

Impact of Time Granularity on Histories Binary Correlation Analysis

Paolo Mengoni[1(✉)], Alfredo Milani[1], and Yuanxi Li[2]

[1] Department of Mathematics and Computer Science, University of Perugia, Perugia, Italy
paolo.mengoni@dmi.unipg.it, alfredo.milani@unipg.it
[2] Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong
csyxli@comp.hkbu.edu.hk

Abstract. Activities taken by students within a Virtual Learning Environment (VLE) can be represented by using binary student histories. Virtual Learning Environments allow educators to track most of the students' individual activities that can be used to elicit the students social communities. In this work, we analyse the impact of granularity in the social community elicitation. Granularity can be seen as the resolution of the student history vectors where each time slot is directly dependent from this value. Indeed, the higher is the resolution of the students histories the more precise is the representation of their actions within the VLE. When comparing the histories using various similarity measures to elicit the students' groups, we find the optimal granularity and demonstrate that there is a resolution limit where the similarity measures will not help to distinguish the social communities.

Keywords: Data analysis · Learning analytics · Cluster analysis · Community elicitation

1 Introduction

By using the Virtual Learning Environments (VLE) educators gained the ability to track and analyse the online activities of students. In this scenario, courses can be delivered by teachers by using various mixtures of online and offline activities.

Learning Designs (LD) can be improved by knowing the students' communities. This can help also to identify situations potentially at risk, which will need the teacher intervention, e.g. when there can be risk of exam failure and/or more serious cases of dropping from studies. In online courses, such as MOOCs or SPOCs, all the student's activities can be analysed, including the student's communications. In the blended learning scenario the major part of communications among the students take place using side channels, such as social networking and instant messaging applications, that are out of the control scope of the teacher and don't leave traces to be analysed.

© Springer Nature Switzerland AG 2019
S. Misra et al. (Eds.): ICCSA 2019, LNCS 11620, pp. 323–335, 2019.
https://doi.org/10.1007/978-3-030-24296-1_27

Using the Histories Binary Correlation Analysis (HBCA) [35], we demonstrated in previous work that is possible to discover the latent social communities of students. This methodology makes use of information about the students' activities within a VLE to create the students' history vectors. The similarity between student's histories is the used to gain knowledge about the social interactions of students and their communities.

Similarity measures used to compare student's histories are highly dependant on how the features are extracted from the raw data. One of the parameters that impacts the creation of student's history vectors is the *granularity* of observations. For low granularity values (i.e. at high resolution) the histories are more accurate in representing the exact timing of the activities, while for high granularity values (i.e. at low resolution) more activities are grouped together as on single value of the history vector. The scope of this work is to analyse this parameter impact on the quality of social communities elicited by using the HBCA approach.

This paper is organised as follows. In Sect. 2 we review the related work in Community Detection in networks. In Sect. 3 we present the History Binary Correlation Analysis to find communities of students in VLE. In Sect. 4 we introduce the experimental setting and methodology. In Sect. 5 we discuss the experimental results. Finally, in Sect. 6, we draw conclusion and present future works.

2 Related Work

Nodes of a network that are similar considering specific characteristics can be grouped in communities, whose connections can be used to determine the community structure.

Different classes of problems related to communities have been studied in literature, mainly to find solutions for community search and community detection. The former class of algorithms tries to identify the community to which belongs a given node [46]. The latter tries to identify the all the communities included in a given network [30].

Starting from a given set of partitions, the strategies to elicit the various communities in a network make use of various approaches, and also their combinations, that merge, split and change the assignment of elements from the starting communities. The survey in "Community detection in graphs" [14] includes an extensive review of the methodologies used to find solutions for the community detection class of problems, that exploit various characteristics of the network and its elements.

Network partitioning and clustering are examples of traditional methods used for community detection. Network partitioning is in general a NP-Hard problem that can be solved with some approximate algorithms. It requires that the number of partitions as a parameter and uses features of the network edges to find the communities [28]. Clustering algorithms [1,11,20,21] make use of distance measures [8,19,22,23,34] to group together different nodes of the network. Different community discovery strategies have been developed, such as spectral [13], partitional [32] and hierarchical [26].

To assess the quality of the elicited community and as heuristic to find approximate solutions to the community detection problem in literature is widely used the Q modularity value, in origin used in Girvan-Newman divisive algorithm [24] as stopping criterion. The modularity maximization optimisation problem has been therefore solved by using greedy [5], simulated annealing [25], evolutionary computation [2,44], probabilistic graphical models [3] and other heuristic methods [6].

The interactions of the students with the VLE have been studied in our previous works [35,36] to elicit the students' communities. Monitoring tools, such as SNAPP [4,10] and Dynamical Morphing Interfaces [16], have been developed to understand the social interactions between the students. SNAPP is a tool that uses discussion forum threads to elicit the social network of the students, while Dynamic Morphing Interfaces are used by the instructors to asses the quality and study behaviour of the students. In other works, such as Khosravi et al. [29], the connections between the engagement of the student with the VLE and their learning outcomes have been investigated. Clustering methodologies, that exploit the learning outcomes and study behaviour similarities, have been used to elicit the groups of students.

3 Methodology

In previous works [35,36] we developed a framework to discover the communities of students, basing on the information of the actions they take in Virtual Learning Environments (VLE). The History Binary Correlation Analysis (HBCA) [35] uses the eLearning system logs' information that is related to student's actions. The timing of activities within the VLE is used to model the students' histories and elicit their communities [9]. Clustering is then applied to the students histories to find similar user behaviour and group the students in distinct communities [17]. HBCA approach elicits various groups of students by maximising the similarities of the ones grouped together. The first step of the approach is to retrieve the students' histories, then it proceeds by evaluating pairwise student histories' similarity and use this information to assign the cluster labelling.

3.1 Student Histories

Student histories are summarised by using granularity parameter that refers to the continuity of observation. Different types of granularity can be referred as *continuous* (i.e. CCTV cameras), *Periodic* (i.e. cellular line polling), and *event* (i.e. system logs). Granularity can be transformed from event to periodic using *bucketing* technique. Activities are grouped together in single buckets when they are within a certain γ time threshold. In the histories a binary value is recorded to represent the presence of interaction in that bucket. The number of buckets of each history-summarised vector is proportional to the length of the period and the granularity γ parameter.

3.2 Histories Pairwise Similarity

A student's similarity matrix is build basing on the comparison of the history binary summarised vectors using various binary similarity measures.

Similarity measures. To determine the similarity of the history summarised vectors a wide spectrum of binary similarity measures have been tested to emphasise the different aspects of the co-occurrence of interactions [18,31]. The comparison is made bucket by bucket to set the values of the parameters represented in the contingency table shown in Table 1.

Table 1. Contingency table for pairwise vector comparison.

	Vector 1 Element Feature F present	Vector 2 Element Feature F absent
Vector 1 Element Feature F present	a	b
Vector 2 Element Feature F absent	c	d

Among all the various binary similarity measures, for our experiments we considered the following, representative of different presence and absence of parameters in the computation:

Cooccurence. This simple statistic measure takes in account only of the presence of features in both sets.

$$Cooccurrence = a \tag{1}$$

Jaccard Similarity Measure. Introduced by Paul Jaccard at the start of 20th century [27], this measure is also known as the similarity ratio. It does not use the features' absence parameter in the computation.

$$Jaccard = \frac{a}{a+b+c} \tag{2}$$

Russell and Rao (RR) Similarity Measure. This similarity measure, introduced in 1940 by Russell and Rao [43], considers at the denominator the information about the absence of a features in the two sets.

$$RR = \frac{a}{a+b+c+d} \tag{3}$$

Simple Matching (SM) Similarity Measure. Known also as the Rand Index [40], this measure is the ratio of the number of matches to the total number of characteristics.

$$SM = \frac{a+d}{a+b+c+d} \tag{4}$$

Dice Similarity Measure. Dice similarity measure [12], also known as Czekanowski or Sorenson, is similar to the Jaccard index, but give more emphasis to the presence of common features between the compared sets.

$$Dice = \frac{2a}{2a + b + c} \tag{5}$$

Sokal and Sneath (SS1) Similarity Measure n. 1. In their work, Sokal and Sneath [45], presented various similarity measures. This measure enhances the concurrent presence and absence of features in the two sets.

$$SS1 = \frac{2\,(a + d)}{2\,(a + d) + b + c + d} \tag{6}$$

Ochiai Similarity Measure. Also known as Fowlkes-Mallows index [15], the Ochiai similarity measure [39] is the binary form of the cosine vector similarity.

$$Ochiai = \sqrt{\frac{a}{a + b} \cdot \frac{a}{a + c}} \tag{7}$$

3.3 Cluster Detection

The different history summarised vectors similarity matrices are used to group the students in different clusters. The approach is general and independent on the clustering algorithm. The different labels assigned to the students by the clustering algorithms are used to determine the students' groupings.

Agglomerative. The hierarchical clustering class of algorithms include bottom up approaches to merge the elements in different clusters. The agglomerative technique starts from clusters containing single elements and proceeds to cluster the most similar ones, e.g. the ones that minimise the sum of the square differences between the clusters distances.

Gaussian Mixture. The Gaussian Mixture, also called Expectation-Maximization (EM), is a statistical iterative algorithm for cluster detection. EM iterations switches between expectation (E) computation and maximisation (M) step. The latter step computes the parameters that maximise the expected likelihood. The algorithm stops when the E step does not increase the likelihood anymore.

K-Means. K-Means clustering is used to partition n elements in k clusters where the distance between the elements is minimal considering a specific distance. Although the problem is, in general, NP-hard some approximations exist that make use of heuristics to converge to a local optimum.

Spectral. Spectral clustering is used to reduce the dimensionality of the problem and then cluster in lower dimensions. The algorithm makes use of the spectum (i.e. the set of eigenvalues) of the similarity matrix to reduce the dimensionality. This matrix can be used in combination with all the previously mentioned clustering algorithms to detect the clusters of elements.

4 Experiments

4.1 Dataset

For this study we use a dataset containing anonimised students' activity logs from University of Perugia Moodle eLearning system. The students' activities have been anonymised by using peusonymisation techniques, extensively studied to manage containing medical data [37,38]. Pseudonymisation balances the need of privacy, by replacement of identifiers with pseudonyms, and the reversibility of the process under strict conditions.

Data from the course "Ingegneria del Software" (Software Engineering), taught with blended learning mode (i.e. with online and offline activities). Among all the various online and offline activities, the most important one for our study is the group project that the students were required to complete for successfully complete the course.

In the dataset are recorded all the students' actions and their timings taken with regards of the content available on the Virtual Learning Environment.

4.2 Ground Truth

As introduced in previous section, the students were required to develop a group project and submit to the instructor the related grouping information. This group composition information is then used to compare the performance of the different clustering algorithms as we can assume that students that work together to develop the project were studying together the course material.

4.3 Granularity

Students' histories are extracted from the dataset described in 4.1 and summarised at various granularity levels.

Similarly to the session time that each student spend while interacting with the VLE, granularity used for bucketing the time slots can not be clearly determined. In sessions, the starting action of each session can be easily identified. On the other hand, the end of session is difficult to be determined as the students tend to close the web browser instead of using logout function. In literature there is no agreement on how long it should be the threshold for session expiry. Early works suggest 30 min for web browsing sessions [7], while later works context-based and user-based thresholds have been suggested [33].

In this work, we will explore the impact of granularity parameter γ used to extract the student histories (previously described in Sect. 3.1). Different time slot lengths, starting from 1 h and its multiples and fractions, have been considered. Higher granularity values imply lower resolution when summarising the students actions, while lower granularity values will imply an higher resolution and detail of the students' histories.

All the combinations of measures and clustering algorithms, introduced in Sects. 3.2 and 3.3, have been tested to elicit the groups of students. In total, 28 combination of measure and clustering were experimented for each granularity setting.

4.4 Evaluation Criteria

The qualitative evaluation of the cluster label assignments have been conducted, with respect of the ground truth as defined in Sect. 4.2. Qualitative assessment of the cluster labelling have been conducted by using the following measures:

Homogeneity. High homogeneity is reached if the clusters contain only elements of a single class in the ground truth assignment [42].

Completeness. High completeness values are achieved if all the elements of a single class in the ground truth are assigned to the same cluster [42].

V-measure. Is a measures that balances the satisfaction of both homogeneity and completeness [41], by computing the harmonic mean of the previously cited scores [42].

5 Experiment Results Discussion

To identify the impact of granularity in community elicitation using HBCA approach, we select to use various preset levels representative of the continuous nature of time of actions. We decided to use some values that can be connected to some human activities. The granularity values, expressed in seconds, that we experimented are: 60, 300, 600, 900, 1800, 3600, 7200, 14400, 28800, and 86400. All the granularity levels have been experimented in combination with each measure and clustering algorithm introduced in previous sections, totalling to 280 combinations.

The best results for each clustering algorithm and similarity measure are summarised in Table 2. Agglomerative clustering is performing at best, in terms of V-measure score, when combined with the Jaccard similarity measure at 900 s of granularity. Gaussian Mixture best performance is obtained in combination with the Dice similarity measure at 600 s of granularity. For KMeans clustering the best measure is Ochiai at 900 s of granularity. Finally, Spectral clustering is performing at best when combined with the Dice similarity measure at 1800 s of granularity.

In Fig. 1 are reported detailed scores for each combination of clustering algorithm and similarity measure score, grouped by granularity value. From this figure we can see that with granularity set at 60 and 300 s some measures (specifically RR, SM, and SS1) cannot give enough information to the clustering algorithms to discriminate the communities of students. Moreover, we can see obtain the best results at 600 and 900 s of granularity. Starting from 1800 s the scores are generally dropping in terms of V-measure value and flattening to a score ≈ 0.5.

As shown in Fig. 2 the best results are, in general, obtained when setting the granularity between 600 and 900 s. At this resolutions the time slots in the history summarised vectors are bucketing the interactions between 10 and 15 min, representing the length of the online study session of students. For measures that take in account the absence of activities d (such as RR, SM, SS1), from this figure we can also see that the V-measure score increases with the increasing of granularity. Moreover, V-measure score of Gaussian Mixture combined with this set of measures, at high resolution (i.e. low granularity), is not able to distinguish the different cluster. The very low number of clusters leads to high completeness scores and extremely low homogeneity scores that result in low V-measure scores.

Table 2. V-measure best results for each clustering algorithm and similarity measure combination

	Measure						
	Cooccurrence	Dice	Jaccard	Ochiai	RR	SM	SS1
Agglomerative	0.493	0.654	**0.663**	0.655	0.493	0.496	0.508
Gaussian mixture	0.490	**0.664**	0.640	0.615	0.490	0.519	0.492
KMeans	0.509	0.572	0.582	**0.594**	0.509	0.497	0.523
Spectral	0.430	**0.635**	0.632	0.634	0.563	0.521	0.530

Dice similarity measure has, in general, the best performance followed by from Ochiai and Jaccard measures. For these measures, the best granularity setting can be found at 600 and 900 s, where the various clustering algorithms find communities that are close to the ground truth grouping of students.

It has been confirmed, as in previous works, that the common trait of the best performing measures is that the concurrent absence of interactions is not taken in account. The presence of interactions has very sparse values in the history summarised vectors when compared with the absence of interactions. From the results, we can find that the low resolution vectors (i.e. high granularity values), that are less sparse and with less number of slots when compared to the high resolution ones, help these similarity measures to lower the gap from the best ones given the same granularity setting.



Fig. 1. V-measure score evaluation: comparison of clustering and similarity measures for each granularity level.



Fig. 2. V-measure score evaluation: comparison of measures at different granularity levels for each clustering algorithm.

6 Conclusions and Future Works

In this work we analysed the impact of granularity when defining the students' history summarised vectors to be used in conjunction with History Binary Correlation Analysis. Different granularity settings lead to different behaviours of similarity measures. In general higher resolution (i.e. lower granularity) histories, until a certain level, help to improve the social community elicitation. Measures that take in account the simultaneous absence of activities have less impact on their performance when lowering the granularity value.

From the experiments, we found that the best granularity values are of 600 and 900 s. With these settings, the best performing similarity measures (specifically Dice, Ochiai, and Jaccard) obtain the best values in terms of V-measure score. This means that the social communities elicited from the approach are close to the ground truth grouping of students.

Future works will include the analysis of student groups' evolution during the course using time windows to evaluate the social communities evolution during time.

References

1. Akarsh, S., Kishor, A., Niyogi, R., Milani, A., Mengoni, P.: Social cooperation in autonomous agents to avoid the tragedy of the commons. Int. J. Agric. Environ. Inf. Syst. **8**(2), 1–19 (2017). https://doi.org/10.4018/IJAEIS.2017040101

2. Baioletti, M., Milani, A., Santucci, V.: A new precedence-based ant colony optimization for permutation problems. In: Shi, Y., et al. (eds.) SEAL 2017. LNCS, vol. 10593, pp. 960–971. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68759-9_79
3. Baioletti, M., Milani, A., Santucci, V.: Learning bayesian networks with algebraic differential evolution. In: Auger, A., Fonseca, C.M., Lourenço, N., Machado, P., Paquete, L., Whitley, D. (eds.) PPSN 2018. LNCS, vol. 11102, pp. 436–448. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99259-4_35
4. Bakharia, A., Dawson, S.: SNAPP. In: Proceedings of the 1st International Conference on Learning Analytics and Knowledge - LAK 2011, p. 168. ACM Press, New York, USA (2011). https://doi.org/10.1145/2090116.2090144
5. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech.: Theory Exp. **2008**(10), 10008 (2008). https://doi.org/10.1088/1742-5468/2008/10/P10008
6. Castellano, C., Cecconi, F., Loreto, V., Parisi, D., Radicchi, F.: Self-contained algorithms to detect communities in networks. Eur. Phys. J. B **38**, 311–319 (2004). https://doi.org/10.1140/epjb/e2004-00123-0
7. Catledge, L.D., Pitkow, J.E.: Characterizing browsing strategies in the World-Wide web. Comput. Netw. ISDN Syst. (1995). https://doi.org/10.1016/0169-7552(95)00043-7
8. Chan, S.W., Franzoni, V., Mengoni, P., Milani, A.: Context-based image semantic similarity for prosthetic knowledge. In: 2018 IEEE First International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), pp. 254–258. September 2018. https://doi.org/10.1109/AIKE.2018.00057
9. Chiancone, A., Franzoni, V., Niyogi, R., Milani, A.: Improving link ranking quality by quasi-common neighbourhood, pp. 21–26. IEEE Press (2015). https://doi.org/10.1109/ICCSA.2015.19
10. Dawson, S., Bakharia, A., Heathcote, E.: SNAPP : realising the affordances of real-time SNA within networked learning environments. In: Proceedings of the 7th International Conference on Networked Learning, pp. 125–133 (2010)
11. Deng, J.J., Leung, C.H.C., Mengoni, P., Li, Y.: Emotion recognition from human behaviors using attention model. In: 2018 IEEE First International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), pp. 249–253. September 2018. https://doi.org/10.1109/AIKE.2018.00056
12. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945). https://doi.org/10.2307/1932409
13. Donath, W.E., Hoffman, A.J.: Lower bounds for the partitioning of graphs. IBM J. Res. Dev. (1973). https://doi.org/10.1147/rd.175.0420
14. Fortunato, S.: Community detection in graphs (2010). https://doi.org/10.1016/j.physrep.2009.11.002
15. Fowlkes, E.B., Mallows, C.L.: A method for comparing two hierarchical clusterings. J. Am. Stat. Assoc. **78**(383), 553–569 (1983). https://doi.org/10.1080/01621459.1983.10478008
16. Franzoni, V., Mengoni, P., Milani, A.: Dimensional morphing interface for dynamic learning evaluation. In: 2018 22nd International Conference Information Visualisation (IV), pp. 332–337. July 2018. https://doi.org/10.1109/iV.2018.00063
17. Franzoni, V., Milani, A.: Pming distance: a collaborative semantic proximity measure. vol. 2, pp. 442–449. IEEE Press (2012). https://doi.org/10.1109/WI-IAT.2012.226

18. Franzoni, V., Milani, A., Pallottelli, S., Leung, C., Li, Y.: Context-based image semantic similarity, pp. 1280–1284. IEEE Press (2016). https://doi.org/10.1109/FSKD.2015.7382127
19. Franzoni, V., Leung, C.H.C., Li, Y., Mengoni, P., Milani, A.: Set similarity measures for images based on collective knowledge. In: Gervasi, O., et al. (eds.) ICCSA 2015. LNCS, vol. 9155, pp. 408–417. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-21404-7_30
20. Franzoni, V., Li, Y., Mengoni, P.: A path-based model for emotion abstraction on facebook using sentiment analysis and taxonomy knowledge. Proceedings of the International Conference on Web Intelligence - WI 2017, pp. 947–952 (2017). https://doi.org/10.1145/3106426.3109420
21. Franzoni, V., Li, Y., Mengoni, P., Milani, A.: Clustering facebook for biased context extraction. In: LNCS, vol. 10404, pp. 717–729 (2017). https://doi.org/10.1007/978-3-319-62392-4_52
22. Franzoni, V., Mencacci, M., Mengoni, P., Milani, A.: Heuristics for semantic path search in wikipedia. In: Murgante, B., et al. (eds.) ICCSA 2014. LNCS, vol. 8584, pp. 327–340. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-09153-2_25
23. Franzoni, V., Mencacci, M., Mengoni, P., Milani, A.: Semantic heuristic search in collaborative networks: measures and contexts. In: 2014 IEEE/WIC/ACM WI-IAT, vol. 1, pp. 141–148. IEEE, August 2014. https://doi.org/10.1109/WI-IAT.2014.27
24. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Nat. Acad. Sci. **99**(12), 7821–7826 (2002). https://doi.org/10.1073/pnas.122653799
25. Guimerà, R., Sales-Pardo, M., Amaral, L.A.: Modularity from fluctuations in random graphs and complex networks. Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics (2004). https://doi.org/10.1103/PhysRevE.70.025101
26. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning (2009). https://doi.org/10.1007/b94608
27. Jaccard, P.: Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin del la Société Vaudoise des Sciences Naturelles 37, January 1901, 547–579 (1901). https://doi.org/10.5169/seals-266450
28. Kernighan, B.W., Lin, S.: An efficient heuristic procedure for partitioning graphs. Bell Syst. Tech. J. (1970). https://doi.org/10.1002/j.1538-7305.1970.tb01770.x
29. Khosravi, H., Cooper, K.M.: Using learning analytics to investigate patterns of performance and engagement in large classes. In: Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education - SIGCSE 2017, pp. 309–314 (2017). https://doi.org/10.1145/3017680.3017711
30. Lancichinetti, A., Fortunato, S.: Community detection algorithms: a comparative analysis. Phys. Rev. E **80**, 1–12 (2009). https://doi.org/10.1103/PhysRevE.80.056117
31. Leung, C.H.C., Li, Y., Milani, A., Franzoni, V.: Collective evolutionary concept distance based query expansion for effective web document retrieval. In: Murgante, B., et al. (eds.) ICCSA 2013. LNCS, vol. 7974, pp. 657–672. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39649-6_47
32. Macqueen, J.: Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability (1967). citeulike-article-id:6083430

33. Mehrzadi, D., Feitelson, D.G.: On extracting session data from activity logs. In: Proceedings of the 5th Annual International Systems and Storage Conference, pp. 3:1–3:7. SYSTOR 2012, ACM, New York, NY, USA (2012). https://doi.org/10.1145/2367589.2367592
34. Mengoni, P., Milani, A., Li, Y.: Multi-term semantic context elicitation from collaborative networks. In: 2018 IEEE First International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), pp. 234–238. September 2018. https://doi.org/10.1109/AIKE.2018.00053
35. Mengoni, P., Milani, A., Li, Y.: Clustering students interactions in elearning systems for group elicitation. In: Gervasi, O., et al. (eds.) ICCSA 2018. LNCS, vol. 10962, pp. 398–413. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-95168-3_27
36. Mengoni, P., Milani, A., Li, Y.: Community graph elicitation from students interactions in virtual learning environments. In: Gervasi, O., et al. (eds.) ICCSA 2018. LNCS, vol. 10962, pp. 414–425. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-95168-3_28
37. Neubauer, T., Heurix, J.: A methodology for the pseudonymization of medical data. Int. J. Med. Inform. **80**(3), 190–204 (2011)
38. Neubauer, T., Riedl, B.: Improving patients privacy with pseudonymization. Stud. Health Technol. Inform. **136**, 691 (2008)
39. Ochiai, A.: Zoogeographical studies on the soleoid fishes found in Japan and its neighhouring regions-II. Nippon Suisan Gakkaishi **22**(9), 526–530 (1957). https://doi.org/10.2331/suisan.22.526
40. Rand, W.M.: Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. **66**(336), 846–850 (1971). https://doi.org/10.1080/01621459.1971.10482356
41. Rijsbergen, C.V.: Information retrieval. Information Retrieval, p. 208 (1979)
42. Rosenberg, A., Hirschberg, J.: V-measure: a conditional entropy-based external cluster evaluation measure. In: Proceedings of EMNLP-CoNLL 2007) 1, June 2007, 410–420 (2007). https://doi.org/10.7916/D80V8N84
43. Rusell, P.F., Rao, T.R.: On habitat and association of species of anopheline larvae in south-eastern madras. J. Malaria Inst. India **3**(1), 153–178 (1940)
44. Santucci, V., Milani, A.: Particle swarm optimization in the EDAs framework. In: Gaspar-Cunha, A., Takahashi, R., Schaefer, G., Costa, L. (eds.) Soft Computing in Industrial Applications, pp. 87–96. Springer, Berlin (2011)
45. Sneath, P.H.A., Sokal, R.R.: The principles and practice of numerical taxonomy. Taxon **12**(5), 190 (1963). https://doi.org/10.2307/1217562
46. Sozio, M., Gionis, A.: The community-search problem and how to plan a successful cocktail party. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 2010, p. 939 (2010). https://doi.org/10.1145/1835804.1835923