Text Mining in Bovine Diseases

R.Santhanalakshmi
Research Scholar,
Dept of MCA, ComputerCenter,
Madurai Kamaraj University,Madurai.

Dr.K.Alagarsami
Associate Proffessor
Dept of MCA, ComputerCenter,
Madurai Kamaraj University,Madurai.

ABSTRACT
Bovine diseases are the common infection occurs in cattle sector, which comes in varieties of forms. Getting the information about these diseases is tedious process. To get the information and optimize the query result of the user, we used Information Extraction in Text Mining. Text mining is the process of analyzing unstructured, natural language texts in order to discover information and knowledge that are difficult to retrieve directly. Information extraction is one of the most important techniques used in text mining, which the process of scanning text for information relevant to some interest, including extracting entities, relations, and events.

Keywords: Bovine Diseases, Information Extraction, Text Mining, Hierarchical clustering.

1. INTRODUCTION

1.1 Bovine diseases:
Bovine leukemia virus (BLV) is a bovine virus closely related to HTLV-I, a human tumour virus. BLV is a retrovirus which integrates a DNA intermediate as a provirus into the DNA of B-lymphocytes of blood and milk. It contains an oncogene coding for a protein called Tax. Nevertheless in its natural host the cattle leukemia is rare. Because the oncogenic properties of the virus were discovered early, a search for evidence of pathogenicity humans started soon after discovery. Mostly farm workers drinking raw milk were tested for disease, especially for leukemia. But neither leukemia nor other signs of infection could be detected. So many in many states it was not tried to get rid of this infection.

Testing strategies have recently changed since the virus was first detected in Cows; "Only very recently have currently available and highly sensitive assays such as Western blot and ELISA been employed in testing human sera. Buehring et al (2003) detected antibodies against BLV p24 capsid antigen in 74% of human sera tested using Western blot, while none of the samples that had given the most intense reaction was positive when tested with one of the earlier techniques."

High Prevalence of virus was found from testing by USDA. "As part of the 2007 dairy study, bulk tank milk was collected from 534 operations with 30 or more dairy cows and tested with an Enzyme Linked-Immunosorbent Assay (ELISA) for the presence of antibodies against BLV. Results showed that 83.9 percent of U.S. dairy operations were positive for BLV."

Many potential routes of BLV transmission exist. Transmission through procedures that transmit blood between animals such as gouge dehorning, vaccination and ear tagging with instruments or needles that are not changed or disinfected between animals is a significant means of BLV spread. Rectal palpation with common sleeves poses a risk that is increased by inexperience and increased frequency of palpation. Transmission via colostrums, milk, and in utero exposure is generally considered to account for a relatively small proportion of infections. Embryo transfer and artificial insemination also account for a small number of new infections as long as common equipment and/or palpation sleeves are not used. While transmission has been documented via blood feeding insects, the significance of this risk is unclear. The bottom line appears to be that transmission relies primarily on the transfer of infected lymphocytes from one animal to the next and that BLV positive animals with lymphocytes are more likely to provide a source for infection.
In general BLV causes only a benign mononucleosis-like disease in cattle. Only some animals later develop a B-cell leukemia called enzootic bovine leukosis. Under natural conditions the disease is transmitted mainly by milk to the calf. Infected lymphocytes transmit the disease too. So for artificial infection infected cells are used or the more stable and even heat resistant DNA. Virus particles are difficult to detect and not used for transmission of infection. It is possible that a natural virus reservoir exists in the water buffalo.

In Europe attempts were made to eradicate the virus by culling infected animals. The first country considered to be free of infection was Denmark**. Soon the United Kingdom followed. Like the North American states, those of the Eastern block in Europe did not try to get rid of the virus. But the Eastern Europe states started to become leukosis free after the political changes at the end of the last century. A very disturbing quote from a USDA fact sheet, "The high individual animal prevalence of BLV reported in the Dairy 1996 study suggests that testing and culling seropositive animals may not be a cost effective method to control the disease. Instead, preventing disease transmission by implementing preventive practices would likely be more cost-effective.

Natural infection of animals other than cattle and buffalo are rare, although many animals are susceptible to artificial infection. After artificial infection of sheep most animals succumb to leukemia. Rabbits get a fatal AIDS like disease similar to rabbit-snuffles, different from the benign human snuffles. But it is not known whether this naturally occurring rabbit disease is linked to BLV infection. "Although several species can be infected by inoculation of the virus, natural infection occurs only in cattle (Bos taurus and Bos indicus), water buffaloes, and capybaras. Sheep are very susceptible to experimental inoculation and develop tumours more often and at a younger age than cattle. A persistent antibody response can also be detected after experimental infection in deer, rabbits, rats, guinea-pigs, cats, dogs, sheep, rhesus monkeys, chimpanzees, antelopes, pigs, goats and buffaloes.

Some long term studies may be necessary, as there appears to be a correlation in instances of cancer among butchers and slaughterhouse workers.*-1 "Several studies have been carried out in an attempt to determine whether BLV causes disease in humans, especially through the consumption of milk from infected cows. There is, however, no conclusive evidence of transmission, and it is now generally thought that BLV is not a hazard to humans.

1.2 Text Mining:

Text mining is a new and exciting research area that attempts to solve the information overload problem. It uses many techniques from data mining, but since it deals with unstructured data, a major part of the text mining process deals with the crucial stage of preprocessing the document collections (using techniques such as text categorization, term extraction, and information extraction). The process also involves the storage of the intermediate representations, techniques to analyze these intermediate representations (such as distribution analysis, clustering, trend analysis, association rules etc).

A typical text mining system begins with collection's of raw documents, without any labels or tags. Documents are then automatically tagged by categories, terms or relationships extracted directly from the documents. Next, extracted categories, Entities and relationships are used to support a range of data mining operations on the documents.

1.3 Information Extraction:

Information Extraction (IE) concerns locating specific pieces of data in natural-language documents, thereby extracting structured information from unstructured text. One type of IE, named entity recognition, involves identifying references to particular kinds of objects such as names of people, companies, and locations [4].

2. PROPOSED METHOD

a) Information Gathering
b) Information processing
c) Information Analysis

2.1 Information Gathering:

In information gathering step we have to collect the all information regarding to the user query. This paper specifically focus the Bovine diseases in biomedicine .so we have to collect the information belong to the biomedicine from various domain as well as various resources. The process of text gathering in biomedical literature involves Pub Med Open Access Initiative [5, 7], Medline, National Library of Medicine, MeSH databases, MDB etc. All the above databases contain more than 12,000,000 references of biomedical publications. Most of the information's in Google and journals are PDF format. To eliminate this hazard we have to convert PDF form into html form using some converters. Even we collect all the information regarding to the domain most of the information's are unwanted. From the collected information we have to predict which data be the best according to the user query for that we go for information processing step to eliminate the unwanted information's.

2.2 Information Processing:

From the collected information we have to fetch the optimized result with user request. For that we are handling the following steps:

- Generate the unique words
- Punctuation removing
- Preposition removing
- Root identification
- Identification of most interesting terms

Instead of handling the plain text we have to form or generate the unique words to easy process handling like a token generation in compiler design. Every word, marks, dots, etc. generated as separate unique words. Generating unique words is very useful things in further steps to classify the information as well as processing.

Punctuation removing, the name itself expose the meaning. All punctuation marks will eliminate in this step. Case sensitive problems will eliminated in this step. We are having our own algorithm for implement this step.
Preposition removing, in this step we are eliminating the preposition words to find out the sentence and optimize the results. For example on, at, too, to, from, above, below, far, many, as, the , a, an, off, etc.

Root word identification, in this we have to identify the root words for the user query. In this paper we are using porter algorithm to find out the root word identification.
The most common algorithm for stemming English, and one that has repeatedly been shown to be empirically very effective, is *Porter's algorithm* .The entire algorithm is too long and intricate to present here, but we will indicate its general nature. Porter's algorithm consists of 5 phases of word reductions, applied sequentially. Within each phase there are various conventions to select rules, such as selecting the rule from each rule group that applies to the longest suffix. In the first phase, this convention is used with the following rule group:

SSE->SS
IES->I
SS->SS
S->

Example:
Caresses->caress
ponies->poni
caress->caress
cats->cat

Sample Text:
Such an Analysis can reveal features that are not easily visible from the variations in the individual genes and can lead to a picture of expression that is more biologically transparent and accessible to interpretation,

Porter stemmer:
Such an analysi can reveal feature that ar not easily visibl from the vaiat in the individu gene and can lead to a picture of express that is more biolog transpar and acces to interpret.

Rather than using a stemmer, you can use a lemmatize , a tool from Natural Language Processing which does full morphological analysis to accurately identify the lemma for each word. Doing full morphological analysis produces at most very modest benefits for retrieval. It is hard to say more, because either form of normalization tends not to improve English information retrieval performance in aggregate - at least not by very much. While it helps a lot for some queries, it equally hurts performance a lot for others. Stemming increases recall while harming precision. As an example of what can go wrong, note that the Porter stemmer stems all of the following words:

operate operating operates operation operative operatives operational to oper.
However, since operate in its various forms is a common verb, we would expect to lose considerable precision on queries such as the following with Porter stemming:

operational and research
operating and system
operative and dentistry

2.3 Information analysis:

After getting the processed information .we has to ordered the processed information that is very important compare with other things. Exact matched result will come first and further deviations come under that. To achieve this we used the hierarchical clustering.
In statistics, hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types:

- **Agglomerative**: This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
- **Divisive**: This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

In general, the merges and splits are determined in a greedy manner. The results of hierarchical clustering are usually presented in a dendrogram. For our implementation we are using Divisive.
In order to decide which clusters should be combined (for agglomerative), or where a cluster should be split (for divisive), a measure of dissimilarity between sets of observations is required. In most methods of hierarchical clustering, this is achieved by use of an appropriate metric (a measure of distance between pairs of observations), and a linkage criteria which specifies the dissimilarity of sets as a function of the pairwise distances of observations in the sets.

The choice of an appropriate metric will influence the shape of the clusters, as some elements may be close to one another according to one distance and farther away according to another. For example, in a 2-dimensional space, the distance between the point (1,0) and the origin (0,0) is always 1 according to the usual norms, but the distance between the point (1,1) and the origin (0,0) can be 2, $\sqrt{2}$or 1 under Manhattan distance, Euclidean distance or maximum distance respectively.
In this paper we are using Euclidean distance as a distance vector. Distances will calculate using this formula

$$\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$$

3. RESULT ANALYSIS:

Fig1 shows how the hierarchical clustering evaluation will made.

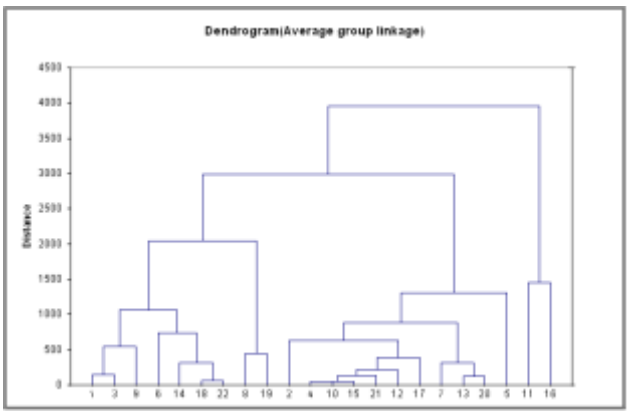


Fig1

Proposed methods Result:

For Example take user query as:

Symptoms of Bovine leukemia

1) Initial cluster formed like this:

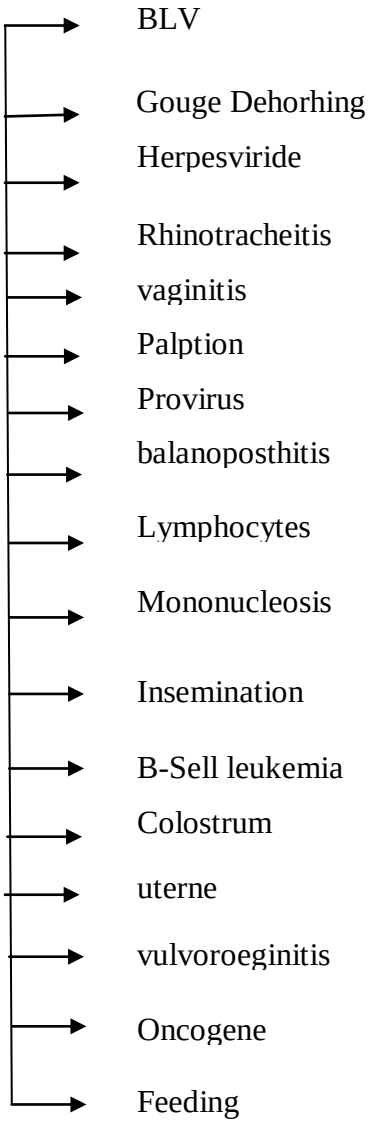


Fig: 2

2) After applying porter and Hierarchical clustering middle phase:

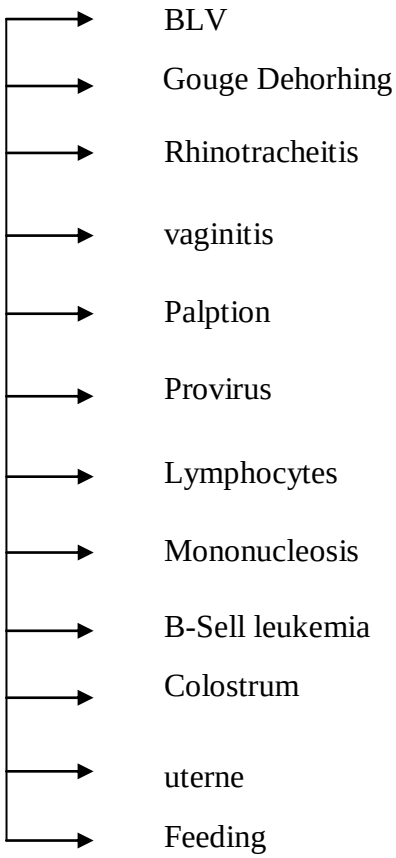
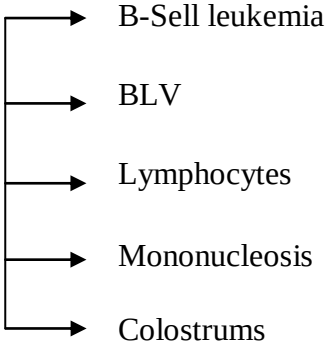


Fig: 3

3) Final Output of Query:



4. FUTURE WORK

In this paper we used porter stemming and hierarchical clustering algorithm to identifying and grouping. In feature we are going to try lovins, paice/husk algorithms with different clustering mechanism to introduce the effective text mining scheme. Comparison analysis will made in feature with different stemming schemes due to that we can derive effective mining scheme.

5. CONCLUSION

Information extraction is the one important scheme in text mining. Using IE we derived an effective algorithm for bovine diseases. Porter algorithm gives an effective root node identification solution. Hierarchical clustering used to form an effective solution group. This work will really help to the medicine sectors.

6. REFERENCES

[1]Jung-Hsien Chiang,IEEE member and Hsu-Chun YU"Literature extraction of protein functions using sentence pattern mining " IEEE Transactions on Knowledge and Data Engineering, vol 17, no 8, August 2005.

[2] Ananiadou, Sophia, Goran Nenadic, Dietrich Schuhmann, and Irena Spasic, 2002. Term-Based Literature Mining from Biomedical Texts", in *Proceedings*, Intelligent Systems for Molecular Biology, ISMB, Text Data Mining SIG, Edmondton, Canada.

[4] D. M. Bikel, R. Schwartz, and R. M. Weischedel. An algorithm that learns what's in a name. Machine Learning, 34:211–232, 1999.

[5]PubMedCentralOpenAccessInitiative,http://www.pubmedcentral.nih.gov/about/openftplist.html

[7]PubMed, http://www.ncbi.nlm.nih.gov/PubMed/, 2004