Informatics in Medicine Unlocked 16 (2019) 100226

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: http://www.elsevier.com/locate/imu

A cloud based knowledge discovery framework, for medicinal plants from PubMed literature

Behera Niyati Kumari [*], G.S. Mahalakshmi

Department of Computer Science and Engg, Anna University, India

ABSTRACT

MOBILE healthcare systems focus to achieve two specific goals: Making e-health applications and medical information available anywhere and anytime, and the invisibility of computing details. In this paper a cloud based mobile application has been initiated to identify herbal medicines of interest to the general population. This system responds to queries posted by users on medicinal plants relating to a particular disease by analysing research articles from PubMed via PubMed CLOUD and MeSH thesaurus. In this proposal, we particularly focus on applying text mining techniques to the biomedical literature, to unearth information concerning the curing of disease based upon the phyto-chemical properties of medicinal plants. The framework also includes an ontological structure to buffer the retrieved information, and thereby enhances the overall system efficiency.

1. Introduction

Herbs or medicinal plants are natural products with a long empirical history for curing disease at low cost. Their minimal or complete lack of side effects as compared with certain Western medicines have spurred researchers to consider and initiate several studies related to medicinal plants. In the present-day scenario, research to improve the medical efficacies of medicinal plants in target diseases, and to develop new herbal drugs for the treatment of various diseases, have become a fascinating and evolving task. Almost every medicinal plant has special chemical properties which endow it to ameliorate various diseases. For example, Ginger, Turmeric, “Green tea, and Aloe Vera have anti-inflammatory properties, which are helpful to remedy swelling and pain of arthritis patients. Similarly Pomegranate has an antioxidant capacity which reduces the oxidative stress in the human body and provides protection against some forms of cancer, other diseases, and cognitive impairment. Hence, identifying these chemical properties in herbs can be an exciting exploration but also challenging because of the unavailability of publically available databases related to medicinal plant-chemical property-disease relationships. In this era of electronic media, from the availability of a towering collection of biomedical literature like PubMed [9] which is a huge warehouse of biomedical articles, it becomes possible to extract important relationships from texts using a text-mining method. An illustration of the massive volume of research articles in PubMed is given in Fig. 1 (data as collected in September 2017).

Presently, investigators find cloud environment as a boon whether it is resource sharing or knowledge sharing. Healthcare domain is undoubtedly benefited by this as it allows physicians to institute the collaborative care of a patient. Android applications have added new points to the list of cloud benefits. As shown in the below figure, the huge amount of available data has become an overwhelming challenge for many health organizations, and the cloud helps to overcome this issue by reducing the in-house storage need. The PubMed Cloud [16] is a mobile application which can assist for easy search of PubMed research articles.This paper is a novel attempt to provide healthcare benefits to common people via mobile application. With the reason mentioned above for choosing the herb domain, this work proposes a cloud-based mobile application that can be used to popularize herbal medicines among the global population.

2. Related work

Medicinal Plants have a long empirical history for healing diseases with relatively less side effects. Since early days, several studies have been performed to identify the therapeutic values of those plants. It includes analysis of their chemical constituents, as these chemicals play a vital role in regulating biological activities in the human body that cause diseases. From those efforts has been coined a new term “Alternative Medicine” which is the idea of using plants for medical purposes. Unlike early studies, availability of a huge collection of biomedical literature in the present day public databases such as PubMed have

DOI of original article: https://doi.org/10.1016/j.imu.2018.04.006.

[*] Corresponding author.

E-mail addresses: niyatibehera@yahoo.co.in (B. Niyati Kumari), gsmaha@annauniv.edu (G.S. Mahalakshmi).

https://doi.org/10.1016/j.imu.2019.100226

Available online 20 August 2019

added a new dimension to the quality of knowledge discovery. For example, Choi and Lee [6] have proposed a framework to prove the medical efficacies of medicinal plants in target diseases and developing new drugs for the treatment of various diseases. Applying text mining along with machine learning for this literature has become a new era research trend. Gonzalez et al. [18] have discussed recent advances in the biomedical domain and how text data mining has been used to extract, analyse and/or evaluate information from the biomedical literature. Rong et al. [23] has applied text mining for large scale accurate drug and disease pair extraction from the literature. Wonjun et al. [11] have applied a text mining approach to prepare a corpus of relationships between plants and chemicals. Jensen et al. [12] have applied text mining and a Naive Bayes classification to identify plant–disease associations from PubMed abstracts. Li et al. [13] prepared a manually annotated corpus for disease, chemicals and disease–chemical interaction and have dedicated the same for the text mining research community. Similarly Karthikeyan et al. [28] have presented a manually curated database of Indian Medicinal Plants, Phytochemistry, And Therapeutics (IMPPAT). Wijaya et al. [14] have proposed a framework for Indonesian Medicinal Plants to predict relationships between disease and plants using supervised clustering and network analysis. A comprehensive and organized database for Herb Ingredients' Targets (HIT) has been constructed by Hao. et al. [15] to store herbal ingredients with protein target information. Though ample of literature is available in the herb domain, there is hardly any publically available resource for preserving the plant chemical property-disease relationships. Mining information from the huge literature has been another challenging task. Text mining methodology has been successfully employed to discover and analyse novel herbs or plants or formulas, from either the traditional Chinese medicine [24] or Korean Medical Literature [25] or historical electronic databases, for different diseases such as vascular dementia [19], chronic cough [20], diabetic nephropathy [21], obesity [27]. Similarly Selvaraj et al. [22] have applied text mining on literature collected from various data sources like PubMed/MEDLINE, Scopus, Science Direct (SciDir) and Wiley to recommend medicinal plants for diabetic treatment. Taking all these as a motivation, herein is proposed a novel attempt to develop a mobile based application for analysing the chemical activities of plants helpful for treating different health issues from PubMed articles. The following sections in this paper discuss the general architecture of our proposed framework, evaluation details, issues, and future enhancements.

3. Cloud based knowledge discovery framework for medicinal plants (CKFMP)

3.1. General architecture

As illustrated in Fig. 2, the entire process comprises four steps: First, the PubMed database is queried with the scientific names of medicinal plants commonly used for a particular disease over Android mobile. For example "Aloe Barbadensis", "Curcuma Longa", "Zingiber Officinale", "UncariaTomentosa", "Eucalyptus Globules", "Camellia Sinensis" and "Piper Nigrumlinn" are a few widely recommended medicinal plants for Rheumatoid Arthritis. The application forwards the request to the PubMed cloud via the Internet. Table 2 illustrates the details of medicinal plants chosen for this study. The cloud returns a set of related articles which are further processed to extract the required information. The retrieved information is stored in the buffer for future reference and is also returned to the user through a mobile interface. The objective of using the buffer is to avoid unnecessary interaction with the cloud if the user persistently asks the same query, and thereby increases the retrieval efficiency.

3.2. MeSH thesaurus category selection

MeSH is a Taxonomical structure designed by the National Library of Medicine (NLM) [10] that is used for searching health-related information and indexing documents in the biomedical domain. MeSH includes over 27,000 pre-defined descriptors covering all aspects of healthcare and medicine.

Among the sixteen categories of MeSH descriptor as shown in Table 1, only one category 'Chemicals and Drugs' (prefix D) has been chosen for our system. As our aim is to identify the name of the chemical property or agents which result in the treatment, prevention, cure or diagnosis of disease, the sub-category 'Pharmacologic Actions' under 'Chemicals and Drugs' has been selected. This sub-category lists a wide range of chemical actions and their uses that are helpful in the treatment, cure or diagnosis of disease. This sub-category includes drugs and chemicals that can alter normal body functions, and also the effect of various chemicals on the environment. As MeSH descriptors are updated annually, in this study, we have selected the 2017 version of MeSH.

3.3. Candidate term selection from biomedical literature

Candidate term selection from biomedical literature Several studies have been proposed to examine the utility of MeSH terms in document clustering [1,2], information retrival [3,4], PubMed query refinement etc. Authors of LitLinker [5] have experimentally found that using MeSH

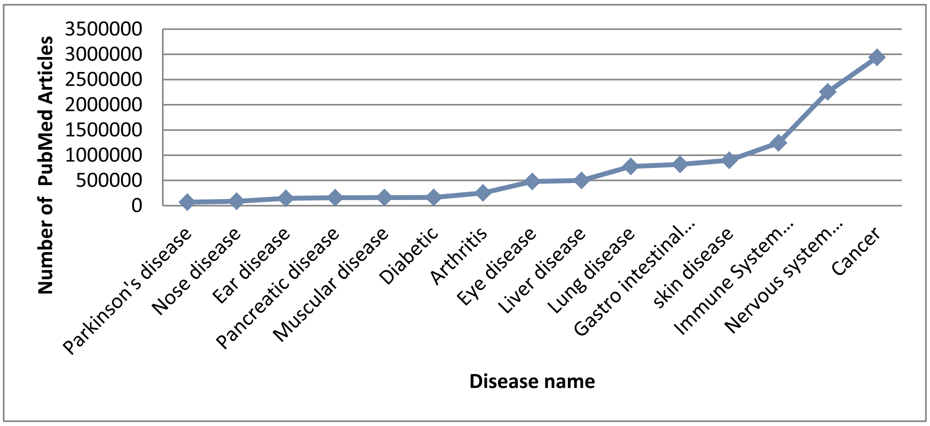

Fig. 1. Sample illustration of the number of PubMed articles for selected diseases (data as collected in September 2017).

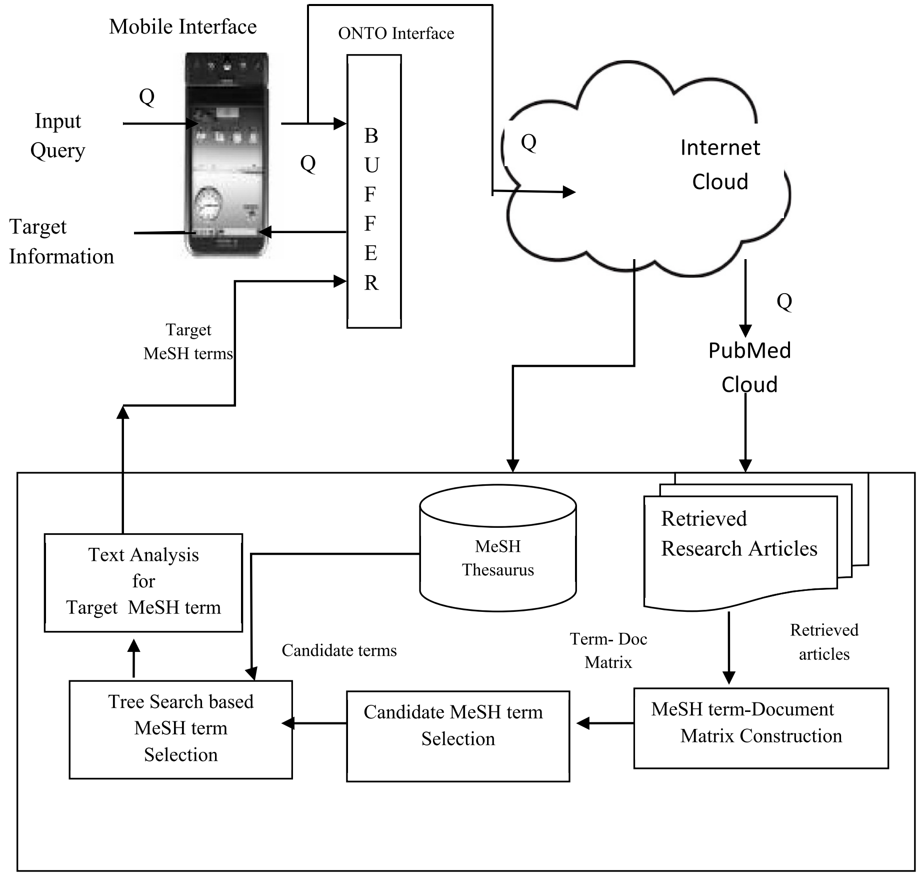

Fig. 2. Overall process details of knowledge discovery framework for Medicinal Plants.

Table 1
MeSH category details.

MeSH category	Selected	MeSH categories	Selected
Anatomy (A)	No	Organisms(B)	No
Diseases (C)	No	**Chemicals anddrugs (D)**	**Yes**
Analytical, diagnosticand therapeutic techniques andequipment(E)	No	Psychiatry andpsychology(F)	No
Biological Science (G)	No	Physical Science (H)	No
Anthropology, education, sociology and social phenomena(I)	No	Technology, industry, Agriculture(J)	No
Humanities (K)	No	InformationScience(L)	No
Named Groups (M)	No	Health Care (N)	No
PublicationCharacteristics(V)	No	Geographical (Z)	No

Table 2
Data Set Used for feature extraction.

Sl. No	Type of Relationship	No of Sentences
1	Treat/Cure for Disease	830
2	Prevent	63
3	Side Effect	30
4	Disonly	629
5	Treatonly	169
7	vague	37
8	To see	75
9	No Treat/Cure for dis	4
10	None	1818
Total		3655

terms to represent document can be computationally inexpensive compared to the conventional NLP method of representation. With motivation from previous work, here we present a framework to establish correspondence between MeSH term and a particular disease. The candidate term selection process begins with searching the MEDLINE database with the common as well as botanical name of medicinal plants recommended for a target disease. Our system then filter MeSH terms belonging to the target MeSH category i.e. Chemicals and drugs (Prefix-D). Then using the MeSH terms indexed in document, a term-document matrix is created for each literature. In this paper, we have used the term literature to define a set of documents retrieved from database for a given query as medicinal plant name curing a particular disease. For instance, in this study 15 Indian medicinal Plants recommended for arthritis have been considered. So we retrieve 15 sets of documents from MEDLINE database where each set is defined as a literature. A major task of our text mining approach is to identify MeSH terms a strong correlation or association with the target disease. Conventional interpretation to this problem would be to find the term frequency in the local literature but this approach not necessarily revel the correlation. For example, the association between ‘arthritis’ and ‘plant extract’ is not very interesting than the association between ‘arthritis’ and ‘Anti-inflammatory agent, nonsteroidal’ though ‘plant extract’ appeared in 22 documents of turmeric literature where as ‘Anti-inflammatory agent, non-steroidal’ appeared in 24 documents.

To address this issue, we have considered term probability in the literature as compared with term frequency. We computed the MeSH term frequency in a literature by the number of documents in the literature which contained the term by the total number of documents in the literature. However, it was observed that the probability distribution of a term over the entire literature set can reveal more interesting correlations. To calculate the mean probability distribution, first we find the probability P of a MeSH term m, in a literature l as given in equation (1) below.

$$P_l^m = \frac{D_l^m}{L_l} \tag{1}$$

Where D_l^m is the number of documents with the MeSH term m in literature l and L_l is the number of documents in literature l. We then calculate the mean probability of the MeSH term m in all of the literatures using the following formula:

$$\overline{P^m} = \frac{\sum_{l=1}^{N} P_l^m}{N} \tag{2}$$

Finally the MeSH terms with mean probability above threshold value are selected as candidate terms for further processing. The threshold value used here is the average mean probability (AMP) of all of the MeSH terms.

3.4. Selection of target MeSH terms

As our aim is to identify the phyto-chemical properties associated with a given disease, we have fixed the target node as "**Pharmacological Action**" on the MeSH hierarchy to initiate the search algorithm. In the MeSH hierarchy, concepts are ordered from broad to specific prospective. This descriptor includes a range of category descriptors which represent chemical actions and uses that result in the prevention, treatment, cure or diagnosis of disease. It also lists drugs and chemicals that act by altering normal body functions and effects of chemicals on the environment.

Fig. 3 gives a detailed outline of the algorithm for pruning of candidate MeSH terms which are generated in the pre-processing step. The algorithm takes three parameters as input:

Candidate MeSH term set {C}
MeSH Tree (T)
Target node (n)

During this subtask, a subset of MeSH descriptor from the list of candidate terms C, which have strong association with target node, are selected. Firstly, a simple tree search technique is applied on T to identify all the child nodes of target node 'n' i.e. Pharmacological Action, and store them in a list 'L1'. Then for each candidate term 'Ci', list L1 is searched to check if it is a descendant of the target node. The reason behind choosing only the children nodes instead of ancestor, siblings and parent of the candidate term is that always the ancestors describe a broad and too general aspect of any target node. If 'Ci' is present in the list, then it is selected as a target MeSH term, else the loop continues to check for other candidate terms in the set.

3.5. Text mining based target selection from biomedical literature

3.5.1. Biomedical data set for feature extraction

We used the standard biomedical text corpus/data set that is obtained from Ref. [26]. This data set is prepared from Medline 2001 abstracts. It is a collection of 3655 sentences annotated with eight possible types of relationships between "Disease" and "Treatment". Among those, only 923 sentences belonging to three relation types "Treat/Cure for Disease", "Prevents" and "Side Effect" have been taken into account for feature generation. Our motivation to identify sentences containing information about phyto property and its influence on a disease is the main reason behind this selection process. Hereafter throughout the paper these sentences are referred to as Informative Sentences (IS) and feature extraction is restricted from these sentences only.

3.5.2. Feature extraction

We have considered Part_of_Speech (POS) tagged unigram features as the basis for feature extraction process. Words with five POS groups like Verbs(v), Nouns(n), Adjectives(j), Adverbs(a) and Preposition(p) were used to extract unigram features from the selected informative sentences. Regular NLTK POS tagger has been used in this stage for feature extraction. For simplicity we grouped NLTK POS tags into the following equivalents as shown in Table 3.

In the literature, though stop words have mostly been removed during the pre-processing step prior to the feature extraction stage, we found that certain stop words such as 'of', 'for', 'by', 'in', 'with','-on','against' having POS tag 'IN' contributed to 22%–25% of informative sentences. Hence, we included these unigram features to improve the information retrieval efficiency of our proposed system.

As discussed in the literature, noun phrases and verb phrases convey more important information for biomedical relation identification;

Table 3
Grouping of NLTK POS tags.

POS Group	NLTK POS tags
v	VB,VBD,VBG,VBN,VBP,VBZ
n	NN,NNP,NNS,NNPS
j	JJ,JJR,JJS
a	RB,RBR,RBS
p	IN

```
Algorithm
Input   : List of candidate MeSH terms C=($C_1$,$C_2$,.........$C_m$),
          MeSH tree (T),
          Target node in T, n where n= "Pharmacological Action"
Output:   List of Target MeSH terms $C_i$ associated with 'n'

1.  Locate target node 'n' in T
      $ID_n$ ← locate(T,n)
2.  List of Child( L1) ← NULL
      L1 ← Retrieve child nodes($ID_n$)
3.  NewL ← NULL
4.  For each $C_i$ in C do
          If $C_i$€ L1 Then
          NewL ← NewL U { $C_i$}
5.  Return NewL
```

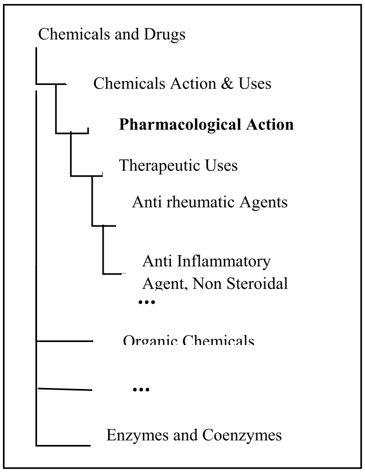

Fig. 3. Selection based on MeSH (SBoM) algorithm and part of MeSH tree.

therefore we considered these multi-gram phrasal features along with unigram features (selected stop words). We used text analysis tools to extract all the noun and verb phrases from informative sentences.

Then the next stage mainly focuses upon choosing the most significant features from the feature set generated in the previous stage. Initially we filtered the features by word size >2 except for the selected stop words, and then by corpus feature frequency >3. Remaining features were used to train the classifier.

Feature weight calculation is an integral part of the classification process. Literature suggests several weight techniques like 'Term frequency',' Term Presence' and 'TF-IDF' etc. We used probabilistic term frequency to assign feature weight (both unigram and multi-gram).

$$\mathrm{W_f} = \frac{n(\mathrm{f}|\mathrm{IS})}{\sum_{k=1}^{m} n(W(k)|IS)} \tag{3}$$

Where.

$\mathrm{W_f}$ = feature weight n(f|IS) = occurrence of feature in informative class
n(W(k)|IS) = sum of occurrence of all features in informative class
IS = informative sentences pertaining to cure/treat disease, side effect and prevent relationship type

3.5.3. Sentence classification

This phase focuses on selecting sentences having evidence that a phyto property is helpful in treating a particular disease based upon the features selected in earlier stages. Thus, given a training set of sentences (each labelled with a class) we trained a multinomial binary sentence classification model.Two vocabularies V1 and V2 were defined, where the number of words in the vocabulary defines the dimension of the feature vectors depending on the position of the feature in the training sentence. One vocabulary considered the feature position as anywhere in the sentence (FAS), whereas the other vocabulary considered only the features that occur between the target entities, i.e., disease and treatment (FBE).

To classify an unlabelled sentence S, we estimate the posterior probability for each class in terms of words or features ' u' which occur in the sentence:

$$\mathrm{P}\left(\mathrm{Ck}|\mathrm{S}\right) = \mathrm{P}\left(\mathrm{Ck}\right)\prod_{\mathrm{j=1}}^{\mathrm{len\ S}}\mathrm{P(uj|Ck)} \tag{4}$$

where $\mathrm{u_j}$ is the j'th word in sentence S.

3.6. BUFFER: an onto interface

Buffer, as the name reveals, is meant for storing information. The proffered system includes an ontological structure as a part of buffer. The ontology is progressively enriched with the information retrieved from the literature. The ontological concepts have been reused from the Philippine Medicinal Plants ontology [17]. Design of this ontology is based upon available information that can be consistently extracted from web documents. This information includes the therapeutic properties of the plant, such as illness it can be applied to, the instruction for medicinal preparation, the body part it affects/cures, and the plant component that is used in the healing process. An accuracy of 85.71% has been claimed in this paper. Following are the defined classes of the considered ontology:

- Medicinal Plant: medicinal plant instances in the ontology.
- Illness: health issues or diseases treated by the plants.
- Body Part: body parts affected by the medicinal plants and illnesses.
- Location: geographical provinces where the medicinal plants are cultivated.
- Plant Part: the plant parts utilized in the medicinal preparations.
- Preparation: preparation for the medicinal plant given an disease.

Through this interface as illustrated in Fig. 4, we plan to populate the ontological concepts with information extracted from the PubMed literature. For example: the instances of the class "MedicinalPlant" are to be populated with data properties like "has taste" and "has colour", and the data object property " helpful phyto-property" to be added to the class "Illness"(which is our objective in this framework).

4. Results and discussion

We have performed experiments for three diseases, 'arthritis', 'diabetic' and 'cancer'. The user queries the system with a 'disease name' via the Interface. We have a set of recommended medicinal plants for each disease in the background. Initially, the buffer is searched for the given query information. If it is available, an immediate response is sent to the user; otherwise the query is forwarded to the Internet cloud for further search with a recommended set of medicinal plants. Table 4 provides a brief description of literature collected from PubMed (Data as collected on September 2017).

To explain the final term selection process, consider the disease 'arthritis'. In the literature, our interest is to identify the properties which contribute to the effect on the disease. The findings of our experiments revel that 'anti-inflammatory agent, nonsteroidal' and 'antioxidants' are closely associated with arthritis, as these properties are helpful to reduce or prevent the disease. The proposed framework is evaluated based on two factors with the help of Web Experts [7,8].

- Whether the considered medicinal plants possess these phyto properties.

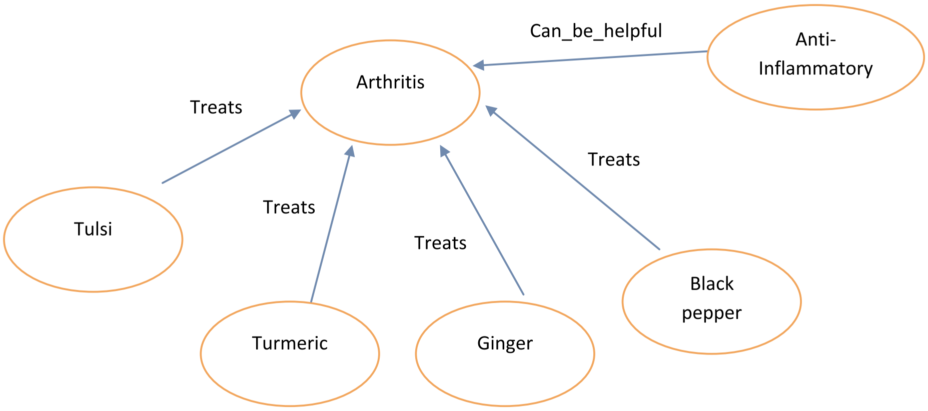

Fig. 4. Populating Ontological Concepts with Object and Data Property values.

Table 4
Details of literature collected from PubMed.

Disease Name	No of Medicinal Plants considered	No of collected documents	Medicinal Plants name
Arthritis	15	253	Black Pepper, Ginger, Aloe Vera, Indian Olibanum, Green Tea, Liquorice, Turmeric, Pomegranate, Indian long pepper, Tulsi,Blackboard tree, Cinnamon,Cinchona, Coriander, Indian Tinospora
Diabetic	15	251	Turmeric, Bitter gourd, Neem, Fenugreek,AloeVera,Mango, Holy basil, Indian Tinospora, Cumin, Cowplant,stone apple, Garlic,cluster fig,Indian Kino Tree, Amla
cancer	10	1582	Turmeric, Ashwagandha, Guggul,Kanchnaar, Tulsi, Indian Echinacea, Wheat grass, Black-cumin, Garlic, Cardamom

- Whether these phyto properties play a major role in suppressing or preventing the disease.

We refer to Dr. Duke's Phytochemical and Ethnobotanical databases [7] to check the first criteria. This database provides the user with a platform for.

- obtaining a list of chemicals and activities for a specific plant of interest, using either its scientific or common name
- finding plants with chemicals known for a specific biological activity
- displaying a list of chemicals with their LD toxicity data
- displaying a list of plants for a given ethnobotanical use
- determining which plants have the highest levels of a specific chemical

We manually check the chemical details of each plant considered for the specific disease, to verify whether they include the above-mentioned bioactivity. We interestingly found that all the plants, taken as a test dataset, possess at least one of these chemical activities in their chemical details. Our analysis excludes the plants listed as ethno botanical plants (e.g. Indian Tinospora, Kanchnaar, Cluster Fig and Indian Long Pepper), as their chemical details are not provided in the database. Results of part of the proposed framework meant for extracting knowledge from the PubMed literature are given in Table 5.

Table 5
Details of Retrieved phyto -properties.

Disease name	Target MeSH terms
Arthritis	Anti Inflammatory Agent, Nonsteroidal, Antioxidants
Diabetic	Antioxidants, Hypoglycemic Agent
Cancer	Anti neoplastic Agents, Anti Inflammatory Agent

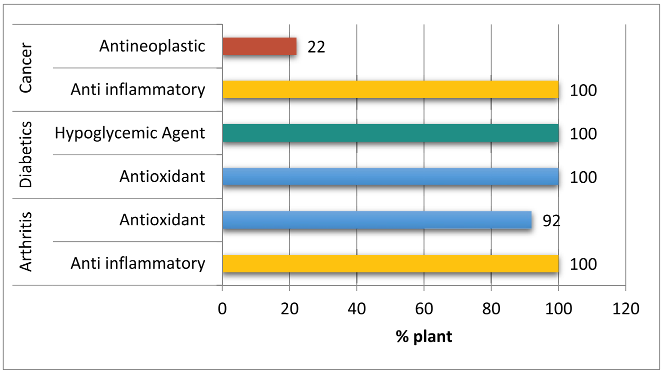

Fig. 5. Summary of plants possessing the retrieved phyto-chemical activities.

Table 6
Plants with the retrieved phyto-chemical activities.

	Anti Inflammatory	Antioxidant	Anti neoplastic	Hypoglycemic Agent
Turmeric	Yes	Yes	-	Yes
Aloe-Vera	Yes	Yes	Yes	Yes
Ginger	Yes	Yes	Yes	Yes
Black Pepper	Yes	Yes	-	Yes
Pomegranate	Yes	Yes	-	Yes
Indian Olibium	Yes	Yes	-	-
Green Tree	Yes	Yes	-	Yes
Blackboard Tree	Yes	-	-	-
Cinnamon	Yes	Yes	-	Yes
Chincona	Yes	Yes	-	Yes
Coriander	Yes	Yes	Yes	Yes
Liquorice	Yes	Yes	Yes	Yes
Tulsi	Yes	Yes	-	Yes
Bitter Gourd	Yes	Yes	-	Yes
Neem	Yes	Yes	-	Yes
Fenugreek	Yes	Yes	-	Yes
Mango	Yes	Yes	-	Yes
Cumin	Yes	Yes	-	Yes
Cowplant	Yes	Yes	-	Yes
Stone Apple	Yes	Yes	-	Yes
Garlic	Yes	Yes	Yes	Yes
Indian Kino Tree	Yes	Yes	-	Yes
Amla	Yes	Yes	Yes	Yes
Aswagandha	Yes	Yes	-	Yes
Guggul	Yes	Yes	Yes	Yes
Indian Ehinacea	Yes	Yes	-	-
Wheat Grass	Yes	Yes	-	Yes
Black Cumin	Yes	Yes	-	Yes
Cardamom	Yes	Yes	-	Yes

Our experimental result includes 3 dataset involving 30 distinct medicinal plants (excluding the plants listed as Ethnobotanical plants in Dr. Duke's database) for diseases like 'arthritis', 'diabetes' and 'cancer'. An exhaustive analysis of chemical activity details of these plants revealed that all of the plants, taken as a test dataset, possess at least one of the retrieved chemical activities in their chemical details. Fig. 5 summarizes the percentage of medicinal plants having the retrieved properties in the considered data set. Existence or non-existence of the retrieved chemical activities in all of the plants that have been considered here for study have been depicted in Table 6. The MeSH term "Anti Inflammatory agent, Nonsteroidal" is one of the target terms identified in connection with Arthritis. In addition to anti-inflammatory actions, it has analgesic, antipyretic, and platelet-inhibitory actions.

The text mining based target selection task takes these phyto property names as well as the disease name, as input, and retrieve their child nodes from the MeSH thesaurus. Then sentences containing these entity pairs (phyto property name-disease name) were identified in the text corpus collected by downloading abstracts from PubMed in the initial phase. Based on the feature sets generated from the reference dataset as discussed in the previous sections, these sentences were further classified as informative or non-informative. Informative sentences contain a clear affirmation that a particular phyto chemical property is helpful in treating the target disease.

For example:

> *"Ginger has a long history of medicinal use, particularly as an **anti-inflammatory agent** for a wide variety of diseases such as **arthritis**.* "[PMID 15750374]

The non-informative sentence, though it includes both entities, does not confirm the usefulness of one on the other. For example:

> *"Medicinal herbs have been effectively used for their anti-inflammatory activity, but their exact role has not yet been documented in the scientific literature for the management of Osteoarthritis (OA)."* [PMID 24228609]

All the retrieved sentences are then pre-processed, and using text mining tools Pattern and Text Blob, all verb and noun phrases were extracted. Then a trained probabilistic classifier was used to classify the sentences as IS and NIS. We conducted experiments for the disease 'arthritis' and achieved a classification accuracy of 66% for FBE and 73% for FAS. Apart from the shortcomings of the text mining tools used in conducting experiments, another major reason for decrease in system performance can be attributed to the fact that the system follows a kind of pipelined process, i.e., sentence extraction followed by sentence classification. Therefore, the performance of the first stage affects the performance of the following stage. In due course of the experiment, it was found that in the test dataset, though there were entities which were actually related, they didn't appear in the same sentence.

> *"Curcumin derived from the tropical plant Curcuma longa has a long history of use as a dietary agent, food preservative, and in traditional Asian medicine. It has been used for centuries to treat biliary disorders, anorexia, cough, **diabetic** wounds, hepatic disorders, **rheumatism**, and sinusitis. The preventive and therapeutic properties of curcumin are associated with its **antioxidant**, **anti-inflammatory**, and anticancer properties."* [PMID 22996381]

As our system considers MeSH terms as the backbone for all selection processes, one of major drawbacks of this work is the fact that the MeSH browser includes only limited medical terms. In the future, we plan to include another large medical thesaurus to improve system performance. For instance, phyto activity details of plants in Duke's database don't contain any term "antineoplastic agent" with respect to cancer treatment, but they contain the term "anticancer", which MeSH browser doesn't include. Our system fails to recognize the semantic similarity between these terms.

As a part of a further extension, we plan to include a few additional frameworks to extract more information about medicinal plants from the available literature, like the chemical compound that possesses the above mentioned chemical activities. For instance:

> *"**Curcumin** (diferuloylmethane) is an orange-yellow component of turmeric (Curcuma longa), a spice often found in curry powder . Traditionally known for **its anti-inflammatory effects**, curcumin has been shown in the last two decades to be a potent immunomodulatory agent that can modulate the activation of T cells, B cells, macrophages, neutrophils, natural killer cells, and dendritic cells . This suggests that **curcumin's** reported beneficial effects in **arthritis**, allergy, asthma, atherosclerosis, heart disease, Alzheimer's disease, **diabetes**, and **cancer** might be due in part to its ability to modulate the immune system."* [PMID 17211725]

5. Conclusion

This paper, a simple yet novel attempt to propose a mobile cloud based framework to analyse PubMed literature on cloud and identify the biochemical activity of medicinal plants involved in soothing disease. We have completed the information retrieval part without involving mobile. In future we plan to deploy the application in android environment to prove the real significance. In due course of information retrieval process, it was found that combining statistical method based on word probability distributions with a knowledge-based approach can very well enhance system effectiveness. Making extracted information available at fingertip on phone can help to bring the of herbal domain one step closer to a larger community.

References

[1] Zhu S, Zeng J. Enhancing MEDLINE document clustering by incorporating MeSH semantic similarity". Mamitsuka H, Bioinformatics. 2009;25(15). Aug 1.

[2] Gu J, Feng W, Zeng J, Mamitsuka H, Zhu S. Efficient semisupervised MEDLINE document clustering with MeSH –semantic and global –content constraints. IEEE Trans Cybern 2013 Aug;43(4):1265–76.

[3] Sang-Jun Yea, BoSeok Seong, Yunji Jang, Chul Kim. A data mining approach to selecting herbs with similar efficacy: targeted selection methods based on medical subject headings (MeSH). J Ethnopharmacol April 2016;182(22):27–34.

[4] Yu Z, Bernstam E, Cohen T, Wallace BC, Johnson TR. Improving the utility of MeSH terms using the Topical MeSH representation. J Biomed Inform 2016 Jun;61: 77–86.

[5] Yetisgen-Yildiz M, Pratt W. Using statistical and knowledge-based approaches for literature-based discovery. J Biomed Inform 2006 Dec;39(6):600–11.

[6] A text mining approach for identifying herb-chemical relationships from biomedical articles", Wonjun Choi, Hyunju Lee, in Proceedings of the ACM ninth international workshop on data and text mining in biomedical informatics, pp.25-25.

[7] Dr Duke's Phytochemical and ethnobotanical databases, " https://phytochem.nal.usda.gov/phytochem/search.

[8] https://www.webmd.com/default.htm.

[9] https://www.ncbi.nlm.nih.gov/pubmed/.

[10] https://meshb.nlm.nih.gov/search.

[11] Choi Wonjun, Kim Baeksoo, Cho Hyejin, Lee Doheon, HyunjuLee. A corpus for plant-chemical relationships in the biomedical domain. BMC Bioinf 2016;17:386.

[12] "Integrated text mining and Chemoinformatics analysis associates diet to health benefit at molecular level", Jenson K, Panagiotou G, Kouskoumvekaki I. PLoS Comput Biol 2014;10(1):1003432.

[13] Li J, Sun Y, Jonhnson R, Sciaky D, Wei C, Leaman R, Davis AP, Mattingly C, Wiegers T, Lu Z. Annotating chemicals, diseases and their interactions in biomedical literature. In: In: proceedings of the fifth BioCreative challenge evaluation workshop. Sevilla: BioCreative Organizing Committee; 2015. p. 173–82.

[14] Wijaya Sony Hartono, Husnawati Husnawati, Afendi Farit Mochamad, et al. Supervised Clustering Based on DPClusO: Prediction of Plant-Disease Relations Using Jamu Formulas of KNApSAcK Database. BioMed Research International 2014;2014:831751. 15 pages, https://doi.org/10.1155/2014/831751.

[15] Ye Hao, Ye Li, Kang Hong, Zhang Duanfeng, Lin Tao, Tang Kailin, XuepingLiu, Zhu Ruixin, Qi Liu, Chen YixueLi YZ, ZhiweiCao. HIT: linking herbal active ingredients to targets. Nucleic Acids Res 1 January 2011;Volume 39:D1055–9. Issue suppl_1.

[16] http://www.vqol.com/2017/07/03/pubmed-cloud-448611222.html.

[17] Lim-Cheng Nathalie Rose, Junn Richmond C Co, Christa Hannah S Gaudiel, Umadac Darah F, Victor Nadine L. Semi-automatic population of ontology of philippine medicinal plants from on-line text. In: Presented at the DLSU research congress. Manila, Philippines: De La Salle University; March 2014. p. 6–8.

[18] Gonzalez GH, Tahsin T, Goodale BC, Greene AC, Greene CS. Recent advances and emerging applications in text and data mining for biomedical discovery. Briefings Bioinf 2016;17:33–42.

[19] Feng S, Ren Y, Fan S, Wang M, Sun T. Discovery of acupoints and combinations with potential to treat vascular dementia: a data mining analysis. Evid Based Complement. Altern Med 2015;2015:310591.

[20] Shergis JL, Wu L, May BH, Zhang AL, Guo X. Natural products for chronic cough: text mining the East Asian historical literature for future therapeutics. ChronRespir Dis 2015;12:204–11.

[21] Zhang L, Li Y, Guo X, May BH, Xue CC. Text mining of the classical medical literature for medicines that show potential in diabetic nephropathy. Evid Based Complement. Alternat Med 2014;2014:189125.

[22] Indian medicinal plants for diabetes: text data mining the literature of different electronic databases for future therapeutics. Bhanumathi Selvaraj, Sakthivel Periyasamy, Biomedical Research 2016:S430–6. Special Issue.

[23] Xu Rong, Quan Qiu Wang. Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing. BMC Bioinf 2013; 14:181.

[24] Xuezhong Zhou, Yonghong Peng, Baoyan Liu. Text mining for traditional Chinese medical knowledge discovery: a survey. J Biomed Inform August 2010;43(4): 650–60.

[25] Choi Moo Jin, Choi Byung Tae, Shin Hwa Kyoung, Shin Byung Cheul, Han Yoo Kyoung, Baek Jin Ung. Establishment of a comprehensive list of candidate antiaging medicinal herb used in Korean medicine by text mining of the classical Korean medical literature,"Dongeuibogam," and preliminary evaluation of the antiaging effects of these herbs. Evidence-Based Complementary and Alternative Medicine 2015;2015:873185. 29 pages, https://doi.org/10.1155/2015/873185.

[26] Rosario B, Hearst Marti A. Classifying semantic relations in bioscience text. In: ACL '04 proceedings of the 42nd annual meeting on association for computational linguistics; 2004. Article No. 430.

[27] Anbarkhan Samar, Stanier Clare, Sharp Bernadette. Text mining approach to extract associations between obesity and Arabic herbal plants. In: The international conference on advanced machine learning technologies and applications (AMLTA2018); 2018. p. 211–20.

[28] Mohanraj Karthikeyan, Karthikeyan Bagavathy Shanmugam, VivekAnanth RP, Bharath Chand RP, Aparna SR, Mangalapandi P, Samal Areejit. IMPPAT: curated database of Indian medicinal plants, Phytochemistry and therapeutics. https://doi.org/10.1101/206995; 2017.

Further reading

[29] Frunza Oana, Inkpen Diana, Tran Thomas. A machine learning approach for identifying disease-treatment relations in short texts. IEEE Trans Knowl Data Eng june 2011;23(6).