

Efficient Dynamic Clustering: Capturing Patterns from Historical Cluster Evolution

Binbin Gu
University of California, Irvine
binbing@uci.edu

Saeed Kargar
University of California, Santa Cruz
skargar@ucsc.edu

Faisal Nawab
University of California, Irvine
nawabf@uci.edu

ABSTRACT

Clustering aims to group unlabeled objects based on similarity inherent among them into clusters. It is important for many tasks such as anomaly detection, database sharding, record linkage, and others. Some clustering methods are taken as batch algorithms that incur a high overhead as they cluster all the objects in the database from scratch or assume an incremental workload. In practice, database objects are updated, added, and removed from databases continuously which makes previous results stale. Running batch algorithms is infeasible in such scenarios as it would incur a significant overhead if performed continuously. This is particularly the case for high-velocity scenarios such as ones in Internet of Things applications.

In this paper, we tackle the problem of clustering in high-velocity *dynamic scenarios*, where the objects are continuously updated, inserted, and deleted. Specifically, we propose a generally dynamic approach to clustering that utilizes previous clustering results. Our system, DynamicC, uses a machine learning model that is augmented with an existing batch algorithm. The DynamicC model trains by observing the clustering decisions made by the batch algorithm. After training, the DynamicC model is used in cooperation with the batch algorithm to achieve both accurate and fast clustering decisions. The experimental results on four real-world and one synthetic datasets show that our approach has a better performance compared to the state-of-the-art method while achieving similarly accurate clustering results to the baseline batch algorithm.

1 INTRODUCTION

Clustering is an unsupervised learning method which groups the objects based on some similarity inherent among them into clusters. The clustering problems, such as k-means [34], DBSCAN [20] and correlation clustering [14], have been addressed in many contexts such as anomaly detection [13], database sharding [17], record linkage [26], recommendation systems [25] and others. Although a lot of clustering methods have been proposed and studied (surveyed in [16, 43]), some of them cluster data from scratch—which we will call the *batching approach* hereafter—making them incur a high overhead.

In practice, databases are dynamic. The data in the databases is continuously updated, added and removed from databases making previous clustering results stale. This makes the batching approach unsuitable for these workloads as they incur significant overhead. This is especially the case for high-velocity workloads in scenarios such as Internet of Things (IoT) and large-scale Internet and cloud applications.

Although some dynamic approaches have been proposed for specific clustering methods, these approaches are not general and

may incur high time complexity. For instance, Gan et. al. [21] propose a dynamic algorithm which is customized to the specific density-based method ρ -approximate DBSCAN. Gruenheid et. al. [26] and Whang et. al. [48] focus on finding properties of the specific algorithms that could be leveraged to make lightweight changes to clustering without having to run a batch algorithm from scratch. The main limitation of these approaches is that each variant relies on specific properties and assumptions of a certain clustering algorithm, such as locality and monotonicity. This makes a proposed incremental method not applicable to other instances of the corresponding clustering algorithm and certainly not applicable to other clustering algorithms.

In this work, we make the observation that a batch clustering algorithm makes clustering changes to a specific workload in predictable patterns. Specifically, the pattern of how an existing clustering changes in reaction to an add, delete, or modify operation, can be learned to predict future changes without involving the original batch clustering algorithm. To this end, we propose DynamicC, a machine learning-based system that is augmented with an existing batch clustering algorithm. DynamicC undergoes a training phase, where it observes the cluster evolution patterns while using a batch clustering algorithm. Then, DynamicC is used—instead of the original batch clustering algorithm—to make clustering decisions in reaction to data operations. The benefits of using DynamicC—compared to the original batching algorithm—is that it can be smaller and faster because it learns the clustering patterns for a specific workload. Additionally, DynamicC dynamically learns the cluster evolution patterns without making stringent assumptions about the used batching algorithm. Therefore, DynamicC can be augmented with a wide-range of clustering algorithms with minimal changes unlike existing incremental methods.

The main challenges we tackle in this work is the design of the training and prediction phases of DynamicC. Specifically, it is not trivial to represent cluster evolutions because the space of information that are needed to make decisions and the space of decisions are large. The space of information is the similarity relations between all nodes, and the decision space is all the possible clustering permutations of these nodes. In the paper, we present the details of how we represent the clustering information and decisions efficiently and compactly while enabling us to make accurate predictions. At a high level, we represent clustering information as the changes applied to the database in addition to global similarity metrics. Also, we represent cluster changes in terms of binary decisions on whether a cluster would merge or split in reaction to the new changes on the database.

The compact information and decision representation in DynamicC enables tractable training and prediction of the machine learning model. However, it also leads to the need to do more processing to interpret the general decisions (merge or split a cluster) to specific actions (which nodes should move from one cluster to another.) To this end, we propose heuristics to translate general decision from the machine learning model to specific actions to be

© 2022 Copyright held by the owner/author(s). Published in Proceedings of the 25th International Conference on Extending Database Technology (EDBT), 29th March-1st April, 2022, ISBN 978-3-89318-085-7 on OpenProceedings.org. Distribution of this paper is permitted under the terms of the Creative Commons license CC-by-nc-nd 4.0.

applied to clusters during the prediction phase. These heuristics use the clustering objective function to make the decisions of which nodes should leave or join specific clusters. A nice feature of this heuristic is that using the objective function enables us to know whether the predicted change is accurate or erroneous (if the change leads to a lower objective function score, then we know that the prediction is not accurate and we avoid it.)

In the long run, DynamicC can adjust its machine learning model by retraining continuously on new objects and observing the erroneous predictions during operation. This can be performed while running the original batching algorithm occasionally to establish a baseline for accuracy.

To summarize, DynamicC addresses the the dynamic clustering problem using a ML model that can be augmented and trained with a wide-range of batching algorithms. This overcomes the challenges of prior work that are either batch algorithms that are too expensive for high-velocity dynamic workloads, or incremental algorithms that are restricted to specific clustering algorithms and cannot be extended to other algorithms. Our experiments on real-world and synthetic datasets show that our method can be 85% faster than the state-of-the-art method while retaining accurate clustering results that are within 2% (in terms of F1) of results from the baseline batching algorithm.

The rest of the paper is organized as follows: We first present an overview and related work in Section 2. Then, we state our problem formally and present relevant background in Section 3. We introduce DynamicC’s method to monitor and capture historical clustering evolution in Section 4. We propose DynamicC’s ML model that learns the patterns from the captured historical cluster evolution in Section 5. In Section 6, we introduce the DynamicC algorithms that use the ML model. The experimental evaluation is covered in Section 7. We conclude in Section 8.

2 OVERVIEW AND RELATED WORK

2.1 Overview

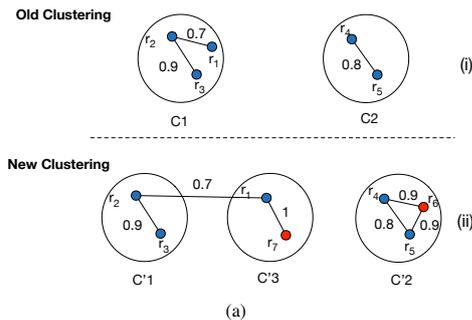


Figure 1: A Motivating Example

In this section, we present an example of using clustering algorithms and our dynamic clustering method to provide a better intuition for the rest of the paper.

The example setup is shown in Figure 1. Assume that there are 5 objects $r_1 - r_5$ initially, and objects r_6 and r_7 are added later in the example.

Clustering can be used to group objects based on their similarity. The “Old Clustering” in Figure 1, for example, is the grouping for the initial five objects based on their similarity. The edges between the objects represent the similarity score, and the absence of an

edge between two objects represents non-similarity. The clustering groups $r_1 - r_3$ together in C1 and groups r_4 and r_5 together in C2.

In the dynamic scenario, objects may be added, modified, or deleted. Consider that after the “Old Clustering” of the first five objects, two objects, r_6 and r_7 , are added. The clustering of objects might change in reaction to the insertion of these two objects. One option to do this re-clustering is to perform a batch clustering algorithm that would compute the groupings from scratch for all objects. This, however, incurs significant overhead, especially for high-velocity dynamic workloads where data operations are continuous. An alternative is to incrementally assign objects to the clusters that are most similar to them—but without changing the structure of clusters. By computing the similarity, we may assign r_7 to C1 and r_6 to C2. However, such a method may not lead to an optimal solution. For example, it is better to split $\{r_1, r_2, r_3, r_4\}$ into C'1 and C'3 as illustrated in the lower part of Figure 1 named “New Clustering”. Existing methods may overcome such inaccurate incremental clustering decisions, however, they require specific knowledge about the properties of the underlying batch algorithm.

DynamicC proposes a dynamic machine learning-based method to allow both fast and accurate clustering decisions while not restricting the design to specific batch algorithms. Specifically, DynamicC would start by running an underlying batch algorithm that makes the clustering decisions during a training phase while observing its clustering decisions. After the training phase, DynamicC have built a model that given the current clustering and the incoming operations would predict the new clustering changes. In this example, assuming that DynamicC is already trained by the time r_6 and r_7 arrive; DynamicC would take as input global similarity metrics of the current clustering—of C1 and C2—and the newly arrived objects, r_6 and r_7 . DynamicC would then make a prediction—using its model—of the changes that would apply to the clusters. For example, DynamicC starts by predicting that r_7 should merge with C1 and r_6 should merge with C2. Then, DynamicC predicts that C1 will split into C'1 and C'3. After this prediction, our heuristics algorithm makes the decisions on which objects from the original C1 would go to C'1 and which objects would go to C'3. In this scenario, as shown in Figure 1 under “New Clustering”, r_2 and r_3 are in C'1 and r_1 and r_7 are in C'3.

A nice feature of DynamicC’s heuristics is that it can check whether a prediction to merge or split would lead to an improvement in clustering accuracy. This is done by computing the underlying batching algorithm’s objective function for the predicted change. If the change yields an improvement in the objective function, then the change is made, otherwise, it is skipped. For example, consider the case in Figure 1 where C1 is predicted to split. The heuristic algorithm checks whether the candidate split clusters would yield a better score for the objective function before applying the prediction.

In the rest of this paper, we provide the details of how DynamicC trains its ML model based on a batching algorithm and how the model is then used to make predictions and clustering decisions.

2.2 Related Work

Clustering (see [16, 43] for a survey) is an unsupervised learning method which groups the objects based on some similarity inherent among them into clusters. Traditional clustering techniques are mainly categorized as partitioning [12, 32, 34], density-based

[10, 20, 44], and hierarchical clustering [29]. Note that this categorization of clustering is neither straightforward, nor canonical, and these categorization groups overlap in practice.

Partitioning Clustering. Partitioning clustering generally constructs a set of k partitions, where each partition represents a cluster which may be represented by a centroid or a cluster representative. Two renowned partitioning clustering are k -means [34] and k -medoids [32]. These two clustering methods differ in how to represent the value of the objects in the cluster.

Density-based Clustering. The idea of density-based methods is to continue growing as long as the number of objects in the neighborhood exceeds some thresholds. Popular density-based clustering methods include DBSCAN [20], OPTICS [10], DP [44] and so on.

Hierarchical Clustering. Hierarchical clustering [29] is a method of cluster analysis which works by building a hierarchy of clusters. Hierarchical clustering methods fall into two types: agglomerative and divisive. While agglomerative clustering proceed by series of merge operations on all the objects into groups until certain conditions are satisfied, divisive methods separate the objects successively into finer (smaller) groups.

Grid-based Clustering. Grid-based clustering quantizes the object space into some segments (also cube, cell, or region) and then do space partitioning based on these segments. A typical example of the grid-based approach is STING [47]. This category is eclectic: it contains both partitioning and hierarchical algorithms.

Objective-based Clustering. Objective-based clustering [12, 14, 26, 34] finds clusters that minimize or maximize an objective function. Typical objective functions in clustering formalize the goal of attaining high intra-cluster similarity and low inter-cluster similarity. The popular objective function based clustering includes k -means [34], balanced k -way partitioning clustering [12], correlation clustering [14], DB-index clustering [26], and so on. There are a lot of variants of the objective function based clustering as the objective function often depends on specific applications or problems. However, finding the optimal solution for such clustering is often NP-hard or intractable. This makes it hard to find a good dynamic algorithm for such clustering problems.

Incremental Clustering and Dynamic Clustering. Incremental clustering is less studied than the general batching approach. In incremental clustering, the focus is on performing clustering incrementally as new objects are added to the database. This is similar to our goal in dynamic clustering where we aim to perform clustering in response to changes (updates, inserts, and deletes) in the database.

Existing dynamic approaches only work for specific clustering methods and may have high time complexity. For instance, Gan et. al. [21] and Ester et. al. [19] propose some dynamic algorithms which are customized to the specific density-based method DBSCAN [20]. Some work (e.g. [15, 38, 46]) focuses on designing new dynamic algorithms that efficiently deal with different tasks. Nevertheless, these works actually belong to “batch algorithms” in our scenario since they do not rely on other existing batch clustering algorithms. Most similar to our work is the proposal by Gruenheid et. al. [26]. That work proposes incremental approaches for re-clustering that are light-weight in comparison to an underlying batching method. However, for their approach to work, they require special properties on the objective function of the underlying clustering algorithm, such as locality and monotonicity. This restricts the applicability of these incremental methods to objective functions satisfying these requirements and excludes

widely-used clustering approaches such as k -means, balanced k -way partitioning and DB-index—that does not satisfy the locality, monotonicity, exchangeability, and separability properties.

Our work, DynamicC, does not make stringent assumptions about the properties of the underlying clustering algorithms. This makes it applicable to a wide-range of clustering solutions unlike existing incremental methods. Some clustering methods are naturally incremental (e.g. hierarchical clustering and grid-based clustering). That is, the incremental or dynamic algorithms for these clustering methods can be straightforward since new updates will not affect previous clustering result. For this reason, we do not evaluate these kinds of clustering methods in our evaluation section and focus on problems that are more challenging to make dynamic.

3 DYNAMIC CLUSTERING

In this section, we begin by formally defining the dynamic clustering problem. Then, we introduce some background about the underlying batching-based clustering problems that will be used in the discussion and evaluation of our work.

3.1 Problem Definition

Clustering Problem. Given a set of objects \mathbf{D} , clustering aims to group data \mathbf{D} into different clusters, where objects in a cluster are similar to each other.

Operations. We define three kinds of possible operations on data and discuss how they lead to triggering dynamic clustering:

- **Adding:** When a new object is added, it may lead to one of the following re-clustering events: the object may join an existing cluster, it may be in a cluster by itself, or it may lead to splitting an existing cluster and then joining one of the resulting split clusters.
- **Removing:** Removing a data object changes the similarity relations between objects in the corresponding cluster. This might lead to splitting or merging with another cluster.
- **Updating:** When an object is updated, this can potentially change the similarity scores between it and other objects. An update has a similar effect to removing data and adding new data which may lead to merging or splitting corresponding clusters.

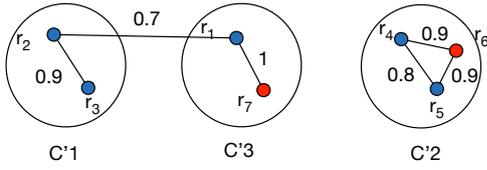
We now formally define the problem of dynamic clustering:

Definition 3.1 (Dynamic Clustering). Given the dataset \mathbf{D} and its clustering results $\mathcal{L}_{\mathbf{D}}$, let \mathbf{D}_{new} be a new set of objects with some updates on \mathbf{D} where $\mathbf{D} \cap \mathbf{D}_{\text{new}} \neq \emptyset$. Dynamic clustering computes the clustering results of \mathbf{D}_{new} based on $\mathcal{L}_{\mathbf{D}}$. The dynamic clustering is denoted by F and its clustering result is denoted as $F(\mathbf{D}, \mathbf{D}_{\text{new}}, \mathcal{L}_{\mathbf{D}})$.

The goal of dynamic clustering is to find an efficient dynamic algorithm F such that we can get the same or similar results compared to that of the batch algorithm. Since an objective function is defined to evaluate the clustering results in this work, the goal of dynamic clustering is to make $F(\mathbf{D}, \mathbf{D}_{\text{new}}, \mathcal{L}_{\mathbf{D}}) \approx B(\mathbf{D}_{\text{new}})$, where $B(\cdot)$ returns the clustering result of the batch algorithm. Dynamic clustering becomes challenging when new updates lead to changes in the clustering structure (clusters merging and/or splitting). Otherwise, the dynamic solution is straightforward as it only involves adding the new data to a cluster.

3.2 Preliminary

Our work in this paper is applicable to any batch clustering algorithm. This is because we do not make assumptions about the



Initial Clustering $\{r_1\}, \{r_2\}, \{r_3\}, \{r_4\}, \{r_5\}, \{r_6\}, \{r_7\}$:
 Step 1: r_2 merges with r_3 , forming $C'1$
 Step 2: r_4 merges with r_5
 Step 3: r_1 merges with r_7 , forming $C'3$
 Step 4: $\{r_4, r_5\}$ merges with r_6 , forming $C'2$

Figure 2: Possible clustering of seven objects from scratch.

objective function and build the machine learning model in DynamicC by observing (and learning) the cluster evolution.

To better illustrate our method, we show how it is integrated with clustering methods that are based on optimizing an objective function (we call these methods *objective-based*). But, we also evaluate our method on other kinds of clustering approaches in the experiment section to show its effectiveness. We focus on objective-based clustering for three reasons. First, these clustering methods (e.g. k-means [34] and correlation clustering [14]) are widely used in various applications. Second, there are a lot of variants for objective-based clustering. This means that generating a dynamic method is needed due to the infeasibility of designing specialized methods for all variants. Third, the batch algorithm of this kind of clustering is often intractable (NP hard or complete) or has high complexity, therefore, a light-weight method for dynamic clustering would offer a significant performance advantage.

In the following, we briefly describe the clustering properties for objective-based clustering that we use in our formulations.

- 1) **Intra-cluster similarity:** For each cluster C , the intra-cluster similarity is the sum of the similarity between objects in the same cluster: $S_{intra}(C) = \sum_{r, r' \in C} sim(r, r')$, where r and r' are any two different objects in the cluster C .
- 2) **Inter-cluster similarity:** For each pair of distinct clusters C and C' , the inter-cluster similarity is the sum of the similarity between objects across the clusters: $S_{inter}(C, C') = \sum_{r \in C, r' \in C', C \neq C'} sim(r, r')$.

Objective Function. To better explain our method with examples, we list the objective function of correlation clustering below. The objective function of correlation clustering—which we will use in some of our example scenarios—is below:

$$F(\mathcal{L}_R) = \sum_{C \in C_R} (1 - S_{intra}(C)) + \sum_{C \in C_R, C' \in C_R, C \neq C'} S_{inter}(C, C') \quad (1)$$

Correlation clustering aims to minimize the complement of intra-similarity and the inter-similarity.

Although different clustering methods use different metrics (e.g. mean) to evaluate their objective function, we focus on the general idea of clustering that (1) objects in the same cluster are similar to each other with intra-similarity, and (2) objects in different clusters are not similar to each other with inter-similarity.

4 MONITORING CLUSTER EVOLUTION

In this section, we present our method to represent and monitor cluster evolution relations while running the underlying batching clustering algorithm. This step is important as cluster evolution information is going to be used to train our ML model that will predict future clustering changes as we show in Section 5. We begin by an overview of cluster evolution operations. Then, we introduce how to extract and represent the cluster evolution information from one round of batch execution from scratch. After that, we present how to represent cluster evolution across batch clustering rounds.

4.1 Cluster Evolution Operations

We define two types of evolution operations: *merge* evolution and *split* evolution. The *merge* evolution means that two clusters merge into one cluster. The *split* evolution means that a cluster is split into two clusters. One evolution can involve more than two clusters. For example, more than two clusters may merge into a cluster when executing a batch algorithm. In such a situation, a change can be divided into multiple merge evolution operations that only involve two clusters. As an example, suppose r_1, r_2 and r_3 are merged together into cluster C . We can represent this merge of three clusters as merging r_1 and r_2 into C' and then merging r_3 with C' . Thus, using an abstraction of a merge/split operation involving two clusters only is sufficient to represent merges and splits involving more than two clusters.

There is another type of evolution called *move* evolution. The *move* evolution means that some objects move from one cluster to another cluster. The *move* evolution is equivalent to the combination of the *split* and *merge* evolution. For example, moving r_1 from $C1$ to $C2$ in the “old clustering” in Figure 1 is equivalent to splitting $C1$ into $\{r_1\}$ and $\{r_2, r_3\}$ first and then merging $\{r_1\}$ with $C2$. Therefore, *merge* and *split* evolution operations are sufficient to express all the possible changes. There is a special situation that splitting a cluster does not improve the score but the move operation does. However, we seldom find such instances which leads to such changes being insignificant for model training.

4.2 Cluster Evolution From Scratch

Here, we begin by showing how cluster evolution can be represented when running a batch algorithm from scratch (the evolution starts from an initial clustering where every object is in a cluster by itself.) Reaching the final clustering result is often a step-by-step process. As illustrated in Figure 2, a batch algorithm may take 4 steps to cluster seven objects from scratch towards the final clustering $\{C'1, C'2, C'3\}$. Each step moves towards a better clustering result. To judge whether a clustering outcome is successful, we compute a score by an objective function. When a batch clustering algorithm terminates, it means that it reached an optimal solution.

Example 4.1. Consider the objects in Figure 2 and apply correlation clustering to them. As described in Section 3.2, the goal is to minimize an objective function F . The similarity among objects is also shown in Figure 2. Having no edge between two objects represents a similarity score of 0. Initially, we have the initial clustering $\mathcal{L}_1 = \{\{r_1\}, \{r_2\}, \{r_3\}, \{r_4\}, \{r_5\}, \{r_6\}, \{r_7\}\}$ (each object is in its own cluster) and $F(\mathcal{L}_1) = 0.9 * 3 + 0.8 + 0.7 + 1 = 5.2$. Suppose r_1 merges with r_7 forming $C'3$, we now have the clustering $\mathcal{L}_2 = \{C'3, \{r_2\}, \{r_3\}, \{r_4\}, \{r_5\}, \{r_6\}\}$. Since $F(\mathcal{L}_2) = 0.9 * 3 + 0.8 + 0.7 + (1 - 1) = 4.2 < 5.2 = F(\mathcal{L}_1)$, it indicates that \mathcal{L}_2 is a better clustering than \mathcal{L}_1 . This process continues until the clustering reaches the optimal clustering.

4.3 Cluster Evolution Across Rounds

As objects are added, removed, and updated, the clustering of objects might become stale. To overcome this, a new round of the clustering algorithm should be performed. A batch clustering algorithm would process every round by clustering from scratch. However, we want to come up with a representation of cluster changes that captures the differences between the cluster results of the current and new clustering. The reason for this is that we want to train our machine learning model to mimic these changes from one round to the next round without having to re-cluster from scratch.

For instance, consider we are monitoring the evolution of a batch clustering algorithm that went through two rounds. In the initial round, the batch algorithm clustered objects r_1 to r_5 and resulted in the clustering shown in Figure 1. This initial clustering can be represented using the method we presented above for clustering from scratch. However, when objects r_6 and r_7 are added, the batch algorithm clusters all the objects again from scratch to arrive at the clustering in Figure 2. The challenge here is that we do not want to capture the steps in this second round of clustering from scratch—instead, we want to capture the steps of the difference between the clustering from Figure 1 to the clustering in Figure 2.

To represent the cluster evolution of the changes only from one clustering to the next, we come up with an algorithm that generates a series of clustering steps (merges and splits) that would transform the old clustering to the new one. The goal of this algorithm is to find a small number of steps to represent this evolution from a clustering to the next efficiently. The algorithm proceeds in the following two phases:

Phase 1: We look for the changes that are relevant to the objects that were added, modified, or removed in this round (e.g., r_6 and r_7 in our example). We take that change and add it to the list (called *transformation list*, of steps to transform from the old clustering to the new one. If an object has undergone multiple changes during a round, then we only add the latest one to the transformation list.

Phase 2: The algorithm checks the rest of the changes that pertain to the remaining (old) objects (objects that were not added, updated, or deleted in this round.) For each cluster c in a change (a merge change contains two clusters and a split change only contains one original cluster and two new clusters), if c contains one or more old objects and it does not exist in the old clustering, we then split the n clusters which have overlapping objects with c in the old clustering. A cluster c' in the old clustering is split into $c' \cap c$ and $c' - (c' \cap c)$. If $c' \subset c$, c' is split into c' and \emptyset . Therefore, we do not really split c' in this case. Finally, we merge these intersection clusters one by one, forming the cluster c and get $n - 1$ merge changes.

The following is an example of how we derive these clustering steps across rounds.

Example 4.2. Consider an example of trying to come up with clustering steps to represent the evolution from the clustering in Figure 1 to the clustering in Figure 2.

Phase 1: Since only steps 3 and 4 are relevant to new objects r_6 and r_7 , we keep them and get rid of steps 1 and 2. Thus, we have the following two changes initially.

Change 1: r_1 merges with r_7 , forming $C'3$.

Change 2: $\{r_4, r_5\}$ merges with r_6 , forming $C'2$.

Phase 2: Change 1 contains two clusters r_1 and r_7 and Change 2 contains two clusters $\{r_4, r_5\}$ and r_6 . Since r_6 and r_7 are not old objects, we do not need to consider them. Because r_1 does not

exist exactly in the old clustering $\{C1, C2\}$, we then split $C1$ into r_1 and $\{r_2, r_3\}$ as $C1 \cap r_1 = r_1$ and $C1 - r_1 = \{r_2, r_3\}$. Therefore, we have the following change.

Change 3: Split $C1$ into r_1 and $\{r_2, r_3\}$.

Since $\{r_4, r_5\}$ is the same as $C2$, i.e. $\{r_4, r_5\}$ exists exactly in $\{C1, C2\}$, we do not need to generate another change.

Note that our algorithm derives the steps to make the transformation from the old to the new clustering but it does not necessarily come up with the right ordering for these steps. This is because such ordering is not necessary for training the machine learning model that trains by observing the changes independently (as we observe in Section 6). In addition, as can be seen in the next section, deriving the order of cluster evolution is intractable or does not help improve the efficiency. Therefore, we adopt a simpler solution to derive the steps of the transformation across rounds.

5 DynamicC ML MODEL

In this section, we present the details of training our machine learning model using cluster evolution history information. Then, we discuss the effectiveness of machine learning models in our scenario. Lastly, we explain how to guarantee that the machine learning model is accurate. We emphasize that our goal is not to design new ML-based batch clustering models or new feature selection methods. Rather, our goal is to propose a light-weight machine learning model to be augmented with an existing batch clustering algorithm to expedite the process of making clustering decision.

5.1 DynamicC Feature Selection

Here, we present the features we use in DynamicC's machine learning model. These features are independent from the underlying batch clustering algorithm and depend only on the observed characteristics of the resulting clustering of objects. We picked these features to describe global characteristics of the clustering to capture the current state without using too much information. The following are these features:

Intra similarity. This feature represents the average similarity between attributes in the same cluster. This measure gives a score of the *cohesion* of a cluster. Therefore, it is a good feature to predict whether a cluster would split or merge.

Maximal inter similarity. This feature represents the maximal similarity among all the average similarity scores between a cluster and the other different clusters. This is a measure of correlation between different clusters. This feature mainly facilitates predicting whether the similarity between clusters would trigger a merge.

Cluster size. This feature represents the size of a cluster. This measure can transform the average similarity of a cluster into the sum of the similarity of a cluster by multiplying by the cluster size.

5.2 Training the Model

The training phase of DynamicC aims to build a machine learning model to enable predicting future cluster changes. This training is performed on a dataset of historical cluster evolution changes that are collected by running a batch clustering algorithm on the desired workload—using the methods in Section 4. All evolution steps are represented via **merge** and **split** operations that involve two clusters only. As we have shown in Section 4, these two operations are sufficient to represent all cluster evolutions.

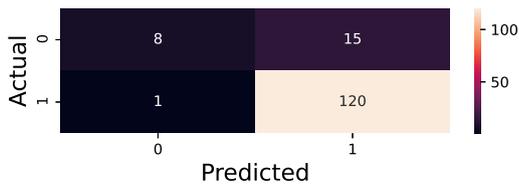


Figure 3: Heatmap of prediction performance on the testing data

An important question in designing and training the machine learning model is how to represent data points in terms of the input to the model and the decisions to be made. A straight-forward representation given our problem is to train the machine learning model to predict whether every pair of clusters should be merged (or split). The problem with this representation is that the number of pairs is high— $n * (n - 1)/2$, where n is the number of clusters. This will produce an intractable machine learning model that is difficult to train as the state space to be learned is large.

Instead, to make the training the machine learning model tractable, we train the model to predict whether a cluster C should be merged or split only. The model needs not predict which clusters should be merged with C or how a cluster C should split (later, we present algorithms to make these decisions given the machine learning outcome and introduce how to guarantee the quality of these solutions in Section 5.4 and Section 6).

Formally, for the *Merge* model, each value in the training data is a 5-dimensional vector $(d_1, d_2, d_3, d_4, d_5)$ about a cluster C :

- 1) d_1 is the average intra-similarity of C and $d_1 \in [0, 1]$
- 2) d_2 is the maximal inter-similarity of C and $d_2 \in [0, 1]$.
- 3) d_3 is the size of C and $d_3 \in [1, +\infty)$.
- 4) d_4 is the size of the cluster C' , where the average similarity between C and C' is maximal for C and $d_4 \in [1, +\infty)$.
- 5) d_5 is the output label which indicates if C should be merged or split and $d_5 \in \{0, 1\}$ where 0 indicates that C should not be merged and 1 indicates that C should be merged.

For the *Split* model, each value in the training data is a 4-dimensional vector (d_1, d_2, d_3, d_5) about a cluster C (what each entry in the vector represents matches the definitions above for *Merge*). The input of *Split* does not have d_4 because split evolution refers to one cluster.

The model is trained separately based on merge evolution and split evolution. That is, the training data of the merge model only contains the merge evolution and the training data of the split model only contains the split evolution.

5.3 Negative Sampling

Instances of cluster evolution are defined as positive samples where the label is set to 1 for training in our case. In order to learn an effective ML model, we need negative samples for training. The negative samples are those clusters which are unchanged when the batch algorithm is performed. Compared with the amount of positive samples, the number of negative samples is often much larger. But we care more about those clusters involved in the connected components since these clusters are checked more times during the execution of a batch algorithm. We call these clusters “active” clusters. Our strategy is to uniformly sample from all negative instances but we give higher weight to those “active” clusters (the weight is set to 0.7 for active clusters and set to 0.3 for non-active clusters in our experiments). We set the number of negative samples to be equal to that of the positive samples for

effective training. In addition, we remove those old samples when the size of training data become too large. This is because the old samples become less and less important in the dynamic scenario and reducing the size of training data could keep the ML model efficient and effective.

5.4 The Objective of the ML Model

In this section, we discuss the objective when training DynamicC’s ML model. Typical ML models train to optimize accuracy. However, we observe that training for recall leads to better results for our problem. In the rest of this section, we discuss this observation and its implications to the training of the ML model.

First, we present the *quality measures* that are relevant to this discussion: accuracy, precision and recall.

Accuracy and Precision. While accuracy means the ratio of the number of correct predictions to the total number of input samples, precision represents how precise your model is out of those predicted positive. As an example, Figure 3 shows a heatmap for model prediction of 144 clusters. Label 0 represents negative predictions and label 1 represents positive predictions. Since 16 (1 + 15) clusters are predicted wrongly and 128 clusters are predicted correctly, the accuracy is $128/(16 + 128) = 0.889$. 15 negative clusters are predicted incorrectly, so the precision is $120/(120 + 15) = 0.89$.

Recall. Recall calculates how many of the true positives (or actual positives) the model captures correctly. So the recall in our scenario is $120/(1 + 120) = 0.992$.

Although we can not guarantee the quality of any of the three metrics of a ML model, these metrics inspire us to think differently of how to apply ML to our problem. We only need to find all the clusters which are ought to change (and we are not affected by finding false positives as we can swiftly check whether these are false positives using the objective function, which we describe next). That is, we are trying to find all positive clusters. This means that recall is the most relevant metric for our problem and we use this observation to influence how we train the DynamicC ML model. In the rest of this section, we present the details of how we avoid false positives and how this leads us to the observation that we ought to train DynamicC’s ML model to maximize recall instead of accuracy.

Avoiding False Positives. Incorrectly predicting the negative clusters as positive clusters can be amended with a swift verification after the prediction. Specifically, we can verify whether we should apply the change that is suggested by the ML model by computing whether the objective score will improve with this change; if it will not improve, then we do not apply that change and the incorrect prediction is avoided. For instance, in Figure2 suppose the ML model predicts that r_7 should merge with $\{r_2, r_3\}$, we can assign r_7 to $\{r_2, r_3\}$ and compute an objective score for the new suggested clustering $\{r_7, r_2, r_3\}, \{r_1\}, C'2$ and compare it with $F(\{C'1, C'2, C'3\})$. Computing an objective function adds an overhead but guarantees the correctness of the decision of a merge change or split change.

Maximizing Recall Instead of Accuracy. Now, our ML model problem becomes about maximizing recall rather than maximizing overall accuracy. Actually, we can reach high recall if there are no other constraints. (As an extreme example, we can make a ML model predict all the samples as positive such that the recall is 100%. However, this strategy would lead to bad accuracy and performance due to having to verify and amend all incorrect predictions.) The strategy we adopt is that we train the ML model

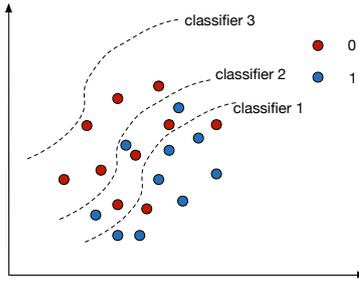


Figure 4: Illustration of different classifiers in 2-D space

to have nearly 100% recall while maximizing accuracy. Training such a model is not straight-forward with machine learning techniques. Next, we explore this question of training for both recall and accuracy.

Trade-off Exploration. We now study how to control the trade-off between the quality and efficiency of clustering.

Usually, a ML model predicts if an item t belongs to a class c with a probability $P(t = c)$. The following equation illustrates how a ML model predicts a binary classification task.

$$\text{Label}(t) = \begin{cases} 1 & \text{if } P(t = 1) \geq \theta \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where t is the item to be predicted. In our case, an item t is a cluster C , therefore we will use $P(C = c)$ instead of $P(t = c)$ in the rest of the paper. Based on this mechanism, we can check which probability values those positive samples have on the training data and set θ as the minimal value of those positive samples. In this way, the probabilities of all positive samples are no less than θ , leading to 100% recall on the training data.

The trade-off between the accuracy and efficiency can be done by setting a different θ value for prediction. The smaller the θ is, the more clusters we need to check. Moreover, the smaller θ often means higher recall and lower precision. Note that we do not change θ value when training the ML model. We manipulate θ only when using the model to control the trade-off between accuracy and efficiency.

Example 5.1. As shown in Figure 4, classifier 1 is the best model in terms of the accuracy, but three positive clusters (or nodes) are missed. Classifier 2 is the best model in terms of recall. It finds all positive clusters and only needs to check 6 more clusters than classifier 1. The recall of classifier 3 is also 100%, but it needs to check 11 more clusters than classifier 1.

6 DYNAMIC ALGORITHM

In this section, we present the algorithm for dynamic clustering that uses the model we developed in Section 5. The algorithm handles three types of operations: add, remove, and update. When an operation is called, the corresponding object incurs initial processing to include it in a cluster. Then, the model runs and predicts whether a merge or split algorithm should be performed. We present the details of initial processing and the merge and split algorithms, followed by the *full algorithm* that puts them all together.

6.1 Initial Processing

When a data operation is performed, we first include the new object information within the clusters based on the type of operation:

- **Adding new objects:** If new objects are added, then we consider each object as a new cluster with one object only.
- **Removing an object:** This operation leads to removing the object from their current cluster.
- **Updating an object:** When an object is updated, it is removed from its current cluster and placed in a new cluster with no other objects. This scenario can be represented as a remove-add sequence.

After the three types of operations are done, we perform the merge algorithm followed by the split algorithm as we describe in the rest of this section.

6.2 Merge Algorithm

The challenge with the merge algorithm is that the ML model only predicts what clusters are candidates for merge operations. However, the model—for tractability reasons we mentioned in Section 5.2—cannot directly predict which cluster it should merge with as the model is a binary classifier. Checking the pairwise clusters takes quadratic time (i.e $O(n^2)$) with the number of clusters. Thereby, pairwise checking can incur significant overhead when the number of candidate clusters has not been reduced significantly.

To reduce this overhead, we make an observation that if two clusters are ought to merge, it is likely that they are both predicted to be merged in our model. This observation leads to reducing the space of candidate clusters. That is, the clusters predicted as “merge” can only merge with those clusters which are also predicted as “merge”. Therefore, the merge algorithm starts by constructing a set Cl_{merge} that contains all clusters that are predicted to merge. Next, we determine which clusters in Cl_{merge} should be merged together. To find this out, we start with a cluster C in Cl_{merge} and find the best candidate to merge with it. This can be done by checking whether merging with every other cluster in Cl_{merge} would lead to a new stable cluster. Specifically, a stable cluster would be one that is not likely to merge again. This can be checked by observing $P(C_{new} = 1)$, where C_{new} is the result of merging C and another cluster C' . The merge that minimizes $P(C_{new} = 1)$ would be the best as it means that it results in the most stable clustering. Checking the effect on $P(C_{new} = 1)$ can be obtained from a ML model in a lightweight way through the coefficients of ML models. In our experiments, we find that the maximal inter similarity and the size of the clusters have respectively high weights in the learned ML model when we do merge predictions.

In the following, we describe the merge algorithm. Denote $Cl_{initial}$ as the initial clustering and Cl_{merge} as the cluster set in which clusters are predicted as 1 by the merge model. The merge algorithm is performed as follows:

- For each cluster $C \in Cl_{merge}$, find the cluster $C' \in Cl_{merge}$ which can maximally reduce the probability $P(C_{new} = 1)$ by merging C and C' . Evaluate if merging C with C' can generate a better clustering by checking the objective score. If so, remove C and C' from Cl_{merge} and merge them in the new clustering. Otherwise, remove C from Cl_{merge} .

Algorithm 1 shows how the merge algorithm works. First, the merge algorithm uses the *merge* model to predict which clusters will be merged and put these clusters in the set Cl_{merge} (line 2). For each cluster $C \in Cl_{merge}$, it checks whether C can merge with the selected cluster in Cl_{merge} except itself (line 5). If C merges with a cluster $C' \in Cl_{merge}$, then C and C' are removed from

Algorithm 1: Merge Algorithm

```
1  $Cl_{merge} \leftarrow \emptyset$ , change  $\leftarrow$  false;
2 Put the clusters predicted as 1 by merge model into
    $Cl_{merge}$ ;
3 while  $Cl_{merge} \neq \emptyset$  do
4   dequeue  $C \in Cl_{merge}$ ;
5   if  $C$  can merge with the selected cluster  $C'$  in  $Cl_{merge}$ 
6     then
7       merge  $C$  and  $C'$ ;
8       remove  $C$  and  $C'$  from  $Cl_{merge}$ ;
9       change  $\leftarrow$  true;
10    end
11  else
12    remove  $C$  from  $Cl_{merge}$ ;
13  end
```

Cl_{merge} (line 6 – 8). Otherwise, it removes C from Cl_{merge} (line 11). This process continues until Cl_{merge} is empty (line 3 – 13).

After the merge algorithm terminates, we perform the split algorithm. We do merge algorithms first because each new object is taken as a new single cluster initially. It is more likely that these clusters can be merged with others rather than being split.

6.3 Split Algorithm

If cluster C is predicted as 1 (i.e. “split”) by the split model, we then check whether splitting it into two clusters would generate a better clustering. Before we introduce the detailed strategy, we first describe the possible scenarios when an existing cluster C will split in response to a new data operations. When a new object arrives, it first joins the closest cluster. Then, the cluster becomes “too big”, containing diverse objects. Therefore, a data r (which is the object most different than the rest of the cluster) in C is increasingly different from objects in the same cluster as C changes into a larger cluster C' . In this case, r should be split out from C . So the heuristic is that the objects which are most different from the other objects in the same cluster will be prioritized to split. If a cluster C is predicted as 1 by the split model, we try to split it in the following steps.

- (1) For each object $r \in C$, denote the inter similarity between $\{r\}$ and $C - \{r\}$ as the weight of r . Rank the objects in C in a decreasing order with their weights and put them in the queue \mathbf{Q} sequentially.
- (2) Check the first object r in \mathbf{Q} , evaluate if splitting r out of C generates a better clustering. If it does, then go to step 3. Otherwise, remove r from \mathbf{Q} and repeat step 2.
- (3) Create a new cluster $C' = \{r\}$ and remove r from C .

If a cluster C is split, then it will generate two clusters $\{r\}$ and $C - \{r\}$. Each time, we just split one object from a cluster for two reasons. First, the *Split* model can be used to predict possible split changes many times in the latter rounds so that we need not worry that some clusters have not been split completely. Second, we observe that most clusters split into two clusters where one of them often has a small size. This is because if new updates cause many objects splitting out from its original cluster C , then this suggests that C does not have high cohesion, thus would have been split earlier.

Algorithm 2 shows how the split algorithm works. The split algorithm first leverages the *Split* model to predict which clusters should be split and put them in the set Cl_{split} (line 2). For each cluster $C \in Cl_{split}$, check whether C be split into two clusters r and $C - \{r\}$ where r is selected according to the strategy presented in the split algorithm (line 5). If yes, then it removes C from Cl_{split} and makes the change status as true (line 6 – 8). Otherwise, it just removes C from Cl_{split} (line 11). This process continues until Cl_{split} is empty (line 3 – 13).

Algorithm 2: Split Algorithm

```
1  $Cl_{merge} \leftarrow \emptyset$ , change  $\leftarrow$  false;
2 Put the clusters predicted as 1 by split model into  $Cl_{split}$ ;
3 while  $Cl_{split} \neq \emptyset$  do
4   dequeue  $C \in Cl_{split}$ ;
5   if  $C$  can split into two clusters  $\{r\}$  and  $C - \{r\}$  then
6     split  $C$  into  $\{r\}$  and  $C - \{r\}$ ;
7     remove  $C$  from  $Cl_{split}$ ;
8     change  $\leftarrow$  true;
9   end
10  else
11    remove  $C$  from  $Cl_{split}$ ;
12  end
13 end
```

6.4 Full Algorithm

The full algorithm of *DynamicC* is shown in Algorithm 3. At a high level, *DynamicC* performs the merge and split algorithms alternately. Initially, it processes the updates according to the approach introduced in Section 6.1 (line 1). This way the algorithm starts with an initial clustering where each new or updated object is assigned to a cluster. Next, it executes the merge algorithm (line 5) and split algorithm (line 6). If some clusters are merged or split, it perform the merge and split algorithms again (line 4 – 7). *DynamicC* terminates when no clusters can be merged or split.

Algorithm Properties. The *DynamicC* algorithm has two properties. First, the *DynamicC* algorithm converges because both of the merge and split algorithms are decreasing the objective score of the clustering. Since there is a minimal objective score for the clustering result and the number of different clustering results are finite, *DynamicC* converges. Second, the complexity of the *DynamicC* is $O(2kn)g(|\mathcal{L}_R|)$, where k is the number of iterations that *DynamicC* needs for converging, n is the number of objects and $g(|\mathcal{L}_R|)$ is the time of evaluating the objective function on clustering \mathcal{L}_R . Note that for both of the merge algorithm and split algorithm, each of them takes $O(n)g(|\mathcal{L}_R|)$ time for each round. Therefore, *DynamicC* algorithm takes $O(2n)g(|\mathcal{L}_R|)$ time. In practice, the number of clusters that are predicted as “merge” or “split” is much smaller than n .

7 EXPERIMENT

We present the experimental results on four real-world datasets and a synthetic dataset. The results show that our method successfully speeds up clustering compared to the underlying batch algorithm while maintaining a clustering accuracy close to the ones obtained by the batching algorithm. We compare with three methods: a batching solution, a naive incremental solution, and a state-of-the-art incremental solution.

Algorithm 3: DynamicC Full Algorithm

Input : Similarity matrix for all the objects;
Initial clustering and the updates;
Merge and split model: $\mathcal{M}_M, \mathcal{S}_M$.

Output : The new clustering Cl_{new} .

```
1 Process the updates (see the approach introduced in the
   beginning of Section 6.1) ;
2  $Cl_{merge} \leftarrow \emptyset, Cl_{split} \leftarrow \emptyset$ ;
3  $change \leftarrow true$ ;
4 while  $change$  do
5   | Execute merge algorithm (Algorithm 1);
6   | Execute split algorithm (Algorithm 2);
7 end
8 return  $Cl_{new}$  ;
```

7.1 Setup

Datasets: We experimented on five datasets. The first real-world dataset, Cora [1], contains 1879 publication objects and has been widely used for clustering. The second real-world dataset, Music Brainz (Music for short) [2], is a collection about music datasets. It contains 19375 objects. The third real-world dataset, Amazon Access Samples (Access for short) [3], is an anonymized sample of access provisioned within the company. It contains 30000 instances. The fourth real-world dataset, 3D Road Network (Road for short) [4], includes 3D road network with highly accurate elevation information. It contains 434874 instances. The fifth dataset is a synthetic dataset. We use the *Febrl* [5] data generator to generate the synthetic dataset. This tool generates a dataset with a number of original and duplicate objects specified by the user. The distribution of the duplicate objects is also set by the user. To experiment with different distributions, we generate three datasets with uniform, poisson and zipf distributions for duplicates. In our experiments with *Febrl*, we start with a dataset consisting of 18K original objects and 30K duplicate objects.

Table 1 summarizes information about the three datasets we use in this section. Our workloads mimic dynamic processes including three kinds of updates: Adding new objects (Add for short), Removing existing objects (Remove for short) and Updating existing objects (Update for short). We explain the details in later sections for each dataset.

Implementations: The experimental environment is a computer with Quad-Core Intel Core i5 processor, 8GB memory, running macOS Catalina. The ML model is implemented in Python with scikit-learn [41] library. The core algorithm is implemented in Java and we execute the python scripts in the java programs. We use the Logistic Regression Model as the default ML model for comparison unless we mention otherwise.

Comparison: We evaluate our method on three kinds of clustering methods of increasing complexity: DBSCAN, k -means and DB-index. We briefly describe the three kinds of clustering problems and then show the reasons about why we evaluate based on the three kinds of clustering.

- **Density-based clustering:** The idea of density-based methods is to continue growing as long as the number of objects in the neighborhood exceeds some thresholds.
- **k -means:** k -means clustering [34] is a classical unsupervised learning clustering method which aims to partition objects into k clusters in which each object belongs to the cluster with the nearest mean. Finding the optimal solution for

k -means clustering is computationally difficult (NP-hard). Although there are many heuristic algorithms [33, 42] for k -means clustering, we use a more robust batch algorithm to test the effectiveness of our dynamic method.

- **DB-index:** Davies-Bouldin index (DB-index) [26] was originally defined for Euclidean space [18]. With some adjustment for the definition of distance, it is adapted to the problem of record linkage [26, 28]. Unlike k -means clustering, DB-index does not assume the number of clusters a priori. Finding the optimal solution for DB-index is also intractable. Compared to correlation clustering [14], DB-index clustering is more challenging because it has no special properties for optimizing clustering [26].

We choose the three clustering problems for two reasons. First, they are representative of difference clustering approaches: density-based, partitioning and objective-based clustering. Second, the complexity of the three clustering methods increases and there is no straightforward dynamic strategy for them. For the above three kinds of clustering problems, we implement two batch algorithms: **DBSCAN** and **Hill-climbing**. Also, we implement one naive incremental method **Naive**, and one state-of-the-art incremental method **Greedy** [26] which is the most similar to our work.

- **DBSCAN:** DBSCAN [20] uses a minimum density level estimation, based on a threshold for the number of neighbors, $minPts$, within the radius ϵ . Objects with more than $minPts$ neighbors within this radius are taken as a core point.
- **Hill-climbing:** This is a general batch algorithm which can be used for any objective function based clustering method [45]. It examines all immediate neighbors (potential migrations) and selects the clustering update providing the highest improvement. This method is applicable to many other clustering methods as it does not require any ad-hoc rules or properties of the clustering methods.
- **Naive:** This is the baseline incremental algorithm. It compares each new object with existing clusters and then assign an object to the closest cluster or a new cluster. This method does not compute the objective score for the clustering. Its decisions are only based on heuristics such as similarity threshold.
- **Greedy:** This method is proposed in [26]. It uses three operators to determine a candidate clustering which makes it able to terminate in polynomial time.

DBSCAN is only applied for density-based clustering. **Hill-climbing**, **Naive** and **Greedy** are applied to k -means and DB-index clustering in the following comparisons.

The accuracy of **DynamicC** depends on the clustering that it starts with. For this reason, we provide evaluations with two scenarios: (1) *GreedySet* scenarios: in this case, the starting point for each round of updates is set to the results of **Greedy** from the previous clustering so that **Greedy** and **DynamicC** start with the same clustering; (2) *DynamicSet* scenarios: in this case, the starting point is the dynamic clustering results from the previous clustering. In a real deployment, **DynamicSet** matches the typical execution, while **GreedySet** matches the case when the underlying batch algorithm is run occasionally. To generate the initial cluster evolution for **DynamicC**, we use the Hill-climbing method.

For correlation clustering, its property has been well studied in [26]. The **Greedy** approach proposed in [26] shows its effectiveness with the correlation clustering and most of the experimental

Table 1: Experimental settings on the datasets

| Dataset | similarity/distance | # of initial objects | # of final objects | data type |
|--------------------------------|--------------------------------|----------------------|--------------------|-----------------------|
| Cora | Jaccard [40] | 279 | 1879 | textual and numerical |
| Music Brain (Music) | Cosine Trigram similarity [39] | 4K | 15375 | textual |
| Amazon Access Samples (Access) | Euclidean distance | 1K | 20208 | numerical |
| 3D Road Network (Road) | Euclidean distance | 100K | 344768 | numerical |
| Synthetic | Levenshtein [49] and Jaccard | 10K | 43K | textual and numerical |

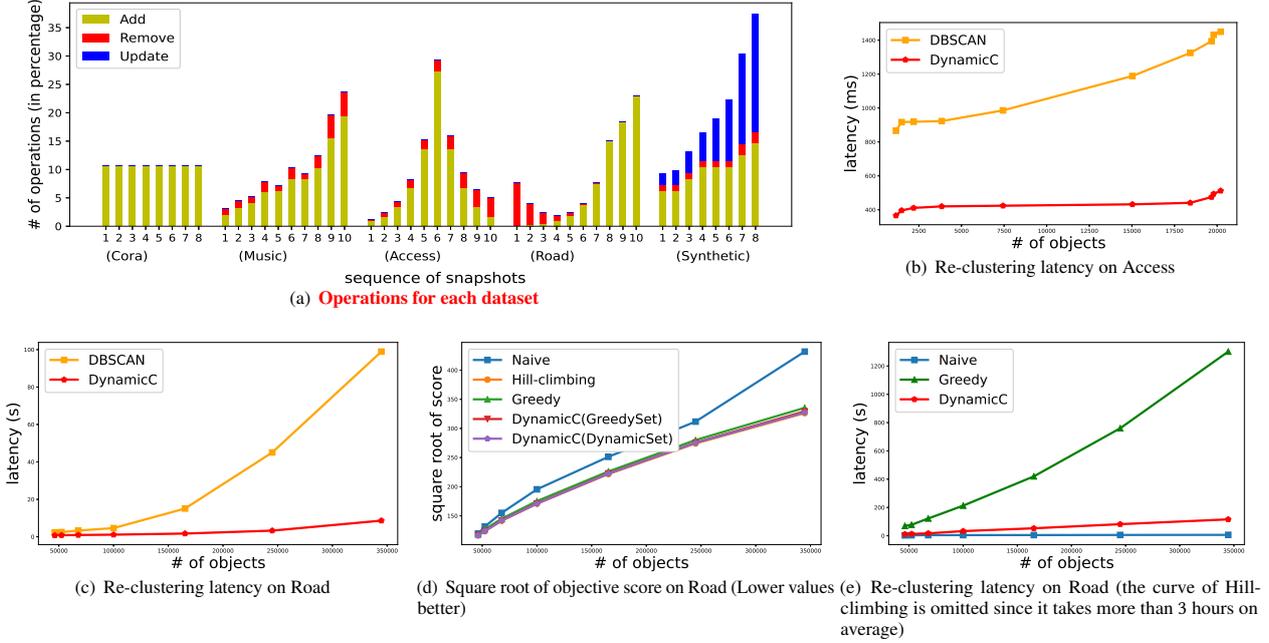


Figure 5: Dynamic process on the datasets and the results for density-based and k-means clustering

results are based on the correlation clustering. However, as described in [26], DB-index clustering is more challenging because it does not have some desirable properties that can be used for optimization.

Measurement: We evaluate our method in terms of latency and quality of clustering results. For latency, we report the average execution time of 20 runs of the experiments. In each round, we focus on the latency of re-clustering as well as the model re-training time. In our experiments, the reported latency is of both the re-training and re-clustering. For measuring the quality of clustering, we report the objective score and pair counting $F1$ measure [7], precision, recall, purity and inverse purity. As the goal of dynamic clustering is to get the similar clustering result with that of the batch methods, we take the clustering result of the batch method as the ground truth.

7.2 Quality and Performance

More details about the experimental settings on each dataset is presented in Table 1. To mimic the dynamic process, we randomly select some objects from the original dataset as the initial dataset in the beginning. Next, at each snapshot (round), we randomly add, remove or update some objects and start re-clustering. The number of added, removed and updated objects for each dataset is shown in Figure 5(a). We only perform the Update operations on the Synthetic dataset by the Febrl Generator since Febrl allows us to generate similar objects as well as do modifications to attribute

values. Update operation can be factorized into Remove and Add operation, so it is fine with only Remove and Add operations. Since DBSCAN and k-means often work for numerical data, we only report the results on Access and Road datasets for them. And we report the results on Cora, Music and Synthetic datasets for DB-index clustering.

7.2.1 Results for DBSCAN. Although **DynamicC** is designed with objective function-based clustering in mind, we test **DynamicC** to show that our method can be augmented with many other clustering methods with minor changes. Since DBSCAN has no objective function, we judge if the clustering is good or not by checking whether the relevant previous core points are stable.

In DBSCAN, changes in the clusters are based on two parameters: minPts and ϵ for each new or old object. We notice that **DynamicC** does not work well when the objects in the final cluster are very sparse (i.e. objects in a cluster are not compact). Different minPts and ϵ settings can significantly affect the time cost and clustering result for DBSCAN. We report the average $F1$ and latency with 10 groups of parameters. The average $F1$ scores on Access and Road are 0.988 and 0.976, respectively, across all the snapshots for **DynamicC**. **DynamicC** does not achieve the same results as DBSCAN since we do not check each object and do not follow its exact rules. However, **DynamicC** can still get good $F1$ score faster as shown in Figure 5(b) and (c). In addition, **DynamicC** does not always get better clustering result than manually designed methods like [19] (while [21] focuses on ρ

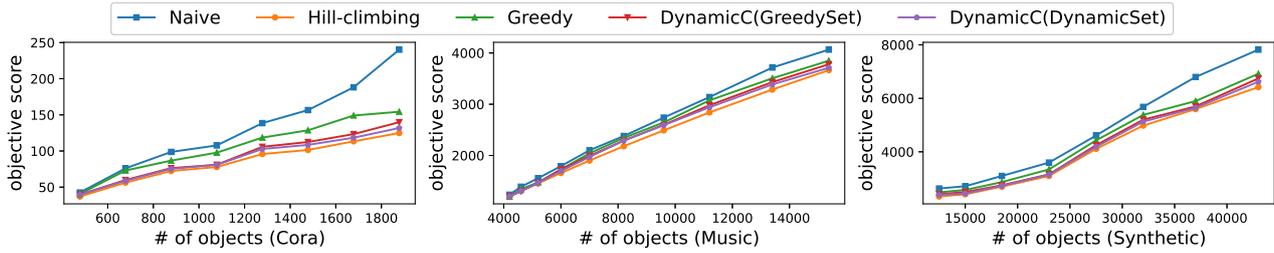


Figure 6: Quality on Cora, Music and Synthetic datasets (Lower values are better)

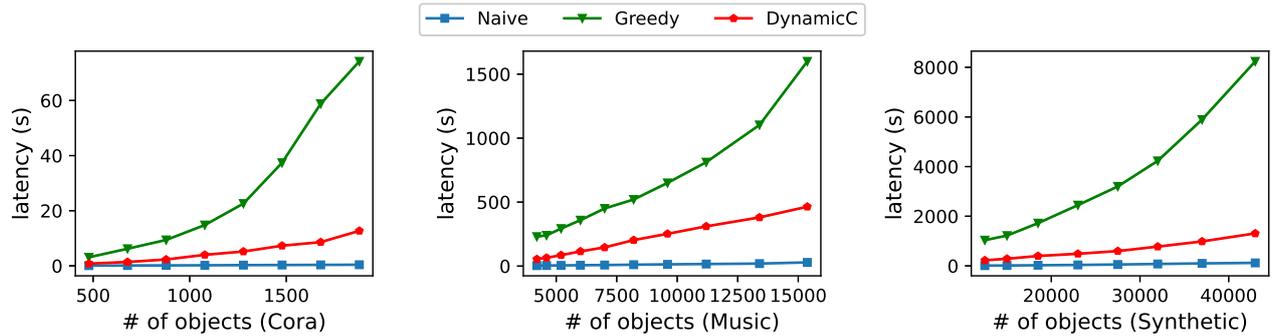


Figure 7: Re-clustering latency on Cora, Music and Synthetic datasets (the curve of Hill-climbing is omitted since it takes more than 4 hours on average for each dataset)

approximate DBSCAN and use another metric to evaluate their method instead of F1). However, DynamicC saves around 40% to 60% time while reaching F1 scores that are close to the optimal.

7.2.2 Results for k -means. The F1 scores of most snapshots are 1 or very close to 1 on Access and Road datasets. This can also be seen from Figure 5(d) which shows the square root of objective score on Road dataset (the results on Access dataset are similar to that of Road dataset). Naive method gets worse objective scores when more updates arrive. But the other methods get almost the same objective score with that of the batch algorithm. The reason is that when the number of clusters is fixed (i.e. the value of k), it is easy for the other methods to converge to a stable clustering which is often a near optimal clustering. However, DynamicC significantly reduces the latency as shown in Figure 5(e).

7.2.3 Results for DB-index. The quality evaluation of five approaches is shown in Figure 6 for DB-index clustering. The experimental results on three datasets are similar. The quality of Naive method decreases when the number of objects increases to be the worst out of the other methods. The reason for this is that the “merge-only” strategy applied in Naive can not work well when the clustering structure changes.

Hill-climbing achieves the best objective score among the five approaches as it searches for a better clustering in a brute-force way. Greedy achieves better score than Naive, but gets worse score than DynamicC. DynamicC reaches better objective score in DynamicSet than the GreedySet, showing that the quality of the initial clustering can make DynamicC obtain better objective score. Table 2 shows the F1 measures of Naive, Greedy and DynamicC (we only show the results of 5 out of 8 (for Cora and Synthetic) or 10 snapshots (for Music) due to space limit). Naive gets worse and worse F1 measure when the number of updates increases. DynamicC gets a little higher F1 measure than that of Greedy can obtain similar result with that of the batch algorithm since its F1 measure is very close to 1. Table 3 shows the performance of three methods with precision, recall,

purity [50] and inverse purity [9]. Due to space limit, we just show the results based on the last round for each dataset. As can be seen from the table, DynamicC can get the best precision and recall. And because purity and inverse purity focus on the cluster with maximum precision and recall respectively, DynamicC can also get the best performance in terms of purity and inverse purity. We do observe that DynamicC occasionally gets lower precision but higher recall than Greedy during the whole dynamic process. Nevertheless, DynamicC gets best F1 scores in most cases which can be seen from Table 2.

For latency, the results are shown in Figure 7 (the points of each line corresponds to the snapshots during the dynamic process). DynamicC has similar execution time in both GreedySet and DynamicSet scenarios. So, we take their average execution time as the time cost of DynamicC. DynamicC significantly reduces the time cost especially when the scale of the dataset is large. Also, Figure 7 demonstrates that the latency of Greedy increases significantly compared to Naive and DynamicC when the dataset becomes larger. We also observe that DynamicC saves significantly more overhead than Greedy on the Synthetic dataset when the dataset becomes larger. This indicates that DynamicC can better address complex data structure as the average number of similar neighbor objects in the Synthetic dataset is larger than that of Cora and Music Brainz datasets.

In summary, DynamicC dramatically decreases latency compared to Greedy and Hill-climbing while retaining or approaching the quality of clustering even when the dataset is large.

7.3 ML Model Evaluation

In this section, we evaluate the effectiveness of ML models in DynamicC and their retraining time.

Table 4 shows the accuracy and recall of Logistic Regression, SVM and Decision Tree models. When the number of training samples increases, the accuracy and recall of the three models all reach satisfactory values. The training time of these models is

Table 2: F1 measure for DB-index clustering

| Dataset/Snapshot | | 1 | 2 | 3 | 4 | 5 |
|------------------|----------|--------------|--------------|--------------|--------------|--------------|
| Cora | Naive | 0.943 | 0.912 | 0.908 | 0.871 | 0.843 |
| | Greedy | 0.998 | 0.985 | 0.984 | 0.981 | 0.981 |
| | DynamicC | 1 | 0.988 | 0.991 | 0.983 | 0.984 |
| Music | Naive | 0.982 | 0.976 | 0.963 | 0.945 | 0.932 |
| | Greedy | 1 | 0.991 | 0.987 | 0.986 | 0.989 |
| | DynamicC | 1 | 0.996 | 0.994 | 0.991 | 0.993 |
| Synthetic | Naive | 0.931 | 0.871 | 0.864 | 0.831 | 0.815 |
| | Greedy | 0.995 | 0.985 | 0.991 | 0.984 | 0.979 |
| | DynamicC | 0.998 | 0.997 | 0.989 | 0.994 | 0.992 |

Table 3: Other metrics for DB-index clustering

| Dataset/Metrics | | precision | recall | purity | inverse purity |
|-----------------|----------|--------------|--------------|--------------|----------------|
| Cora | Naive | 0.884 | 0.806 | 0.914 | 0.842 |
| | Greedy | 0.992 | 0.970 | 0.994 | 0.984 |
| | DynamicC | 0.996 | 0.972 | 0.997 | 0.988 |
| Music | Naive | 0.913 | 0.952 | 0.943 | 0.976 |
| | Greedy | 1 | 0.978 | 1 | 0.992 |
| | DynamicC | 1 | 0.986 | 1 | 0.994 |
| Synthetic | Naive | 0.835 | 0.796 | 0.879 | 0.861 |
| | Greedy | 0.987 | 0.971 | 0.976 | 0.986 |
| | DynamicC | 0.990 | 0.994 | 0.999 | 0.992 |

Table 4: Evaluation for different ML models on Cora dataset (the training time is less than 1s for all the three models when the number of samples is 20K)

| # of new objects/ (# of training samples) | | 200 (97) | 400 (248) | 600 (420) | 800 (745) | 1000 (1077) |
|--|----------|-------------|--------------|--------------|--------------|----------------|
| LogisticRegression | Accuracy | 0.77 | 0.82 | 0.88 | 0.90 | 0.93 |
| | Recall | 0.25 | 0.98 | 1.0 | 1.0 | 1.0 |
| SVM | Accuracy | 0.77 | 0.81 | 0.87 | 0.89 | 0.92 |
| | Recall | 0.25 | 0.95 | 0.96 | 1.0 | 1.0 |
| Decision Tree | Accuracy | 0.86 | 0.76 | 0.86 | 0.93 | 0.95 |
| | Recall | 0.75 | 0.80 | 0.97 | 0.96 | 1 |

Table 5: The performance of Logistic Regression with default settings on three datasets

| fraction of training samples | | 5% | 10% | 20% | 40% | 80% |
|------------------------------|----------|------|------|------|------|------|
| Cora | Accuracy | 0.62 | 0.74 | 0.83 | 0.90 | 0.98 |
| | Recall | 0.15 | 0.18 | 0.98 | 1.0 | 1.0 |
| Music | Accuracy | 0.84 | 0.87 | 0.94 | 0.96 | 0.97 |
| | Recall | 0.56 | 0.93 | 1.0 | 1.0 | 1.0 |
| Synthetic | Accuracy | 0.73 | 0.85 | 0.88 | 0.89 | 0.93 |
| | Recall | 0.47 | 0.81 | 0.92 | 0.95 | 0.98 |

negligible as it is less than 1 second for all the three models when the number of samples is less than 20K. The ML models do affect the latency of DynamicC. The larger precision and recall of the models, the less latency of DynamicC. Nevertheless, DynamicC gets almost the same F1 score with different ML models finally. This is because DynamicC chooses a proper θ value (described in Section 5.4) to make sure that the model will get high recall for new objects.

We evaluate the performance of Logistic Regression model in terms of the fraction of training data with respect to the three datasets. Table 5 shows the results with default parameter settings. For each round of evaluation, 20% of the whole objects are added for model evaluation. Initially, the model has low accuracy and recall since it has little information about the clustering result. But the model becomes experienced with whether the clustering is good or not when more samples are provided. The accuracy of the models will reach stability when sufficient samples are fed. This indicates that these models are effective when more samples are provided.

8 CONCLUSION

In this paper, we tackle the problem of efficient clustering. We observe that existing solutions incur a significant overhead to achieve good quality. To overcome this overhead, we propose DynamicC, a machine learning-based solution that can be augmented with existing batching solutions for clustering. DynamicC advances the state of incremental clustering by proposing a more general dynamic system model where objects are added, deleted, and modified. Furthermore, DynamicC does not make assumptions about the underlying batching mechanisms which allows it to be augmented with a wider range of batching solutions unlike existing incremental methods. Our evaluation shows that DynamicC can achieve a significant performance improvement compared to other methods while retaining their quality.

Our future directions based on this work include applying the proposed efficient clustering methods to real problems. In particular, DynamicC is suitable for environments where there is a need for fast response from a machine learning model. This includes problems that utilize machine learning in the path of execution of storage and transactional solutions [22–24, 27, 30, 31]. This also includes problems where there is an asymmetry between resources, such as edge and cloud environments [35, 36]. Finally, DynamicC allows deriving smaller models from bigger ones which can be useful in blockchain decentralized applications where there is a need to reduce the storage and processing footprint of applications due to cost and performance overhead [6, 8, 11, 37]

9 ACKNOWLEDGMENTS

This research is supported in part by the NSF under grant CNS-1815212 and a gift from Facebook.

REFERENCES

- [1] [n.d.]. <https://www.cs.utexas.edu/users/ml/riddle/data.html>.
- [2] [n.d.]. https://dbs.uni-leipzig.de/en/research/projects/object_matching/benchmark_datasets_for_entity_resolution.
- [3] [n.d.]. <https://archive.ics.uci.edu/ml/datasets/Amazon+Access+Samples>.
- [4] [n.d.]. [https://archive.ics.uci.edu/ml/datasets/3D+Road+Network+\(North+Jutland,+Denmark\)](https://archive.ics.uci.edu/ml/datasets/3D+Road+Network+(North+Jutland,+Denmark)).
- [5] [n.d.]. <https://sourceforge.net/projects/febrll/>.
- [6] Daniel Abadi, Owen Arden, Faisal Nawab, and Moshe Shadmon. 2020. Anylog: a grand unification of the internet of things. In *Conference on Innovative Data Systems Research (CIDR '20)*.
- [7] Elke Ahtert, Sascha Goldhofer, Hans-Peter Kriegel, Erich Schubert, and Arthur Zimek. 2012. Evaluation of clusterings—metrics and visual support. In *2012 IEEE 28th International Conference on Data Engineering*. IEEE, 1285–1288.
- [8] Maha Alaslani, Faisal Nawab, and Basem Shihada. 2019. Blockchain in IoT systems: End-to-end delay evaluation. *IEEE Internet of Things Journal* 6, 5 (2019), 8332–8344.
- [9] Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information retrieval* 12, 4 (2009), 461–486.
- [10] Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. 1999. OPTICS: Ordering points to identify the clustering structure. *ACM Sigmod record* 28, 2 (1999), 49–60.
- [11] Maha Aslani, Osama Amin, Faisal Nawab, and Basem Shihada. 2020. Rethinking blockchain integration with the industrial Internet of Things. *IEEE Internet of Things Magazine* 3, 4 (2020), 70–75.
- [12] Amel Awadelkarim and Johan Ugander. 2020. Prioritized Restreaming Algorithms for Balanced Graph Partitioning. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1877–1887.
- [13] Peter Bailis, Edward Gan, Samuel Madden, Deepak Narayanan, Kexin Rong, and Sahaana Suri. 2017. Macrobases: Prioritizing attention in fast data. In *Proceedings of the 2017 ACM International Conference on Management of Data*. 541–556.
- [14] Nikhil Bansal, Avrim Blum, and Shuchi Chawla. 2004. Correlation clustering. *Machine learning* 56, 1-3 (2004), 89–113.
- [15] Nathalie A Barbosa, Louise Travé-Massuyès, and Victor H Grisales. 2016. A novel algorithm for dynamic clustering: properties and performance. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 565–570.

- [16] Pavel Berkhin. 2006. A survey of clustering data mining techniques. In *Grouping multidimensional data*. Springer, 25–71.
- [17] Carlo Curino, Evan Philip Charles Jones, Yang Zhang, and Samuel R Mad-den. 2010. Schism: a workload-driven approach to database replication and partitioning. (2010).
- [18] David L Davies and Donald W Bouldin. 1979. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence* 2 (1979), 224–227.
- [19] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Michael Wimmer, and Xiaowei Xu. 1998. Incremental Clustering for Mining in a Data Warehousing Environment. In *VLDB'98, Proceedings of 24rd International Conference on Very Large Data Bases, August 24-27, 1998, New York City, New York, USA*, Ashish Gupta, Oded Shmueli, and Jennifer Widom (Eds.). Morgan Kaufmann, 323–333. <http://www.vldb.org/conf/1998/p323.pdf>
- [20] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise.. In *Kdd*, Vol. 96. 226–231.
- [21] Junhao Gan and Yufei Tao. 2017. Dynamic density based clustering. In *Proceedings of the 2017 ACM International Conference on Management of Data*. 1493–1507.
- [22] Samaa Gazzaz, Vishal Chakraborty, and Faisal Nawab. 2021. Croesus: Multi-Stage Processing and Transactions for Video-Analytics in Edge-Cloud Systems. *arXiv preprint arXiv:2201.00063* (2021).
- [23] Samaa Gazzaz, Vishal Chakraborty, and Faisal Nawab. 2022. Croesus: Multi-Stage Processing and Transactions for Video-Analytics in Edge-Cloud Systems. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*. IEEE.
- [24] Samaa Gazzaz and Faisal Nawab. 2019. Collaborative edge-cloud and edge-edge video analytics. In *Proceedings of the ACM Symposium on Cloud Computing*. 484–484.
- [25] Songjie Gong. 2010. A collaborative filtering recommendation algorithm based on user clustering and item clustering. *J. Softw.* 5, 7 (2010), 745–752.
- [26] Anja Gruenheid, Xin Luna Dong, and Divesh Srivastava. 2014. Incremental record linkage. *Proceedings of the VLDB Endowment* 7, 9 (2014), 697–708.
- [27] Binbin Gu, Zhixu Li, An Liu, Jiajie Xu, Lei Zhao, and Xiaofang Zhou. 2019. Improving the Quality of Web-Based Data Imputation With Crowd Intervention. *IEEE Transactions on Knowledge and Data Engineering* 33, 6 (2019), 2534–2547.
- [28] Songtao Guo, Xin Luna Dong, Divesh Srivastava, and Remi Zajac. 2010. Record linkage with uniqueness constraints and erroneous values. *Proceedings of the VLDB Endowment* 3, 1-2 (2010), 417–428.
- [29] Stephen C Johnson. 1967. Hierarchical clustering schemes. *Psychometrika* 32, 3 (1967), 241–254.
- [30] Saeed Kargar, Heiner Litz, and Faisal Nawab. 2021. Predict and Write: Using K-Means Clustering to Extend the Lifetime of NVM Storage. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. IEEE, 768–779.
- [31] Saeed Kargar and Faisal Nawab. 2021. Extending the lifetime of NVM: challenges and opportunities. *Proceedings of the VLDB Endowment* 14, 12 (2021), 3194–3197.
- [32] Leonard Kaufman and Peter J Rousseeuw. 1990. Partitioning around medoids (program pam). *Finding groups in data: an introduction to cluster analysis* 344 (1990), 68–125.
- [33] Stuart Lloyd. 1982. Least squares quantization in PCM. *IEEE transactions on information theory* 28, 2 (1982), 129–137.
- [34] James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Vol. 1. Oakland, CA, USA, 281–297.
- [35] Natasha Mittal and Faisal Nawab. 2021. CoolSM: Distributed and Cooperative Indexing Across Edge and Cloud Machines. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. IEEE, 420–431.
- [36] Faisal Nawab. 2021. Wedgechain: A trusted edge-cloud store with asynchronous (lazy) trust. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. IEEE, 408–419.
- [37] Faisal Nawab and Mohammad Sadoghi. 2019. Blockplane: A global-scale byzantizing middleware. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE, 124–135.
- [38] Markus Nentwig, Anika Groß, and Erhard Rahm. 2016. Holistic entity clustering for linked data. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*. IEEE, 194–201.
- [39] Markus Nentwig and Erhard Rahm. 2018. Incremental clustering on linked data. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 531–538.
- [40] Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn, and Supachanun Wanapu. 2013. Using of Jaccard coefficient for keywords similarity. In *Proceedings of the international multicongference of engineers and computer scientists*, Vol. 1. 380–384.
- [41] Fabian Pedregosa and journal=the Journal of machine Learning research volume=12 pages=2825–2830 year=2011 publisher=JMLR. org Varoquaux, Gaël and Gramfort, Alexandre and Michel, Vincent and Thirion, Bertrand and Grisel, Olivier and Blondel, Mathieu and Prettenhofer, Peter and Weiss, Ron and Dubourg, Vincent and others. [n.d.]. Scikit-learn: Machine learning in Python. [n.d.].
- [42] Dan Pelleg and Andrew Moore. 1999. Accelerating exact k-means algorithms with geometric reasoning. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. 277–281.
- [43] Pradeep Rai and Shubha Singh. 2010. A survey of clustering techniques. *International Journal of Computer Applications* 7, 12 (2010), 1–5.
- [44] Alex Rodriguez and Alessandro Laio. 2014. Clustering by fast search and find of density peaks. *science* 344, 6191 (2014), 1492–1496.
- [45] Stuart Russell and Peter Norvig. 2002. Artificial intelligence: a modern approach. (2002).
- [46] Victor S Sartório and Thaís CO Fonseca. 2020. Dynamic clustering of time series data. *arXiv preprint arXiv:2002.01890* (2020).
- [47] Wei Wang, Jiong Yang, Richard Muntz, et al. 1997. STING: A statistical information grid approach to spatial data mining. In *VLDB*, Vol. 97. 186–195.
- [48] Steven Euijong Whang and Hector Garcia-Molina. 2014. Incremental entity resolution on rules and data. *The VLDB Journal—The International Journal on Very Large Data Bases* 23, 1 (2014), 77–102.
- [49] Li Yujian and Liu Bo. 2007. A normalized Levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence* 29, 6 (2007), 1091–1095.
- [50] Ying Zhao and George Karypis. 2001. Criterion functions for document clustering: Experiments and analysis. (2001).