ITC 3/53

Information Technology and Control
Vol. 53 / No. 3/ 2024
pp. 758-771
DOI 10.5755/j01.itc.53.3.35101

Detecting the Medical Plant Association from PubMed Using Hypergraph-based Clustering with Dominating Set

Received 2023/09/13 | Accepted after revision 2024/07/12

HOW TO CITE: Sampath, P., Jomy, E., Kalyanaraman, R., Shanmuganathan, V., Gonzalez Crespo, R., Chakrabarti, P. (2024). Detecting the Medical Plant Association from PubMed Using Hypergraph-based Clustering with Dominating Set. *Information Technology and Control*, 53(3), 758-771. https://doi.org/10.5755/j01.itc.53.3.35101

Detecting the Medical Plant Association from PubMed Using Hypergraph-based Clustering with Dominating Set

Pradeepa Sampath, Elizabeth Jomy, Ramya Kalyanaraman

SASTRA Deemed University, Thanjavur, Tamilnadu, India; emails: pradeepa.pradee@gmail.com, elizabethjomy.mk@gmail.com, ramyakal21@gmail.com

Vimal Shanmuganathan

Department of Artificial Intelligence and Data Science, Sri Eshwar college of Engineering, Kinathukadavu, Coimbatore, Tamilnadu, India; email: svimalphd@gmail.com

Ruben Gonzalez Crespo

Department of Computer Science and Technology, Universidad Internacional de La Rioja, Logroño, La Rioja, Spain; email: ruben.gonzalez@unir.netm

Prasun Chakrabarti

Department of Computer Science and Engineering, Sir Padampat Singhania University, Udaipur 313601, Rajasthan, India; email: drprasun.cse@gmail.com

Corresponding author: ruben.gonzalez@unir.net

Medicinal plants provide immunity against diseases and can also be taken in a precautionary sense against them. It is pivotal to know the benefits of these plants against various ailments. The identification of these plants' essential properties can give a great impact on medicinal research and practice. This research focuses on identifying the cardinal properties of five plants namely- Aloe Vera, Fennel, Fenugreek, Mint, and Tulsi by using the concept of text analytic features and NLP functions. Text data on medicinal plants are extracted from the biomedical literature dataset. Text mining is used for the extraction of the implicit relations between medicinal plants and their biomedical properties. The intricate relationship between the keywords and the medicinal plants is captured using hypergraph clustering and dominating sets. The visualization of the correlation between the keywords and the plants is carried out by clustering. With an emphasis on their potential in preventative and medical care, this model lists the common characteristics and health advantages of medicinal plants. Strong clustering is indicated by the modularity score of 0.577, with five separate communities each reflecting a unique set of features. In order to facilitate future studies, these findings offer a methodical and data-driven viewpoint.

KEYWORDS: Deep learning, Text mining, Medical Plants, Text Datasets, Apriori algorithms, Hypergraph clustering, data visualization.

1. Introduction

Traditionally, plants with medical properties have been used to treat various human ailments. While there has been significant progress in allopathic medicine, treatment using medical plants still prevails [4], [8]. Medicinal plants are an important natural resource, providing natural therapists and raw materials for the production of traditional and modern medications [31]. The World Health Organization (WHO) estimates that about 70–80% of the world's population relies on nonconventional medicines in their healthcare [22]. Countries like India place a considerably greater value on medicinal plants economically than the rest of the globe [26]. Extensive research work on ethnomedicine has already been performed to identify indigenous medical plants and study their uses [33]. Allopathic medicine is also shown to be plant dependent and about 20-25% of drugs are plant reliant [32]. Plants synthesize biochemical molecules in their barks, fruits, seeds and other parts that are used for treatment [2], [23].

Extensive biomedical research has made the extraction of information a cumbersome process [6], [35], [38]. The fragmented traditional knowledge, inconsistent data quality, intricate keyword-property connections, and the divide between traditional and biomedical knowledge all provide challenges to medicinal plant analysis. Literature-based discovery is utilized to link biomedical terms with medicinal plants [38]. Knowledge and statistical-based methodologies are used for mining the potential benefits of medicinal plants from the repository of biomedical research data, PUBMED [17]. Text mining is utilized to access the biomedical knowledge of the phytochemical properties of medicinal plants. Text mining is used to automate the process of text extraction [35].

The study aims to represent the correlation between the medicinal plant's keywords present in PUBMED abstracts to the relevant medicinal plant. This process is carried out by utilizing text mining for the extraction of the PUBMED abstracts focusing on five plants, namely, Tulsi, Aloe Vera, Mint, Fennel, and Fenugreek. The connecting link between the keywords and the medicinal plants is visualized in the form of a network analysis.

2. Related Works

The study of medicinal plants and their therapeutic uses has attracted more attention in recent years [30]. According to their similarities, characteristics, and possible therapeutic uses, plant species are frequently grouped and categorized in this subject. Clustering approaches allow plants to be organized and classified into meaningful groups, allowing for the examination of their shared properties and potential synergistic effects.

Researchers can learn about medicinal plants' historic uses, chemical composition, and pharmacological qualities by grouping them together, which can help them develop evidence-based herbal treatments and alternative treatment choices [12]. A symbolic strategy approach for the classification of plant leaves based on the Modified Local Binary Patterns (MLBP) has been proposed [21]. Clustering is used to select multiple class representatives and to capture intra-cluster variations obtained from interval-valued symbolic features. This approach faces difficulties in species of leaves with higher intra-class variations. A focused investigation on the potential of Arabic Herbal medicine as an alternative medicine has been studied [3]. The study identifies the relationship between obesity and the potential benefits of Arabic herbal medicinal plants. Based on the complicated multipartite network of medicinal plants, multi-chemicals, and many targets, a novel technique to analyze the interactions between the chemicals in medicinal plants and various targets was performed [16]. The chemical compounds found in plants and their biological effects on the targets were combined to create the multipartite network. The target potency score (TPS) was developed to assess the efficacy of plant compounds on a protein target of interest. The analysis can reveal distinct chemical profiles from each plant group, which can then be used to uncover new alternative therapeutic compounds [16]. The resultant multipartite network could be applied to plant and chemical combinations to further investigate the connection between them.

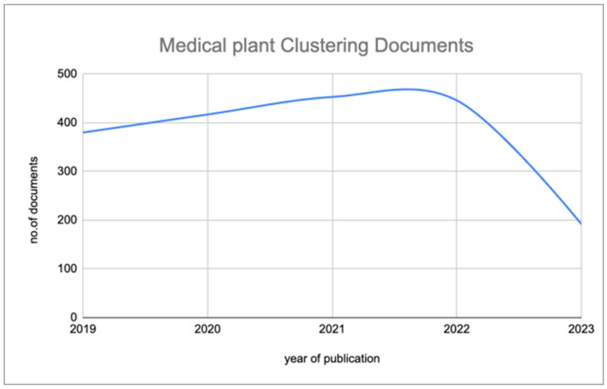
Varieties of machine-learning approach, clustering, for classifying herbal plant species from photographs were described on the method Herbal plant analysis based on leaf features using K-means clustering [20] which concentrated on six Malaysian herbal plants and k-means algorithm was used by testing with different cluster sizes ranging from two to three, four, and five [20]. However, this model required improvement on the feature extraction method. In the model conducted [14], the author suggested an automated system for identifying plants using CNN. Changes in leaf characteristics can be used to undertake plant comparison research. As a result, their automated approach will aid in the identification of medicinal plants and assist agronomists in identifying suitable herbs [14].

A simple pattern-based semi-supervised approach to extract health information about medicinal plants has been carried out using NLP techniques [5]. This study focused on multi-term phrases and complex sentences for a collection of web documents on medicinal plants.

Keyword clustering is essential for knowledge organization and retrieval to explore and navigate the vast information in biomedical databases. State-of-the-art keyword clustering techniques have advanced significantly, but they are limited in their scalability in handling vast biomedical databases. The databases comprise both structured and unstructured data which poses a challenge to the clustering algorithms. There is an absence of ground truth data for evaluation with a benchmarked dataset. With the continuously evolving plant discoveries, the terminology in the biomedical field is updated and the clustering techniques must adapt to these shifts in terminology. Figure 1 shows the statistics on the number of documents on medical plant clustering in PUBMED which is plotted by collecting the number of documents on medical plant clustering published in a given year.

Figure 1

Statistics on no. of documents on medical plant clustering in PUBMED



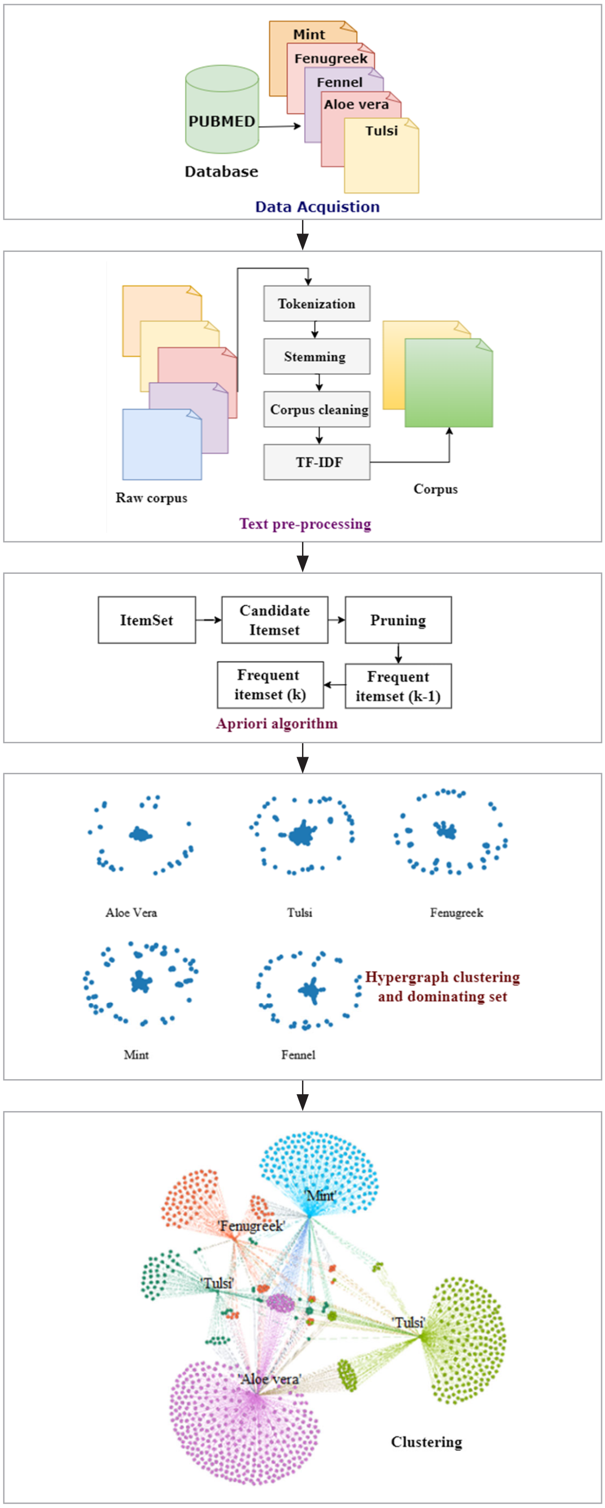
3. Methodology and Description

The proposed model is divided into four sections which consist of a compilation of the research articles in PUBMED that are related to mint, fenugreek, fennel, aloe Vera and tulsi then defining the top words for each of the documents in the dataset which is followed by a visualization of the hypergraph clustering and domination set for the 5 plants and finally performing cluster analysis of the mapping of keywords to medicinal plants.

There are other clustering models and methodologies that are being used for data analysis, but this model uses Hypergraph because of its ability to handle large data types and extract complex knowledge from the PUBMED articles. For handling complex connections present between the keywords, Hypergraph is chosen as it involves multiple higher-order links. Thus, this model enables a thorough evaluation of the health benefits of medical plants, advancing medical research and practice.

Figure 2 shows the architecture diagram of the model. The methodology followed throughout this research and the description is cataloged here. For the keyword

Figure 2
Architecture diagram of the model



extraction and clustering of the properties of medicinal plants, the following processes are carried out:

3.1. Data Acquisition

The dataset is acquired from PUBMED articles and papers through a literature search. The specific medicinal plant is selected as the keyword in the search [17]. The PubMed IDs obtained in the search query are fetched by employing the PubMed fetcher function in the Metapub library [37]. The abstracts of the articles are retrieved corresponding to the PubMed IDs. The obtained abstracts comprise the relevant research papers and articles that map with the medicinal plant as the keyword.

3.2. Text Pre-processing

Text pre-processing is carried out from modules present in the Natural Language Toolkit (NLTK) [34].

The extracted abstracts are unstructured free text that is chaotic and noisy. The following processes are carried out to obtain a clean and consistent text dataset.

3.2.1. Tokenization and Stemming

The text dataset is initially in the form of sentences that are split into tokens in the form of words or terms. The tokenize module present in NLTK is applied to tokenize the dataset. Stemming is the process of extracting the root form of a token after stripping the prefixes and suffixes. A clear and concise dataset is obtained after performing stemming with the stem module in NLTK.

3.2.2. Corpus Cleaning

The dataset contains commonly used words in English that are insignificant to the dataset. The elimination of these words enables focusing on important words that are related to a medicinal plant. The insignificant words grouped under stopwords are filtered and cleaned by using the corpus module of NLTK.

3.2.3. Term Frequency-Inverse Document Frequency Algorithm

To clean up textual material, raw preprocessing employs morphology and syntactic features. Raw feature reduction and numerical feature reduction methods are employed. Because text mining methods use vector space models as input to reduce data to its most significant properties, TF/IDF is applied.

Term Frequency-Inverse Document Frequency (TF-IDF) is a statistical model used for retrieving information that evaluates the significance of a term in a document or corpus of documents [12]. Term Frequency (TF) is the frequency of occurrence of a term in a document and the proportion of documents that contain a term is referred to by Inverse Document Frequency (IDF).

$$TF = \frac{number\ of\ times\ the\ term\ occurs\ in\ the\ document}{total\ number\ of\ terms\ in\ the\ document} \quad (1)$$

$$IDF = \log\left(\frac{number\ of\ documents\ in\ the\ corpus}{number\ of\ documents\ in\ the\ corpus\ with\ the\ term}\right) \quad (2)$$

$$TF - IDF = TF * IDF \quad (3)$$

The TF-IDF score identifies terms that frequently occur in a given document but are less frequent in the corpus as a whole.

Algorithm 1: **TF-IDF**

Purpose: The algorithm determines the relevant words in a specific document among a collection of documents. The keywords pertaining to a document are extracted.

1 Construct a feature set for each document comprising individual words
2 Compute the TF-IDF score for each word.
3 Normalize the TF-TDF by document size using relative frequency for the word-in-document ratio.
4 Construct a TF-IDF matrix that represents the document-term matrix, with rows representing the terms, the columns containing the documents and the values as the TF-IDF scores.

3.3. Apriori Algorithm

The apriori algorithm generates the frequent itemset and determines association rules based on the frequent itemset and confidence measures. It utilizes prior knowledge of frequent itemsets to find the k+1 itemset from the k-frequent itemset through a level-wise search. The apriori property follows that all non-empty subsets of a given frequent itemset are also frequent and the supersets of an infrequent itemset are infrequent. The apriori algorithm upholds the anti-monotonicity of support measure.

Association rule mining comprises (i) finding the frequent itemset in the corpus with minimum support and (ii) constructing further association rules from a frequent itemset for a confidence value [1]. The association rules are used to establish the relationship between the occurrence of data X with the occurrence of data Y [11].

The statement of the association rule mining in a dataset is given as follows: Let $Z = \{i_1, i_2, ..., i_m\}$ be a set containing literals with each term representing an item. The transaction set comprises M transactions T with $T \subseteq Z$. Each transaction has a unique identifier. The transaction T contains X which is a subset of Z given $X \cup Y$. An association rule is constructed such that with $X, Y \subset Z$ and $X \cap Y = \phi$. The formulated association rule holds true in the transaction set M with confidence c given that $c\%$ transactions in M contain both X and Y. The support s in transaction set M withholds if $s\%$ of transactions in M follows $X \cup Y$. All the association rules with support and confidence greater than the minimum support and minimum confidence have to be generated respectively [29].

The apriori algorithm in text mining is used to generate frequent itemsets with increasing lengths, removing infrequent itemsets and generating association rules. The support is a measure of the frequency of occurrence of an itemset in a dataset.

$$\text{support}(X) = \frac{number\ of\ transactions\ containing\ X}{total\ number\ of\ transactions}, \quad (4)$$

where the support(X) is the support value of item X.

Confidence is a measure of the possibility of a given association rule being true. The conditional probability of the consequent is calculated based on the antecedent.

$$confidence(X \Rightarrow Y) = \frac{\text{support}(X \cup Y)}{(\text{support}(X) * \text{support}(Y))}, \quad (5)$$

where X and Y are items in the dataset.

Algorithm 2: **Apriori**

Purpose: This algorithm generates the frequent items in a dataset with association rules generated based on the items. The association rules depict the probability of a relationship between items in the dataset.

Input: *itemset* => a set , *min_confidence* and *min_support* threshold value

Output: *frequent_item* list and *association_rule* list

1 Define *generate_1_frequent_items*

1.1. Parameters: *itemset* and *1*

1.2. Initialise list *frequent_itemset*

1.3. for item -> combination(*itemset*, 1)

1.3.1. append item into *frequent_item* list

1.4. return *frequent_itemset*

2 Define *generate_association_rules*

2.1. parameters: *frequent_itemset* , min confidence

2.2. Initialise list *association_rule*

2.3. for *itemset*->*frequen_itemset*

2.3.1. for i -> range(1, len(*itemset*)); calculate subset as combination(*itemset*, l)

2.3.2. iterate subset to calculate non-empty subset of *frequen_item* list

2.3.3. $$confidence = \frac{support(itemset)}{support(itemset)} \tag{6}$$

2.3.4. if confidence >= *min_confidence* then append it into *association_rule*

3 Define *apriori*

3.1. Initialize k as 1

3.2. Extend *frequent_itemset* to length k

3.3. while *frequent_itemset*

3.3.1. Increment k by 1

3.3.2. *potential_itemset*: set of combination

3.3.3. Initialize empty *pruned_itemset* list

3.3.4. for *itemset* ->*potential_itemset* ; subset ->combination(*itemset*, k-1); Eliminate *potential_itemsets* when it not used mostly in apriori property.

3.3.5. Calculate support value as potential *itemset* by number of transaction containing *itemset*

3.3.6. Extend *frequent_itemset* if support>=*min_support* using list compression

3.4. *association_rules* ->*generate_association_rules*(*frequent_itemset*, *min_confidence*)

3.5. return *frequent_itemset*, *association_rules*

3.4. Hypergraph Clustering

A hypergraph is a graph with the formula $G = (V, E)$, where $V = \{v_1, v_2, ..., v_n\}$ denotes the set of graph nodes including a d-dimensional attribute vector, and $E = \{e_1, e_2, ..., e_n\}$ denotes the set of hyperedges. Graph clustering is the division of a graph into multiple sets of nodes so that similar nodes are grouped under the same cluster [15]. Hypergraph clustering identifies densely connected components in a hypergraph and ensures the edge structure of the graph by populating the edges within a cluster [25]. The weight of each cluster is the sum of the vertex weights [36].

By taking into account higher-order links between data points, hypergraph clustering expands on conventional clustering techniques. Vertices are used to represent data points, while subsets of those points are used to depict higher-order relationships. Based on their connectedness in the hypergraph, the data points are divided into clusters during the procedure. The clustering objective has been optimized using a variety of approaches. Additionally, hypergraph clustering can produce results that are more reliable and precise when dealing with noisy or imperfect data. Hypergraph clustering can reduce the impact of noise and enhance the separation of separate clusters by taking higher-order links into account.

Algorithm 3: Hypergraph clustering

Purpose: This algorithm uses hypergraphs to find patterns and connections in large datasets, which improves and strengthens the data interpretation and analysis in various fields.

Input: *hyperg* -> dictionary where *keys* => hyperedges and values => list of words in hyperedge

Output: *clust* => clustered obtained after application of cluster algorithm

1 Import networkx and matplotlib libraries.

2 Initialize dictionary *hyperg* to store the hypergraph

3 Initialize the list *allwords* to store the words in the hypergraph.

4 **Iterate** through all the hyperedges in the hypergraph:

4.1 Append hyperedge to the *listofnodes* list

4.2 Append each word in the hyperedge to the *allwords* list

5 Remove the duplicates from *allwords* list and store them in *setofwords* set

6 Initialize the dictionary *rank* to store the rank of each word

7 Initialize the rank to 0 in the *rank* dictionary for each word in the *setofwords* set

8 Create the dictionary *dictofwords* to map a word to its index in the *setofwords* set.

9 **Iterate** through all the hyperedges in the hypergraph:

9.1 Initialize the list *words* to store the words in the hyperedge

9.2 **For** words present in the *dictofwords* dictionary and the hyperedge add it to the *words* list

9.3 Add hyperedge index and words to the *hyperg* dictionary

10 Create a hypergraph G with the *networkx* library

11 **For** each pair of hyperedges between node1 and node2 present in the *listofnodes* list:

11.1 Initialize *newset* to store the common words in the selected hyperedges

11.2 If the intersection of the words present in node1 and node2 is not empty

11.2.1 Add edge between node1 and node2 in the hypergraph G

11.2.2 Calculate the common words and store them in the *newset* list

11.2.3 Increment the rank in the *rank* dictionary of all the words in the *newset* list

11.3 Else

11.3.1 Add node1 and node2 in the hypergraph G without a hyperedge

12 Sort and store the words in the *rank* dictionary based on decreasing rank order in the sortedwords list

13 Apply spectral clustering algorithm on the adjacency matrix of the hypergraph G to obtain cluster assignments

14 Create the list *clust* to store n clusters

15 Assign a hyperedge to its corresponding cluster based on the cluster assignment for all the hyperedges

16 Create the colour mapping dictionary *colormap* to store the colours of the hyperedges based on their cluster assignments.

17 Visualize the hypergraph G with the node colours from the cluster assignments.

18 Store the clusters

The hypergraph clustering involves the construction of a hypergraph, calculation of the rank of words based on occurrence, application of clustering algorithm to the hypergraph and visualization of the clusters.

Spectral clustering is a technique of hypergraph clustering where the data points are considered as graph nodes and the clustering problem is transformed into a graph-partition problem. Applying this technique and utilizing the eigenvalue decomposition method, the clusters are extracted. Spectral clustering is performed on the adjacency matrix of the graph by eigenvalue decomposition of the Laplacian matrix. Listed below are hypergraph clustering terms:

(i) Hypergraph Laplacian matrix:

Hypergraph Laplacian matrix is based on the hypergraph structure for spectral clustering

$$L = D - A, \tag{7}$$

where is the L Laplacian matrix, D is the degree matrix and A is the adjacency matrix of the hypergraph.

(ii) Hypergraph Normalized cut:

The normalized cut is the measure of the quality of hypergraph clustering.

$$N_{cut} = \frac{cut(H1, H2)}{vol(H1)} + \frac{cut(H1, H2)}{vol(H2)}, \tag{8}$$

where $cut(H_1, H_2)$ is the cut between two hypergraph partitions and $vol(H)$ is the volume of partition of hypergraph.

(iii) Hypergraph random walk:

The random walk on a hypergraph model is used to derive the cluster assignment based on which clustering takes place.

3.5. Hypergraph Dominating Set

A dominating set is formulated from the graph obtained in the hypergraph clustering process. The dominating set for an undirected graph $G = (V, E)$ where V and E are the edge and vertex set respectively D is such that $D \subseteq V$.A vertex v is considered to dominate itself and all its neighbors, for $v \in V$ the neighbors, are given by $N[v\{u \mid u = v$ or $uv \in E\}$.For every vertex $v \in V$, either $v \in D$ or a hyperedge $e \in E$ exists such that v is adjacent to e. The domination number for graph G is symbolized as $(G) = \min\{|D|\}$ (9). where D is a dominating set of the graph. The smallest dominating set in the hypergraph is a measure of the problem's complexity. The minimum cardinality of dominating set-in hypergraph H is represented as the domination number.

Finding a hypergraph dominating set involves figuring out the smallest set of vertices that can encompass the hypergraph's hyperedges. The intricacy or effectiveness of the solution is gauged by the size of the dominant set. It is computationally difficult to find the ideal answer promptly since finding the minimal hypergraph dominating set is an NP-hard task. To identify effective answers or approximate solutions to the problem, numerous heuristic and approximation methods have been presented.

Algorithm 4: ***Dominating Set***

Purpose: The aim of this algorithm is to investigate and depict the dominant set. This method locates and emphasizes the important nodes, which helps in network analysis and decision-making situations where comprehending the importance of certain nodes is crucial.

Input: Hypergraph *'G'* and dominating set on hypergraph 'G'

Output: support value and relative support value of each node in dominating set visualization of the dominating set using Matplotlib.

1 Import networkx and matplotlib libraries

2 Apply nx.dominating_set function from networkx library on the hypergraph G

3 Return the nodes that form the dominating set.

4 Initialize an empty *supplist* list to store the support value of each node

5 for node ->*dominating_set*

5.1 for hyperedge, words ->*hypergraph.items*()

5.1.1 if node -> words then Append support value to the *supplist* list

6 for *support_value* ->*supplist*:

6.1 relative_support => support_value / total_support

6.1.1 Append the *relative_support* value to supplist

7 Visualize the dominating set with Matplatlib

3.6. Cluster Analysis

Cluster analysis is a basic data exploration and segmentation approach that seeks to reveal hidden structures and patterns within a dataset. These clusters are defined using a similarity or distance metric between data points, with the aim of maximizing homogeneity within clusters and heterogeneity between them. In order to create clusters, cluster analysis locates naturally occurring groups within a dataset. The data points that are densely connected contribute to a cluster. Sparse regions represent weakly connected data points. The cluster analysis is used to identify the frequently used keywords related to a given medical plant. This helps in studying the relationship between the keywords and a medicinal plant and also a keyword with multiple medicinal plants. The clusters are represented in the form of a network that shows the spatial distribution of keywords mapped to the corresponding medicinal plants [19], [28].

4. Experimental Results

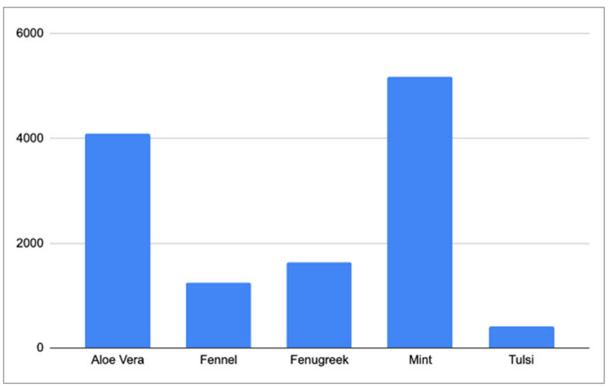
The methods followed in this research describe a complete approach to grouping medical plant keywords that incorporates text mining, hypergraph clustering, and data visualization approaches [24], [27]. Experiments were carried out using a diverse dataset of medical plant keywords to assess the success of the strategy. The analysis and interpretation of the data have crucial implications for future research and practical applications in clustering and text mining. This section presents some major results that demonstrate the usefulness of Hypergraph clustering.

4.1. Medical Plant Datasets: Data collection

In the dataset, the abstracts of five medicinal plants: Aloe Vera, Fennel, Fenugreek, Mint, and Tulsi were retrieved from PUBMED [7]. Figure 3 depicts a bar

Figure 3

Number of abstracts of each plant taken for this study



graph representation of the number of abstracts obtained from the PubMed database, which includes around 4092 Aloe Vera documents, 1246 Fennel materials, 1629 Fenugreek documents, 5175 Mint documents, and 408 Tulsi documents. Each of these abstracts was stored in a CSV file, and any rows with null values were eliminated as well.

4.2. Text Pre-processing and Feature Reduction

Text processing techniques are used to extract data for the analysis [13]. The data is cleaned, and then frequent stopwords and non-ASCII characters are eliminated before tokenizing it into words. The TF-IDF values for every term in the dataset were computed. The top terms from each document are arranged according to their TF-IDF scores. The words with scores exceeding a 0.02 threshold are retained, and discarding the remaining words, a concise list of significant terms for each document is obtained. This analysis provides a refined dataset for the task of keyword extraction. Figure 4 shows the essential words from the abstract that are obtained after pre-processing the datasets.

Figure 4
Word Cloud of the dataset after Pre-Processing

4.3. Clustering based on Apriori Algorithm

The Apriori algorithm is a well-known approach for association rule mining, which seeks to identify common item sets in a dataset [10]. The Apriori method creates a JSON file containing all the items along with their related support values, confidence scores, and lift values by setting a minimum support threshold of 0.005, a minimum confidence threshold of 0.5, and a minimum length of 1. The Apriori method is applied individually to all 5 plants to generate their respective item sets. The itemset retrieved for mint comprises 2053 items, whereas the itemset retrieved for fennel has 435. In the case of aloe vera, fenugreek, and tulsi, the items retrieved are comprised of 1106, 310, and 4732 items, respectively.

4.4. Hypergraph Clustering

The study used a hypergraph-based approach to understand the relationships between words in a dataset. The dataset was read from a JSON file and a set of unique words was created. A rank dictionary was initialized to track word appearances. The data was organized into a hypergraph, with each hyper edge representing a set of words. The hypergraph was converted into a NetworkX graph, and edges were added between sets that shared at least one word. The intersection of words between these sets updated the rank of individual words. The NetworkX library was used to visualize the graph, and spectral clustering was applied to partition the hypergraph into clusters. Figure 5 shows Clusters of the selected five plants.

Figure 5
Clusters of the selected five plants

4.5. Hypergraph Dominating Set

The evaluation identified a dominating set of clusters within the hypergraph derived from plants, with every non-included cluster connected to another one. By mapping these clusters back to their original items, support values were obtained for each cluster, revealing the significance of specific items across them. These findings provide valuable insights into the prevalence and importance of certain items in the dataset. Support values were visualized using a bar chart for easy identification. Figures 6-10 show a Bar graph of the dominating set of these plants. Here, the x-axis refers the support value for the collected dominating set.

4.6 Analyzing Keyword Clusters and Modularity

In the context of clustering, it can be viewed as a multi-objective optimization problem [16]. An undirected graph is created to examine keyword clusters. The nodes in the network indicate distinct keywords that were taken from a combined dataset. Based on the keywords that were extracted from each plant, clusters are to be assigned labels. The quantity of these clusters reveals the contribution of a specific plant to the biomedical nomenclature. Nodes between clusters represent benefits that two or more plants share in common.

The undirected graph in this study has 890 nodes and 1104 edges. With a resolution of 1.0 and randomized parameters, the graph's modularity is determined. The calculated modularity score is 0.577, which shows that the graph has a respectable amount of clustering structure. The graph also shows five communities, each of which represents a unique cluster or set of keywords with related properties. Figures 11-13 shows the keywords clustered for the different plants.

The multi-objective optimization perspective is used in this method to provide a systematic manner to find keyword clusters and their associated advantages. The communities that were discovered in the undirected graph reflect significant clusters of keywords that provide light on the connections and traits that various plants have in common in terms of their advantages.

Traditionally cluster analysis algorithms like k-means and hierarchical clustering rely on the pair-

Figure 6

Plant 1: Fennel Bar graph of dominating set

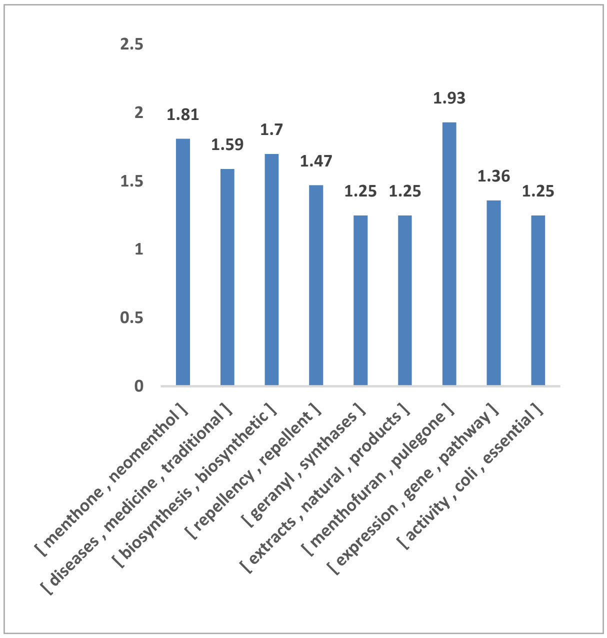


Figure 7

Plant 2: Fenugreek Bar graph of dominating set

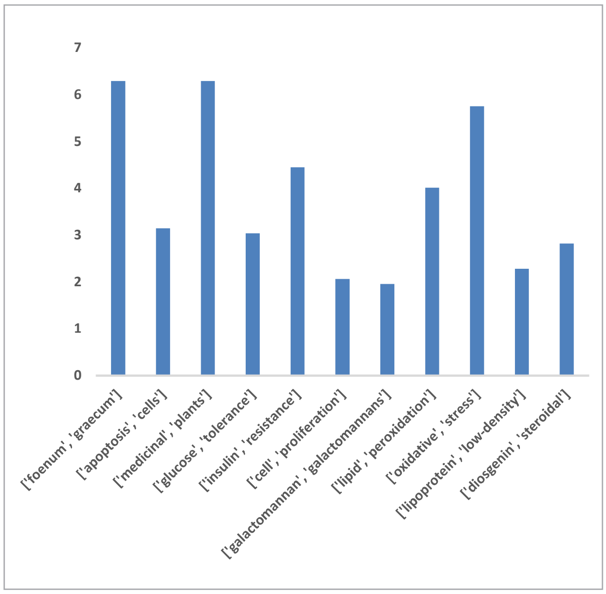


Figure 8

Plant 3: Mint Bar graph of dominating set

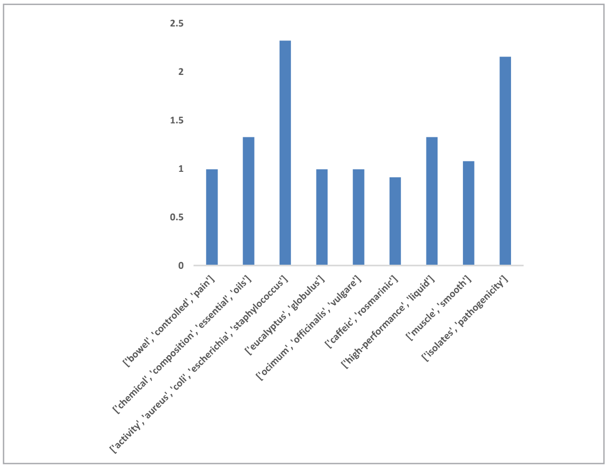


Figure 9

Plant 4: Tulsi Bar graph of dominating set

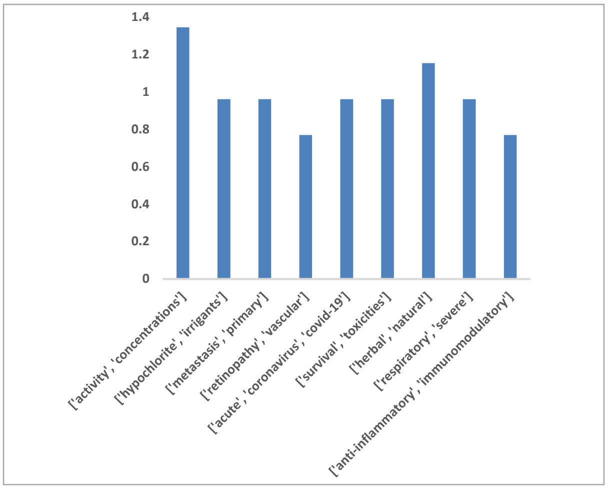


Figure 10

Plant 5: Aloe Vera Bar graph of dominating set

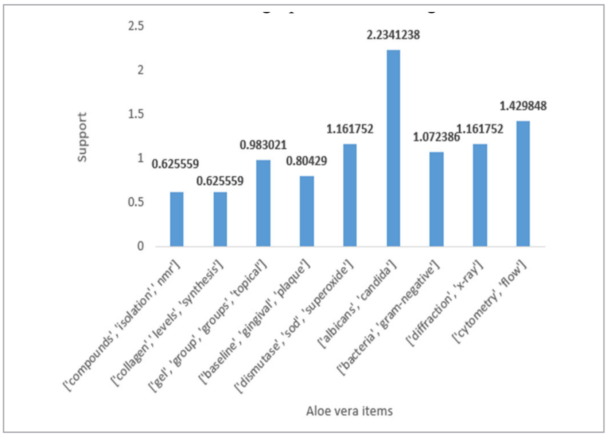


Figure 11

Keyword clustered of all five plants formulated using an undirected graph with percentage density

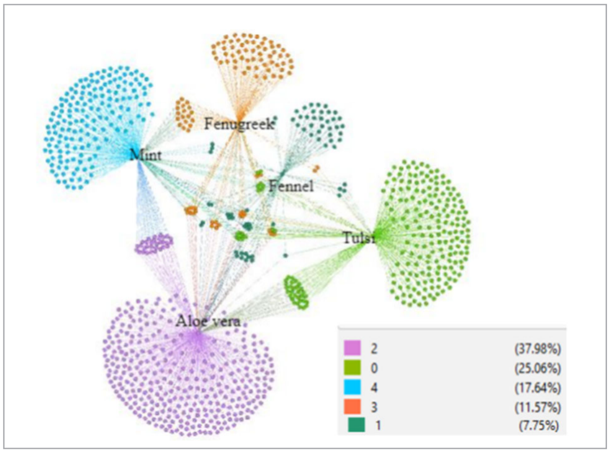


Figure 12

Keyword clustered of Aloe Vera and Fenugreek with percentage density

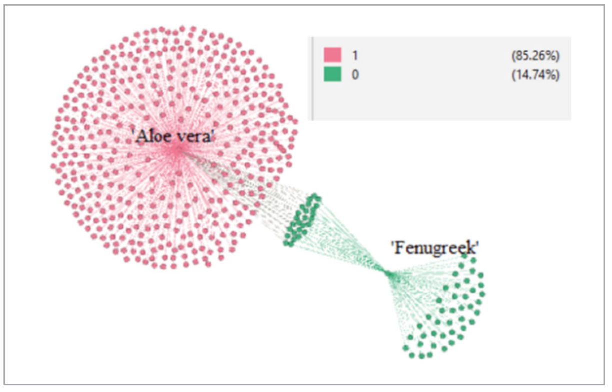
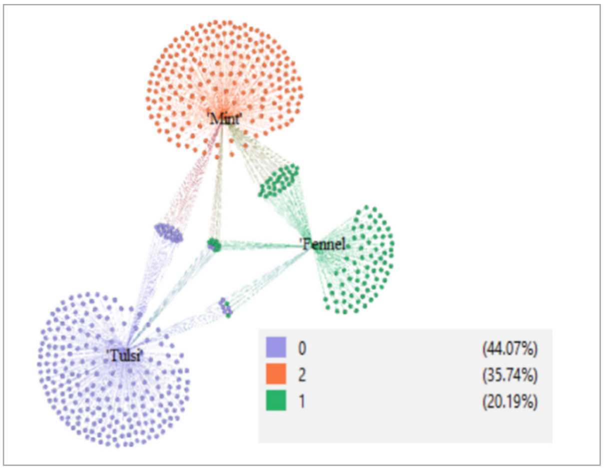


Figure 13

Keyword clustered of Mint, Fennel and Tulsi with percentage density



wise relationship between two data points [18]. The limitation of the binary relationships results as a hindrance to the discovery of complex relationships between the plants and the keywords. Thus, multiway association using hypergraph clustering is utilized to explore higher dimensional associations among the data points.

The cluster analysis of the keyword and plant association is a representation of the multiway association. As depicted in Figure 13, nodes of extracted keywords have edges from one or more than one plant. It can be inferred that Aloe Vera has its keywords closely linked to Tulsi and Mint and Mint has its keywords closely linked to Fenugreek and Aloe Vera.

Hypergraph-based clustering might generate less populated edges due to data sparsity in the high dimensional data. The complex cluster structures can also result in interpretability being more challenging than traditional clustering methods.

5. Conclusion

Identifying relationships among medicinal plants in published documents is crucial for advancing medicine. This paper introduces a keyword extraction model, Hypergraph-based Clustering with Dominating Set, comprising four phases: Text preprocessing, hypergraph construction, clustering, and dominating set determination. The Hypergraph is constructed, treating documents as edges and words within the edges as vertices. Clustering is then executed on the graph, and the dominating set property in the graph illustrates relationships among the extracted words. The study concludes by presenting all results through a relationship graph for better comprehension. This research significantly contributes to the expanding body of knowledge in medical plant research by providing a robust method for systematically clustering terms associated with medicinal plants. Beyond data organization, this framework not only enhances our understanding of medicinal plant relationships but also fosters new opportunities for investigation, creativity, and progress in the fields of natural and herbal medicine.

Conflict of interest

The authors declare that they do not have any conflict of interest. This research does not involve any human or animal participation. All authors have checked and agreed on the submission.

References

1. Agrawal, R., Imieliński, T., Swami, A. Mining Association Rules Between Sets of Items in Large Databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, June 1993, 207-216. https://doi.org/10.1145/170035.170072
2. Anand, U., Tudu, C. K., Nandy, S., Sunita, K., Tripathi, V., Loake, G. J., Dey, A., Proćków, J. Ethnodermatological Use of Medicinal Plants in India: From Ayurvedic Formulations to Clinical Perspectives-A Review. Journal of Ethnopharmacology, 2022, 284, 114744. https://doi.org/10.1016/j.jep.2021.114744
3. Anbarkhan, S., Stanier, C., Sharp, B. Text Mining Approach to Extract Associations Between Obesity and Arabic Herbal Plants. The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2018), 2018, 211-220. Springer International Publishing. https://doi.org/10.1007/978-3-319-74690-6_21
4. Athanasiadou, S., Githiori, J., Kyriazakis, I. Medicinal Plants for Helminth Parasite Control: Facts and Fiction. Animal, 2007, 1(9), 1392-1400. https://doi.org/10.1017/S1751731107000730
5. Behera, N. K., Mahalakshmi, G. S. Medicinal Plant Information Extraction System-A Text Mining-Based Approach. In Progress in Advanced Computing and Intelligent Engineering: Proceedings of ICACIE, 2017, 2, 215-226. Springer Singapore, 2019. https://doi.org/10.1007/978-981-13-0224-4_20
6. Behera, N. K., Mahalakshmi, G. S. A Cloud Based Knowledge Discovery Framework for Medicinal Plants from PubMed Literature. Informatics in Medicine Unlocked, 2019, 16, 100105. https://doi.org/10.1016/j.imu.2018.04.006
7. Canese, K., Weis, S. PubMed: The Bibliographic Database. The NCBI Handbook, 2013, 2(1).
8. Cheikhyoussef, A., Shapi, M., Matengu, K., Mu Ashekele, H. Ethnobotanical Study of Indigenous Knowledge on Medicinal Plant Use by Traditional Healers in Oshikoto Region, Namibia. Journal of Ethnobiology and Ethnomedicine, 2011, 7, 1-11. https://doi.org/10.1186/1746-4269-7-10
9. Cho, G., Park, H. M., Jung, W. M., Cha, W. S., Lee, D., Chae, Y. Identification of Candidate Medicinal Herbs

for Skincare via Data Mining of the Classic Donguibogam Text on Korean Medicine. Integrative Medicine Research, 2020, 9(4), 100436. https://doi.org/10.1016/j.imr.2020.100436

10. Cong, Y. Research on Data Association Rules Mining Method Based on Improved Apriori Algorithm. In IEEE 2020 International Conference on Big Data Artificial Intelligence Software Engineering (ICBASE), October 2020, 373-376. https://doi.org/10.1109/ICBASE51474.2020.00085
11. Darwish, S. M., Essa, R. M., Osman, M. A., Ismail, A. A. Privacy Preserving Data Mining Framework for Negative Association Rules: An Application to Healthcare Informatics. IEEE Access, 2022, 10, 76268-76280. https://doi.org/10.1109/ACCESS.2022.3192447
12. Hamilton, A. C. Medicinal Plants, Conservation and Livelihoods. Biodiversity and Conservation, 2004, 13, 1477-1517. https://doi.org/10.1023/B:BIOC.0000021333.23413.42
13. Kadhim, A. I. An Evaluation of Preprocessing Techniques for Text Classification. International Journal of Computer Science and Information Security (IJCSIS), 2018, 16(6), 22-32.
14. Kadiwal, S. M., Hegde, V., Shrivathsa, N. V., Gowrishankar, S., Srinivasa, A. H., Veena, A. Deep Learning Based Recognition of the Indian Medicinal Plant Species. In IEEE 2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA), 2022, 762-767. https://doi.org/10.1109/ICIRCA54612.2022.9985746
15. Kumar, T., Vaidyanathan, S., Ananthapadmanabhan, H., Parthasarathy, S., Ravindran, B. Hypergraph Clustering by Iteratively Reweighted Modularity Maximization. Applied Network Science, 2020, 5(1), 1-22. https://doi.org/10.1007/s41109-020-00300-3
16. Lee, N., Yoo, H., Yang, H. Cluster Analysis of Medicinal Plants and Targets Based on Multipartite Network. Biomolecules, 2021, 11(4), 546. https://doi.org/10.3390/biom11040546
17. Liao, K. Y., Wang, Y. H., Li, H. C., Chen, T. J., Hwang, S. J. COVID-19 Publications in Family Medicine Journals in 2020: A PubMed-Based Bibliometric Analysis. International Journal of Environmental Research and Public Health, 2021, 18(15), 7748. https://doi.org/10.3390/ijerph18157748
18. Likas, A., Vlassis, N., Verbeek, J. J. The Global K-Means Clustering Algorithm. Pattern Recognition, 2003, 36(2), 451-461. https://doi.org/10.1016/S0031-3203(02)00060-2
19. Martín Merino, M., López Rivero, A. J., Alonso, V., Vallejo, M., Ferreras, A. A Clustering Algorithm Based on an Ensemble of Dissimilarities: An Application in the Bioinformatics Domain. International Journal of Interactive Multimedia and Artificial Intelligence, 2022, 7(6), 6-13. https://doi.org/10.9781/ijimai.2022.09.007
20. Mutalib, S., Hasbullah, N. H., Abdul-Rahman, S., Shamsuddin, M. R., Ab Malik, A. M. Herbal Plant Analysis Based on Leaf Features Using K-Means Clustering. In IOP Conference Series: Earth and Environmental Science, 2022, 1019(1), 012026. IOP Publishing. https://doi.org/10.1088/1755-1315/1019/1/012026
21. Naresh, Y. G., Nagendraswamy, H. S. Classification of Medicinal Plants: An Approach Using Modified LBP With Symbolic Representation. Neurocomputing, 2016, 173, 1789-1797. https://doi.org/10.1016/j.neucom.2015.08.090
22. Okoye, T. C., Uzor, P. F., Onyeto, C. A., Okereke, E. K. Safe African Medicinal Plants for Clinical Studies. In Toxicological Survey of African Medicinal Plants, 2014, 535-555. Elsevier. https://doi.org/10.1016/B978-0-12-800018-2.00018-2
23. Petrovska, B. B. Historical Review of Medicinal Plants' Usage. Pharmacognosy Reviews, 2012, 6(11), 1. https://doi.org/10.4103/0973-7847.95849
24. Requena, S. H., Nieto, J. M., Popov, A., Delgado, I. N. Human Activity Recognition from Sensorised Patient's Data in Healthcare: A Streaming Deep Learning-Based Approach. International Journal of Interactive Multimedia and Artificial Intelligence, 2023, 8(1), 23-37. https://doi.org/10.9781/ijimai.2022.05.004
25. Rostami, M., Oussalah, M., Farrahi, V. A Novel Time-Aware Food Recommender-System Based on Deep Learning and Graph Clustering. IEEE Access, 2022, 10, 52508-52524. https://doi.org/10.1109/ACCESS.2022.3175317
26. Salmerón-Manzano, E., Garrido-Cardenas, J. A., Manzano-Agugliaro, F. Worldwide Research Trends on Medicinal Plants. International Journal of Environmental Research and Public Health, 2020, 17(10), 3376. https://doi.org/10.3390/ijerph17103376
27. Sandoval, A. M., Díaz, J., Llanos, L. C., Redondo, T. Biomedical Term Extraction: NLP Techniques in Computational Medicine. International Journal of Interactive Multimedia and Artificial Intelligence, 2019, 5(4), 51-9. https://doi.org/10.9781/ijimai.2018.04.001
28. Seal, A., Herrera Viedma, E. Performance and Convergence Analysis of Modified C-Means Using Jeffreys-Di-

vergence for Clustering, 2021. https://doi.org/10.9781/ijimai.2021.04.009

29. Shawkat, M., Badawi, M., El-Ghamrawy, S., Arnous, R., El-Desoky, A. An Optimized FP-Growth Algorithm for Discovery of Association Rules. The Journal of Supercomputing, 2022, 1-28.
30. Sher, H., Aldosari, A., Ali, A., de Boer, H. J. Economic Benefits of High-Value Medicinal Plants to Pakistani Communities: An Analysis of Current Practice and Potential. Journal of Ethnobiology and Ethnomedicine, 2014, 10, 1-16. https://doi.org/10.1186/1746-4269-10-71
31. Singh, D. B., Pathak, R. K., Rai, D. From Traditional Herbal Medicine to Rational Drug Discovery: Strategies, Challenges, and Future Perspectives. Revista Brasileira de Farmacognosia, 2022, 32(2), 147-159. https://doi.org/10.1007/s43450-022-00235-z
32. Smith-Hall, C., Larsen, H. O., Pouliot, M. People, Plants and Health: A Conceptual Framework for Assessing Changes in Medicinal Plant Consumption. Journal of Ethnobiology and Ethnomedicine, 2012, 8, 1-11. https://doi.org/10.1186/1746-4269-8-43
33. Sofowora, A., Ogunbodede, E., Onayade, A. The Role and Place of Medicinal Plants in the Strategies for Disease Prevention. African Journal of Traditional, Complementary and Alternative Medicines, 2013, 10(5), 210-229. https://doi.org/10.4314/ajtcam.v10i5.2
34. Spring, R., Johnson, M. The Possibility of Improving Automated Calculation of Measures of Lexical Richness for EFL Writing: A Comparison of the LCA, NLTK and SpaCy Tools. System, 2022, 106, 102770. https://doi.org/10.1016/j.system.2022.102770
35. Van Landeghem, S., De Bodt, S., Drebert, Z. J., Inzé, D., Van de Peer, Y. The Potential of Text Mining in Data Integration and Network Biology for Plant Research: A Case Study on Arabidopsis. The Plant Cell, 2013, 25(3), 794-807. https://doi.org/10.1105/tpc.112.108753
36. Wang, Z., Chen, J., Rosas, F. E., Zhu, T. A Hypergraph-Based Framework for Personalized Recommendations via User Preference and Dynamics Clustering. Expert Systems with Applications, 2022, 204, 117552. https://doi.org/10.1016/j.eswa.2022.117552
37. White, J. PubMed 2.0. Medical Reference Services Quarterly, 2020, 39(4), 382-387. https://doi.org/10.1080/02763869.2020.1826228
38. Yetisgen-Yildiz, M., Pratt, W. Using Statistical and Knowledge-Based Approaches for Literature-Based Discovery. Journal of Biomedical Informatics, 2006, 39(6), 600-611. https://doi.org/10.1016/j.jbi.2005.11.010