IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 3, Ver. IV (May-Jun. 2014), PP 86-91
www.iosrjournals.org

Improved Text Analysis Approach for Predicting Effects of Nutrient on Human Health using Machine Learning Techniques

Dhanasekaran Kuttiyapillai[1], Rajeswari Ramachandran[2]
[1]*(Computer Science and Engineering, Info Institute of Engineering/Anna University, India)*
[2]*(Electrical and Electronics Engineering, Government College of Technology/Anna University, India)*

Abstract : *A text analysis method is introduced which processes the unstructured information from document collection in order to support efficient text classification and information extraction. The Information extraction helps the system to discover useful knowledge for the users. The keyword based information search method fails due to the generation of many false positives. It leads to inaccurate generation of user expected result for many applications where context sense plays a major role while identifying the relevant features. The method developed can solve this problem by using combination of techniques from machine learning and natural language processing. The text contents are preprocessed and applied with entity tagging component. The proposed method performs accurate information extraction by improving precision and recall value. The automatic generation of rules provides an easy way to predict the effects of nutrient on malnutrition. The bottom-up scanning of key-value pairs helps to speed up the content finding to generate relevant items to the task. Our method can outperform other methods by eliminating sub tree which has an overestimating heuristic. Our approach will be useful to prevent errors generated through misinterpretation of available facts and improves the accuracy of information extraction in many text analysis applications.*

Keywords: *Classification, Knowledge discovery, Machine learning, Semantic relevance, Term extraction*

I. Introduction

The natural language processing plays a major role in many text processing applications. The extraction system takes as input the text content, it fails to perform better due to misinterpretation of facts, irrelevant features, and sometimes inadequate amount of data for further analysis. The English syntax contains at least two sentence structure rules for a given sentence. In general, the first one Noun Phrase (NP) introduces a sentence under the domination of an NP.The second one Verb Phrase (VP) involves a sentence under the domination of a VP.

The automatic generation of these rules using machine learning techniques for a complex sentence [1][2] helps the computational system to interpret the meaning of information in a context very fast as compared as keyword based retrieval.

The sentence structure rules are used as follows:
S->NP VP
NP->Noun|Pronoun|Name|NP PP|Article Noun|NP RelClause
VP->Verb|VP NP|VP Adjective|VP PP|VP Adverb
PP->Preposition NP
RelClause->that VP

In context-based applications, the sentence structure optimization helps us to make decision accurately. The support for reasoning and learning are realized with optimistic generation of solution. To improve accuracy in information retrieval or text mining, the term frequency-inverse document frequency (tf-idf) is used as a method. This method evaluates the importance of word in a document [3].

In order to have an insight into our basic model, we need to elaborate algebraic model which represents textual information as a vector. The components of the vector could represent the importance of a term or the absence or presence of it in a document classical VSM (Vector Space Model) [4][5].Initially a dictionary of terms present in documents is created for modeling the document in a vector space. In order to do that, all terms in the document are selected and converted to a dimension in the vector space. There are some kinds of words (stop words) in almost all documents; extracting important features from document can identify them among other similar documents. So, in this paper, the terms like “the, is, at, on, etc” is ignored in order to improve the retrieval speed. So we ignore them in the information extraction process.

II. Previous Research

1. Introduction

In feature-based mining [6][7][8], relevant features are grouped to support efficient information extraction. The positive and negative attributes are identified to produce an aggregated data set. The features of climate analysis are, e.g., “very high”, “too cool”, “very low” and “too hot”. People use different words or phrases to describe the same situation. Reference [5] has investigated the efficiency of feature extraction.

Grouping of those features is critical for feature-based extraction in an application based on natural language processing. Although WordNet and other dictionaries can support this task, they are not sufficient.

First, many words and phrases may not be synonyms in a dictionary but they may refer to the same attribute in an application domain. For example, “sweet” and “pleasant” are synonyms in communication summary, but they are not synonyms in food analysis summary.

The task of determining the logical expressions indicate the same feature dependent on the user’s need. Grouping feature expressions manually into appropriate groups is time consuming process since there are hundreds of feature expressions. Reference [9][10] has analyzed the performance of feature grouping compared various methods of supervised learning and unsupervised learning. As discussed, the previous algorithm does not produce expected result because of increasing complexity of data analysis.

Due to the vast amount of data on the information world, people focus on getting the useful and relevant message from the information sources. This created interest among researchers to find efficient method for extraction, classification and prediction task in data mining and natural language processing[11][12][13]. Feature extraction is one method which supports this kind of data analysis and visualization. In general, the feature extraction method can be classified into five methods, document frequency, information gain, mutual information, $\chi 2-$ text, and cross entropy. Each method has its own positives and negatives.

The popular methods like neural network, Support Vector Machine (SVM), KNN has been applied to solve problems. But they are not suitable in some cases where complex nonlinear functions need to be represented. Reference [14] has used parameter-free semi-supervised learning, has shown applicability of gene classification.

SVM and KNN cannot efficiently support both natural selection and classification task. A single layer network has a very simple learning algorithm with less expressive power. Multilayer networks, on the other hand, are much more expressive. Although they can represent nonlinear functions, it gets stuck at local optima and because of the high dimensionality of the weight space the training may be very hard[15].

The advanced text analysis method and classification algorithm can help clearly expressing its knowledge well without creating ambiguity. Our extraction techniques applied with dynamically changing constraints to predict the correct class label of given samples in food safety prediction domain.

The goal of this paper is to design and text analysis system for predicting effects of nutrient on human health. It can overcome the problems of clustering-based and the association-based classification method. Instead of using the sequential extraction to find the useful rules, our method extracts a set of highly relevant phrase-based rules which is found and stored in a data store based on corpus-based analysis. These rules are entity-tagged elements and it gives correctness on retrieving most relevant and task-specific sentence sequences for accurate decision-making with regard to the given training instances.

Specifically, given a training data which consists of set of articles or documents and a threshold-value for the group of sentences, the task is to discover useful matching information for each of the instances in the training data set and then build an efficient classifier from these rules. The training instances must satisfy the stopping condition in order to assure consistency in information retrieval. For constructing our model, this paper adds transformation technique to convert categorical data into numerical data based on weighting method. The weight adjustment technique in this paper supports more efficient retrieval of k highest sentence rules more specific to the information extraction task.

The heuristics proposed are as follows:
H1: Instance in the same rule set cannot have same content ID.
If exists, add to low frequency item list. For RL (Relevant List), apply mapping between training data e.g. noun and the test data e.g. noun then test whether they include the feature expression in the same set.
H2: Select neighbor node with the highest rank value to support efficient feature maximization. The rank computed based on weighting method.

Our approach used different number of rules (v) for each voting group and each execution used a different random seed to create the initial population. Thus there exists interdependency between executions. We set v is equal to minimum 3 in a voting group. For multiclass problem, the value of v is changed to 5 for making decision.

III. Research Method

1. Introduction

The input to the proposed constraints-based classification method include: A set of sentences S, a set of feature expressions from S.Constraints include: Include items (II), Exclude Items (EI), soft constraints, hard constraints, and dynamic constraint for changing needs.

Firstly, the user will assign feature expressions to the predefined group label C.Then the algorithm assigns the remaining discovered feature expressions (F2) to C.In information extraction stage, the gene learning performed to learn terms, synonyms, concepts, concept hierarchies, relations, and grammar rules.

The current Named Entity Recognition (NER) has been used to populate general-purpose knowledge base. This method is only effective for recognizing instances of general concepts such as “animals” and “plants”. It limits its applicability to more specific and abstract domains. Our method extends its usability for more specific domains as well.

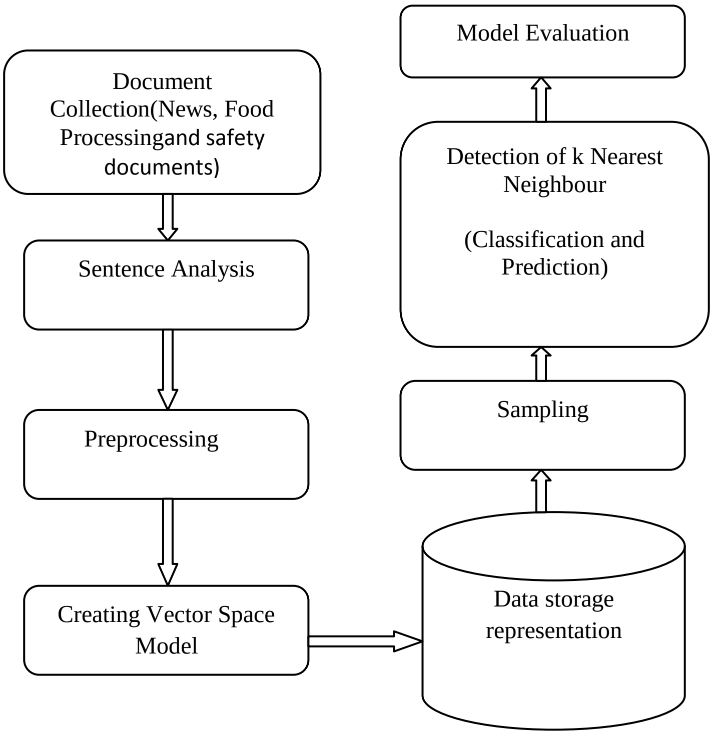
Model Evaluation
Document Collection(News, Food Processingand safety documents)
Detection of k Nearest Neighbour
(Classification and Prediction)
Sentence Analysis
Sampling
Preprocessing
Creating Vector Space Model
Data storage representation

Figure.1 overall architecture of proposed Text Analysis system

The process steps in our approach are: Tokenization, Preprocessing and Entity Attachment, Classification based on KNN, similarity content finder, Phrase Extraction, and visualizing result. The major components of proposed system are discussed in the following sections. The overall diagram of our proposed method is shown in Fig.1

1.1 Tokenization

The tokenization process determines sentence boundaries, and then separates the text into a stream of tokens (or terms). During the process it removes unwanted tokens by means of preprocessing a corpus of text documents so that the extracted items are passed to the classification model in order to support accurate prediction task. It uses single space to delimit token and the double spaces are used to delimit sentences. In order to support the keyphrase extractor to parse the text efficiently, the document texts are tokenized correctly.

1.2 Pre-Learning stage

The analyzer in our approach extracted the noun phrases using a sentence structure rules, it is composed of different sequences of parts-of-speech tags. For the experimental purpose, the limit for the longest recognizable noun phrase pattern was set to fifteen words in length. We consider the shortest pattern at the length of one, the single noun. The fifteen-word limit can lead to some unexpected error, since the tagger is likely to misuse noun phrases longer than the specified limit, which may or may not be useful terms for our extraction process.

For example, the sentence, “We proposed this notion as evidence of classification of weather condition for the cultivation area.” which gives the noun phrase “notion as evidence of classification of weather condition for the cultivation”, but consequently it will lead to the phrase “cultivation area” being lost. The tagged words used some rules using a window of fifteen words in size. This window moved over the words of the text, the window contents are used with set of noun phrase patterns. The delimiters are used to truncate the window.

The weight-based parallel comparison method is used to determine the best key phrase using noun in the text, because some rules are subsets of other rules. If a noun phrase is located, then window will slide to the next word following the phrase reading the contents of a new fifteen-word window.

1.3 Finding k Nearest Neighbor using Text Classification method

Our paper used grammar rule of natural language as a base to formulate constraints like IC (Include Constraint) and EC (Exclude Constraint), soft constraints, and hard constraints in some cases.

The feature expressions sharing some common words are likely to belong to the same group. For example, “plant life” and “plant power”. Similarly, the feature expressions that are synonyms are likely to belong to the same group. For example, “weather” and “climate”.

The heart of our feature grouping is POS tagging which adopts latent semantic mapping (LSM).The steps include 1) for each input sentence S, LSM is used to construct a neighbors of globally relevant training sentences.2) Extraction of associated POS sequences, and 3) Leveraging the targeted evidence to inform the tagging process.

In our approach, function prediction is divided into two parts: 1) Extraction of neighbor context features.2) Prediction based on those features.

The sensitivity of the contents in functional classification tree for a target gene plays a major role in the neighborhood feature extraction step. The implementation is based on the similar neighborhood hypothesis that is formulated above.

Since the support vector machine has successful applications in many complex, real-world problems such as text and image classification, hand-written recognition, data mining, bio-informatics, medicine and bio-sequence analysis, stock market analysis, we used KNN, so no need of random restart search to perform search efficiently and the algorithm is not affected by local maxima. Mapping Function implemented in our approach predicts the feature accurately. The failure rate is reduced to an acceptable level by using efficient mapping constraints

Our algorithm generalizes well on unseen data. The ultimate aim is to find the global minimum or minimal target state where the goal should approach the positive result by satisfying its parsing criteria. The parsing method of our approach used a library of standford parser to search through a state space tree on a bottom-up manner and it finds appropriate place to replace with a suitable grammar rule.

Our approach proceeds with required inputs. It uses score and similarity measure to check if the feature expression is a constrained one. Initially the edges are added for each feature expression u_m associated with feature expression u_n, where u_m, $u_n \epsilon U$.Subsequently the connected components tree are gathered. For each of these components, for each feature group we repeat the loop to find score.

Finally the connected components are added to RL.In the next stage, a function for checking constraints is called with two parameters x1, x2.The main operation is to check if both x1and x2 are synonyms by ensuring single word expression. If it does not satisfy the condition, we add that x2 is a phrase.

The outer part of a condition checking ensures that x1 is a phrase when inner condition is false. For similarity checking on phrases we use function to calculate similarity.

IV. Result And Discussion

In this section of paper we present analysis of the results of our method. For assessing the performance of classifier, the nutrient articles in the dataset is divided into two sets. One is called a training set and the other one is called a test set. The percentage of data selection is varied in order to find the result with various combinations of data which are selected from training set.

Some set of data selected from the test data necessarily satisfied the requirements of training process so that the relationship between unknown and known items are clearly established in the model by means of our computational model for finding the correct boundary between the different categories of data.

When we considered 70% of data from the training set and 30% from the test data, our process shows 94.66% accuracy. We generated the development test data to avoid overfitting of data during estimation of accuracy. If the accuracy drops then we go for tuning the test data. The metrics used are defined as follows:

$$\text{Recall}= \frac{\text{Number of words information that are correctly predicted}}{\text{Number of word items that are present in the sample}}$$

$$\text{Precision} = \frac{\text{Number of consequent that are correctly predicted}}{\text{Number of consequent that are predicted}}$$

$$\text{F1} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision} \times \text{Recall}}$$

At least one informative class should be associated with an informative instance. When T is given all informative instances in the tree and is considered as contained in the training instance set. It has been observed that the relevant information makes the predictor functions more specific in a given situation. The result generated in our process is shown below:

The nutrient dataset contains 150 instances, which corresponds to three frequent class of nutrients. Each instance contains four attributes: calcium in mg, protein in mg, zinc in mg, Iron in mg.

RWeka was used to automatically find the best value for k which range between 1 and 10. Evaluating classifier using nfold=10, the following results were obtained.

```
Correctly classified instance              142          94.6667%
Incorrectly classified instances     8              5.3333%
Kappa statistic                            0.92
Mean absolute error                        0.041
Root mean squared error              0.1414
Relative absolute error              9.2339%
Root relative squared error          29.9987%
Total number of instances            150
Confusion matrix:
a    b    c    <-- classified as
50   0    0    a = potassium
0    46   4    b = iron
0    4    46   c = protein
```

The above result shows that a kNN classifier makes few mistakes in a dataset. The confusion matrix shows that all misclassifications are between instances of protein and iron.Weka was used to evaluate the classification algorithm developed, which makes machine learning, and prediction, easier to experiment with the parameters.

Many subsets of features including development set items considered for limiting the training set size. In this case, the target gene is set as unknown and then a predictor computable function executes using the remaining information in a classification model.

The score distribution of all methods are necessary for every class and then the candidate functions or populations selected for calculating the accuracy for different values of different candidate functions. The proposed method shows somewhat promising result in some combination of data selection.

V. Conclusion

Since the information world has increasing collection of documents, the problems in information retrieval due to the complex syntactic and semantic structure, need to be addressed to find the right solution for the problem. Our proposed method performs sensitive text analysis and extracts useful information with the help of formulated heuristics. The effects of nutrient for malnourishment can be identified in a health care unit and it can be prevented by predicting situation-specific event. Similarly the safety criteria can be presented to prevent the future consequences.

The two most important tasks in information extraction from the Web are webpage structure understanding and natural language sentence processing. However, little work has been done toward an integrated statistical model or classification model for understanding and processing natural language sentences within the more specific domain.

Our approach introduced a new approach for extraction using a combination of natural language tools and efficient text classification method. Text finder in our approach improves classification accuracy by utilizing optimization constraints and the built-in library of RWeka.

References

[1]. Wang Wei, Payam Barnaghi and Andrzej Bargiela, Probabilistic Topic Models for Learning Terminological Ontologies, IEEE Transactions on Knowledge and Data Engineering, 2010, pp.1028-1040.

[2]. Qinbao Song, Jingjie Ni and Guangtao Wang,A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data, Journal IEEE Transactions on Knowledge and Data Engineering, 10(10), 2011,pp.1-14.

[3]. Hong Huang and Hailiang Feng,Gene Classification Using Parameter-free Semi-supervised Manifold Learning,IEEE/ACM Transactions on Computational Biology and Bioinformatics, 10(10),2011,pp.1-13

[4]. Ahmed Rafea, Hesham A. Hassan, Mohamed Yehia Dahab, TextOnto Ex: Automatic Ontology Construction from Natural English Text,International conference of Artificial Intelligence and Machine Learning,2006.

[5]. Topon Kumar Paul and Hitoshi Iba,Prediction of Cancer Class with Majority Voting Genetic Programming Classifier Using Gene Expression Data,IEEE/ACM transactions on computational biology and bioinformatics, 6,2009, pp.353-367.

[6]. Michele Carenini, Angus Whyte, Lorenzo Bertorello, Massimo Vanocchi,Improving Communication in E-democracy Using Natural Language Processing,IEEE Intelligent Systems,2007,pp.20-27.

[7]. Rile Hu, Chengqing Zong and Bo Xu,An Approach to Automatic Acquisition of Translation Templates Based on Phrase Structure Extraction and Alignment,IEEE Transactions on Audio Speech, and Language Processing, 14(5),2006,pp.1656-1663.

[8]. Smith Aree Thunkijjanukij, Asanee Kawtrakul, Supamard Panichsakpatana, Uamporn Veesommai,Lesson learned for ontology construction with Thai rice case study, World Conference on agricultural information and IT, 2008, pp.495-502, in press.

[9]. Antonio M. Rinaldi,An Ontology-Driven Approach for Semantic Information Retrieval on the Web, In ACM Transactions on Internet Technologies, 9(10), 2009.

[10]. Elena P. Sapozhnikova,Multi-label classification with art neural networks,In Second International Workshop on Knowledge Discovery and Data Mining, 2009, pp. 144-147, “in press”.

[11]. Shubin Zhao Ralph Grishman,Extracting Relations with Integrated Information Using Kernel Methods,In Proceedings of the 43rd Annual Meeting of the ACL,2005, pp. 419–426, “in press”.

[12]. Hongbo Liu1, 3, Ajith Abraham2, and Benxian Yue3,Nature Inspired Multi-Swarm Heuristics for Multi-Knowledge Extraction, Advances in Machine Learning II, 2010,pp. 445–466, “in press”.

[13]. Christine W. Chan,From Knowledge Modeling to Ontology Construction, International Journal of Software Engineering and Knowledge Engineering (IJSEKE),2004.

[14]. Harith Alani,Position paper: Ontology construction from online ontologies,” Proceedings of the 15th international conference on World Wide Web ,2006,pp. 491 – 495.

[15]. Xu Binfeng,Luo, Xiaogang Peng Cenglin, Huang Qian, Based on ontology: construction and application of medical knowledge base”, IEEE International conference on complex medical engineering, 2007,pp.586- 589.