Engineering Journal issn: 0125-8281

Article

Prescription Based Recommender System for Diabetic Patients Using Efficient Map Reduce

Ritika Bateja[1,a,*], **Sanjay Kumar Dubey**[1,b], **and Ashutosh Bhatt**[2,c]

1 Department of Computer Science and Engineering, Amity University Uttar Pradesh, Noida, India
2 School of Computer Sciences and Information Technology, Uttarakhand Open University, Haldwani, Nainital, Uttarakhand, India
E-mail: [a,*]ritikalpr@gmail.com (Corresponding author), [b]skdubey1@amity.edu, [c]ashutoshbhatt123@gmail.com

Abstract. Healthcare sector has been deprived of leveraging knowledge gained through data insights, due to manual processes and legacy record-keeping methods. Outdated methods for maintaining healthcare records have not been proven sufficient for treating chronic diseases like diabetes. Data analysis methods such as Recommendation System (RS) can serve as a boon for treating diabetes. RS leverages predictive analysis and provides clinicians with information needed to determine the treatments to patients. Prescription-based Health Recommender System (HRS) is proposed in this paper which aids in recommending treatments by learning from the treatments prescribed to other patients diagnosed with diabetes. An Advanced Density-Based Spatial Clustering of Applications with Noise (DBSCAN) clustering is also proposed to cluster the data for deriving recommendations by using winnowing algorithm as a similarity measure. A parallel processing of data is applied using map-reduce to increase the efficiency & scalability of clustering process for effective treatment of diabetes. This paper provides a good picture of how the Map Reduce can benefit in increasing the efficiency and scalability of the HRS using clustering.

Keywords: Recommendation, map reduce, patient-centric healthcare, DBSCAN, HRS.

ENGINEERING JOURNAL Volume 26 Issue 10
Received 6 June 2022
Accepted 18 October 2022
Published 31 October 2022
Online at https://engj.org/
DOI:10.4186/ej.2022.26.10.85

DOI:10.4186/ej.2022.26.10.85

1. Introduction

The Recommendation system (RS) provide users with personalized online product or service recommendations to handle the growing online information overload problem and improve customer relationship management. Collaborative filtering (CF) techniques have been developed for several years as the key thrust behind recommending systems and continue to be a hot field in both industry and academics [1]. RS can be described as programs that attempt to suggest specific users (companies or individuals) the most relevant items (services and products) by predicting a user's interest in an object based on user and user-to-item experiences and related item details [2]. The goal of RS development is to minimize the abundance of information by retrieving the most important information and resources from a vast volume of data and thereby delivering customized services. The most significant aspect of a recommendation framework is the ability to "guess" the needs and desires of a user by evaluating the user's actions and/or other user's behavior to produce customized recommendations. In data-intensive applications, recommendation systems are quite useful when they evaluate vast volumes of data and make better recommendations [3].

RS are widely used in e-commerce/social networks, and can be equally important in healthcare domain as well. In healthcare services, RS can be used by clinicians to administer medication to patients based on patient's profile and illness diagnosis. RS can be used by patients as well as by the physicians while doing self-study of the disease and in selecting appropriate treatment. There are three types of recommendation algorithm used: Collaborative filtering, Content-based filtering, Hybrid filtering as discussed in [4]. Also the combination of various approaches is used for creating recommender system to enhance the performance of recommendations [5].

In this paper, prescription based HRS is implemented over healthcare data for diabetic patients using advanced DBSCAN clustering and map reduce. Conventional HRS suffers from efficiency and scalability issues, as they are unable to manage vast amounts of health data with reliability and speed. To solve this problem, the map reduce architecture is used which distributes a load of large health-related datasets over many nodes, operating on commodity hardware, in order to allow parallel processing. The other features provided by the map reduce paradigm are ease of use, fault tolerance, and flexibility [6]. Apart from this, map reduce architecture also supports distributed health care approach which eliminates the overload of data by using it with the RS.

For unsupervised data mining, clustering is one of the strategies that deal with obtaining classes in a collection of unlabeled data. It is used to segment the data sets into different clusters in such a way that the objects in the cluster of the same group are strongly related to each other and different from the objects in another group [7]. In various applications, such as customer purchase pattern analysis, pattern recognition, and image processing, clustering analysis has helped extensively. For clustering healthcare data, clustering algorithms in a huge number are currently available, but to select the right clustering algorithm, is very difficult for people with little knowledge of data mining [8]. This paper aims to study current clustering approaches in healthcare and assess the efficiency of DBSCAN clustering and map reduce algorithms to overcome this problem. One of the most common clustering algorithm that is often cited in scientific literature is DBSCAN. DBSCAN [9] [10] is Density-Based Spatial Clustering of Applications with Noise. It is a density-based algorithm used to locate the clusters with noisy data and arbitrary form. DBSCAN holds some desirable properties similar to other clustering techniques. First, it breaks data into clusters of random forms. Clusters that are absolutely surrounded by another cluster, for example, can be identified. Second, the number of clusters a priori is not needed by DBSCAN. Third, the order of the points in the dataset is oblivious to this.

To measure the accuracy of created clusters using DBSCAN clustering algorithm, Confusion Matrix is used. Usually confusion matrix is used with classification where target variable is known, but it can also be used with clustering for performance evaluation. The confusion matrix is more generally referred to as the contingency table under which the matrix may be arbitrarily big. The sum of diagonals in the matrix is the number of correctly categorized instances; while the remaining instances in the matrices are classified incorrectly. A confusion matrix includes information from a classification system regarding predicted and actual classifications. Using the data in the matrix, the output of such structures is generally evaluated. This paper contributes in providing effective recommendations in the form of prescriptions on large set of diabetes data by grouping patients having similar symptoms under one cluster and labelling cluster name on the basis of frequently used keywords. It will not only help in early prediction and diagnosis of this severe disease but also guide patients for controlling it at initial stage itself.

The rest of the paper is organized as follows: Section 2 discusses the related work, where recommendation systems using different clustering techniques are discussed. Section 3 covers the techniques used. It basically gives details about the similarity measure used for clustering the data, clustering technique used and role of map reduce framework in distributing the load for efficient processing. The extended DBSCAN algorithm is detailed out in section 4 followed by details on datasets. In section 5, implementation details are covered. Section 6 analyses the result/performance of the proposed algorithm followed by the conclusion in section 7.

2. Related Work

Various clustering-based recommender systems were developed in the literature providing recommendations on

food, hospitals, movies, diet, books, etc. In [11], a food recommendation system is proposed for diabetic patients based on nutrition value using Self Organizing Maps (SOM) and k-means clustering techniques. It recommends the substituted food based on its nutrition values and food characteristics. In [12], a recommender system for u-Commerce is developed using a new clustering method based on item category with weight and Bayesian probability based on item preference. A hospital recommender system is discussed using correlation and clustering [13]. In [14], a book recommender system using k-mean clustering and collaborative filtering is discussed which uses ratings and scores for recommending books. A movie recommender system is built using the k-nearest neighbor algorithm and k-means clustering based on different root mean squared error (RMSE) values calculated for different cluster values in [15]. In [16], a personalized nutrition recommender system was proposed for diabetic patients using improved krill-herd optimization and k-means. AI book based recommender system is developed in [17] using matrix factorization from collaborative filtering and AI based lexile level measurement from content filtering making it a hybrid model. [18] proposed News Recommender System(NRS) using Deep Neural Networks. Table 1 below, discussed the drawbacks of related works that has been carried out.

Table 1. Drawbacks of the related work.

Author	**Type of RS**	**Technique**	**Proposed work**	**Drawback**
Phanich, Pholkul and Phimolt are,	Food recommendation system	SOM, K-means	Recommends the substituted food on the basis of its nutrition values and food characteristics.	SOM has no mechanism to determine the number of clusters, initial weights and stopping conditions.
Cho, Park and Ryu	U-commerce recommender system	RFM method	Reflect the importance of an item by frequently changing trends of purchase pattern.	It only focuses on the best customers and only use limited number of selection variables.
Tabrizi et al.	Hospital recommender system	Unsupervised data-driven methodology	Correlations are extracted, validated, ranked and converted to recommendations	Due to size & dimensions of datasets, it is difficult to identify factors for better satisfaction.
Rani et al.	Book recommender system	k-mean clustering, and collaborative filtering	Variety of books are offered by book recommendation system, it display the results based on the search of user.	Not appropriate because k-mean clustering only calculates those users who scores for items is used.
Ahuja, Solanki and Nayyar	Movie recommender system	k-means clustering and k-nearest neighbor algorithm	Predicts the user's preference of a movie on the basis of different parameters.	High Computation cost, requires large memory proportional, Low accuracy rate.
Devi, Bhavithra, Saradha	Personalized nutrition recommender system	k-means and krill-herd optimization	Suggest the best diets according to patient's health situation.	The concepts and principles about nutrition incorporated in the systems are not deep enough.
Milcah and Moorthy	AI Book based Recommende	Matrix Factorization and hybrid recom	Provides personalized recommendations on	Suffers cold start problem

	r System	menda tion approa ch	books as well as predict list of books that are not visited or viewed by specified users.	
Raza and Ding	News Recom mende r System (NRS)	Deep Neural Netwo rks	Focus is on the consequenc e of NRS on User's behaviour and suggest possible solutions to overcome the challenges faced by NRS.	Not able to ensure that all popular news are credible and truly popular.

Various recommender systems have been built in past using different techniques and each technique is associated with some drawbacks. We also proposed DBSCAN cluster based recommender system using map reduce using winnowing algorithm as a similarity measure and threshold value as a parameter to cluster the instances.

3. Techniques Used

3.1. Similarity Measure

Measuring similarity is important for getting only relevant information and is effective in various text mining applications like information retrieval, document clustering, text classification, text summarization, checking plagiarism, etc. [19]. There are four measures to compute the similarity of texts: Hybrid similarities, string-based, Corpus-based, and Knowledge-based [20]. String-based measures are further divided into two categories: character-based and term-based. Winnowing algorithms, used in this paper for computing similarity are character-based and that uses n-grams (sequences of substring) to find fingerprints [21]. It uses the Dice coefficient to measure the similarity between the symptoms of two patients. Dice coefficient is defined as twice the number of common terms in the strings that are being compared then divided by the total number of terms in both the strings.

$$\text{Dice}=\frac{2\times|f(A)\cap f(B)|}{|f(A)|+|f(B)|} \quad (1)$$

where f(A) and f(B) are fingerprints of symptoms of patient A and patient B. Fingerprints are created from the hash values calculated on N-grams tokens using the MD5 function. The similarity between the symptoms of two patients is calculated by creating their fingerprints from the hash value generated using MD5 as follows:

- The symptom for Patient A = Unintended Weight Loss
- The symptom for Patient B = Fatigue and Weakness

5-grams tokens for the symptoms of patient A "unintended weight loss" are created and their corresponding hash value is calculated using MD5. It is then converted into bytes' form by converting each digit into bits and then combining 8-bits into bytes. Some of the byte's representations of hash values are as follows.

Unint="134159941449661201208143311311855773171157"
ended="221381192081274114920732361894217918415023 5"
weigh="177399023310325318937209105471652262457122 4"
tloss="321515746113381481111192216156128130536 4"

Finally, hypothetically the first byte in the byte array is taken as the hash value for the corresponding n-gram tokens and it is shown in Table 2.

Table 2. 5-grams and final hash-values as Ist byte for symptoms of Patient A.

5-gram Tokens	Hash Value (32-digit hexadecimal number)	Final Hash Values
Unint	869F5E90603DC9D08F1F83B93949AB9D	134
Ninte	5346CBDC183342431F50FA5872EBA735	83
Inten	AB5855ED8358CB92080655DE160850EA	171
Ntend	39ADDC7E127DAC48063960E06E777002	57
Tende	C5BE12E1008814E2C9CDECAFA96A175B	197
Ended	DD2677D07F29951449ECBD2AB3B896EB	221
Nded w	277D9CA90BC97646D270E062CA886A7F	39
Dedw e	3E3328A3F3F7D77AE79AE05D74BBD4B0	62
Edwei	83ADBEBC898464E0D5215BB03700D686	131
Dweig	A6807D1DA89E945BAEB4C73399B2DE25	166
Weigh	B1275AE967FDBD25D1692FA5E2F547E0	177
Eight	24D27C169C2C881EB09A065116F2AA5C	36
Ightl	E906181FA656B53BE492552287718C3B	233
Ghtlo	10E4A3ED1C575710CB8AE00BBD59673E	16
Htlos	7C3932A5BC2F8DD2631961F86576D547	124
Tloss	200F9D2E7126946F7702D89C80823540	32

The window created for hashes of size 4 is shown in Table 3.

Table 3. Window size, W= 4.

Windows created	
134,83,171,57	83,171,57,197
171,57,197,221	57,197,221,39
197,221,39,62	221,39,62,131
39,62,131,166	62,131,166,177
131,166,177,36	166,177,36,233
177,36,233,16	36,233,16,124
233,16,124,32	

The value of W=4 and N=5 is chosen depending upon the length of the string, No. of N-grams formed, etc. The length of window size should be less than the value of N in N-grams.

As $(L - N + 1)$ =No. of N-gram (2)

=> 20-N+1=16,

=>N=21-16=5. (3)

Therefore, N=5 is chosen. So, Fingerprints created for symptoms of "unintended weight loss" of "patient A" are: [57 39 62 36 16]. Similarly, the fingerprints are calculated for 5-grams tokens on symptoms of "patient B" and "fatigue and weakness" are- [18 126 80 89]. Then the similarity using Dice coefficient is calculated using the formula, where the value of f(A) and f(B) are:

f(A) = [57 39 62 36 16] = 5 (4)

f(B) = [18 126 80 89] =4 (5)

So, the Dice coefficient between symptoms of Patient A and Patient B is:

2*0/9=0. (6)

Thus, the value of the dice-coefficient will be 0 if symptoms have nothing in common and it will be 1 when both patients are having identical symptoms [22]. This means if two documents are symptoms like "unintended weight loss", then their Dice coefficient will be "1". The winnowing algorithm is more stable and gives better performance than fingerprint [23] and cosine similarity. It also aims to select the minimum hash value calculated using MD5 from each window. The flowchart for calculating the similarity between symptoms of patients is shown in Fig. 1 and the flow of the winnowing algorithm is shown in Fig. 2.

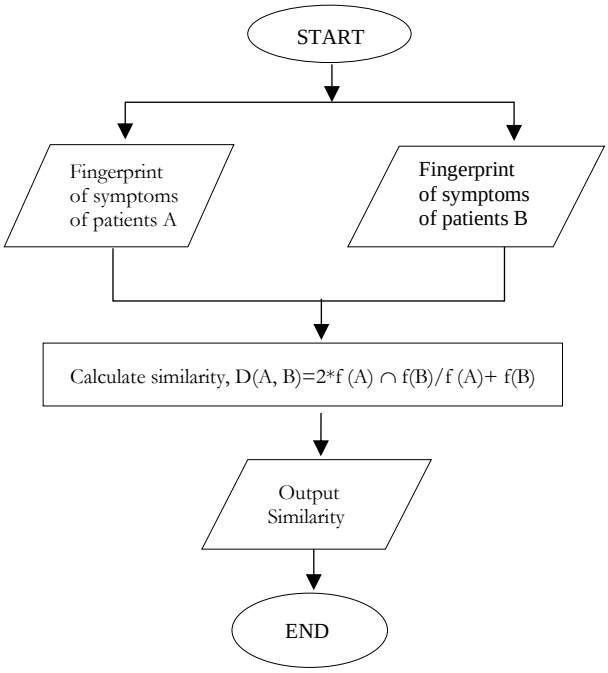

Fig. 1. Illustration of similarity calculation.

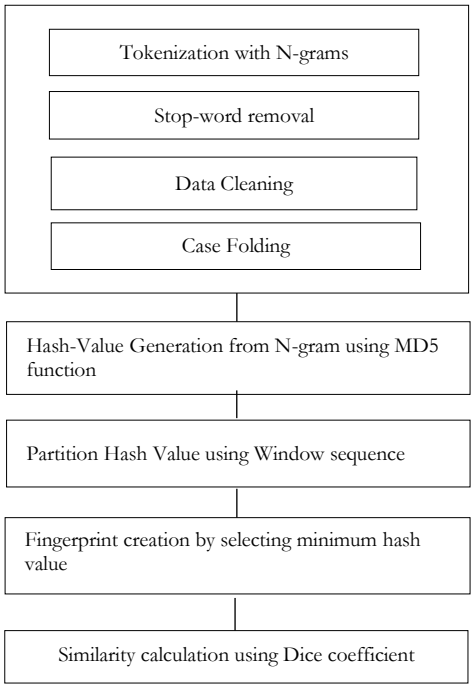

Fig. 2. Flow of Winnowing Algorithm.

3.2. Clustering Techniques Used

Clustering is the process of dividing the large datasets into groups so that similar objects are placed in the same group. Employing clustering algorithms on recommender systems reduces sparsity and improves scalability by reducing the response time due to the application of the algorithm on clustered datasets only [24]. Various types of clustering algorithms exist in the literature like hierarchical clustering, partitioning-based, density-based, grid-based

clustering algorithm. Every clustering algorithm is associated with certain advantages and disadvantages. Selection of clustering algorithm depends on various parameters like nature of data, cost, computational speed, accuracy, etc. Although K-Means algorithm is one of the most widely used clustering algorithm, but it performs well on numerical datasets only. However, the patient dataset contains attributes like symptoms and prescriptions which is in textual form and likely going to result in clusters of different size/shape. It will be impractical to use such clusters for the recommender system as the resulting recommendation would not be effective. Due to these limitations, it is not practical to use K-mean clustering for the solution being proposed in this paper.

Density-based algorithms i.e. DBSCAN (Density-based spatial clustering of application with noise) clustering is used and further extended in our solution for improving the results. DBSCAN clustering has various advantages like it forms arbitrary shape clusters and it does not need to input the number of clusters in advance, more flexibility in terms of shape and size of clusters, also the identification of outliers as noise rather than classifying them as clusters. It works on two parameters only that can be used in detecting outliers and efficiently separate clusters with high densities from the clusters with low density [25]. One of the biggest disadvantages of the DBSCAN clustering algorithm is that it is inefficient in processing a large volume of data. In order to overcome this limitation, a parallel computing Map-Reduce Framework is used as discussed in [26]. The dice coefficient resulting from the winnowing algorithm will be used as a similarity measure for clustering the data using Advanced DBSCAN clustering. Figure 3 below shows the overall steps used in Advanced DBSCAN Algorithm.

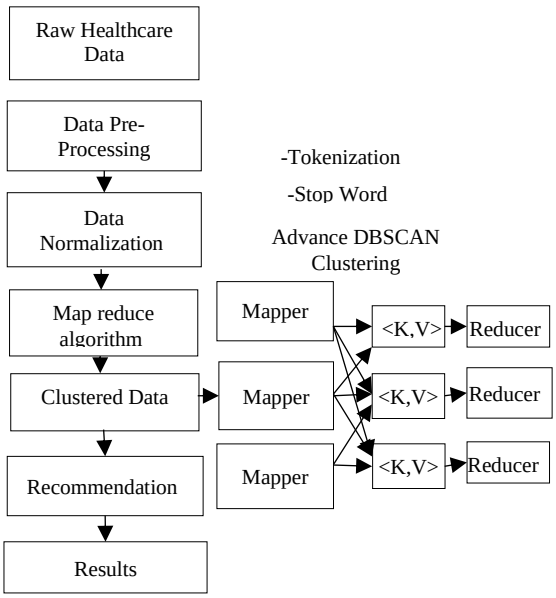

Fig. 3. Advanced DBSCAN algorithm.

4. The Proposed Algorithm

Algorithm begins with cleaning and pre-processing of raw datasets using techniques such as tokenization using n-gram and stop-word removal. Weights of the processed data(tokens) is then scaled in the range from 0 to 1 using normalization. Winnowing algorithm is used to find similarity based on patient symptoms. It is a lightweight, highly efficient, flexible and more reliable algorithm [27]. It uses Dice coefficient as a similarity measure which is more efficient in finding text-based similarity as compared to jaccard coefficient, cosine similarity and berg coefficient [28]. MD5 hashing method is also used in order to obtain fingerprint of healthcare record. MD5 is faster and one of the most widely used algorithm for creating one-way hash. It is resistant to collisions and produces 128 bits fixed length hash values [29]. It is followed by an extended version of original DBSCAN clustering algorithm.

4.1. Pre-Processing Stage

This stage reduces the dimensions and noise from the data. Various pre-processing steps performed in this solution are- Tokenization with N-grams, filtering like Stop words removal, removing unnecessary symbols, whitespaces, etc. followed by case folding i.e. converting all letters to lowercase. Various types of data pre-processing techniques are discussed in [30].

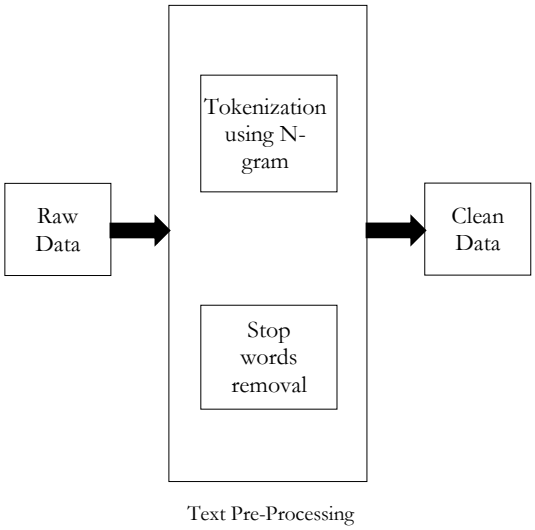

Fig. 4. Healthcare data pre-processing techniques.

In Fig. 4 above, the pre-processing methods applied to the diabetic datasets are explained as follows:

- Tokenization with n-grams- In this method, n-grams tokens are created by splitting the symptoms and prescription from the healthcare record into the various permutation and combinations of all possible symbols, words and phrases called tokens, which are further used for mining.

- Stop Words Removal- Stop words are frequently used words, which unnecessary takes a lot of space and adds overhead to the text processing. Removing these common words used by clinicians in prescriptions reduces the dimensionality which eliminates noise in the healthcare data.

Pre-processing stage results in data in reduced form, which is further normalized to get the weights of n-gram tokens between 0 and 1.

4.2. The Normalization Stage

Normalization brings the value of patient records in the dataset to the common range for effective distribution to clusters. Before normalizing the data, all the n-grams tokens resulted from the data pre-processing stage are assigned weight equals to 1 using an equal weighting scheme. This is done so that all the tokens are given equal importance. Normalization is performed to scale all the tokens in the healthcare dataset between ranges of 0 to 1. Normalized data for a particular token value is achieved by dividing the weight of that single n-gram token by the combined weights of all the tokens formed from the symptoms of the patient in the dataset. The following formula is used for Normalization:

$$\text{Normalization} = \frac{\text{weight of single n-gram token}}{\text{sum of weights of all tokens of patient's symptoms}} \quad (7)$$

Normalization is explained below with help of an example.

- Before Normalization: N-gram tokens formed from a patient record with "Extreme Hunger" symptoms are: Extreme, Hunger, Extreme Hunger with the weight of each token is initialized to 1 using an equal weighting scheme.
- After Normalization: Normalization scales the tokens in the range of 0 to 1 as shown below:

$$\text{Weight of Extreme} = 1/3 = 0.333333 \quad (8)$$

Results from the normalization process are further used to cluster the data which is explained in the next section.

4.3. The Clustering Stage

As clustering is being performed on the symptoms of the patients which is in textual form, therefore DBSCAN, a density-based clustering algorithm is chosen for grouping the data into clusters based on the density of the neighborhood. Density-based spatial clustering of applications with noise (DBSCAN) depends on two concepts: density reachability and density connect ability [31]. This algorithm helps in scanning the data space with high-density regions, separated by lower density points regions. Key elements in the original DBSCAN algorithm used for clustering text are:

- Epson – It defines the radius around a data point to find similar neighbors. Two points are considered as neighbors if the distance between them is lower or equal to Epson. For text documents, Epson will be the value of the distance measure used to compute the similarity to obtain the neighbors.
- Min points- Min points are the minimum number of neighbors points that are required to form a cluster. It is basically a threshold value used to form a cluster.

To get more efficient clusters, an enhanced version of DBSCAN i.e. Advanced DBSCAN algorithm is used, which uses a winnowing algorithm as a similarity measure. It not only calculates the similarity using dice-coefficient, which will be the Epson, but also generate clusters of leaders and followers based on the min-points i.e. the threshold value. It also further optimizes the clusters by finding the similarity between the leaders and putting them in the same cluster. The step-by-step process is explained below in Algorithm 1.

Algorithm 1: DBSCAN Algorithm with input (datasets D, Epson ->Dice coefficient (between 0 and 1), Min points -> threshold value τ to get Leaders and Followers):

a) Scan all the records in the patient's dataset D.
b) Define threshold, τ = 0.9(based on experiment/analysis)
c) If record R is already visited, then move to next Record.
d) For each non-visited Record R, follow steps (e – k), compute similarity-matching with every other record in the dataset, using the winnowing algorithm.
e) Get fingerprint of each record using n-gram tokens.
f) Calculate the hash of each record by using the MD-5 function on N-gram tokens.
g) Get Window sequence of hash value using (N-grams=5, window size=4) resulted from MD5 function.
h) Do fingerprint matching by using Dice Coefficient,

$$\text{Dice}=\frac{2\times|f(A)\cap f(B)|}{|f(A)|+|f(B)|}$$

where f(A) and f(B) is the fingerprints of symptoms of patient A and B find similarity.
i) Discard the record (outlier), if similarity resulted from Dice coefficient is less than threshold i.e. Dice < τ.
j) Flag Record R as Leader (l) and matched record as Follower (f) if similarity resulted from Dice coefficient is greater than threshold i.e. Dice> τ.

k) Flag/Label Leaders, l and Followers f, as visited so that they are ignored in subsequent iterations of scanning the records.

We got leaders and their followers based on similarity calculated using the winnowing algorithm which are further clustered on the basis of certain parameters as shown in Algorithm 2.

Algorithm 2: Advanced DBSCAN Algorithm (D, τ_{min}=0.3, Min Points ≥ 3 to get matching Leaders):

a) Define min Threshold, τ_{min}=0.3.
b) Compute similarity for leaders having followers, f>=3, repeat steps (c) - (g) in above algorithm,
c) **If** similarity> τ_{min}, i.e. $||l1\text{-}l2|| > \tau_{min}$, **Then** merge l1 followed by followers f1,f2,f3 and l2 followed by followers f4,f5,f6 and form a cluster.

The corresponding leaders, cluster of leaders, followers and their clusters are shown in Fig. 5.

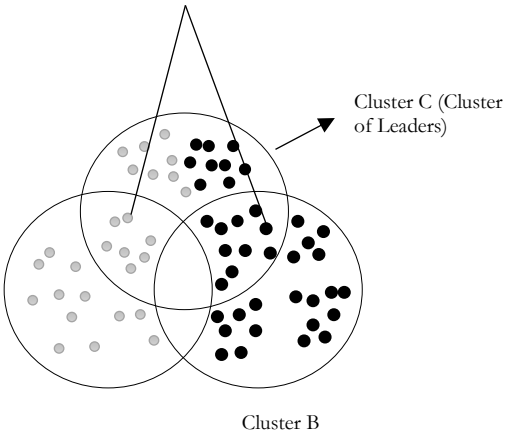

Fig. 5. Clustering of Leaders.

Then the similarity will be recalculated, but this time it is calculated between leaders and using a lesser threshold value than the previous one. The value of minimum objects is also defined to select a leader and it is based on the count of their followers, if any leader is having followers less than the count of minimum objects then that leader will be discarded. In the final stage, the clusters and labels to each cluster are defined. The label will be defined with the keywords whose occurrence is maximum in the data of that cluster.

4.4. Recommendation using Map-Reduce

The diabetes dataset meets most of the characteristics of Big Data which are volume, veracity, and variety. For faster processing of healthcare dataset, Map Reduce framework is applied for distributed and parallel processing. Map Reduce is a fault-tolerant, efficient and scalable framework that enables data parallelization, load balancing, and data distribution [32]. It is used for mining large datasets and restricted to be applied on algorithms that use <key, value > pairs [33] [34].

In this algorithm, a map reduce framework is used to process the clusters resulting from Advanced DBSCAN clustering, which are created based on symptoms of diabetes datasets. As mentioned in the last section, the data is already clustered based on the similarity of the symptoms and prescription, using Advanced DBSCAN clustering. Clusters are iteratively passed to the mapper nodes for processing. In mapper nodes, content-based filtering is applied over the data to extract recommendations. Speed of map reduce process is optimized by passing clustered (instead of un-related) data to mapper nodes which limits processing needed by reducer nodes in combining the results. The Steps of Recommendation Algorithm are explained in Algorithm 3.

Algorithm 3: Recommendation Algorithm

a) Keywords of patient's symptoms and other information like Plasma glucose concentration, Body Mass Index (BMI), Diabetes Pedigree Function, Age and, Probable symptoms were used.
b) Compute similarity of keywords with cluster labels.
c) If sim > τ, use that cluster to recommend the prescription for the patient.

5. Implementation

5.1. Dataset Details

The key driver of our research is to control diabetes by recommending relevant treatments/prescription to patients in the early stages. This is achieved by providing prescriptions in the form of recommendations to the patient when the patient enter his/her basic attribute details or symptoms on the portal. Datasets used here uses all the attributes of the PIMA Indians diabetes database of the National Institute of Diabetes and Digestive and Kidney Diseases [35]. Symptoms and Prescriptions are the two additional attributes that are included to comprehend the dataset as per our research need. These two attributes include highly confidential information which is not easily available. Due to which frequent symptoms are found in diabetic patients and the treatments prescribed to them are explored and added to the resulting dataset. Schema and attributes of datasets is listed below.

- Dataset: National Institute of Diabetes and Digestive and Kidney Diseases [35]
- Number of Attributes: 10+2=12
- Number of Records: 768

The attributes and its details are shown in Table 4.

Table 4. Attributes details.

S. NO.	Attribute	Details
1	PatientId	Id of the patient
2	Preg	Number of times pregnant
3	Plas	Glucose level for Plasma
4	Pres	blood pressure
5	Skin	skin thickness
6	Insu	serum insulin
7	Mass	Body Mass Index
8	Pedi	Pedigree function
9	Age	Patient age in a year as of the date of treatment
10	Class	The classifier to determine whether a patient is tested positive or negative for diabetes
11	Symptoms	Symptoms demonstrated by diabetic patients
12	Prescriptions	Prescription recommended to patients for treatment of diabetes

On analyzing the dataset, it was found that the most common symptoms of diabetic patients are - Frequent Urination, Extreme Hunger, Unintended weight loss, Increased thirst, Fatigue and weakness. Generally suggested prescriptions for these symptoms based on this dataset are:

- The patient was advised to use rapid-acting insulin analogs to reduce hypoglycemia risk.
- Metformin is the preferred initial pharmacologic agent for the treatment.
- Educated patient on matching prandial insulin doses to anticipated physical activity, Carbohydrate intake, and pre-meal blood glucose levels.
- The early introduction of insulin was considered as there was evidence of ongoing catabolism (weight loss).

In this section, the results of step-by-step implementation of the multi-step recommendation, an algorithm is demonstrated using tabular representation. Table 5 below represents the first stage i.e. the pre-processing stage, where n-grams tokens based on the symptoms of different patients are achieved followed by stop-word removal.

Table 5. Preprocessing Stage.

Token id	Token	Weight	Patient ID
1	Frequent	1	1
2	Urination	1	1
3	Frequent_Urination	1	1
4	Extreme	1	2
5	Hunger	1	2
6	Extreme_Hunger	1	2
7	Fatigue	1	3
8	Weakness	1	3
9	Fatigue_and	1	3
10	Fatigue_and_Weakness	1	3

Table 6 below shows the normalization stage where the weights of N-gram tokens are scaled in the range of 0 to 1. As shown in equation (7) and (8), for the token "Extreme" from the symptom "Extreme Hunger", weight is 1/3 = 0.3333 after normalization.

Table 6. Normalization Stage.

Token id	Token	Weight after Normalization	Patient ID
1	Frequent	0.3333..	1
2	Urination	0.3333..	1
3	Frequent_Urination	0.3333..	1
4	Extreme	0.3333..	2
5	Hunger	0.3333..	2
6	Extreme_Hunger	0.3333..	2
7	Fatigue	0.1666..	3
8	Weakness	0.1666..	3
9	Fatigue_and	0.1666..	3
10	Fatigue_and_Weakness	0.1666..	3

Table 7 below shows Leaders and followers achieved after threshold value 0.3.

Table 7. Leaders and Followers for clustering.

S No	Patient ID	Followers
1	1	10,107,109,113…
2	100	7,102,106,110…
3	130	104,105,111,118,..
4	134	11,23,56,112,….

DOI:10.4186/ej.2022.26.10.85

Table 8 below show clusters formed from sample diabetic dataset used in our solution.

Table 8. Clusters formed using Advanced DBSCAN.

S No	Label	Cluster
1	Weakness Fatigue	XXX
2	Fatigue Weakness	YYY
3	Extreme Hunger	ZZZ

Figure 6 below shows recommendation provided to patients based on details input by them, which are extracted from processed data resulting from processing the clusters from previous steps using Map Reduce framework.

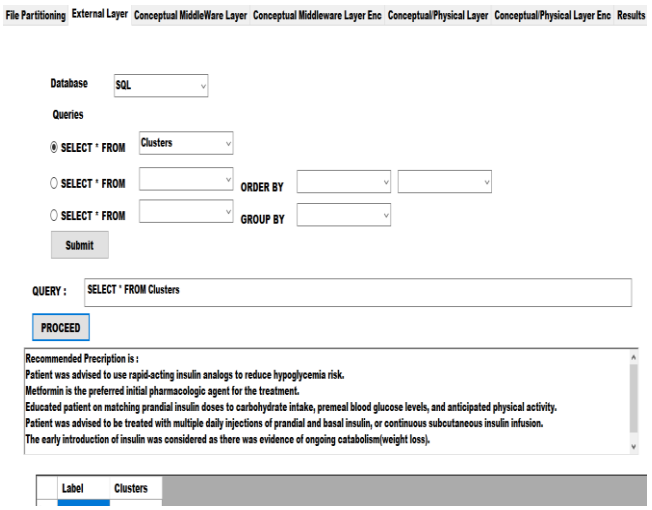

Fig. 6. The final recommendations based on the patient's details.

6. Results and Analysis

For evaluating the proposed recommendation system, the information for patient 1 and 2 has been taken as depicted in Table 9 below.

Table 9. Patient details.

S.no.	Plasma glucose concentration	BMI	Diabetes Pedigree function	Age	Symptoms
1.	182	30.5	0.345	28	Fatigue and weakness
2.	162	24.3	0.178	50	Extreme hunger

On the basis of provided inputs, the proposed recommendation system recommended the following prescriptions for the patients as shown in Fig. 7 and Fig. 8 below for Patient 1 and Patient 2:

- Patients were advised to use rapid-acting insulin analogs to reduce hypoglycemia risk.
- Metformin is the preferred initial pharmacologic agent for the treatment.

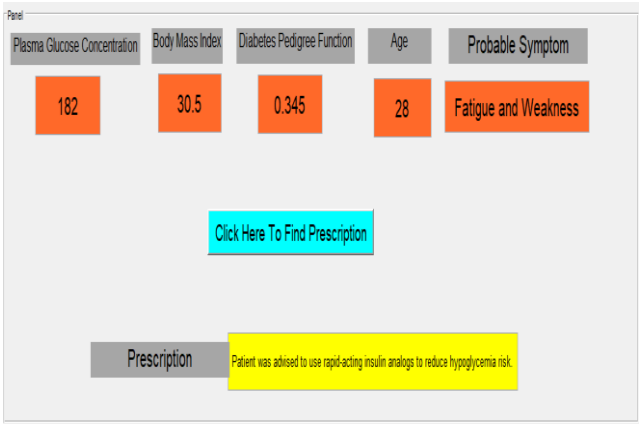

Fig. 7. Recommendation for patient 1.

Fig. 8. Recommendation for patient 2.

For evaluating the performance of model, the metrics computed are Confusion Matrix, Precision, Recall, F1-Score. The test dataset consists of 295 people. The Confusion Matrix of test dataset for 154 patients is:

$$\begin{matrix} 47 & 3 & 0 & 2 & 3 \\ 0 & 54 & 2 & 5 & 4 \\ 0 & 2 & 56 & 4 & 5 \\ 1 & 2 & 5 & 42 & 3 \\ 2 & 1 & 2 & 3 & 47 \end{matrix}$$

Based on the confusion matrix, the following values are calculated.

Mean absolute error is 0.2236827153664131 (9)

Mean Squared error is 0.05503617517905924 (10)

Root mean square error is 0.2937463335436398 (11)

Accuracy of Model is 83.334754576 % (12)

Figure 9 below shows the comparison graph for the same.

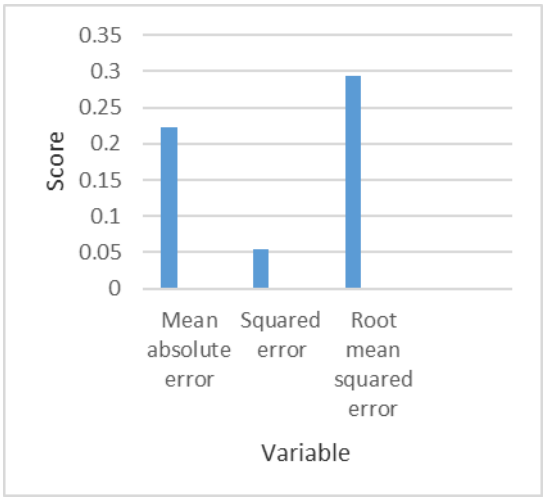

Fig. 9. Comparison graph between mean absolute error and squared error.

F1 score, precision, and recall, provides better insights into the prediction as shown in Fig. 10 below.

F1 Score: 0.65869555555 (13)
Recall: 0.623666666 (14)
Precision: 0.6646444444 (15)

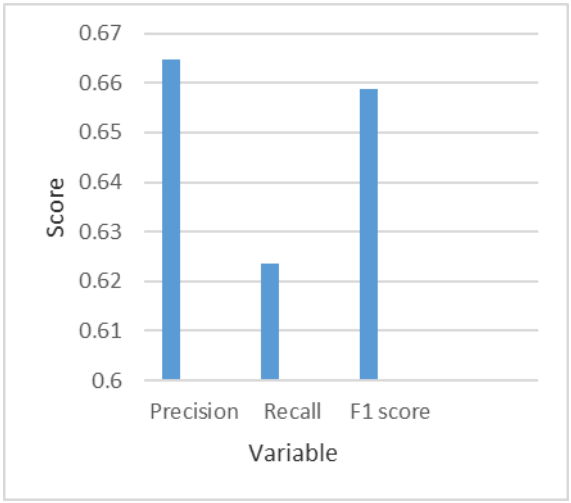

Fig. 10. Performance graph for better insight on precision, recall and F1 score.

Furthermore, to analyze the result of the proposed solution, the performance in terms of time taken is also measured:

- Time taken to create the clusters
- Response time to recommend the treatment to patients

Findings from the analysis are detailed in the following sub-sections.

6.1. Efficiency in the Time Taken to Cluster Data

The analysis on time taken to cluster the data by normal

DBSCAN clustering and Advanced DBSCAN clustering was done. As resulting from our analysis, it was found that Advanced DBSCAN clustering leads to time efficiency by reducing the time taken to create the clusters. In sample dataset, there are 768 records. Sample dataset was provided as input to measure the time taken to create the resulting clusters. This process of providing the dataset as input and measuring the time taken to create clusters was repeated for 5 cycles. In each cycle number of records in the dataset were varied to measure the performance with different volume of data. Each cycle was also repeated with both DBSCAN and Advanced DBSCAN clustering as developed in this paper.

In Fig. 11, the orange line shows the time taken by Advanced DBSCAN and the Blue line shows the time taken by DBSCAN. As clearly articulated in the chart Advanced DBSCAN clustering takes much lesser time than DBSCAN cluster. Total time taken to create clusters is reduced by 40%-50% with Advanced DBSCAN clustering which proves it's more efficient than normal DBSCAN clustering.

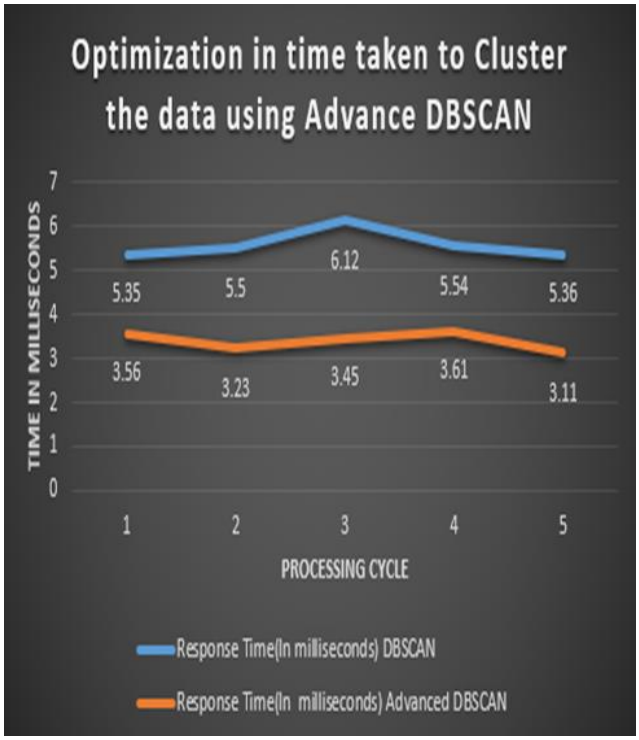

Fig. 11. Optimization in the time taken to create the clustering with Advanced DBSCAN.

6.2. Efficiency in Response Time in Providing Recommendation to Patients

The graph below in Fig. 12 articulates the benefits of providing recommendations to the patients using Advanced DBSCAN Clustering. To match inputs provided by patients, without clustering all the records in the dataset would need to be scanned to find a suitable match and recommend treatment. For example, in the sample dataset of 768 records used in the paper, assuming it takes 100 milliseconds to match per record basis, it would have taken us 76800 milliseconds without clustering. But with clustering, as records are grouped into clusters based on symptoms and prescription, the time take to recommend treatment to the patient is vastly reduced. Processing time varies, depending on with which cluster patient inputs are matched. In this analysis, the processing time reduces by 85% as instead of matching with 768 records, only 115 records are needed to match with to provide recommendations.

In the Fig. 12 below, the orange bar shows the number of records that would have to be scanned without clustering and the grey bar shows a number of records to be scanned using our solution, to recommend treatment to patients. The yellow trend line shows an overall improvement in processing items. As clearly depicted in the diagram processing time can be improved up to 85% which varies based on different permutation/combination of details being input by a patient looking for treatment.

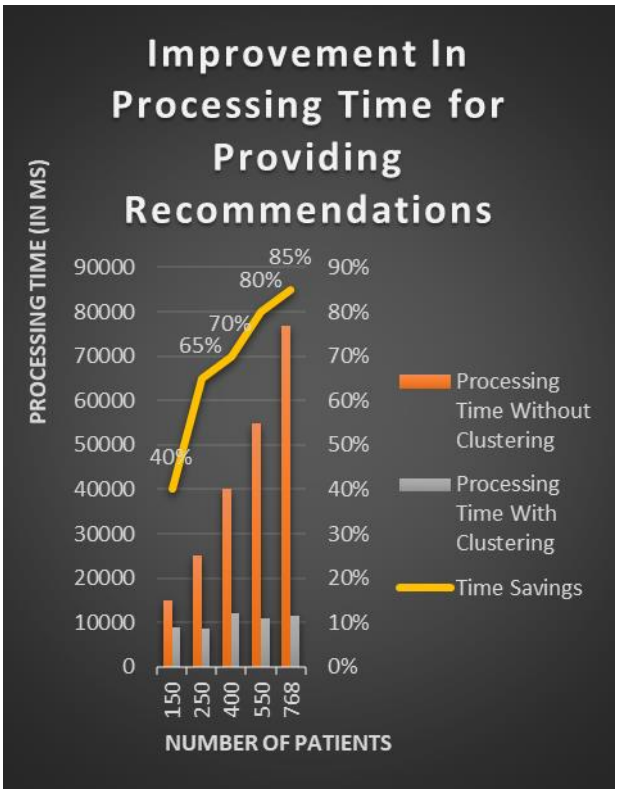

Fig. 12. Efficiency achieved in recommendations using cluster-labels in Advanced DBSCAN clustering.

7. Conclusion

Presently, due to lack of analysis of diabetes data, if patients are looking for advice, they search the results on the web and are provided with overloaded information as those results are provided without considering any context. Those search results are more based on keywords being searched instead of patients searching the information. Our solution truly enables patient-centric healthcare for diabetic patients as instead of relying on various unrelated searches, out-of-context searches, our system understands the attributes of patients and provides them with recommendations based on the medical history of other patients having similar attributes. Apart from this, clustering using Advanced DBSCAN produces good quality clusters using an efficient similarity measure based on confusion matrix and optimizes the data processing by using map reduce and extract the recommendations from the raw data. Also using Map Reduce not only improves the efficiency, scalability and flexibility in terms of clustering the data, but also provide fast recommendations by processing the datasets in parallel.

References

[1] T. Chen et al., "SVD feature: A toolkit for feature-based collaborative filtering," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 3619-3622, 2012.

[2] J. Bobadilla, F. Ortega, A. Hernando, and Gutiérrez, "A recommender systems survey," *Knowledge-Based Systems*, vol. 46, pp. 109–132, 2013.

[3] S. Saravanan, "Design of large-scale content-based recommender system using Hadoop MapReduce framework," in *2015 Eighth International Conference on Contemporary Computing (IC3), IEEE Computer Society,* Noida, India, 2015, pp. 302-307, doi: 10.1109/IC3.2015.7346697.

[4] R. Bateja, S. K. Dubey, and A. Bhatt, "Health recommender system and its applicability with MapReduce framework," in *Soft Computing: Theories and Applications. Advances in Intelligent Systems and Computing*, M. Pant, S. Rawat, K. Ray, T. K. Sharma, and A. Bandyopadhyay, Eds. Singapore: Springer, 2018, vol. 584, pp. 255-266.

[5] T. Phuksenga and S. Sodseeb, "Recommender system based on expert and item category," *Engineering Journal*, vol. 22, no. 2, pp. 157- 168, 2018.

[6] I. A. T. Hashem, N. B. Anuar, A. Gani, I. Yaqoob, F. Xia, and S. U. Khan, "MapReduce: Review and open challenges," *Scientometrics*, vol. 109, pp. 389–422, 2016. [Online]. Available: https://doi.org/10.1007/s11192-016-1945-y.

[7] J. Han, M. Kamber, and J. Pei, "Cluster analysis-10: Basic concepts and methods," in *Data Mining*, 3rd ed., J. Han, M. Kamber, and J. Pei, Eds. Boston: Morgan Kaufmann, 2012, pp. 443–495.

[8] P. N. Tan, M. Steinbach, and V. Kumar, "Data mining cluster analysis: Basic concepts and

algorithms," *Introduction to Data Mining.* 2013, pp. 526-533.

[9] Y. He et al., "MR-DBSCAN: An efficient parallel density-based clustering algorithm using map reduce," in *2011 IEEE 17th International Conference on Parallel and Distributed Systems*, Tainan, Taiwan, 2011, pp. 473-480, doi: 10.1109/ICPADS.2011.83.

[10] M. Ester, H. P. Kriegel, J. Sander, and and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96),* 1996, pp. 226–231.

[11] M. Phanich, P. Pholkul, and S. Phimoltares, "Food recommendation system using clustering analysis for diabetic patients," in *2010 International Conference on Information Science and Applications*, Seoul, 2010, pp. 1-8, doi: 10.1109/ICISA.2010.5480416.

[12] Y. S. Cho, H. W. Park, and K. H. Ryu, "Clustering method using item category with weight and bayesian probability based on item preference for recommendation in u-commerce," presented at *The 2nd FTRA International Conference on Ubiquitous Context-Awareness and Wireless Sensor Network ITME 2014*, KAIS-CITA 2014.

[13] T. S. Tabrizi, M. R. Khoie, E. Sahebkar, S. Rahimi, and N. Marhamati,"Towards a patient satisfaction-based hospital recommendation system," in *2016 International Joint Conference on Neural Networks (IJCNN*), Vancouver, BC, 2016, pp. 131-138, doi: 10.1109/IJCNN.2016.7727190.

[14] R. Rani and R. Sahu, "Book recommendation using k-mean clustering and collaborative filtering," *International Journal of Engineering Sciences and Research Technology*, vol. 6, no. 11, pp. 145-150, Nov. 2017.

[15] R. Ahuja, A. Solanki, and A. Nayyar, "Movie recommender system using k-means clustering and k-nearest neighbor," in *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence),* Noida, India, 2019, pp. 263-268, doi:10.1109/CONFLUENCE.2019.8776969.

[16] K. R. Devi, J. Bhavithra, and A. Saradha, "Personalized nutrition recommendation for diabetic patients using improved k-means and Krill-Herd optimization," *International Journal of Scientific & Technology Research*, vol. 9, no. 3, pp. 1076-1083, Mar. 2020.

[17] Y. Mercy Milcah and K. Moorthi, "AI based book recommender system with hybrid approach," *International Journal of Engineering Research & Technology (IJERT)*, vol. 09, no. 02, Feb. 2020.

[18] S. Raza and C. Ding, "News recommender system: A review of recent progress, challenges, and opportunities," *Artif Intell Rev* , vol. 55, no. 1, pp. 749–800, Jan. 2022. [Online]. Available: https://doi.org/10.1007/s10462-021-10043-x.

[19] W. H. Gomaa and A. A. Fahmy, "A survey of text similarity approaches," *International Journal of Computer Applications*, vol. 68, no. 13, pp. 13-18, 2013. [Online]. Available: https://doi.org/10.5120/11638-7118

[20] D. D. Prasetya, A. Wibawaand, and T. Hirashima, "The performance of text similarity algorithms," *International Journal of Advances in Intelligent Informatics*, vol. 4,no. 1, pp. 63-69, 2018. [Online]. Available: https://doi.org/10.26555/ijain.v4i1.152

[21] T. Khuat, N. D. Hung, and L. T. M Hanh, "A comparison of algorithms used to measure the similarity between two documents," *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, vol. 4, no. 4, pp. 1117-1121, 2015.

[22] M. Afzali and S. Kumar, "Comparative analysis of various similarity measure for finding similarity of two document," *International Journal of Database Theory and Application (IJDTA*), vol. 10, no. 2, pp. 23-30, 2017.

[23] A. T. Wibowo, K. W. Sudarmadi, and A. M. Barmawi, "Comparison between fingerprint and winnowing algorithm to detect plagiarism fraud on Bahasa Indonesia documents," in *2013 International Conference Information and Communication Technology (ICoICT)*, 2013, pp. 128-133.

[24] B. M. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Recommender systems for large scale E-commerce: Scalable neighborhood formation using clustering," in *Proceedings of the Fifth International Conference on Computer and Information Technology*, 2002, vol. 1, pp. 291-324.

[25] H. Shah, K. Napanda, and L. D'mello, "Density based clustering algorithms," *International Journal of Computer Sciences and Engineering (IJCSE)*, vol. 3, no. 11, pp. 54-57, 2015.

[26] X. Hu, L. Liu, N. Qiu, D. Yang, and M.A. Li, "Map Reduce-based improvement algorithm for DBSCAN," *Journal of Algorithms & Computational Technology*, vol. 12, no. 1, pp. 1-9, 2018.

[27] V. Gurusamy and S. Kannan, "Pre-processing techniques for text mining," *International Journal of Computer Science & Communication Networks*, vol. 5, no. 1, pp. 7-16, 2014.

[28] X. Duan, M. Wang, and J. A. Mu, "Plagiarism detection algorithm based on extended winnowing," *MATEC Web of Conferences*, vol. 128, p. 02019, 2017.

[29] R. Sutoyo *et al.*, "Detecting documents plagiarism using winnowing algorithm and k-gram method," in *2017 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)*, 2017, pp. 67-72, doi: 10.1109/CYBERNETICSCOM.2017.8311686.

[30] C. G. Thomas and R. T. Jose, "Comparative Study on Different Hashing Algorithms," *International Journal of Innovative Research in Computer and Communication Engineering(ijircce),* vol. 3, issue 7, pp.170-175, 2015.

[31] L. Sharma and K. A. Ramya, "Review on density based clustering algorithms for very large datasets,"

International Journal of Emerging Technology and Advanced Engineering, vol. 3, no. 12, pp. 398-403, 2013.

[32] S. N. Khezr and N. J. Navimipour, "Map reduce and its applications, challenges, and architecture: A comprehensive review and directions for future research," *Journal of Grid Computing*, vol. 15, no. 3, pp. 295-321, 2017. [Online]. Available: https://doi.org/10.1007/s10723-017-9408-0

[33] A. F. Gates, O. Natkovich, S. Chopra, P. Kamath, S. M. Narayanamurthy, C. Olston, B. Reed, S. Srinivasan, and U. Srivastava, "Building a high level dataflow system on top of Map Reduce: The pig experience," in *Proc. of Very Large Data (VLDB)*, 2009, vol. 2, no. 2, pp. 1414-1425.

[34] P. P. Anchalia, A. K. Koundinya, and N. K. Srinath, "Map Reduce design of k-means clustering algorithm," in *2013 International Conference on Information Science and Applications (ICISA),* Pattaya, Thailand, 2013, pp. 1-5, doi: 10.1109/ICISA.2013.6579448.

[35] "Pima Indians Diabetes Database," https://www.kaggle.com/uciml/pima-indians-diabetes-database

Ritika Bateja is currently working as an Assistant Professor in B. S. Anangpuria Institute of Technology & Management, Alampur, Haryana, Faridabad. She has more than 12 years of teaching experience in Manav Rachna International Institute of Research and Studies, Faridabad. Her area of research is Data Mining, Big Data Analytics and Healthcare. She has published various research papers in international journals and conferences.

Dr. Sanjay Kumar Dubey is Ph. D. in Computer Science and Engineering. He is Associate Professor in Department of Computer Science and Engineering in Amity School of Engineering and Technology, Amity University Uttar Pradesh, India. He has 19 year of teaching experience. He is a member of IET, ACM, IEANG, CSI and other Engineering professional bodies. He has rich academics & research experience in various areas of Computer Science and Engineering. His research areas include Artificial Intelligence, Soft Computing, HCI, and Data Mining. He has published 90 research papers in SCOPUS/SCIE indexed International Journals and in the proceedings of the reputed International Conferences. These publications have good citation records. He has authored 3 books also. He is guiding 6 Ph. D students. In addition, he has also served as a Session Chair and Technical Program Committee Member of several reputed International/National conferences.

Dr. Ashutosh Bhatt is currently working as an Associate Professor in School of Computer Science and Information Technology, Uttarakhand Open University, Haldwani, Distt- Nainital (Uttarakhand). Dr. Bhatt has completed Ph. D. in Computer Science in 2009. His work area of research is Artificial Neural Network. He has more than Seventeen years of teaching and research experience in various organizations of repute for PG & UG courses of Computer Science & IT. He is also associated with many renowned National/International Journals as Lead Guest Editor/reviewer/editorial board member. Around 10 SCIE/ESCI and more than 50 Scopus indexed/UGC Approved/other reputed National/International Journal research publications are credited to him. He was member of Board of Studies and Academic/Research Degree Committee of many universities. He is life member of CSI (Computer Society of India). He had served as State Student Coordinator of Region I Uttarakhand of CSI (for three year) and also served as Secretary of Uttarakhand ACM Professional Chapter.