ISSN:0975-9646

An Enhancing approach in MEDLINE & PubMed using Text Mining

R.Santhanalakshmi,Dr.K.Alagarsamy,
Dept of MCA, ComputerCenter, Madurai Kamaraj University,Madurai.

Abstract:
Text mining is a new and exciting research area that attempts to solve the information overload problem. Information extraction using text mining has been applied to many different areas of bio-medicine such as clinical diagnosis and protein structure prediction. This is the first proposal to utilize this technique for Bovine disease diagnosis. MEDLINE & PubMed are the famous medical databases. In this paper we applied searching mechanism to find out Bovine diseases symptoms and suggestions using modified hits algorithm in those databases.

Keywords: Bovine Diseases, HITS Algorithm, MEDLINE, PubMed.

1. Introduction:

1.1 Bovine Diseases:

Bovine herpes virus 1 (BHV-1) is a virus of the family Herpesviridae that causes several diseases worldwide in cattle, including rhinotracheitis, vaginitis, balanoposthitis, abortion, conjunctivitis and enteritis. BHV-1 is also a contributing factor in shipping fever. It is spread through sexual contact, artificial insemination, and aerosol transmission. Like other herpesviruses, BHV-1 causes a lifelong latent infection and shedding of the virus. The sciatic nerve and trigeminal nerve are the sites of latency. There is a vaccine available which reduces the severity and incidence of disease.

The respiratory disease caused by BHV-1 is commonly known as infectious bovine rhinotracheitis. Symptoms include fever, discharge from the nose, cough, difficulty breathing, and loss of appetite. Ulcers commonly occur in the mouth and nose. Mortality may reach 10 percent. The genital disease causes infectious pustular vulvovaginitis in cows and infectious balanoposthitis in bulls. Symptoms include fever, depressions, loss of appetite, painful urination, a swollen vulva with pustules and discharge in cows, and pain on sexual contact in bulls. In both cases lesions usually resolve within two weeks. Abortion and still births are occurring one to three months post infection. BHV-1 also causes a generalized disease in newborn calves, characterized by enteritis and death.

Bovine malignant catarrhal fever (BMCF) is a fatal lymphoproliferative disease caused by a group of ruminant gamma herpes viruses including Alcelaphine Herpes Virus 1 (AlHV-1) and Ovine Herpes Virus 2 (OHV-2) These viruses cause in apparent infection in their reservoir hosts, (sheep with OHV-2 and wildebeest with AHV-1) but are usually fatal in cattle and other ungulates such as deer, antelope, and buffalo. BMCF is an important disease where reservoir and susceptible animals mix. There is a particular problem with Bali cattle in Indonesia, bison in the USA and in pastoralist herds in Eastern and Southern Africa.

Disease outbreaks in cattle are usually sporadic although infection of up to 40% of a herd has been reported. The reasons for this are unknown. Some species appear to be particularly susceptible, for example Peer David's deer, Bali cattle and bison, with many deer dying within 48 hours of the appearance of the first symptoms and bison within three days. In contrast, post infection cattle will usually survive a week or more.

Symptoms:

The most common form of the disease is the head and eye form. Typical symptoms of this

form include fever, depression, discharge from the eyes and nose, lesions of the buccal cavity and muzzle, swelling of the lymph nodes, opacity of the corneas leading to blindness, in appetence and diarrhea. Some animals have neurological signs, such as ataxia, nystagmus, and head pressing. Per acute, alimentary and cutaneous clinical disease patterns have also been described. Death usually occurs within ten days. The mortality rate in symptomatic animals is 90 to 100 percent. Treatment is supportive only.

1.2 MEDLINE

MEDLINE is the U.S. National Library of Medicine's (NLM) premier bibliographic database that contains over 18 million references to journal articles in life sciences with a concentration on biomedicine. A distinctive feature of MEDLINE is that the records are indexed with NLM's. The great majority of journals are selected for MEDLINE based on the recommendation of the Literature Selection Technical Review Committee (LSTRC), an NIH-chartered advisory committee of external experts analogous to the committees that review NIH grant applications. Some additional journals and newsletters are selected based on NLM-initiated reviews, e.g., history of medicine, health services research, AIDS, toxicology and environmental health, molecular biology, and complementary medicine, that are special priorities for NLM or other NIH components. These reviews generally also involve consultation with an array of NIH and outside experts or, in some cases, external organizations with which NLM has special collaborative arrangements.

MEDLINE is the primary component of PubMed, part of the Entrez series of databases provided by NLM's National Center for Biotechnology Information (NCBI). MEDLINE may also be searched via the NLM Gateway.

1.3 Pub med:

MEDLINE is the largest component of PubMed, the freely accessible online database of biomedical journal citations and abstracts created by the U.S. National Library of Medicine. Approximately 5,400 journals published in the United States and more than 80 other countries have been selected and are currently indexed for MEDLINE. A distinctive feature of MEDLINE is that the records are indexed with NLM's controlled vocabulary, the Medical Subject Headings (MeSH).

In addition to MEDLINE citations, PubMed also contains:

- In-process citations which provide a record for an article before it is indexed with MeSH and added to MEDLINE or converted to out-of-scope status.
- Citations that proceed the date that a journal was selected for MEDLINE indexing (when supplied electronically by the publisher).
- Some OLDMEDLINE citations that have not yet been updated with current vocabulary and converted to MEDLINE status.
- Citations to articles that are out-of-scope (e.g., covering plate tectonics or astrophysics) from certain MEDLINE journals, primarily general science and general chemistry journals, for which the life sciences articles are indexed with MeSH for MEDLINE.
- Citations to some additional life science journals that submit full text to PubMed Central and receive a qualitative review by NLM.
- Citations to author manuscripts of articles published by NIH-funded researchers. m
- Citations for a subset of books available on the NCBI Bookshelf (a citation for both the book and each chapter or section of the book).

One of the ways users can limit their retrieval to MEDLINE citations in PubMed is by selecting MEDLINE from the Subsets menu on the Limits screen.

Other PubMed services include:

- Links to many sites providing full text articles and other related resources
- Clinical queries and Special queries search filters
- Links to other citations or information, such as those to related articles
- Single citation matcher
- The ability to store collections of citations, and save and automatically update searches
- A spell checker
- Filters to group search results

NLM distributes all but approximately 2% of all citations in PubMed to those who formally lease MEDLINE from NLM [6].

2. Proposed method:

One of the interesting points that he brought up was that the human perspective on how a search process should go is more complex than just compare a list of query words against a list of documents and return the matches. Suppose we want to buy a car and type in a general query phrase like "the best automobile makers in the last 4 years", perhaps with the intention to get back a list of top car brands and their official web sites.

When you ask this question to a computer that is running a text based ranking algorithm [1], things might be very different. That computer will count all occurrences of the given words in a given set of documents, but will not do intelligent rephrasing for you. The list of top pages we get back, while algorithmically correct, might be very different than what expected. One problem is that most official web sites are not enough self descriptive. They might not advertise themselves the way general public perceives them.

It would be of course great if computers could have a dictionary or ontology, such that for any query, they could figure out synonyms, equivalent meanings of phrases. This might improve the quality of search, nevertheless, in the end; we would still have a text based ranking system for the web pages. We would still be left with the initial problem of sorting the huge number of pages that are relevant to the different meanings of the query phrase. We can easily convince ourselves that this is the case.

Even if trying to find pages that contain the query words should be the starting point, a different ranking system is needed in order to find those pages that are authoritative for a given query

Jon Kleinberg's algorithm [2] called HITS identify good authorities and hubs for a topic by assigning two numbers to a page: an authority and a hub weight. These weights are defined recursively. A higher authority weight occurs if the page is pointed to by pages with high hub weights. A higher hub weight occurs if the page points to many pages with high authority weights. In this paper we are trying to enhance the quality of the hits algorithm adding some extra parameters while searching. Hits algorithm developed for searching the content in internet in effective manner.

2.1 Modified Hits Algorithm:

Here we are going to search in PubMed & MEDLINE database. It wills the optimized one thing to search and also we suggest some thing to enhance the quality of databases because those mainly focus the human diseases not veterinary. MEDLINE & PubMed are the authorized sites by government, which provides only true information. Getting the right information is very important than the getting information. The optimized results should be the valid information. HITS algorithm mainly based on the concept of Hub & Authority. This mainly algorithm designed for web content searching but here we are minimize the searching content in specific authorized database only. Due to that we need some modification in that. First of all those databases contain the medicine journals information’s.

We have to add some more information regarding to the veterinary. Both of them contain the many research articles from the various domains. Additionally we have to add the information for every disease. If we give any query or ask any questions the result should show the symptoms of diseases. Many sites are providing the wrong information.

HITS algorithm mainly focusing hub and authority values but in our modification gives flexibility of searching. First one is ranking based web pages. How to provide the ranking we have to maintain the separate hash table for alphabets that means hash key table contains A-Z values then this key value redirect the content of corresponding letter. Whenever user type query in search box the corresponding matches will fetch from the database. While fetching us can use follow the hits hub and authority methods. In that whenever hub choosing we have to add the counting value for that hub into the hash table. According to the counting value we can allocate the ranking to the corresponding web pages and also categorization achieved by hash table itself due to that searching make it easy and effective. Authority makes the keywords present in the corresponding web pages. Keywords presented authority & frequent used web pages will be fetched for every user queries.

3. Result Analysis:

Implementation has done in C using data structures because implementation of linked list will more comfortable in c with data structure. Table 1 shows sample output.

4. Conclusion:

In this paper we provide an efficient search mechanism in MEDLINE & PubMed. Those databases are government authorized sites to get information regarding to the medicine. We tried to improve the quality of database and enhance the searching mechanism to get the fast and right information retrieval using modified hits algorithm. Hash table enhance the searching speed based on the pointer concept. It is our belief this paper will help to enhance the quality of MEDLINE & PubMed.

Table 1

Hash Key	Hub	Access counting
A	ataxia	3
B	Bovine leukaemia virus	134
C	cattle leukaemia	567
D	DNA viruses	21
H	Herpes Virus 2	897
l	leukaemia	117
p	Poxviridae	343
Y	Young	23
Z	-	

5. References:

[1] J. Wang, Z. Chen, L. Tao, W. Ma, and W. Liu. Ranking user's relevance to a topic through link analysis on web logs.
WIDM, pages 49–54, 2002.
[2] J. Kleinberg. Hubs, Authorities, and Communities. ACM Computing Surveys, 31(4es, Article No.5), 1999.
[3]S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, and .Rajagopalan. Automatic resource compilation by analyzing hyperlink structure and associated text. In *Proc.* 7th International World Wide Web Conference, Brisbane,Australia, 1998.
[4]http://www.cs.helsinki.fi/u/goethals/software/index.html#apriori.
[5] http://www.nlm .nih.gov.
[6]http://www.nlm.nih.gov/databases/databases_medline.html.