ISSN 2229-5518

Ontology Based Text Categorization - Telugu Documents

Mrs.A.Kanaka Durga, Dr.A.Govardhan

Abstract— In this paper, we introduce a new method of ontology based text classification for Telugu documents and retrieval system. Many of the text categorization techniques are based on word and/or phrase analysis of the text. Term frequency analysis signifies the importance of a term within a document. Two terms within a document can have the same frequency, but one term may contribute more to the meaning of the sentence compared to the other term. Our aim is to capture the semantics of a text. The model we worked enables to capture the terms that presents the concepts in the text and thus identifies the topic of the document. We have introduced the new concept based model which analyzes the terms on the sentences and documents level. This concept-based model effectively discriminates between non-important terms with respect to sentence semantics and terms which hold the concepts that represent the sentence meaning. The limitations of key-word based search are overcome by usage of Ontology which is a motivation of semantic IR. The retrieval model is based on an adaptation of the classic vector-space model. The concept of ontology is associated with the related words and their weights from the pre-classified documents as a learning stage. In the main process, the words and their mutual relations are extracted from the target documents. The concept of Ontology is used to map the target document. A detailed description of the test results is illustrated in the paper and we explained thoroughly how the concept based classification is far more superior when compared to the word based classification for telugu documents.

Index Terms— Concept-based model, IR, Ontology, Retrieval model, Term frequency, Text categorization and Telugu documents,

---

1. INTRODUCTION

In the current paper we have focussed our efforts on electronic documents of Telugu Language.

The Telugu Language: Telugu language is the second most spoken languages after Hindi in India. Telugu belongs to the South Central Dravidian subgroup of the Dravidian family of languages. It has been recently awarded the Classical status. Telugu has been the language of choice for lyrical compositions for its vowel endings words, rightly called the "Italian of the East". Words in Dravidian languages, especially in Telugu are long and complex. Telugu, like other Dravidian languages is highly rich in morphology and hence agglutinative in nature. Telugu has 16 vowels and 40 consonants.

Text Categorization: (TC) is the classification of documents with respect to a set of one or more pre-existing categories. TC is a hard and very useful operation frequently applied to assign subject categories to documents, to route and filter texts, or as a part of natural language processing systems. In the past, several methods proposed for text categorization were typically based on the classical Bag-of-Words model where each term or term stem is an independent feature. The disadvantages of this classical representation are:

a) The ignorance of any relation between words, as a result of which learning algorithms are restricted to detect patterns in the used terminology only, while conceptual patterns remain ignored.

b) The big dimensionality of the representation space. In this article, we propose a new method for text categorization, which is based on the use of the Word Net ontology to capture the relations between the words.

In this approach terms are merged with their associated concepts extracted from the used ontology to form a hybrid model for text representation. We have undertaken a series of experiments on Telugu documents which highlight the positive contribution of this approach.

Ontology: Ontology is not necessarily norms on the Construction or definition or expression. A conceptual description of ontology including concept, attribute, entity, association description and the main purpose for knowledge sharing and reuse is given by Jade Goldstein [2]. Ontology is the concept (concepts, classes) of abstract sets and attributes (properties, attributes) is for the characteristics of objects and entities (individuals, instances) is a real thing and association (relations) will attribute is used for the titles of the two concepts or entities.

Ontology is a formal, explicit specification of a shared conceptualization. Jade [2] defined that ontology is a conceptual description, including concept, attribute, entity and association description with the main purpose of knowledge sharing and reusing knowledge. In the context of knowledge sharing, we will use the term ontology to mean a specification of a conceptualization. That is, ontology is a description of the concepts and relationships that can exist for an agent or a community of agents. This definition is consistent with the usage of ontology as set-of-concept-definitions, but more general.

The Proposed work is an efficient way of extracting text from the Telugu Documents and performing Information Retrieval from that Telugu Document.

Related works in this area have been explained in Section 2. Our Proposed Work and its layout have been explained in Section 3. Results and Performance are dealt in Section 4. Section 5 states the Conclusions and further work to be done

LITERATURE SURVEY

Semantics has been introduced at various linguistic levels, word level, sentence level and document content extraction level and at various stages of Information Retrieval such as query and document

representation, and in indexing. Any attempt to bring in semantics needs to balance the amount of complex natural language processing required, with the increase in retrieval performance. It is important to note that the pre-processing done for document representation is an offline one-time process which would every time provide improvement in the retrieval performance. The main modules of IR are pre-processing, indexing and retrieval. A set of documents is given as input to the pre-processing phase where the stop words and punctuation are removed. The parts of speech of the content words are determined by the POS (Part of Speech) tagger after the stemmer stems the content words resulting in root words. Basically a document can be represented with a bag of words using Boolean model. The bag of words however does not provide ranking of the retrieved documents.

To overcome this limitation, keyword-based search has been put forward where precision and recalls are improved but this also giving some ambiguous results. The use of ontology is the motivations of the Semantic information retrieval. Semantic search engine is viewed as a tool that gets formal ontology-based queries (e.g., in RDQL, RQL, SPARQL, etc.) from a client, executes them against a knowledge base (KB), and returns ontology values that satisfy the query. These techniques typically use Boolean search models based on an ideal view of the information space as consisting of non-ambiguous, non-redundant, formal pieces of ontological knowledge.

Conceptual search, i.e., search based on meaning rather than just character strings, has been the motivation of a large body of research in the IR field long before the Semantic information retrieval emerged. This drive can be found in popular and widely explored areas such as Latent Semantic Indexing , linguistic conceptualization approaches or the use of thesaurus and taxonomies to improve retrieval.

Those proposals are commonly based on shallow and sparse conceptualizations, usually considering very few different types of relations between concepts and low information specificity levels. The model we proposed considers a much more detailed and densely populated conceptual space in the form of an ontology-based KB. Though it is difficult to obtain such a rich conceptual space, this is one of the major targets addressed by the Semantic Web research community.

Our approach combines the flexibility and generality of an IR model for unstructured search spaces. The expressiveness and detail of a structured relational model describes some of the knowledge involved in the unstructured information space, in a structured and formal way, with powerful and precise data querying facilities. Ontology-based approach can be relied since it enables further inferencing capabilities that can be exploited to enhance the retrieval process. By building upon an ontology-based layer, our model benefits from semantic data integration facilities.

Mayfield and Fin in combine ontology-based techniques and text-based retrieval in sequence. We share with Mayfield et al. the idea that semantic search should be a complement of keyword-based search as long as not enough ontologies and metadata are available.

3. DETAILS OF THE WORK CARRIED OUT

To start with, each text document is tokenized so that it gives raise to the set of wordsThe efficiency levels are low when we apply any of the conventional classifier methodology. These low efficiency levels are attributed to the inflated form of words. In order to overcome this defect, we use morphological analyzer tool to get the root words. As a next step, domain specific key words are identified. Text classifier is applied on the key words selected from the telugu document and found that the classifier efficiency of key words are better w.r.t. to the words. We found that when we applied the ontological classification, there is an enormous amount of improvement in the classifier efficiency. Here we have used the Ontology_Dictionary (Wordnet - telugu) developed by Centre for Advanced Linguistics and Ttransliteration Studies (CALTS-UOH), university of Hyderabad(Central University) for feature grouping. All such words which are grouped based on the features are termed as word class/concept. With the help of ontology, terms that are found in and around the same concept are mapped into one dimension. This will help in excluding or disambiguating the terms that are present in many concepts due to the semantic ambiguity.

3.1 An Illustration Using Ontology Based Classification for Telugu Document

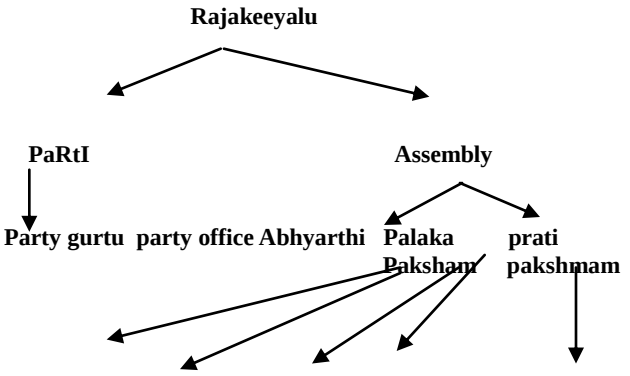
"AarDika mMtrito, **kAryadarsi**to muKhya mMtri assembly lo mMtaNalu" - Telugu ("Chief Minister discussed with the Finance Ministerand secretary in the assembly" – English)
Words: { AarDika , **mMtrito**, muKhya, **mMtri**, assembly lo, mMtaNalu}

Root word: mMtri

Key words: {mMtri, **assembly}**

Feature Grouping: {mMtri, muKhya**mMtri, kAryadarsi }**

Word Class/Concept: {mMtri, muKhya**mMtri, kAryadarsi }**



On Pre-processed document, root words are extracted through Ontology at concept level. Noun words are identified and their frequency computed and preserved in the data bank. On the nouns thus retrieved, feature-matrix clusters developed. We have calculated a representative feature vector for each concept node in an Ontology. We have then measured the similarity of the two of those class vectors by a simple cosine measure.

3.2 Equations

Algorithm:

1. Start
2. Morph Analysis (Finding base words)
3. Apply Ontology
4. Find the Sub-category
5. Recognize parent-node as a category of the respective document
6. If the parent has a child- repeat the process (iterate the process from 3-5)
7. Otherwise take parent as the final category

3.3.Vector Space Model:

We used vector space model to weigh terms and calculate feature vectors

Weight of a term is given as : $w_{ik} = tf_{ik} \times idf_k$

where

tf_{ik} is the number of occurrences of term t_k in document i and

idf_k is the inverse document frequency of the term t_k in the collection of documents.

A commonly used measure for the inverse document frequency is:

$idf_k = \log(N / n_k)$

where

N is the total number of documents in the collection, and

n_k is the number of document which contains a given term

.

Ontology based classification is carried in three steps: -

Step I : Ontology creation

Step II : Calculating relevance score

Step III : Text classification

Experiments were conducted on a small sample of 400 Telugu documents which were broadly categorised into two categories namely rajakeeyalu (Politics), Aatalu (Sports) .Out of these 400 documents 80% were used as training docments and the balance 20% were used as testing documents.

3.4. Hypothesis:

H1: On ontology based text categorization, more distinctive features are mapped towards right of the ontology scale.

H0: On ontology based text categorization, less distinctive features are mapped towards left of the ontology scale.

If H1 is true, accept the fact that the efficiency levels are better.

To measure the performance of these measures, we calculated recall rate and precision rate.

Recall rate= a/b and precision rate = a/c

where a = No. of documents which are classified into category correctly.

b= No. of documents of category in the testing data.

c=No. of documents which are classified into category.

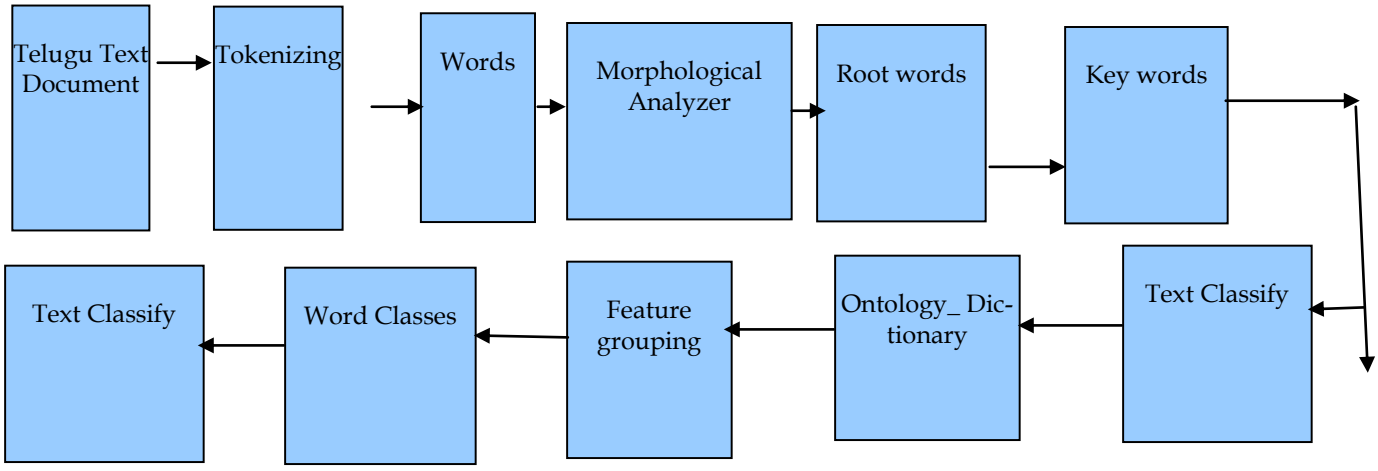
Table 1. Groups of Misclassification

Result type	No.of texts
Texts assigned to the subclass category	20
Texts assigned to the sperclass category	4
Texts assigned to the other category	40

.

.

3.5 FIGURES: FLOW PROCESS OF ONTOLOGY BASED TEXT CATEGORIZATION FOR TELUGU DOCUMENTS



4. CONCLUSION:

Literature on earlier research have proven that in conventional methods, misclassified items are not accessible. Further it is also easy to develop weakly thesauruses than conventional methods. In our paper, we have proposed a nd proven that the efficiency of text classification of the term is better when we used the Ontology model for Telugu documents when compared to the conventional methods.

5. ACKNOLEDGMENTS:

We sincerely thank Dr.G.Uma Maheswara Rao for providing Ontology_dictionary for Telugu and Morphoogical Analyzer tool.

6.REFERENCES

[1]Sebastiani F., "Machine Learning in Automated Text Categorization," ACM Computing Surveys, vol. 34, no. 1, pp. 1-47, 2002.

[2] Jade Goldstein, Mark Kantrowitz, Vibhu Mittal, and Jaime Carbonell (1999), Summarizing Text Documents: Sentence Selection and Evaluation Metrics, In ACM SIGIR 1999, pp.121-128, 1999.

[3] Dr.G.Uma Maheswara Rao, Morphological Analyser, at the centre for ALTS, University of Hyderabad.

[4] Dr.G.Uma Maheswara Rao and research team "Ontology_Dictionary-Telegu",at the centre for ALTS, University of Hyderabad.

[5] A. karthikeyan et al.,"An Novel Approach sing Semantic Information retrieval For Tamil documents", International Jornal of Engineering Science and Technology,vol.2(9),2010,4424-4433.

[6].S.MChaware et al.,"A survey:Issues of semantic Matching for Indian Languages Using Ontology",International Journal of Information echnology and knowledge Management,vol.2(2).pp.351-354,2010.