Clustering, Universalities, and Evolutionary Schema Design

Issei Fujishiro[1], Naoko Sawada[1] and Makoto Uemura[2]

[1]Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanawaga 223-8522, Japan
[2]Hiroshima University, 1-3-2 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8511, Japan

Abstract

Exploring data features using visual clustering is a significant challenge of big data analytics. In this vision paper, we focus primarily on the relationship among visual data clustering, the discovery of universalities, and the design of an evolutionary database to propose an inter-disciplinary method for scientific data management. The feasibility of the proposed method is empirically proven through application to a practical visual analytics environment for time-varying multi-dimensional datasets of blazar observations.

Keywords

visual data clustering, universality, evolutionary database, schema design

1. Introduction

Feature exploration is a significant challenge of big data analytics. In response, visual data clustering [1] has become a useful approach for such a task, because it enables the identification of salient features coupled with appropriate user intervention. Careful visual data clustering can lead to the discovery of universalities hidden in target datasets. In this vision paper, we strive to demonstrate how evolutionary database design [2] can fully support this kind of valuable scientific activity.

2. Evolutionary Schema Design

This section proposes our evolutionary schema design in relation to the visual discovery of universalities. We use Universal Modeling Language (UML) [3] class diagrams for conceptual design, followed by translations into corresponding relational schemas.

2.1. Sample Class

A data matrix (multi-dimensional data samples) can be formulated as the class `Samples`, consisting of n_s attributes, as shown in Fig. 1. Samples have observational relationships with each other, and these can be abstracted by a recursive *association*, called `Samples_Transit`, also shown in Fig. 1.

The corresponding relational schema consists of the following two third normal form (3NF) relation schemas:

```
Samples(sample-ID, sa-1, sa-2, ..., sa-n_s)
Samples_Transit(sample-ID_s, sample-ID_d, t-info).
```

Actual instances of `Samples` and mutual relationships between the instances clearly form a weighted directed graph and are usually visualized with a node-and-link diagram. In the case of many `Samples` and dense mutual relationships, such a diagram often suffers from visual clutter artifacts.

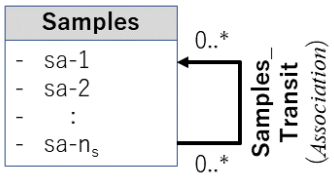


Figure 1: `Samples` class and its recursive association `Samples_Transit`

2.2. Cluster Class

Next, let us shift our attention to clustering, which allows each of the samples to belong to a cluster. These relationships can be modeled using a database abstraction called *aggregation*, as shown in Fig. 2, where each cluster is described by n_c attributes, many of which can be derived from `Samples` attributes via the aggregation `Belong_to`.

Proceedings of the 6th International Workshop on Big Data Visual Exploration and Analytics co-located with EDBT/ICDT 2023 Joint Conference (March 28-March 31, 2023), Ioannina, Greece
fuji@ics.keio.ac.jp (I. Fujishiro); naoko.sawada@fj.ics.keio.ac.jp (N. Sawada); uemuram@hiroshima-u.ac.jp (M. Uemura)
https://fj.ics.keio.ac.jp/en/member/issei-fujishiro (I. Fujishiro); https://fj.ics.keio.ac.jp/en/member/naoko-sawada (N. Sawada); https://home.hiroshima-u.ac.jp/~uemuram/ (M. Uemura)
0000-0002-8898-730X (I. Fujishiro); 0000-0002-9281-441X (N. Sawada); 0000-0002-7375-7405 (M. Uemura)

In normal visualization, visual clutter artifacts cannot be resolved. It is because each cluster may be accentuated by an ellipse, while the original inter-instance links usually remain unchanged.

Here, we consider making explicit the *universalities* found in the Samples instances. Specifically, if associations between Samples instances can commonly be observed in the same pair of Clusters, we propose to upgrade the mutual relationships between Samples to mutual associations between Clusters, also shown in Fig. 2.

At this point, provided that an *evolutionary* data management environment is available, the corresponding relational schema can be re-formulated using the following three 3NF relation schemas:

Samples(sample-ID, cluster-ID, sa-1, sa-2, ..., sa-n_s)
Clusters(cluster-ID, ca-1, ca-2, ..., ca-n_c)
Clusters_Transit(cluster-ID_s, cluster-ID_d, meta_t-info).

Note that the aggregation Belong_to is realized via the foreign key cluster-ID in the new definition of the relation schema Samples. Note also that the relation schema Clusters_Transit has meta_t-info, which can be derived from the t-info values of the belonging Samples. It would be interesting to describe the occurrence probability as an attribute of meta_t-info. As a by-product of such a universality specification, the number of inter-cluster associations can drastically be reduced, resulting in a simplified visualization.

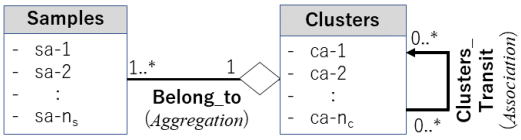


Figure 2: Clusters class and its recursive association Clusters_Transit

2.3. Subsample Class

For each instance of Clusters, *idiosyncratic* attributes may have to be specified. To manage such attributes efficiently, we propose to define a new class, Subsamples, as a *specialization* of Samples, as shown in Fig. 3.

The corresponding relational schema consists of the following $(n_c + 1)$ 3NF relation schemas:

Samples(sample-ID, sa-1, sa-2, ..., sa-n_s)
Subsamples$_k$(sample-ID, ssa-1, ssa-2, ..., ssa-n_{ss_k})
$(k = 1, \ldots, n_c)$.

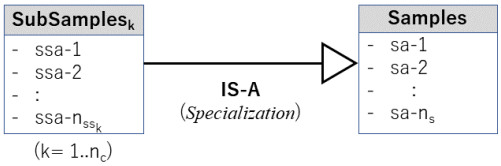


Figure 3: Subsamples class

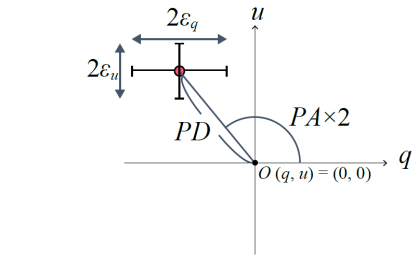
Note that the specialization IS_A is naturally realized by the common primary key sample-ID in the relation schemas. The idiosyncratic attributes of Subsamples$_k$ may be used to derive new attributes of Clusters. From the viewpoint of big data visual analytics, a remarkable advantage of idiosyncratic attribute separation lies in its ability to avoid the explosion of *inapplicable* null values in single relation Samples.

3. Case Study

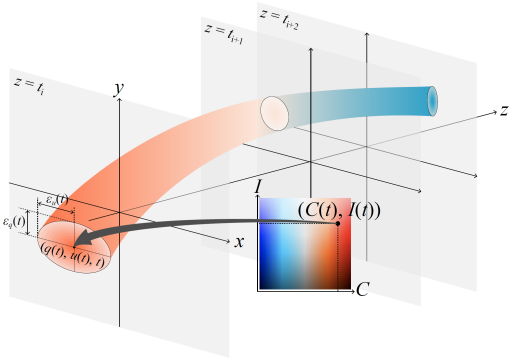
Blazars are the brightest and most energetic objects in the universe. To demystify the physics of the magnetic field within a relativistic jet ejected from a central black hole of a blazar, the light from a blazar is regularly observed. The Hiroshima Astrophysical Science Center (HASC) has scrutinized optical photo-polarimetric and near-infrared observation datasets to identify characteristic blazar behaviors, such as light bursts (i.e., *flares*) and rotated polarization (i.e., *rotation*), to explore recurring time-variation patterns. TimeTubesX [4, 5] is an integrated visual analytics environment that allows blazar researchers to analyze efficiently and in detail long-term, multi-dimensional blazar observation datasets. This section strives to apply the evolutionary schema design in Sec. 2 to sophisticated data management in the TimeTubeX system.

3.1. Data

The HASC has observed the polarization, intensity, and color (C) of the light from a blazar, where the linear polarization is described by three Stokes parameters, Q, U, and I, with I denoting the total intensity of the polarized and unpolarized components, Q the intensity of the linear horizontal or vertical polarization components, and U the intensity of the linear $+1/4\pi$ or $-1/4\pi$ polarization components, respectively. Instead of Q and U, we mainly utilize q and u, which can be obtained by dividing Q and U by I, because q and u explain blazar behaviors better than Q and U. The observation errors of q and u are described as ϵ_q and ϵ_u, respectively. The space spanned by q and u is termed the *Stokes plane* (Fig. 4a). When analyzing time variations in the Stokes



(a) Stokes plane



(b) TimeTube

Figure 4: Visual encoding for blazar dataset

plane, blazar researchers pay careful attention to time variations in the radial distance and polar angle on the Stokes plane. The radial distance on the Stokes plane is termed the *polarization degree* (*PD*), while one half of the polar angle is the *polarization angle* (*PA*).

As illustrated in Fig. 4b, a blazar dataset can be encoded geometrically as a single 3D volumetric tube, called a *TimeTube*, which helps blazar researchers recognize intuitively the time variations of and correlations among the variables described above. The quantities q and u are assigned to the x and y axes of the 3D visualization domain, respectively, and time t is assigned to the z axis. Thus, the four parameters related to polarization $(q, u, \epsilon_q, \epsilon_u)$ at each timestamp t are naturally encoded as an ellipse located at the point $(x, y, z) = (q(t), u(t), t)$ with a width of $2\epsilon q(t)$ and a height of $2\epsilon u(t)$. See [6] for more details.

3.2. Visual Clustering

To enable blazar researchers to examine universalities in blazar datasets, TimeTubesX provides them with time-varying multi-dimensional subsequence clustering methods [5], together with a designated set of visual analysis methods, including the advanced sample retrieval functionalities *query-by-example* and *query-by-sketch* [4]. The clustering methods extract subsequences of various lengths from a long-term observation dataset, considering missing data and observation frequencies, and then they filter subsequences with overlapping features. The clustering methods consider correlations among variables and compute means of subsequences without smoothing out their features.

The timeline view of TimeTubesX in Fig. 5 summarizes the temporal distributions of six found clusters of different stripe colors.

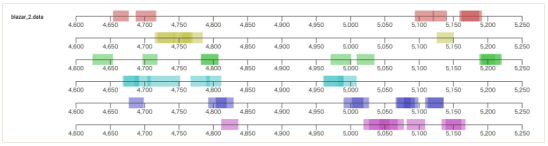


Figure 5: Timeline view of TimeTubesX

3.3. Inter-flare Cluster Transitions

Figure 6 shows a class diagram of TimeTubesX, where another type of database abstraction, called *composition*, is used to specify `Is_composed_of` for composing a new class `Subsequences` as a time series of `Samples`, whose `time`-dependent attributes are the six variables described in Sec. 3.1. Actual flare analysis detects the timing and size of the time interval of each flare. `Subsequences` includes as its attributes the `length`, `cor` (the center of rotation), and `angle` (the total amount of rotation) of the subsequences, all of which can be derived through the composition. Note that `flareID` comes directly from the flare analysis. The samples within the flare time interval often require more detailed analysis; thus, more attributes should be immediately available, such as PD, PA, q, and u. These idiosyncratic attributes are described separately by a specialized class `FlareSamples`. Meanwhile, `Clusters` includes as its attributes the number of subsequences and the cluster prototype, and `Is_followed_by` is characterized by its transition probability.

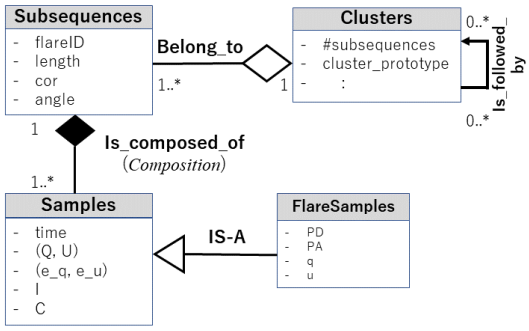


Figure 6: Class diagram of TimeTubesX

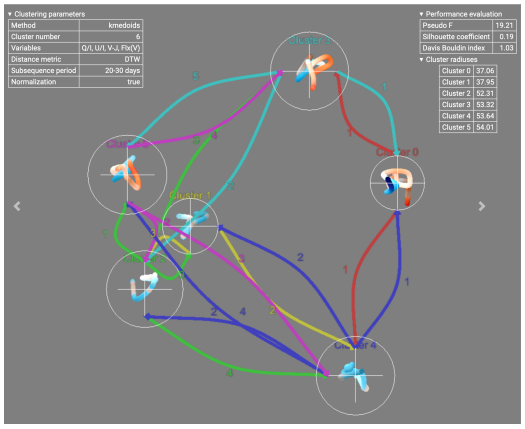


Figure 7: Time-series motifs and their transitions in TimeTubeX cluster feature view

The corresponding relational schema consists of the following five 3NF relation schemas, where the composition is naturally realized via the foreign key `ss-ID` in the relation schema `Samples`:

`Samples(sample-ID, ss-ID, time, Q, U, e_q, e_u, I, C)`
`FlareSamples(sample-ID, PD, PA, q, u)`
`Subsequences(ss-ID, cluster-ID, flareID, length, cor, angle)`
`Clusters(cluster-ID, #subsequences, cluster_prototype)`
`Is_followed_by(cluster-ID_s, cluster-ID_d, transit-prob).`

The cluster feature view of TimeTubesX in Fig. 7 is intended to summarize the universalities discovered in the dataset. It displays the features of cluster prototypes (time-series motifs) and their uninterrupted transitions. The comprehensibility of the node-link diagram stems mostly from the arrangement of the cluster prototypes in terms of their similarities in *multi-dimensional scaling* [7], the cluster prototypes encoded by TimeTubes at nodes, and a limited number of intercluster transition links. The node color is automatically selected to maximize the color difference between labels in a L*a*b* color space, while the link color encodes from the cluster at which the transition starts. The radius of a white ring around a TimeTube indicates the cluster radius (i.e., within-cluster separation), which signifies the maximum distance between a cluster prototype and all subsequences in the cluster. The number beside the link denotes how many transitions are observed between the two clusters, where the thresholding could contribute to further distillation of the universality.

4. Concluding Notes

In this paper, we demonstrated the possibility of bridging three worlds, i.e., visual analytics, universality discovery, and database refactoring. Through the application of the present methodology to the practical problem of blazar observation, we empirically proved that universality identification based on visual data clustering is strongly supported by evolutionary schema design.

Acknowledgments

This work has been partially supported by the Grant-in-Aid for Challenging Research (Pioneering) JP20K20481.

References

[1] M. Sips, Visual clustering, in: Encyclopedia of Database Systems, Springer, Boston, MA, 2009, pp. 3350–3360. doi:10.1007/978-0-387-39940-9_1124.

[2] S. W. Ambler, P. J. Sadalage, Refactoring Databases: Evolutionary Database Design, Addison-Wesley, 2006.

[3] G. Booch, J. Rumbaugh, I. Jacobson, The Unified Modeling Language User Guide, 2nd ed., Addison-Wesley, 2005.

[4] N. Sawada, M. Uemura, J. Beyer, H. Pfister, I. Fujishiro, Timetubesx: A query-driven visual exploration of observable, photometric, and polarimetric behaviors of blazars, IEEE Transactions on Visualization and Computer Graphics 28 (2022) 1917–1929. doi:10.1109/TVCG.2020.3025090.

[5] N. Sawada, M. Uemura, I. Fujishiro, Multi-dimensional time-series subsequence clustering for visual feature analysis of blazar observation datasets, Astronomy and Computing 41 (2022) 100663:1–100633:12. doi:10.1016/j.ascom.2022.100663.

[6] I. Fujishiro, N. Sawada, M. Nakayama, H.-Y. Wu, K. Watanabe, S. Takahashi, M. Uemura, Timetubes: Visual exploration of observed blazar datasets, in: Proceedings of International Meeting on High-Dimensional Data-Driven Science, volume 1036 of *Journal of Physics: Conference Series*, IOP Publishing, 2018, pp. 012011:1–012011:13. doi:10.1088/1742-6596/1036/1/012011.

[7] W. S. Torgerson, Multidimensional scaling: I. theory and method, Psychometrica 17 (1952) 401–419. doi:10.1007/BF02288916.