ISSN: 2623 - 4629

Journal of Integrated Information & Management

e-Journal

ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

UNIVERSITY OF WEST ATTICA

Volume 9 - Number 2 / Jul - Dec 2024

https://ejournals.epublishing.ekt.gr/index.php/jiim

ISSN: 2623 - 4629

Journal of Integrated Information & Management

e-Journal

Volume 9 - Number 2 / Jul - Dec 2024
https://ejournals.epublishing.ekt.gr/index.php/jiim

Journal of Integrated Information Management

Vol. 9 – No 2
Jul – Dec 2024

ISSN 2623-4629 (on-line)

Publisher

Department of Archival, Library and Information Studies, School of Management, Economics and Social Sciences, University of West Attica

Table of contents

Research papers

Editorial message

Dear Colleagues,

JIIM is an international, multidisciplinary, blind, peer-reviewed electronic open-access journal that publishes research efforts on all aspects and issues regarding Information Science and Integrated Information Management. JIIM is the official journal of the Department of Archival, Library & Information Studies, University of West Attica (UNIWA), and it is available through the Greek National Documentation Centre (EKT) ePublishing platform for electronic journals: https://ejournals.epublishing.ekt.gr/index.php/jiim.

The current issue publishes research articles about information as a physical reality or a human tool, a literature review on the citation indexes, physical and digital heritage experiences of Asia Minor refugees in 1922 in Attica, Greece, and graph databases graph neural networks.

In the beginning, the nature of information is explored. The perspective of information in the realm of reality is explored in natural, physical, and measurement settings. Firstly, the information on natural signals versus noise is analyzed through physical measurements of speech experiments. Secondly, the information in classifying relative things into groups is analyzed by clustering flowers. A method for the selection of the appropriate number of clusters in the K-Means clustering process is presented.

The following paper is a literature review on citation indexes. The published comparisons between Web of Science, Scopus, and Google Scholar have been explored through a literature survey covering the era between 2004 and 2021. The majority of works utilize multiple citation indexes in their research taking advantage of their unique features.

The third paper is about the physical and digital storytelling of the Asia Minor refugees after the end of the Greco-Turkish war in 1922 in Attica, Greece. The project aims to deliver the refugees' experiences through guided tours and approaches for younger audiences, such as a coloring book, a narrative card game, and an activity book. The phygitality, the hybrid experience with augmented reality has been used as a delivery tool of heritage information.

Finaly, graph databases and neural networks of large-scale data derived from social networks are presented. An IMDB dataset of movies and people contributing to a film has been used in the data modeling, querying, and graph representation paradigm.

We welcome special Issues proposals that should be emailed to the Associate Professor Dimitrios Kouis (dkouis@uniwa.gr) or Assistant Professor Artemis Chaleplioglou (artemischal@uniwa.gr). We expect your contributions, active support, remarks, and points of improvement. The JIIM Editorial Team would like to wish to all of you Happy Holidays and a Happy New Year.

Assistant Professor - Editor

Artemis Chaleplioglou

Department of Archival, Library and Information Studies University of West Attica

Agiou Spyridonos Str., 12243 Aegaleo, Athens, Greece

Journal of Integrated Information Management electronic journal

Information: A physical reality or a humanly tool? From the model order to the appropriate number of clusters

Markos Dendrinos

Professor of Informatics and History & Philosophy of Science, Department of Archival, Library and Information Studies, University of West Attica, Athens, Greece

mdendr@uniwa.gr [ORCID: 0000-0001-5675-3069]

Article Info

Article history:

Received 06 May 2024

Received in revised form 12 June 2024

Accepted 25 June 2024

http://dx.doi.org/10.26265/jiim.v9i2.39839

Abstract:

This paper is a presentation of two important types of information regarding natural signals and groups of relative things. The measure of the first information type is the order of the system that produces the signal, while the second one is the appropriate number of distinct clusters for the most effective classification of a certain group into narrower and more representative subgroups. Firstly a review is given concerning the two approaches, and then a certain method is proposed for the identification of the correct number of clusters to be used in K-Means clustering process. The two information measures, model order and number of clusters, can be considered as two equivalent views of an inherent natural element, an objective order behind any physical system.

Index Terms — information, information measure, information detection, model order, K-Means, clustering.

I. INTRODUCTION

Heisenberg (1927) has formulated the principle of uncertainty. According to this principle, the combined accuracy of measurement of the position and the momentum (velocity) of a particle has a lower bound, which is of the order of magnitude of the Planck constant h. This principle constitutes an alarm against the traditional belief of the unavoidable scientific progress in the description of the reality. The empiricists emphasize that Heisenberg's principle of uncertainty concerns mainly nanoscale measurements and its power seems to be eliminated at the macroscopic level. This fact does not override the ultimate scientific truth that the humanly observations and measurements have certain limits and we must face the reality in other terms. Heisenberg and the radical group of the quantum physicists unsettled the scientific background even more, as they denied the physical reality at all.

Thus a fundamental question has been set: Do we have the right to assign a value to the momentum of an electron just before its final measurement? Is it real? According to Heisenberg it is not. Before the final measurement, the best we can attribute to the electron is some unsharp or fuzzy momentum. The ontological sense of the elementary units of matter has been overturn[1].

If we add to the whole image the role of the observer in the double slit experiment, where the nature and the effects of the beam of electrons shot against the slit changes depending on the presence or not of an observer, then, indeed, we can say that reality has lost all its supposed compactness becoming a "smoke of probabilities". We must also take into account the complex not exclusive nature of the matter, which is particle-like and wave-like simultaneously, as well as the quantum entanglement between any couple of particles, where the measurement of any characteristic of the first affects in no time the value of the second. These properties were the stimulus for quantum computing where the basic unit of information, the qubit has not a distinct certain value 0 or 1 but it is a superposition of both of them, obtaining a value only in the case of its use, affecting also any other qubit entangled with it.

From all the above we can see that the structure of the universe as well as the nature of the information has radically changed. I'll try in the subsequent sections to find a suitable place for an ontological structure of information within this extraordinary contemporary view of the reality. In such a perspective the information could claim the role of the ultimate element of reality substituting the particle-like or wave-like matter.

Information has a lot of significances: a measure of expectancy towards the most economical representation (information entropy), a language system for human communication along with intentionally created symbolical languages (propositional and predicate calculus), a number of harmonics (resonances) of a natural signal related to the order of the underlying system, classification of an object to a predefined or unknown number of clusters. The last two cases will be the subject of this paper, in an attempt to find any common ground between them and a way to shed a light to the open question of the objectivity or human

[1] https://plato.stanford.edu/entries/qt-uncertainty/

dependent nature of the information. Model order detection and clustering processes will be used as the tools to this end.

II. Information vs Noise

In this section we'll make a review in the literature concerning order detection of synthetic harmonic signals as well as natural signals, such as speech signals and respiratory signals. The order of these signals is meant as their harmonic content. In the case of the natural signals this is an objective information indicator, corresponding to the anatomical characteristics of the system that produces them (phonetic tube, lung structure etc.). In the case of synthetic harmonic signals the order is twice their number of harmonics; the order detection is problematic mainly in the case of a signal embedded in white or colored noise and it proves to be of low success when the Signal to Noise Ratio (SNR) is low. The general idea is to reduce the dimensionality of a signal retaining only the dimensions that contain the real information part through applying Principal Component Analysis [3].

In this frame SVD methodology can be applied in the data or autocorrelation matrix of the signal in order to find the singular values (a generalization of eigen-values) of the matrix and remove all the unnecessary values which are supposed to correspond to the noise. The required order is the number of the remaining singular values, which consist the information subspace. The identification of this number is crucial, since this number shows the orthogonal axes into which the real information is decomposed, that is the signal subspace against the noise subspace. An analysis by synthesis procedure based on the SVD methodology has been proposed in [1] for synthetic signals and in [8] for speech signals, which has been a basic reference for many articles about order detection, speech enhancement and noise cancellation since then.

Bakamidis, Dendrinos and Carayannis [1], in the frame of decomposing the synthetic harmonic signals through SVD Principal Component Analysis, defined a new criterion NEE (standing for Noise Error Estimation), superior to the Rao criterion ([12]), aiming at detecting the number of harmonics in the presence of noise [1]. Consecutive reconstructions are performed and the resulting error power is compared to the noise variance in order to get the best approximation of the original non corrupted signal. The number of the singular values corresponding to a reconstruction error power as close as possible to the noise variance gives the parsimonious order. The existence of such a criterion is important for both high quality reconstruction and accurate spectral analysis. Various spectral estimation techniques used on the reconstructed signal give the possibility to retrieve harmonics in highly noisy environment along with availability of very short data lengths.

In this frame, a number of successively reconstructed signals, derived through this analysis by synthesis procedure, were used for determination of the harmonics order of synthetic signals in hard conditions: limited number of samples, low SNR, and close frequencies. The introduced criterion shows not only the number of the harmonics (complex sinusoids) but also the best order for reconstruction (leading to a signal as close to the pure one as possible). This methodology has been applied in the field of enhancement of noisy speech with great success.

Some indicators of the success rates of the proposed method are the following: Case A. Signal length 100, SNR 0 db, sinusoidal frequencies: 0.4 f0, 0.43 f0 (f0 is the Nyquist rate): Success rate: 96%. Case B. Signal length 100, SNR -3 db, sinusoidal frequencies: 0.4 f0, 0.43 f0: Success rate: 80%. This proved to be a significant improvement compared to the 83% and 53% of Rao criterion (Rao et. al., 1988), in spite of the shorter used data records.

NEE criterion has been also applied in real conditions, as in speech enhancement from noise and detection of Obstructive Sleep Apnea (OSA) in medicine. In both cases the detection of the exact order of the signal is the key factor for the correct reconstruction of the signal and the abolishment of noise along with the minimum distortion of the initial signal.

Dendrinos, Bakamidis and Carayannis [8] proposed a speech enhancement technique, based on principal component analysis and a new criterion for the selection of the parsimonious number of real signal components aiming at noise-free signal regeneration [8]. Both isolated phoneme and continuous speech experiments were presented. The results have been evaluated by informal listening and SNR computations, which show that the methodology had an improved performance compared to other widely used methods. The estimated number of singular values to be retained is supposed to keep all the necessary information content of the speech signal, while the omitted ones correspond to the additive noise.

In this context various phoneme and phrases embedded in noise have been reconstructed after recovering the correct model order of the natural signal.

As you can see in Fig. 1 (a), the order of a noisy pronounced phoneme [e] was selected where the NEE measure takes its lower value (order=9). The reconstructed signal corresponding to this value is the signal with the greater segmental SNR, showing the best possible noise elimination. Similarly, in (b) the order of a noisy pronounced phoneme [n] was selected where the NEE measure takes its lower value (order=6).

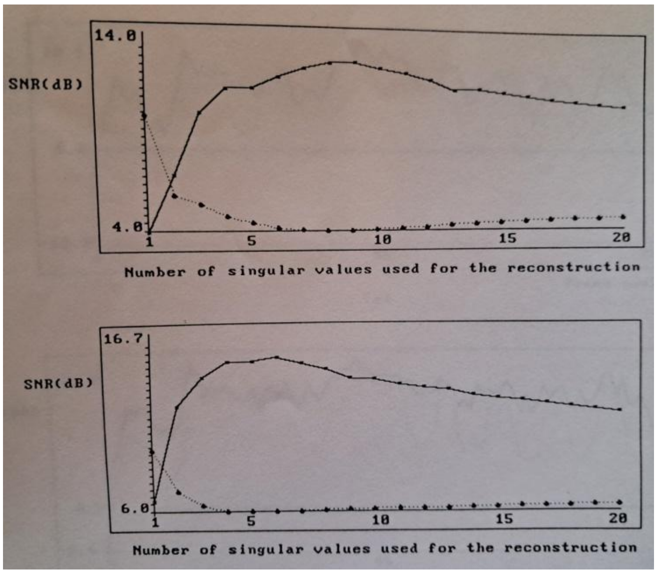
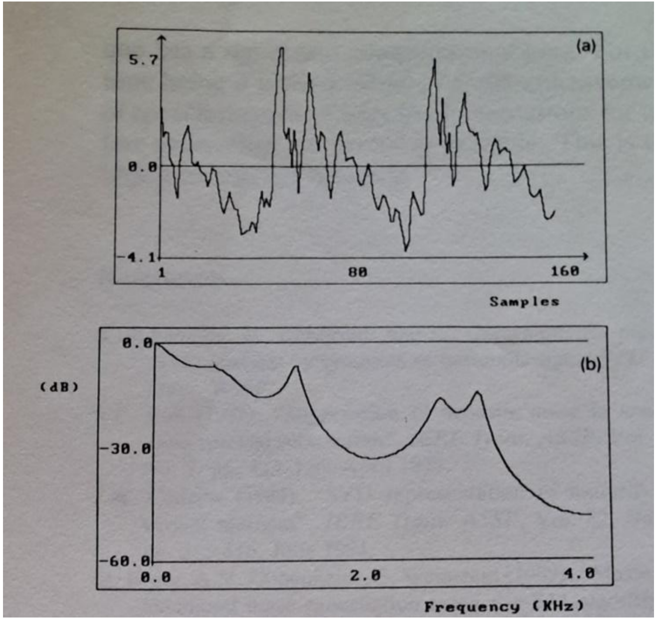


Fig.1. (a) The segmental SNR's (continuous line) and the NEE criterion values (dotted line) for the phoneme [e] embedded in noise 10 dB, (b) The same for the phoneme [n].

Next figure shows the recovered spectral characteristics of a Greek dictated phrase, which had been hidden due to the noise. The phrase was divided in many overlapping frames, each of which was analyzed. After selecting the parsimonious order for each frame, a reconstructed signal is derived. The correct order selection for each frame is of great importance in this process.



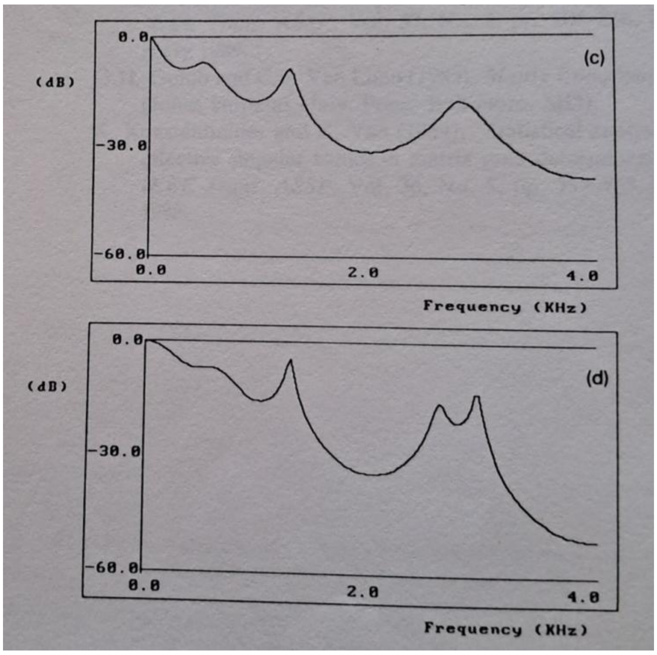


Fig. 2. (a) An examined signal frame [No. 21] of a dictated sentence in Greek ('i mana mou emene me mena') embedded in noise with global SNR = 20 dB, (b,c,d) The LPC (Linear Predictive Coding) spectral densities of the original, the noisy and the reconstructed signals, respectively.

Since then, various alternative approaches, taking into account more or less the NEE technique, have been suggested, which try to enhance further the order detection efficacy. Besides, a lot of speech enhancement methods based more or less in the above methodology have also been suggested in the meantime.

Wang et al (2015) [15] propose an efficient framework and algorithm for the one dimensional harmonic retrieval problem in the case of either additive colored Gaussian or non-Gaussian noise, in the hard condition of very closely spaced harmonic signals in the frequency domain. The wavelet packet (WP) method to the blind source separation (BSS) based harmonic retrieval model is utilized.

Tong et al. (2017) [14] study the problem of parallel waveform enhancement via the multi-sensor fusion technology. They represent the observed multiple noisy observations as a 3-D tensor, and propose two novel approaches in the time domain, i.e. the transforming and filtering (TAF) approach and the direct multidimensional filtering (DMF) approach, for parallel waveform recovery and interference suppression. The system can produce an estimate of the clean waveform in each sensor channel simultaneously, as it is implied by the term "parallel". In the TAF approach the observed tensor is transformed into a different domain where the noise can then be filtered by discarding the insignificant coefficients. Noise reduction is achieved in the DMF approach directly by applying multidimensional filtering on the observed tensor. Both approaches are "blind" since they do not require a priori frequency responses between the desired source and distributed sensors.

Deppisch et al. (2023) [9] propose a similar subspace method that decomposes SRIRs (Spatial room impulse responses) into a direct part, which comprises the direct sound and the salient reflections, and a residual, to facilitate enhanced analysis and rendering methods by providing individual access to these components. The proposed method is based on the generalized singular value decomposition and interprets the residual as noise that is to be separated from the other components of the reverberation. Large generalized singular values are attributed to the direct part, which is then obtained as a low-rank approximation of the SRIRs.

Balasubramanian et al. (2023) [2] suggest a combined method for speech enhancement, based on both speech cochleagram and visual cues using Audio-Visual Multichannel Convolutional Neural Network (AVMCNN). Several researchers have already shown that speech enhancement using visual data as an additional input along with the audio data is more effective in minimizing the acoustic noise present in the speech signal. This work proposes a novel CNN-based audio-visual Ideal Ratio Mask (IRM) estimation model.

Christensen et. al. (2016) [6] investigate the potential performance of generalized subspace filters for speech enhancement in cocktail party situations with very poor signal/noise ratio, e.g. down to -15 dB. Performance metrics such as output signal/noise ratio, signal/distortion ratio, speech quality rating and speech intelligibility rating are mapped as functions of two algorithm parameters. The paper results to a recommended trade-off between noise, distortion and subjective performances.

Dendrinos (1998) [6] has used spectral characteristics of the respiratory signals in order to discriminate between normal (just breathing or snoring) and pathological (apnea) cases. Research has been held for the selection of the suitable features of the examined signals. The features which were proved to be information distinctive were the spectral energy between 0 and 700 Hz and the one between 700 and 1400 Hz. The spectral ration between them takes high values in the cases of apneas and pauses, whereas it takes much lower values during breathing or snoring periods.

Steinhaouer et al. (2019) [13] present a system to detect symptoms of allergic rhinitis remotely by using uttered speech and by exploiting its specific spectral characteristics. Based on the principles of adaptive modelling and fundamental frequency variations (jitter) as well as speech analysis by means of acoustic models, the proposed technique achieves an efficient classification of patients from uttered speech. A Singular Value Decomposition based iterative approach is used for the accurate estimation of the jitter and Hidden Markov Models are implemented to model the 32 phonemes. The final decision is derived by optimally combining the individual estimates providing a tool for the automatic diagnosis of allergic rhinitis.

III. Information as a classification tool

The model order, conceived as the order of a system that produces natural signals, such as speech or respiratory signals, is an indicator of an objective information, independent of the humanly intervention. This can be considered as information hidden in nature, which, under the appropriate approaches, as those developed in the previous section, can be revealed.

In this section we'll deal with a continuous implicit or explicit process in the frame of human understanding of our environment: the classification of things or events in categories aiming at an easier confrontation of the 'information' that incessantly shoots us. Data classification gives the human the first phase of the sense of the things, as they are classified into categories (or classes) according to some carefully selected features. The origin of this endeavor goes back to ancient times, in the philosophical context of platonic ideas, which were seen as representatives of groups of things with similar characteristics. There is a great discussion about the objectivity or subjectivity of the process of categorization, which can be thought of as a part of the open question of the nature of the information itself.

Data analysis and information generation is a critical component of intelligence. The formulation of questions, collection and preparation of data, analysis and interpretation which in turn may lead to responses by the intelligent agent are not restricted to humans. Several living organisms have the ability to obtain their environment, extract conclusions, and produce responses. They may also predict events or count or cluster similar objects.

In the frame of classification there is a great branch of statistical informatics offering us tools for supervised or unsupervised clustering, that is division of a number of given vectors of data to a number of clusters. The data here are the features of the objects to be discriminated and they are either numerical or nominal. The number of clusters (classes) where the objects will be classified is critical and it reminds the model order selection problem, discussed in the previous section. One widely used method is the trade-off between minimizing the within-cluster sum of squares (a measure of how tight each cluster is) and maximizing the between-cluster sum of squares (a measure of how separated each cluster is from the others)[2].

In this context we are seeking a method for absolutely automatic unsupervised clustering leading to the estimation of model order (the parsimonious number of clusters). The results are compared to the experts' estimation of that order, since we suppose that an experts group has a more accurate perception of a certain area; for example a botanist may conclude differently to iris flowers attributes, due to a 'positive bias' extracted from his exhausting measurements, whilst a non-expert may not. In this way objective and subjective information detection seem to mutually converge.

The method proposed in this paper is the consecutive application of K-means clustering algorithm, through increasing one by one the number of the clusters. The used environment is the open source data mining platform WEKA, We compute in each case the difference of the 'within cluster sum of squared errors' SSE value from the

[2] https://towardsdatascience.com/10-tips-for-choosing-the-optimal-number-of-clusters-277e93d72d92

corresponding one in the previous case (current number of clusters-1). Then the rhythm of decrease (difference of differences) is estimated as a discrete measure of second derivative of SSE. The maximum value of the sequence of the 'rhythm of decrease' values gives the optimal number of clusters.

Two classification cases are presented. The one case concerns a group of iris flowers characterized by 4 metrical features of their petals and sepals. The SEE criterion gives the true number of categories-clusters, that is 3, the same with the a-priori classification of the flower instances given by the plant scientists into Iris Setosa - Iris Versicolour - Iris Virginica. The second case concerns a group of image segmentation data extracted from 7 distinct categories of outdoor images: brickface, sky, foliage, cement, window, path, grass. The SEE criterion gives again the true number of categories-clusters, that is 7.

A. *Iris Plants Database*

- **Creator:** R.A. Fisher.
- **Donor:** Michael Marshall (MARSHALL%PLU @io.arc.nasa.gov), July, 1998.
- **Number of Instances:** 150 (50 in each of three classes).
- **Number of Attributes:** 4 numeric attributes and the class.
- **Attribute Information:** (1) sepal length in cm, (2) sepal width in cm, (3) petal length in cm, (4)] petal width in cm, (5) class: Iris Setosa - Iris Versicolour - Iris Virginica.
- **Missing Attribute Values:** None.

You can see below the data extracted after the consecutive K-means applications.

Table 1. Iris flower. K-Means clustering in WEKA environment. (a) Number of clusters. (b) SSE: within cluster sum of squared errors. (c) Difference between consecutive SSE values. (d) Difference rhythm between consecutive SSE values.

Number of clusters	SSE	SSE diff	SSE diff rhythm
2	**12,14368828**	5,145574	
3	**6,998114005**	1,465283	3,680291
4	**5,532831003**	0,402046	1,063237
5	**5,130784647**	0,443769	-0,04172
6	**4,687015166**	0,929425	-0,48566
7	**3,757589924**	0,34968	0,579745
8	**3,40790992**	3,40791	-3,05823

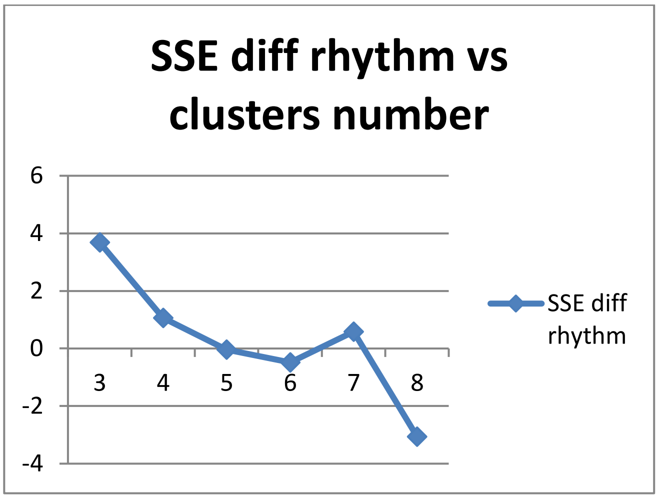


Fig.3. SEE difference rhythm vs clusters number for Iris-flower data.

We can see from the above diagram the optimal number of clusters: 3, corresponding to the higher SSE difference rhythm value.

B. *Image Segmentation data.*

- Creators: Vision Group, University of Massachusetts.
- Donor: Vision Group (Carla Brodley, brodley@cs.umass.edu), November 1990.
- Relevant Information: The instances were drawn randomly from a database of 7 outdoor images. The images were hand-segmented to create a classification for every pixel. Each instance is a 3x3 region.
- Number of Instances: Training data: 210 (30 instances per class). Test data: 2100 (300 instances per class).
- Number of Attributes: 19 numeric attributes and the class.
- Attribute Information: (1) region-centroid-col: the column of the center pixel of the region. (2) region-centroid-row: the row of the center pixel of the region. (3) region-pixel-count: the number of pixels in a region = 9. (4) short-line-density-5: the results of a line extractoin algorithm that counts how many lines of length 5 (any orientation) with low contrast, less than or equal to 5, go through the region. (5) short-line-density-2: same as short-line-density-5 but counts lines of high contrast, greater than 5. (6) vedge-mean: measures the contrast of horizontally adjacent pixels in the region. This attribute is used as a vertical edge detector (mean). (7) vegde-sd: (see 6, standard deviation), (8) hedge-mean: measures the contrast of vertically adjacent pixels. Used for horizontal line detection (mean). (9) hedge-sd: (see 8, standard deviation). (10) intensity-mean: the average over the region of (R + G + B)/3. (11) rawred-mean: the average over the region of the R value. (12) rawblue-mean: the average over the region of the B value. (13)

rawgreen-mean: the average over the region of the G value. (14) exred-mean: measures the excess red: (2R - (G + B)). (15) exblue-mean: measures the excess blue: (2B - (G + R)). (16) exgreen-mean: measures the excess green: (2G - (R + B)). (17) value-mean: 3-d nonlinear transformation of RGB. (Algorithm can be found in: Foley, James D. and Van Dam, Andries, 1982). (18) saturatoin-mean: (see 17). (19) hue-mean: (see 17).

- Class Distribution: brickface, sky, foliage, cement, window, path, grass.
- Missing Attribute Values: None

You can see below the data extracted after the consecutive K-means applications.

Table 2. Image Segmentation data. K-Means clustering in WEKA environment. (a) Number of clusters. (b) SSE: within cluster sum of squared errors. (c) Difference between consecutive SSE values. (d) Difference rhythm between consecutive SSE values.

Number of clusters	SSE	SSE diff	SSE diff rhythm
4	361	40	
5	321	31	9
6	290	27	4
7	263	14	13
8	249	10	4
9	239	21	-11
10	218	16	5
11	202	10	6
12	192	3	7
13	189	10	-7
14	179	2	8
15	177	9	-7
16	168	5	4
17	163	3	2
18	160	4	-1
19	156	3	1
20	153	1	2
21	152	8	-7
22	144	5	3
23	139	5	0

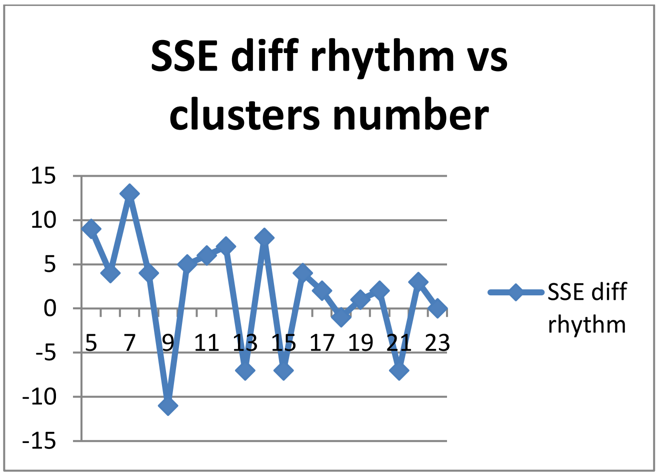


Fig.4. SEE difference rhythm vs clusters number for Image segmentation data.

We can see from the above diagram the optimal number of clusters: 7, corresponding to the higher SSE difference rhythm value.

IV. Conclusion

This article deals with the open question of the nature of information. Is the information inherent element of the things or just a humanly tool aiming at the better understanding of our environment? In this frame two types of information were examined. The first information type concerns the real informative part of a natural signal after eliminating the noisy components. The second information type concerns the classification of things or events, members of a wide group, into smaller groups, so that they both inherit the general characteristics and they are also, as members of the distinct subgroups, discriminated between each other. Distinction is an inherent feature of nature, while at the same time the classification process satisfies the humanly need for better understanding through specification. The outcome of the first information type is the number of harmonics of the signal, which identifies with the order of the underlying system. The outcome of the second information type is the number of clusters where a group of things is classified. These two measures seem to be related, since each of them shows the number of independent projection axes of the examined features. The innovative part of this paper is the presentation of a method for the selection of the appropriate number of clusters in the K-Means clustering process.

Capurro gives a very interesting interpretation of the concept of information through its etymological connection to the concept of form. Information in this sense is the process of forming a piece of matter or, metaphorically, human knowledge (Capurro, 1996, 256-270) [5]. Shannon and Weaver in their mathematical theory of communication establish a neutral, independent of the human comprehension, meaning of information content, that of the information entropy. Carl Friedrich von Weizsaker relates this neutral substance, which is neither matter nor energy, to the Platonic eidos and the Aristotelian form [5].Therefore, information content could

be meant as an autonomous substance, a pattern archetype for in-forming something (Capurro, 1991) [4].
Under this view, that extracted number, either as the model order or as the multiplicity of clusters, seems to be an inherent element of a natural process, either in the frame of signals to be eliminated from noise or in the frame of a group of things to be divided through some distinctive features. Thus the harmonics of a signal or the clusters of an overall group appear as non-temporal characters of the temporal phenomena. These characters remind us of the platonic ideas, a general concept of genera and species/ eide/ ideas/ forms. We could suppose that a genus itself, such as the iris flower studied in the previous section, is a part of the nature corresponding to a certain design, and holding an objective character apart from the human endeavor to categorize it as a means of better insight and understanding.

REFERENCES

[1] Bakamidis, S.; Dendrinos, M. and Carayannis, G., "SVD Analysis by Synthesis of Harmonic Signals", Acoustics Speech and Signal Processing (ASSP), IEEE, Vol. 39, No.2, pp. 472-477, Feb. 1991, https://ieeexplore.ieee.org/document/80831

[2] Balasubramanian, S.; Rajavel, R.; Kar, Asuthos, "Ideal ratio mask estimation based on cochleagram for audio-visual monaural speech enhancement", *Applied Acoustics,* Volume 211, August 2023.

[3] Carayannis, G. and Gueguen, C. "The factorial linear modelling : a Karhunen-Loeve approach to speech analysis", *Proc. IEEE Int. Conf. Acoust. Speech Signal Process., ICASSP,* pp. 489-492, 1976.

[4] Capurro, R., "Foundations of Information Science, Review and Perspectives", *International Conference on Conceptions of Library and Information Science, University of Tampere, Tampere, Finland, 26-28 August 1991*. Also available at www.capurro.de/tampere91.htm

[5] Capurro, R., "On the genealogy of information", *International Conference: Information. New Questions to a Multidisciplinary Concept, organized by the Chair for Philosophy of Technology at the Technical University of Cottbus held from March 1st to 3rd, 1994*. Berlin: Akademic Verlag Berlin, pp. 259-270, 1996, also available at www.capurro.de/cottinf.htm

[6] Christensen, Knud B.; Christensen, Mads G.; Boldt Jesper B.; Gran, Fredrik, "Experimental study of generalized subspace filters for the cocktail party situation", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 20-25 March 2016.

[7] Dendrinos, M., "Obstructive Sleep Apnea (OSA) Detection through Energy and Spectral Measures", International Conference on Signal Processing Applications and Technology (ICSPAT-98), IEEE, Toronto, Sept. 1998.

[8] Dendrinos, M.; Bakamidis, S. and Carayannis G., "Speech Enhancement from Noise : A Regenerative Approach", Speech Communication, Vol.10, No.1, pp.45-57, Feb.1991 http://www.sciencedirect.com/science/article/pii/016763939190027Q

[9] Deppisch, Thomas; Amengual, Sebastià V.; Calamia, Garí Paul; Ahrens, Jens "Direct and Residual Subspace Decomposition of Spatial Room Impulse Responses", *IEEE/ACM Transactions on Audio, Speech and Language Processing,* Volume 31, pp. 927–942, 2023.

[10] Fisher, R.A. "The use of multiple measurements in taxonomic problems", *Annual Eugenics,* 7, Part II, 179-188, 1936; also in *Contributions to Mathematical Statistics*, John Wiley, NY, 1950.

[11] Foley, James D. and Van Dam, Andries, *Fundamentals of Interactive Computer Graphics*, Addison-Wesley Publishing Company, 1982.

[12] Rao, S. S. and Gnanapmkasam, D. C., "A criterion for identifying dominant siịgu1ar values in the SVD based method of harmonic retrieval," in *Proc. iCJ5SP 88* (New York, NY), E.6.5, pp. 2460- 2463, 1988.

[13] Stainhaouer, Gregory; Bakamidis, Stelios; Dologlou, Ioannis, "Automatic Detection of Allergic Rhinitis in Patients", *International Conference on Computational Science and Computational Intelligence (CSCI),* 5-7 Dec. 2019.

[14] Tong, Renjie; Ye, Zhongfu, "Data fusion over localized sensor networks for parallel waveform enhancement based on 3-D tensor representations", Signal Processing, Volume 141, pp. 249-260, December 2017.

[15] Wang, Fasong; Wang, Zhongyong; Li, Rui; Zhang, Linrang "An efficient algorithm for harmonic retrieval by combining blind source separation with wavelet packet decomposition", Digital Signal Processing, Volume 46, pp. 133-150, November 2015.

Journal of Integrated Information Management electronic journal

A Literature Review on Research Indexes

Konstantina Christopoulou[1], Evangelia Triperina[1], Angeliki Antoniou[1], Manolis Wallace[2], Dimitrios Kouis[1]

[1]University of West Attica Department of Archival, Library and Information Studies

[2]University of Peloponnese Department of Informatics and Telecommunications

kchristopoulou@uniwa.gr [ORCID: 0000-0003-4164-2993], evatrip@uniwa.gr [ORCID: 0000-0003-4282-2259], angelant@uniwa.gr [ORCID: 0000-0002-3452-1168], wallace@uop.gr [ORCID: 0000-0002-4629-5946], dkouis@uniwa.gr [ORCID: 0000-0002-5948-9766]

Article Info

Article history:

Received 01 March 2024

Received in revised form 30 May 2024

Accepted 15 July 2024

http://dx.doi.org/10.26265/jiim.v9i2.37724

Abstract:

Purpose – *In this article, we have conducted a literature review (LR) on citation indexes to evaluate their acceptance and usage within the scientific community and the tools and metrics most frequently employed.*

Design/methodology/approach – *The presented LR followed the PRISMA framework and the methodology described by Kitchenham. Based on a set of research questions, several queries were made on the most prominent citation databases to retrieve the respective publications.*

Findings – *According to the outcomes of our LR, researchers utilised all three research databases, showing a preference for Scopus.*

Originality/value – *The paper presents a literature review of the publications related to research databases to gain insights about the current state of searching, retrieval, evaluation, and exploitation of the research publications and the related information by academics.*

Index Terms —citation resource comparison, Google Scholar, citation database, Scopus, Web of Science.

I. INTRODUCTION

Citation indexes or databases provide a reliable source of information for academics and researchers. A large number of citation indexes covers various disciplines **[1]**, while there are also many discipline-focused indexes **[2].** Nonetheless, there are several more that are widely used among the research community. The consistent cataloguing and presentation of the publications and the respective data, as well as the related metrics that accompany them, either in citation indexes as merely a reference or in databases with the full-text publication available, have contributed to the field of Bibliometrics. BBibliometrics corresponds to a set of methods for quantitative analysis of academic outputs and scholarly communications **[3]**, which can be used for books, websites, monographs, conference proceedings, policy statements, and even patents **[4]** and is mainly utilised to find the impact of research publications. At the same time, informetrics has been defined as the discipline that studies information through a quantitative perspective **[5]** by producing, disseminating, and using all forms of information, paying no attention to its formation or origination **[3]**. Likewise, the quantitative study of the field of science **[6]** is referred to as Scientometrics, and it deals with the impact of science on a greater scale. The measurement of research activity and collaboration and its depiction in research metrics facilitate the evaluation of the quality of research. Research activity is captured with metrics such as citation count **[3]**, H-index **[7]**, impact factor, and i10-index **[8]**. In contrast, research collaboration is measured by metrics, such as collaborative index, etc., and presented with co-authorship networks and graphs. The examination of the abovementioned data is of vital importance **[1]**.

Our motivation is to contribute to a more efficient management and assessment of research publications. We are conducting a literature review on the use and acceptance of the citation databases, the involved metrics, and the tools used by the research community to understand the current state of the research publications environment.

In this literature review, we have focused our search on three of the most prominent research indexes: Scopus, Google Scholar, and Web of Science. Scopus was released in 2004 **[9]** and constitutes a citation database, providing some of its services for free. Nevertheless, full access to its content is available only through subscription. The search for scientific publications through Scopus is an effortless task, while it also offers many tools and allows for personalised services during information retrieval.

Google Scholar was launched in 2004 **[10]**, and it is available for free; therefore, it has an ever-increasing popularity internationally. It returns a larger amount of results than the other citation indexes, mainly due to its extended volume of content being indexed (see information provided below). Google Scholar provides the user with a much-simplified experience, with fewer advanced search options in comparison to its counterparts. More particularly, the interface of Google Scholar has an advanced search option allowing the user to choose the words or phrases that will be included or excluded, in either the title or in both the title and the content of the article. It also allows filtering the results according to the authors or the Publisher, or the date range in which it was published.

Web of Science is one of the most popular citation databases. It was established in 1993 **[11]** and provides a set of filters and many criteria for content retrieval. The records returned from Web of Science are detailed and of high quality. Also, users can store/export the records in multiple formats.

As mentioned above Google Scholar has the largest number of content items indexed compared to the other two citation databases. Even though it is difficult to assess the magnitude of citation indexes due to their constant growth and increasing popularity, it is estimated that Google Scholar had 389 million records in 2018 and as a direct consequence, it is the most comprehensive academic search engine, while the study was conducted **[12].** More specifically, Google Scholar provided access to 389.000.000 records, Scopus to 72.212.354, whereas Web of Science to 105.519.854 **[12]**. Since then, more than 2.4 billion cited references are available through Scopus (December 2023), providing a time span from 1970 until now **[13]**. However, Clarivate and Google seem to have no up-to-date data regarding the statistics of the available records announced on their websites.

As far as it concerns the methodology followed in this research, the search and the aggregation of the related publications from two out of the three aforementioned citation databases, namely Scopus and Google Scholar, can be facilitated by tools such as the Publish or Perish tool **[14]**, which is freely available. In any case, the search and the retrieval from Scopus and Web of Science, through any tool, let alone Publish or Perish requires a subscription. This particular tool was used because of its reliability, its popularity amongst the academic community, and its openness.

For the screening process of the search results, we followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) reporting guideline **[15]** and the methodology described by Kitchenham **[16]**. The PRISMA guideline is an evidence-based methodology **[17]**, designed to assist the authors in improving the reporting of systematic reviews, enhancing the transparency of their research methods **[18]**. Kitchenham's systematic review activities are organised into three phases, i) the planning (identification of the need for a review and development of the review protocol), ii) the conducting (identification of research, primary studies selection, study quality assessment, data extraction and monitoring, and data synthesis) and iii) the reporting of the review. Several stages of each phase are not necessarily sequential but can be intertwined **[16]**. The publications that are reviewed are the outcome of a set of research questions that be thoroughly described in the Methodology of the paper.

The remaining of the paper is structured as follows: In Section II, the related literature is presented, whereas in Section III, the methodology of our LR on citation databases is thoroughly described. Section IV analyses the results and discussion of the LR, while the conclusions and future work are outlined in Section V.

II. Methodology

The literature review was conducted on three databases Web of Science, Scopus, and Google Scholar. The time frame selected for this review is between 2004 to 2021, following the procedures for conducting a systematic literature review described by Kitchenham (2004) **[16]**, as well as the PRISMA framework **[15]**. This literature review aims to find papers referring to citation databases and their comparison. Due to their interdisciplinary nature, we have focused our research on the most widespread and comprehensive citation databases in the academic community, which correspond to Scopus, Web of Science, and Google Scholar. The review follows the procedures proposed by Moher **[15]**, and more specifically, the process corresponds to the identification of the records, the screening, as well as the assessment of their eligibility, followed by the evaluation of the records according to the inclusion/exclusion criteria, resulting in the research corpus.

A. Planning - Research protocol - Research strategy

In the context of the object of the literature review, the searches carried out in each database were a combination of the keywords "Scopus", "Web of Science", and "Google Scholar". For each database, three queries were carried out (Table 1).

Table 1 – Research keywords per research database

QUERY ID	KEYWORDS
QUERY 1	"Scopus" AND "Google Scholar"
QUERY 2	"Google Scholar" AND "Web of Science"
QUERY 3	"Scopus" AND "Web of Science"

Based on the research questions (Table 2), the inclusion and exclusion criteria were determined to ensure that the review was inclusive and thorough and to eliminate any results that did not satisfy our specific research requirements.
The inclusion criteria were:

- Papers published between 2004 and 2021.
- Papers published in English.
- Papers focused on the comparison of citation databases.

The exclusion criteria were the following:

- Master Dissertations
- Presentations

Table 2 – Research questions

ID	Research Question (RQ)
RQ1a:	How many articles provide a comparison between all three citation databases? (Scopus, Web of Science, and Google Scholar)
RQ1b:	How many articles compare two out of three citation indexes?
RQ1c:	What is the percentage of those articles that compare Scopus with other citation indexes?
RQ1d:	What is the percentage of those articles that compare Web of Science with other citation indexes?

RQ1e: What is the percentage of those articles that compare Google Scholar with other citation indexes?

RQ2a: How many articles indicate a preference for Scopus?

RQ2b: How many articles indicate a preference for Web of Science?

RQ2c: How many of those articles indicate a preference for Google Scholar?

RQ3: Which bibliometrics were used most commonly?

RQ4: Which tools helped the authors while carrying out the research for their articles?

B. Data extraction

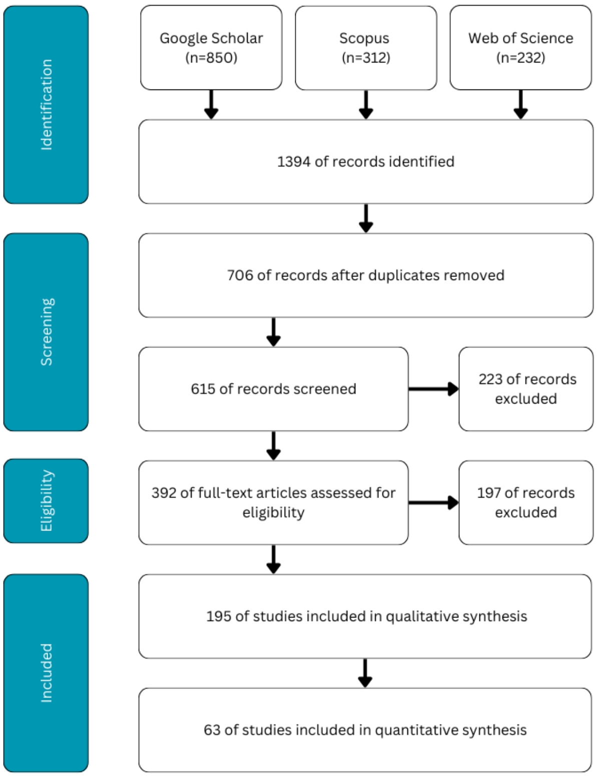

Identification
Google Scholar (n=850)
Scopus (n=312)
Web of Science (n=232)
1394 of records identified
Screening
706 of records after duplicates removed
615 of records screened
223 of records excluded
Eligibility
392 of full-text articles assessed for eligibility
197 of records excluded
Included
195 of studies included in qualitative synthesis
63 of studies included in quantitative synthesis

Figure 1 – Selection process of research publications based on the PRISMA framework

Table 3 - Results per keyword search in the field "paper title" per research database

Search ID	Keywords	Nº of results
S1	"Scopus" & "Google Scholar"	62
S2	Web of Science" & "Google Scholar"	57
S3	"Web Of Science" & "Scopus"	193
GS1	"Scopus" & "Google Scholar"	144
GS2	"Google Scholar" & "Web of Science"	122
GS3	"Scopus" & "Web of Science"	584
WoS1	"Scopus" & "Google Scholar"	42
WoS2	"Google Scholar" & "Web of Science"	45
WoS3	"Scopus" & "Web of Science"	145
	Total papers	1394

As mentioned before, the search and the accumulation of the research corpus were performed via Publish or Perish software, for Google Scholar and Scopus, and directly from Web of Science, for research publications spanning from 2004 to 2021. The results per keyword combination in each citation index are presented in Table 3. For the sake of brevity, they will be mentioned as S1, S2, and S3 when referred to Scopus, G1, G2, and G3, when referred to Google Scholar and W1, W2, and W3, respectively, for Web of Science. The queries were performed using the field "paper title".

The results were exported in Excel format for better management and exploration. During the initial search, 1394 papers were retrieved from the three citation indexes mentioned above (850 from Google Scholar, 312 from Scopus and 232 from Web of Science). After applying the deduplication process, we were left with 706 papers. After the screening of the remaining publications for corrupted, non-accessible papers and publications written in other languages, we narrowed them to 392. In the eligibility phase, the papers were reviewed, and based on their relevance to our research queries in terms of both abstract and full text, we retained 195 papers, which were included in the qualitative synthesis (see Figure 1). During the final phase, the remaining publications were examined in relevance to the research questions, and 132 records were excluded, resulting in 63 research publications for analysis.

The publications within the defined time range were divided chronologically into three categories, as shown in Table 3. The first category includes articles from 2004 to 2009, with at least 50 citations. The second category comprises articles published between 2010 and 2015, with at least 10 citations. The third category corresponds to the articles published between 2016 to 2021. In the last category, all the retrieved publications were considered irrespective of their citations.

Table 4 – Categorisation of the selected publications

Period	Citations	Publications
2004-2009	≥50	8
2010-2015	≥10	6
2016-2021	-	49

As shown in Figure 2, most of the publications have been published in journals, whereas several papers were published in conference proceedings. There were fewer publications in other forms, such as chapters in books, reports, preprints, and PhD dissertations. Furthermore, it is evident that from 2016 to 2021, there was an increase in publications concerning citation databases.

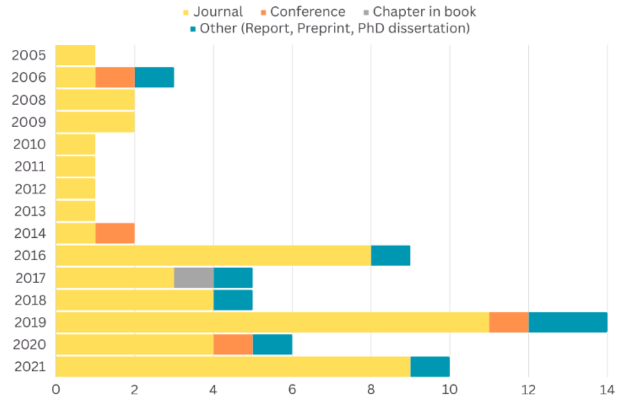


Figure 2 – Distribution of the papers included in the quantitative synthesis

III. Results and Discussion

One of the most important outcomes of our literature review is an overview of the papers that concern citation database research. In the following section, we answer each research question according to the information that was evident in the research corpus.

RQ1a How many articles provide a comparison between all three citation databases?

Most of the research publications focused their comparison on two citation index databases (see below). Nevertheless, 34.92% of the papers (22 out of 63) included all three citation indexes in their comparisons (see Figure 3).

RQ1b How many articles compare two out of three reference indexes?

In addition, 41 out of the 63 papers that were studied performed a comparison between two citation indexes, which is 65.08% of the total papers (see Figure 3).

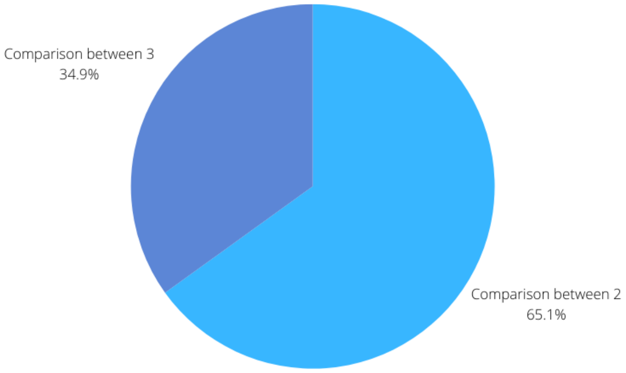


Figure 3 - Comparison between research databases in the research corpus

RQ1c How many articles were compared to Scopus with other citation indexes?

Scopus was the most compared citation index, with 60 articles including it in their study. This is estimated at 95.24% of the total papers studied, which depicts the reception of this particular research database from the research community.

RQ1d How many articles compared Web of Science with other citation indexes?

Web of Science is the second most frequently compared citation index, with a percentage of 93.65%, which means that 59 papers compared WoS to one or more citation indexes.

RQ1e How many articles compared Google Scholar with other citation indexes?

The citation index that was the least compared to the others was Google Scholar, with 29 papers including it in their comparisons, giving a 46.03%, which was expected, given the fact that it is newer compared to the other two.

RQ2a Which citation index is the most preferred by the scientific community?

Scopus seems to be the citation index that received the most positive feedback, with 20 papers (31.75%) indicating their preference, and stating its positive features and the quality of the search results in their publications.

RQ2b How many articles indicate a preference for Web of Science?

and

RQ2c How many of those articles indicate a preference for Google Scholar?

Researchers of 17 publications expressed their preference for the Web of Science, scoring the same percentage with Google Scholar (26.98%).

Meanwhile, it is worth mentioning that 29 papers do not mention any preference, with many of them advising the researchers to utilise as many citation indexes as possible to get the best results.

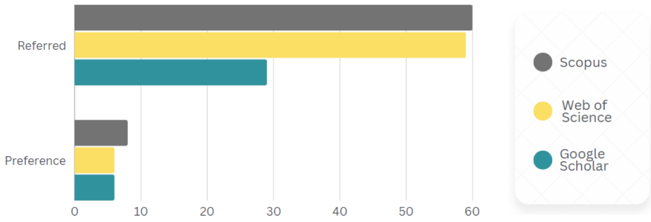


Figure 4 – An overview of the total publications that included a research database in their comparison and the respective indication of preference between the alternatives

RQ3 Which bibliometrics were used most commonly?

The most popular metric that was taken into consideration in the papers that were analysed, was the h-index, with 10 records mentioning it in their content. H-index, which was proposed by J.E. Hirsch in 2005 [7], is considered to be reliable for the qualitative evaluation internationally. Other bibliometric indicators that have been referred by the papers are the Mentor-Index, the SCIE and SSCI Indexes, H-classics, the Relative Citation Ratio (RCR) and the JIF and SJR indicators.

RQ4 Which tools helped the authors while carrying out the research for their articles?

A variety of tools were mentioned in the research corpus. Among them, excel and SPSS were the most used software as far as it concerns the statistical analysis. Publish or Perish and Classic Papers by Google were also popular tools regarding the collection, the organisation, and the study of publications. Other tools discussed in the publications were CiteSearch by the ACM Digital Library, HistCite by Clarivate, VOSviewer (a software tool for constructing and visualising bibliometric networks), the Sapiro-wilk test of data set normality, and several APIs.

IV. Conclusions and Future Work

The LR was based on the PRISMA framework, using a set of research questions for all the citation indexes that were taken under consideration (Scopus, Web of Science, and Google Scholar). After having analysed the research findings, we concluded that there is no clear indication as to which is the best in its field. This is understandable, as each citation index has its unique features, filtering methods, and search criteria. For this reason, the majority of the researchers employ a variety of citation indexes during their research, to benefit from their unique features. However, we should keep in mind that these citation indexes were created in different years, affecting their credibility, their acceptance, as well as their usage over the past years. Another consideration to bear in mind is that all citation indexes are constantly enriching their content and improving the search tools that they provide.

Furthermore, we discovered that different languages affected our search outcomes more than expected. Articles written in different languages appeared in the search results, even when in some cases we have specifically searched for articles written only in English. In the before mentioned cases, the abstracts were written in English. Consequently, that indicates a weakness into categorising the content properly when it comes to language preference. Future work lies in the study of the influence of open science in conducting, capturing, and disseminating research.

REFERENCES

[1] Bar-Ilan, J., Levene, M., & Lin, A. (2007). Some measures for comparing citation databases. *Journal of Informetrics, 1*(1), 26-34. https://doi.org/10.1016/j.joi.2006.08.001.

[2] Torres-Salinas, D., Robinson-Garcia, N., Miguel Campanario, J., & Delgado Lopez-Cozar, E. (2014). Coverage, field specialisation and the impact of scientific publishers indexed in the Book Citation Index. *Online Information Review, 38*(1), 24-42, https://doi.org/10.1108/OIR-10-2012-0169.

[3] Das, A. K. (2015). *Research evaluation metrics* (Vol. 4). UNESCO Publishing, ISBN 978-92-3-100082-9.

[4] Cooper, I. D. (2015). Bibliometrics basics. Journal of the Medical Library Association: JMLA, 103(4), https://doi.org/217. 10.3163/1536-5050.103.4.013.

[5] Qiu, J. P. (2007). Informetrics. Hubei: Wuhan Publisher. https://doi.org/10.1007/978-981-10-4032-0.

[6] Hess, D. J. (1997). Science studies: An advanced introduction. NYU press. ISBN 9780814790953.

[7] Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences, 102*(46), https://doi.org/10.1073/pnas.0507655102.

[8] Noruzi, A. (2016). Impact Factor, h-index, i10-index and i20-index of Webology. *Webology, 13*(1), 1-4, Available at: http://www.webology.org/2016/v13n1/editorial21.pdf.

[9] Deis, L., & Goodman, D. (2005). Web of Science (2004 version) and Scopus. The Charleston Advisor, 6(3). Retrieved from http://www.charlestonco.com/comp.cfm?id=43.

[10] Mayr, P., & Walter, A. K. (2007). An exploratory study of Google Scholar. *Online information review, 31*(6), 814-830. https://doi.org/10.1108/14684520710841784.

[11] Qiu, J., & Lv, H. (2014). An overview of knowledge management research viewed through the web of science (1993-2012). Aslib Journal of Information Management, 66(4), 424-442. https://doi.org/10.1108/AJIM-12-2013-0133.

[12] Gusenbauer, M. Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. Scientometrics 118, 177–214 (2019). https://doi.org/10.1007/s11192-018-2958-5

[13] Elsevier. (n.d.). www.elsevier.com. https://www.elsevier.com/products/scopus/content

[14] Harzing, A.W. (2007) Publish or Perish, available from https://harzing.com/resources/publish-or-perish

[15] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, and PRISMA Group*, "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement", Annals of internal medicine, vol. 151(4), pp.264-269, 2009. https://doi.org/10.7326/0003-4819-151-4-200908180-00135.

[16] B. Kitchenham, "Procedures for performing systematic reviews", Keele, UK, Keele University, vol. 33, pp. 1-26, 2004. ISSN:1353-7776.

[17] PRISMA statement. (n.d.). PRISMA Statement. https://www.prisma-statement.org/

[18] Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & McKenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. bmj, 372, https://doi.org/10.1136/bmj.n160.

Konstantina Christopoulou is a Ph.D. researcher and a member of the Information Management Research Lab at the Department of Archival, Library, and Information Studies of the University of West Attica. She graduated from the Department of Informatics and Telecommunications of the University of Peloponnese, earning her BSc diploma in Informatics and Telecommunications, majoring in Informatics. Furthermore, she continued her studies in the same Department, earning her MSc diploma in Computer Science, with merit. She has also gained her Pedagogy and Teaching Competence (PDE) certification from the Hellenic Open University and is currently working as a High School Computer Science Teacher.

Evangelia Triperina holds a PhD in Computer Science from the University of Limoges (France), with a thesis entitled "Visual interactive knowledge management for multicriteria decision making and ranking in linked open data environments". She is a Department of Computer Engineering graduate of TEI of Athens. She holds an MSc in Information Technology, Image Synthesis and Computer Graphics from the University of Limoges (France). She has worked in European research projects at GRNET, Agro-Know Technologies and the University of West Attica. She is currently a PostDoc Researcher at the the Department of Archival, Library, and Information Studies of University of West Attica.

Angeliki Antoniou is an Assistant Professor in the Department of Archival, Library & Information Studies at the University of West Attica, a collaborating researcher at the ATHENS Research Center and at the Department of Information Studies at the University College London. She holds a degree in Preschool Education (University of Athen), a degree (BSc) in Clinical with Social Psychology (University of Kent at Canterbury, UK), a postgraduate degree (MSc) in Human Interaction with Ergonomics (University College London, UK), a music degree, a Piano Diploma, and a PhD in Educational Technologies (University of Peloponnese, Department of Computer Science and Technology).

Dr. Manolis Wallace [http://gav.uop.gr/wallace/], Associate Professor at the Department of Informatics and Telecommunications and Director of the Knowledge and Uncertainty Research Laboratory (ΓAB LAB) [http://gav.uop.gr/], has a long experience in the management of educational and research organisations from his previous tenures at the University of Indianapolis and at the Foundation of the Hellenic World. He now transfers this experience to Tripolis and the University of Peloponnese where he leads an interdisciplinary group of researchers specialising on cultural and educational informatics.

Dimitrios Kouis received his Diploma in Computer Engineering and Informatics from the University of Patras and his PhD from National Technical University of Athens (NTUA) in 1994 and 2004 respectively. His scientific interests include Library Networks, Digital Publishing, Scholarly Communication topics, Software development, Content Management, IT middleware platforms, meta-data modelling etc. He has been involved in several European and national projects and has published more than 30 articles in journals and conferences. Currently, he is an assistant professor at the Department of Archival, Library and Information Studies, University of West Attica.

Journal of Integrated Information Management electronic journal

Phygital Heritage Experiences in Refugee Attica

Angeliki Antoniou, Despoina Lampada, Afroditi Kamara, Daphne Kyriaki-Manessi
Department of Archival, Library & Information Studies, University of West Attica, Aegaleo, Greece, ICI Paper-Social Enterprise, Cholargos, Greece, Time Heritage, Papagos, Greece, Department of Archival, Library & Information Studies, University of West Attica, Aegaleo, Greece
angelant@uniwa.gr [ORCID: 0000-0002-3452-1168], info@ici-paper.com, info@timeheritage.gr [ORCID: 0000-0002-6380-6402], dkmanessi@uniwa.gr [ORCID: 0000-0002-3310-6616]

Article Info

Article history:
Received 29 July 2024
Received in revised form 30 September 2024
Accepted 10 October 2024

http://dx.doi.org/10.26265/jiim.v9i2.38540

Abstract:

The Greco-Turkish war's aftermath led to a significant refugee crisis in 1922, with over 1.5 million fleeing Asia Minor for safety in Greece. The Digistoryteller project aims to document and share narratives of these refugees' struggles to establish homes in Attica, using digital storytelling and crowdsourcing features. This project, through its database and mobile apps, allows for city exploration and contributions from experts and the public. A key focus of the project is the concept of "phygitality," which combines physical and digital experiences. Phygitality encompasses various combinations, including augmented reality, 3D printing, and holograms. In cultural heritage, phygitality offers new ways to engage with historical sites and enhance cultural experiences. Different phygital products developed within the framework of the project will be presented. In the case of Vyronas, a municipality in Attica founded as the first urban refugee settlement, the project introduces phygital objects like paper reconstructions of historical buildings. These objects, like the Old Town Hall, provide educational and touristic value by allowing users to assemble them and access augmented reality information about the building's history. Initial user testing has shown promising results, with plans to integrate these objects into educational programs and museum shops. The municipality of Vyronas intends to produce these objects for both educational and touristic purposes.

Index Terms — This is where the keywords should be placed (up to six terms). **Cultural Heritage, Phygitality, Augmented Reality**

I. THE DIGISTORYTELLER PROJECT

The Greco-Turkish war in Asia Minor came to an end in 1922. Following the conclusion of that conflict, a massive refugee crisis forced over 1.500.000 people to flee their homes in Asia Minor and Thrace and seek safety in Greece. A large number of these individuals settled in the prefecture of Attica. Over 100 years later, the goal of the Digistoryteller project (https://digistoryteller.eu), which is devoted to the refugee crisis, is to share narratives about the difficulties and attempts of refugees to establish homes in the Attica region. The project created a rich database and mobile apps to support city exploration with the use of digital storytelling and crowdsourcing features which allow contributions from experts and the public.

More specifically, the Digistoryteller includes a repository of information related to the arrival, settlement, and gradual integration of Asia Minor refugees in Attica. The reference period is from 1914 to 1949, with a focus on the period 1922-1928. Despite the various commemorations of the Asia Minor Catastrophe, primarily by Asia Minor organizations, the emphasis usually lies on the catastrophe itself and the development of municipalities with predominantly refugee populations, a phenomenon occurring from 1934 onwards. Research into the initial phases of these people's settlement, their continuous relocations until finding suitable living and working conditions, their entrepreneurial activities, the lack of access to basic goods, and the support networks and social organizations they created, has only recently begun. Many aspects regarding urban planning, its political ramifications, relations with the indigenous population, and the process of economic integration are illuminated through archival research and the development of new inquiries.

The guiding function is developed on two levels:

In guided tours using documents collected, selected, and entered by the project's research team, forming a route within each municipality, supplementing historical information for each point of interest with multimedia material (photos, music, recorded testimonials, video interviews, and recipes).

In the participatory crowdsourcing system, where the public (residents, inhabitants, visitors) can upload their own documents and information from their family archives or even simple testimonials from grandparents, grandmothers, or parents, allowing previously unknown material to emerge.

In addition, a large part of the project focused on the creation of relevant phygital products which would be used to provide reconstructions of monuments and elements of cultural heritage involving users in unique ways and allowing them to explore the past. Such products included (indicatively):

a) A coloring book and an adjoint series of 6 coloring cards, based on authentic photos which were provided by the Asia Minor Society of Egaleo “Nees Kydonies” (Figure 1);

the original photos are accessible through the AR application, which works both on the black-and-white and colored-in pictures.

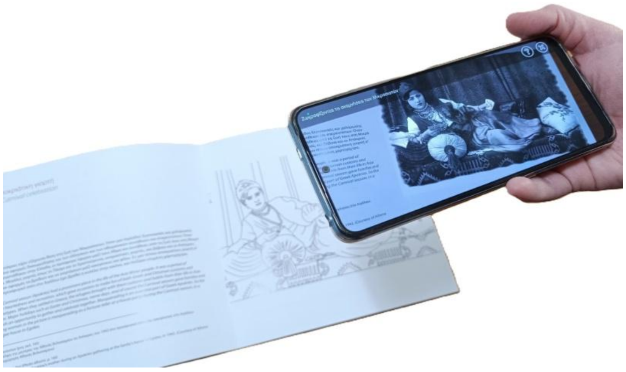


Fig. 1. The coloring book with the painting card series and the AR application.

b) A narrative card game which comes together with an original audiobook, titled A Day in Kastraki: Stories from the Refugee Settlement of Drapetsona, 1922-1960. The game is based on the power of narration and storytelling, and can be played either in conjunction with the story of the audiobook or autonomously. The AR application accompanying the game activates the camera of the mobile device and scans each of the cards to provide access to a different excerpt of the audiobook each time, allowing for various playful and educational applications.

c) An activity book for use by children aged 9-14, based on the approach of "mind maps", to help children approach aspects of Asia Minor memory and cultural heritage as shaped through the refugee experience. The organization of the content and the structure of the book invite each reader to complement it with their own multimedia material (photos, audio recordings, videos), mapping their family and local history in their own way. To facilitate the recording and organization of user-generated multimedia material, the accompanying application has been designed, which activates the recording applications of the mobile device on which the application is installed and allows for the management of files in accordance with the organization of the book's sections-collections (Figure 2).

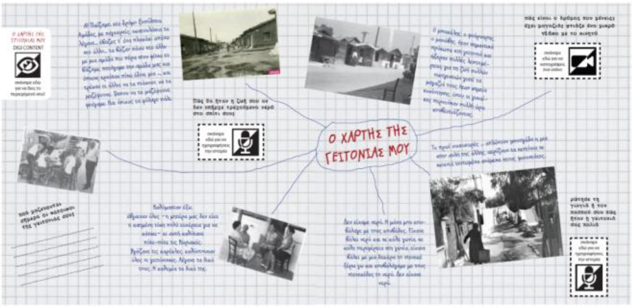


Fig. 2. The activity book with the augmented mind map.

II. Phygitality

The term phygitality was introduced to describe new forms of hybrid experiences which combine physical aspects of the experience with augmented reality, 3D printing, holograms, etc. This is an umbrella term which generically describes any such combinations without being defined in concrete ways. Thus, phygitality refers to an ecosystem of various combinations of physical and digital experiences [1]. A recent literature review revealed four main dimensions of phygitality: phygital objects and applications, phygital spaces, phygital user journey (user decisions making is assisted through the coexistence of the physical and virtual), phygital experience (the resulted enhanced cultural experience) [2].

Literature review also identifies five main areas that phygitality emerges: 1. Marketing, 2. Education, 3. Social issues and Politics, 4. Technical and Legal issues, and 5. Cultural Heritage, Tourism and Urban Development [3]. Regarding cultural heritage, there are already research efforts that wish to introduce phygitality in the domain of heritage and first attempts provided positive results [4]. Although many efforts focus on new forms of cultural tourism [5, 6], others focus more on the cultural experience as such [7]. For example, phygitality was considered as a new way to experience one of Portugal most known cultural sites, Quinta da Regaleira [8]. In addition, museum visitors that received phygital experiences in the form of Mixed Reality showed higher perceptions of authenticity and reported higher quality of the cultural experience [9]. Finally, phygitality was used to promote a cultural site and enhance heritage communication [10, 11].

The current work uses phygital objects and phygital experiences to engage users in cultural experiences. As mentioned in the previous section, in the context of Digistoryteller, different phygital products were developed and tested with visitors and policy makers. In the following section, we will focus on the phygital 3D papercraft of the old Town Hall of Vyronas, which represents an effort to implement the concept of phygital and game-based storytelling experiences for a built landmark, rather than for aspects of intangible refugee heritage.

III. The Case of the Old Town Hall of Vyronas

A. *Focusing on the physical aspect of "phygital"*

During the last decades, there has been a notable expansion of the use of immersive digital technologies, such as Virtual and Augmented Reality, in order to make learning about cultural assets more attractive and accessible to wider audiences. The combination with game-based approaches, such as Serious Games, has proved very effective in various areas of education related to cultural heritage. However, most examples of such experimentation focus on digitized assets within a Virtual Reality environment, maintaining that in this way the contact with the assets appears to be more direct and realistic [12].

Our approach came from a different point of departure: that playful experience of built cultural assets and landmarks can be served by focusing on the physicality of the educational object. Tangible interactions with educational games, especially with games incorporating analysis and documentation, are essential for turning cultural assets, including historical buildings, into educational resources for expanding the learning skills and perspectives, with a special focus on school-age audiences [13]. In this, we have drawn from the experience from previous work of the project's partners with paper reconstructions of archaeological and architectural monuments, which have proven very popular with Museum visitors in all major archaeological Museums in Greece.

Moreover, the cross-curricular value of papercraft projects in educational contexts is significant: for one, they are a playful way of understanding how geometric shapes can be handled to form 3D structures, thus familiarizing users with basic engineering and spatial relationship understandings. At the same time, papercrafts entail building from templates, concentration and fine motor skills, which obliges users to actively engage with the papercraft subject and can thus build positive associations with it, sparking curiosity for further exploration. Digitally enhanced storytelling can therefore serve as an essential part of an extended and enriched experience with the cultural asset addressed. Based on this premise, we decided to test this approach for at least one historical building related to the refugee experience in Attica.

B. *Focusing on the Old Town Hall of Vyronas*

The choice of the specific historical building was due to a number of reasons:

a) Vyronas is one of the most known refugee neighborhoods in Attica, and its old Town Hall is an emblematic landmark related to the refugee settlement: originally it was the Red Cross Polyclinic (1924) and for many decades to follow the Town Hall of Vyronas (until 1996). It has since served as municipal cultural center for almost 20 years, and it is now going to host the newly established and soon-to-be-implemented History Museum of the Municipality of Vyronas. Additionally, the Municipality was looking for creative synergies with Digistoryteller, in order to promote its own activities promoting local history and refugee heritage among its citizens. All these factors made the old Town Hall building a suitable choice within the context of Digistoryteller.

b) Contrary to other historical buildings with similar advantages, in the case of the old Town Hall of Vyronas we had access to the building's detailed architectural plans, due to the recent museography and restoration blueprints for the transformation of the building into a city-museum. This part is essential for the 3D papercraft to be closer to a paper reconstruction, however simplified.

c) The plans of the Municipality to turn the building into a History Museum permitted us to conceptualize the product as part of the Municipality's awareness and expectation raising campaign. Also, we could draw from the narrative developed for the forthcoming Museum's exhibition, in order to build our content and storytelling around it. This ensured more focused testing sessions for our product with residents and Municipality employees.

C. *The phygital products for the new History Museum*

The principal product is the 3D papercraft, which consists of color printed pre-cut pieces on 3 A4 sheets, plus a color base on an extra sheet. By assembling the pieces, the user can build the paper reconstruction of the old Town Hall and soon-to-be History Museum (Figure 3). The key points of the building's history are presented on the information sheet included in the packaging. Through the AR application, the user who has assembled the model can be guided through a more detailed version of the building's history, including some accounts and anecdotes about the centrality of this landmark for the everyday life of Vyronas residents through the decades. The user also gains access to the VR video about the forthcoming Museum and its first exhibition.

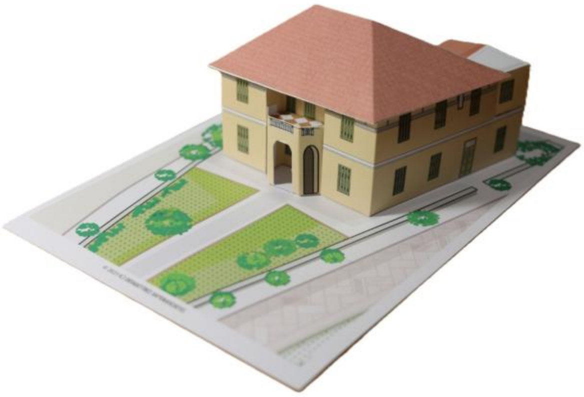


Fig. 3. Paper reconstruction of the Old Town Hall of Vyronas.

The application is available for free installation on a phone or tablet from the Google Play Store. A QR code on the cover (packaging) of the paper construction leads directly to the application, for easier installation. The application activates the camera of the portable device and scans the object. Each side of the building opens different text bubbles, through which the history of the building unfolds in brief. Reading the back side of the building opens the VR video created by the Municipality of Vyronas for the planned Museum, guiding the user to see the next page of the building's history.

Together with the reconstruction of the historical buildings, Vyronas' visitors will be also able to use an 8-fold informational brochure for the Vyronas Municipality History Museum, double-sided (Figure 4). The interior includes a map of the wider area of the Museum (Old Town Hall) in Vyronas. Through the same AR application (but by choosing a different menu), additional visual and auditory content is projected onto the map, for points of interest near the Museum building - the Old Town Hall, which are related to the establishment and inauguration of the refugee settlement in 1924. By selecting a POI, the user can see a brief description, listen to a more detailed audio description, and view additional multimedia material (photos and videos) about the landmark.

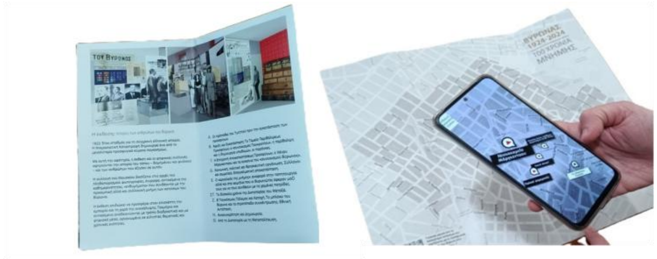


Fig. 4. Augmented brochure for with map of the wider area of the Old Town Hall.

IV. Perspectives and conclusions

The phygital objects of the Old Town Hall were presented to the Mayor of Vyronas, to municipal employees and locals in December 2023, and were given to them for user testing. Initial user testing revealed the educational potential of the 3D papercraft, as well as the commercial potential of both products, which have been picked up as saleables at the new Museum's shop. In fact, the Municipality already expressed its intention to proceed with the production of the item for use in the educational programs of the Museum, after the start of its operation, while it is also considering the possibility of production for use in the educational programs of its schools already before the opening of the Museum.

Specifically with regard to the 3D papercraft of the old Town Hall, it is noteworthy that it is addressed to younger audiences, who have no recollection of the Museum premises used as Town Hall. Therefore, it was considered a playful and creative means to familiarize younger audiences with the historicity of a building which they are only going to know as Museum, already remodelled inside and rebranded as a landmark. In this sense, the old Town Hall of Vyronas has provided a meaningful testbed for the use of digitally enhanced paper reconstructions as tools in a game-based approach to cultural and local heritage education, especially when it comes to raising awareness and re-inventing the relevance of historical buildings.

V. Acknowledgements

The Digistoryteller project was implemented within the framework of the call "RESEARCH AND INNOVATION SYNERGIES IN THE REGION OF ATTICA" in the Operational Program "Attica 2014-2020".

REFERENCES

[1] P.P. Klaus, “Phygital–the emperor’s new clothes?”, Journal of Strategic Marketing, vol.23, pp.1-8, 2021.

[2] C. Mele, T.R. Spena, M. Marzullo, and I. Di Bernardo, “The phygital transformation: a systematic review and a research agenda”, Italian Journal of Marketing, vol. 3, pp. 323-349, 2023.

[3] P. Del Vecchio, G. Secundo, and A. Garzoni, “Phygital technologies and environments for breakthrough innovation in customers' and citizens' journey. A critical literature review and future agenda”, Technological Forecasting and Social Change, vol. 189 (122342), 2023.

[4] E. Lupo, “Design and innovation for the Cultural Heritage. Phygital connections for a Heritage of proximity”, AGATHÓN| International Journal of Architecture, Art and Design, vol. 10, pp. 186-19, 2021.

[5] R. Baratta, A. Bonfanti, M.G. Cucci, and F. Simeoni, “Enhancing cultural tourism through the development of memorable experiences: the “Food Democracy Museum” as a phygital project”, Sinergie Italian Journal of Management, vol. 40, pp-1153-1176 , 2022.

[6] C.O. Moreira, R. Ferreira, and T. Santos, “Smart tourism and local heritage: Phygital experiences and the development of geotourism routes”. In: Handbook of Research on Cultural Heritage and Its Impact on Territory Innovation and Development, pp. 206-232. IGI Global ,2021.

[7] M.L. Turco, and E.C. Giovannini, “ Towards a phygital heritage approach for museum collection”, Journal of Archaeological Science: Reports, vol.34 (102639), 2020.

[8] J.G. Andrade, and P. Dias, “A phygital approach to cultural heritage: Augmented reality at regaleira”, Virtual Archaeology Review, vol. 11(22), pp. 15-25, 2020.

[9] S.F. Çiftçi, and B. Çizel, “Exploring relations among authentic tourism experience, experience quality, and tourist behaviours in phygital heritage with experimental design”, Journal of Destination Marketing & Management, vol. 31 (100848), 2024.

[10] K. Muangasame, and E. Tan, “ Phygital rural cultural heritage: a digitalisation approach for destination recovery and resilience. Worldwide hospitality and tourism themes”, vol.15(1), pp. 8-17, 2023.

[11] E. Nofal, M. Reffat, and A. Vande Moere, “Phygital heritage: An approach for heritage communication”. In: Immersive learning research network conference, pp. 220-229 ,2017.

[12] A. Roumana, A. Georgopoulos, and A. Koutsoudis, “Developing an Educational Cultural Heritage 3D Puzzle in a Virtual Reality Environment”, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 43(B2-2022), pp. 885-891, 2022.

[13] D.E.D. Heidtmann Junior, and A.C. Chiarello, “Cultural Heritage Mediators: A Creative Approach in Historical Cities of Southern Brazil”. In: Bandung Creative Movement 2015 - 2nd International Conference on Creative Industries “Strive to Improve Creativity”, pp. 17-24, 2015.

Journal of Integrated Information Management electronic journal

Graph Databases and Graph Neural Networks

Stratos Tsolakidis[1], Anastasios Tsolakidis[2], Evangelia Triperina[2], Nikitas N. Karanikolas[1], Christos Skourlas[1]

[1]University of West Attica Department of Informatics and Computer Engineering
[2]University of West Attica Department of Archival, Library and Information Studies
etsolakidis@uniwa.gr [ORCID: 0009-0000-0860-4579], atsolakid@uniwa.gr [ORCID: 0000-0001-7364-4542], evatrip@uniwa.gr [ORCID: 0000-0003-4282- 2259], nnk@uniwa.gr [ORCID: 0000-0003-1777-892X], cskourlas@uniwa.gr [ORCID: 0000-0003- 4464-5305]

Article Info

Article history:
Received 24 October 2024
Received in revised form 29 November 2024
Accepted 09 December 2024

http://dx.doi.org/10.26265/jiim.v9i2.39198

Abstract:

Purpose *- Nowadays, social networks, online media sharing and e-commerce platforms generate a vast amount of data, which, among other information, capture the interactions among the users. Storing, analyzing and exploiting the aforementioned information allow the exploration of hidden and unstructured patterns.*

Design/methodology/approach *- The associations among the users during their visit in a platform construct a graph network which capture all the generated data. Graph Neural Networks are applied in these data models, to make suggestions based on their topology. In the presented research, Graph Databases and Graph Neural Networks are utilized for data exploration and analysis in graph databases networks.*

Findings *- In this study, we compare the use of graph databases with relational databases for large-scale databases and we present that the use of graph neural networks over graph databases can be used efficiently to apply machine learning tasks for those datasets.*

Originality/value *- Thus, in this paper, we present the benefits of applying graph neural networks and graph databases for data analysis in large-scale data from social networks. Also, we examine to the efficiency of using graph databases over relational databases for analyzing those networks.*

Index Terms — Graph Data, Graph Database, Relational Database, Graph Neural Network, Social Recommender System.

I. INTRODUCTION

The Relational Database Management Systems (RDBMS) are the most popular technology for data management, which is widely used in academic research and industry. However, graph databases are designed to manipulate dataset which include interaction among the entities. Multiple studies show the advantages of using graph databases, due to their inherent property of handling big data, comparing to non-graph database when applied to connected data [1].

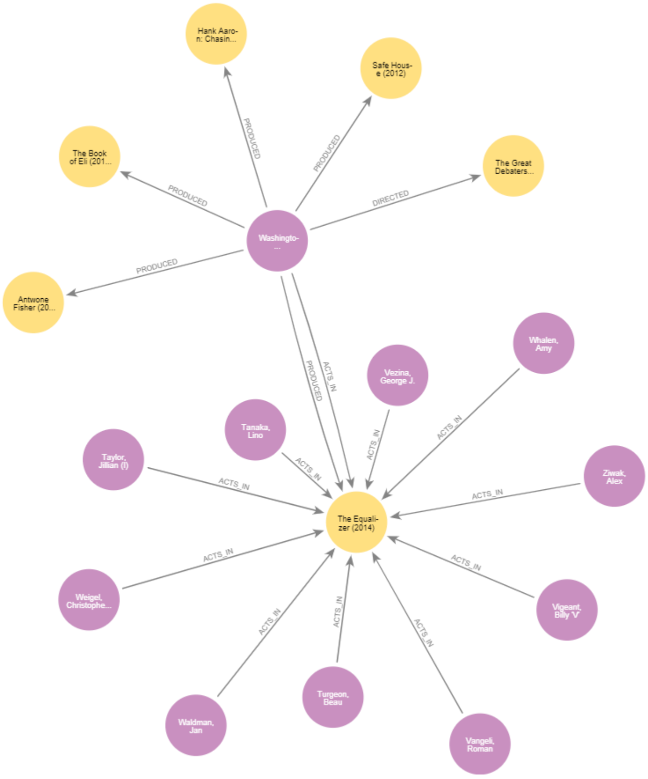


Figure 1 Graph database as a basis for a social network based on IMDB data

Graphs constitute a type of data structure that represents a set of objects or entities (nodes) and their relationships (edges). Their expressive power has attracted the attention of researchers in areas such as social networks. Figure 1 illustrates an example of graphs and graph database as a basis for a social network derived by IMDB data. In graph databases, nodes (circles) represent objects; every node has properties, which correspond to pairs of names and values. The directional relationships, represented by arrows, connect the nodes, and denote actions. They also have properties, which are in pairs of names and values, like mentioned before. Figure 1 shows that Denzel Washington produced and acted in the film “The Equalizer (2014)”. Both “Denzel Washington and “The Equalizer (2014)” are nodes. All the films where D. Washington acted are presented in the graph, as well as the other actors that also participated in the film “The Equalizer (2014)”.

There are two major research approaches focused on developing frameworks for graph data: the representation of graph data with Graph Databases and the application of machine learning methods on graph data with Graph Neural Networks (GNNs).

The technology of big data and the spread of its applications offer an opportunity to re-evaluate the technology of relational databases. GNNs are machine learning models designed for graph data that utilize the graph topology. Further investigation is needed to understand the perspectives of the two above approaches whereas a review of the two research approaches at a theoretical and a practical level is important. More specifically, the current study addresses the following research questions:

RQ1. Is there in the literature any satisfactory theoretical and experimental verification of the benefits of applying graph neural networks and graph databases to managing large-scale data in social networks?

RQ2. Are there any comparative advantages of using graph databases over relational databases in managing large-scale data in social networks?

The paper is structured as follows: the first section is the introduction, whereas the second section corresponds to the literature review. Section 3 outlines a comparative presentation of handling graph data and relational data by examining the data modeling, the queries and the representation of the data. The fourth section demonstrates how graph neural network can be applied in social recommendation and finally, in section 5, there is the conclusion.

II. LITERATURE REVIEW

Although Relational Database Management Systems (RDBMS) are considered the most established technique for data management, graph databases seem to attract the interest of researchers, because of their ability to manage big data. According to Kumar Kaliyar, R. (2015) [2], “most of the real-world applications can be modeled as a graph and one of the best real-world examples is social network”. According to Bhattacharyya & Chakravarty [3], the evolution of relational DBMS will be the Graph Databases with NoSql methodologies, “which is emerging as beyond of relational model”, while Tian [4] referred to the growth of big network data in industry that demands graph technologies, and denoted that the research projections showed a significant growth of the global market for graph databases in the subsequent years. Xirogiannopoulos & Deshpande [5] stated that the analysis of the graph structure among the underlying entities (or objects) in a dataset deliver meaningful information and value in various application domains and they mentioned the development of various graph databases, e.g., Neo4j, which address these needs.

Both Graph Databases and GNNs for graph data have been the focus of research and according to literature many benefits derive from their application. The benefits that may occur from their combination and from the proposal and the use of a unified framework are yet to be examined. According to the literature there is a research gap in the before mentioned research area. Among the efforts made to this direction, Besta et al. [6] combined GNN models with graph databases but , limited work has been done in this research area.

GNNs [7], [8] are deep learning-based methods that operate on graphs, in which the edges connecting the nodes express the underlying topology. GNNs can exploit this topology and combine it with features on the nodes, in order to provide predictions. According to Zhou et al., there are many variants of GNNs include graph convolutional network (GCN), graph attention network (GAT) and graph recurrent network (GRN). Wu et al. [9] provide an overview of GNNs, a new taxonomy, as well as a set of evaluation techniques for GNNs, including open-source codes and benchmark data sets. Xu et al. [10] highlight that there is limited understanding of the representational properties and the limitations of GNNs, describing an aggregation scheme, in which the representation vector of a node is calculated by accumulating recursively and transforming the representation vectors of the adjacent nodes. Eventually, they present a theoretical framework for analyzing the ability of GNNs to capture different graph structures.

Spectral approaches of GNNs have their origin in signal processing, while working with a spectral representation of graphs. A graph signal is initially transformed by the graph Fourier transform and then the convolution operation is carried out, leading to the resulting signal, which is transformed once again using the inverse graph Fourier transform. According to Zhu & Koniusz [11], despite the great significance of Graph Convolutional Networks (GCNs) for learning, there is also the need for special architectures. As a direct consequence, a Simple Spectral Graph Convolution (S2GC) was proposed, so as to achieve the target performance. In general, a GNN-based technique could be applied on data extracted from popular social networks applications. For example, Behún [12] conducts an overview and a comparison of such GNN-based techniques, which are used to address problems of learning on graph data from the Tripadvisor website.

Based on the literature, Graph Databases are considered the most suitable technique for graph data, because of their ability to manage large-scale data from social networks. In addition, GNN can effectively be used for predictions by analyzing the topology of the graph. Nevertheless, there is a gap in the literature about the interoperation of GNN models to graph databases, which will be discussed at the sections below.

III. A COMPARATIVE PRESENTATION OF HANDLING GRAPH DATA AND RELATIONAL DATA

Choosing and developing performant database, which will support operational functionalities and decision-making activities, seems to be a challenging task, especially when big datasets and complex data structures are involved. Conventional databases, such as mysql and oracle are designed to store relational data. The data are retrieved using joins among the tables. The performance of a query is directly associated with the complexity of the data model and the number of tables, used for the execution of the query. Those databases support codifying tabula structures and paper forms. Robinson et al. [13] pointed out the significance of implementing and utilizing database infrastructures for large datasets with complex relationships. Nevertheless, issues occurred when the relationships among the entities increase, or / and the complexity of the queries augments. When the data are stored on graph databases, users manage data more naturally, because they are represented as network of nodes - entities. In such cases, Neo4j significantly outperforms relational databases, since it models the data as a graph network, while they are retrieved by querying the graph. One of the main advantages of relational databases is that they can eliminate the redundancy, and ensure the integrity of data, which originated from the normalization process. On the other hand, the relational databases are not so flexible on changes that may occur to the model. Sandell et al. [14] conducted a performance comparison between state-of-the-art graphs and relational databases, evaluating their efficiency and capabilities. They conducted Performance Comparison Analysis of ArangoDB, MySQL, and Neo4j and the results indicate that Neo4j performs faster in querying connected data than MySQL and ArangoDB.

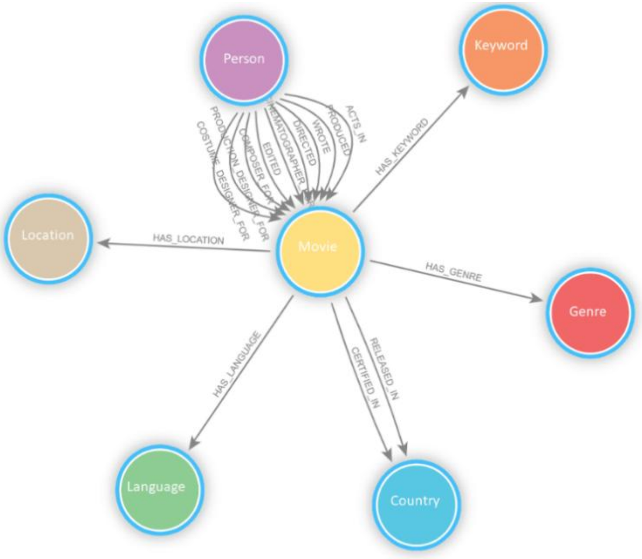


Figure 2 Graph Database Model

In this section, mysql and neo4j are compared, using an IMDB dataset of movies and people (actors, directors, etc.). In this dataset, there are seven entities (Country, Genre, Movie, Person, Language, Location and Keyword) and 15 relationships (Acts_In, Has_Keyword, Has_Genre,

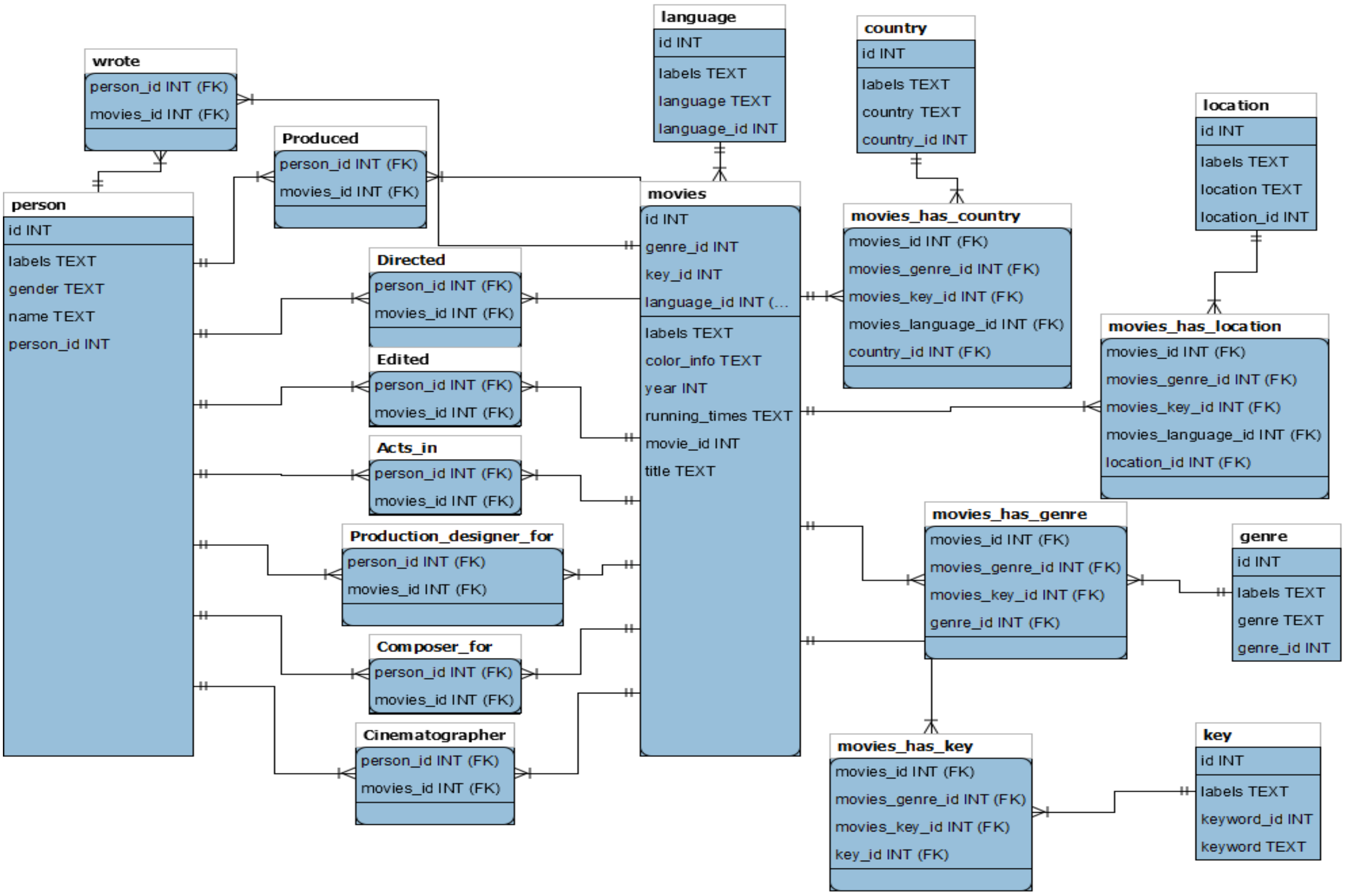


Figure 3 Relational Data Model

Released_In, Produced, Wrote, Directed, Has_Language, Cinematographer_For, Edited, Has_Location, Composer_For, Certified_In, Production_Designer_For, Costume_Designer_For). The comparison will be held on the following categories:

- Data Modeling
- Querying of Data
- Data Representation

A. Data Modeling

In figures 2 and 3, the models of graph databases (neo4j) and relational databases (mysql) are presented respectively. As it is obvious from the figures, the relational schema is more complex. Despite the fact that relational databases provide structured and high-quality data without inconstancies, due to the normalization process, the relationships and the constraints that exist among the entities have to be considered. The schema of a relational database cannot efficiently incorporate changes on the relationships on the database, as the schema is inflexible. On the other hand, because graph database is scalable and flexible it is easy to adopt any change on the relationships among the entities

B. Querying the Data

The performance of a query's execution is of vital importance and is associated with the data complexity and diversity. In relational databases, the performance is getting worst in case JOIN operations applied to large tables in contrast to graph databases. When the data are generated using two to three hops across tables, the RDBMS provide results in satisfactory response time. However, when more hops are required, then the response time is amplified, while in several cases, some tables will be locked, as they will wait the completion of the execution of the submitted queries.

C. Data Representation

The data representation is associated with the decision-making process, which depends on the selection of the database selected for data retrieval. The two main criteria that are used in order to decide the appropriate visualizations are:

- the cost to produce the results and
- the way the representations lead to informative decisions.

Figures 4 and 5 present the data in Neo4j and in MySQL. The graph-based representation lead to decisions that are more accurate and provide more functionalities so as to identify in detail the associations among the entities. For instance, in figure 6 the movie with title "Mystery of Maya" is selected and the entities associated with that movie are displayed. Nevertheless, in order to examine all the entities related to the selected movie in a relational database, too many queries, have to be executed that cause high demands on resources and increase the required time.

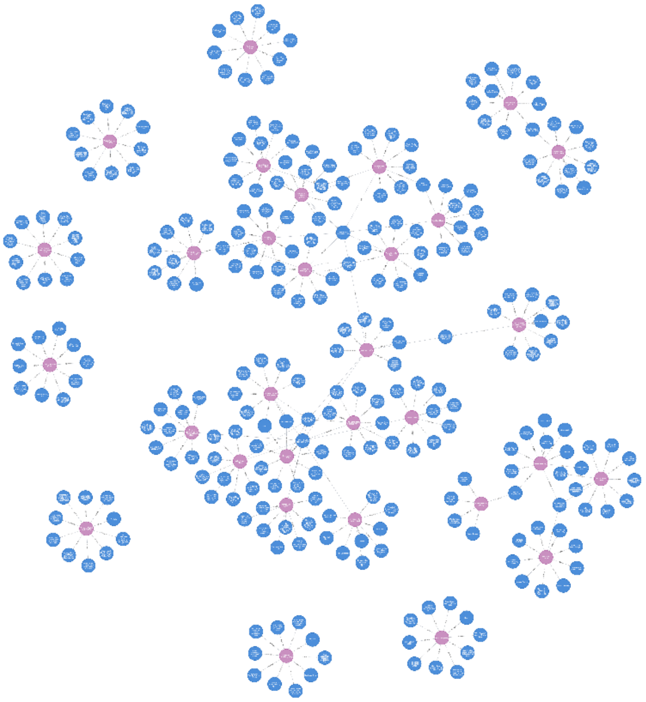
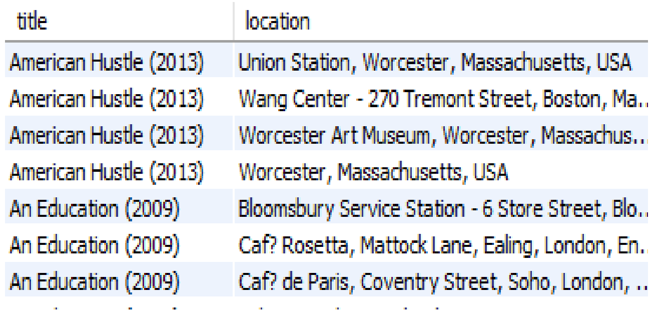


Figure 4 . Neo4j Graph Representation



title	location
American Hustle (2013)	Union Station, Worcester, Massachusetts, USA
American Hustle (2013)	Wang Center - 270 Tremont Street, Boston, Ma..
American Hustle (2013)	Worcester Art Museum, Worcester, Massachus..
American Hustle (2013)	Worcester, Massachusetts, USA
An Education (2009)	Bloomsbury Service Station - 6 Store Street, Blo..
An Education (2009)	Caf? Rosetta, Mattock Lane, Ealing, London, En.
An Education (2009)	Caf? de Paris, Coventry Street, Soho, London, ..

Figure 5 Mysql Tabular Representation

Table 1 illustrates some IMDB data used in the implementation process for both DBMSs.

Rank	1.0	2.0
Title	Guardians of the Galaxy	Prometheus
Genre	Action,Adventure,Sci-Fi	Adventure,Mystery,Sci-Fi
Description	A group of intergalactic criminals are forced to work together to stop a fanatical warrior from taking control of the universe.	Following clues to the origin of mankind, a team finds a structure on a distant moon, but they soon realize they are not alone.
Director	James Gunn	Ridley Scott
Actors	Chris Pratt, Vin Diesel, Bradley Cooper, Zoe Saldana	Noomi Rapace, Logan Marshall-Green, Michael Fassbender, Charlize Theron
Year	2014.0	2012.0
Runtime (minutes)	121.0	124.0
Rating	8.1	7.0
Votes	757074.0	485820.0
Revenue (Millions)	333.13	126.46
Metascore	76.0	65.0

TABLE 1 IMDB DATASET

IV. GRAPH NEURAL NETWORKS IN SOCIAL RECOMMENDATION

In recent years, there is a vast development in social network applications (such as Facebook, X, Amazon, etc.) that cover diverse areas of consumer interests and needs. All these applications keep the users' data stored. In most cases, they also record their social connections, as well as their interactions over time. These data is also known as Big Data, due to their massive volume. In industry, there is an enormous interest in exploiting valuable information from the before mentioned data, especially for their application in Social Recommendation. Social Recommender Systems are used in order to predict new connections or actions for a person using their social activities, connections and interactions with the other users in their social network. Those systems are valuable for the industry and especially in e-commerce, to offer more personalized recommendations to the users, based on their previous behavior or their interactions in their social network.

In Social Recommendation, ignoring social relations would lead to loss of valuable information. Social relations that are apparent in a social network have an effect on the preferences of an individual. As McPherson M. et al. [15] suggest, social homophily indicates that a user is more likely to connect to another user with similar

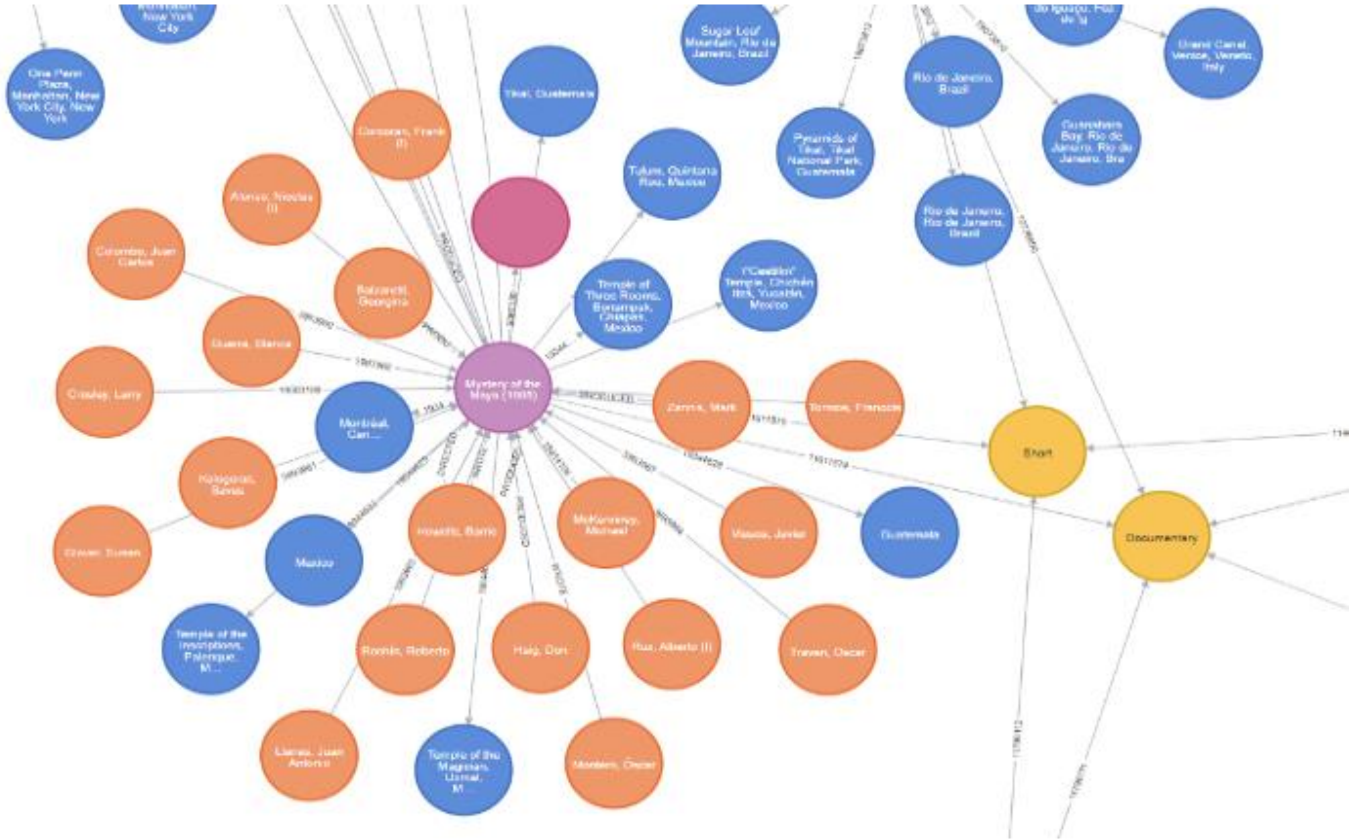

One Penn Plaza, Manhattan, New York City, New York
New York City
Corcoran, Frank (I)
Tikal, Guatemala
Alonso, Nicolas (I)
Tulum, Quintana Roo, Mexico
Colombo, Juan Carlos
Balzarotti, Georgina
Temple of Three Rooms, Bonampak, Chiapas, Mexico
'Castillo' Temple, Chichén Itzá, Yucatán, Mexico
Guerra, Blanca
Crosby, Larry
Mystery of the Maya (1995)
Montréal, Can...
Zanna, Mark
Tomson, Francois
Short
Kalogeras, Savas
Mexico
Hovanitz, Barrie
McKenney, Michael
Vasquez, Javier
Guatemala
Documentary
Temple of the Inscriptions, Palenque, M...
Rodrigo, Roberto
Haig, Don
Ruiz, Alberto (II)
Traven, Oscar
Llamas, Juan Antonio
Temple of the Magician, Uxmal, M...
Montero, Oscar
Sugar Loaf Mountain, Rio de Janeiro, Brazil
Rio de Janeiro, Brazil
Grand Canal, Venice, Veneto, Italy
Pyramids of Tikal, Tikal National Park, Guatemala
Guanabara Bay, Rio de Janeiro, Rio de Janeiro, Bra
Rio de Janeiro, Rio de Janeiro, Brazil
PRODUCED

Figure 6 Results on Graph Database

attributes and preferences. Likewise, Marsden P., et al. [16] talk about social influence, which indicates that users which are connected to each other, tend to have similar preferences. It is evident that social relations among people are of significant interest and provide valuable insights to analysts. Many different models have been developed and as Gao C., et al. [17] and by Sharma K., et al. [18] mentioned in their article, the first efforts to build Social Recommender Systems was the Matrix Factorization (e.g. [19] and [20]), followed by the deep learning models (e.g. [21], [22]). Among the deep learning models, various GNNs models are proposed. A plethora of such models are proposed in literature, with an ever-increasing rate over the last recent years [18], due to the fact that GNNs utilize more effectively the high order connectivity in social networks. Often GNNs models for Social Recommendation outperform the previous proposed models on public benchmark datasets [23].

According to literature, there are surveys that cover different aspects of GNN based Social Recommender Systems and referencing many of the proposed models (e.g. [17], [18]). The two main aspects covered, are the input types and the architectures of these systems. The input types refer to the graph types that model the data input and can be homogeneous graphs (i.e. edges connect only two nodes and there is one type of nodes and edges), heterogeneous graphs (i.e. edges connect only two nodes and there may be more than one type of nodes and edges), and hypergraphs (i.e. edges may connect more than two nodes). The input data are used by the Social Recommender System to be trained. As was suggested by Sharma, et al. [18], the architectures of these Social Recommender Systems have three key components: the encoder, the decoder and the loss function. The encoder is a GNN model that produces node embeddings (users and items). The GNN encoder must be suited for the corresponding graph of input data (e.g. homogenous, heterogenous). Decoders base their predictions on the embeddings, which are given by the encoders. Ultimately, the loss function trains the model so as to achieve accurate predictions.

Various GNN models are proposed as the encoder in the previous mentioned architecture, such as Graph Convolutional Networks [24], GraphSAGE [25], Graph Attention Networks [26], Gated Graph Neural Networks [27] and

Hypergraph Neural Networks [28]. Graph Convolutional Networks operate on spatial, or on spectral domain. On spatial domain, a neighbor of nodes is selected for each node and a convolution is applied in its neighbor. On spectral domain, the signal is transformed from spatial to spectral domain, a convolution is applied on the signal, and it is transformed back to the spatial domain. Either way, Graph Convolutional Networks leverage information from their neighboring nodes.

Social Recommender Systems have attracted research interest, since they have many possible application domains in industry and cover needs with high demand. Many GNNs based architectures have been proposed over the years. But GNNs have been also proven their efficiency for graph data in domains other than social networks, such as physical systems [29] and protein structure [30]. GNNs have been increasingly applied and evolve within machine learning for graph data, as they effectively leverage the complex, higher-order connectivity inherent in graph structures.

V. CONCLUSION

In this study, the research questions RQ1, RQ2 were affirmatively addressed through a comprehensive review and analysis of relevant literature findings. Concerning the first research question whether there is any satisfactory theoretical and experimental verification in the literature of the benefits of applying graph neural networks and graph databases to managing large-scale data in social networks, a large increase in publications of research results in the topics of GNN and Graph database is identified and mentioned in the paper. In particular, a remarkable rise of research outcomes has been observed since 2020. There is also a continuous reference to big data management and applications. Graph databases are widely used to manage Big Data and social network data. GNNs are increasingly being proposed as machine learning techniques for graph data, particularly in the aforementioned context. Consequently, GNNs and Graph Databases can be incorporated to represent, query and make predictions in big network data. Regarding the second research question if there are any comparative advantages of using graph databases over relational databases in managing large-scale data in social networks, the answer was experimentally validated, as the technology behind relational databases has reached maturity and is extensively utilized to meet business requirements. Relational databases are particularly well-suited for applications involving structured data, such as transaction processing and customer data management, while ensuring data integrity. Although the schema of a relational database is fixed, it allows for easy modifications. These databases can be scaled vertically with relative ease; however, their performance may decline as the size of the datasets increases. In contrast, graph database technology is relatively new, yet it excels at managing large datasets. These databases offer the advantage of intuitive usage. Additionally, they easily scale horizontally through partitioning. Moreover, graph databases give emphasis on supporting the relationships (edges) between objects (nodes), and, therefore, are well-suited for semantic search and recommendation engines.

VI. REFERENCES

[1] Almabdy, S. (2018). Comparative analysis of relational and graph databases for social networks. In 2018 1st International Conference on Computer Applications & Information Security (ICCAIS) (pp. 1-4). IEEE. https://doi.org/10.1109/CAIS.2018.8441982

[2] Kumar Kaliyar, R. (2015, May). Graph databases: A survey. In International Conference on Computing, Communication & Automation (pp. 785-790). IEEE. https://doi.org/10.1109/CCAA.2015.7148480

[3] Bhattacharyya, A., & Chakravarty, D. (2020, January). Graph database: A survey. In 2020 International Conference on Computer, Electrical & Communication Engineering (ICCECE) (pp. 1-8). IEEE. https://doi.org/10.1109/ICCECE48148.2020.9223105

[4] Tian, Y. (2023). The world of graph databases from an industry perspective. ACM SIGMOD Record, 51(4), 60-67. https://doi.org/10.1145/3582302.3582320

[5] Xirogiannopoulos, K., & Deshpande, A. (2017, May). Extracting and analyzing hidden graphs from relational databases. In Proceedings of the 2017 ACM International Conference on Management of Data (pp. 897-912). https://doi.org/10.1145/3035918.3035949

[6] Besta, M., Iff, P., Scheidl, F., Osawa, K., Dryden, N., Podstawski, M., Chen, T., & Hoefler, T. (2022, December). Neural graph databases. In Learning on Graphs Conference (pp. 31-1). PMLR.

[7] Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2008). The graph neural network model. IEEE transactions on neural networks, 20(1), 61-80. https://doi.org/10.1109/TNN.2008.2005605

[8] Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., & Sun, M. (2020). Graph neural networks: A review of methods and applications. AI open, 1, 57-81. https://doi.org/10.1016/j.aiopen.2021.01.001

[9] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S.. (2020). A comprehensive survey on graph neural networks. IEEE transactions on neural networks and learning systems, 32(1), 4-24. https://doi.org/10.1109/TNNLS.2020.2978386

[10] Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2018). How Powerful are Graph Neural Networks?. In International Conference on Learning Representations.

[11] Zhu, H., & Koniusz, P. (2021, May). Simple spectral graph convolution. In International conference on learning representations.

[12] Behún, M. (2023). Graph neural networks and their application to social network analysis [Master's thesis, Charles University]. Charles University Digital Depository.

[13] Robinson, I., Webber, J., and Eifrem, E. (2015). *Graph Databases: New Opportunities for Connected Data* (2nd ed.). O'Reilly Media.

[14] Sandell, J., Asplund, E., Ayele, W. Y., & Duneld, M. (2024). Performance Comparison Analysis of ArangoDB, MySQL, and Neo4j: An Experimental Study of Querying Connected Data. Hawaii International Conference on System Sciences 2024

[15] McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. Annual review of sociology, 27(1), 415-444. https://doi.org/10.1146/annurev.soc.27.1.415

[16] Marsden, P. V., & Friedkin, N. E. (1993). Network studies of social influence. Sociological Methods & Research, 22(1), 127-151. https://doi.org/10.1177/0049124193022001006

[17] Gao, C., Zheng, Y., Li, N., Li, Y., Qin, Y., Piao, J., Quan, Y., Chang, J., Jin, D., He, X. & Li, Y. (2023). A survey of graph neural networks for recommender systems: Challenges, methods, and directions. ACM Transactions on Recommender Systems, 1(1), 1-51. https://doi.org/10.1145/3568022

[18] Sharma, K., Lee, Y. C., Nambi, S., Salian, A., Shah, S., Kim, S. W., & Kumar, S. (2024). A survey of graph neural networks for social recommender systems. ACM Computing Surveys, 56(10), 1-34. https://doi.org/10.1145/3661821

[19] Jamali, M., & Ester, M. (2010). A matrix factorization technique with trust propagation for recommendation in social networks. In Proceedings of the fourth ACM conference on Recommender systems (pp. 135-142). https://doi.org/10.1145/1864708.1864736

[20] Yang, B., Lei, Y., Liu, J., & Li, W. (2016). Social collaborative filtering by trust. IEEE transactions on pattern analysis and machine intelligence, 39(8), 1633-1647. https://doi.org/10.1109/TPAMI.2016.2605085

[21] Deng, S., Huang, L., Xu, G., Wu, X., & Wu, Z. (2017). On deep learning for trust-aware recommendations in social networks. IEEE transactions on neural networks and learning systems, 28(5), 1164-1177. https://doi.org/10.1109/TNNLS.2016.2514368

[22] Krishnan, A., Cheruvu, H., Tao, C., & Sundaram, H. (2019, November). A modular adversarial approach to social recommendation. In Proceedings of the 28th ACM international conference on information and knowledge management (pp. 1753-1762). https://doi.org/10.1145/3357384.3357898

[23] Wu, S., Sun, F., Zhang, W., Xie, X., & Cui, B. (2022). Graph neural networks in recommender systems: a survey. ACM Computing Surveys, 55(5), 1-37. https://doi.org/10.1145/3535101

[24] Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In International Conference on Learning Representations.

[25] Hamilton, W., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. Advances in neural information processing systems, 30.

[26] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2018). Graph attention networks. In International Conference on Learning Representations.

[27] Li, Y., Tarlow, D., Brockschmidt, M., & Zemel, R. (2016). Gated graph sequence neural networks. In International Conference on Learning Representations.

[28] Feng, Y., You, H., Zhang, Z., Ji, R., & Gao, Y. (2019, July). Hypergraph neural networks. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 3558-3565). https://doi.org/10.1609/aaai.v33i01.33013558.

[29] Battaglia, P., Pascanu, R., Lai, M., & Jimenez Rezende, D. (2016). Interaction networks for learning about objects, relations and physics. Advances in neural information processing systems, 29.

[30] Fout, A., Byrd, J., Shariat, B., & Ben-Hur, A. (2017). Protein interface prediction using graph convolutional networks. Advances in neural information processing systems, 30.

VII. AUTHORS

Stratos Tsolakidis is a PhD candidate in the Department of Informatics and Computer Engineering of the University of West Attica. His topic is Graph Neural Networks and applications on Social Networks. His previous studies include a graduate degree in Mathematics from the National and Kapodistrian University of Athens and an MSc in Computer Science from the Athens University of Economics and Business. Currently, he works at the Ministry of Education, Religious Affairs and Sports in Greece.

Dr. Anastasios Tsolakidis received his PhD degree in computer science from the University of Limoges, France, in 2015 and is currently an Assistant Professor at the Department of Archival, Library and Information Studies at the University of West Attica. His research interests lie in the fields of Visual Analytics, Decision Support Systems, Business Intelligence and E-Government . He has previously worked as a research associate in many European and Greek projects and as a Business Intelligent Analyst at "e-Government Center for Social Security (IDIKA SA)" at the sector of E-Health.

Evangelia Triperina is a PostDoc researcher at the Department of Archival, Library and Information Studies of University of West Attica. She holds a PhD in Computer Science from the University of Limoges (France), with a thesis entitled "Visual interactive knowledge management for multicriteria decision making and ranking in linked open data environments". She holds an MSc in Information Technology, Image Synthesis and Computer Graphics from the University of Limoges (France). She worked in European research projects at GRNET, Agro-Know Technologies and the University of West Attica.

Nikitas N. Karanikolas is a professor at the Department of Informatics and Computer Science of the University of West Attica. He was Head of the Technological Educational (1996) Institute's Library (TEI of Athens Library). One year later, November 1997, he got the position of chief in the Department of Informatics and Organization of the Aretaieio University Hospital until November 2004. He participates as a coordinator and/or key researcher in European and nationally funded research and development projects. His research work has been published in international journals and conference proceedings.

Christos Skourlas is an emeritus professor at the Department of Informatics and Computer Science of the University of West Attica. He was an analyst-programmer and head of the systems with the National Documentation Centre of Greece (1983- 89) and a research assistant with the Nuclear Research Centre "Demokritos" (1977-82). He was head of the research lab "Data, Information and Knowledge Management (InfoDat_KM)". He participates as a coordinator and/or key researcher in European and nationally funded research and development projects. His research work has been published in international journals and conference proceedings.

Journal
of Integrated Information
& Management

Volume 9 - Number 2 / Jul -Dec 2024
https://ejournals.epublishing.ekt.gr/index.php/jiim