2009

Concept-Based Automatic Amharic Document Categorization

Meron Sahlemariam
Addis Ababa University, nahommer@gmail.com

Mulugeta Libsie
Addis Ababa University, mlibsie@cs.aau.edu.et

Daniel Yacob
The Ge'ez Frontier Foundation, dyacob@gmail.com

Concept-Based Automatic Amharic Document Categorization

Meron Sahlemariam
Addis Ababa University
nahommer@gmail.com

Daniel Yacob
The Ge'ez Frontier Foundation
dyacob@gmail.com

Mulugeta Libsie (PhD)
Addis Ababa University
mlibsie@cs.aau.edu.et

ABSTRACT

Along with the continuously growing volume of information resources, there is a growing interest toward better solutions for finding, filtering and organizing these resources. Automatic text categorization can play an important role in a wide variety of flexible, dynamic, and personalized information management tasks. The aim of this research work is to make use of concepts as a way of improving the categorization process for Amharic[1] documents. In recent years, ontology-based document categorization method is introduced to solve the problem of document classification. Previous works on keyword-based document categorization miss some important issues of considering semantic relationships between words. In order to resolve the existing problems, this research proposed a framework that automatically categorizes Amharic documents into predefined categories using concepts. The research shows that the use of concepts for an Amharic document categorizer results in 92.9% accuracy.

Keywords

Ontology, Keyword-based Document Categorization, Concept-based Document Categorization, Text Classification.

INTRODUCTION

Due to the growth and availability of information in mass volume, the need for a new approach to manage information is certain. For the last few decades, the development of the World Wide Web (WWW) provided an easy way for accessing information. The volume of information on the Web tripled between the years 2000 and 2003 (Wiley, 2006). This fact shows that there must be some technique to extract and organize information. One of the most successful paradigms to organize information is classifying documents into different meaningful categories, called Automatic text categorization.

Automatic document categorization can be done using keywords or concepts. Keyword-based text categorization uses keywords which are extracted from the text to identify the category of a given document. That means a document that is going to be categorized should contain a specific keyword that matches the represented document to be categorized into the predefined categories. The major problem with this approach is that it ignores semantic relationship between the document's content and the designated category. Using only keywords affects the classification result since authors may use different keywords.

The other approach is to use the concept that the document represents. A concept captures the "semantics" or meaning of terms. Terms are words that describe concepts or act as synonyms for concepts. For example, the terms Football and Soccer have the same meaning; that means the concept behind both terms is the same or both terms are about the same thing.

Hence, concept-based text categorization allows classification of documents based on meaning rather than keywords. This method extracts concepts from the document and uses those concepts to categorize a document (Smith, 2004). In order to use concepts, the concepts should be represented in a knowledgebase using ontologies. An ontology is a systematic formalization of concepts, definitions, relationships, and rules that captures the semantic content of a domain in a machine-readable format (Gomez-Perez, Fernandez and Corcho, 2004).

[1] Amharic uses the Ethiopic script and is the official language of the Federal Government of Ethiopia.

In this paper, we proposed a model for concept-based automatic Amharic document categorization. In order to implement the proposed model, we developed the "News" ontology that contains the appropriate knowledgebase with several concepts and comprises of categories and sub-categories of News items. In addition to the knowledgebase, we developed an Amharic document categorizer. During the classification process, all the documents pass through pre-processing stages. Then index terms are extracted from a given document which are mapped onto their corresponding concepts in the ontology. Finally, the selected document is classified into a predefined category.

RELATED WORK

Many researches on text categorization have been done for different languages such as English (Gongde, Hui, Bell, Bi, and Kieran, 2007), Chinese and Japanese (Shih-Hung, Tzong-Han, and Wen-Lian, 2003; Corcho, Fernandez-Lopez, and Gomez-Perez, 2002) and Arabic (El-Kourdi, Bensaid, and Rachidi, 2004). These researches categorize a given document into predefined classes using keywords. In recent years, Amharic language related researches have led to an increasing awareness of Amharic language resources processing and digital information access. To this end, some works have been carried out on Automatic Text Categorizations for the Amharic language (Sintayehu, 2001, Afework, 2008; Teklu, 2003). Sintayehu (2001) attempted to develop an Amharic News Classifier (ANC) that has the capability of classifying Amharic news items into predefined classes automatically using statistical techniques. Teklu (2003) investigated the application of machine learning techniques Naïve Bayes (NB) and K-Nearest Neighbors (KNN) to automatic document categorization of Amharic news items. Afework (2008) has also carried out research on developing a toolkit for preprocessing Amharic text for document categorization.

All researches were carried out based on keywords. Therefore, there are still a number of issues that are not addressed. The foremost issue is that all of the previous studies depend only on keywords to categorize documents to a certain category. The major problem with this approach is that it ignores semantic relationship between the document's content and the designated category. A document consists of one or more ideas. It is this central idea of the document that makes it interesting to the user. Organizing the document collection using the central idea or the concept of each document will make the process of classification accurate. As opposed to keyword-based technique, this approach guarantees robust classifier as it is not influenced by word variations.

BACKGROUND

Text Categorization

Text categorization is the process of automatically classifying a set of documents into predefined categories. There are two main steps to categorize documents automatically: pre-processing and the actual classification. The pre-processing incorporates the following activities: lexical analysis, normalization, stop-word removal, stemming, and index term selection. In the classification phase, categorization can be done using keywords or concepts. There are a number of standard machine learning techniques which have been applied on keyword-based text categorization. For concept-based text classification, ontology is used as a way of representing knowledge in a specific domain.

Knowledge Representation

Knowledge representation involves abstraction and interpretation of real world knowledge using formal theories and reasoning procedures. Therefore, knowledge has to be conceptualized and represented in machine understandable mode. It means that knowledge is formalized and conceptualized in a symbolic form which can be interpreted.

Conceptualization is the process of representing knowledge in a machine understandable way. It consists set of objects, concepts, and other entities about which knowledge is being expressed and of relationships that hold among them. The process of conceptualization can be achieved using ontologies, which allow the representation of knowledge in machine understandable form (Gomez-Perez et al., 2004). Typically, the knowledge has been acquired to categorize a given document into predefined categories; it is then required to represent the knowledge in an ontology language.

Ontology

According to Gomez-Perez et al. (Gomez-Perez et al., 2004) a commonly agreed definition of ontology is: "An ontology is an explicit and formal specification of a conceptualization of a domain of interest''. This definition stresses two key points: the conceptualization is formal and permits reasoning by computer, and a practical ontology is designed for some particular domain of interest.

Formally, an ontology consists of terms, their definitions, and axioms relating them. As a formal description, ontologies consist of concepts known as classes, relations or properties, and instances. Formalized ontologies are instruments for capturing the meanings of concepts so that they may be used for improved and automated management of information. Ontologies may cover very general concepts or represent specific and restricted domains. The selection of concepts and their level of detail depend on the characteristics of the domains to be covered and the operations needed.

The main advantage of ontology is knowledge representation and reusability that allows reusing and sharing application domain knowledge using common vocabulary. Ontologies are used to organize knowledge in a structured way and they are the preferred ways of knowledge representation in semantic technology. Ontologies are useful to avoid building applications right from scratch and provides common mode of communication among the agents. Moreover, they facilitate representation of machine accessible information formally and explicitly with reasoning capability (Gomez-Perez et al., 2004).

This paper is concerned about the ontology used in intelligent computer applications. Throughout this paper the term ontology refers to a systematic formalization of concepts, definitions, relationships and rules that capture the semantics content of a domain in a machine readable format.

Basic Components of an Ontology

To formalize knowledge, ontologies use basic modeling component types. The two most important kinds of components in an ontology are the classes in which individuals can be categorized, and the relations that are used to create links between classes (Gomez-Perez et al., 2004).

Classes/Concepts can be anything something is said about and also it can be the description of a task, function, action, strategy, reasoning process, and the like. It is used to capture knowledge about a kind of thing and represent characteristics that may apply to many individuals. Classes are like generic nouns that are applied to distinct and named individuals such as a machine, human, dog or company.

Predicates are the most important type of relations in an ontology. Predicate terms explicitly represent relations that may link two or more items. For example, two people are related by the predicate father:

George H. W. Bush is the father of George W. Bush.

Predicates may connect classes, an individual and a class, or multiple individuals:

Republicans are a type of politician.

George W. Bush is a Republican.

George W. Bush leads the United States.

Individuals, also called instances, are those described using the concepts of an ontology. Typically, individuals are described as being members of some class. An individual member of a class has the general character of the class and it may have other characteristics and relationships as well (Gomez-Perez et al., 2004). For example Haile G/ Selassie, the famous Ethiopian runner, is an instance of the class of Athlete. Individuals may also have name strings ("Haile G/Selassie"), attributes ("Male", "1.65m"), and relations to other individuals "Haile G/Selassie is the husband of Alem Haile".

Attributes in general denote simple qualities that are secondary characteristics of objects (e.g. red, solid, short), in contrast to the essential properties represented by classes. In many cases, qualities represented by attributes include physical states of matter, colors, size, and the like (Gomez-Perez et al., 2004).

Functions are special cases of relations in which the n^{th} element of the relationship is unique for the n-1 preceding elements. Examples of functions are Mother-of and Price-of-a-used-car that calculates the price of a second-hand car depending on the car-model, manufacturing date and number of kilometers (Gomez-Perez et al., 2004).

CONCEPT-BASED AUTOMATIC AMHARIC DOCUMENT CLASSIFIER

As mentioned earlier, the aim of this paper is to make use of concepts as a way of improving the categorization process for Amharic documents. The paper also presents the structural design of the system in order to attain the objective. Figure 1 shows the general architecture of the concept-based automatic Amharic text categorization system. It is structured into three modules based on the data and process flow between the components. The pre-processing module is responsible to the target document processing. Domain knowledge is represented in the knowledgebase module which includes the reasoning process.

The classification module selects a specific category out of the list of concepts that represent categories. The input of the system is a document and the output will be a concept which is also the predicted category of the target document.

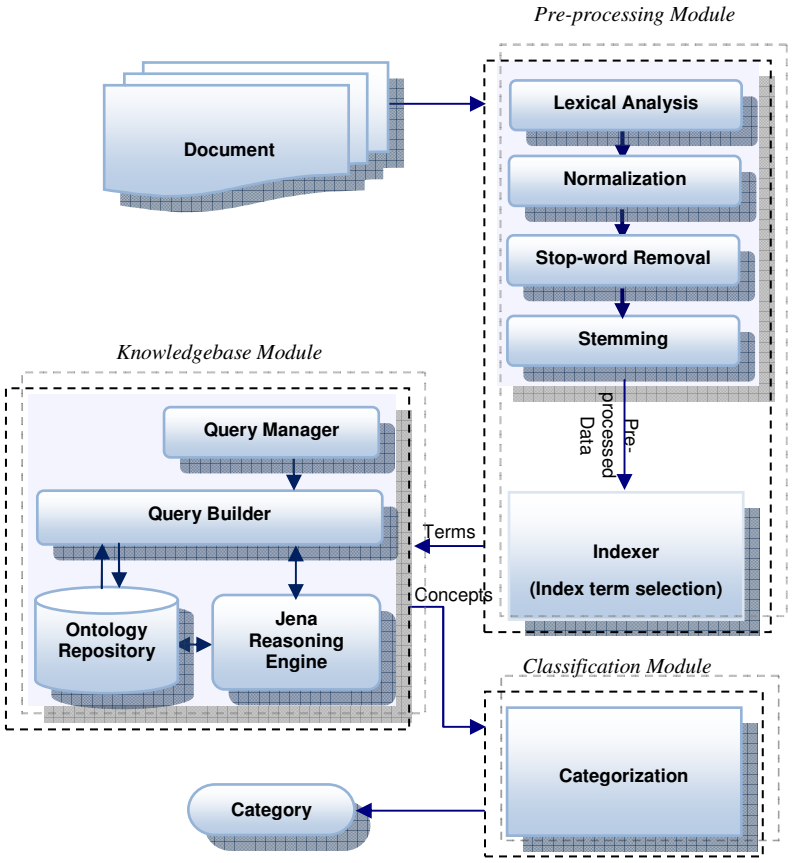

Figure 1. The general architecture of concept-based automatic Amharic text categorizer

Pre-processing Module

The responsibility of this module is to accept the input document and produce a set of index terms after carrying out lexical analysis, normalization, stop-word removal, stemming, and index-term selection. Index terms or terms that have the capability to represent the document are selected using Lucene. Lucene is a mature, free, open source, high performance, and scalable information retrieval library.

The Knowledgebase Module

This module serves as a knowledgebase to categorize a given document into predefined categories. To represent knowledge in the ontology, concepts are the main components. Each concept is formalized in the ontology using classes, sub-classes and relationships between classes. Words are used to represent concepts and a single concept can have list of words/terms i.e., a concept can be represented using multiple words. It is possible to say that the ontology is a database of all concepts, words, and categories and how they relate to each other.

Therefore, the ontology contains the represented knowledge in a specific domain. In order to represent the required knowledge in the ontology classes, instances and relationship between classes are organized into hierarchies. In the News ontology around 375 classes, 1811 instances and 119 properties are used.

Having that, there must be some way of accessing and using the knowledgebase in the ontology. To do so, it is necessary to have a means towards mapping terms on top of the ontology concepts and relations between concepts. The content of the document is better represented by these related mapped concepts. Using these related concepts, it is possible to capture the semantic relations found among the words in the text.

The result of this module is a concept; when the pre-processing module requests the knowledgebase to get the concept this module returns the corresponding concept to the classification module. The ontology maps the terms with the corresponding concepts and returns a specific concept that a term represents.

Reasoning

So far, pre-processing with the aim of extracting document representative index terms and ontology formulation to represent the knowledge is presented. However, having the represented knowledge and pre-processed data is not sufficient to categorize the document successfully. Besides the knowledge representation and pre-processing, reasoning is desirable to make use of concepts in the classification process.

The idea behind the reasoning process is to apply an inference engine to enable the classifier with a reasoning capacity. There are various inference engines that derive additional information in an abstract processing way. The most known and commonly used are Pellet, SWRL and Jena Rules, Jean Reasoner, and SwiftOWLIM. During this research Jena semantic framework is used to implement the reasoning capability. Jena is an open source toolkit for processing RDF, OWL and other semantic web data and it is composed of RDF Processing API, OWL Processing API, a rule-based reasoning engine and SPARQL query engine.

Classification Module

The list of concepts from the knowledgebase module serve as input for the classification module and identification of where the document belongs to; that is, the target category is selected on a specific concept from list of concepts depending on the weight of the concept. An input document for this process is accepted from the user which is going to be categorized.

Once the index terms are identified for the selected document, the next step is to inquire the knowledgebase to get concepts. The index terms that are extracted from a given text will be mapped onto their belonging concepts in the ontology. In the knowledgebase, concepts are extracted based on terms using the ontologies. The pre-processing module queries the ontology by passing index terms and then the knowledgebase returns the concepts to the classification module where the term belongs to.

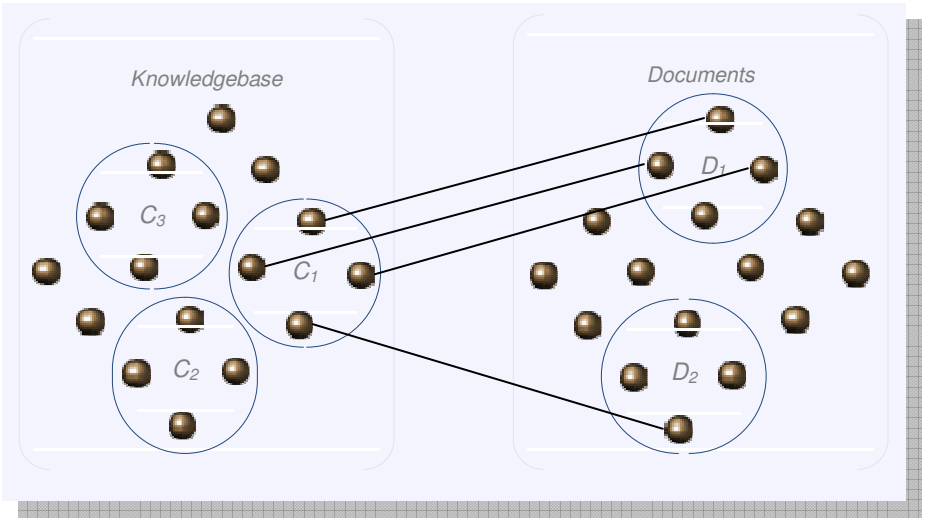

Figure 2. Mapping between Documents and Concepts

As depicted in Figure 2, each index term is mapped to the corresponding concepts in the knowledgebase. The knowledgebase contains *n* number of concepts including the concept terms. For example, document D_1 contains n number of index terms D_1 (DT_1, DT_2 DT_3…DT_n), and the concept C_1 is represented in n number of concept terms C_1 (CT_1, CT_2, CT_3…CT_n) in the knowledgebase. So, the mapping is done from the document term DT_i to the concept term CT_j.

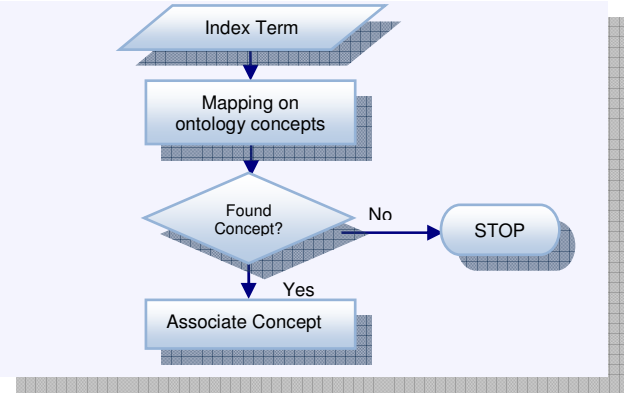

Figure 3. The flowchart of the mapping between terms and the News ontology concepts

As shown in Figure 3, the system checks whether or not the index term exists in the knowledgebase. If the concept is found, it associates the document term with the corresponding concept. However, there is a possibility that the index term may not be able to be mapped onto its corresponding concept if there is no such concept available in the News ontology. This situation requires an alternative way to map the index term onto the external knowledge.

In order to discriminate between the important and the less important concepts, a weight is assigned to each concept. The determinant that influences a weight given to a concept is the occurrence of a concept based on the number of term frequency. The term frequency indicates how frequent a particular concept is mentioned in the document. The higher the frequency, the more important the concept is considered to be.

Hence, to weight concepts, this paper considers concepts, generic concepts and term frequency. There are sets of generic concepts Gc, such as Accidents, Economy, Science and Technology, Environmental Preservation and Weather Condition, and Sport.

$Gc = \{Gc_1, Gc_2, Gc_3 \ldots Gc_n\}$

$C = \{C_1, C_2, C_3 \ldots C_n\}$

Where Gc is the set of generic concepts and C is a set of concepts

To determine weight for a concept C_i, the classifier first checks whether or not the concept is a generic concept. If the knowledgebase returns a generic concept, then the classifier ignores the concept; otherwise it weighs the concept. If term T_j belongs to the concept C_i then the weight of concept C_i will be the sum of term frequency of the j^{th} term of a document d, $f(T_j)$.

If ($C_i = G_{ch}$) then

Do nothing

Else if (T_j belongs (C_i))

$$WC_i = \sum_{i=1}^{n} f(T_j)$$

Based on the above method, the selected document concepts should be weighted to determine the category of the document. In order to select a specific concept, this research used maximum weight of the concepts ± some constant *k*, where *k* is decided through the experiment. The reason for inclusion of ± some constant *k* is to categorize the document into multiple categories. In the case of single-label classification only one class is assigned for the document, but in a multi-label classification each document can be assigned an arbitrary number of multiple labels of multiple possible classes. This is because a document may contain multiple concepts.

$C_o = \text{Max } \{WC_1, WC_2, WC_3 \ldots WC_n\}$

$C_s = C_o \pm k$

Where C_o is the maximum of weighted concept WC_i and C_s is the selected concept or category of a given document, which is $C_o \pm k$. Based on the above concept of weight mechanism, multiple concepts can be selected with values equal to C_s.

EXPERIMENTAL RESULTS

The source data for Amharic News items was the Ethiopian News Agency (ENA). For the experimental purpose, 975 Amharic news documents are prepared. The data was manually classified into categories and sub-categories. The collection is categorized and sub-categorized under concepts of: Accidents, Economy, Science and Technology, Environmental Preservation and Weather Condition, and Sports.

Evaluation of the classifier is done with the evaluation parameter that compares the number of documents which are classified correctly and incorrectly. Typically, the comparison is done amid the document classified using the automatic classifier and that of the manually classified documents.

Precision and recall, which are the evaluation parameters of Information Retrieval, are used in text classification. Precision, P, is the ratio of the number of documents classified correctly to the total number of documents in a given category.

$$P = \frac{TC}{TC + FC},$$

Where, TC denotes the number of documents which are classified correctly and FC denotes the number of documents which are classified incorrectly.

Recall, R, is the ratio of TC and the whole documents belonging to the category,

$R = \frac{TC}{TC + MC}$, where MC denotes the number of documents which are missed by the classifier, that is, documents neither classified correctly nor incorrectly.

Description	No of Input document	TC	FC	MC	P	R	Accuracy %
Accidents	49	39	0	10	1	0.79	79.5
Economy	575	547	18	10	0.96	0.98	95.1
Environmental Preservation and Weather Condition	76	64	0	12	1.0	0.84	84.2
Science and Technology	88	74	4	10	0.94	0.88	84.0
Sport	187	182	3	2	0.98	0.98	97.3
Total	975	906	25	44	0.97	0.95	92.9

Table 1. Classified documents per category

As depicted in Table 1, it was found that the classifier gave a correct result on the average for 92.9% of the test documents. That is, from the 975 documents 906 of them are correctly classified; the remaining 69 are classified as incorrect and missed by the classifier.

The result primarily depends on the knowledge represented in the ontology. From the wrongly classified documents, most of the concepts contained in the documents are similar with each other. However, index terms which are extracted from the documents are included in some other concepts. As a result, the reasoner tries to find related concepts and maps those index terms onto the related concepts but not the exact concept. For example, document "Econ332" is wrongly classified, because the document contains index terms such as "Farmer". In the knowledgebase, the concept "Agriculture" contains a concept term "Farmer". Hence, the reasoner maps it into the concept "Agriculture", which is the wrong category. Due to these reasons, the categorizer incorrectly classified the above documents.

However, it can be seen that larger number of News items are classified correctly. This shows that category concepts for correctly classified documents are plainly represented in the knowledgebase that distinguishes it from other categories. Therefore, it is apparent that the classification process is primarily governed by the represented knowledge in the ontology.

CONCLUSION

The techniques of automatic classification, using concepts for Amharic document categorization correctly classified 92.9% of the test collection documents. However, the technique needs to be supported with extended knowledge. The classification process is primarily governed by the represented knowledge in the ontology. In general, having a complete and clearly stated knowledge for each category and sub-category decides the final result of the categorization process. This shows that as the knowledgebase gets richer, the performance of the system will be enhanced considerably.

RECOMMENDATIONS

The results found in this research showed that classification can be done automatically for Amharic documents using concepts. However, it is also learnt that further research and developmental effort is needed so as to enable the complete exploitation of this technology. The ontology development and deployment environment is rich with ideas that could further improve the process of knowledge based systems construction. In this section, a number of such ideas are listed. Those ideas deal with future research issues and some features that are not of a research nature but that are needed to provide a better result.

Extending into multi lingual documents: the main characteristics of the ontology are being sharable and reusable. To enable knowledge sharing and reuse, it is necessary to represent concepts and relations in multiple languages. Instead of automatically categorizing only Amharic documents, it is possible to incorporate other languages such as English and to make it multi lingual.

External Knowledge: In the process of extracting concepts from the knowledgebase, index terms are mapped on the corresponding concepts of the ontology. However, there is a possibility that the term may not exist because of the limited number of concepts available in the News ontology. This situation requires an alternative way of mapping onto the external knowledgebase concept. The alternative way is to use the extended concept in order to map between the external concept and the existing knowledgebase. Having such external knowledge makes the existing knowledgebase more powerful to incorporate possible concepts.

Amharic lexical database: having Amharic lexical database that offers information related to various semantic relationships among words is essential. It has a potential to be used in order to represent concepts rather than words and the relationships between concepts including the corresponding synonyms and antonyms. Having the feature representations with this database information is believed to result in a significant improvement of the classification process.

REFERENCES

1. Afework, Y. (2008) Preprocessing Toolkit for Amharic Text Categorization, *the case of Ethiopian News Agency*, School of Computer Science, Addis Ababa University, Addis Ababa.
2. Corcho, O., Fernandez-Lopez, M., and Gomez-Perez, A. (2002) Methodologies tools and languages for building Ontologies, Laboratorio de Inteligencia Artificial Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo sn. Boadilla del Monte, 28660. Madrid, Spain.
3. El-Kourdi, M., Bensaid, A., and Rachidi T. (2004), Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm, *20th International Conference on Computational Linguistics,* Geneva August 28th.
4. Getahun, F. and Atnafu, S. (2007) The Use of Semantic-based Predicates Implication to Improve Horizontal Multimedia Database Fragmentation, *International Multimedia Database Conference*, Augsburg, Bavaria, Germany.
5. Gomez-Perez, A., Fernandez, M., and Corcho, O. (2004) Ontological Engineering, *with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*, 403 p. 159.
6. Gongde, G., Hui, W., Bell, D., Bi, Y., and Kieran, G. (2007) An KNN Model-based Approach and Its Application in Text Categorization, *Proceedings of the 12th WSEAS International Conference on Applied Mathematics*, Cairo Egypt, pp. 228-233.
7. Pretorius, A. J. (2004) Ontologies - Introduction and Overview, *Semantic Technology and Applications Research Laboratory*, Vrije Universiteit Brussel, Pleinlaan 2, B-1050 Brussels, Belgium.
8. Shih-Hung, W., Tzong-Han, T., and Wen-Lian, H. (2003) Text Categorization Using Automatically Acquired Domain Ontology, Institute of Information Science Academia, Sinica Nankang, Taipei, Taiwan.

9. Sintayehu, Z. (2001) Automatic classification Amharic news items, *the case of Ethiopian News Agency,* School of Information Studies for Africa, Addis Ababa University, Addis Ababa.
10. Smith, B. (2004) Beyond Concepts: Ontology as Reality Representation, Department of Philosophy, *Proceedings of FOIS International Conference on Formal Ontology and Information Systems*, Turin, 4-6 November, University at Buffalo, USA.
11. Teklu, S. (2003) Automatic categorization of Amharic news document, *a machine learning Approach*, School of Information Studies for Africa, Addis Ababa University, Addis Ababa.
12. Wiley, J. (2006) Semantic Web Technologies*, the Atrium*, Southern Gate, Chichester, West Sussex, England.