Mathematical Biosciences and Engineering

MBE, 18(6): 8661–8682.
DOI:10.3934/mbe.2021429
Received: 01 July 2021
Accepted: 16 September 2021
Published: 11 October 2021

http://www.aimspress.com/journal/MBE

Research article

Big data integration enhancement based on attributes conditional dependency and similarity index method

Vishnu Vandana Kolisetty[1] **and Dharmendra Singh Rajput**[2,*]

[1] SCOPE, Vellore Institute of Technology, Vellore 632014, India
[2] SITE, Vellore Institute of Technology, Vellore 632014, India

* **Correspondence:** Email: dharmendrasingh@vit.ac.in.

Abstract: Big data has attracted a lot of attention in many domain sectors. The volume of data-generating today in every domain in form of digital is enormous and same time acquiring such information for various analyses and decisions is growing in every field. So, it is significant to integrate the related information based on their similarity. But the existing integration techniques are usually having processing and time complexity and even having constraints in interconnecting multiple data sources. Many of these sources of information come from a variety of sources. Due to the complex distribution of many different data sources, it is difficult to determine the relationship between the data, and it is difficult to study the same data structures for integration to effectively access or retrieve data to meet the needs of different data analysis. In this paper, proposed an integration of big data with computation of attribute conditional dependency (ACD) and similarity index (SI) methods termed as ACD-SI. The ACD-SI mechanism allows using of an improved Bayesian mechanism to analyze the distribution of attributes in a document in the form of dependence on possible attributes. It also uses attribute conversion and selection mechanisms for mapping and grouping data for integration and uses methods such as LSA (latent semantic analysis) to analyze the content of data attributes to extract relevant and accurate data. It performs a series of experiments to measure the overall purity and normalization of the data integrity, using a large dataset of bibliographic data from various publications. The obtained purity and NMI ratio confined the clustered data relevancy and the measure of precision, recall, and accurate rate justified the improvement of the proposal is compared to the existing approaches.

Keywords: integration attributes dependency; similarity index; big data

1. Introduction

Data from different domains are collected and integrated at an alarming rate across various data storage available at diverse distributed networks. This enormous growth of information and massive data repository requires innovative methodologies and intelligent tools to comprehend the relationships among each document's property to extract useful information. Due to today's unstructured distribution, developing efficient data aggregation and indexing of relevant information is one of the biggest challenges in the field of information science. Several methodologies are based on predictive networks for document clustering [1], semantic short text analysis [2] and clustering new products through collaborative decision-making [3,4]. Transforming data into knowledgeable information from various structured and unstructured information is a major challenge in the information age. However, most modern business systems often contain a huge number of data records with multiple groups of properties for daily operations performance analysis. However, in business and real-time data analytics, information is still disseminated to multiple sources, which is extremely expensive and makes it more difficult to integrate these sources of multiple data because of the time constraints [6–8]. The subsequent challenge with a large amount of data is to identify structured data objects for quick and accurate access to queries.

The study of data integration provides accurate and computationally efficient data extraction methods to classify data. Depending on the amount, volume, and complexity of data attribute classification and indexing, a big database should be developed to provide a list of tags or names for quick and accurate access to consolidated and structured data. In [5,9,10], several other methods have been proposed to facilitate integration through efficient clustering. The difficulty in grouping these multi-level objects is indicated by the probability distribution that occurs under different conditions.

The major challenges today data integration systems are facing are due to the unstructured form of data and variety of data sources. The diverse forms of data sources are being present at the schema level and these data sources represent to describe the same domain, it is also available at the entity level where the diverse sources can symbolize the identical physical entity in diverse manners. Let's consider a collection of bibliographic data published as an article characterizes with a variety of information, but is substantively associated with topic categories. In the past, several techniques are proposed for analyzing the heterogeneous form of data for integration employing schema mapping [15], document associations with entity references, and entity corresponding, however, integrating multiple source data objects with unstructured data objects can still be challenging majorly due to the there is information description [16] and diversity of properties [17]. To enhance the search operations over vast and complex datasets like big data indexing techniques are being utilized in the scalable distributed storage [18,19]. Manually retrieving these records is impractical and incompetent, so to achieve better and accurate results indexing mechanism more effective for various querying operations [20]. Therefore, effective indexing mechanisms are needed to efficiently retrieve the related information from big data, but the variety of data properties presents create difficulties in accurately indexing data objects.

In recent works [12,21] various authors' utilized un-supervised methods for extensive text processing and attribute discovery. Generally, terminology in text documents is expressed in form of words or phrases which need a lot of memory to store. To select the unidentified attribute it has to depend on certain conditions over a primary space of terminology in terms of subset selection criteria. The unsupervised methods have described preserving the functions of information integration and similarity to gain better performance [14,22–25]. In order to convey similarity in the primary data finest approach is to select the conventional spatial structure of space. In such a case, if the information is identified close to the base information node point, then the point of information closer to the selection is identified concerning their closeness of the properties.

The legitimacy of an efficient, integrated method for on-the-fly processing of big data queries motivates us to propose new approaches for improving big data retrieval. The primary goal of the integration of data from real-time information obtained from multiple fields is to create precious and significant data. Existing data integration technologies use an "extract, transform and load (ETL)" procedure to process large amounts of data. As information sources from many heterogeneous data sources evolve into different qualities of big data, integration with existing data integration technologies presents some challenges. Studies on attribute extraction [14] and selection [11,26] have proposed effective means of integrating un-supervised data into meaningful groups and supporting semantic index-based approaches [27].

This paper aims to propose an efficient integration method called ACD-SI based on the calculation of attribute conditional dependency (ACD) and similarity index (SI) by utilizing data attributes. This proposal will implement an attribute conversion and attributes choosing a method to select the most preferred attribute to perform an efficient integration through multiple attribute value (MAV) analysis of the data. The main contributes the proposal is to do the following achieve enhancement in integration:

1) It mainly implements the existing k-means algorithm to build the most relevant cluster data according to the MAV mean similarity ratio, and the results further use the ACD method to generate the most influential properties to be considered for better integration.

2) It modified the Bayesian mechanism to calculate probabilistic similarities between properties by understanding the conditional dependencies between these properties.

3) It will improve on-the-fly accurate clustering of un-supervised data by providing sparsity in attribute selection and reducing irrelevant attributes in MA data.

4) Later, it implements an attribute-based latent-semantic analysis (LSA) technique to have appropriate data indexing on the aggregated data. Here, it learns data semantic associations with classified data and identifies identified patterns of data attributes as key sets of mapping data attributes, and indicates with similarity indexes to facilitate refinement and accurate searches on big data.

The bibliographic data dataset of research papers collects big data from multiple publication sources and proposes several operations for consolidation [28,29] and indexing [30] to simplify data access. Due to the heterogeneous information in the articles, it seems extremely hard to explore these associated documents of articles because of the improper indexing and integration. In this regard, integration is presuming one of the most important tasks in the field of digital libraries' bibliographic data [31]. In literature, the method of integration is suggested based on attribute patterns and the indexing utilizing LSA by various researchers and scientists as a solution to identify the latest research trends through efficient literature search and articles-associated mining.

Further sections of the paper are presented in 5 sections. The Section 2 of the paper presents the past associated works in relevance and the importance of attribute selection. The Section 3 presents the proposed working principal and modules, and Section 4 presents the evaluation study of the proposed work. In the final Section 5, the conclusion of the paper is discussed.

2. Related works

Data mining based on relevant text format structures seems to be an effective way to determine document compliance, i.e., whether it is relevant or not applicable to the field of data acquisition. Modern technology [32,33] is based on a long-term mechanism, in which training packages are provided to identify relevant features, but there is a limitation of a low level of support. The field of information quality relies on the data integrity and supervision of uncertain information. So, to improve information quality integration is the primary function in the distributed data across various sources. Traditional data collection systems are a combination of limited resources with time-consuming and often have complex design and execution steps.

As discussed in [15], data aggregation systems need to deal with the degree of uncertainty in semantic mapping between the data source and the schema that mediates the indexing of keywords for effective data search efficiency. This means discovering maps in terms of the meaning of the attributes denoting the attributes of schema elements, but it has many challenges in understanding dependable attributes [34] and it has an association for the exact integration of big data [35]. Over the years, cluster analysis has been extensively researched for the effective integration of information with a major focus on distance-based mass spectrometry using "k-means", "k-media" and other methods. However, these methods usually work well when the data are small, but they are often costly or even impossible when dealing with uncertain MAV data [36]. Therefore, it is difficult to find a collection of data documents with the meaning of many properties in an uncertain space. Especially since this data can be very irregular and distorted.

In particular, k-means algorithms [19,37] are frequently used to identify objects based on their simple equal similarity and classification distance, and objects are associated with a "mean" rather than a group cluster. A group of domains whose clusters are assigned to a collection of similar or close properties, such as data points that identify the individual properties. Additionally, the k-means algorithm allows large databases to be consolidated into multiple databases using statistical and categorical values. It is important to note that a cluster can have more than one mode in a k-medium cluster, but the algorithm is highly dependent on the choice of cluster procedure mode.

The various mutable values or properties that illustrate information frequently grow in numerous fields due to the variety of data arrangements such as text objects, photographs, medical, and other mining information [38]. The distribution of information grows by creating multiple values of various properties monitored in various fields. The majority of these attribute values may be exclusive and may define a few relationships but some may be unnecessary and noisy, which greatly affects the choice [39]. This feature is considered as an upper-order quality of conventional learning forms, which is accompanied by deficiencies due to over-equipment, lower effectiveness, and low achievement. Consequently, it is also difficult to find a way to achieve the accuracy and confidence in the desired outcome to decrease pointless and recurring overly related features that need to be identified from different collections of attribute.

Integration is a powerful and computationally accurate method in information mining through data classification. It has to deal with huge information sets concerning the complexity, dimensions and volume of the information attributes [40]. To support efficient and accurate integration various methods have been proposed for multiple domains [15] and uncertain data [34]. The grouping of mutable data objects according to the probability of their similarity distribution was presented in [23]. It shows how difficult it is to combine difficult uncertain data objects according to probability distributions occurring in various circumstances. The distribution-based attribute relationship problem can be efficiently overcome with attribute selection [21,41], and attribute extraction methods [14,26]. This reduces the attribute dimension in group lookup and selection of meaningful attribute sets.

2.1. Significance of attributes and values in integration

The process of attribute selection generally utilizes the method of supervised if learn data available else implements the unsupervised methods. The supervised methods [41] identify the association between the attributes through learning the relevant class information from a collection of attributes. Hence, the majority of research attempts to mine appropriate values for investigation. But the distribution of various data models and grouping of these interrelated data makes it difficult to choose precise attributes [42].

The method of identifying suitable and affluent output attributes is an important value and it is a widely employed method in diverse sources of data analysis and extraction [43,44]. The various applications for learning for selecting attributes consider the techniques of machine learning. Mostly the algorithms can be classified into "supervised algorithms" [45], "semi-supervised algorithms" [46] and "un-supervised algorithms" [15,41] according to the use of information labelling methods [44]. Several past techniques based on sparsity [22] are used to understand relationships between properties. However, the data found can be staggering or very complex for humans to perform. It can be effortlessly classified the various datasets which might have less number of tagged or without tag data. A proposed small-label sample problem [47] illustrates the problem for the supervised algorithms.

This is usually because they do not include sufficient information about records class and simple supervision methods that unintentionally remove specified attributes or fail with improperly selected attributes. Therefore, "semi-supervised property selection" was developed concurrently to utilize unnamed and labelled data. If an appropriate label is not available, it can discover the attribute through selecting the un-supervised element. It is further added complex to evaluate an explicit problem. This attribute evaluates the property of interest in the ability of an element to hold certain properties. High-dimensional markers for identification face labelling costs and, consequently, require competent development and highly autonomous techniques to select un-supervised traits [48–50]. In many real-world applications, the unavailability of classified data and the fast growth of advanced resources make attribute recognition difficult. The concluding result is a capable and automated classification requirement for the selection of un-supervised attributes to address integration issues.

2.2. Data integration using un-supervised attributes

Integration using un-supervised properties, which is difficult to define due to the lack of class/label information defining some of the specific cluster properties according to the criteria specified in the most complex clustering criteria [51,52]. The clustered attribute identification technique uses a functional concept to create a virtual tag and select attributes and create pseudo tags for duplicate data. Most of the difficulties in selecting attributes that are not supervised are solved by using probabilistic model methods. By categorizing these methods into a group of labels using "Grouping by attribute selection" based on passive attribute variables. Visual attributes and user class attributes separately for the origin model for grouping high-dimensional related data are proposed [24]. In the probabilistic model for global integration function and un-supervised attribute selection discussed by the authors [26]. In [14] the authors proposed a transformational inference framework for an “un-supervised non-Gaussian” method for attribute selection.

The studying and selecting MAV solves the problems identified by the description of the attribute vector. Here, each instance of a property is associated with a set of category labels. Current methods [47,49] learn from MAV data by compiling a set of symmetric attributes (e.g., the representation used to identify each RC label). However, each tag has its characteristics, so this general strategy may not be the best option for your solution.

Z. Zhao et al [44] suggested an approach that uses naming attributes to acquire the benefit of segregation in various category class labels called “LIFT” to study MAV data. Initially, it identifies the labels states as positive or negatives for analyzes and then implements the process of learning and testing of the grouped outcomes to builds each label’s specific attributes. A detailed study of 17 datasets proved the benefits of the proposed LIFT and the effectiveness of nominal attributes in comparison to many well-known learning algorithms.

E. Brodley et al. [46] used the mechanism for spectral collection and pseudo-label data to identify attributes for evaluation performance simultaneously. Y. Yang et al. [12] presented the label predicted that a discriminate analysis to form a general structure could be predicted by a “linear classifier decision” to take an example of information input that combines the “$l_{2.1}$-norm minimization” integrated framework for selecting an independent attribute. There is also an assumption based on attributes to enhance the cluster quality, and Y. Tao et al. [53], also perform pseudo-label creation, which involves a different analysis to select unobserved attributes.

M. L. Zhang et al. [43] propose new forces for fuzzy clustering of large and multidimensional data that are convenient for document classification. The main idea of taking into account large and large dimensions when solving problems is to combine confusing collections specially designed for documents to combine effective schemas to solve big problems. Three representative schemes have been identified in a complex group to manage large-scale data such as single-pass, divide-ensemble and sampling-extension. A variety of studies have been conducted in large databases in real-time using these methods, and the results show that these methods are consistently better in document categorization approaches.

An analysis conducted by L. Chen et al. [52] showed that this difficulty reduced the eminence of the aggregate outcomes and suggested a novel correlation technique that encourages standard matrices by recognizing unusual items through cross-group assessment in the dataset. In particular, it is recommended that it accomplish a method based on correlations be utilized to verify essential similarity. They are working to improve the collection of confidential information among different clusters, but as a result, they are competing with existing technologies. Unfortunately, these technologies continue to generate payloads based on incomplete data. Unlikely, these mechanisms create a lot of overheads due to the inconsistency of data information.

The primary matrix information category will give as anonymous relationships despite several dataset records are anonymous. Apply the split chart technique to a weighted double-sided chart consisting of iterative matrices to get the final aggregated result. The core matrix data cluster presents unidentified information sharing, leaving multiple pieces of pieces of evidence in an anonymous database. It uses a two-dimensional weight diagram containing the matrices iterative to obtain a concise result of the split graph method.

Xiang et al. [56] propose a fast unsupervised heterogeneous data learning algorithm in the big data analytics scenarios, namely a two-stage unsupervised multiple kernels extreme learning machine known as TUMK-ELM. It receives information from several sources and learns to display identical information with closed-loop solutions.

It indicates that it is suitable for uncontrolled learning from multiple sources and that other uncontrolled big data that allows for better performance can be adopted in analytics scenarios. It also shows that it improves learning speed by achieving better or more consistent clustering accuracy compared to current clustering methods.

Even though several inconsistencies and separations were acknowledged in the examination of the information group based on the purpose and nature of the data. But there are fewer proposals for analyzing uncertain MAV data [54]. It is due to the unconditional growth of data in multidimensional in terms of its attributes increases the complexity in the prediction of the relation among them and raises a concern to the effective integration among them. This document is intended to provide solutions to improve the integration of data on MAV data for use in retrieving various data for big data use. This paper intends to provide a solution to advance the integration of data about MAV data for use in various data outlets for large data applications. Therefore, the existing method of integration will limit the demand for relevant information, as well as increase the processing advantage and take more time. Thus, determining the appropriate integration model for uncontrolled documents that distribute MAV data is an open issue in big data analysis software.

3. Proposed big data integration methods

This section describes the proposed methods for combining big data based on the calculation of the attribute conditional dependence (ACD) and similarity index (SI) called ACD-SI. It analyzes MAV data to calculate conditional dependencies to predict attribute relationships. The integration process should create a matrix data analysis based on data attribute relationships and introduce new criteria for attribute selection to understand object-based data similarity analysis and data characteristics. Figure 1 shows how to combine data from different sources using the required attributes generated using the ACD and SI methods.

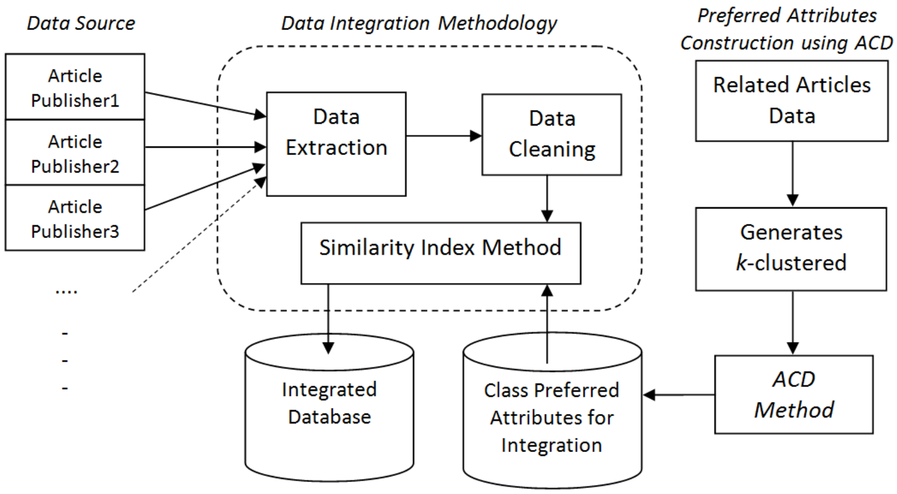


Figure 1. System architecture.

The method of the proposed mechanism functions in two parts. In the first part, it implements *k*-mean and ACD methods to generate the preferred attributes for each class. In the second part, the process of obtaining data from various sources is used and data retrieval and cleaning are initially and later using the SI method to perform data integration utilizing the bibliographic data set. Data sources are collected from distributed sources and are cleaned for the purpose of being stored in large, scalable, and capable storage. The transformation data is performed by using an attribute construction method. It supports the construction of attributes of data transformation to help the mining process by adding new and relevant attributes sets.

3.1. Preferred attributes construction using ACD method

To create the required attributes for bundling it has compiled a relevant dataset for each class based on two or more similar MAV objects. The default grouped data object is used to generate G-number clusters based on the defined set of classes as specified in Table 1.

Table 1. Article class table.

G. Id.	Article Class
G1	*Applied Numerical Mathematics*
G2	*CAD*
G3	*Computer Architecture*
G4	*Computer Graphics*
G5	*Cryptography and Security*
G6	*Genetic Algorithms*

Initial k-cluster results are first constructed for the group of articles given in Table 1 using a traditional k-means algorithm in terms of data relationship properties, and the results obtained are utilized to compute the conditional dependency-based matrix to generate the most it's possible with desirable properties for integration.

The definition of a correlation matrix with the integration method [55] explains the association among individual properties where every property establishes a mutually valuable result. It is generally a connection between a structure formed by a set of graph nodes and a couple of vertices illustrating as edges of groups. It must be kept in mind that graphs with edges require comparative thought among edges and graphs which will collect many clusters within each group to define the vertex as a group category.

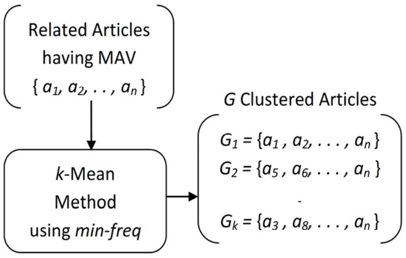

Related Articles having MAV
$\{ a_1, a_2, . . , a_n \}$
k-Mean Method using *min-freq*
G Clustered Articles
$G_1 = \{a_1, a_2, . . . , a_n\}$
$G_2 = \{a_5, a_6, . . . , a_n\}$
-
$G_k = \{a_3, a_8, . . . , a_n\}$

Figure 2. Clustering of related articles.

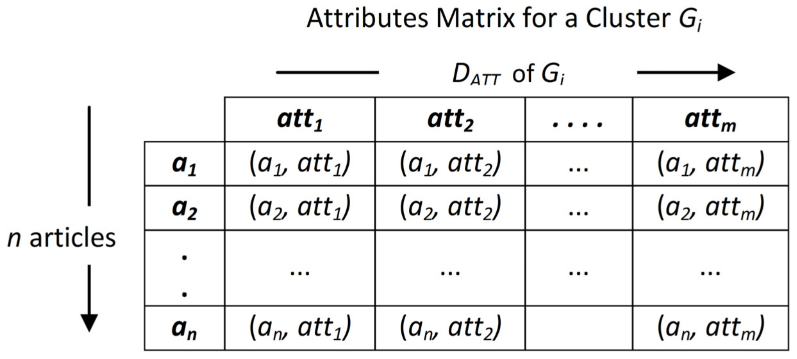
To create, a group of clusters from the MAV datasets employs the conventional *k*-mean algorithm. It takes a collection of data records as input and as per their attribute values and performs an iteration of comparison till it falls into any of the cluster categories. The *k* value of the method is predefined as per the count of grouped categories as *G*. The process of creation is shown in Figure 2.

The above construction of clusters creates a set of data with their relevant attributes concerning the means of similarity among the attribute values. The computation of nearest means similarity among the attributes measures the ratio of attributes similarity between every data record. Let's consider *X* and *Y* are two data records having few sets of unique values then the mean ratio of frequency is given by Eq (1).

$$\text{min\_freq}(X,Y)=\frac{|X_{MAV}\cap Y_{MAV}|}{|X_{MAV}|\cdot|Y_{MAV}|} \tag{1}$$

The outcome of these clustering might have many outliers' records as the mean frequency among these data records might have a low frequency and high differences in comparison to other attribute values. In such cases, it is important to further learn the cluster data to facilitate accurate data in the right cluster it employs the ACD method to extract the most preferable attributes to classify the data objects.

The method ACD constructs an attribute matrix (AM) for each clustered data created using the *k*-mean method. It builds up a set as D_{ATT} of the most frequent and distinct attributes from the cluster having minimum frequency support of more than 1. Now each article attribute of a cluster performs a relation analysis with D_{ATT} as shown in Figure 3.



Attributes Matrix for a Cluster G_i

D_{ATT} of G_i

n articles

	att_1	att_2	....	att_m
a_1	(a_1, att_1)	(a_1, att_2)	...	(a_1, att_m)
a_2	(a_2, att_1)	(a_2, att_2)	...	(a_2, att_m)
. .	...	...	...	...
a_n	(a_n, att_1)	(a_n, att_2)		(a_n, att_m)

Figure 3. Article with attributes relation matrix.

Here it considered two data records supposed to be adjoining if the defined threshold value of probability similarity is met. So, the clustering quality relies on sensitivity to the threshold changes to support the selected correlation approximation. It is specifically is to find a collection of extended and overlapping better-quality attributes that are optimal for the scope of specified test situations, it might be necessary to implement an especially minimum support threshold at the time of the attributes generating period.

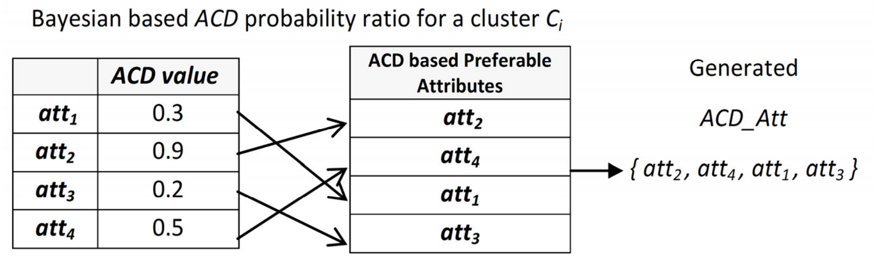
Let's considered that each Cartesian product between a_n and att_m is given by a value $M_{(1,\ 1)}$ to $M_{(n,\ m)}$ using the Eq (2). The product of "data row × data col" will be (0, 1) in the matrix.

$$prob\left(M_{(n,m)}\right) = \prod_{i=0}^{m} \frac{prob\left(a_n(att_m) \cap D_{ATT}\right)}{prob\left(D_{ATT}\right)} \tag{2}$$

where, $p\left(M_{(n,m)}\right) = \begin{cases} 1, & if\ a_n(att_m) \in D_{ATT} \\ 0, & otherwise. \end{cases}$. Now, to compute the probability of ACD concerning each attribute a Bayesian probability ratio is calculated using Eq (3).

$$prob\left(ACD\left(att_m\right)\right) = \frac{\sum_{x=1}^{n} p\left(M_{(n,m)}\right)}{Number\ of\ articles} \tag{3}$$

So, on computing each ACD value using Eq (3) it recognized the most preferable attributes needed for clustering for different MAV data records. The attribute which is having the highest ACD is being considered to have the most preferable attribute as *ACD_Att* for the integrating and can act as a centre point in a relation graph with other attributes. According to Figure 4, the highest preference attribute is att4 and others are dependent attributes which inclusion in order will improve the purity and NMI of the clustering data.


Bayesian based *ACD* probability ratio for a cluster C_i
Generated
ACD_Att
{ att_2, att_4, att_1, att_3 }

	***ACD* value**
att_1	0.3
att_2	0.9
att_3	0.2
att_4	0.5

ACD based Preferable Attributes
att_2
att_4
att_1
att_3

Figure 4. Attributes preference order using CD value.

Based on knowledge of the attribute probability relation it will assess the efficiency of clustering for the different number of feature selection in MAV datasets.

Most of the associated methods in the literature proposed associated the documents based on the frequency of attributes but they didn't evaluate the dependency of attributes on each other. So, the method of ACD using Eqs (2) and (3) construct the required *ACD_Att* set to perform the SI method. It solves the problem of clustering of articles having heterogeneous information, which seems to be complex in the association of the attributes. In the subsequent part, based on an order of *ACD_Att* set it will determine the MAV datasets clustering and utilized to perform classification using SI method to facilitate the integration process as discussed next section.

3.2. Data integration using SI method

To integrate the data records It extracts the attributes to carry out the attribute conversion and selection. The preferred attributes generated using the selected *ACD_Att* are related to classifying. The data record of the article comprises metadata and other details such as title, author and keywords.

To successfully integrate a data record it extracts words from keywords and in an invisible field to form a set of words that will be used to perform SI to classify the article category for merging. The similarity mechanism based on the concept [29] is derived for the data records associations. The conception of such expressions and the metadata of the article are looking to determine its relevance, and it attempts to relate to other elements of the article class.

Let's assume an article class presents a collection of words as, $R = \{w_1, w_2, ..., w_n\}$ for a data record *R*. To initially finding the relation association among these article words and the article class generated *ACD_Att* through a TF (term frequency) computation as, *tf(w)* among the words to another word utilizing the Eq (4).

$$tf(w) = \frac{S_n}{Z}, S \in \text{ACD\_Att}_k \tag{4}$$

where, *S* represents article class A_k associated words, and *Z* represents the summation of words extracted from the record. Later, it calculates the probability of SI for each data record by evaluating the words associated with the class of articles using Eq (5).

$$SI(R{:}\,ACD\_Att_k) = Prob\left(\frac{(w \cap ACD\_Att_k)}{|ACD\_Att_k|}\right) \tag{5}$$

The mechanism of the Naive Bayes (NB) classifier illustrates the dependency of the attribute. It suggests the Condition-Independence of class attributes is completely independent of the variations of the values.

The probability of association among the words *t* of an article is derived as, $pr(t_{(1\to n)} \cap G_{(1\to k)})$, where the *G* is a set of article classes. The maximum probable association of each *G* is given by *pr(G)* from the obtained $pr(G_{1\text{-}k})$ to categorize the record of articles. It utilizes a modified NB to compute the probability of SI in association with terms of *G* for classifying the datasets. Algorithm-1 presents the method of SI function below.

Algorithm-1: Similarity Index Method

```
Input:
Data Record → R;
Group of article class → G [ ];
W [ ] = extractTerm (R) ;
G_Size = sizeOf ( G [ ]);
SI = 0;
for (I = 0;  i < G_Size; i ++)
{
    G_CName = G [ i ] ;
    ACD_Att [ ]  = getCTerms (G_CName );
    //-- Compute the SI value using Bayes probability
    v= prob(ACD_Att [ ] ∩ W [ ] ) ;
    if  ( v >  SI  )
    {
        G = G_CName;
        SI = v;
    }
}
R classified as, G.
```

The method LSA is most frequently being in many indexing applications, so the proposed method of SI inherits the mechanism of LSA and derives the Eqs (4) and (5) to identify the highest similarity of a document is relevant to a class. It will enhance the accuracy of the integration of articles concerning their mapping class using this SI method. The obtained *G* as the output of the algorithm is being utilized to accurately classify data record *R*, which will, in turn, enhance the process of integration.

As per the above discussions, ACD-SI is sensitive to two parameters as the frequency of attributes and SI. The frequency of value states the scope of an article in a cluster whereas SI states the closeness of an article to a particular class. So, the preciseness of an article classification can be controlled by defining each class attributes accurately, because SI computation is dependent on class attribute similarity. In algorithm-1 it computes the SI value using Bayes probability as "*v= prob* (*ACD_Att [ ]* ∩ W [ ])" where *v* is the probable parameter of SI based on which article class is decided. So, the value of *v* can be tuned to have a variation in classification purity in ACD-SI.

Table 3. Collection of articles.

DId	Class	Article Title
c1	CA	A review of dynamic memory with enhanced data access
c2	CA	Implementation of dynamic aspects of the processor symbol hardware compiler
c3	CA	A fault-tolerant multiprocessor architecture for real-time control applications
c4	CA	Functional memory techniques applied to the microprogrammed control of a dynamic associative processor
c5	CA	A machine-oriented memory and processor management architecture
s1	CS	Concerning Certain Linear Transformation Apparatus of Cryptography and Security
s2	CS	Advanced Cryptography Techniques for Computers for Data Encryption
s3	CS	Cryptography Key Management Scheme for Implementing the Data Encryption Standard for Computers Security
s4	CS	Cryptography Architecture for Information Security
s5	CS	A Method for Obtaining Digital Signatures and Public Key Cryptography

Let's consider a dataset having 10 articles. The article has more than one attribute of frequency are categorized to form a cluster. It considers the two-class group as "computer architecture (CA)" and "cryptography and security (CS)" and utilizing the k-mean algorithm it initially creates a cluster of the article as given in Table 2. The next section presents the experimental evaluation of the proposed integration mechanism.

From these articles, attributes are extracted and an attributes relation matrix is formed to build the frequency of occurrence of each article using Eq (2) and probability of ACD concerning each attribute a Bayesian probability ratio is calculated using Eq (3) as given below in Table 4. The ACD ratio values obtained for each article attribute are utilized for constructing the preference order of attributes. So, to perform the accurate integration of articles it later computes the probability of SI in comparison to classes as given in Table 5.

Table 4. Articles with attributes relation matrix.

AId	Attributes	c1	c2	c3	c4	c5	s1	s2	s3	s4	s5	∑	ACD Ratio
a1	Dynamic	1	1	0	1	1	0	0	0	0	0	4	0.8
a2	Memory	1	0	0	1	1	0	0	0	0	0	3	0.6
a3	Architecture	0	0	1	0	1	0	0	0	0	0	2	0.4
a4	Processor	0	1	0	1	1	0	0	0	0	0	3	0.6
a5	Control	0	0	1	1	0	0	0	0	0	0	2	0.4
a6	Cryptography	0	0	0	0	0	1	1	1	1	1	5	1
a7	Computer	0	0	0	0	0	0	1	1	0	0	2	0.4
a8	Key	0	0	0	0	0	0	0	1	0	1	2	0.4
a9	Encryption	0	0	0	0	0	0	1	1	0	0	2	0.4
a10	Security	0	0	0	0	0	1	0	1	1	0	3	0.6

Table 5. Article probability of SI in comparison to classes.

	SI value using Bayes probability (*v*)					
DId	G1	G2	G3	G4	G5	G6
d1	0.214	0.458	0.981	0.124	0.124	0.199
d2	0.127	0.471	0.347	0.199	0.997	0.263
d3	0.487	0.547	0.598	0.547	0.471	0.145
d4	0.347	0.225	0.287	0.158	0.912	0.142
d5	0.588	0.263	0.440	0.192	0.814	0.287

4. Experiment evaluation

4.1. Datasets

It collects the required Bibliography dataset from information extraction and synthesis laboratory. The dataset consists of BibText files approximately 3500 in the count, and a total of 5 million technical article documents published by the different publishers. Processing and analysis are easy when the data is limited, but processing and analysis become very complex when it is large, dissimilar and unstructured. Therefore, the received data file is converted to the particular class format for preprocessing and implies a cleanup method to remove the scrambled data and fill in the missing data.

The acquired data were pre-processed and loaded to the hadoop framework data cluster to perform the experimental analysis. The pre-processing data are stored in the Hadoop cluster and the mechanism of ACD-SI is implemented using Java in Eclipse IDE environment to read and compute the evaluation measures. It initially implements the ACD method to generate the required preferred attributes and later runs the integration method utilizing the SI method. To evaluate the enhancement it measures the following measures in comparison to the existing methods given below.

4.2. Evaluation measures

In the past various works have utilized the purity and NMI measures to evaluate the qualification of the integrated clustering proposals. To measure the transparency and independence of information in the integrated data cluster it also employs the purity and NMI measure [5,54]. These measuring procedures are utilized to evaluate integration precision based on how the cluster objects are related to their essential information of the actual class.

By employing the integration methodology it builds a set of the cluster as |C| having j unit clusters and employing a preferred attribute relation it constructs an exclusive attribute partition as |P| having *i* unit portions over an N number of datasets.

Purity measure: It is a measure to quantify the accuracy of data in clusters to its related class. For each cluster its identify the available data points based on the most preferable attributes and compute the number of data records that are accurately allotted with the exact match to the class. If the outcome of the purity shows high, then the generated clustered is accurate. It is calculated using Eq (6) as given below.

$$Purity(C,P)=\frac{1}{N}\sum_{j}\max_{i}|C_j\cap P_i| \tag{6}$$

NMI measure: It is extensively being used to measure the similarity among the clusters. The range of values of NMI lies between 0 and 1. The highest value indicates the closer similarity of the values and higher reduction of uncertainty, which means 1, is considered as best correlated and perfectly integrated, and 0 indicates no mutual relation with the highest independent of attributes. The NMI measure between integrated clusters can be calculated using Eq (7).

$$NMI(C,P)=\frac{\sum_{i=1}^{I}\sum_{j=1}^{J}|C_j\cap P_i|\log\frac{N|C_j\cap P_i|}{|C_j||P_i|}}{\sqrt{\sum_{j=1}^{J}|C_j|\log\frac{|C_j|}{N}\sum_{i=1}^{I}|P_i|\log\frac{|P_i|}{N}}} \tag{7}$$

It was seen that various data objects in each group of domain differ from other groups in terms of the number of items, attributes, and its values. If the numbers of items and their values in the integrated cluster have a closer relation to class then the measures of NMI and purity performance shows enhanced and accurate to their class category.

To facilitate the effectiveness of the proposed integration technique it measures the estimations of probabilities of documents relevant concerning PR (precision), RC (Recall) and ACC (Accuracy) using the following ratio given below.

$$R=\frac{|\sum True\ Positive\ result\ |}{|\sum True\ Positive\ with\ false\ positive\ Result|}$$

$$RC=\frac{|\sum True\ Positive\ result\ |}{|\sum True\ Positive\ with\ false\ negative\ Result|}$$

$$ACC=\frac{|\sum True\ Positive\ with\ True\ Negative\ result\ |}{|\sum True\ Positive\ and\ Negative\ with\ False\ positive\ and\ Negative\ Result|}$$

4.3. Result analysis

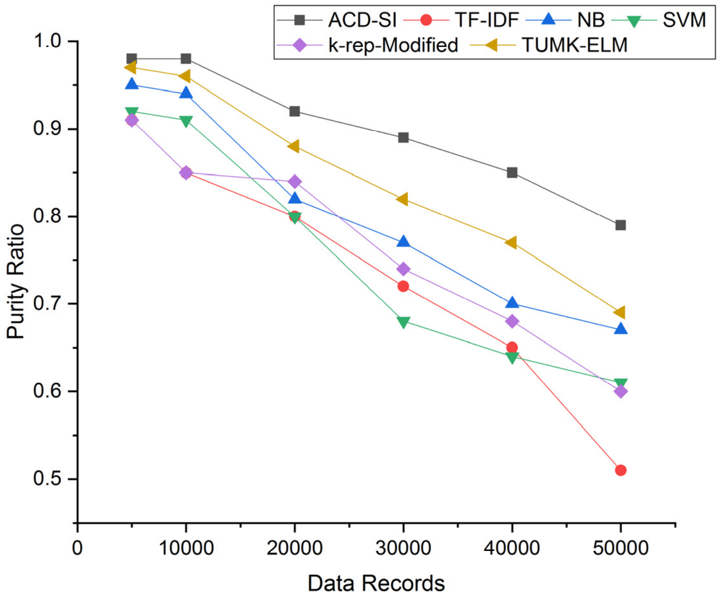


Figure 5. Purity analysis of the integration.

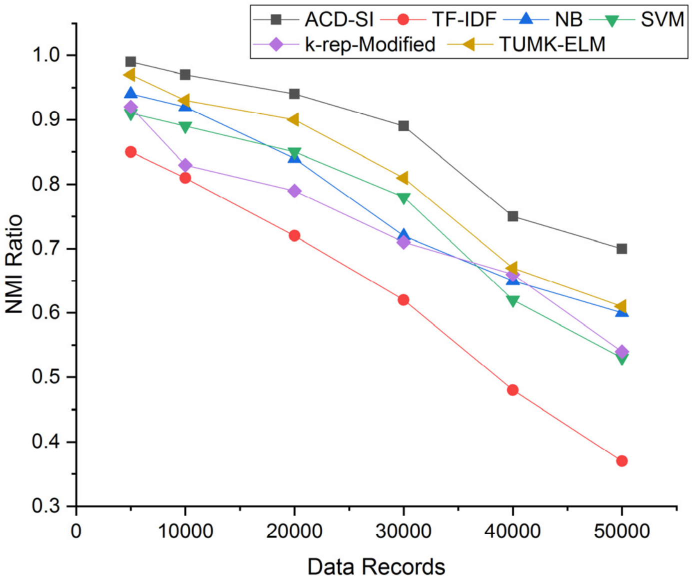


Figure 6. NMI analysis of the integration.

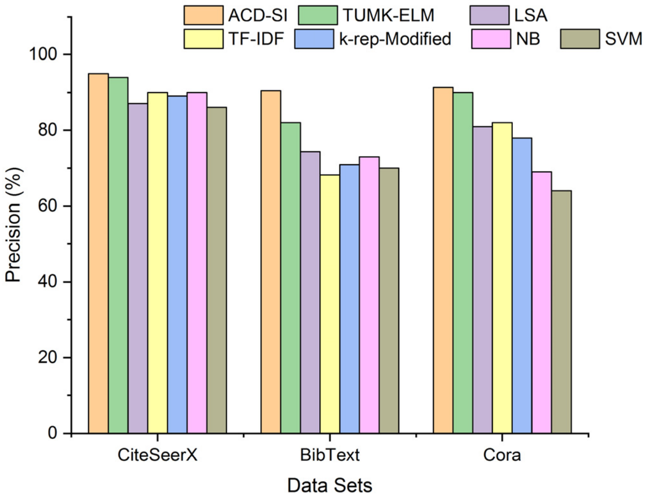


Figure 7. Precision analysis with different datasets.

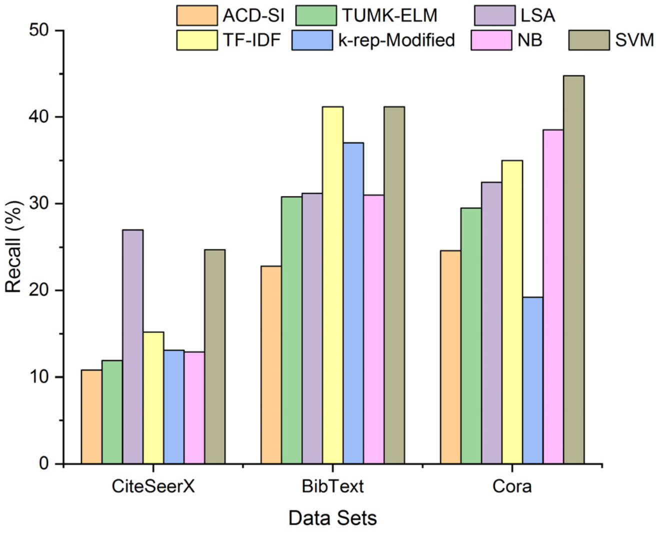


Figure 8. Recall analysis with different datasets.

In this segment, it is discussed that, the evaluation output investigation of the proposed ACD-SI with integration comparing with semantically associating approaches based on the state-of-art method TF-IDF, *k*-representatives-Modified [9], SVM (support vector machines) and NB (naive bayes) to compute the purity and NMI and later discuss the measure of the estimations of probabilities in terms of PR, RC and ACC with the state-of-art method along with LSA and TUMK-ELM method [56] to show the effectiveness of the proposal.

Figure 5 shows the comparison analysis measurement of purity and Figure 6 shows the comparison analysis measurement of NMI. The proposed ACD-SI demonstrated an average of 10% better purity evaluated with existing classifier methods and 5% better than TUMK-ELM based on record association instructions. In the case of increasing evaluation records SVM, TF-IDF and *k*-representatives-Modified methods show a consistent decrease in purity, but the NB and TUMK-ELM method shows a close value of purity in comparison to ACD-SI as both works on the probability of association to predict a class. The enrichment in purity is because of the learning accuracy of correlating the most desired attributes between data records and article classes through probabilistic and semantic similarity attribute models.

The NMI measures the common information being associated among the data record with words learned from the classes of articles in ACD-SI, and the utilization of semantic association mechanism for the process of data integration. The compared method SVM, *k*-representatives-Modified, and TF-IDF shows a drop in NMI because of the high variance of independent words with increasing data records. In case of TUMK-ELM shows a difference of 5% variance on the NMI with increasing records. The TUMK-ELM also generates clusters similar to ACD-SI by the kernel *k*-means. It allows learning a clear cluster boundary for the different classes in the datasets and enhancing clustering performance measures.

The quantify the accuracy of data in clusters initially achieves the most related class due to which it shows a better purity and NMI ratio. The utilization is of ACD able to build a cluster based on the most preferable attributes and compute the number of data records that are accurately allotted with the exact match to the class. The method of ACD measures the similarity among the clusters and data records to indicate the closer similarity of the values and higher reduction of uncertainty. The probabilities of the proposed ACD-SI regarding the relevance of agreements with article classes will increase the amount of data sharing and contributing to achieving better NMIs in comparison. Due to the creation of the most required attributes and the most accurate clustering using the SI method, the anonymous record association is more precise to achieve. Better purity and NMI in comparison.

To have research efficiency and ability of the proposed ACD-SI based integration it compares the different methods in different kinds of datasets related to bibliography as, BibTex, Cora and CiteSeerX, To analyze the efficiency of integration it poses a set of the query to retrieve the data records in relevant to the query. The volume of results obtained is being evaluated statistically in a confusion matrix to visualize the performance of the methods. It represents the values in TP, TN, FP and FN over the top 100 results. So, utilizing these measures it computed the PR, RC and ACC to justify the fairness of improvement.

Figures 7 and 8 show the precision and recall performance for the ACD-SI with others methods. It shows that the proposed ACD-SI shows better precision in comparison. It shows an average of 6% improvisation with CiteSeerX, 15% improvisation with BibText, and 14% improvisation with Cora, similarly in case of recall also ACD-SI shows the least in comparison to others. An average of 5% less with CiteSeerX, 18% less with BibText, and almost 21% less with Cora. This confirms the better precision of integration against all the datasets. Even in comparison to TUMK-ELM, it shows better as ACD-SI qualifies to understand the relation between attributes more precisely than TUMK-ELM. The TUMK-ELM relates the attributes-based probabilities of association which provide a range of gaps towards accurate mapping of the class and affect the performance.

To acquire the outcome of the proposed ACD-SI with other methods has shown in Figure 9. It shows that an average of 9% of improvisation with datasets CIterSeerX, 12% with BibText and 26% with Cora. The improvisation is accomplished due to the extraction of a greater number of relevant results, which introduces the enhancement of the integration and from the relevant results and the total number of actual results indicates accuracy. In comparison to TUMK-ELM and LSA, it shows better accuracy. It is because the information extraction to a query needs to identify the k-nearest objects in relevance. The relevancy of data highly depends on how closely it is mapped to its class. Since the proposed ACD-SI achieves a better purity in the cluster performance, so the accuracy of mapped related data is also high. An experimental result against various bibliographic data sets demonstrates the enhancement of the proposed ACD-SI. So, the improvisation on the accuracy level in comparison to the state-of-art methods with different datasets ensures the fairness of the proposed ACD-SI.

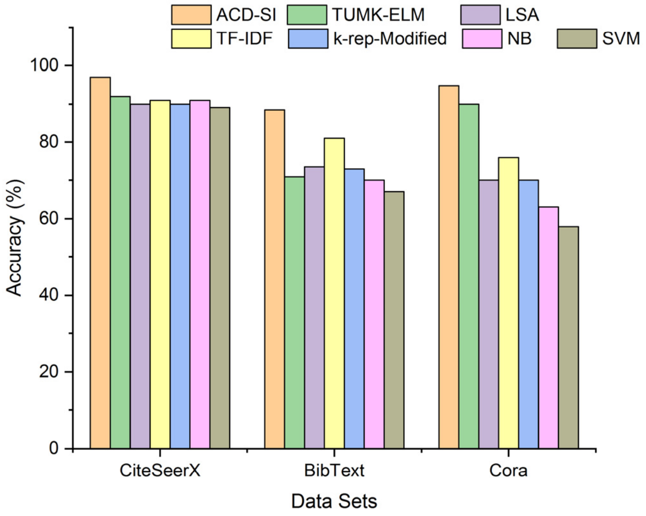


Figure 9. Accuracy analysis with different datasets.

The enhancement of the accuracy in terms of principal component analysis (PCA) is due to the effectiveness of the attribute extraction by reducing the least informative attributes dimension. It has analyzed the PCA over three datasets of CIterSeerX, BibText and Cora using the ACD method to extract the most dependent attributes that are highly correlated to the articles class. It is quite predictable from the obtained measures that the extraction of dependable attributes enhances the improvisation of the integration of heterogeneous data documents.

To illustrate the cost-effectiveness of the proposed ACD-SI method it measures the time complexity using $O(n^2)$ notation. Since the iteration of the execution is depends on the number of attributes extracted and the number of the class available. For example, if an array has 1 element then it has to go with each one iteration of operation, so for 10 classes having every 10 attributes it has to go for 100 iterations for each class. So, with increasing attributes and classes the process complexity with an $O(n^2)$ times and increases linearly. Even though the proposed ACD-SI shows better accuracy results in comparison to others, but with the increasing size of data records, it shows a limitation with a drop fall in the purity. This is due to the high variation of data attributes which makes it more complex to associate to the accurate class and might affect the integration.

The development of unstructured and semi-structured data has entered the rapid information age, where a vast amount of information is from internet sources. This rich and disparate data needs to be properly analyzed and mapped to their related domain to improve the accuracy of data access.

Therefore, there is an immediate need for a technology that can process such structured and unstructured big data effectively and efficiently for the emerging information in this new era. The proposed ACD-SI method which has worked on such structured and unstructured collection datasets shows effectiveness in comparison to the compare classification methods and even shows better integration accuracy. The enhancement in integration is achieved is due to accurately predicting the most preferable attribute and its similarity index. The ACD-SI method of retrieving data in which it identifies data records most similar to related groups formed by integration. This enhancement plays a major contribution towards accessing accurate data mining for information extraction.

5. Conclusions

The two key skills needed to quickly organize and access information in today's big data environment are integration and indexing. This paper proposes a competent integration technique based on conditional dependence (ACD) and similarity index (SI) methods. Selecting the most desired attributes and precisely linking them to SI support to optimize integrations in big data. It generates the optimal attributes for each set of classes using ACD so that the corresponding valid attributes are grouped. It utilizes attribute selection and transformation mechanisms to relate and group data for integration and analyzing the contextual relationships of data attributes using LSA and other methods to provide accurate and appropriate data retrieval. An experimental analysis of the proposal compared to various datasets from various publication sources demonstrates improvisation and NMI purity in the integration, and also shows improvements in accuracy and retrieval accuracy. It shows an average of 10% better purity in compared to state-of-methods and an average precision of 6% with CiteSeerX, 15% with BibText and 14% with Cora along with a low average recall of 5% with CiteSeerX 18% with BibText, and 21% with Cora. It's outperform is justify with the result of accuracy which shows an improvisation of 9% with datasets CIterSeerX, 12% with BibText and 26% with Cora. In the future, the integration testing performance method can be extended unconditionally under the un-supervised method selection requirement and data indexing to simplify integration.

Conflict of interest

The authors declare no competing interests.

References

1. J. Brockmeier, T. Mu, S. Ananiadou, J. Y. Goulermas, Self-tuned descriptive document clustering using a predictive network, *IEEE Trans. Knowl. Data Eng.*, **30** (2018), 1929–1942.
2. W. Hua, Z. Wang, H. Wang, K. Zheng, X. Zhou, Understand short texts by harvesting and analyzing semantic knowledge, *IEEE Trans. Knowl. Data Eng.*, **29** (2017), 499–512.
3. H. Jaber, F. Marle, M. Jankovic, Improving collaborative decision making in new product development projects using clustering algorithms, *IEEE Trans. Eng. Manage.*, **62** (2015), 475–483.
4. K. Yu, L. Tan, L. Lin, X. Cheng, Z. Yi, T. Sato, Deep-learning-empowered breast cancer auxiliary diagnosis for 5GB remote e-health, *IEEE Wirel. Commun.*, **28** (2021), 54–61.

5. T. Iwata, T. Hirao, N. Ueda, Topic models for unsupervised cluster matching, *IEEE Trans. Knowl. Data Eng.*, **30** (2018), 786–795.
6. W. Wang, N. Kumar, J. Chen, Z. Gong, X. Kong, W. Wei, et al., Realizing the potential of internet of things for smart tourism with 5G and AI, *IEEE Network*, **34** (2020), 295–301.
7. Y. Zhang, Y. Sun, R. Jin, K. Lin, W. Liu, High-performance isolation computing technology for smart iot healthcare in cloud environments, *IEEE Internet Things J.*, (2021).
8. W. Wang, X. F. Zhao, Z. G. Gong, Z. K. Chen, N. Zhang, W. Wei, An attention-based deep learning framework for trip destination prediction of sharing bike, *IEEE Trans. Intell. Transp. Syst.*, **22** (2020), 4601–4610.
9. T. Nguyen, V. N. Huynh, A k-means-like algorithm for clustering categorical data using an information theoretic-based dissimilarity measure, in *Folks*, Spring, (2016), 15–130.
10. L. Tan, K. Yu, F. Ming, X. Cheng, G. Srivastava, Secure and resilient artificial intelligence of things: a honeynet a roach for threat detection and situational awareness, *IEEE Consum. Electr. Mag.*, (2021).
11. Z. Li, Jing Liu, Yi Yang, X. Zhou, H. Lu, Clustering-guided sparse structural learning for unsupervised feature selection, *IEEE Trans. Knowl. Data Eng.*, **26** (2013), 2138–2150.
12. Y. Yang, H. T. Shen, Z. Ma, Z. Huang, X. Zhou, L2, 1-norm regularized discriminative feature selection for unsupervised learning, in *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, Spring, (2011), 1589–1594.
13. W. Fan, N. Bouguila, D. Ziou, Unsupervised hybrid feature extraction selection for high-dimensional non-Gaussian data clustering with variation inference, *IEEE Trans. Knowl. Data Eng.*, **25**(2012), 1670–685.
14. H. A. Mahmoud , A. Aboulnaga, Schema clustering and retrieval for multi-domain pay-as-you-go data integration systems, in *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, (2010), 411–422.
15. A. Gani, A. Siddiqa, S. Shamshirband, F. Hanum, A survey on indexing techniques for big data: taxonomy and performance evaluation, *Knowl. Inf. Syst.*, **46** (2016), 241–284.
16. F. Amato, A. De Santo, F. Gargiulo, V. Moscato, F. Persia, A. Picariello, et al., Semtree: an index for supporting semantic retrieval of documents, in *2015 31st IEEE International Conference in Data Engineering Workshops (ICDEW)*, (2015), 62–67.
17. C. Liu, R. Ranjan, X. Zhang, C. Yang, D. Georgakopoulos, J. Chen, Public auditing for big data storage in cloud computing a survey, in *IEEE 16th International Conference on Computational Science and Engineering*, (2013), 1128–1135.
18. J. Wang, S. Wu, H. Gao, J. Li, B. C. Ooi, Indexing multi-dimensional data in a cloud system, in *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, (2010), 591–602.
19. B. B. Cambazoglu, E. Kayaaslan, S. Jonassen, C. Aykanat, A term-based inverted index partitioning model for efficient distributed query processing, *ACM Trans. Web*, **7** (2013), 1–23.
20. Z. Li, Y. Yang, J. Liu, X. Zhou, H. Lu, Unsupervised feature selection using nonnegative spectral analysis, in *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, (2012), 1026–1032.
21. L. Wolf, A. Shashua, Feature selection for unsupervised and supervised inference: The emergence of sparsity in a weight-based a roach, *J. Mach. Learn. Res.*, **6** (2005), 1855–1887.

22. B. Jiang, J. Pei, Y. Tao, X. Lin, Clustering uncertain data based on probability distribution similarity, *IEEE Trans. Knowl. Data Eng.*, **25** (2013), 751−763.
23. D. S. Rajput, S. M. Basha, Q. Xin, T. R. Gadekallu, R. Kaluri, K. Lakshmanna, et al., Providing diagnosis on diabetes using cloud computing environment to the people living in rural areas of India, *J. Amb. Intel. Hum. Comp.*, **4** (2021), 1−12.
24. K. Yu, Z. Guo, Y. Shen, W. Wang, J. C. Lin, T. Sato, Secure artificial intelligence of things for implicit group recommendations, *IEEE Int.Things J.*, **8** (2021).
25. Y. Guan, M. I. Jordan, J. G. Dy, A unified probabilistic model for global and local unsupervised feature selection, in *International Conference on Machine Learning*, (2011), 1073−1080.
26. A. Duric, F. Song, Feature selection for sentiment analysis based on content and syntax models, *Decis. Support Syst.*, **53** (2012), 704−711.
27. T. Do, D. Lam, T. Huynh, A framework for integrating bibliographical data of computer science publications, in *2014 International Conference on Computing, Management and Telecommunications*, (2014), 245−250.
28. T. Huynh, H. Luong, K. Hoang, Integrating bibliographical data of computer science publications from online digital libraries, in A*sian Conference on Intelligent Information and Database Systems*, Springer, (2012), 226−235.
29. K. W. Lim, W. Buntine, Bibliographic analysis with the citation network topic model, in *Asian conference on machine learning*, (2015), 142−158.
30. S. A. Salloum, M. Emran, A. A. Monem, K. Shaalan, Using text mining techniques for extracting information from research articles, in *Intelligent Natural Language Processing: Trends and Alications*, Spring, (2018), 373−397.
31. R. Zhao, K. Mao, Fuzzy bag-of-words model for document representation, *IEEE Trans. Fuzzy Syst.*, **26** (2018), 794−804.
32. V. V. Kolisetty, D. S. Rajput, A review on the significance of machine learning for data analysis in big data, in *Jordanian Journal of Computers and Information Technology (JJCIT)*, (2020).
33. N. Ayat, H. Afsarmanesh, R. Akbarinia, P. Valduriez, Uncertain data integration using functional dependencies, *Amsterdam: Informatics Institute, University of Amsterdam*, (2012).
34. A. Kadadi, R. Agrawal, C. Nyamful, R. Atiq, Challenges of data integration and interoperability in big data, in *IEEE International Conference on Big Data*, (2014), 38−40.
35. X. Pei, C. Chen, W. Gong, Concept factorization with adaptive neighbors for document clustering, *IEEE Trans. Neur. Net. Lear. Syst.***, 29** (2018), 343−352.
36. J. Wu, H. Liu, H. Xiong, J. Cao, J. Chen, K-means-based consensus clustering: a unified view, *IEEE Trans. Knowl. Data Eng.*, **27** (2015), 155−169.
37. J. Zhu, K. Wang, Y. Wu, Z. Hu, H. Wang, Mining user-aware rare sequential topic patterns in document streams, *IEEE Trans. Knowl. Data Eng.*, **28** (2016), 1790−1804.
38. X. He, D. Cai, P. Niyogi, Laplacian score for feature selection, in *Proceedings of the 18th International Conference on Neural Information Processing Systems,* (2005), 507−514.
39. G. T. Reddy, M. P. K. Reddy, K. Lakshmanna, R. Kaluri, D. S. Rajput, G. Srivastava, et al., Analysis of dimensionality reduction techniques on big data, *IEEE Access*, **8** (2020), 54776−54788.
40. D. Cai, C. Zhang, X. He, Unsupervised feature selection for multi-cluster data, in *Proceedings of the 16th ACM SIGKDD International Conference of Knowledge Discovery and Data Mining,* (2020), 333−342.

41. S. M. Basha, D. S. Rajput, A supervised aspect level sentiment model to predict overall sentiment on tweeter documents, *Int. J. Metadata Semantics Ontologies*, **13** (2018), 33–41.
42. J. P. Mei, Y. Wang, L. Chen, C. Miao, Large scale document categorization with fuzzy clustering, *IEEE Trans. Fuzzy Syst.*, **25** (2016), 1239–1251.
43. M. L. Zhang, Lei Wu, LIFT: multi-label learning with label-specific features, *IEEE Trans. Pattern Anal. Mach. Intell.*, **37** (2014), 107–120.
44. Z. Zhao and H. Liu, Spectral feature selection for supervised and unsupervised learning, in *Proceedings of the 24th international conference on Machine learning*, (2007), 1151–1157.
45. X. Li, Y. Pang, Deterministic column-based matrix decomposition, *IEEE Trans. Knowl. Data Eng.*, **22** (2009), 145–149.
46. E. Brodley, J. G. Dy, Feature selection for unsupervised learning, *J. Mach. Learni. Res.*, **5** (2004), 845-889.
47. A. M. Almalawi, A. Fahad, Z. T. Muhammad, A. Cheema, I. Khalil, kNNVWC: An efficient k-nearest neighbors a roach based on various-widths clustering, *IEEE Trans. Knowl. Data Eng.*, **28** (2016), 68–81.
48. D. Ienco, R. G. Pensa, R. Meo, From context to distance: learning dissimilarity for categorical data clustering, *ACM Trans. Knowl. Discov. Data*, **6** (2012), 1–25.
49. O. M. San, V. N. Huynh, Y. Nakamori, An alternative extension of the k-means algorithm for clustering categorical data, *Int. J. Ap. Mat. Comp. Sci.*, **14** (2004), 241–247.
50. L. Chen, Q. Jiang, S. Wang, Model-based method for projective clustering, *IEEE Trans. Knowl. Data Eng.*, **24** (2012), 1291–1305.
51. Natthakan I. On, T. Boongeon, S. Garrett, C. Price, A link-based cluster ensemble a roach for categorical data clustering, *Knowl. Data Eng.*, **24** (2012), 413–425.
52. J. Tang, X. Hu, H. Gao, H. Liu, Discriminat analysis for unsupervised feature selection, in *Proceedings of the SIAM International Conference on Data Mining*, (2014), 938–946.
53. Y. Tao, R. Cheng, X. Xiao, W. K. Ngai, B. Kao, S. Prabhakar, Indexing multi-dimensional uncertain data with arbitrary probability density functions, in *Proceedings of the 31st international conference on VLDB*, (2015), 922–933.
54. X. He, M. Ji, C. Zhang, H. Bao, A variance minimization criterion to feature selection using Laplacian regularization, *IEEE Trans. Pattern Anal. Mach. Intell.*, **33** (2011), 2013–2025.
55. L. Xiang, G. Zhao, Q. Li, W. Hao, F. Li, TUMK-ELM: a fast unsupervised heterogeneous data learning a roach, *IEEE Access*, **6** (2018), 35305–35315.

AIMS Press