# SIMILAR HERB SELECTION USING DATA MINING

A.Linda Sherin , PG Scholar,
Dr.K.Kumar, Assistant Professor,
Dept. of Computer Science and Engineering,
Government College of Technology,
Coimbatore.
linz15sherin@gmail.com

***Abstract***

***Ethno-pharmacological relevance of Natural products has long been the most important source of ingredients in the discovery of new drugs. Moreover, since the Nagoya Protocol, finding alternative herbs with similar efficacy in traditional medicine has become a very important issue as it proved to be less effective; therefore, this project proposes a novel targeted selection method using data mining approaches in the MEDLINE database to identify and select herbs with a similar degree of efficacy. Phytochemicals are non-nutritive plant chemicals that have protective or disease preventive properties. It is well-known that plant produce these chemicals to protect themselves but recent research demonstrate that they can also protect humans against diseases. The main objective of this project is to provide the performance evaluation of the phytochemicals present in herbs thus, by collecting articles from the MEDLINE database. These articles are analysed and the required phytochemical structures are obtained. Finally the candidate herbs are found out which match the phytochemicals present in the target herb.***

***Keywords – Herbs, Efficacy, MeSH, Data mining, targeted selection, k-means clustering, MEDLINE., Pharmacology***

## I.INTRODUCTION

Data mining is knowledge discovered from data. The data mining processes include expressing a term, collecting data, performing preprocessing, estimating the model, and clarifying the model and draw the conclusions. It is the process of analyzing and summarizing data from different perspectives and converting it into useful information. Thus Data mining Data mining has been defined as the nontrivial extraction of implicit, previously unknown, and potentially useful information from databases/data warehouses. It uses machine learning, statistical and visualization techniques to discover and present knowledge in a form, which is easily comprehensive to humans. Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help user focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviours, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. With the obtained phytochemicals the best fit alternative herb is found out with the help of data mining clustering algorithms. Here, the list of candidate herbs competing with the target herb is found out by comparing the phytochemicals present in the target herb and the others present in our database. As a result a list of candidate herbs with

similar phytochemical structures are obtained. Using the clustering algorithms several clusters are formed thereby analyzing the structure of the phytochemicals. This produces the best fit herb with similar efficacy among the candidate herbs.

## II. RELATED WORKS

BrendanCoolsaet, Tom Dedeurwaerdere and John Pitseys [1] The Nagoya Protocol on Access and Benefit sharing, or ABS, is the latestprotocol to the Convention on Biological Diversity (CBD). Its origin goes back to the early days of the Convention as its core objective is to further the implementation of the third objective of the CBD, namely the fair and equitable sharing of benefits arising from the utilization of genetic resources. The questions underlying this survey are therefore the following: what are the consequences of the multi-level governance character of the Nagoya Protocol for its implementation in Belgium? How does the multi-level character of the implementation impact the choice between a self-regulatory approach and a normative institution a list one? In addition, to what extent does the multi-level character restrain the political opportunity to move beyond a mere self-regulatory approach to implementation? These are answered in this research as First, the internationalization and institutional interplay with the rules of the global economy might generate strong pressure to adopt a minimalist implementation approach close the self-regulatory implementation process. Second, countries are increasingly confronted by powerful corporate interests (denationalization). Third, the global reinforcement of the role private stakeholders (i.e., the destatization) will exacerbate the unequal distribution of power and resources necessary to conclude fair and equitable agreements.

Sang-Jun Yea, Chul Kim, IckTae Kim, BoSeok Sung [2] Ethno-pharmacological relevance: Natural products have long been the most important source of ingredients in the discovery of new drugs. Moreover, since the Nagoya Protocol, finding alternative herbs with similar efficacy in traditional medicine has become a very important issue. Although random selection is a common method of finding ethno-medicinal herbs of similar efficacy, it proved to be less effective; therefore, this paper proposes a novel targeted selection method using data mining approaches in the MEDLINE database in order to identify and select herbs with a similar degree of efficacy. From among sixteen categories of medical subject headings (MeSH) descriptors, three categories containing terms related to herbal compounds, efficacy, toxicity, and the metabolic process were selected. In order to select herbs of similar efficacy in a targeted way, we adopted the similarity measurement method based on MeSH. In order to evaluate the proposed algorithm, we built up three different validation datasets which contain lists of original herbs and corresponding medicinal herbs of similar efficacy.

Rey G. Tantiado [3] This study aimed to categorize the diversity of medicinal plants in Tigbauan, Iloilo, Philippines based on their taxonomic rank and document the traditional uses, preparations and applications of medicinal plants (ethnopharmacology) by the local community and determine the distribution, morphological forms, habitat and values of indigenous medicinal plant resources in Tigbauan, Iloilo; and lastly identify and enumerate the medicinal uses of each identified indigenous plants. Ethno-pharmacological and taxonomic data of indigenous medicinal plants were collected in the study site through semi-structured interview and snowball sampling methods among knowledgeable elders, gardeners, healers, and traders. The taxonomic classification of the indigenous medicinal

plants in Tigbauan, Iloilo was based on Cronquist's System of classification. A total of 101 species, grouped within 92 genera, 44 families and 27 orders. The medicinal plants were described according to preparation techniques, mode of application, administration route, growth forms, habitat distribution, abundance and medicinal uses. Results showed a diversity of medicinal plants, traditional and ethno-pharmacological knowledge about the uses, preparations and applications present and maintained among the Tigbauenos. This study allowed the identification of many high value and high priority medicinal plant species, indicating high potential for economic development through sustainable collection and trade.

Kawo, A.H., Mustapha, A., Abdullahi, B.A., Rogo, L.D., Gaiya, Z.A., Kumurya, A.S. [4] A comparative preliminary study on the phytochemistry and antibacterial effects of ethanol and aqueous extracts of the leaves and latex of Calotropis procera on four pathogenic clinical bacterial isolates namely Escherichia coli, Staphylococcu saureus, Salmonella species and Pseudomonas species was carried out using paper-disc diffusion and broth dilution techniques. The results obtained revealed that ethanol was the best extractive solvent for a fraction with antibacterial properties of the C. procera leaves and latex. Gene rally, the aqueous extracts showed no activity on the isolates. Generally, the antibacterial effects of the plant parts revealed that theleaf extracts had stronger activity in comparison with those of the latex.

Jyoti Yadav, Monika Sharma [5] Cluster analysis is a descriptive task that seek to identify homogenous group of object and it is also one of the main analytical method in data mining. K-mean is the most popular partitional clustering method. In this paper, we discuss standard k-mean algorithm and analyze the shortcoming of k-mean algorithm. In this work three dissimilar modified k-mean algorithms are discussed which remove the limitation of k-mean algorithm and improve the speed and efficiency of k-mean algorithm. First algorithm removes the requirement of specifying the value of k in advance practically which is very difficult. This algorithm results in optimal number of 14 cluster. Second algorithm reduce computational complexity and remove dead unit problem. It selects the most populated area as cluster center. Third algorithm use simple data structure that can be used to store information in each iteration and that information can be used in next iteration. It increases the speed of clustering and reduce time complexity.

The basic idea of this algorithm is two keep two simple data structure to store the information of each iteration and the information can be used in next iteration. First data structure can be used to store the label of cluster. Second data structure can be used to store the distance of each data item to the nearest cluster center in each iteration this information can be used in next iteration. In second iteration we calculate the distance between data item and the new cluster center. After that we compare the distance between data item and new cluster with distance stored in previous iteration. If the new distance is smaller than or equal to older center then data item stay in its cluster that was assigned in previous iteration. Now there is no need to calculate distance between data item and remaining k-1cluster center. Some data item will remain in original cluster in each iteration so there is no need to calculate the distance and it will reduce the computational complexity.

Samir Kumar Sarangi and Dr. Vivek Jaglan [6] The abundance of data in business, research, industry, science and in many fields makes it very difficult to handle them. It is complicated to explore any valuable information, needed to take any important decision, but problem is how to discover this precious information. The effective solution may be data mining, which is a

very popular topic at present research. Two main techniques of data mining are clustering and classification, which are basically studied as individual approach till now. In this paper we integrated both (clustering and classification) techniques. After combine application of most frequently used clustering (k-means) algorithm with classification (J48, Multilayer Perceptron, BayesNet, NavieBayes) algorithms, the results were compared and the WEKA data mining tool was used. Here four different classifiers are integrated with the simple k-means clustering algorithm and this integration technique were applied on "Diabetes Diagnosis" data set. From our observation and analysis it was concluded that the integration of K-means (clustering) + J48 (classification) have zero MAE and RMSE error and it also takes less time to build the model. So the performance of K-means +J48 is better than other algorithms. There are large numbers of classifiers present and many other data mining tools are present. So the future work will be based on other classifiers that can be applied on the data set and also to apply other data mining tools on the data set such that the best techniques can be identified.

## III.EXISTING SYSTEM

Most of the existing work is focused only on the Nagoya protocol for the collection of various herbs structure. Also this protocol had various drawbacks for implementation. The Nagoya Protocol on Access and Benefit sharing, or ABS, is the latest protocol to the Convention on Biological Diversity (CBD). Its origin goes back to the early days of the Convention as its core objective is to further the implementation of the third objective of the CBD, namely the fair and equitable sharing of benefits arising from the utilization of genetic resources. The questions underlying this survey are therefore the following: what are the consequences of the multi-level governance character of the Nagoya Protocol for its implementation in Belgium? How does the multi-level character of the implementation impact the choice between a self-regulatory approach and a normative institution a list one? In addition, to what extent does the multi-level character restrain the political opportunity to move beyond a mere self-regulatory approach to implementation? These are answered in this research as First, the internationalization and institutional interplay with the rules of the global economy might generate strong pressure to adopt a minimalist implementation approach close the self-regulatory implementation process. Second, countries are increasingly confronted by powerful corporate interests (denationalization). Third, the global reinforcement of the role private stakeholders (i.e., the destatization) will exacerbate the unequal distribution of power and resources necessary to conclude fair and equitable agreements.

## IV. PROPOSED SYSTEM

Many valuable drugs of today came into use through the study of indigenous remedies. Chemists continue to use plant-derived drugs as prototypes in their attempts to develop more effective and less toxic medicinal. In order to select herbs with similar efficacy in targeted ways, we propose an algorithm of similarity based on MeSH (SoM), which exploits data mining methods in the MEDLINE database. The data that is obtained from the MEDLINE database is sorted and is stored in a cloud which allows the access to all the users over the globe, thus by overcoming the data access problem from the previous system. The proposed model improves the accuracy in finding out the alternative herbs when compared to that of the existing model.

## A. DATA COLLECTION AND PROCESSING

The basic input for this proposed work is the scientific name or botanical name and the common name of the various herbs and plants. Fetching out articles in MeSH are done only with the help of the botanical or scientific names of the herb rather than the common name of the particular herb. Deriving data from these articles feed the project work with required stuffs.

## B. SOM ALGORITHM

In order to select herbs with similar efficacy in targeted ways, we propose an algorithm of Similarity based on MeSH (SoM), which exploits data mining met hods in the MEDLINE database. The SoM algorithm comprises four steps: First, the scientific names of the original plants designated as target herbs and candidate herbs are extracted from the reference database, which contains data of medicinal herbs and the corresponding original plants. Second, the extracted scientific names are used to search articles in the MEDLINE, and the MeSH terms of searched articles forms the MeSH vector for each herb. Third, the MeSH thesaurus and similarity calculation, are adopted to draw similarity scores between the target herb and each candidate herb using the MeSH vectors built in the previous step

## C. REFERENCE DATABASE

The initial step of this project work deals with the obtaining of scientific name (botanical name) for the various medicinal plants herbs. These data are obtained from the open source data available from the web, from Agriculture University and from Botanical Survey of India (BOI).

## D. SIMILARITY MEASUREMENT

After the required phytochemicals are retrieved from the MEDLINE database, the calculation of the similarity scores are done for the target herb (original herb).And the corresponding candidate herbs (herb which compete for similar efficacy). These are done by using k-means algorithm which eventually clusters the similar phytochemicals of the herbs. The basic step of k-means clustering is simple. In the beginning, we determine number of cluster k and we assume the centroid or center of these clusters.

# V IMPLEMENTATION

## A.DATA COLLECTION

Similar herb extraction deals a lot with the botanical name of a particular plant. Therefore the basic input for this proposed work is the scientific name or botanical name and the common name of the various herbs and plants. Obtaining these names are important as we are dealing with the Medical Subject Heading (MeSH), fetching out articles are done only with the help of the botanical or scientific names of the herb rather than the common name of the particular herb. Here a database is created, which holds the common name and the botanical name of a plant.


**Figure 5.1 Reference Database**

## B.ARTICLE SELECTION

The next step deals in selecting articles from the Medline database with the scientific name of the plant. Deriving data from these articles feed the project work with required data based on the similar

efficacy. The MeSH Browser allows users to search directly for MeSH terms, and conduct text-word searches of the Annotation, and Scope Note fields of records. The Registry Number (RN) and Relative Registry Number (RR) can be also be searched to find Chemical headings. To search the MeSH Browser, locate a vocabulary term using any word in an expression or using the complete expression.

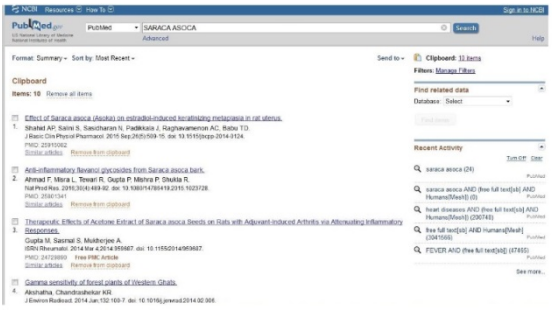

**Figure 5.2. Article Selection**

## CONCLUSION

In this project, the identification of the scientific name of the various herbs was done from vivid sources. These data are stored in a reference database. Using these scientific names, the articles which hold the phytochemicals were located in the MeSH database. These articles are stored in the cloud which helps in easy retrieval. Using these phytochemicals the candidate herbs of a particular target herb is found out. Moreover the phytochemical structures present in the herbs are also screened.

## REFERENCES

**[1]** Brendan Coolsaet, Tom Dedeurwaerdere and John Pitseys, "The Challenges for Implementing the Nagoya Protocol in a Multi-Level Governance Context: Lessons from the Belgian Case", ISSN 2079-9276, Resources 2013, 2, 555-580; doi: 10.3390/resources2040555

**[2]** Sang-Jun Yea, Chul Kim, IckTae Kim, BoSeok Sung,"Picking out herbs with analogous efficacy based on MeSH semantic similarity",2014 IEEEInternational Conference on Bioinformatics and Biomedicine.

**[3]** Rey G. Tantiado , "Survey on Ethnopharmacology of Medicinal Plants in Iloilo, Philippines.", International Journal of Bio-Science and Bio-Technology Vol. 4, No. 4, December, 2012.

**[4]** Y. U. Dabai, A. H. Kawo and R. M. Aliyu, "Phytochemical screening andantibacterial activity of the leaf and root extracts of Senna italica", African Journal of Pharmacy and Pharmacology Vol. 6(12), pp. 914-918, 29 March, 2012.

**[5]** Jyoti Yadav, Monika Sharma, "A Review of K-mean Algorithm.", International Journal of Engineering Trends and Technology (IJETT) – Volume 4 Issue 7- July 2013.

**[6]** Samir Kumar Sarangi and Dr. Vivek Jaglan, "Performance Comparison of Machine Learning Algorithms on Integration of Clustering and Classification Techniques", International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS) 2013.