# Ontology based Categorization of News Article Documents

Phyo Wai Thaw
*University of Computer Studies, Yangon*
*phyoweithaw@gmail.com*

## Abstract

*Nowadays, the number of electronically available information and knowledge from the internet is rapidly growing. Therefore it becomes difficult to find relevant information users need. The most successful example to organize this mass of information is classifying different text documents according to their topics. Text categorization is a general tool for information retrieval, knowledge management and knowledge discovery. In this paper, ontology based news articles documents categorization is introduced. In the process of categorization the incoming news document is categorized into possible concept(s) from a predefined set of concepts. The ontology provides a hierarchical structure where each concept can be treated as a category. Then news can be retrieved by keyword and concept later. Top down method according to the ontology structure is used for classification.*

## 1. Introduction

Electronic online documents started appearing at about the same time that the Internet became public. With increasing amount of documents, the need for an effective mechanism to organize not only information, but also knowledge becomes critically important. Text categorization is the task of assigning texts to one or more pre-defined categories based on their content. The applications range from automatic document indexing for information retrieval systems, document organization, text filtering, word sense disambiguation categorization of web pages and most recently, spam filters. [3].

Text classification organizes information by associating a document with the best possible concept(s) from a predefined set of concepts. One of the examples is the categorization of incoming news into predefined set of concept categories such as entertainment, politics, business, sports, etc. The traditional categorization methods are based on machine learning and probabilistic approaches. These including comparison between vector representations of the documents (Support Vector Machines, k-Nearest Neighbor, Linear Least-Squares Fit), use of the joint probabilities of words being in the same document (Naïve Bayesian), decision trees, and neural networks. Some of these approaches include implementing unsupervised learning algorithms like Latent Semantic Analysis and using AI rule-base trees to compute the conceptual relevancy of search results [1].

As described by the World Wide Web Consortium (W3C), ontology is used to describe and represent an area of knowledge. Ontologies are used by people, databases, and applications that need to share domain information. Ontology is a data model that represents a set of concepts within a given domain and the relationships between those concepts. It is used to reason about the concepts within that domain.

In this paper, a system that is able to categorize news into concept ontology and extract information using concept and category is proposed. Concepts are able to update into the ontology.

The rest of this paper is structured as follows: Section 2 presents the related works. In section 3 the architecture of the system is present. In Section 4 conclusion is given.

## 2. Background Theory

### 2.1 Ontology

Ontology is an explicit specification of a conceptualization. Ontology has been used in data base and information retrieval to support distributed and heterogeneous data sources interoperability. In the domain of Artificial Intelligence (AI), ontology is an engineering artifact. It contains a glossary set, which describes some conditions and an axiom set. So the ontology is always a lexicon and a description of terms [4]. A typical ontology composes of five parts: classes, relations, functions, objects and axioms.

Ontology construction is an approach to utilize computer for the structure representation of domain knowledge. Ontology-based computer systems do not interact directly with the real world but rather with internal models of the relationships between concepts and objects in the real world. Such models represent problem domains, and the development of such models in computers is referred to as ontology building.

Concepts represented by an ontology can usually be clearly depicted by a natural language because the ontology and the natural language function similarly (i.e., describing the world). Ontologies can be built by specifying the semantic relationships between the terms in a lexicon. One example of such ontology is Senses [Knight 94], a taxonomy featuring over 70,000 nodes. The OntoSeek system [Guarino 99] uses this ontology for information retrieval from product catalogs. [1].

The advantage of ontology is that it can be read, interpreted and edited by human. Another advantage of allowing human editing is that the ontology produced can be shared by various applications such as from a QA system to a knowledge management system. Ontologies can be very helpful with the reuse of domain knowledge, and for the separation of domain knowledge and software code that performs operations on that knowledge.

Developing ontology includes

- Defining classes in the ontology,
- Arranging the class in a taxonomic (subclass-super class)hierarchy,
- Defining slots and describing allowed values for these slots,
- Filling in the values for slots instances.

We can then create a knowledge base by defining individual instance of this classes filling in specific slot value information and additional slots restriction.

## 2.2 Stemming or Lemmatization

In all languages, suffixes may be added to words to indicate a particular form of that word: the past of a regular verb (”-ed” suffix), the adding of ”-s” or ”es” suffix for the plural form, the comparison (adding ”-er” or ”-est” suffix) and the name or the adverb derived from a verb (”-er” and ”-ely” suffix). Usually the linguists call *stem* or *lemma* the base form of the word and the process reporting a word to its base form is called *stemming* or *lemming*.

The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. Moreover the occurrences count of a word should be assigned to its base form, increasing the score for it. A document dealing with cars and similar objects could contain many occurrences of the verb ”drive”, the past form ”driven” or the *-ing* form ”driving”; if no reduction is performed all of them are an entry of the vocabulary while they should considered all as occurrence of ”drive”.

Many strategies for suffix stripping have been reported in the literature all working on the English Language and they vary depending on whether a stem dictionary is used.

### 2.2.1 Porter stemmer

The most used stemming tool is the Porter stemmer. It is a rule-based algorithm. The algorithm is a process for removing the common morphological and in flexional ending words for in English. The Algorithm consists of five phases for word reductions and applied sequentially. For Example, the word “Caresses” will change to “caress” according to the rules. Some rules and examples are shown in table1.

| Rule | Example |
|---|---|
| SSES->ss | Caresses->caress |
| IES->i | Ponies->poni |
| SS->ss | Caress->caress |
| S->s | Cats->cat |

**Table 1: Some rules of porter stemmer**

## 3. System Architecture

In this section, the general architecture of our system, which consists of two portions search and concept creation, is introduced. In first portion, anyone can search news without authority levels but the second one required authority permission.

In first portion, user can search news by concept or keyword. The system searches these in ontology. Then the system will display information about news.

In second one, the user must be the administrator who understands news ontology concept. This portion contains concept creation and adding of news. Concept creation creates only concept for class levels. For adding news, users can type necessary information of news. And then the system will save this information in the ontology concept and keyword database.

## 3.1 Concept Creation

Concept creation is a necessary step for developing and maintaining ontology. There are three strategies for the creation of ontology: top down, bottom-up and a combination of two. Ontology creation is an iterative process of modeling the given domain, by choosing the most important concept and identifying the most important concepts and identifying the most relevant relations between them. The system used top-down development process for the creation of ontology.

To develop ontology, OWL (Web Ontology Language) is used. OWL is a W3C recommendation language. OWL is a language for defining and

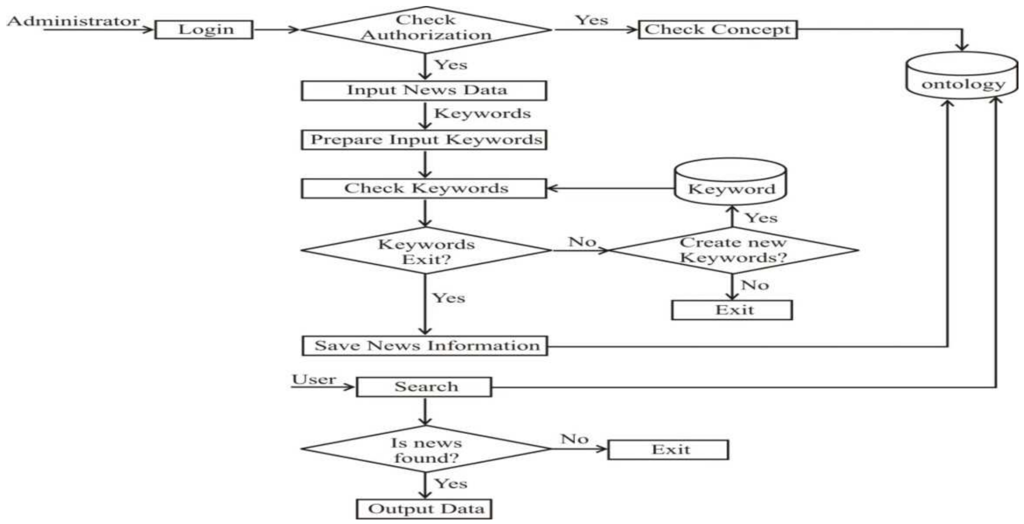

**Figure 1: Flow of System**

maintaining ontology. OWL may include descriptions of classes, properties and instances. OWL provides standard syntax for writing ontologies and modeling perspective.

News concepts ontology includes class with a maximum three-depth, as shown in fig2. For example, some of the first level concepts of news ontology are Economic Business and Finance, Health, Lifestyle and Leisure. Some of the second level concepts of Economic Business and Finance are Agriculture, Computing and Information Technology. Some of the third level concepts of Agriculture are Livestock farming and Fishing Industry.

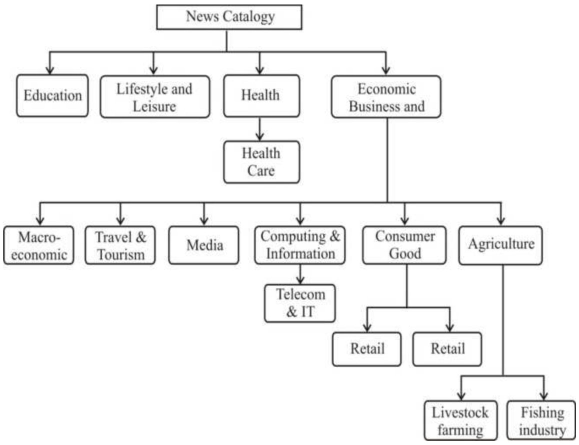

**Figure 2: Part of news concept hierarchy**

## 3.2 News Adding

After adding necessary information for news with keyword, the keyword is matched with keyword in keyword database. If the keyword is found, the system searches the keyword and its category in ontology using it. If keyword and its category exist in ontology concept, the system will input data according to this keyword and the category.

### 3.2.1 Keyword Categorization

Keyword categorization is pre-mapping for ontology based system. The input keywords been stemmed using Porter Stemming Algorithm. Then keywords will be mapped with keyword in keyword database that contains keyword with their corresponding concept. There is a possibility that the keyword may not be able to be mapped onto its corresponding concept because there is no such keyword in database for the concept. In this situation new keyword can be inserted with its corresponding concept.

## 3.3 Searching News

Using keyword and concept, the user is able to search the news for items of interest. When the user uses keyword for searching, first the input keyword is stemmed by porter stemming algorithm. And then, the system matches keyword in keyword database. If keyword is found, the system searches in the ontology using query language which allows selecting from ontology instances. If the keyword is found in the ontology, the system will show information about news. The input keyword may concern with more than one concept. In this case, the system finds all news that contains the input keyword. If user use concept, the system directly find in Ontology.

## 4. Conclusion

This paper describes about the system that can be used for creating and maintaining a new concept into the ontology and to insert documents into the ontology. The ontology classes are used as classification categories. The system is very useful in concept creation with ease and fastness. In addition, it can search news exactly. Users can exactly access the necessary information.

## References

[1] B. B. Wang, R I. (Bob) McKay, H. A. Abbass, M. Barlow, Learning Text Classifier using the Domain Concept, School of Computer Science, University College, ADFA, University of New South Wales, Canberra, ACT 2600.

[2] D. Raviandan, and S. Gauch, Exploting Hierarchical Relationships in Conceptual Search, EECS Development, University of Kansas Lawrence, KS66045.

[3] H. Bacan, I. S. Pandzic, and D. Gulija, Automated News Item Categorization, Faculty of electrical engineering and computing, University of Zagreb, Unska 3, HR-10000 Zagreb, Croatia, Croatian News Agency HINA Marulicev trg 16, HR-10 000 Zagreb, Croatia.

[4] H. Gu1, Kuanjiu Zhou, “Text Classification Based on Domain Ontology”, Journal of Communication and Computer, ISSN1548-7709, Volume 3, No.5 (Serial No.18), USA, May 2006, pp. 29-32.

[5] K. Schouten, P. Ruijgrok, J. Borsje, A Semantic Web-Based Approach for Personalizing News Erasmus University Rotterdam, PO Box 1738, NL-3000, Rotterdam, the Netherlands.

[6] L. Tenenboim, B. Shapira, P. Shoval, “Ontology-Based Classification of News in an Electronic Newspaper”, International Book Series "Information Science and Computing", pp. 89-97.

[7] M. H. Seddiqui, and M. Aono, Use of Ontology in Text Classification , Toyohashi University of Technology, Aichi, Japan.

[8] Noy, Natalya P. and Mcguinness, Deborah L., A Guide to Creating Your First Ontology, Stanford University, Stanford, CA, 94309.

[9] V. Sanfonov, A. Kovikov, A. Smolyakov,”knowledge .Net Ontology-based knowledge management toolkit for Microsoft.Net”, .Net Technology 2006, Plzen, Czech Republic, ISBN 80-86943-12-7.