Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF
AGRICULTURAL STATISTICS 67(1) 2013 89-96

Extracting Concepts using Linguistic Ontology in Agriculture Domain

Aditi Sharan, Nidhi Malik and Vajenti Mala
School of Computer Science & Systems Sciences, Jawaharlal Nehru University, New Delhi

Received 18 October. 2011; Revised 21 May 2012; Accepted 19 June 2012

SUMMARY

With the widespread and increasing availability of text documents in electronic form and need for managing the information that resides in the vast amount of available text documents, the field of *Text Mining* is receiving a lot of attention. For the same reason text mining is also becoming important task in agriculture domain. Traditional text mining was based on keyword based extraction which has some limitations , as this method does not consider inherent semantic relationship among the words. Semantic text mining can overcome these limitations. Semantic text mining aims at discovering the hidden information from the text documents based on relationships of the terms occurring in them. Linguistic Ontology is one of the most widely used tools for semantic text mining. The objective of this paper is to highlight the role of linguistic ontology in mining textual information with a focus on agriculture domain. Specifically, we present an algorithm for extracting concept clusters from text documents using WordNet ontology. We have taken documents from the agriculture field and performed experiments on them. The results are encouraging and encourage us to explore this area further.

Keywords: Information retrieval, Semantic relations, Concept cluster, WordNet ontology, Concept based information retrieval, Agriculture.

1. INTRODUCTION

With the widespread and increasing availability of text documents in electronic form, large amount of unstructured information is available in digital form. Huge scale and unstructured nature of electronic text causes information overload for user. Due to the always increasing need for managing the information that resides in the vast amount of available text documents, the field of *Text Mining* (Kochandy 2006) is receiving a lot of attention. For the same reason there is a vast potential of text mining based research in agriculture domain. Traditional text mining approaches are based on keyword based extraction, having some inherrent limitations. Main limitation being that these approaches are based on lexicographic similarity and does not take into account syntax or semantics of natural language text. Semantic text mining can overcome these limitations. Semantic text mining aims at discovering the hidden information from the text documents based on relationships of the terms occurring in them (Tavrianou 2007). Ontology is one of the most widely used tools for semantic text mining. Since ontology can provide semantic patterns from text documents, ontology based text mining is coming up as an important research area. The objective of this paper is to highlight the role of linguistic ontology in mining textual information with a focus on agriculture domain. Specifically emphasis is on using ontology for extracting concepts from text documents in Agriculture domain.

The paper is divided in 6 sections. In Section 2 we present role of ontology in context of semantic text mining. In Section 3 we present our ontology based model for concept extraction. Section 4 presents our algorithm for identifying concept clusters and assigning semantic weights to the concepts used in the documents.

Corresponding author : Aditi Sharan
E-mail address : aditisharan@gmail.com

With the help of an example, different concept clusters for a document in Agriculture domain are shown. In Section 5 we present some applications of semantic text mining. Finally we conclude in Section 6.

2. ONTOLOGY BASED APPROACH IN CONTEXT OF SEMANTIC TEXT MINING

Before going into details of ontology based approach in context of Text mining, we first look into what is ontology.

"Ontology" is originally used as philosophical word, which means the branch of metaphysics that deals with the nature of being (Gruber 1993). But currently, in the field of context of knowledge sharing, the term "Ontology" means a specification of a conceptualization (Andreas 2003). That is, "Ontology" is a description of the concepts and relationships that can exist for a community or a particular field, or we may say that Ontology is a model that represents a set of concepts within a domain and the relationships between those concepts. It is used to reason about the objects within that domain.

Ontologies can be seen as a special kind of graph describing the entities, their properties and relationships between them. The basic building block of ontologies is concepts and relationships between them. Concepts can be thought of as sets and appear as nodes in ontology, edges represent relationship between these concepts. In ontology concepts are treated as classes generally described by one or more terms so a term need not only represent one concept in ontology.

Ontology can be used for finding semantic similarity between two terms (concepts). To find similarity between terms using ontology, terms to be matched are mapped onto ontology and the similarity based on various structural properties of ontology is computed. Ontology are very well being used in area of information retrieval; these are used effectively by many retrieval models. Retrieval models which are based on ontology may use semantic measures that rely on various structural properties of ontology like depth, density, link strength etc.

Though the idea presented in this paper can be generalized for any ontology, we have focused on use of linguistic ontology. Specifically we have used [Wordnet], which is a lexical ontology of English words, developed under direction of George A. Miller at Princeton's university. In Wordnet nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each representing one underlying distinct lexical concepts. Synsets are linked by means of variety of lexical and semantic relations such as synonym,antonym, hypernym-hyponym, meronym-holonym etc. Although each relation has its own importance, Wordnet is heavily grounded on its taxanomic structure that employs the IS-A (hypernym-holonym) inheritance relation. Wordnet can be viewed as a graph where synsets are vertices and relationships form edges.

3. ONTOLOGY BASED MODEL FOR CONCEPT EXTRACTION

Kang *et. al.* (2005) have provided a method for exploiting ontology for concept extraction. Using Kang's approach a document can be represented as a collection of concept clusters. *A concept cluster is a weighted lexical chain that represents one aspect of the meaning of a document and expresses the degree of relatedness among the semantic terms within a document*. Some limitations of this model are as follows:

1. There can be different possible representations for same concept cluster.
2. Due to ambiguity in the representation of concept cluster, we cannot develop a well defined algorithm for constructing concept cluster.
3. There is redundancy in resulting concept clusters as same node can appear multiple times.

A modified representation of clusters has been suggested in Sharan (2011) which is able to overcome some of the limitations of Kang's approach. This modified approach has been used to construct concept clusters in this paper.

Representation of Concept Cluster

Concept clusters are extracted using Wordnet ontology. A concept cluster is represented through a graph. Each node of the graph represents a unique term. Each term has an attribute giving frequency of the term in the document. This frequency actually captures identity relationships among the terms. Further this model facilitates calculation of weight for each term, indicating importance of each term. To be more specific, concept based model is represented as follows:

Let T={(t_1,f_1),(t_2,f_2), . . . , (t_m,f_m) } be the set of terms and their frequency in a concept cluster, where t_i is term and f_i is frequency of *i*th term. Let R = {identity, synonyms, hypernym-hyponym, meronym-holonym} be the set of lexical relations. Let $M(r_k, t_g)$ be the sum of frequency of all the terms linked to term $t_g \in T$ through relation $r_k \in R$ and let $W_{(rk)}$ be the weight of relation r_k. Then the score $S_{\text{Term}}(t_g)$of term t_g in a concept cluster is defined as:

$$S_{\text{Term}}(t_g) = (f_g - 1) \times W(r_1) + f_g \times \sum_{k=2}^{4} M(r_k, t_g) \times W(r_k)$$

$$1 \le g \le m \qquad (1)$$

where r_1, r_2, r_3, r_4 are identity, synonym, hypenym-hyponym and meronym-holonym relations respectively.

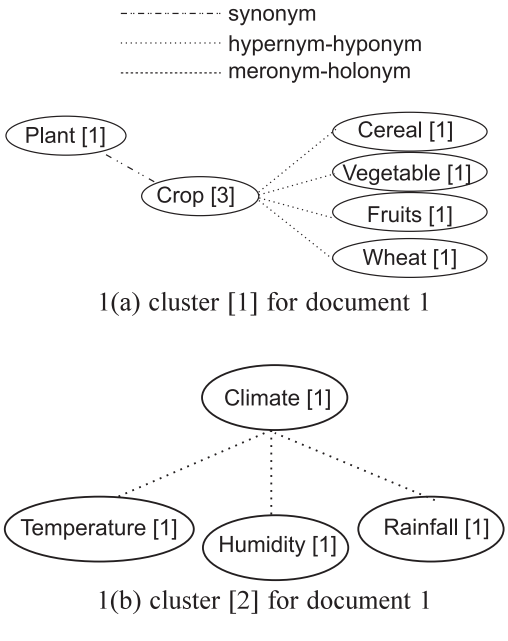
4. ALGORITHM FOR IDENTIFYING CONCEPT CLUSTERS

Before giving the algorithm for identifying concept clusters we start with defining concept clusters in terms of a graph. Each document can be represented as graph where nodes represent terms and edges represent semantic relations between the terms. Each concept cluster then represents a connected component of the graph. Wordnet ontology can be used for extracting semantic relationships between the terms present in the document. Now, we present simplified algorithm for extracting concepts from the text documents.

```
1. Initialize weights $W(r_k)$ of semantic relations $r_k$.
2. For a document $D$ find $T$, where $T$ represents the set of all nouns in document $D$ (use Part of Speech (POS) tagger (Mucherino et al. 2009).
3. Repeat
   3(a) for document $D$ identify a concept cluster represented as a graph $C$ ($V$, $E$)
   3.1 Start with a term $x$ □ $T$ and add $x$ as vertex $V$ to cluster $C$.
   3.2 Assign $f_x$ as frequency of $x$ in $D$.
   3.3 Assign $T = T - x$.
   3.4 For each term $y$ □ $T$ related to any $x$ □ $V$ through any relation $r_k$
       (a) Assign $f_y$ as frequency of $y$ in $D$
       (b) Use $y$ to extend the graph $C$ by adding $y$ to $V$ as vertex and relation $r_k$ as edge between $x$ and $y$.
       (c) Assign $T = T - y$.
       Until $T$=$\phi$.
4. Find weight of each term $x$ using concept cluster
   4.1 Find the concept cluster to which term $x$ belongs.
   4.2 Repeat following steps within concept cluster.
       4.2.1 For node $x$, identify the relations using the edges connected to $x$ and calculate weight of $x$ using formula given in equation (1).
5. Represent document by the concept clusters identified along with weights assigned by concept based model.
```

Now, we present a small example that explains working of algorithm. We have considered few documents from agriculture domain as shown in Appendix. The weights considered for the semantic relations are 0.6 for identity, 0.3 for hypernym-hyponym and 0.1 for meronyms.

From this algorithm, concept clusters identified for Doc 1 are:



1(a) cluster [1] for document 1

1(b) cluster [2] for document 1

Fig.1 Clusters for document 1
(Inside the oval, the numeral shows the frequency of the term in the document.)

In document1 one can observe that emphasis is on concept crop, however there is one more cluster for the concept climate. Concept clusters that are being generated after applying this algorithm is also depicting that term crop is the most prominent one in this document, having maximum weight. After constructing concept clusters we can calculate weight of each term using equation (1). Accordingly weight of the term crop is equal to 6.8. Though there are some other concept clusters also which are isolated so they are omitted. Similarly we can generate concept clusters for other documents also. Concept clusters for other documents are shown in the Appendix 2.

Our second experiment consists of comparing the results of concepts based IR with the existing tf based IR technique. For this experiment we have taken five documents from agriculture domain (see Appendix 1) and the query considered is 'crop'. An intuitive judgement shows that the expected results in ranked order should be : doc 1, doc 2. doc 5, doc 4, doc 3.

Table 1 shows the weights assigned to the query using the *tf* technique and Table 2 shows semantic weights assigned to the same query using the algorithm discussed above.

Table 1. Weights based on *tf* technique

Doc/Term	Crop
Doc 1	0.17
Doc 2	0.14
Doc 3	0.06
Doc 4	0.33
Doc 5	0.18

Table 2. Weights based on concept based model

Doc/Term	Crop
Doc 1	6.8
Doc 2	18.0
Doc 3	0.4
Doc 4	2.4
Doc5	6.0

Table 3. Ranked relevant documents for the query "crop" based on both the models and expected results

Term Frequency	Concept based	Expected Results
Doc 4	Doc 2	Doc 1
Doc 5	Doc 1	Doc 2
Doc 1	Doc 5	Doc 5
Doc 2	Doc 4	Doc 4
Doc 3	Doc 3	Doc 3

Table 3 shows the ranked results of relevant documents retrieved using the tf technique and the concept based technique. The result shows that the documents retrieved by the concept based techniques are more accurate. It can be clearly observed that this algorithm is capturing semantic similarity between the terms occurring in the documents.

5. APPLICATIONS

Text mining has its applications in various areas of research including NLP, biomedical sciences, telecommunications, biotechnology etc. In this paper our focus is on agriculture domain. Till now in agriculture most of the data mining techniques are implemented on structured data. Comparatively, limited amount of work is done on mining unstructured data. Considering the availability of large amount of data in textual domain, text mining is an open research area. Moreover semantic text mining is a very new and challenging field in text mining. Therefore, it has immense potential applications.

Some of the general applications of semantic text mining are:

(i) Finding Features/Concepts

The ultimate goal of text mining is analysis of textual content (Mucherino 2009). In order to analyze the text document and to extract relevant information from these documents, it is important to extract proper features from the documents (Liritano 2001). In traditional information retrieval, features are represented by the terms present in the document (Sharan 2011). Semantic text mining allows relationship between these terms to be used as features. So, it provides more meaningful features that can deal with the meaning and relationship between the terms.

As seen previously ontologies can be very useful for extracting semantic features from the text documents. Terms which are semantically related can be further helpful for concept extraction. Ontologies can also be used to provide expert background knowledge about a domain (Nagai 2008).

(ii) Text Categorization

Text categorization is simply classifying a set of documents into some predefined categories. Keyword based techniques have the limitation that they only consider the actual content of the documents. Due to absence of semantic aspects these techniques fail to give good results. With the help of ontologies text can be categorized into specific categories based on their semantic features. These semantic features involve meaning and relationships of words. Many algorithms are available for Categorization (Nicholas 2007) which aims to discover natural grouping in a collection of documents.

(iii) Improving Retrieval Efficiency

As traditional information retrieval techniques are based on keyword based similarity, these techniques give results only on the basis of lexicographic similarity. Traditional information retrieval is based on vector space model which is not able to capture the semantic aspects of the documents. Most of the times, the user is interested in retrieving the documents giving the meaning of the query he/she has entered instead of the number of documents containing just the query term. In such cases, concept based representation of information using ontologies can be very helpful. With the help of semantic text mining, this gap can be fulfilled and efficiency of information retrieval can be improved.

(iv) Topic Discovery

There is no doubt that a document collection organized into a hierarchy is very helpful for users. However, it is not enough. Users also need to determine at a glance whether this information is of their interest by means of some kind of summary.

There are two key issues here (i) how to discover the structure of topics in a collection, that is, how to identify not only the topics but also the subtopics they comprise and (ii) how to properly summarize these topics and their subtopics (Nicholas 2007). With the help of different types of relationships that can be obtained with the help of ontologies it becomes very easy to show the hierarchical relationships in the documents.

(v) Text Clustering

Cluster is a collection of data objects which are similar to one another within the same cluster but dissimilar to objects in other clusters (Nicholas 2007). Clustering is beneficial in information retrieval in the sense that documents can appear in multiple subtopics ensuring that a useful document will not be omitted from the search results (loh 2000). It is different from text categorization mainly because it does not provide standard labeled classes. Therefore classification is done based on similarities between data items. Defining similarity itself is very crucial to clustering. Traditional similarity approaches are based on finding overlap between the keywords within the documents. Again these similarity measures are unable to capture semantic similarity aspects. Ontologies can provide various relationships which can be used to find semantic similarity between the documents/terms. Therefore clustering algorithms based on semantic similarities can result in building more meaningful clusters in comparison to keyword based similarities.

CONCLUSION

Semantic text mining differs from traditional text mining in the way that it is able to capture the meaning and relationships between the words present in the document. With the help of ontologies it becomes easier to handle the complex semantic relationships among word/terms in the text. By tracking the concept hierarchy ontologies make the text mining approaches recognize the relationship between two terms for measuring the semantic similarity between documents.

Semantic text mining techniques can be very beneficial in agriculture domain also. Due to availability of large amount of textual data in Agriculture domain, it is important to develop semantic approaches that can automatically extract concepts from such data. In this paper, we have suggested the use of ontologies for concept extraction. Specifically, we have presented a way to represent a document as a collection of concept clusters. Further an algorithm has been developed for finding concept based clusters and for assigning semantic weights to the terms in a document. The working of the algorithm has been shown for extracting concepts from documents in agriculture domain. The implementation has been shown using WordNet ontology which is a linguistic ontology for English language. It has been shown that even this general ontology can be useful for finding semantics of documents in agriculture domain. The use of agriculture based ontologies can further improve the results of our approach. Experiments have been performed on some documents taken from agriculture domain. In future the results need to be further investigated on larger data set to see the extent of improvement. Further our next attempt is to give a more computationally efficient algorithm.

REFERENCES

Andreas, H., Andreas, N. and Gerhard, P. (2005). *A Brief Survey of Text Mining*. LDV Forum - GLDV. *Journal for Computational Linguistics and Language Technology*, **(20)1**, 19-62.

Tavrianou, A., Andritsos, P. and Nicoloyannis, N. (2007). Overview and Semantic issues of Text Mining. *SIGMOD Record*, **36(3)**, 23-34.

Gruber, T.R. (1993). A translation approach to portable ontologies. *Knowledge Acquisition*, **5(2)**, 199-220.

WordNet (1998). *An Electronic Lexical Database*. edited by Christiane Fellbaum , MIT Press.

Kang, B., Kim, D. and Kim. H. (2005). Fuzzy information retrieval indexed by concept identification. *LNAI* **3658**, 179-186.

Kochandy, Manu. *Text Mining*. Charles River Publications.

Sharan, A., Joshi, M. and Pandey, A. (2011). Exploiting ontologies for concept based information retrieval. *ICISIL 2011*, **139(1)**, 157-164.

Loh, S., Wives, L.K. and de Oliveira, J.P.M. (2000). Concept-Based Knowledge Discovery in Texts Extracted from the Web. SIGKDD Explorations, ACM SIGKDD, July 2000, **1(1)**, 29-39.

http://nlp.stanford.edu/software/tagger.html, POS Tagger

Mucherino, A., Papajorgji, P.J. and Pardalos, P.M. (2009). *Data Mining in Agriculture.* Springer Dordrecht Heidelberg, London, New York.

Nagai, M., Horanont, T., Sunithi, T., Kawatrakul, A., Prathumchai, K. and Shibasaki. R. (2008). Development of ontological Information for Agriculture in Thailand. IAALD AFITA WCCA 2008, 479-483.

Nicholas O. Andrews and Edward A. Fox (2007). Recent developments in document clustering. *Virgina Tech.*, Blackburg, VA 24060.

Liritano, S. and Ruffolo, M. (2001). Managing the knowledge contained in electronic document : a clustering method for text mining. *IEEE*, Italy. 454-458,

http://www.icar.org.in/

APPENDIX 1

Document 1

When plants of the same kind are grown and cultivated at one place on a large scale, it is called a **crop**. For example, crop of wheat means that all the plants grown in a field are that of wheat. Crops are of different types like cereals, vegetables and fruits. These can be classified on the basis of the season in which they grow. India is a vast country. The climatic conditions like temperature, humidity and rainfall vary from one region to another.

Document 2

There is a rich variety of crops grown in different parts of the country. Despite this diversity, two broad cropping patterns can be identified. These are:

(i) Kharif : The crops which are sown in the rainy season are called kharif crops. The rainy season in India is generally from June to September. Paddy, maize,cotton, etc., are kharif crops.

(ii) Rabi : The crops grown in the winter season are called rabi crops. Their time period is generally from October to March. Examples of rabi crops are wheat, gram and linseed etc. Besides these, pulses and vegetables are grown during summer at many places.

Document 3

Because of its nutritional excellence and ease of cultivation, cabbage deserves greater use and research attention. Its high content of proteins, vitamins and minerals needs to be communicated to consumers in order to stimulate the demand for its production. This crop has C4-type photosynthesis and is therefore productive under hot and dry conditions, a trait of increasing value in the face of climate change.

Document 4

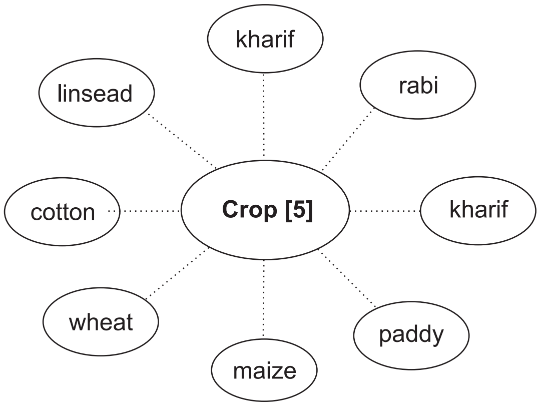
Rice is the main grain crop of India. India ranks second in the world in production of this crop. About 34% of the total cultivated area if the nation is under rice cultivation. Out of the total production of food grains, production of rice is 42%. Rice is cultivated in areas having annual average rainfall of 125 cm and average temperature of 23 degree Celsius. Major cultivating areas for this crop are north east India, eastern and western coastal regions and river basin of Ganga. West Bengal, Punjab and Uttar Pradesh are the major rice producing states. Besides, Tamil Nadu, Karnataka, Orissa, Haryana, Bihar, Chhattisgarh, Assam and Maharashtra also produce rice.

Document 5

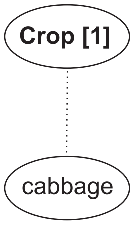
Crop yields for some farms within India are within 90% of the best achieved yields by farms in developed countries such as the United States and in European Union. No single state of India is best in every crop. Indian states such as Tamil Nadu achieve highest yields in rice and sugarcane, Punjab enjoys the highest yields in wheat, Karnataka does well in cotton, Bihar does well in pulses. These differences in crop productivity within India is a function of local infrastructure, soil quality, micro-climates, local resources, farmer knowledge and innovations. However, one of the serious problems in India is the lack of rural road network, storage, logistics network, and efficient retail to allow free flow of farm produce from most productive but distant Indian farms to Indian consumers. Indian retail system is highly inefficient. Movement of agricultural produce within India is heavily and overly regulated, with inter-state and even inter-district restrictions on marketing and movement of agricultural goods.

APPENDIX 2 :

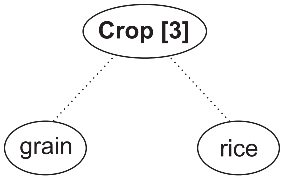
Concept cluster for document 2



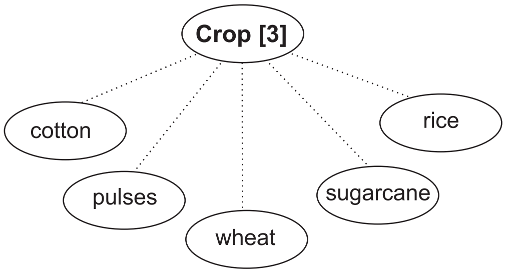
Concept cluster for document 3



Concept cluster for document 4



Concept cluster for document 5


Crop [3]
cotton
rice
pulses
sugarcane
wheat