considerable amount of glottal stops. The major representatives of sources of the sound in a language are the vowels and consonants. Afaan Oromo is a phonetic language, which means that is spoken in the way it is written (Wakshum, 2000). The Afaan Oromoo vowels represented by letters (a, e, o, u and i) are called “Dubbiftu/Dubbachiftu” in Afaan Oromo and the consonants known as “dubbifamaa” in Afaan Oromo.

Punctuation Marks

All punctuation marks that are used in English are also used for the same purpose in Afaan Oromoo except the apostrophe. Unlike its use to show possession in English, it is used as a symbol to represent a hiccup (called *hudhaa*) sound in Afaan Oromoo writing. An apostrophe, and less commonly a hyphen, is used “’” represent this sound in writing. Sometimes an H, which represents the closest glottal sound, is also used in place of an apostrophe (Ibrahim, 2015), for a reason to be apparent later, the apostrophe will be considered as a distinct symbol (say, as the 27th letter of the alphabet).

2.4.2. Morphology

Natural Language Processing is the application of computational models to tasks involving human language text) (Ibrahim, 2015). Automatic document clustering and text preprocessing therefore requires the studying of the morphology of the language on which it applies. The smallest linguistic sign is the morpheme, a meaningful form that cannot be divided into smaller meaningful parts (Algeo, 2010) (Ibrahim, 2015). In every language, whether it is spoken or written, every meaningful pattern has its own structure and the elements of language related to each other in understandable manner. Words are the basic elements of a language and are formed from morphemes which constitute the smallest meaningful unit of speech in a language; this is also true for Afaan Oromoo which has its own rules of words and/or sentence structure (Assefa, 2005).

Morphology is a way of studying the language words structure (Algeo, 2010). It is about the way words are put together, their internal structure. It is a level at which the structure of language is analyzed. It deals with the analysis and examination of meaningful units of forms which make up sentences. The smallest meaningful units of forms are called "morphemes" which are either "free" or "bound" (Richard, 2003). A free morpheme can occur on its own whereas bound morphemes do not occur alone. Bound morphemes are of three types, these

are, prefix attached to the initial positions, infix inserted in the middle and suffix attached to the final position of the word. All the three types of bound morphemes are called "affixes".

The Afaan Oromo Morphology

Every language has its own morphological structure that defines rules used for combining the different components the language may have (Wakshum, 2000). The Afaan Oromo language for instance is basically different in its morphological structure from French, Arabic, English and other languages that use Latin characters.

There are a number of word formation processes in Afaan Oromo (Wakshum, 2000), affixation and compounding are among these word formation processes.

Affixation is generally described as the addition of affixes at the beginning, in between and/or at the end of a root/stem depending on whether the affix is prefix, infix or suffix (Wakshum, 2000).

Afaan Oromo word is composed of two parts (Ibrahim, 2015): (1) the root (base morpheme), which generally consists of basic sound and provides the basic lexical meaning of the word, and (2) the pattern, which consists of prefixes and/or suffixes and gives grammatical meaning to the word. Thus, the root /Bar/ combined with the pattern /-e/ gives Bare ‘learned,’ whereas the same root combined with the pattern /-te/gives Barte ‘she learns’.

Compounding is the joining together of two linguistic forms, which functions independently suffix (Wakshum, 2000). Examples compound nouns include; abbaa-buddenaa ‘step father’ from abba- ‘father’ and buddena ‘food’.

2.5. Related Works

Antoniol et al. (2002) have published a series of papers on recovering code to documentation traceability. They use Information Retrieval as well, however with another approach. They rely on external documentation as text corpus, and then they query the documentation with identifiers taken from source code to get matching external documents (Adrian Kuhn et al, 2006).

Previous research on cluster-based retrieval has been inconclusive as to whether it does bring improved retrieval effectiveness over document-based retrieval (Xiaoyong, 2010). Recent developments in the language modeling approach to IR have motivated them to re-examine

this problem within this new retrieval framework. They propose two new models for cluster-based retrieval and evaluate them.

The study used cluster based document clustering model, it viewed as a mixture model of three sources: the document, the cluster/topic the document belongs to, and the collection. A relevant document assumed being generated by this mixture model. Both partitioning and hierarchical agglomerative clustering algorithms have been studied in the context of IR. They used a three-pass K-means algorithm as an example of partitioning methods in their static clustering experiments, primarily motivated by its efficiency.

To experiment their work they used over six data sets taken from TREC. Two sets of experiments are performed in this study. The first set of experiments investigates whether a simple language model of clusters can be used to rank clusters. And the second set of experiments examines the effectiveness of cluster-based retrieval using CBDM model in the context of query likelihood retrieval and the relevance model (RM), for both static clustering and query-specific clustering. Both experiments have given promising result to clustering based retrieval is more effective than the traditional research engine.

2.5.1. Local Related Research Works

There were several research works done in Afaan Oromo language Information Retrieval to enhance the retrieval performance of the system. These research works are discussed below based on their chronological order.

Abera, (2009) attempted to develop and adopt processing tools for Afaan Oromo text classification and investigate the application of machine learning techniques for automatic classification of Afaan Oromo news items. The data source for this research was Afaan Oromo news items obtained from Radio Fana Share Company.

In this research, tools for pre-processing Afaan Oromo news items such as tokenization, removal of extraneous characters, removal of stop-words and removal of affixes from the words was prepared to facilitate the experimentation process for the automatic classifiers. The result of the experimentation is encouraging, the best result (accuracy) from all applied to classify a given document.

Tesfaye, (2010) was conducted research on IR, designed and developed search engine for Afaan Oromo language the performance evaluation of this search engine is conducted using selected set of documents and queries. According to precision-recall measures employed, 76%

precision on the top 10 results and an average precision of 93% are obtained. Experiment on some specific features of the language against the design requirements is also made.

In this case the average precision for the top 10 results is 76%. This indicates that displaying only the first few top results, top 10 in this case, entails better user satisfaction as compared to displaying all the results. When a query is posed to the search engine, the search was take place using the default OR operator of Lucene search on the terms in the query. This causes more number of documents to be retrieved. In effect, the precision was being negatively affected.

Daniel, (2011) explored the development of a corpus-based CLIR system which makes use of word based query translation for Afaan Oromo-English language pairs and evaluation of the system on a corpus of test documents and queries prepared for this purpose. In this study documents are collected from Bible chapters, legal and some available religious documents.

Evaluation of the system was conducted by both monolingual and bilingual retrievals. In the monolingual run, the Afaan Oromo queries are given to the system and Afaan Oromo documents are retrieved while in the bilingual run the Afaan Oromo queries are given to the system after being translated into English to retrieve English documents. For the bilingual run translation of Afaan Oromo queries into their English equivalent is done by using bilingual dictionary constructed from the collected parallel corpora. Maximum average precision and recall value of this study was 0.468 and 0.316 are obtained for the Afaan Oromo and English documents respectively.

Gezehagn, (2012) further intended to make possible retrieval of Afan Oromo text documents by applying techniques of modern Information Retrieval system. Information Retrieval is a mechanism that enables finding relevant information material of unstructured nature that satisfies information need of user from large collection. Afaan Oromo text retrieval developed in this study has indexing and searching parts. Basically Vector Space Model of Information Retrieval system was used to guide searching for relevant document from Oromiffa text corpus. The model is selected since Vector space model is the widely used classic model of Information Retrieval system. The index file structure used is inverted index file structure.

In this study experimental result shows that the performance is on the average (57.5%) precision and (62.64%) recall. The challenging tasks in the study are absence of standard

corpus, handling synonymy and polysemy, inability of the stemmer algorithm to all word variants, and ambiguity of words in the language.

Eyob, (2013) tried to design and develop a corpus based Afaan Oromo–Amharic cross lingual Information Retrieval system so as to enable Afaan Oromo speakers to retrieve Amharic information using Afaan Oromo queries. This approach selected to be followed in the study is corpus based, particularly parallel corpus. For this study parallel documents including news articles, bible, legal documents and proclamations from customs authority were used. The system is tested with 50 queries and 50 randomly selected documents. Two experiments were conducted, the first one by allowing only one possible translation to each Afaan Oromo query term and the second by allowing all possible translations. The retrieval effectiveness of the system is measured using recall and precision for both monolingual and bilingual runs.

Accordingly, the first experiment returned a maximum average precision of 0.81 and 0.45 for monolingual (Afaan Oromo queries) and bilingual (translated Amharic queries) run. The result of the second experiment showed better result of recall and precision than the first experiment. The result obtained in the second experiment is a maximum average precision of 0.60 for the bilingual run and the result for the monolingual run remained the same. From these results, he was concluded that, cross lingual Information Retrieval for two local languages namely Afaan Oromo and Amharic could be developed and the performance of the retrieval system could be increased with use of larger and clean corpora.

Daniel, Ramesh and Dereje, (2015) attempted to develop Afaan Oromo-English CLIR system which enables Afaan Oromo native speakers to access and retrieve the vast online information sources that are available in English by writing queries using their own (native) language. Evaluation of the system was conducted by both monolingual and bilingual retrievals. The performance of the system was measured by recall and precision. As they was mentioned in their study larger documents were retrieved for the monolingual run (i.e. for the retrieval of documents by using baseline queries of Afaan Oromo) than for the bilingual run (i.e. for the retrieval of English documents using Afaan Oromo queries after being translated into English). The performance of the system propose was highly affected by the size, reliability and correctness of the corpus used for the study. A maximum average precision of 0.468 and 0.316 for Afaan Oromo and English was obtained respectively.

Recently Wegari, (2017) conducted a rule based root generation system for Afaan Oromo. In their investigation; they have been shown that rule-based method can be used to develop root generation system for Afaan Oromo. The system mainly used as starting point to develop a complete morphological analysis and information retrieval for the target language. The experimental results show that the methodology proposed is effective in identifying root boundaries.

3. METHODOLOGY OF THE STUDY

The methodology section describes actions to be taken to investigate a research problem and the rationale for the application of specific procedures or techniques used to identify, select, process, and analyze information applied to understanding the problem, thereby, allowing the reader to critically evaluate a study's overall validity and reliability (Hevner and Chatterjee, 2015). A methodology is "a system of principles, practices, and procedures applied to a specific branch of knowledge." Such a methodology might help researchers to produce and present high quality science research. In order to realize this study, the following methodology will be applied, including review other related literatures to have background knowledge on the domain area of the study and the problem solving methods.

3.1. Research Design

The development of an IR system involves various techniques and general principles of information retrieval system design, so as the system aims to overcome the semantic problems in indexing and clustering relevant documents.

Basically the generic information retrieval system incorporates two major components: indexing and searching. However, this semantic document clustering based indexing study will have three major components; semantic indexing, semantic document clustering, and searching. Under each components there are number of constituent elements so as to make it works well. Design science research is inherently a problem-solving process (Hevner and Chatterjee, 2015). The fundamental principle is that knowledge and understanding of a design problem and its solution are acquired in the building and application of an artifact. Design science research approach will be used in this research because of the following reasons: design science research, which is centered towards practical problem solving, includes prescriptive or solution-oriented knowledge, it is important in a discipline oriented to the creation of successful artifacts (Ken Peffers et al, 2007), it follows a pragmatic research paradigm that calls for the creation of artifact to solve real-world problems and to address promising opportunities.

3.2. Corpus Preparation

In this study both primary and secondary data collection method will be employed in order to collect the required corpus which written Afaan Oromo language. Since carefully study of linguistic features is needed for Afaan Oromo Information Retrieval system, an interview and consultation will be made with the language experts. Several preprocessing computation will be applied to collected data in order to design and develop proposed system. Generic IR system first, including; corpus preparation, lexical analysis, elimination of stop words, stemming.

After document preprocessing indexing will be the next step with representing in information retrieval system. This logical representation of documents using its content bearing words is inverted file (Christopher, 2011). Inverted index, or sometimes inverted file, has become the standard term in information retrieval. Each item in the list, which records that a term appeared in a document is conventionally called a posting. So as to have fast retrieval time the researcher need to build the index in well organized and structured manner. Here after index terms are extracted, the semantic relations between the terms and concepts contained in an unstructured collection of text will be identified, then semantic document clustering will be applied to retrieve clustered relevant documents.

3.3. Implementation Tools

Python programming language and Windows environment will be used to develop the intended prototype. Python was developed by Guido van Rossum in the late eighties and early nineties at the National Research Institute for Mathematics and Computer Science in the Netherlands (Halterman, 2011) and it is derived from many other languages, including ABC, Modula-3, C, C++, Algol-68, SmallTalk, UNIX shell, and other scripting languages. Higher-level programming languages like Python allow programmers to express solutions to programming problems in terms that are much closer to a natural language (Halterman, 2011). The reason to use Python programming language is; on one hand, the exposure of the researcher to the language and on the other hand, due to the fact that python is used for develop clustering, indexing and searching in IRS. Not only these but also Python is dynamic programming language that is used in a wide variety of application domains. It is simple,

strong, involves natural expression of procedural code, modular, dynamic data types, and embeddable with in applications as a scripting interface (Python Software Foundation, 2012).

3.4. Evaluation Procedure

The proposed retrieval system will be tested and evaluated to ensure the performance of the system whether it meets the objective of the study or not. To do this, corpus is prepared, queries are constructed and relevance judgment is done for evaluating effectiveness of this work, in Information Retrieval system corpus can be used for evaluation of the system. In addition to this, the standard effectiveness measures; precision, recall and F-measure will be used to evaluate the performance of the system in retrieving relevant documents.

4. WORK PLAN

4.1. Work plan (2017/18)

No	Activities	Nov. 2017	Dec. 2017	Jan. 2018	Feb. 2018	Mar. 2018	Apr. 2018	May. 2018
1	Literature review							
2	Writing and defense proposal							
3	Requirement analysis							
4	Data collection							
5	Corpus preparation							
6	Modeling, Design architecture and prototyping							
7	System testing (unit, functional and integration testing)							
8	Writing over all summarization, analysis, interpretation and conclusion							
9	Thesis submission							
10	Thesis defense							

5. BUDGET BREAKDOWN

The budget proposal is done on the bases of lump sum and the budget required for the completion of the research is listed in the table below.

5.1. Stationary

No	Items	Unit	Quantity	Unit price (Birr)	Total price (Birr)
1	Printing paper	Packet	3	100.00	300.00
2	Pen	Packet	1	100.00	100.00
3	Pencil	No	32	32.00	32.00
	Marker	Packet	1	100.00	100.00
4	Note book	No	3	50.00	150.00
5	Compact disc (CD-RW)	No	2	25.00	50.00
6	Flash disc (8 GB)	No	2	200.00	400.00
7	Storage Device (Hard disk)	No	1	2450.00	2450.00
8	Scientific calculator	No	1	200.00	200.00
	Subtotal				**3782.00**

5.2. Transportation Cost

No	Person	From	To	No of trip(s)	Unit price (Birr)	Total cost (Birr)
1	Researcher	Harar	Addis Ababa	1	192.00	192.00
2	Researcher	Addis Ababa	Jimma	1	200.00	200.00
3	Researcher	Jimma	Addis Ababa	1	200.00	200.00
4	Researcher	Addis Ababa	Meda- welabu	1	300.00	300.00
5	Researcher	Meda- welabu	Addis Ababa	1	300.00	300.00
6	Researcher	Addis Ababa	Harar	1	192.00	192.00
	Subtotal					**1384.00**

5.3. Miscellaneous Expenses

No	**Description**	**Unit price (Birr)**	**Quantity**	**Total cost (Birr)**
1	Corpus printing	1.00	1200	1,000.00
2	Thesis printing	1.00	6	480.00
3	Thesis binding	-	-	600.00
5	Subtotal			**2080.00**

5.4. Personnel Costs

No	Title	Purpose	No of Person	No of days	Rate/day (Birr)	Total exp. (Birr)
1	Researcher	Data collection and Corpus preparation Per dime	1	20	206.00	4120.00
2	Day Worker	Data organizing/ Encoding	13	5	100.00	6500
3	Advisors	Advisor per dime	2	-	-	4120.00
4	Advisors	Supervision fee	1	-	-	3,000.00
	Subtotal					**17740.00**

5.5. Summary of Expenses

No	Description	Total expenses (Birr)
1	Stationery	3782.00
2	Transport cost	1384.00
3	Miscellaneous expenses	2080.00
4	Personnel cost	17740.00
5	Total Expenses	**24986.00**

Budget source: MOE

6. REFERECE

Andreas et al. (2003). Wordnet Improves Text Document Clustering. *n Proc. of the SIGIR 2003 Semantic Web Workshop*.

Chris et al. (1998). Optimizations of Inverted Vector Searches. *SIGIR*.

Abera. (1988). Long Vowels in Afaan Oromo: A Generative Approach. *MSc Thesis*.

Abera, D. (2009). Automatic Classification of AFaan Oromo News Text:In Case of Radio Fana. *Msc Thesis Addis Ababa University*.

Abey. (2011). Semantic Based Query Expansion Technique for Amharic IR. *MSc Thesis, School of Information Science*.

Adrian Kuhn et al. (2006). Semantic Clustering: Identifying Topics in Source Code. *Journal of Information Systems and Technologies*.

Antoniol et al. (2002). Recovering Traceability Links Between Code and Documentation. *IEEE Transactions on Software Engineering, 28*(10), 970-983.

Assefa, W. (2005). Development of Morphological Analyzer for Afaan Oromo. *Master thesis at Faculty of Informatics*.

Atelach and Lars. (2009). An Amharic Stemmer : Reducing Words to Their Citation Forms. *Journal of KTH*.

Aurélie et al . (2008). Automatic Indexing of Specialized Documents:Using Generic vs. Domain-Specific Document Representations. *National Library of Medicine*.

Bader et al. (2009). Document Clustering of Scientific Texts Using Citation Contexts. *Springer Science Business Media, LLC*.

Baeza-Yates and Ribeiro-Neto. (2000). *Information Retrieval.* US Department of Education Grant.

Baeza-Yates and Ribeiro-Neto. (1999). Modern Information Retrieval. *Journal of ACM*.

Baeza-Yates et al. (2004). Web Dynamics, Structure and Page Quality. *Springer*, 93-109.

Barbara. (2000). Latent Semantic Indexing: An Overview. *INFOSYS 240 Spring*.

Bethelem. (2002). N-gram-Based Automatic Indexing for Amharic Text. *Msc thesis school of information science Addis Ababa*.

Bjornar Larsen and Chinatsu Aone. (1999). Fast and Effective Text Mining Using Linear-time Document Clustering. *SRA International, Inc.*

Bo-Yeong. (2013). A Novel Approach to Semantic Indexing Based on Concept, Department of Computer Engineering. *Kyungpook National University*.

Bruce. (1995). *What Do People Want from Information Retrieval?'*. Retrieved October 2017, from http://www.dlib.org/dlib/november95/11croft.html

Cambridge . (2009). The Term Vocabulary and Posting Lists.

Cambridge. (2009). Flat Clustering. *Online edition (c) Cambridge UP*.

Cambridge. (2010). Evaluation of Information Retrieval System.

Choudhary and Bhattacharyya. (2002). Text Clustering Using Semantics. *in Proc of the 11th International World Wide Web Conference.*

Christoper, I. (2010). Document Clustering. *Master of Science Thesis, University of Gothenburg*.

Christopher et al. (2008). *Introduction to Information Retrieval.* New York: Cambridge.

Christopher et al. (2009). Introduction to Information Retrieval. *Cambridge Online edition* .

Christopher, D. (2011). An Introduction to Information Retrieval. *online edition*.

Clarck and Cormack. (1995). Dynamic Inverted Index for a Distributed Full text Retrieval System. *Journal of Tech. Rep MT-95-01, University of Waterloo*.

Daniel, B. (2011). Afaan Oromo-English Cross Lanaguage Information Retrieval (CLIR): A Corpus Based Approach. *Msc Thesis*.

Daniel, Ramesh and Dereje. (2015). A Cross Lingual Information Retrieval (CLIR) System for Afaan Oromo-English using a Corpus Based Approach. *International Journal of Engineering Research & Technology (IJERT), 4*(5).

Daphe Koller et al. (1997). Hierarchically Classifying Documents Using Very Few Words. *Proceedings of the 14th International Conference on Machine Learning (ML)*, (pp. 170-178).

David et al. (2013). An Ontology-Based Information Retrieval Model,. *Universidad Autónoma de Madrid, Campus de Cantoblanco, 11*.

Deerwester et al. (1989). Indexing by Latent Semantic Analysis. *Journal of the ACM*.

Deerwester et al. (n.d.). Indexing by Latent Semantic Analysis.

Demie, F. (1996). Historical Challenges In The Development Of The Oromo Language. *Journal of oromo studies*, 18-25.

Ding. (1999). A Similarity-based Probability Model for Latent Semantic Indexing. *Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 59-65). ACM SIGIR.

Edie, R. (2015). Clustering Algorithm. *Journal of American Society for Information Science, 35*, 268-276.

Etnologue. (2009). Show Language. *Online edition*.

Eyob, N. (2013). Afaan Oromo-Amharic Cross Lingual Information Retrieval: Corpus Based Approach. *Msc. Thesis Addis Ababa University School of Graduate Studies* .

Frakes and Baeza-Yates. (1992). Information Retrieval Data Structures and Algorithms. 504.

Fuhr and Buckley. (1991). A Probabilistic Learning Approach for Document Indexing. *ACM Transactions on Information Systems, 9*(3), 224-248.

Gezehagn, G. (2012). Afaan Oromo Text Retrieval System. *Msc Thesis, ADDIS ababa Universiity* .

Halterman, R. (2011). *Learning to Program with Python.*

Hevner and Chatterjee. (2015). Design Science Research in Information Systems. *Association for Information Systems*.

Hofmann, T. (1999). Probabilistic Latent Semantic Analysis. *Journal of UAI'99.*

Ibrahim, B. (2015). The Origin of Afaan Oromo: Mother Language. *Global Journal of Human Social Science: Linguistics and Education, 15*(12).

Juban and Falguni. (2007). Search Algorithms an Aid to Information Retreival in Digital Libraries. *5th International CALIBER ,* 401-414.

Judit Bar-Ilan and Tatyana Gutman. (2003). How do Search Engines Handle Non-English Queries? - A case study. *In Proceedings of the Alternate Papers Track of the 12th.* Hungary: Budapest.

Khaled and Mohamed. (2010). Collaborative Document Clustering. *Journal of Machine Intelligence Research Group, University of Waterloo,* 451-461.

Ken Peffers et al. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems, 24* (3), 45-78.

Kothari, C. (2004). *Research Methodology.* New Age International Publishers.

Kowalski, G. (1997). Information Retrieval Systems – Theory and Implementation. *Kluwer Academic publisher*.

Kula and Vama. (2007). Evaluation of Oromo-English Cross-Language Information Retreval. *In IJCAI 2007 Workshop on CLIA.*

Kula et al. (2007). Evaluation of Oromo-English Cross-Language Information Retrieval. *Journal of IJCAI*.

Latifur et al. (2013). Retrieval Effectiveness of an Ontology-Based Model for Information Selection. *National Library of Medicine*.

Lei Shi. (2005). Putting Pieces Together:Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing. *In Computational Linguistics and Intelligent Text Processing*, 100-111.

Linda. (2010). Performance of Two Statistical Indexing Methods, With and Without Compound-word Analysis. *University Press, Cambridge*.

Loulwah et al. (2007). Local Semantic Kernels for Text Document Clustering. *In Workshop on Text Mining, SIAM International Conference on Data Mining.*

Maitri et al. (2015). A Survey on Semantic Document Clustering. *Researchgate Conference Paper*. IEEE.

Marco and Kathleen. (2009). An Approach to Semantic Indexing and Information Retrieval. *Rev. Fac. Ing. Univ, 48*, 165-187.

Marco Suárez and Kathleen Salinas. (2009). An Approach to Semantic Sndexing and Information Retrieval. *Journal of Information Retrieval and Semantic Indexing*.

Marti. (2008). Discovery of WordNet Relations. *MIT Press*.

Michael Steinbach, George Karypis and Vipin Kumar. (2009). *A Comparison of Document Clustering Techniques.* University of Minnesota.

Morka, m. (2001). Text-to-Speech System for Afaan Oromo. *Master Thesis at School of Information Studies for Africa*.

Moukdad, H. (2003). Lost In Cyberspace: How Do Search Engines Handle Arabic Queries? *The 12th International World Wide Web Conference*.

Mulualem, W. (2013). Semantic Indexing and Document Clustering for Ahmaric Information Retrieval. *Msc Thesis, ADDIS ABABA UNIVERSITY*.

Nagma et al. (2016). A Survey of Document Clustering Using Semantic Approach. *International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, 2555-2562.

Neepa and Sunita. (2012). Document Clustering: A Detailed Review. *International Journal of Applied Information Systems (IJAIS)*, 2249-0868.

Neepa and Sunita. (2012). Semantic Based Document Clustering: A Detailed Review. *International Journal of Computer Applications, 52*(5), 42-52.

Oren Zamir et al . (1997). Fast and Intuitive Clustering of Web Documents. *Journal KDD*, 287-290.

Pankaj, J. (2008). Document Clustering. *PhD diss Indian Institute of Technology Kharagpur*.

Paralic and Kostial. (1999). Ontology-based Information Retrieval . *journal Web Technologies*.

Paul et al. (2009). Addressing Morphological Variation in Alphabetic Languages. *In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval.*

Potsdam. (2007). A Quantitative Evaluation of the Enhanced Topic-Based Vector Space Model. *Journal HPI(Hasso planttner Instut)*.

Recardo. (1999). Modern Information Retrieva. *Journal of ACM*.

Richard, H. (2003). *Encyclopedia of English*. Retrieved October 20, 2017, from An encyclopedia of English Grammar and Word Grammar: http://www.phon.ucl.ac.uk/home/dick/enc-gen.htm

Rijsbergen, C. .. (1979). *Information Retrieval.* American Elsevier Inc., New York.

Sajendra and Ram Kumar. (2012). Ontology Based Semantic Indexing Approach for Information Retrieval System. *International Journal of Computer Applications, 49*(12).

Salton, G. (1983, September). Introduction to Modern Information Retrieval. *McGraw-HillComputer Science Series*.

Sathiyakumari et al. (2011). A Survey on Various Approaches in Document Clustering. *International Journal of Computer Technology and Application (IJCTA), 2*(5), 1534-1539.

Singhal, A. (2008). *Modern Information Retrieval*. Retrieved October 23, 2017, from A Brief Overview, Google: http://ilps.science.uva.nl/ Teaching/0405/AR/part2/ir_ overview. pdf

Tesfaye, G. (2010). Afaan Oromo Search Engine. *Msc Thesis, School of Graduate Study Addis Ababa University*.

Tewdros. (2003). Amharic text retrieval: an Experiment Using Latent Semantic Indexing (LSI) With Singular Value Decomposing (SVD). *Msc Thesis, School of Graduate Studies of Addis Ababa University*.

Thomas, H. (1999). Probabilistic Latent Semantic Indexing. *Journal of SIGIR*.

Tilahun, G. (2008). Afaan Oromoo. *Journal of Oromoo Studies*.

Turtle and Croft. (1992). A Comparison of Text Retrieval Models. *Computer Journal, 35*(3), 279-290.

Wakshum, m. (2000). Development of a Stemming Algorithm for Afaan Oromo Tex*t, Master Thesis at School of Information Studies for Africa,*. Addis Ababa: Addis Ababa University.

Wegari, G. (2017, March). OroRoots: Rule-Based Root Generation System for Afaan Oromo. *International Journal of Scientific & Engineering Research, 8*(3).

Xiao, Y. (2010). Survey of Document Clustering Techniques Comparison of LDA and moVMF.

Xiaoyong. (2010). Cluster-Based Retrieval Using Language Models. *Center for Intelligent Information Retrieval*.

APPROVAL SHEET

HARAMAYA UNIVERSITY

POST GRADUATE PROGRAM DIRECTORATE

SEMANTIC DOCUMENT CLUSTERING BASED INDEXING FOR AFAAN OROMO LANGUAGE INFORMATION RETRIEVAL SYSTEM

Submitted by:

_____	_____	_____
Name of Student	Signature	Date

Approved by:

1.	_____	_____	_____
	Major Advisor	Signature	Date
2.	_____	_____	_____
	Co-Advisor	Signature	Date
3.	_____	_____	_____
	Chairman, DGC/SGC	Signature	Date
4.	_____	_____	_____
	Dean, SGC	Signature	Date