Semi-Automatic Literature Mapping of Participatory Design Studies 2006-2016

Ari Tuhkala
University of Jyväskylä
P.O. Box 35 (Agora)
Jyväskylä FI-40014, Finland
ari.tuhkala@gmail.com

Tommi Kärkkäinen
University of Jyväskylä
P.O. Box 35 (Agora)
Jyväskylä FI-40014, Finland
tommi.karkkainen@jyu.fi

Paavo Nieminen
University of Jyväskylä
P.O. Box 35 (Agora)
Jyväskylä FI-40014, Finland
paavo.j.nieminen@jyu.fi

ABSTRACT

The paper presents a process of semi-automatic literature mapping of a comprehensive set of participatory design studies between 2006-2016. The data of 2939 abstracts were collected from 14 academic search engines and databases. With the presented method, we were able to identify six education-related clusters of PD articles. Furthermore, we point out that the identified clusters cover the majority of education-related words in the whole data. This is the first attempt to systematically map the participatory design literature. We argue that by continuing our work, we can help to perceive a coherent structure in the body of PD research.

CCS CONCEPTS

• **Information systems** → **Clustering**; *Clustering and classification*; • **Human-centered computing** → **Participatory design**; • **Computing methodologies** → **Natural language processing**;

KEYWORDS

Systematic literature mapping, clustering, participatory design

ACM Reference format:
Ari Tuhkala, Tommi Kärkkäinen, and Paavo Nieminen. 2018. Semi-Automatic Literature Mapping of Participatory Design Studies 2006-2016. In *Proceedings of Participatory Design Conference, Hasselt & Genk, Belgium, August 20–24 (PDC'18),* 5 pages.
https://doi.org/10.1145/3210604.3210621

1 INTRODUCTION

The number of published scientific articles is increasing with high pace and new publication venues, such as journals, conferences, and open access archives, are emerging [19, 28]. To base research on existing body of knowledge, researchers use academic search engines and databases, such as ACM Digital Library. However, the problem is that there is no single search engine that would cover all different publication venues. Consequently, staying up to date with all published research has become arduous.

This problem accumulates in interdisciplinary fields, such as participatory design (PD). The domains, where PD is carried out, have extended from mere offices to schools, hospitals, and other contexts. Consequently, Participatory Design Conference proceedings represent only a fraction of all PD research and a vast amount of research is published in discipline-specific journals and conferences. Thus, researchers need to use several search engines and go through a large number of studies. This raises a question: could computational methods assist researchers in mapping previous knowledge and locating relevant literature?

This paper describes the process of a semi-automatic literature mapping and demonstrates the applicability of our method. First, we built a comprehensive set of all PD literature, that is published between 2006-2016, by systematically collecting studies from 14 different search engines and databases. Then, we conducted a semi-automatic mapping, similarly to the approach in Nieminen et al. [21], of this set of PD studies. We applied unsupervised learning to automatically find a division of topics and a thematic structure in a body of literature. Although, preprocessing the article dataset, determining the number of article clusters in recursive clustering, and interpreting the article clusters were done semi-automatically by the authors. To demonstrate the applicability of our method, we scrutinised the clusters to identify education related PD literature. In addition, we analysed the proportions of education-related words within these clusters. With our method, we managed to identify six education related clusters that covered the majority of education-related words.

This is the first attempt to systematically map PD research literature. For example, Halskov and Hansen [14, p.83] focused on Participatory Design Conference proceedings between 1990-2012. Nunes et al. [22, p.408] focused on IEEE and ACM databases. In line with Halskov and Hansen [14], we propose that systematic mapping of PD literature provides better understanding of where, how, and why PD has been carried out. This helps PD adjuncts to build more solid bases for their work by locating relevant studies and serving as a pre-stage for the actual literature review. For this, our study is an encouraging step. However, we remind that the quality, or relevance to a certain topic, of individual studies should not be based solely on the mapping. Instead, the method is useful when perceiving structure in a large amount of studies.

2 METHOD

2.1 Data Collection

The data collection took place in January 2017 and encompassed 14 databases: ACM Digital library, Bielefeld Academic Search Engine, EBSCOhost Research Databases, ERIC Institute of Education Sciences search, IEEE Xplore Digital library, JSTOR, ProQuest, SAGE Journals, ScienceDirect, Scopus, SpringerLink, Taylor and Francis

PDC'18, August 20–24, Hasselt & Genk, Belgium

ACM ISBN 978-1-4503-5574-2/18/08…$15.00
https://doi.org/10.1145/3210604.3210621

Online, Wiley Online Library, and Thomson Reuters Web of Science. Criteria for database selection were the possibility to 1) export multiple references, 2) export references in Mendeley supported format, and 3) include abstracts in meta data. Thus, we excluded Google Scholar, Semantic Scholar, and CiteSeerX.

From the selected databases, we extracted references where title, abstract, or keywords contained keyword pair "participatory design", publication year was between 2006-2016, and publication type was either journal or conference article. We excluded book chapters, reports, reviews, theses, lectures, and patents, if it was possible in the search engine.

We combined the references in Mendeley to build a PD reference database. First, we removed the references that were obviously faulty, such as references with only empty meta-information. We inspected all references to ensure that the needed metadata (title, abstract, author, year, and publication venue) were included. If the reference had missing fields, we retrieved them manually. If it was not possible to retrieve the missing metadata with any of the search engines, we removed the reference. Some journals, such as MIT Design Issues, did not provide abstracts in meta data, but we could retrieve them from full texts.

We removed initial duplicates with the Mendeley's "Check for Duplicates" tool. We found that the tool cannot identify a duplicate, if an author's name is written with non-latin characters in one reference and with latin characters in another. We resolved these cases manually. After the duplicates were removed, the database consisted of 2939 articles.

2.2 Data processing

We exported the references as a single RIS file, which is a standard tag format by Research Information Systems. We exported the file to our preprocessor, which splits the titles and abstracts of each reference into words around whitespace boundaries and converts the words to lowercase. We removed the common English language stopwords by using the default English stopwords of the Natural Language Tool Kit (NLTK) corpus and stemmatised the words with NLTK Snowball stemmer [7]. The data contained 310013 non-stopword stems, of which 10194 were unique. Hereby, we refer to the word stem simply as *word*.

The top ten frequent words in data were: *design (12826), participatori (4177), use (3660), user (3468), develop (3280), process (2713), research (2525), particip (2330), paper (2142),* and *system (2073).* As can be seen, manual addition of stopwords was needed to guide the algorithm to produce clusters based on the content words, not on the format. Thus, we removed the words from the query string (participatory design): *design (12826), participatori (4177), particip (2330), part (376), pd (844), redesign (100),* and *codesign (54).* We also removed 1515 words based on the knowledge that they are typical research parlance and have no discriminative power, such as: *use (1480), develop (1282), paper (1251), process (1155), user (1130), studi (979), research (949), approach (892), base (848),* and *method (774).* After adding the manual stopwords, the word matrix was ready for clustering the abstracts.

2.3 Data clustering

Clustering means unsupervised classification of observations into groups with a twofold aim: observations within a cluster should be similar to each other, and dissimilar to observations in other clusters [16]. There are various clustering methods and approaches, such as density-based, probabilistic, grid-based, and spectral clustering [1]. The most common clustering methods are hierarchical and prototype-based clustering, of which the basic form of hierarchical clustering is not scalable to the large volume of data because of the pairwise distance matrix requirement [32]. Moreover, many clustering algorithms, including the hierarchical clustering, can produce clusters of arbitrary shape in the data space, which is difficult or even impossible to interpret [25]. Thus, we avoided any dimension reduction technique (see [21]), and used a prototype-based clustering method with a global distance measure to identify groups of similar documents. When each cluster is characterised by a prototypical document, the cluster centroid, it is straightforward to determine a set of most typical documents of a cluster, closest to the centroid, for analysis and interpretation.

Well-known iterative relocation algorithms, most prominently the classical K-means [16, 20], approach clustering in two main steps: *i*) initial generation of K prototypes, *ii*) local search (refinement) of the initial prototypes. In general, initial prototypes should be well separated from each other without being outliers [16, 18]. Lately, the K-means++ algorithm [4], where the random initialisation is based on a density function favouring distinct prototypes, has become the most popular variant. However, because the search phase of these algorithms is locally exploitative, they need repeated restarts in initial prototype regeneration to address global exploration of the best clustering structure [15].

Each document is represented as a bag-of-words (BOW) vector with the number of occurrences of each stemmed word. Because the analysed documents arised from titles, abstracts, and keywords of the articles, we used the so-called *inverse document frequency (idf)* transformation [6] with the scaling function $\log(N/df)$, where N denotes the number of documents and df the overall word frequency. Even when such scaling changes the original data type of integers into real numbers, we still had a strictly discrete set of values with uniform quantization error [26]. Hence, we used the l_1/Cityblock distance, favorable compared in [11], as the distance function and to define the clustering error criterion. This means that we used as the actual clustering algorithm the K-medians++, which is an initialization strategy with the l_1-distance for the density function [4], and median as the document subset prototype within each cluster.

After few initial tests, we noticed the need for recursive application of the document clustering. This was suggested in [29] and successfully used for other application of clustering with c. 500 000 observations by [27]. We hypothesised that this is due to the different shapes and scales of document clusters, as illustrated in [5, Figure 5]. Similarly to the dimension reduction approach in [21], we re-applied the *idf*-scaling at each level of recursive clustering and removed the non-informative words with at most one occurrence within the analysed subset of documents. For the original document set, we had occurrences of 8672 words, which were then reduced to 4760 words of at least two occurrences. During the course of

Algorithm 1 Hierarchical application of prototype-based document clustering

Input: Set of documents with bag-of-words encoding.
Output: Set of prototypes and cluster labels for different refinement levels, through
1: Remove words with at most one instance and apply *idf*-scaling.
2: Cluster the current document set using K-medians++ with 1000 restarts for $K = 2, \ldots, 10$. Check CVAIs and select number of clusters according to the recommendation by one or more cluster indices.
3: Recursively recluster those document clusters which contain more than 300 documents.

recursive clustering, we fixed the threshold of 300 documents as the minimum size of a cluster, which was exposed to refinement.

In an unsupervised clustering scenario, the number of clusters (K) is unknown and needs to be estimated. The quality of the clustering result can be measured with the so-called Clustering Validation Indices (CVAI), which can be divided into three categories [32]: internal, external, and relative. Internal CVAIs, which do not utilise any external knowledge, typically measure the compactness and separation of clusters. To estimate the number of clusters, we used suggestions from the l_1-distance modified set of the best internal CVAIs, as identified in the tests in [15, 17]: K times the clustering error (kCE) [17], Wemmert-Ganciarsky (WG) [12, 13], Ray-Turi (RT) [24], Calinski-Harabasz (CH) [10], and Pakhira-Bandyopadhyay-Maulik (PBM) [23]. As an example, CH and RT suggested six clusters for the original document set.

The experiments were carried out in the Matlab-environment, by using the available *kmeans* clustering algorithm with the 'Cityblock' distance and 1000 repetitions. The overall document clustering method is summarised in Algorithm 1.

2.4 Duplicate identification

After the clustering, we assessed the BOW representation of the documents to identify possible duplicates. We studied the closest document match of the first 100 documents with the Euclidean distance by manually checking whether this was a duplicate. In this way, we detected the first three duplicates, along with their distances to the 'host' document. We used the maximum of these distances, 3.5, as the basis of threshold 3.6 when scanning through all 2939 documents. If we detected a document, of which the distance was less or equal than the threshold, we checked manually both the titles and the abstracts of such documents.

Even when we had used the Mendeley duplicate removal tool in data collection stage, we now identified 86 duplicates. Majority of duplicates occurred because two different search engines had exported titles in a slightly different form. Another source of difference, that was not detected by Mendeley but revealed here, was that the title or abstract of the same article was written in two different languages. However, and slightly alarmingly, we identified cases where the same abstract, or almost the same, was used in a different article, such as: [2] and [3], [8] and [9], [30] and [31].

3 FINDINGS

Figure 1 shows an overview of the total number of 53 clusters and 49 unique articles on 21 refinement levels. The clusters are marked as boxes, where the number represents the size of a cluster (number of articles). The clusters containing more than 300 articles, and thus chosen for re-clustering, are emphasised with bold lines. Unique articles, that were not assigned to any cluster, are emphasised with dotted lines. The clusters that are vertically aligned belong to the same refinement level. For example, clustering of the original document set provided six clusters of sizes 141, 123, 1984, 125, 103, and 377, of which the third and the sixth were reclustered.

3.1 Education-related clusters

Interpretation of document clusters was based on interactive expert analysis of the most frequent words and the most representative documents. First we assessed the 20 most common words of a cluster, how much of the total occurrences in the whole document set they cover (in percentages). Then, similarly to the BOW duplicate detection, we scanned through the titles, abstracts, and keywords of ten documents in the cluster, closest to the cluster prototype. In this way, out of the whole set, we identified six clusters related to the joint theme of PD in education (Table 1).

3.2 Word portions of education-related clusters

The preprocessor provided a list of all words in the data, sorted from the highest frequency to the lowest. From this list, we selected education related words with the frequency of 30 or higher. Table 2 presents these words (Word), word frequency in the whole data (Freq), and number of articles that include the word at least once (AF). Then, the table displays the word portions of all six education related clusters. The number shows how many percentages the cluster covers from the total word frequency. Thus, the final column (Total) shows how many percents the six education related clusters cover from total frequency all together.

4 DISCUSSION

Due to the restricted space of a short paper, this study concentrated on demonstrating the semi-automatic clustering method for the systematic literature mapping of PD studies. In the future, we provide more detailed analysis of the clusters and the most representative articles. Furthermore, we analyse articles that the algorithm could not assign to any cluster, because they may represent some exceptional studies in PD field. In addition, we provide an access to the collected literature by implementing a web interface that uses Mendeley API [1].

The main challenge of the systematic mapping was that data collection stage took a lot of manual work. This was due to the faced usability problems in the used academic search engines. For illustration: ACM Digital Library did not provide abstracts when exporting multiple references, so we copied them manually. In EBSCOhost, references needed to be moved to a folder (50 at a time) and downloaded, with a limit of 150 references at a time. In JSTOR, references needed to be selected by clicking a checkbox, one by one, and then exported. To overcome this kind of defiencies, academic search engines should improve and standardise their database meta-information fields, search protocols, and exporting features.

[1] Mendeley API: http://dev.mendeley.com, retrieved 15.1.2018

Figure 1: Cluster map

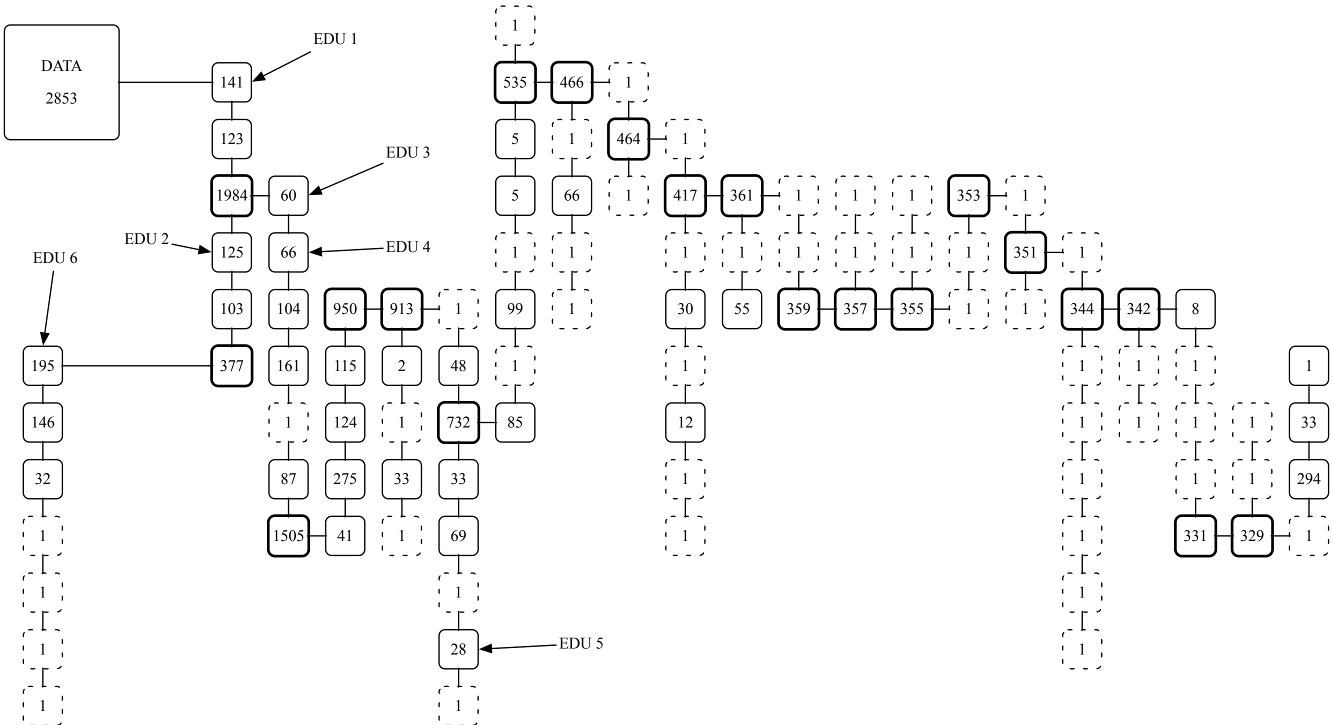

Table 1: PD in education clusters

Cluster	Articles	Journal / conference	Name
EDU 1	141	79 / 62	PD in learning, learning technology including edugames, and learning environment design
EDU 2	125	59 / 66	Designing with children and for children (also with special needs)
EDU 3	60	29 / 31	Game design and gaming in design
EDU 4	66	38 / 28	Teaching PD and PD in learning design
EDU 5	28	18 / 10	PD with students and for educational activities
EDU 6	195	100 / 95	Educational and assistive technology design

Table 2: Word portions of the PD in education clusters

Word	Freq	AF	EDU 1	EDU 2	EDU 3	EDU 4	EDU 5	EDU 6	Total
learn	1372	590	55.4	5.4	1.7	4.3	0.6	4.2	71.6
educ	814	392	16.8	5.4	3.3	13.2	9.0	12.4	60.1
student	791	265	24.9	0.8	1.8	41.3	1.7	5.1	75.6
school	418	184	16.3	15.3	4.1	10.5	4.8	8.2	59.2
teacher	390	151	38.8	7.1	4.5	11.6	2.4	9.2	73.6
teach	172	101	27.9	1.2	4.2	19.4	1.2	6.1	60.0
learner	138	70	56.3	3.7	2.2	5.2	1.5	4.4	73.3
classroom	115	55	36.9	5.8	1.0	12.6	8.7	6.8	71.8
instruc	120	64	32.2	2.5	3.4	11.9	2.5	2.5	55.0
pedagog	45	33	42.2	13.3	4.4	8.9	0.0	15.6	84.4
curriculum	42	30	26.2	0.0	7.1	11.9	4.8	0.0	50.0
pedagogi	30	24	42.9	3.6	0.0	3.6	0.0	0.0	50.1
Mean	370.6	163.3	34.7	5.4	3.2	12.9	3.1	6.2	65.4

REFERENCES

[1] Charu C Aggarwal and Chandan K Reddy. 2013. *Data clustering: algorithms and applications*. CRC press.

[2] Meghan Allen, Joanna McGrenere, and Barbara Purves. 2007. The Design and Field Evaluation of PhotoTalk: A Digital Image Communication Application for People. In *Proceedings of the 9th International ACM SIGACCESS Conference on Computers and Accessibility (Assets '07)*. ACM, New York, NY, USA, 187–194. https://doi.org/10.1145/1296843.1296876

[3] Meghan Allen, Joanna McGrenere, and Barbara Purves. 2008. The Field Evaluation of a Mobile Digital Image Communication Application Designed for People with Aphasia. *ACM Trans. Access. Comput.* 1, 1, Article 5 (May 2008), 26 pages. https://doi.org/10.1145/1361203.1361208

[4] David Arthur and Sergei Vassilvitskii. 2007. K-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 1027–1035.

[5] Sami Äyrämö. 2006. *Knowledge Mining Using Robust Clustering*. Jyväskylä Studies in Computing, Vol. 63. University of Jyväskylä.

[6] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. 1999. *Modern information retrieval*. Vol. 463. ACM press New York.

[7] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

[8] Tone Bratteteig and Ina Wagner. 2010. Spaces for Participatory Creativity. In *Proceedings of the 11th Biennial Participatory Design Conference (PDC '10)*. ACM, New York, NY, USA, 51–60. https://doi.org/10.1145/1900441.1900449

[9] Tone Bratteteig and Ina Wagner. 2012. Spaces for participatory creativity. *CoDesign* 8, 2-3 (2012), 105–126. https://doi.org/10.1080/15710882.2012.672576

[10] Tadeusz Caliński and Jerzy Harabasz. 1974. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods* 3, 1 (1974), 1–27.

[11] John Collins and Kazunori Okada. 2012. A Comparative Study of Similarity Measures for Content-Based Medical Image Retrieval. In *CLEF (Online Working Notes/Labs/Workshop)*.

[12] Bernard Desgraupes. 2013. Clustering indices. *University of Paris Ouest-Lab Modal'X* 1 (2013), 34.

[13] Germain Forestier, Pierre Gançarski, and Cédric Wemmert. 2010. Collaborative clustering with background knowledge. *Data & Knowledge Engineering* 69, 2 (2010), 211–228.

[14] Kim Halskov and Nicolai Brodersen Hansen. 2015. The diversity of participatory design research practice at PDC 2002–2012. *International Journal of Human-Computer Studies* 74 (feb 2015), 81–92. https://doi.org/10.1016/j.ijhcs.2014.09.003

[15] Joonas Hämäläinen, Susanne Jauhiainen, and Tommi Kärkkäinen. 2017. Comparison of Internal Clustering Validation Indices for Prototype-Based Clustering. *Algorithms* 10, 3 (2017), 1–14.

[16] Anil K. Jain. 2010. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters* 31, 8 (2010), 651–666.

[17] Susanne Jauhiainen and Tommi Kärkkäinen. 2017. A Simple Cluster Validation Index with Maximal Coverage. In *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning - ESAINN 2017*. 293–298.

[18] Shehroz S Khan and Amir Ahmad. 2013. Cluster center initialization algorithm for K-modes clustering. *Expert Systems with Applications* 40, 18 (2013), 7444–7456.

[19] Peder Olesen Larsen and Markus von Ins. 2010. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics* 84, 3 (sep 2010), 575–603. https://doi.org/10.1007/s11192-010-0202-z

[20] S. Lloyd. 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory* 28, 2 (1982), 129–137.

[21] Paavo Nieminen, Ilkka Pölönen, and Tuomo Sipola. 2013. Research literature clustering using diffusion maps. *Journal of Informetrics* 7, 4 (oct 2013), 874–886. https://doi.org/10.1016/j.joi.2013.08.004

[22] Eunice P. S. Nunes, Alessandro R. Luz, Eduardo M. Lemos, and Clodoaldo Nunes. 2016. *Approaches of Participatory Design in the Design Process of a Serious Game to Assist in the Learning of Hospitalized Children*. Springer International Publishing, Cham, 406–416. https://doi.org/10.1007/978-3-319-39513-5_38

[23] Malay K Pakhira, Sanghamitra Bandyopadhyay, and Ujjwal Maulik. 2004. Validity index for crisp and fuzzy clusters. *Pattern recognition* 37, 3 (2004), 487–501.

[24] Siddheswar Ray and Rose H Turi. 1999. Determination of number of clusters in k-means clustering and application in colour image segmentation. In *Proceedings of the 4th international conference on advances in pattern recognition and digital techniques*. Calcutta, India, 137–143.

[25] Mirka Saarela, Joonas Hämäläinen, and Tommi Kärkkäinen. 2017. Feature Ranking of Large, Robust, and Weighted Clustering Result. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 96–109.

[26] Mirka Saarela and Tommi Kärkkäinen. 2015. Analysing Student Performance using Sparse Data of Core Bachelor Courses. *JEDM-Journal of Educational Data Mining* 7, 1 (2015), 3–32. http://www.educationaldatamining.org/JEDM/index.php/JEDM/article/view/JEDM056

[27] Mirka Saarela and Tommi Kärkkäinen. 2015. Do country stereotypes exist in educational data? A clustering approach for large, sparse, and weighted data.. In *Proceedings of the 8th International Conference on Educational Data Mining (EDM 2015)*. 156–163.

[28] Richard Van Noorden. 2014. Global scientific output doubles every nine years. (2014).

[29] P. Wartiainen and T. Kärkkäinen. 2015. Hierarchical, prototype-based clustering of multiple time series with missing values. In *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning - ESAINN 2015*. 95–100.

[30] Y. Yu and Z. Liu. 2006. Integrated scenario-based design method for interactive online teaching system. In *2006 International Conference on Computational Inteligence for Modelling Control and Automation and International Conference on Intelligent Agents Web Technologies and International Commerce (CIMCA'06)*. 153–153. https://doi.org/10.1109/CIMCA.2006.128

[31] Y. Yu and Z. Liu. 2006. Research on a user-centered design method for interactive online teaching system. In *2006 International Conference on Communication Technology*. 1–4. https://doi.org/10.1109/ICCT.2006.341739

[32] Mohammed J Zaki and Wagner Meira Jr. 2014. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.