A K-Medoid Centered Efficient Knowledge Retrieval System for Agriculture

Isha Kumari* , Dr. Mukesh Kumar**
*Department of Computer Science & Engineering
Rabindranath Tagore University, India
Email- ijha6322@gmail.com
**Department of Computer Science & Engineering
Rabindranath Tagore University, Raisen, Madhya Pradesh, India
Email- goutam.mukesh@aisectuniversity.ac.in

Abstract-

The agricultural information on the internet is increasing information. There is a need for an intelligent information retrieval system that could handle complications like low relevancy rate, unstructured data format, and high computation time. To offset these issues, an Information Retrieval (IR) framework is developed based on farmers specific crop ontology and machine learning techniques. The main objective of this research is to implement fast and effective agricultural information retrieval by incorporating clustering and classification techniques. Farmer-based crop ontology is developed, which will be useful for guiding the farmers by providing instructions and information related to crop cultivation and management of fertilizers based on their soil and crop developing stage. details, and disease structure for paddy crops are gathered. The developed CropOnt Ontology adapted the METHONTOLOGY approach and implemented it in the Protege tool. The performance of the developed

Protocol and RDF Query Language (SPARQL) queries in the Protege SPARQL tab with sample triple The information retrieval s (which means construct, cluster, and classify) algorithm. The intended algorithm has three modules, such as Construction of hash tree, applying clustering techniques, and classification of clustered records. The hash tree is constructed using Closed Frequency Patterns (CFP). After constructing a hash tree for the hash codes, the leaf nodes of the hash tree are clustered using K- Medoid clustering and classified by using the Adaptive Neuro-Fuzzy Inference System(ANFIS). Instead of K-medoid clustering, the proposed Inter Quartile Pruning Range-Hierarchical Divisive Clustering (IQPR-HDC) hybrid clustering method is used to cluster the hash codes and analysed the performance measures. For a given user query, CFP and hash code is calculated, and based on hash codes the results are retrieved from the database.

The proposed system performance is measured and compared with the existing Steiner Tree(ST), IBRI-CASONTO, and Bidirectional-Long Short-Term Memory (BILSTM) techniques. The experimental results reveal that the proposed approach achieved a maximum accuracy for simple queries which have a maximum of three key parameters like crop variety, farmer id and village as well as for complex queries which have complex conditions with more than three key parameters. It is evident that recall, precision, f- score, and accuracy have good values when compared with existing methods. The proposed system using IQPR-HDC for clustering outperforms well in terms of accuracy and relevancy when compared with K-medoid clustering. Farmers will get benefitted from obtaining accurate and relevant information from the proposed information retrieval systems that will help them to make decisions..

Keywords—Agriculture information System,

1. INTRODUCTION

In India, agriculture has a diversity of natural resources such as wider cultivable land, different agro-climatic zones, different soil types, and suitable seasons for cultivating a variety of crops. Indian agriculture is largely reliant on natural resources and suitable weather conditions. Therefore, under a vast agricultural diversity, strong information support and knowledge system are required for the farmer community to achieve a high yield. Accurate and timely information and instruction are vital to the agricultural processes that should be delivered to farmers properly. The available agricultural websites, mobile applications, and software provide general practices to the farmers and fail to offer the solution for specific problems. It is desirable to incorporate agricultural knowledge into a system that understands the requirements of farmers and provides them with the best solutions. So there is a need for an Information Retrieval (IR) system that should give contextual information to the farmers based on geographical area, climatic condition, soil nature, previous experience and current state of the crop.

Significance of Ontology in IR

Ontology was a philosophical concept in the beginning and later became academic concept through knowledge engineering. One of the most conventional concepts of ontology is that it is a categorical prescribed description of common components. It is a validated and indicative representation that involves the terminology to refer the expressions in the specific area, depict what the terms are and how they are interrelated with one another, and

also how a few cannot be related. Hence, Ontology gives a terminology for signifying and collaborating knowledge on certain topics and the relationship they have amongst the terms.

Ontology representation

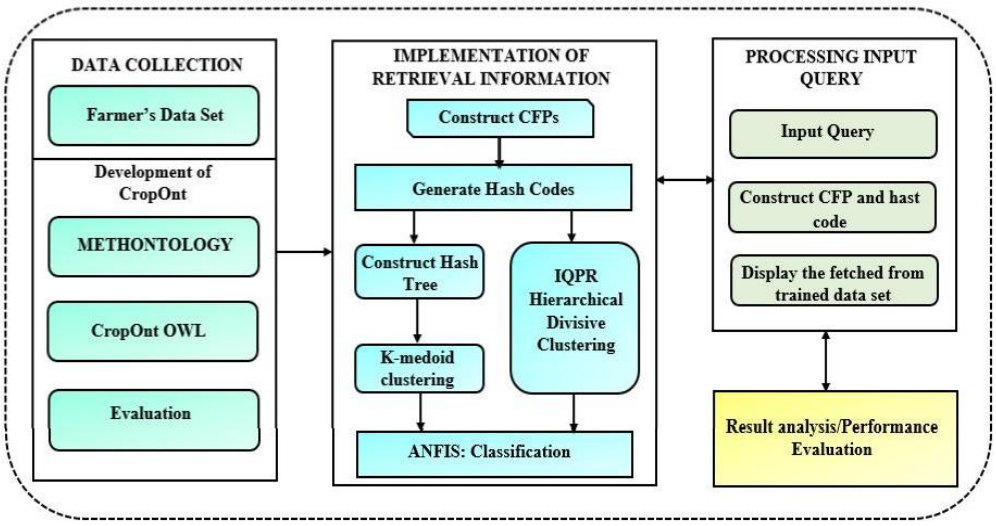
Ontology comprises four major mechanisms: concepts, instances, relations, and axioms (Mohammad Mustafa Taye 2010). A notion also called as a category is an intangible group, set or assortment of objects. It is the central element of the region and generally embodies a group or class wherein the affiliates have a stake in communal properties. This module is signified in categorised graphs so that it appears to be akin to object-directed systems. The notion is demonstrated by super-class and it embodies the upper class or alleged parent class that signifies the subsidiary or alleged child class. An instance (called individual too) is the bottom-level element of an ontology that embodies a particular object or element of a notion or category.

An alliance or relation is used for expressing connection between two perceptions in a said dominion. In particular it depicts the connection between the initial concept indicated in the domain and the next indicated in the range. An axiom that is utilized for imposing restraints on the principles of classes or instances are typically stated using rational centred lexicons like first order logic that is utilized to authenticate the ontology.

Resource Description Framework (RDF) is a W3C standard (RDF Working group, 2004) for depiction of metadata and demonstrating the web resources. The format used is XML format. RDF offers a supple technique to

Crumble any know-hows into triplicates with certain principles on language annotation of those segments. (Leslie F Sikosand Dean Philp 2020) triples are an amalgamation of tri-variables namely Subject, Predicate and the Object. While the objects are forever verbatim, the subject and predicates on the other hand are resources. Resources are recognised by Uniform Resource Identifiers (URI) when the information is from a web source and a group of reports when the information is from documents or interpersonal databases. One drawback of RDF is its restricted vocabulary of conveying information concerning triples.

Ontology-based information retrieval system is proposed specifically in the agriculture domain. The development of Crop ontology with terminologies of crop. The creation of complete farmers cultivation database. The proposed architecture of the IR system using crop ontology and 3c's algorithm is shown in Figure 1.Construction, Cluster and Classification process are applied on the agricultural data and in short it is represented as 3c's algorithm



A. Figure 1 Proposed Architecture for Effective Information Retrieval

By using the concepts and properties defined in Crop ontology, patterns are identified and frequency of each pattern is calculated to generate the Closed Frequency Patterns (CFP). Machine learning algorithms are used to organize agricultural information so that relevant information can be retrieved. The crop cultivation information is retrieved from farmer's database based on CFP and divided into clusters using K-medoid and Inter Quartile Pruning Range Based Hierarchical Divisive Clustering (IQPR-HDC) algorithms. Adaptive Neurofuzzy

Inference System (ANFIS) is a type of artificial neural network classification algorithm based on the Takagi-Sugeno fuzzy inference system which is used to classify the clusters.

The K-Medoids algorithm begins with arbitrarily selecting data items as first medoids to signify the clusters. The entire remaining items are comprised in a cluster that holds its medoid nearest to them. After that, a new medoid is ascertained, which can signify the cluster better. The entire remaining data items are nevertheless allocated to the clusters encompassing the closest medoid. In every one of the iterations, these medoids vary their site. The technique limits the sum of the dissimilarity betwixt all data items and their equivalent medoid. This cycle is recurring until no medoid varies its placement. This marks the finish of the process; in addition, the resulting last clusters with their medoids are attained. Clusters are built centred on the medoids and all the data members are deployed in the suitable cluster centred on the nearest medoid. The K-Medoid Algorithm steps are elucidated below:

Step 1: Let $C=\{C_1, C_2 \ldots C_j\}$ be the set of cluster heads.

Step 2: Assign each cluster node to the nearest cluster heads to get the clusters. Now, Euclidean distance is utilized to calculate the distance between the cluster heads and the other cluster nodes

Step 3: Next, randomly select a cluster node C_{random} to replace cluster head node C_j, with the condition that the residual energy of this random node must be higher than the average residual energy of all nodes.

Step 4: Now, assign the cluster nodes to the nearest new cluster head and obtain the clustering result. Then, calculate the Euclidean distance between the cluster heads and the other cluster nodes using the Equation 5.11. And calculate the sum of distances for all cluster nodes to the cluster heads.

Step 5: If the sum of distances is equal to the previous sum of distances, then stop the algorithm. Otherwise, go back to Step 3.

Now, the clustered values are given as an input for further classification. The pseudo-code for the K-Medoid clustering is staged in Figure 2.

Input : $C(F_i)$, the number of closed frequent itemset
Output : K, clusters

Begin
Randomly choose K data points from E as medoids m_k
Repeat
Assign remaining non-medoid data points to its closest medoid
Compute total distance TD_i between medoids m_i and non medoids e_j
For each medoids do
Select the non-medoid e_j for which the total distance TD is minimal
Compute $TD(e_j \rightarrow m_i)$
If $TD(e_j \rightarrow m_i)$ is smaller than the current TD_i
Swap m_i and e_j
End if
End for
Until no changes
End

Figure2 Pseudocode for K-Medoid Algorithm

The newly suggested system run in the JAVA platform and further experiments were executed on a personal computer that contains Windows 7 Operation system (3.20GHz dual-core computer containing 4 GB RAM working in Intel i5/core i7 processor). Since this platform meets the challenges like efficiency, accuracy, and security, its features are enhanced. Java has a set of packages that will solve all the requirements from the user side.

The input data, such as crop and the farmer's details for 3 years are collected from various resources and represented as a database in Excel Sheet. The details of the farmer, like a Farmer's name, address, and the details of the crop field, like a crop type, duration of crop growth, location of the crop, etc. are collected. The proposed system

takes the rice crops. Crop Ontology is represented using Protégé tool and package for visualizing the concepts are also included.

PERFORMANCE METRICS

Performance parameters such as precision, F-score, True Positive Rate (TPR), recall, precision, accuracy, and False Positive Rate (FPR) are used as a measure in analyzing the performance. The analysis of the performance is done by determining a discrete number of performance parameters. The confusion matrix includes True Positive (P), False positive (FP), False Negative (FN), and True Negative (TN) which are most commonly utilized for the purpose of comparison. The detailed description of each of these parameters is explained below:

True positive (TP): the number of relevant records retrievedcorrectly.

False-positive (FP): the number of records improperly fetchedas relevant.

True negative (TN): the number of records exactly recognizedas irrelevant.

False-negative(FN):the number of records improperlyrecognized as irrelevant.

Precision: Precision (P) is the ratio between relevant retrieved records andretrieved records.

a) Recall:

The recall is defined as the ratio of the number of items retrieved that are relevant to the user and the number of possibly relevant documents that are available in the database. Recall measures how well a system processes a particular query to provide related items that the user wants to see. It is a very useful concept but the denominator cannot be calculated in operational systems. If the system is equipped with total number of relevant entries in the database, then recall is given by

These notions can be made clear by examining the followingcontingency Table 6.1.

B. Table 2 Matrix Table

	Relevant	**Non-relevant**
Retrieved	true positives (tp)	false positives (fp)
Not retrieved	false negatives (fn)	true negatives (tn)

The importance of precision and recall are varied in different scenarios, since these parameters compromise each other in measuring the performance of the IR. Common web surfers would expect every result on the first page to be relevant (high precision) but not have the least curiosity in glancing at every document that is related. In contrast, various proficient searchers like researchers and information analysts are very concerned to get high relevant documents (high recall) as possible and accept comparatively low precision results. Nevertheless, the two quantities clearly trade-off against one another: recall can always be 1 (but very low precision) by retrieving all documents for all queries. The recall is a non-decreasing function of the number of documents fetched whereas precision usually decreases as thenumber of documents retrieved is augmented.

Accuracy: It is the ratio of the true outcome, which is sum of TN and TP. It measures the amount of accurate finding of the system by correctly identifying the relevant document that is retrieved.

FNTN In this sector, the intended 3c's is compared with the prevailing techniques of IBRI-CASONTO, BILSTM, and Steiner tree (ST) regarding precision, recall, f-score, accuracy, returned vs. effective information, retrieved results, and query retrieval time. Such measures are matched for the divergent kinds of input queries, like simple and also complex queries. The aquired outcomes of recommended 3c's technique and prevalent methods are presented in table 6.4.

C. Table 3 Performance outline of suggested 3c's with prevailing practices for simple and Complex Queries

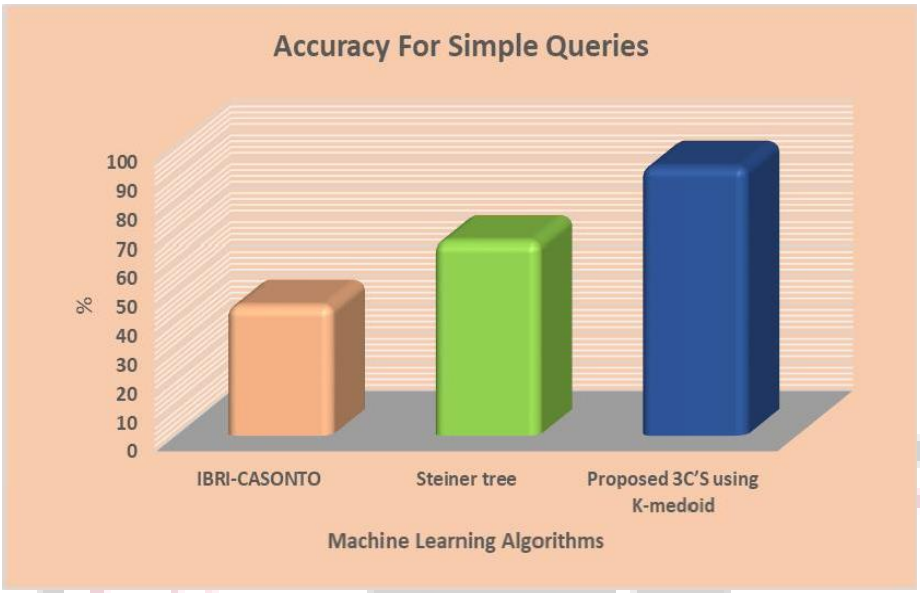
(a)

Performan ceMetrics (%)	**Simple Queries**			
	IBRI-CASONTO	**Steinertree**	**BILSTM**	**Proposed using 3 C's using K-medoid**
Precision	52.49	95.2	98.78	96.56
Recall	75	80.8	98.96	82.77
F-score	68.54	74.32	98.54	84.45
Accuracy	45	67	-	92.77

Table 3 ((b)

Performance Metrics (%)	**Complex Queries**			
	IBRI-CASONTO	**Steiner tree**	**BILSTM**	Proposed using 3 C's **using K-medoid**
Precision	50.12	93.562	89	94.45
Recall	73.56	78.34	72	81.45
F-score	66.34	72.34	79.64	83.44
Accuracy	45	67	-	91.45

Table 3 mentioned above presents the comparison outcomes of suggested 3C's with prevailing IBRI-CASONTO, BILSTM and ST for uncomplicated and intricate queries regarding precision, recall, f-score, and accuracy. For simple and intricate queries, the precision value of the recommended 3C's is 96.56 and 94.45 where the prevailing IBRI, CASONTO, BILSTM and ST provide 52.49, 98.78, and 95.2 for simple queries and 50.12 and 93.56 for intricate queries. The BILSTM is higher performance for simple queries compare to the recommended technique. Figure 6.7 shows the performance of the proposed 3C's with the existing IBRI-CASONTO and Steiner tree based on accuracy metric. Accuracy denotes the correctness of the information retrieval process. The proposed system retrieves 92.77% accurate information in simple queries but the existing IBRI-CASONTO, and Steiner tree have 45 %, and 67 %, accuracy respectively. Moreover, the intended 3C's algorithm has 91.45% accuracy values for complex queries, whereas the prevailing methods have much lesser than the intended method. Shows that the proposed crop information retrieval system using 3c's Thus, the graphical representation achieves higher accuracy than the existing classifiers.



D. Figure 4 Performance Assessment of suggested 3c's with Prevailing procedure regarding Accuracy for (a)Simple Queries and (b) Complex Queries

Table 4 Comparison Table for Returned Vs Related Information forSimple Queries

	Simple Queries			
	Existing IBRI-CASONTO	Existing Steiner tree	Existing BILSTM	Proposed 3C's using K-medoid
Returned Information	23	27	26	28
Related Information	17	23	25	26

The performance of intended 3C'S and prevailing techniques regarding returned Vs effective information, retrieved outcomes, and query retrieval duration is elucidated in Figure 6.7. Figure 6.7 demonstrate the suggested 3C's and prevailing IBRI-CASONTO, BILSTM and ST concerning the measures (a) returned Vs effective information (b) retrieved outcomes, and (c) query retrieval time. Based on 30 records, the recommended 3C's returns 28 records; in that, 26 records are dynamic whereas the prevailing IBRI-CASONTO, BILSTM and ST return 23, 26 and 27 records. But on those records, 17 and 23 records are dynamics. Utilizing the 3c's the system recover 92% of information about crop but the recovered information of the prevailing IBRI-CASONTO,prevailing IBRI-CASONTO, BILSTM and ST takes 12450ms, 10829ms and 11670ms. In this, the IBRI-CASONTO absorbs more duration for retrieving equated with the prevailing IBRI-CASONTO and ST. Hence, the complete 3c's performance is greater for all associated measure.

2. CONCLUSION

In this research we have discussed the implementation of the information retrieval system using 3c's algorithm based on k-medoid clustering. The performance of this system is assessed by comparing it with IBRICASONTO and Steiner Tree. It is evident that recall, precision, f-score, and accuracy have good values when compared with existing methods. The proposed system using information retrieval system using 3C's algorithm based on k-medoid clustering for clustering outperforms well when compared with K-Medoid clustering and Crop ontology development (protégé tool). The table and graph are clearly shows that the proposed method is higher complex queries than the prevailing techniques. However, for simple queries, the

existing BILSTM outperforms with higher precision than the proposed system due to use of less number of key parameters in query and complexity level of the system is low due to non-maintenance of meaningful relationship among entities. The experiment results clearly demonstrated that the proposed system is helpful for the farmers to fetch timely, accurate and required information.

3. REFERENCES

[1].Swinehart, Df 1962, „The Beer-Lambert Law", Journal Of Chemical Education, Vol. 39, Issue 7, 33,. Syamasudha Veeragandham & Santhi H 2020, „A Review On The Role Of Machine Learning In Agriculture", Scalable Computing: Practice And Experience", Vol. 21, Issues. 4, Pp. 583–589, Doi: 10.12694:/Scpe. V21i4.1699.

[2].Sharma, A, Jain, Gupta, P & Chowdary,V 2021, „Machine Learning Applications For Precision Agriculture: A Comprehensive Review",Ieee Access, Vol. 9, Pp. 4843-4873, Doi:10.1109/Access.2020. 3048415.

[3].Sharma, R, Kamble, Ss, Gunasekaran, A, Kumar, V & Kumar, A 2020, „A Systematic Literature Review On Machine Learning Applications For Sustainable Agriculture Supply Chain Performance", Computers And Operation Research,Vol. 119,. Sharma, A, Jain, Gupta, P & Chowdary,V 2021, „Machine Learning Applications For Precision Agriculture: A Comprehensive Review",Ieee Access, Vol. 9, Pp. 4843-4873, Doi:10.1109/Access.2020. 3048415.

[4].Revati, P, Potdar, Mandar, M, Shirolkar, Alok, J, Verma, Pravin, S, More & Atul Kulkarni 2021, „Determination Of Soil Nutrients (Npk) Using Optical Methods: A Mini Review", Journal Of Plant Nutrition, Vol.44, Pp.1826-1839, Doi: 10.1080/01904167.2021.1884702.

[5].Mohammad Nishat Akhtar, Abdurrahman Javid Shaikh, Ambareen Khan, Habib Awais, Elmi Abu Bakar & Abdul Rahim Othman 2021, „Smart Sensing With Edge Computing In Precision Agriculture For Soil Assessment And Heavy Metal Monitoring: A Review,„ Agriculture, Vol. 11, Pp. 1-37,

[6].Hatture, Sm & Yankati, Pv 2021, „Iot-Based Smart Farming Application For Sustainable Agriculture", Ict Systems And Sustainability, Advances In Intelligent Systems And Computing, Vol.1270,

[7].Karishma Mohiuddin & Mirza Mohtashim Alam 2019, „A Short Review On Agriculture Based On Machine Learning And Image Processing", Acta Scientific Agriculture, Vol. 3, Pp.55-59.

[8].Kulkarni, N, Thakur & Rajwal, At 2019, „Smart Soil Nutrients Analysis And Prediction Of The Level Of Nutrients Using A Bot", 3rd International Conference On Recent Developments In Control, Automation & Power Engineering. Ieee, Pp. 663-668.

[9].Jitendra 2020, „Economic Survey 2019-20: Agriculture Growth Stagnant In Last 6 Years", Https://Www.Downtoearth.Org.In/News/Agriculture/ Economic-Survey-2019-20-Agriculture-Growth-Stagnant-In-Last-6-Years.

[10]. Jirapond Muangprathub, Nathaphon Boonnam, Siriwan Kajornkasirat, Narongsak Lekbangpong, Apirat Wanichsombat & Pichetwut Nillaor 2019, „Iot And Agriculture Data Analysis For Smart Farm", Computers And Electronics In Agriculture, Vol. 156, Pp. 467-474, Issn 0168-1699, Doi:Https://Doi.Org/10.1016/J.Compag.2018.12.011.