A Data Mining Approach to Improve Re-Accessibility and Delivery of Learning Knowledge Objects

Sai Sabitha
Research Scholar, ASET
Amity University, NOIDA,
UP, India
saisabitha@gmail.com

Deepti Mehrotra and Abhay Bansal
Amity University, NOIDA,
UP, India
mehdeepti@gmail.com;
abhaybansal@hotmail.com

Abstract

Today Learning Management Systems (LMS) have become an integral part of learning mechanism of both learning institutes and industry. A Learning Object (LO) can be one of the atomic components of LMS. A large amount of research is conducted into identifying benchmarks for creating Learning Objects. Some of the major concerns associated with LO are size, learning outcomes, pedagogical relevance, and amount of information it delivers to learners. With the advent of knowledge enriched learning, there is a need to create Knowledge Objects (KO) as well and combine these with LOs to create Learning Knowledge Objects (LKO), which can be delivered through an LMS, so that a more holistic knowledge bank is provided to the learners. For an effective LMS, creating a high quality LKO using an algorithm that ensures the delivery of appropriate learning material to the learners is the key issue. Smaller and relevant objects can be delivered to the student using data mining approaches, thereby helping advanced learners to improve their higher order thinking skills. Use of hierarchical clustering techniques for identifying LOs based on user needs is already established. In this paper the Shared Density Approach (SDA) is used to get cohesive clusters and handle cluster of different densities. Finding similar learning objects through clustering technique reduces the domain of search. SDA not only helps with delivery of Learning Objects from a relevant cluster, but also helps in finding objects that are closer to one another but belong to a different class. Objects can be delivered based on user learning approaches, thereby have a wider usage and thus improve re-accessibility.

Keywords: Learning Objects, Knowledge Objects, Clustering, Learning Knowledge Object, K-Nearest Neighbour, Shared Nearest Neighbour, Shared Density Approach.

Introduction

The learning environment is making a paradigm shift with e-learning technology by providing learners with web-based dynamic interfaces, through which learners can access appropriate content. Learning Management Systems (LMS), Content Management Systems (CMS), Virtual Learning Environments...
Delivery of LKO

(VLE) and Knowledge Management Systems (KMS) are tools used to design such an environment. Learning Objects (LO) and Knowledge Objects (KO) are some of the basic building blocks of LMS and KMS respectively. For better management of learning, other instructional characteristics, such as learning objectives, strategy, and assessment, can be attached to these objects whilst they are being delivered.

It is a prime objective that when a LO or KO is delivered to a learner, it should be clearly sign-posted and should be worthy and consistent with the goals of the curriculum and appropriate for the given level of learner according to the learner’s pedagogical need. The learner needs to understand the content given and should be given an opportunity to use it to solve complex real-world problems. Different learners have different needs based on their learning styles and learning approaches (Entwistle, 1998; Felder & Brent, 2005; Keefe, 1979). An individual learner has a different interpretation, relation, and response to learning material, which arise because of different aptitude, intelligence, age, gender, attitude, and motivation.

Based on the knowledge attained by learners, their learning approaches can be divided into three categories. The first kind of learners adopts a surface approach by memorizing facts; the second kind of learners takes a deeper approach and does not simply rely on memorization of the course material but focuses on understanding it. The third group of learners adopts a strategic approach, is well organized and is efficient in their learning (Felder & Brent, 2005). Many advanced learners fall into the third category, in that they follow a surface approach by learning superficially, but, if the need arises they may follow a deeper approach (Lublin, 2003). In a wider range of situations and learning environments, different learning orientations may be adopted (Entwistle 1998). Classification of Learning Objects according to the learning style of potential users, based on profile and experiences, may provide better learning results (Mustaro & Silveira, 2006).

An enhanced LO called a Learning Knowledge Object has already been proposed, for improved learning, which uses the K-Nearest Neighbour approach (Sabitha, Mehrotra, & Bansal, 2014a) and an agglomerative clustering technique (Sabitha, Mehrotra, & Bansal, 2014b). The aim of these is to produce more consistent objects with coherent topic coverage that will satisfy various kinds of learners. An algorithm called the Shared Density Approach, which allows sharing of relevant objects among the clusters, is used in the approach adopted in this paper. This approach not only encompasses the delivery of an object, but also shows how they can be used under various learning categories. Simple clustering techniques find relevant objects in a single cluster and may lose some similar objects as they belong to different clusters thereby reducing its visibility. Each object belongs to only one cluster and an object which happens to be relatively close to another object but belongs to a different cluster can also be accessed using the Shared Neighbor Technique, thus increasing the usage of objects. This approach is more helpful when we use LKO, since KO can now be attached to a greater number of LOs, thereby increasing the re-accessibility of these objects for different user queries.

**Literature Survey**

An LMS is a high-level strategic solution for planning, delivering, and managing all learning events within an organization, which includes creation of online content, instructor led courses, and a portal for collaborative learning (L. Greenberg, 2002). Based on American Society for Training & Development (ASTD’s) definition (Ellis, 2009), LMS refers to a group of software applications that automate the administration, tracking, and reporting of training events. Learning content is referred to as LOs, which are reusable, independent small pieces of information used as building blocks for e-learning content (Gallenson, Heins, & Heins, 2002; Hodgins, 2002; IEEE LTSC, 2002). Many structures and key features like learning objectives, metadata, and goals are proposed for Learning Objects (Gallenson et al., 2002; Griffith, 2003; Smith, 2004; Wagner, 2002; Wiley, 2000), and other features are reusability, interoperability, durability, and accessibil-
ity. Reuse of the LO enables an LO to be used in different ways and in different programmes and is referred to as “any entity, digital or non-digital, which can be used, re-used or referenced during technology supported learning” (IEEE, LTSC, 2002). These LOs are stored in a Learning Object Repository (LOR). LORs help users to search for learning material and they usually provide simple and advanced searches. Simple search results are based on keywords. Advanced searches allow users to specify values for metadata elements and filter the learning material according to the needs of the user. Some well-known LORs are MERLOT (Multimedia Educational Resources for Learning and Online Teaching), EDNA (The Education Network Australia), CAREO (Campus Alberta Repository of Educational Objects), HEAL (Health Education Assets Library) (Roy, Sarkar, & Ghose, 2010) and federated searches provide the facility of searching for learning materials from other repositories. In order to get better functionality of the LMS and LOR various metadata standards are designed, such as Dublin Core Metadata (Dublin, 2012), IEEE Learning Object Metadata, and IMS Global Learning Consortium (IMS, 2006). Metadata can be generated manually by creators and authors, automatically using automated tools or by semi automatic methods. Metadata harvesting and metadata extraction have been identified as two methods of automatic metadata generation (J. Greenberg, 2004; J. Greenberg, Spurgin, & Crystal, 2006).

To improve learning these LOs should have knowledge blended with them, and LORs’ functionalities need to be revisited from a Knowledge Management (KM) perspective (Sampson & Zervas, 2013). To extract knowledge in an organization, various tools and techniques of KMS are used. KM is essentially about facilitating the processes by which knowledge is created, shared, and reused in organizations (Servin & De Brun, 2005), so that an organization can obtain the greatest value from the knowledge available to it. In this context, KOs are defined as a record of information that serves as a building block for KMS. It has content, method of organizing a knowledge base, and rules to identify and categorize knowledge. Researchers have proposed many theories for a KO (Horton, 2001, 2006; Jones, 1990; Merrill, 1999). A KO can be considered a tightly integrated bundle of ideas and related information and experiences, which can be used to teach a topic in an e-learning environment. Converging LO and KO are said to improve the overall learning process (Ruffner & Deibler 2008). Various methods of formation of LKOs and their use in a learning model have been proposed (Ruffner & Deibler, 2008; Štuikys & Damaševičius, 2007; Zouaq, Nkambou, & Frasson, 2007a, 2007b).

Delivery of the desired LO to a learner is another important aspect to be considered. There are many techniques for delivery of LOs, like search based on key words, search based on metadata, agent-based federated catalogs (Barcelos, Gluz, & Vicari, 2011), Recommender Systems (Shen & Shen, 2004; Tsai, Chiu, Lee, & Wang, 2006), Clustering (da Silva & Mustaro, 2009; Sabitha et al., 2014b) and Ranking (Ochoa & Duval, 2008; Sabitha & Mehrotra 2012; Sabitha, Mehrotra, & Bansal, 2012; Yen, Shih, Chao, & Jin, 2010), which results in a more personalised dimension for delivery of objects and object oriented approach (Raghunveer & Tripathy, 2012). Metadata and their standardization helps in more adaptive and flexible delivery of learning content (Yordanova, 2007). Data mining techniques like clustering, classification, and association rule mining were used for courses by Valsamidis, Kontogiannis, Kazanidis, and Karakos (2011). Much research has considered clustering of LOs, for example, da Silva & Mustaro (2009) have proposed clustering of LOs with self-organizing maps. Cluster ensembles as a knowledge reuse framework was proposed by Strehl and Ghosh (2003). A way to improve the content of LOs was proposed by converging it with KOs using agglomerative clustering (Sabitha et al., 2014b). Research on semantic understanding-based approach in clustering LOs has also been carried out.

Clustering is one of the popular data mining techniques that is used for finding interesting patterns from a large data set. A cluster can be defined as a collection of data objects, and cluster analysis is a statistical method to identify and group similar objects into classes. A good clustering method produces high quality clusters with high similarity and interclass dissimilarity. Clus-
ters are formed based on “distance” between points. Two major classes of distance measures are Euclidean (Manhattan distance) and non-Euclidean distance (Jacquard, Cosine). To calculate the proximity (similarity) between two objects, measures like Euclidean distance for two dimensional points are used. Sparse data like documents use Jacquard and Cosine similarity measures. Cosine measure is taken in finding the proximity of the contents of LOs and KOs instead of their metadata. It is calculated as, given two vectors of attributes, $A$ and $B$, the cosine similarity, $\cos(\theta)$, is represented using a dot product and magnitude. Refer to eq 1 of Figure 1.

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad \text{..... eq.1.}$$

**Figure 1: Cosine Similarity Measure**

The popular clustering techniques are hierarchical, partition, exclusive, overlapping, fuzzy, complete, and partial. The various clustering methods are partitioning (K-MEAN), hierarchical AGNES (Agglomerative Nesting), DIANA (Divisive Analysis), Density Based Approach (DBSCAN), Grid based (STING), etc. (Han & Kamber, 2006). The choice of clustering algorithm depends on various factors such as type of a cluster, the character of the data set, noise, number of attributes, and the number of data objects. Many algorithms try to solve the clustering problem to optimise an objective function (Tan, Steinbach, & Kumar, 2007). In most clustering methods each object is assigned to one cluster like K-MEAN or AGNES.

Some of the clustering algorithms used in this paper for delivery of LKOs are the following.

A) **K-Nearest Neighbour** (KNN) one of top algorithms in data mining (Wu et al., 2008) is a non-parametric lazy learner. It does not make any underlying assumption on the data distribution. It is a technique which uses K-instances as represented points in a Euclidean space.

i) In KNN classification, an object is classified by a majority vote of its neighbours, and the object is assigned to the class most common among its $K$ nearest neighbours for discrete value

ii) For real value, it returns the mean values of the K nearest neighbours ($K$ is a positive integer, typically small). If $K=1$, then the object is simply assigned to the class of that single nearest neighbor.

KNN Algorithm is as follows:-

- All instances correspond to points in the n-Dimensional space
- The nearest neighbors are defined in terms of Euclidean distance, Dist $(X_1, X_2)$ or similarity measures
- Target function could be discrete- or real-valued
  - For discrete-valued, KNN returns the most common value among the K training examples nearest to an object.
  - KNN for real-valued prediction for a given unknown tuple. Returns the mean values of the K nearest neighbours.

B) The **Shared Nearest Neighbour** (SNN) similarity is based on a number of shared neighbours as long as the two objects are on each other’s nearest neighbor list. The underlying proximity measures can be similarity or dissimilarity measure. SNN works by a principle that if two points are similar to many of the same points, then they are similar to one another, even if a direct meas-
urement of similarity does not indicate it (Tan et al., 2007). Since the neighbors are considered for similarity measure, an object that is relatively close to another object but belongs to another class can also be retrieved along with the objects of its own cluster. SNN also addresses the problem of clusters of varying density. It depends on the number of neighbours that the two objects share, and does not depend on how far they are from each other. The diagrammatic representation of SNN is shown in Figure 2.

SNN Algorithm is as follows:-
- Find KNN of all the points
- If two points x, y are not among the K nearest neighbour of each other similarity \((x, y) = 0\) else Similarity \((x, y) = \text{number of neighbour shared}\).

C) Clustering is an unsupervised learning technique in data mining. We are considering a density based clustering method called DBCAN (Ester, Kriegel, Sander, & Xu, 1996). The high density regions of objects are surrounded by the low density regions. This algorithm uses the concept of core points, border points, and noise points. Density with a point is the number of points \((\text{min pts})\) in a region of specified radius \((\text{Eps})\) around the point. Points with a density above a specified threshold are classified as core points, while noise points are defined as non-core points that don’t have a core point within the specified radius. Noise points are discarded, while clusters are formed around the core points. If two core points are neighbors of each other, then their clusters are joined. Non-noise, non-border points, which are called boundary points, are assigned to the clusters associated with any core point within their radius.

Density Based Scan (DBSCAN) Algorithm is as follows:-
- Arbitrarily select a point \(p\)
- Retrieve all points density-reachable from \(p\) for a given \(Eps\) and \(Min Pts\).
  - If \(p\) is a core point, a cluster is formed.
  - If \(p\) is a border point, no points are density-reachable from \(p\) and DBSCAN visits the next point of the database.
- Continue the process until all points has been processed.

D) Shared Density Approach (SDA): SNN with clustering (DBSCAN) is an efficient method that identifies the objects which are closest to each other and objects of different classes as well (Tan et al., 2007). Shared Nearest Neighbor and clustering (DBSCAN) together is used to achieve the closeness of the objects in a cluster. This approach is called Shared Density Approach (SDA) and the algorithm is as follows:-
- Compute SNN
- Apply DBSCAN with user specified parameters \(Eps\) and minimum points (\(\text{minpts}\))
E) **Cluster validity:** Cluster evaluation or cluster validation is a part of cluster analysis. There are many evaluation measures like cohesion, separation, correlation etc. Correlation is a cluster validity measure that is based on proximity measures. An ideal cluster is one whose points have a similarity value close to 1, which indicates the objects belong to the same cluster are close to each other. The value of ‘r’ always lies in the range [-1, 1]. A positive ‘r’ indicates a positive association between the objects. A negative ‘r’ indicates a negative association between the objects.

Pearson correlation ‘r’ is computed using the formula as shown in Figure 3.

\[
 r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \quad \text{eq 3}
\]

**Figure 3: Pearson correlation**

If we are going to converge LO & KO to get an LKO, it can be achieved by using KNN approach. The objective is to get cohesive clusters by which a small set of relevant objects are delivered to the users using SDA approach. The objects can be reused for various topics based on the choice of the learner. They can also be delivered based on the learning approaches as discussed earlier. Thus data mining algorithms KNN, SNN and DBSCAN are used.

**Proposed Model**

The proposed model caters for the needs of knowledge enriched learning by the user. Ideally, material that a learner receives in a topic for a subject would be blended with enhanced knowledge. KO of KMS can be used as an instructional object, together with LO. Basically LO comprises an asset (image, text, video, web page) and an information object that teaches a single concept. The smallest level of granularity of an object can be a picture or a text and at the largest level can be a set of courses (Wiley, Gibbons, & Recker, 2000). Instead of delivering one big LO, the proposed model suggests developing Learning and Knowledge Objects of smaller size. SDA, as described in the previous section, will result in re-accessing a similar LO or KO from different cluster of various topics, thereby increasing the referencing and overall reusability of the objects. Most of the LORs use metadata for retrieval of a relevant object. If metadata of the LO and KO are taken into consideration for clustering, then finding the proximity measures that give a meaningful cluster centroid like K-Mean can be chosen. There is a need to find the strongest and closest LOs and KOs, so that only very close objects or cohesive clusters are produced. In this work the content similarity of LOs and KOs are considered for finding the closeness of objects within a cluster, and the cosine similarity measure is used.

**Steps of the Proposed Model**

The proposed model of LKO is given in Figure 4. The process is explained below.

**Creation of LO & KO**

LOs are generated and their metadata are created in LMS. KOs are extracted from the user through KMS, stored in a LOR and are extended with a goal, content, and metadata. The user interface, shown in Figure 5 is for developers and interface in Figure 6 is for learners. Next there is generation of Proximity Matrix between objects using cosine similarity measure (See eq.1, Figure 1).
Generation of K-nearest neighbors

Convergence of LO and KO can be done through classification algorithms of data mining such as decision tree, rule based classifier, naïve Bayes classifier, and KNN. Here the LOs and KOs are classified using KNN. Thus, for each LO we may have one or more associated LOs and KOs which can be further considered as a part of an instructional unit. These can be considered as a LKO.

Shared nearest neighbour

The classified objects, using KNN approach, are further used in generating a number of shared objects using SNN technique.

Cluster formation using DBSCAN (Shared Density Approach)

Matrices generated using SNN are used to cluster using DBSCAN clustering technique (SDA). A set of relevant clusters is identified and the objects are delivered.

Experimental Setup

Step 1. Creation of LO & KO through the User Interface

The Learning Knowledge Management system was developed and the user interfaces are given below. Figure 5 shows an interface for adding the learning objects to the Learning Knowledge System. The Knowledge Objects are collected through the innovation link on the interface. The delivery of LKO’s is shown in Figure 6.
As discussed earlier, there are three types of learning approaches, namely, surface, deep, and strategic. The basic or surface learner can be delivered Learning Objects from a relevant cluster and delivered to a deep learner LKOs from a valid cluster. For strategic learner, LKOs are given, and if the need arises the other relevant objects obtained through SNN are provided.

A set of 60 LOs & 15 KOs is considered from the Learning Knowledge System, which are the classified objects based on user queries for the topic “data mining”. A part of the dataset of classified objects under the topic “data mining” is shown in Figure 7. The KOs that are unstructured are converted into structured by adding metadata to them. The metadata considered is as follows:
KO_Id, KO_Creator, KO_Approver, Date of Submission, Review Date, Status (Approved/Rejected), KO_Topic, KO_SubTopic, KO-Types (Content knowledge, Curricular knowledge, Pedagogical Content Knowledge).

The structured LOs and metadata considered are as follows:
LO_Id, LO_Title, LO_Topic, LO_SubTopic, LO_Content, LO_Category, LO_Learner Type (basic/advance learners).

<table>
<thead>
<tr>
<th>KEY</th>
<th>OBJECT</th>
<th>TOPIC</th>
<th>SUBTOPIC</th>
<th>CATEGORY</th>
<th>CONTENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>LO1</td>
<td>INTRODUCTION</td>
<td>DATA_MINING</td>
<td>DEFINITION</td>
<td>Data mining (knowledge discovery from data)</td>
</tr>
<tr>
<td>2</td>
<td>LO2</td>
<td>INTRODUCTION</td>
<td>KNOWLEDGE_DISCOVŔE</td>
<td>DEFINITION</td>
<td>Data mining is a step in Knowledge discovery process</td>
</tr>
<tr>
<td>3</td>
<td>LO3</td>
<td>INTRODUCTION</td>
<td>DATA_WAREHOUSE</td>
<td>DEFINITION</td>
<td>Data warehouse is a repository which stores data from various sources in an unified format</td>
</tr>
<tr>
<td>4</td>
<td>LO4</td>
<td>KNOWLEDGE_DISCOVERY</td>
<td>PATTERN EVALUATION</td>
<td>DEFINITION</td>
<td>A pattern is interesting if it is easily understood by humans, valid on new or test data with some degree of certainty,</td>
</tr>
<tr>
<td>5</td>
<td>LO5</td>
<td>PRE_PROCESSING</td>
<td>DATA_CLEANING</td>
<td>DEFINITION</td>
<td>Data in the real world is dirty, inconsistent and noisy. So cleaning is required</td>
</tr>
<tr>
<td>6</td>
<td>LO6</td>
<td>PRE_PROCESSING</td>
<td>DATA_INTEGRATION</td>
<td>DEFINITION</td>
<td>Integration of multiple databases, data cubes, or filesCombines data from multiple sources into a coherent store</td>
</tr>
<tr>
<td>7</td>
<td>LO7</td>
<td>PRE_PROCESSING</td>
<td>DATA_TRANSFORMATIC</td>
<td>DEFINITION</td>
<td>Converts the data into appropriate forms for mining</td>
</tr>
</tbody>
</table>

Figure 7: Dataset of LO & KO

Step 2. Generation of Similarity Matrix

To calculate the proximity (similarity) between two objects, a cosine similarity measure is used. The two objects are represented by the two vectors in the two dimensional space. The cosine measure returns the cosine value of the angle \( \alpha \) between the two vectors. When the components of all vectors are greater than or equal to zero, the cosine measure returns values of interval \([-1, 1]\). If two objects are equal, then they have equal vectors. The similarity between the object contents of Figure 7 are generated using the formula given in Figure 1 and are shown in Table 1.

Table 1: Similarity Value

<table>
<thead>
<tr>
<th>Object1</th>
<th>Object2</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>0.2877408</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0.04241506</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>0.12719396</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>0.01491909</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>0.06201117</td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>0.03396251</td>
</tr>
<tr>
<td>1</td>
<td>9</td>
<td>0.12719396</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>0.03076241</td>
</tr>
<tr>
<td>1</td>
<td>11</td>
<td>0.08882356</td>
</tr>
<tr>
<td>1</td>
<td>12</td>
<td>0.10597134</td>
</tr>
<tr>
<td>1</td>
<td>13</td>
<td>0.07352384</td>
</tr>
</tbody>
</table>
Based on the similarity values given in Table 1, a 75 X 75 matrix is generated, as shown in Figure 8.

### Figure 8: Proximity measures of LO& KO in matrix

**Step 3. Generation of K-Nearest Neighbors**

A function for the generation of KNN was developed in “python” language. The number of nearest objects required for an object was defined using a positive integer ‘K’. Various values of ‘K’ were taken into consideration. The cosine value is always between –1 and 1, the cosine of a small angle is near 1, and the cosine of a large angle near 180 degrees is close to –1. The cosine similarity is 0 when the angle between x and y is 90 degrees, and they do not share any terms (words).

Since the cosine similarity measure is used between the LOs and KOs content, the highest distance measures i.e. (closer to 1) are chosen. The function and first six nearest neighbours for each object (ie K=6) are shown in Table 2.

### Table 2: Function & Output of KNN

```
def fun1(a,b):
    w=a.split()w1=b.split()
    c=[]
    for i in range(0,75):
        w[i]=float(w[i])
    c1=[]
    for i in range(0,75):
        w1[i]=int(w1[i])
    for i in range(0,75):
        for j in range(i+1,75):
            if(w[i] < w[j]):
                t=w[i];w[i]=w[j]
                w[j]=t
                t1=w1[i]
                w1[i]=w1[j];w1[j]=t1
            return w1

defcommon_elements(list1 , list2):
    result = []
    for element in list1:
        if element in list2:
            result.append(element)
    return result

# to print first 6 nearest knn
k=1
for i in range(0,75):
    aa=fun1(obj[i],ans)
    for j in range(0,6):
        if aa[j]!= k:
            print aa[j],'	',
        k=k+1
```

256
Step 4. Shared Nearest Neighbour

Based on the KNN matrix shown in Table 2, the number of similar objects of any two points is generated. A function to generate shared objects is shown in Table 3. Since many pairs of objects will have an SNN similarity of 0, this can be a sparse graph.

<table>
<thead>
<tr>
<th>Def common_elements(list1, list2):</th>
</tr>
</thead>
<tbody>
<tr>
<td>result = [] for element in list1:</td>
</tr>
<tr>
<td>if element in list2:</td>
</tr>
<tr>
<td>result.append(element)</td>
</tr>
<tr>
<td>return result</td>
</tr>
</tbody>
</table>

| Table 3: Function & Output of SNN |

```
def common_elements(list1, list2):
    result = []
    for element in list1:
        if element in list2:
            result.append(element)
    return result

def fun2(a):
    count = []
    for i in range(0, 75):
        x = common_elements(a, list1[i])
        z = len(x)
        count.append(z)
    return count

# main module
countlo1 = []
for i in range(0, 75):
    z1 = fun2(list1[i])
    countlo1.append(z1)
```

Step 5. Cluster formation using DBSCAN (Shared Density Approach)

The output of SNN in Table 3 are the inputs for the DBSCAN clustering technique. The algorithm automatically determines the clusters. Different values of KNN, Eps and min points were taken and are shown in the Table 4. The value of ‘K’ was taken as 3, 6, 8 and 15.

| Table 4: Various input Parameters |

<table>
<thead>
<tr>
<th>KNN</th>
<th>Eps (Epsilon)</th>
<th>Minpoints</th>
<th>Figure no.</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>.85</td>
<td>2</td>
<td>Figure. 9, Figure. 10</td>
</tr>
<tr>
<td>15</td>
<td>.49</td>
<td>5</td>
<td>Figure. 9</td>
</tr>
<tr>
<td>8</td>
<td>.75</td>
<td>4</td>
<td>Figure. 9</td>
</tr>
<tr>
<td>3</td>
<td>1.1</td>
<td>2</td>
<td>Figure. 9</td>
</tr>
<tr>
<td>15</td>
<td>.85</td>
<td>2</td>
<td>Figure. 10</td>
</tr>
<tr>
<td>8</td>
<td>.85</td>
<td>2</td>
<td>Figure. 10</td>
</tr>
</tbody>
</table>

The scatter plot of clusters for various measures of Table 4 is shown in Figure 9 & Figure 10.
Radius of Clusters (Eps) and Minimum Points (min points within radius Eps) are the parameters given as inputs to DBSCAN, along with the output of SNN. Various input values were considered as shown in Table 4. The clusters formed are depicted in scatter plot as shown in Figure 9 and Figure 10. Large clusters were generated by DBSCAN algorithm when the value of ‘K’ in KNN
was large, and when the value of ‘K’ in KNN was considerably smaller, cohesive clusters of similar objects within a cluster were obtained.

Result Analysis

Choice of K in KNN

The choice of ‘K’ is critical. A small value of ‘K’ means that noise will have a higher influence on the result. A larger value of ‘K’ makes it computationally expensive, and the basic philosophy behind KNN is that a point that is near might have similar densities. Analysis was carried out for different values of ‘K’ in KNN. DBSCAN algorithm generates the clusters automatically. The ‘K’ value of 6 generated cohesive cluster, and each cluster had similar objects under a particular “topic”. For ‘K’ value set at 15, three large clusters were generated. The ‘K’ value set at 8 generated four clusters, and most of the objects were placed within two clusters as shown in Figure 9 and Figure 10. The ‘K’ value of 3 generated many clusters, but many clusters did not have objects of similar topic. Thus, the value of ‘K=6” was considered and it was validated using correlation measure.

Parameter Determination-DBSCAN

The value of ‘K’ was taken as 3, 6, 8, and 15 (Table 4). Some of the Eps and Min points values chosen for the above-mentioned ‘K’ values generated a single cluster. Determination of a suitable Eps and min points is necessary. Thus, the parameter Eps for DBSCAN algorithm was determined by sorting the distance of the closest point for every point of LO & KO for ‘K’ is 6. The plot in Figure 11, shows sorted distance of $K^{th}$ nearest neighbour on the ‘y’ axis and the LOs and KOs on the ‘x’-axis. The reasonable Eps for ‘K’ as six was chosen between 0.8 and 1. The Eps value “0.85” gave well defined clusters. The second parameter min points of higher value resulted in all points in a single cluster. Keeping the granular size of LKO to be delivered at minimum, the min points chosen were less than 5.

Generation of Clusters for K=6

The output of ‘K’ value at 6 and the clusters generated are shown in Figure 12. The cluster_1 to cluster_8 has strongly related points. The points in cluster_0 are not strongly related.
Correlation Matrix

Cluster evaluation or cluster validation is a part of cluster analysis. An ideal cluster is one whose points have a similarity value close to 1, which indicates the objects belonging to the same cluster are close to each other. Since we have taken the similarity matrix of the objects based on the cosine similarity of the contents of LOs & KOs for finding KNN, the clusters generated by itself is valid. We have used correlation as a measure for cluster validity (See eq.2, Figure 3) in Literature Survey, cluster validity). The correlation between the objects for clusters generated from K = '6' and K = '8' are shown in Figure 13a & Figure 13b.

![Figure 12: Clusters & LKO for K=6](image)

**Figure 13a:** Correlation values of clusters with ‘K’ value = 6

![Table](table)
Based on the correlation measure, the clusters, number of objects in a cluster and their corresponding correlation value is given in Figure 13c. It was found that most of the clusters formed for ‘K’ value of 6 had a high correlation, indicating that points that belong to the same cluster are close to each other compared to a ‘K’ value of 8. Thus ‘K’ was chosen to be 6.

**Case Study**

**Delivery & re-accessing of LKO for keywords “data integration” & “preprocessing”**

The Learning Knowledge Management System provides for the users to choose their topic of interest for learning. The case study given below discusses the objects delivered related to the two topics “DATA INTEGRATION” and “PRE PROCESSING” under the domain “DATA MINING” chosen by the user. The outputs are shown in Table 5.

According to cluster 4, the total number of LKOs delivered for the choice of topic as “DATA INTEGRATION” is three (LO_38, LO_49, KO_62). There are two Learning Objects and one Knowledge Object. From the SNN table, the objects closer to (LO_38, LO_49, KO_62) are (LO_6, LO_27) which belong to other clusters, and can also be delivered to the user.
Similarly, the total number of LKOs delivered for the choice of topic as “PREPROCESSING” is three (LO_5, LO_6, LO_27). There are three Learning Objects. From the SNN table, objects closer to (LO_5) are (LO_7, LO_26, KO_62) which belongs to other clusters and are also delivered. Thus the visibility and re-accessibility of the objects (KO_62) is enhanced (Figure 14).

<table>
<thead>
<tr>
<th>Topic: “DATA INTEGRATION”</th>
<th>Topic: “PRE PROCESSING”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster Output</td>
<td></td>
</tr>
<tr>
<td>PRE PROCESSING</td>
<td>LO_5</td>
</tr>
<tr>
<td>DATA</td>
<td>LO_38</td>
</tr>
<tr>
<td>DATA</td>
<td>KO_62</td>
</tr>
<tr>
<td></td>
<td>PRE PROCESSING</td>
</tr>
<tr>
<td></td>
<td>DATA</td>
</tr>
<tr>
<td></td>
<td>DATA</td>
</tr>
</tbody>
</table>

Clusters and number of objects.

CLUSTER_4: Three Objects

Obj_IDS: LO_38, LO_49, KO_62

No. of Learning Objects

2 (LO_38, LO_49)

Related objects in the same cluster

CLUSTER_4: Two Objects

No of Knowledge Objects

1 (KO_62)

LKO & FIRST FIVE NEAREST OBJECTS

<table>
<thead>
<tr>
<th>LKO_ID</th>
<th>FIRST</th>
<th>SECOND</th>
<th>THIRD</th>
<th>FOURTH</th>
<th>FIFTH</th>
</tr>
</thead>
<tbody>
<tr>
<td>LO5</td>
<td>7</td>
<td>26</td>
<td>38</td>
<td>62</td>
<td>49</td>
</tr>
<tr>
<td>LO6</td>
<td>27</td>
<td>62</td>
<td>38</td>
<td>49</td>
<td>23</td>
</tr>
<tr>
<td>LO27</td>
<td>6</td>
<td>62</td>
<td>38</td>
<td>49</td>
<td>46</td>
</tr>
<tr>
<td>LO38</td>
<td>62</td>
<td>49</td>
<td>6</td>
<td>27</td>
<td>5</td>
</tr>
<tr>
<td>LO49</td>
<td>38</td>
<td>62</td>
<td>6</td>
<td>27</td>
<td>30</td>
</tr>
<tr>
<td>KO62</td>
<td>38</td>
<td>49</td>
<td>6</td>
<td>27</td>
<td>5</td>
</tr>
</tbody>
</table>

Objects delivered

FIRST: LO_38, LO_49, KO_62
SECOND: LO_6, LO_27
THIRD: LO_30, LO_5.

Similar objects

4 (LO_6, LO_27, LO_5, LO_30)

A pictorial representation of the objects delivered related to the two topics “DATA INTEGRATION” and “PRE PROCESSING” is shown in Figure 14.
Conclusion & Future Enhancements

In this paper, the SNN based clustering algorithm Shared Density Approach was chosen to deliver the LKO as it provides a cohesive cluster and can even find clusters in the presence of noise and outliers. These clusters have different shapes, sizes, and density (Ertöz et al., 2003). Earlier approaches such as Partitioned based clustering (K-Mean) were used for delivery, but were found to be unable to handle noisy data and outliers and, therefore, not suitable to discover clusters with non-convex shapes (Dunham, 2006; Han & Kamber, 2006). Delivery through the ward’s criteria applied to hierarchical agglomerative clustering produces smaller size, reasonable clusters. It minimizes the total within the cluster variance, but can never undo what was done previously (Sabitha et al., 2014b). The SDA approach can handle clusters having varying density unlike basic DBSCAN algorithm.

The objective of this research is to provide an efficient delivery of Learning and Knowledge Objects for a student during learning and, thereby, provide an enhanced learning environment. Delivery of objects is achieved through similarity of content by using the above data mining techniques, which resulted in identifying better quality and accurate objects. This clustering technique (SDA) was shown to provide more flexibility, as now the learner can access relevant objects first and can even access the similar objects based on content belonging to different clusters. The objects are re-accessed under various search criteria by the users.

In future work the delivered clusters may be categorised based on their learning/knowledge index, so that most suitable clusters are identified and the objects can be ranked as per the learner type. The clustering technique Fuzzy C-mean allows a single point to belong to one or more clusters, so that each instance is not associated within a single cluster, but has a certain degree of membership for each of the existing centroids, and this may be used as a clustering method to get relevant and reusable objects. Other cluster validity measures such as silhouette coefficient, entropy, and purity can also be used to determine good clusters. The work may be extended for multi-objective clustering where different criteria like learner’s objective, content level, and type of content, can be considered as attributes for clustering.

Figure 14 : Objects delivered for topics “Data Integration” & “Pre Processing”
References


Delivery of LKO


**Biographies**

Mrs Sai Sabitha is BE (CS), ME (CS) in Computer Science and engineering. She is Associate Professor (CSE), KEC Ghaziabad. She is pursuing Ph.D. Programming in Computer Science & Engineering in ASET, Amity University. She has published research papers in conferences & journals. She has over 14 years of Academic and industry experiences. She has handled various B. Tech & M. Tech Projects. Her area of interests are e-Learning, Knowledge Management, Data Mining, Artificial Intelligence & Web Technologies.
Dr. Deepti Mehrotra received her Ph.D. from Lucknow University and currently she is working as Professor in Amity school of Engineering and Technology, Amity University, Noida. Earlier she worked as Director of Amity School of Computer Science, Noida, India. She has more than 20 years of research, teaching and content writing experience. She has published more than 60 papers in international refereed Journals and conference Proceedings. She is editor and reviewer for many books, referred journal and conferences. She is regularly invited as resource person for FDPs and invited talk in national and international conferences. She guided Ph.D. and M.Tech students.

Dr. Abhay Bansal is BE (CS), ME (IT), MBA and PhD. He is Professor & HOD (CSE), ASET & Director, DICET, Amity University, Noida. With over 19 years of Industry and Academic Experience, Dr. Bansal has regularly contributed more than 40 papers in various International journals/conferences. He is a fellow of the Institution of Engineering and Technology (U.K), Sr. Member, International Association of Computer Science and Information Technology. He is a member of ISTE, IEEE (USA), ACM, IETE. He is also a Microsoft Certified Professional and Microsoft approved Technical Associate. He has been an active member and Chairman of the programme committees of several National and International conference/seminar. He is also a reviewer of various National/International Journals of repute.