SYMPTOMS BASED PADDY CROP DISEASE PREDICTION AND RECOMMENDATION SYSTEM USING BIG DATA ANALYTICS

S. Nithya[1], S. Savithri[2], G. Thenmozhi[3], Dr. Kalaimani Shanmugham[4]
[1, 2, 3] *UG Students, Department of Computer Science and Engineering*
[4] *Professor/Head of Computer Science and Engineering*
Arasu Engineering College, Kumbakonam, Tamil Nadu, India

***Abstract*— Agriculture is the art and science of growing plant and other crops for rising economic gain. However, the diseases that affect the paddy crops have make a little impact of agriculture production. The paddy usually gets infected by pathogens such as bacteria, fungus and virus. In order to address the above issue a novel recommendation system for earlier detection of paddy diseases. The paddy disease related data's are gathered from heterogeneous websites such as Agropedia and blogs, processed and analyzed through Hadoop, Hive tools and HiveQL. The collected documents are represented in the form vector using Vector Space model and calculate weight of the vector based on the TF-IDF ranking. The cosine similarity measure used to calculate similarity between the document vector and query vector. The disease based similarity measure is used in the proposed method. The outcome of the proposed project is represented through graphical visualization, thereby providing better recommendation solution projecting the high similarity symptoms.**

Keywords*— *Big data, Hadoop, Vector Space Model (VSM), TF-IDF, Latent Dirichlet Allocation, Cosine Similarity.

I. INTRODUCTION

Big data analytics is the process of analyzing massive volume of data or big data that can be gathered from variety of sources like social networks, videos, digital images and etc. It contains both structured and unstructured data. The big data analytics used to predict the hidden patterns and other some useful information to the created users. Traditional systems may small in size and using some traditional software techniques, because it unable to analyze as large volume of data. But the big data can't handle by the traditional software techniques and tools. The sophisticated software programs like Hadoop, MapReduce and NoSQL databases are emerged with the big data. It increases the storage capacity, processing power and availability of data.

Big data can be processed with the software tools frequently used as part of advanced analytics control such as predictive analytics, data mining, text analytics and analysis. Big data analytics that holds the process of examining a large amount of data that contains a variety of data types. The main process of big data analytics is to split up the data in various

systems and create an environment that gathers all the unstructured data to a structured form.

In recent analysis big data help researchers to convert human DNA in minutes, predict where terrorists plan to attack and determine which gene is mostly likely to be responsible for certain diseases. This big data analytics helps to discover various kinds of activities pertaining to customer's behavior, web server logs, survey response, etc.

The characteristics of big data analytics can be explained in following:

Volume (size of the data set):

Volume is one characteristic which needs to be considered while dealing with big data. Volume means to the large amounts of data that are generated every seconds. Size of data performs very difficult role in discovering value out of data.

Variety (based on data types):

Variety refers to heterogeneous sources and the nature of data, both structured and unstructured together such as in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. This variety of unstructured data poses certain issues for storage, mining and analyzing data. The wide variety of data requires a different approach as well as different techniques to store all raw data.

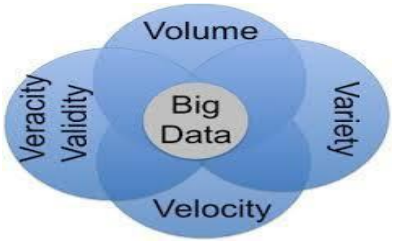

Fig. 1.1 4 V's of the Big data analytics.

Velocity (speed of data):

The Velocity is the speed at which the data is created, stored, analyzed and visualized. Technology allows us now to analyze the data while it is being generated without ever putting it into databases. Computers and servers required substantial time to process the data and update the databases.

II. RELATED WORK

There are various sectors in which big data analytics is used in research purpose. Data in agriculture sector is growing at each and every second and it also enter in the era of big data. In the year of 2015 IBM introduced agricultural big data analytics.

Related research in various field:
"*Agro Advisory System For Cotton Crop*" **Sanjay chaudhary** [2015] describes the power of ICT to improves the agricultural practices and there by the production. An agro advisory system presented in this paper helps to bridge the gap between farmers and the agriculture domain experts. The system consists of three basics components are cotton ontology, web services and mobile application development.

"*Crop Management Using Big Data*" **Michel le page** [2015] discuss about the crops are not getting enough water by rainfall or natural drainages. The paper provides the application of the exact amount of water needed to the plants requires the knowledge `numerous parameters such as soil conditions, plant development and weather.

"*Crop–planning, making smarter agriculture with climate data*" **Fabian mejia** [2015] illustrate the consequence of climate change and climate variability modifying established practices for traditional crops. It is designed to offer easy access to relevant data and updated crop calendars form farmers and also share management's practices from local authorities.

"*New Optimized Spectral Indices For Identifying And Monitoring Winter Wheat Disease*" **Wenjiang huang** [2014] explains indices from hyper spectral data have been shown to be effective for indirect monitoring of plant disease on crops. We aimed to develop new spectral indices that would be useful for identifying different disease and crops. The most and least relevant wavelength for different diseases where first extracted from leaf spectral data using the RELIEF-F algorithm.

III. METHODOLOGY

A. Notations and Preliminaries

D - The dictionary, with the set of keywords, denoted as D= $\{w_1,w_2,....,w_z\}$.

Z - The total number of keywords in *D*.

D_x - The subset of D, representing the keywords in the query.

P - The document collection, denoted as a collection of u documents. P= $\{p_1,p_2... p_u\}$. Each document pin the collection can be considered as a sequence of keywords.

U - The total number of documents in P.

S - The normalized document collection stored in the HDFS server, denoted as S= $\{s_1,s_2,..,s_u\}$.

B. Proposed Architecture

The proposed architecture of the current work is depicted in figure 2. We collect data from heterogeneous websites like Agro net, Agropedia and agriculture info websites. The collected data's are preprocessed by three different techniques and stored in a HDFS.

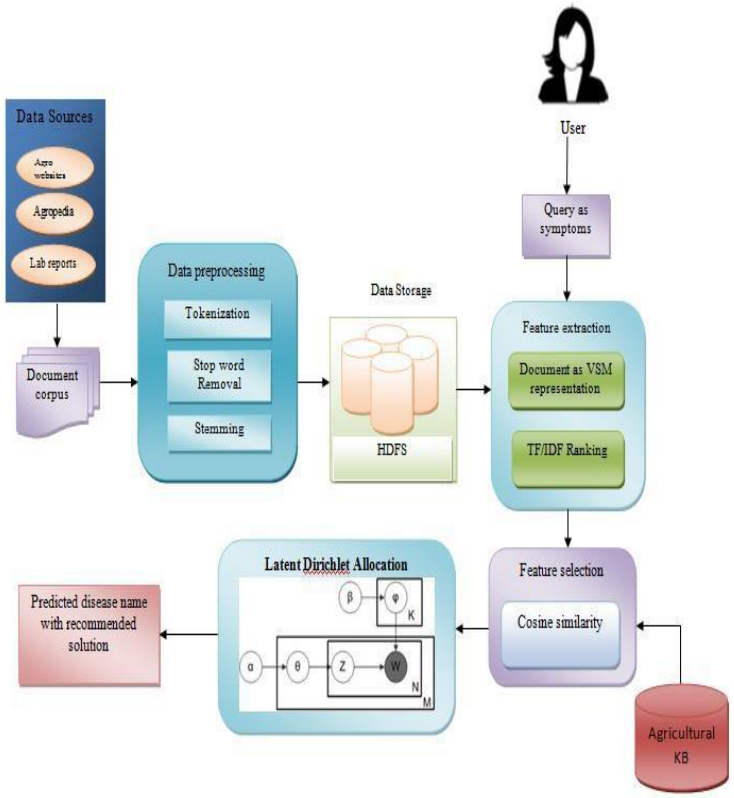

Fig. 3.1 Architecture of oryza recommendation system

When the users give the query in form of symptoms of paddy crops are evaluated from the cultivated lands. Retrieve the documents based on the symptoms and converts into vectors.TF-IDF ranking method used to calculate the weight of the term present in the document and query. Compute the similarity between the document and query vector using cosine similarity techniques.

Finally the recommendation system predicts the disease name and suggests solution based on predictive analysis method of latent Dirichlet allocation.

Data sources are:
Agropedia Blog:

It is a crucial source of data for researchers and knowledgeable persons. The blog holds the data for cereal

crops. It contains a very wide variety of concepts like climate, diseases, Pests, Irrigation, and Pesticides etc.

Agriculture info websites:

These websites act like mentor for farmers. These sites give information related to agricultural economic entity; commonly used pesticides etc. agriculture information websites provide information to farmers about which crop to plant where and when. And suggest solutions to various problems related to crops. By these sites farmers get knowledge about new techniques and tools.

Agriculture department reports:

Using these reports decision making is easy for crops of particular area. These reports are important to provide information regarding particular field of a geographical area.

Data that is collected from above sources is stored on Hadoop distributed file system in the form of text file. Collected data is unstructured and it contains irrelevant data.

Firstly unimportant data is removed and relevant data is extracted from collected data.

Then features are selected and extracted from relevant data and save into text file on hive data warehouse. Hive is used to querying the data in distributed environment. Hive is open source software tool used for data ware housing. To extract data out from Hadoop system Hive provides interface that is similar to SQL interface which is termed as (HIVEQL) HIVE query language.

D. Vector Space Model (VSM)

Vector space model is widely used in Information Retrieval system for document representation. The plaintext data is efficiently represented as multi-dimensional vectors. Similar to the documents, the queries are also denoted as vectors. Term Frequency and Inverse Document Frequency (TF-IDF) is a technique used in this model.

a. Term Trequency and Inverse Document Frequency(TF×IDF)

Term frequency (TF) is simply the number of times a given term or keyword appears within a document, and inverse document frequency (IDF) is obtained through dividing the number of documents in the whole collection by the number of documents containing the term.

$$TTFF_{pp,ii} = \frac{TTFF'_{pp,ii}}{\sqrt{\sum_{ii} (TTF_{pp,ii})^{22}}} \quad (1)$$

where denotes the number of times the word w_i appears in the document p, where $TTFF'_{ww} = 1 + \ln TF_{p,w_i}$

$$IIFF_{ii} = \frac{IIFF'_{ii}}{\sqrt{\sum_{ii} (IIFF'_{ii})^{22}}} \quad (2)$$

where ID_{ww} denotes the inverse document frequency

$$\text{Relevance Score } S = \sum_{ii \in DD} TTFF_{pp,ii} \times IDFF_{ii} \quad (3)$$

Based on the relevance score the documents are ranked in ascending order. For each word, the documents are ordered to form the index. This index score is used for efficient searching.

E. Cosine Similarity

In this paper, we use the cosine similarity measure together with relevance score to provide accurate ranking. In information retrieval, a ranking function is used to calculate relevance scores of matching files to a given search request. The most widely used statistical measurement for evaluating relevance score in the information retrieval community uses the TF × IDF rule. The similarity between query and document are calculated using Cosine function. In equation 4, the similarity function is given.

$$Sim(q,d_u) = \frac{\sum_{j} UUSS_{j,jj} . SS_{j,j}}{\sum_{j=} \sum_{} (\sum^{UU} (SS_{jj})^2) SS (\sum^{UU} (SS_{jj})^2)} \quad (4)$$

where S_p denotes the relevance score of document p and S_q denotes the relevance score of query.

F. Latent Dirichlet allocation

In natural language processing, latent Dirichlet allocation (LDA) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. Each document may be viewed as a mixture of various topics where each document is considered to have a set of topics that are assigned to it via LDA.

This is similar to probabilistic latent semantic analysis , except that in LDA the topic distribution is assumed to have a Dirichlet prior. In practice, this results in more reasonable mixtures of topics in a document. It has been noted, however, that the probabilistic LSA model is equivalent to the LDA model under a uniform Dirichlet prior distribution.

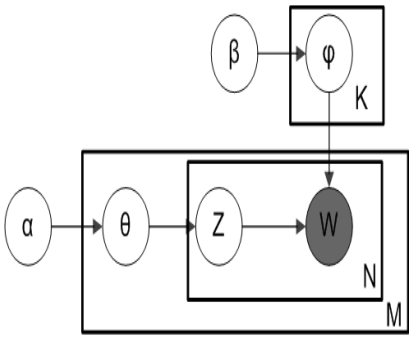

Fig. 3 Latent Dirichlet Allocation

of the word $w_{i,}$, where = ln (1+U/U_{wi}) and N denotes

*II*ئچئچ

the total number of documents. Using TF, IDF values, relevance score is calculated to assign the weight of each term in the document.

α is the parameter of the Dirichlet prior on the per-document topic distributions.

β is the parameter of the Dirichlet prior on the per-topic word distribution.

IV. IMPLEMENTATION

In order to find the similarity of query keyword with the existing document, the vector space model is used. In vector space model both documents and query is represented in the form of vector.

For example,

Text 1: paddy leaf in deep yellow

Text 2: Paddy crops affected by pathogen

Keywords: {Paddy, crops, leaf, pathogen, affected, deep, yellow, in, by}

Number of times each keyword exists in the text as:

Paddy 1 1
Crops 0 1
Leaf 1 0
Pathogen 0 1
Deep 1 0
Yellow 1 0
Affected 0 1
In 1 0
By 0 1

The two vectors correspond to the text are:

j:[1,0,1,0,1,1,0,1,0]

d:[1,1,0,1,0,0,1,0,1]

Using vector space model an 8 dimensional vector is formed.

To retrieve the top-K documents, Cosine similarity is used.

For example:

Consider a small collection of three documents:

D 1:"brown plant leaf hopper"

D 2:" Leaf blast "

D 3:"paddy leaf "

Step 1:Calculation of term frequency

For all the documents, calculate the relevance score for all the terms in C. Assign the score 1 if the keyword appear in that particular document, otherwise assign 0:

TABLE 4.1. TERM FREQUENCIES

	Brown	Plant	Leaf	Blast	hopper	Paddy
Document1	1	1	1	0	1	0
Document2	0	0	1	1	0	0
Document3	0	0	1	0	0	1

Step 2: Calculation of Inverse Document frequency

Number of documents, N=3.therefore, the IDF values are:

Brown $\log_2(3/1)=1.584$

Plant $\log_2(3/1)=1.584$

Leaf $\log_2(3/3)=1$

Blast $\log_2(3/1)=1.584$

Hopper $\log_2(3/1)=1.584$

Paddy $\log_2(3/1)=1.584$

Step 3: Calculation of TF×IDF values

Multiply the TF score by the IDF values by each term.

TABLE 4.2. INVERSE DOCUMENT FREQUENCIES

	Brown	Plant	Leaf	Blast	Hopper	Paddy
D1	1.584	1.584	1	0	1.584	0
D2	0	0	1	1.584	0	0
D3	0	0	1	0	0	1.584

Step 4: Calculation of Cosine Similarity function :

Let the given query be: "very very times", calculate the TF×IDF score for the query, and compute the score of each document in C relative to this query, using the cosine similarity. When computing the TF×IDF values for the query terms, divide the frequency by the maximum frequency and multiply with the IDF values.

The cosine similarity between any two documents are calculated as

Cosine Similarity(D1,D2) = Dot product (D1,D2)/||D1|| * ||D2||

Dot product (D1,D2) = D1 [0] * D2 [0] + D1[1] * D2[1] * … * D1[n] * D2[n]

||D1|| = square root (D1[0]2 + D1[1]2 + ... + D1[n]2)

||D2|| = square root (D2[0]2 + D2[1]2 + ... + D2[n]2)

The query entered by the user can be represented as a vector form

k [0 0 (2/2)*1.584=1 0 (1/2)*1.584=0.792 0]

The length of each document and of the query:

Length of D1 = sqrt(1.584^2+1.584^2+1^2+1.584^2) =2.9201

Length of D2 = sqrt(1^2+1.584^2) =1.9799

Length of D3 = sqrt(1.584^2+1^2) =1.9799

Length of k = sqrt(1.584^2+0.792^2) =1.7709

Then the similarity values are:

Cosine Similarity (D1,k)=(0*0+0*0+1.584*1.584+0*0+1.584*0.792+1.584*0) / (2.9201*1.7709) = 0.7277

Cosine Similarity (D2,k)=(0*0+0*0+1.584*1.584+1.584*0+0*0.792+1.584*0) / (1.9799*1.7709) =0.7156

Cosine Similarity (D3, k) = (1.584*0+1*0+0*1.584+0*0+1.584*0.792+0*0) / (1.9799*1.7709) = 0.3578

According to the similarity values, the final order in which the documents are presented as result to the query will be: D1, D2, and D3.

A. Latent Dirichlet Allocation

We take a corpora of K documents, each representing i documents (e..g tweets), such that will be count of all words in corpora in total of d documents.]

Steps for latent Dirichlet allocation

Step 1: Select ~ Dirichlet (α) where i Є 1... D

Step 2: Select ~ Dirichlet (β) where i Є 1...K

Step 3: For each word where

Choose a topic ~ Multinomial ()

Choose a word ~ Multinomial ( )

Where,

α is the parameter of the Dirichlet prior on the per-document topic distributions.

β is the parameter of the Dirichlet prior on the per-topic word distribution.

Θ_m is the topic distribution for document i.

$\phi\phi_t$ is the word distribution for topic k.

This model writes the corresponding profiled interests from which we can infer undetected topics and user interest topics through learning of model parameters. LDA finds a pre-specified set of |Z| topics within |D| documents. Each term t in a tweet with Ki terms then ends up correlated with a topic z.

Z = {z1, z2, z3,...., }is the set of n latent topics which exemplifies coarseness and resulting final set of topics.

V. PERFORMANCE ANALYSIS

We implemented the proposed model based on two techniques, Latent Dirichlet Allocation technique and Vector Space model (VSM). In the VSM used to calculate the weight of the term that present in the document and query. The vectorized terms in query are compared with the document to find a similarity. Latent Dirichlet Allocation can be used to retrieve the topics from the mixture of documents to predict the disease name and provide the prevention methods such as harmless fertilizers.

VI. RESULT AND DISCUSSION

Considering the developed system and testing performed, we have a system in place is that is accessible from the system without internet connection. The HIVESQL database deployed in the HDFS server. Finally it provide symptoms based disease name and recommend harmless fertilizers.

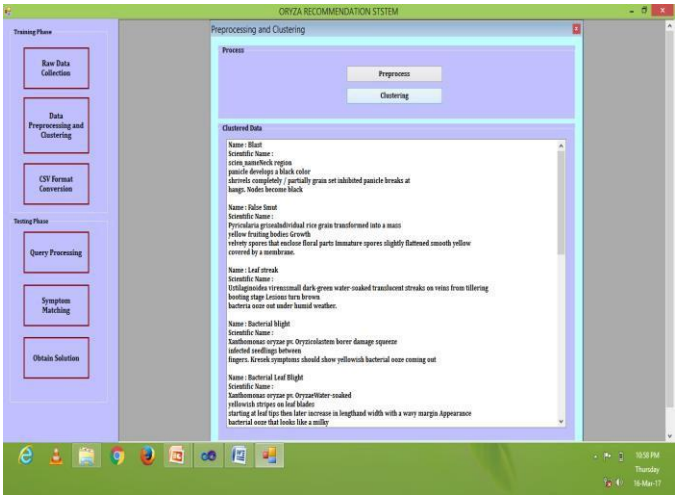

Fig 6.. Screen Shot For Data Preprocessing

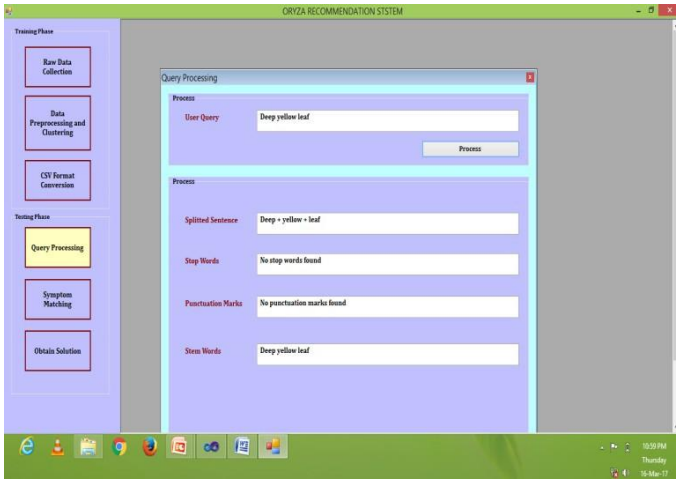

Fig 6.2. Screen Shot For Query Processing

VII. CONCLUSION AND FUTURE WORK

In this paper, we propose an efficient recommendation system to predict the symptoms based paddy diseases with high similarity. Among various multi-keyword search techniques we choose an enhancing one, i.e., vector space model to present the relevance between documents and symptoms. The cosine similarity measure is used to quantitative the similarity between documents related to the symptoms and query. It further achieves an accurate ranked search results. The further enhancement of our ranked search mechanism involves more search semantics. In future we have to improve protection of paddy crops. Instead of latent dirichlet allocation we can use another predictive analysis techniques like deep learning algorithm.

References

[1] Big data in agriculture. Available from: http://www.citethisforme.com/topic-ideas/technology/'Big%20Data'-6678234

[2] Chen X-W, Lin X. Big data deep learning challenges and perspective. IEEE Access. 2014 May; 2:514–22.

[3] Gates A. Programming pig. 1st ed. California, USA: O'reilly Media Inc; 2011.

[4] Hive performance benchmark. Available from: https:// issues.apache.org/jira/browse/HIVE-396

[5] Hunt P, Flavio P. ZooKeeper: Wait-free coordination for internet-scale systems. Proceedings of Usenix Annual Technical Conference; MA. 2010 Jun. p. 11–29.

[6] Introducing apache mahout: Scalable, commercial-friendly machine learning for building intelligent applications. Available from: http://www.ibm.com/developerworks/ library/j-mahout/

[7] Lam C, Warren J. Hadoop in action. 1st ed. Greenwich: Manning Publications; 2010 Dec.

[8] Laney D. 3D data management: Controlling data volume, velocity and variety. Meta Group Inc Application Delivery Strategies; 2001 Feb. p. 1–4.

[9] Lu Q, Li Z, Kihl M, Zhu L, Zhang W. A conceptual framework for big data analytics applications in the cloud. IEEE Access. 2015 Oct; 3(1):944–52.

[10] Marx V. Biology: The big challenges of big data. Nature. 2013 Jan; 498(7453):255–60.

[11] My smart farm. Available from: https://www.kickstarter. Com/projects/1911579744/my's-smart-farm-active-food producer.