Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00206

A knowledge graph-based deep learning framework for efficient content similarity search of Sustainable Development Goals data

Irene Kilanioti[1a], George A. Papadopoulos[b]

[a]School of Electrical and Computer Engineering, National Technical University of Athens, 9 Heroon Polytechneiou St., Zografou Campus Athens 157 80, Greece
[b]Department of Computer Science, University of Cyprus, 1 University Avenue, Aglantzia, CY-2109, Nicosia, Cyprus

Abstract

Sustainable development denotes the enhancement of living standards in the present without compromising future generations' resources. Sustainable Development Goals (SDGs) quantify the accomplishment of sustainable development and pave the way for a world worth living in for future generations. Scholars can contribute to the achievement of the SDGs by guiding the actions of practitioners based on the analysis of SDG data, as intended by this work. We propose a framework of algorithms based on dimensionality reduction methods with the use of Hilbert Space Filling Curves (HSFCs) in order to semantically cluster new uncategorised SDG data and novel indicators, and efficiently place them in the environment of a distributed knowledge graph store. First, a framework of algorithms for insertion of new indicators and projection on the HSFC curve based on their transformer-based similarity assessment, for retrieval of indicators and load-balancing along with an approach for data classification of entrant-indicators is described. Then, a thorough case study in a distributed knowledge graph environment experimentally evaluates our framework. The results are presented and discussed in light of theory along with the actual impact that can have for practitioners analysing SDG data, including intergovernmental organizations, government agencies and social welfare organizations. Our approach empowers SDG knowledge graphs for causal analysis, inference, and manifold interpretations of the societal implications of SDG-related actions, as data are accessed in reduced retrieval times. It facilitates quicker measurement of influence of users and communities on specific goals and serves for faster distributed knowledge matching, as semantic cohesion of data is preserved.

Keywords: Content similarity; Distributed knowledge graphs; Sustainable Development Goals; Hilbert space filling curves; Deep learning

1. Introduction

Sustainable development signifies the enhancement of living conditions in the present without compromising future generations' resources. Sustainable Development Goals (SDGs) [1] were established by the United Nations (UN) in the framework of a 15-year plan, the UN 2030 Agenda, as a measurable international initiative to safeguard the future for the next generations, by eradicating poverty, protecting the environment and maintaining peace and welfare. SDG data

[1]Corresponding author: Irene Kilanioti (Email: eirinikoilanioti@mail.ntua.gr; ORCID: 0000-0002-4157-3900)

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00206

consists of sustainable development goals, targets, indicators and data series for the quantification of their accomplishment [2].

A lot of research has been conducted between 2015-2019 on the content and interactions of the SDGs according to [3], especially about goals associated with responsible consumption, sustainable cities and good health. The collaborative effort to accomplish the goals bears a transformative view of our world and focuses on building a peaceful, equitable society that will ensure protection of the environment and elimination of hunger and poverty. The collective effort to optimally harmonize sustainability goals bears a transformative view of the world and requires the conscious social, fiscal and technological contribution of many societal agents among which sustainable IT can also play a crucial role [4] [5]. "Data which are high-quality, accessible, timely, reliable and disaggregated by characteristics relevant in national contexts" is required (A/RES/70/01)[6].

From the IT perspective, the question of how we can leverage SDG data to estimate the impact of various actions especially in the context of social welfare and sustainability is a highly relevant topic and a challenge of great interest for society. Hence, we need to reduce access times for SDG data analysis and improve semantic cohesion of uncategorized data. Efficient processing and storage solutions for data in this respective field are necessary for practitioners, that entail intergovernmental organizations, government agencies and social welfare organizations, i.e., civic organizations and associations of persons engaged in the promotion of social welfare.

Heretofore, SDG-related information systems have included in essence solely monitoring tools of SDGs data and metadata, and also mechanisms to enhance interoperability across independent information systems: UN [7] showcases use of mappings of terms to the UN Bibliographic Information System (UNIBIS) and the EuroVoc vocabularies and a SDG interface ontology (SDGIO) has already been proposed [8]. In [9] Li et al. focus on the extraction of information to map to an ontology defined in collaboration with sector experts, that will enable the public to meet their knowledge needs related to social-impact funding. In [10] Warchold et al. study the unification of SDG datasets from various sources. However, to the best of our knowledge, a generic storage scheme of SDG data based on their semantic similarity and leveraging the infrastructure of multiple servers across which the data is split, has not been addressed yet. Subsequent improvements in the retrieval of data would facilitate all possible applications for causal analysis, inference, and manifold interpretations of the societal implications of SDG-related actions.

In this work, we propose dimensionality reduction methods to semantically cluster new uncategorised SDG data as well as new indicators with internationally yet unestablished methodology or standards and keep them close in the underlying physical networking environment of a distributed knowledge graph store. We introduce a framework of algorithms for insertion of new indicators and projection on the HSFC curve based on their similarity assessment, for retrieval of indicators and load-balancing. An approach for data classification of entrant-indicators is also described. The algorithms are based on HSFCs as the line of projection where new, gradually more refined, semantic categories are directly mapped onto. Our work proposes and experimentally corroborates the use of HSFCs to efficiently store distributed knowledge graph data, ensuring reduced access times and preservation of semantic closeness.

Section 2 describes the methodology we followed for use of an additional distributed environment layer based on HSFCs to map conceptually close, uncategorised according to existent SDG schema, data. First, in the subsection 2.1 the proposed algorithms for insertion of new indicators and projection on the curve, for filtering and refinement as well as load-balancing, and an approach for data classification of entrant-indicators over this layer are described. Then,

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00206

subsection 2.2 describes a detailed case study in a distributed knowledge graph environment, that experimentally evaluates our algorithm. Dataset and experimental setup are thoroughly discussed. The results are presented in Section 3 and discussed in light of theory and the actual impact they can have for practitioners in Section 4. Section 5 summarizes the paper's contribution and discusses future extensions of our work.

2. Methodology

2.1. Proposed algorithm

2.1.1. Hilbert Space-Filling Curves

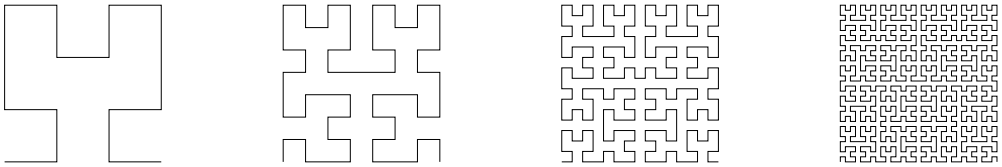

Figure 1: Construction of approximations of the Hilbert curves of increasing order τ=2; . . . ;5 in 2 dimensions.

A true Hilbert curve [11] is the limit of $\tau \rightarrow \infty$ of the τ^{th} discrete approximation to a Hilbert curve. HSFCs of 2 dimensions can be depicted on a NXN grid and the coordinates on the grid range in the space $x, y \in [0, N-1]$.

$$N = 2^{\tau} \tag{1}$$

In Fig. 1 next order curve comprises of four gyrated reiterations of the previous order curve. In the next repetition, quadrants are split up into four sub-quadrants each and so on. The line is repetitively folded in such a way that passes by successive neighboring points without intersecting itself and with infinite iterations of the curve construction algorithm it will not omit any point on a continuous plane. HSFCs are always bounded by the unit square, with Euclidean length exponentially growing with τ. Continuity of the curve ensures that affinity of bins on the unit interval signifies affinity in the unit square as well. Two points (x_1, y_1) and (x_2, y_2) with affinity in HSFC of order τ_1 depict affinity in HSFC of order $\tau_2 > \tau_1$ as well. Hilbert approximations result in more efficient maintenance of local features as opposed to that achieved by linear ordering, while locality properties degrade with the increase of dimensions.

2.1.2. Knowledge Graphs

A knowledge graph comprises of sets of triples that relate a subject entity to an object entity and encode domain and application knowledge. Knowledge graphs complimentarily serve for explainability that cognitively facilitates human-level intelligence. They serve for the representation of generic data interlinked by many relationships as well as for specific domains, such as biomedical research and manufacturing [12, 13]. They cover diverse application fields including search, data governance, question answering and recommendation. Distributed knowledge graphs integrate multiple and heterogeneous data sources, as their data are disseminated in a decentralised way across the web.

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00206

SDG ontology comprises substantially sustainable development goals, targets, indicators and data series for the quantification of their accomplishment [2], and the full taxonomy is accessible as linked open data at [14]. Depending on the grade of development of internationally established methodology and standards as well as regularity of production of relevant data, the afore-mentioned indicators are categorised into tiers. Tier III indicators are not associated with any existent methodology / standards. The distinguishing element between the first two tiers is the fact that data of Tier I indicators are collected on a regular basis for not less than half of the countries and population in every relevant region [15].

Challenges associated with the uptake of distributed knowledge graph technologies include their efficient storage and use at scale [16]. Heretofore, SDG-related systems have included in essence solely monitoring tools of SDGs data and metadata and mechanisms to enhance interoperability across independent information systems, e.g., [7]. However, to the best of our knowledge, a generic storage scheme of SDG data based on their semantic similarity to facilitate all possible applications for causal analysis, inference, and manifold interpretations of the societal implications of SDG-related actions, has not been addressed yet.

2.1.3. Insertion of SDG Tier III indicators in HSFCs

We propose an algorithm (Alg. 1) for the efficient placement of Tier III SDG indicators in the underlying physical networking environment.

Algorithm 1 Algorithm for insertion of SDG Tier III indicators in HSFCs

Input: $HSFC\_dims$, $HSFC\_order$, $indicator\_sentence$, $probe\_sentence$, μ
Output: $M = (indicator, T = (x \text{ tuple}, y) \in \mathrm{N})$
Parameters: bin, bin_size, $indicator$, $indicator\_number$, $T=(x,y) \in \mathrm{N}$, Hilbert_Space_Filling_Curve $HSFC$
Initialisation :

$HSFC \leftarrow ConstructHSFC(\ HSFC\_dims,\ HSFC\_order)$
$bin\_size \leftarrow \frac{|\mu|}{|(2^{HSFC\_order})^2|-1}$
A ←compute embedding for $probe\_sentence$
for $indicator = 1$ to $indicator\_number$ **do**
    B ←compute embeddings for $indicator\_sentence[indicator]$
    compute $s{=}semantic\_similarity = \frac{\cdot A \cdot B}{||A||||B||}$
    if $s(\ probe\_sentence,\ indicator\_sentence[indicator]) \leq Threshold_s$ **then**
        $bin \leftarrow \lfloor \frac{indicator}{bin\_size} \rceil$
        $T \leftarrow ObtainHSFCCoordinates(\ bin,\ HSFC)$
    end if
end for

The order of the HSFCs used defines the range of possible coordinates. We incorporate a binning mechanism to ensure that each new indicator can be projected to a tuple of coordinates in the higher dimension space. Bins hold consecutive elements of the data vector. The suggested mapping represents the indexing mechanism for the data in the distributed knowledge graph storage prototype we develop.

The first layer of the distributed knowledge graph store (Fig. 7) will entail semantic representation of data. In the next layer, which acts as a substrate of the network topology, we split up the indexing area in semantically homogeneous areas through HSFCs. Use of curves in this building block proves beneficial for preserving the neighbourhood property of concepts expressed by

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00206

the indicators, as semantically related terms, more probable to respond to a user query, will be placed in the vicinity. In our suggestion linearization is implemented as an overlay upon existing two-dimensional search structures and the distributed file system, that ensures distribution and sharding that scale. Multidimensional queries upon the distributed knowledge graph can be mapped to two-dimensional queries, that range from the minimum to maximum linearization points of the initial query (Fig. 7).

We are interested in relative positioning that expresses affinity. The reverse process of HSFCs mapping, when a position in the description space for higher-order partitioning needs to be translated into a position in the indicators vector, is not applicable and does not cause any issue in our scenario. The algorithm can be further modified to scale with new entries in terms of targets, goals and other potential refinements of the SDG ontology with corresponding increase in the order of the HSFC.

2.1.4. Similarity Assessment

For the similarity assessment of the indicators, we compute semantic textual similarity as calculated in Sentence-BERT (SBERT) [17], that extracts and compares semantically meaningful sentence embeddings and is based on deep learning transformers model BERT. We quantify the semantic textual similarity of each probe sentence, that is a candidate entrant-indicator, with existent SDG indicators (*indicator*).

Firstly, words in the $sentence_i$ are preprocessed. Then each processed word in the sentence is encoded into vectors v_{ij} of 300 dimensions. Embedding in the vector space is conducted with Word2vec. The vector representation for $sentence_i$ is based on the average of such v_{ij} vector representations for $j = \{i, w\}$, where w is the number of words in the sentence. Sentence embeddings for all existent indicators in the SDG taxonomy are precalculated. They are assumed to be close in the 300 dimensional vector space if they are similar. Computing cosine similary between the (300 dim) vector representation provides ideally score 1 for identical sentences and score 0 for sentences maximally dissimilar to each other.

Therefore, for the regression objective function the cosine-similarity between the two sentence embeddings u and v of two indicators is calculated (Fig. 9), and mean-squared-error loss is used at inference stage as the objective function.

For the computation of the classification objective function, e.g., to tune the model, sentence embeddings u and v of two indicators are concatenated with their element-wise difference and multiplied with the trainable weight $W_t \in \mathrm{R}^{3z \times \omega}$:

$$o = softmax(W_t(u, v, |u - v|)) \tag{2}$$

where z is the dimension of the sentence embeddings and ω the number of labels, and cross-entropy loss is optimized according to [17].

Fig. 8 depicts the transformer architecture for an entrant indicator of Tier III and Fig. 9 visualizes the architecture among existent indicators categorised according to SDG schema. The SBERT Algorithm is implemented with 12 stacked transformer layers and indicators can be depicted according to various available dimensions, including standard demographic info, location, etc.

2.1.5. Data Classification of Entrant-Indicators

We aim to avoid unnecessary congestion of specific subquadrants in the HSFC mapping, thus we team up semantically close entrants, namely indicators of Tier III. In this direction, we aim to

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00206

categorize points in $\mathbb{R}^n$ without exploiting SDG schema features. For the data classification, we use a dataset of N_{DC} individuals described by Q categorical variables and construct the $N_{DC} \times J$ indicator matrix Z, where:

$$J = \sum_{q=1}^{Q} J_q \tag{3}$$

rows denote the datasources, namely nodes of the graph store where data associated with the indicators reside, and columns denote the indicators of the uncategorised SDG data. We calculate a matrix of proportions P where $p_{ij} = n_{ij}/n$ and n is the sample size, summing up all values of N_{DC}. r and c are the sums along the rows and along columns respectively.

Categorization is based on chi-squared distances between two entrant-indicators:

$$dist^2_{\chi^2}(ind_j, ind_{j'}) = \sum_{i=1}^{N_{DC}} \frac{1}{r_i}\left(\frac{p_{ij}}{c_j} - \frac{p_{ij'}}{c_{j'}}\right)^2 \tag{4}$$

The distance is reduced when there is overlapping between individuals belonging to multiple categories. Our aim is to project the points onto a subspace of lower dimensionality, within which the eigenvectors u_k are the result of eigenvalue decomposition of $PD_c^{-1}P^T D_r^{-1}$. So we solve the equation:

$$\frac{1}{Q} Z D^{-1} Z^T u_k = \lambda_k u_k \tag{5}$$

where Z is the indicator matrix, D_r, D_c the diagonal matrix of row and column masses respectively.

2.1.6. Retrieval of indicators

The algorithm for matching k-semantically closest indicators is based on multi-step filtering and refinement, that consecutively removes irrelevant results and narrows the candidate set (Alg. 2). In order to optimally calculate distances, we use the algorithm proposed in [18], that performs optimally as far as the number of distance calculations is concerned, and modify it for HSFC representation. We create a ranking by means of the lower bound $l_{\delta H}$, that for all objects o_1, o_2 ensures that $l_{\delta H}(o_1, o_2) \leq \delta_H(o_1, o_2)$ for a distance function δ_H among HSFC projections. Reranking takes place provided that the lower bound does nor surpass the k^{th}-nearest neighbor distance and the results are updated with objects of smaller distances.

The process of refining multi-dimensional data to answer a query of k-closest semantically indicators after projecting on a HSFC is depicted in Fig. 2. After having reduced dimensionality with application of HSFCs, the query for semantically similar indicators can be handled as a nearest neighbor search and implemented with a multi-step filter-and-refine approach [18] [19] in an efficient way. The main idea is to filter at a later stage results falsely retrieved at first stage. Creating a lower bound with a simple distance function filters out initially irrelevant results, and in the next step evaluation of results returned at the previous stage takes place with the use of the original distance function. There are multiple properties describing each observation (data entry) and their Statistical Data and Metadata eXchange (SDMX)-standardized code equivalents are also provided. Dimensions (standard demographic info, the whole variety of different age profiles, etc.), time periods and area codes, described through the UNM49 standard are available in the dataset for each indicator from 2000 onwards.

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00206

Algorithm 2 Algorithm for filtering similarity search results

Input: $HSFC\_dims$, $HSFC\_order$, $indicator\_sentence$, $probe\_sentence$, k, query q, distances $l_{\delta H}, \delta_H$

Output: result_set S

Parameters: bin, bin_size, $indicator$, $indicator\_number$, T=(x,y) $\in$ N, Hilbert_Space_Filling_Curve $HSFC$

$S \leftarrow \emptyset$
$R_H \leftarrow$ranking(q, l_{δ_H})
$\xi \leftarrow$next value $\in$R$_H$
while l$_{\delta H}(q,\xi) \leq max_{\alpha \in S}\, \delta_H(q,\alpha)$ **do**
  if $|S| < k$ **then**
    $S \leftarrow$S $\cup \xi$
  else
    if $\delta_H(q,\xi) \leq$max$_{\alpha \in S}\, \delta_H(q,\alpha)$ **then**
      $S \leftarrow$S $\cup \xi$
      $S \leftarrow$S $-$ argmax$_{\alpha \in S}\, \delta_H(q,\alpha)$
    end if
  end if
  $\xi \leftarrow$next value $\in$R$_H$
  end
return S

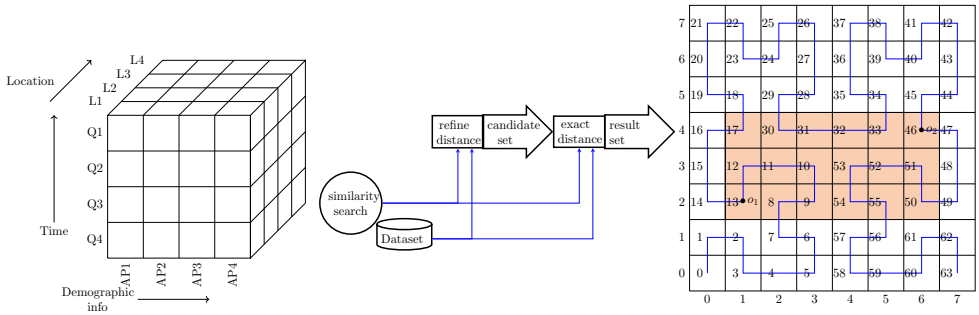

Figure 2: Process of refining SDG multi-dimensional data to answer a query of k-closest semantically indicators after projecting on a HSFC. After having reduced dimensionality (standard demographic info, the whole variety of different age profiles (AP), etc.), time periods and area codes, described through the UNM49 standard) with application of HSFCs, the query for semantically similar indicators can be handled as a nearest neighbor search and implemented with a multi-step filter-and-refine approach. Creating a lower bound with a simple distance function filters out initially irrelevant results, and in the next step evaluation of results returned at the previous stage takes place with the use of the original distance function.

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00206

2.1.7. Load-balancing

Algorithm 3 Algorithm for load balancing

Input: $HSFC\_dims$, $HSFC\_order$, $indicator\_sentence$, $probe\_sentence$
Output: assignment of virtual_nodes
Parameters: bin, bin_size, $Hilbert\_Space\_Filling\_Curve$ HSFC, $HSFC\_node$, $virtual\_node$, $load\_threshold$, $load$
Initialisation :

assign $virtual\_nodes$ to $HSFC\_node$
for $HSFC\_node$=1 to $HSFC\_nodes\_number$ **do**
  $load \leftarrow 0$
  for $virtual\_node$=1 to $virtual\_nodes\_number$ **do**
    $load(HSFC\_node) \leftarrow \sum_{i=0}^{n-1} load(virtual\_node)_i$
    if $load(virtual\_node) > load\_threshold$ **then**
      divide $virtual\_node$ to a set of $virtual\_nodes$
    end if
    if $load(physical\_node) > load\_threshold$ **then**
      assign $virtual\_node$ to $argmin_{HSFC\_node}(load)$
    end if
  end for
end for

Introduction of load balancing mechanisms at runtime or periodical batch-level processing of data ensures that in case of skewed distributions (more occurences of specific indicators or semantic categories) the equivalent subquadrants in the HSFC unit square will not be congested [20], [21]. We run a load-balancing algorithm (Alg. 3) based on existence of HSFC-physical nodes, that correspond to the physical placement of HSFC bins, and HSFC-virtual nodes Y_1, Y_2, ...Y_ϕ, that correspond to the physical node Y. Existence of virtual nodes can be beneficial in terms of fair splitting of computational power for heterogeneous nodes and decentralization in case of failure on a single physical node. Physical nodes can host several virtual nodes and their load is calculated with aggregation of the load of the virtual nodes they host. Exceeding the allowed threshold for a virtual node means that the node will be divided into more than one virtual nodes. When a physical node surpasses the load threshold, some of its virtual nodes will be assigned to less burdened physical nodes.

2.2. Case study

2.2.1. Dataset

We harvested a dataset of 2,21M. entries in total that includes all dimensions (standard demographic info, the whole variety of different age profiles, etc.), time periods and area codes, described through the UNM49 standard available for each indicator from 2000 onwards. We used the API of UN Statistics Division [22] with a set of scripts written in TypeScript and ran in the node.js environment. Dataset was particularly focused on indicators and list of all available SDG indicators was our starting point in the API, providing all available indicators in a self-contained response. Within the indicator related datasets, we collected 3 core datasets, while others were

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00206

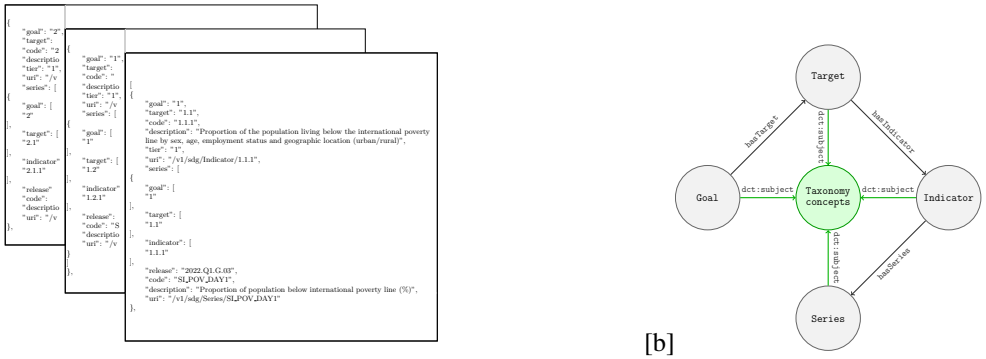

Figure 3: (a) IndicatorData dataset indicative excerpts for indicators 1.1.1, 1.2.1 and 2.1.1 (b) SDG schema of the dataset depicting structure of the UN SDG ontology.

mostly redundant data provided for different data access or interpretation. Our dataset IndicatorData includes 169 targets, 248 indicators (with 13 replicated under two/three different targets), as they were described in the 2022 refinement of the SDGs, as well as 663 data series for the quantification of the SDGs' accomplishment [2]. Since 2022 the classification entails 136 indicators of Tier I, 91 indicators of Tier II and 4 indicators consisting of modules of disparate tiers [15]. The dataset includes series information and goal - target hierarchy with overall 663 series across 248 indicators (Fig. 3). The number of data entries per each indicator is 4150 after removal of 20% of top and tail outliers. There are multiple properties describing each observation (data entry) and their Statistical Data and Metadata eXchange (SDMX)-standardized code equivalents are also provided. Table 3 depicts number of occurences of uncategorised indicators in graph store nodes and Table 4 describes potential uncategorised entrant indicators, that are not included in the dataset we harvested and do not follow SDG schema beforehand.

PivotData dataset returns a list of observations pivoted by year. This dataset contains all data described in IndicatorData aggregated for the whole observation period and showing only pivoting years in the years property of each data entry. The property was serialized and we deserialized it for the convenience of data manipulation.

There are 247.251 entries in total, with 550 entries per indicator after removal of 20% of outliers (top and tail).

2.2.2. Experimental Setup

We evaluate our algorithm in an experimental distributed environment over a key-value store of SDG data, that we collected. We use multiple servers and Hypertext Preprocessor (PHP) clients as APIs to handle cached values in a scheme built on Memcached, an optimized distributed hash map-based mechanism. Placement of data with HSFCs is compared to default placement scheme of the prototype distributed cache mechanism in terms of response time for the executed SELECT queries and in terms of disk I/O. Experimental setup settings are described in Table 2. In order to make clusters for entrant indicators and put their content in close Hilbert areas, we use the Agglomerative Hierarchical Clustering (AHC).

The similarity threshold we choose is minimum to allow augmentation of data with the whole set of entrant indicators. AHC proceeds with combination of clusters from the simple level of clusters-individuals to merging pairs of them with a bottom-up approach. The metric used in our

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00206

setup is the Euclidean distance for pairwise observations.

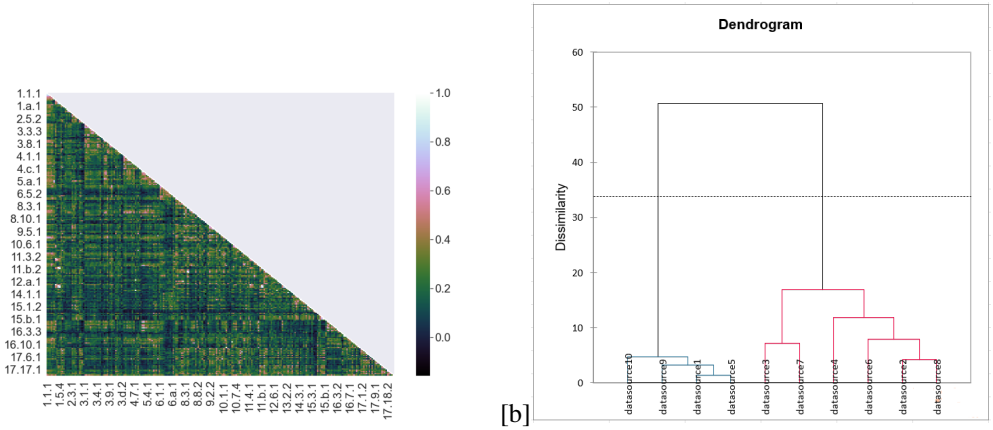

Figure 4: (a) Similarity among existent SDG indicators. (b) Entrant indicators' AHC dendrogram based on their number of occurences in datasources.

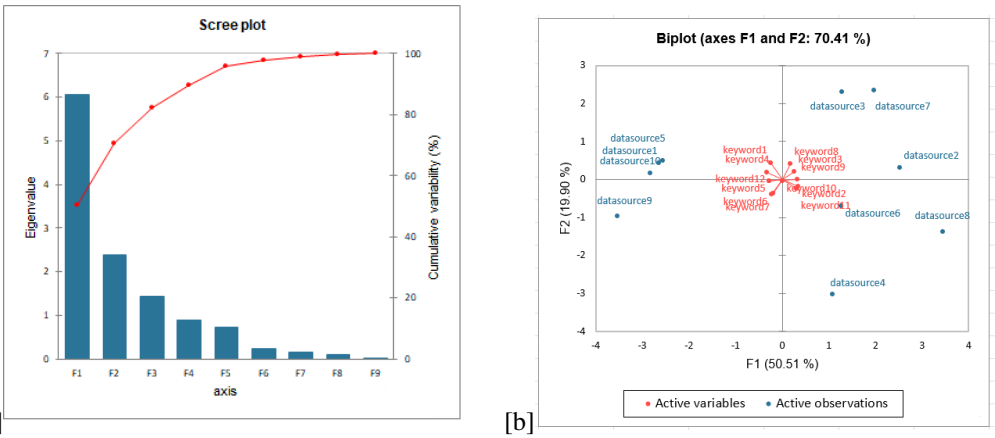

Figure 5: PCA of new indicators. (a) Scree plot with first two axes F1, F2 contributing. The Principal Component Analysis (PCA) scree plot indicates that two dimensions F1, F2 suffice for the visual interpretation of the analysis, since the sum of first two eigenvalues is sufficient percentage of variance. (b) Biplot PCA denoting the suggested division of layer to two Hilbert areas. The biplot verifies the split up of the datasources to two main semantic categories with active observations corresponding to the selected datasources and active variables corresponding to selected indicators.

3. Results

3.1. Cost-aware Data Classification of Entrant Indicators

Firstly, the approach for cost-aware data classification of entrant-indicators is verified. Concerning the uncategorised indicators of Table 4, cutting the dendrogram (Fig. 4) at the height of the dotted line verifies a coarser clustering of two semantic categories, namely of datasources

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00206

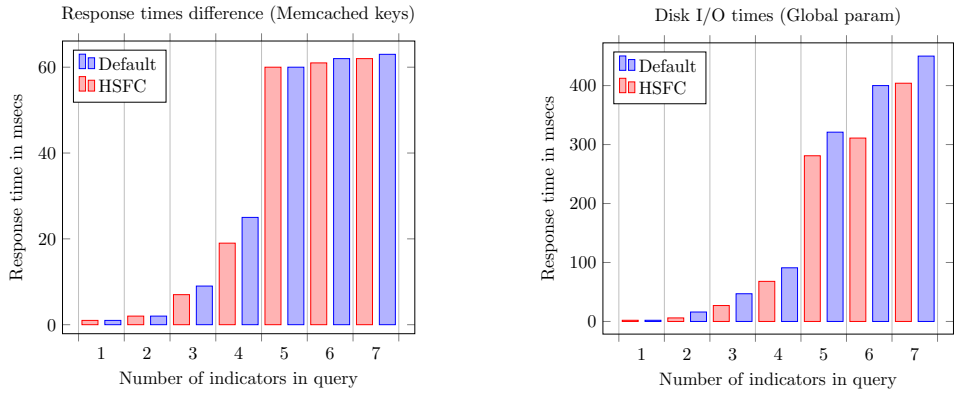

Figure 6: (a) Response time differences for HSFC mapping assigned to Memcached keys and (b) Disk I/O times for HSFC mapping passed as Globalparam. We notice significant reduction in average response times for selection queries of combined indicators. Time difference between HSFC storage scheme and baseline distributed key-value store approach is more obvious in the case of disk I/O times (Global parameters used). There is also improvement in response times when HSFC mapping is loaded into Memcached keys directly, which is more obvious for combinations of sets of up to 4 indicators in our setup.

(1,5,9,10) covering topics (1,4,5,6,7,8,10,12) associated with data and those of the rest datasources (2,3,4,6,7,8) associated with animal issues (2,3,9,11). Explicit reference to terms ("animal", "data") here is irrelevant. Thus, datasources (1,5,9,10) and datasources (2,3,4,6,7,8) should be put in two separate subquadrants in the Hilbert unit square. As for existent SDG indicators, Fig. 4 depicts their comparison in terms of semantic similarity.

The Principal Component Analysis (PCA) scree plot indicates that two dimensions F1, F2 suffice for the visual interpretation of the analysis, since the sum of first two eigenvalues is sufficient percentage of variance. The quality of the fit is measured by the percentage of inertia related to the two-dimensional map, namely the ratio of variance of coordinates of individuals on the axis to the total variance of coordinates of individuals. The quality is high for our dataset of restricted size (10 individuals (datasources) and 14 categorical variables (indicators)) and high data interlinking. With the eigenvalue λ_d equal to the variance of the points of each indicator on d-dimension:

$$(\lambda_1 + \lambda_2)/\sum_{d=1}^{9} \lambda_d = 70,41\% \tag{6}$$

the biplot verifies the split up of the datasources to two main semantic categories with active observations corresponding to the selected datasources and active variables corresponding to selected indicators (Fig. 5).

3.2. Response Time Reduction

We ran multiple sets of queries in an experimental distributed environment over a key-value store of SDG data with multiple servers and PHP clients as APIs to handle cached values in a scheme built on Memcached. After each set of queries the Memcached server was reset. We

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00206

notice significant reduction in average response times for selection queries of combined indicators. Time difference between HSFC storage scheme and baseline distributed key-value store approach is more obvious in the case of disk I/O times (Global parameters used). There is also improvement in response times when HSFC mapping is loaded into Memcached keys directly, which is more obvious for combinations of sets of up to 4 indicators in our setup (Fig. 6). The improvement in terms of memory response times can be further increased with further paging configuration, due to the nature of Memcached custom memory manager (slabs hold objects within specific ranges and slabs contain pages, split up in chunks) and the fact that a single indicator's entries reach up to 20MBs in our detailed dataset.

4. Discussion

In light of the HSFC theory, HSFCs can map multidimensional data to two dimensions maintaining spatial locality, namely affinity in the multidimensional space means relative affinity in the two-dimensional space. Our work is based on this observation and further extends it to harvest the benefits of a framework for similarity search for domain-specific knowledge graphs. Previously, HSFCs [11] have been used along with Gray code and Z-order curves for heuristic multi-dimensional indexing via linearization. The wide spectrum of applications includes image compression, data visualization and peer-to-peer architectures [23, 24, 25]. McSherry et al. [26] observed that edge ordering based on a HSFC substantially improves cache performance for single-threaded PageRank. Schmidt et al. [27] implemented a Distributed Hash Table (DHT)-based Web service discovery system leveraging HSFCs and mapped points of multidimensional space corresponding to service description components to DHT keys. Wang et al. [28] leveraged the spatial locality of HSFCs to store and display on request point-based spatial data in a spatial triple store.

In the suggested framework the locality in the multidimensional space describing the semantically associated indicators indicators is preserved after their mapping, as input items with higher semantic similarity are mapped to nearby addresses. Hence, nearby mapping is leveraged and the placement of conceptually close SDG indicator data on an HSFC as the line of projection indeed reduces retrieval times. The suggested topological mapping scheme is nondisruptive in terms of space and maintains local feature correlations of the original space. Moreover, our framework evades bottlenecks with avoidance of unnecessary congestion of specific subquadrants in the HSFC mapping via the approach of data classification of entrant indicators and the load-balancing mechanism that we suggest.

In our work HSFC points were coarsely equivalent to servers in the experimental distributed environment. Therefore, further refinement at a graph store node level and per server could lead to even better results in terms of response times, because communication cost among servers would be alleviated. The observed improvement in terms of memory response times can also be further increased with further paging configuration, due to the nature of Memcached custom memory manager (slabs hold objects within specific ranges and slabs contain pages, split up in chunks) and the fact that a single indicator's entries reach up to 20MBs in our detailed dataset.

The practical impact of our work is that data retrieval times are reduced for semantically close data, that have not been categorised according to the prevailing SDG schema. Our approach empowers SDG knowledge graphs for causal analysis, inference, and manifold interpretations of the societal implications of SDG-related actions, as data are accessed in reduced retrieval times. The framework facilitates quicker measurement of influence of users and communities

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00206

on specific goals and serves for faster distributed knowledge matching, as semantic cohesion is preserved.

Specifically, the suggested framework's impact on actions of organizations that harvest and process SDG datatakes is based on the consideration that the pursued goals and undertaken actions can be thematically interweaved and mutually influenced: one agent may pool interrelated existing content from federated repositories or one undertaken action may affect currently ongoing activities of other agents. Knowledge graphs are augmented with quicker similarity search, that reduces response times for manifold interpretations of the societal implications of actions that apply to each SDG goal, for association of contributions of actions to concrete SDG targets and quantification of the exerted influence. Concerning SDGs, knowledge graphs can quicker: i) support explainable decisions and insightful recommendations, ii) measure the influence of users and communities and iii) improve the user experience, as they facilitate extraction and organization of knowledge in a distributed manner and serve for quicker distributed knowledge matching.

5. Conclusions

Our work aims to support the efficient processing of SDG data and the seamless integration of novel indicators. An efficient storage scheme is needed for new uncategorised SDG data as well as indicators with internationally yet unestablished methodology and standards. In this paper, we propose a framework of algorithms for insertion of new indicators and dimensionality reduction based on their similarity assessment, for retrieval of indicators and load-balancing along with an approach for data classification of entrant-indicators. The mapping method is based on HSFCs as the line of projection where semantic categories of conceptually close SDG indicator data, uncategorised according to the existent schema, are directly mapped onto. A case study on real SDG data in a distributed knowledge graph store validates that data retrieval time is reduced. The proposed algorithms can be adapted for targets, goals, and potential future refinements of the SDG ontology.

Our approach empowers SDG knowledge graphs for causal analysis, inference, and manifold interpretations of the societal implications of SDG-related actions, as data are accessed in reduced retrieval times. It facilitates quicker measurement of influence of users and communities on specific goals and serves for faster distributed knowledge matching, as semantic cohesion of data is preserved.

We aim, furthermore, to study how increasing order of HSFCs affects performance. In another direction, we intend to explore geolocation features of indicators to leverage multiple HSFCs for spatial joins and range queries, as well as optimize queries to correspond to global search trends on SDG data.

The collective effort to optimally harmonize sustainability goals requires the conscious technological contribution of sustainable IT for timely and reliable data. Our work aspires to contribute in this direction and prove useful for all practitioners gathering and assessing SDG data.

Author Contributions

All authors including I. Kilanioti (eirinikoilanioti@mail.ntua.gr) and George A. Papadopoulos (george@ucy.ac.cy) took part in writing the paper. In addition, I. Kilanioti conceived the idea, designed the algorithms and experiments, collected the data, conducted the experiments and performed the data analysis.

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00206

References

[1] UN, "Sustainable development goals," September 2015. [Online]. Available: https://www.un.org/sustainabledevelopment/sustainable-development-goals/
[2] ——, "Global SDG indicator framework after 2022 refinement," 2022. [Online]. Available: https://unstats.un.org/sdgs/indicators/indicators-list/
[3] "Elsevier (2020). the power of data to advance the sdgs. mappingresearch for the sustainable development goals. in tech. rep. elsevier foundation, relx." 2020.
[4] I. Kilanioti and G. A. Papadopoulos, "An efficient storage scheme for sustainable development goals data over distributed knowledge graph stores," in *Proc. of 16th IEEE International Conference on Knowledge Graph (ICKG) '22*, Orlando, FL, USA, November 2022.
[5] I. Kilanioti, "Teaching a serious game for the sustainable development goals in the scratch programming tool," *European Journal of Engineering and Technology Research, Special Issue of 14th Conference of Informatics in Education CIE, Nov 2022*, vol. 7, no. 7, 2022.
[6] UN, "A/res/70/01," October 2015. [Online]. Available: https://undocs.org/
[7] ——, "Linked sdg," 2022. [Online]. Available: https://linkedsdg.officialstatistics.org/
[8] M. Jensen, "Sustainable development goals interface ontology." in *ICBO/BioCreative*, 2016.
[9] Y. Li, V. Zakhozhyi, D. Zhu, and L. J. Salazar, "Domain specific knowledge graphs as a service to the public: Powering social-impact funding in the us," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 2793–2801.
[10] A. Warchold, P. Pradhan, P. Thapa, M. P. I. F. Putra, and J. P. Kropp, "Building a unified sustainable development goal database: Why does sustainable development goal data selection matter?" *Sustainable Development*, 2022.
[11] D. Hilbert, "Über die stetige abbildung einer linie auf ein flächenstück," in *Dritter Band: Analysis· Grundlagen der Mathematik· Physik Verschiedenes*. Springer, 1935, pp. 1–2.
[12] A. Santos, A. R. Colaço, A. B. Nielsen, L. Niu, M. Strauss, P. E. Geyer, F. Coscia, N. J. W. Albrechtsen, F. Mundt, L. J. Jensen, and M. Mann, "A knowledge graph to interpret clinical proteomics data," *Nature Biotechnology*, vol. 40, p. 692–702, 2022.
[13] K. S. Aggour, V. S. Kumar, P. Cuddihy, J. W. Williams, V. Gupta, L. Dial, T. Hanlon, J. Gambone, and J. Vinciquerra, "Federated multimodal big data storage & analytics platform for additive manufacturing," in *Proc. IEEE Big Data '19*, Los Angeles, CA, USA, Dec. 9-12, 2019, pp. 1729–1738.
[14] UN, "SDG taxonomy," November 2019. [Online]. Available: http://metadata.un.org/sdg/
[15] ——, "Tier classification for global SDG indicators," June 2022. [Online]. Available: https://unstats.un.org/sdgs/iaeg-sdgs/tier-classification/
[16] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, "A survey on knowledge graphs: representation, acquisition, and applications," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 2, pp. 494–514, 2022.
[17] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. Conf. on Emp. Methods in Nat. Lang. Processing and the 9th Int. Joint Conf. on Nat. Lang. Processing (EMNLP-IJCNLP) '19*. Hong Kong, China: ACL, November 2019, pp. 3982–3992.
[18] T. Seidl and H.-P. Kriegel, "Optimal multi-step k-nearest neighbor search," in *Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, 1998, pp. 154–165.
[19] C. Yu, *High-dimensional indexing: transformational approaches to high-dimensional range and similarity searches*. Springer, 2002.
[20] B. Yagoubi and Y. Slimani, "Dynamic load balancing strategy for grid computing," *Transactions on Engineering, Computing and Technology*, vol. 13, no. 2006, pp. 260–265, 2006.
[21] C. Schmidt and M. Parashar, "Squid: Enabling search in dht-based systems," *Journal of Parallel and Distributed Computing*, vol. 68, no. 7, pp. 962–975, 2008.
[22] UN, "SDG API." [Online]. Available: https://unstats.un.org/SDGAPI/swagger/
[23] B. Moon, H. V. Jagadish, C. Faloutsos, and J. H. Saltz, "Analysis of the clustering properties of the hilbert space-filling curve," *IEEE Trans. Knowl. Data Eng.*, vol. 13, no. 1, pp. 124–141, 2001.
[24] J. K. Lawder and P. J. King, "Using space-filling curves for multi-dimensional indexing," in *Proc. British National Conference on Databases '00*. Springer, 2000, pp. 20–35.
[25] M. Ammari, D. Chiadmi, and L. Benhlima, "A semantic layer for a peer-to-peer based on a distributed hash table," in *Proc. Int. Conf. on Informatics Engineering and Information Science (ICIEIS) '11*. Springer, 2011, pp. 102–114.
[26] F. McSherry, M. Isard, and D. G. Murray, "Scalability! but at what COST?" in *Proc. 15th USENIX Conf. on Hot Topics in Operating Systems (HotOS XV)*, May 18-20, 2015. [Online]. Available: https://www.usenix.org/system/files/conference/hotos15/hotos15-paper-mcsherry.pdf
[27] C. Schmidt and M. Parashar, "A peer-to-peer approach to web service discovery," in *Proc. World Wide Web (WWW) '04*, vol. 7, no. 2. Springer, 2004, pp. 211–229.

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00206

[28] C.-J. Wang, "Database indexing for skyline computation, hierarchical relational database, and spatially-aware sparql evaluation engine," Ph.D. dissertation, 2015.

Appendix A Appendix

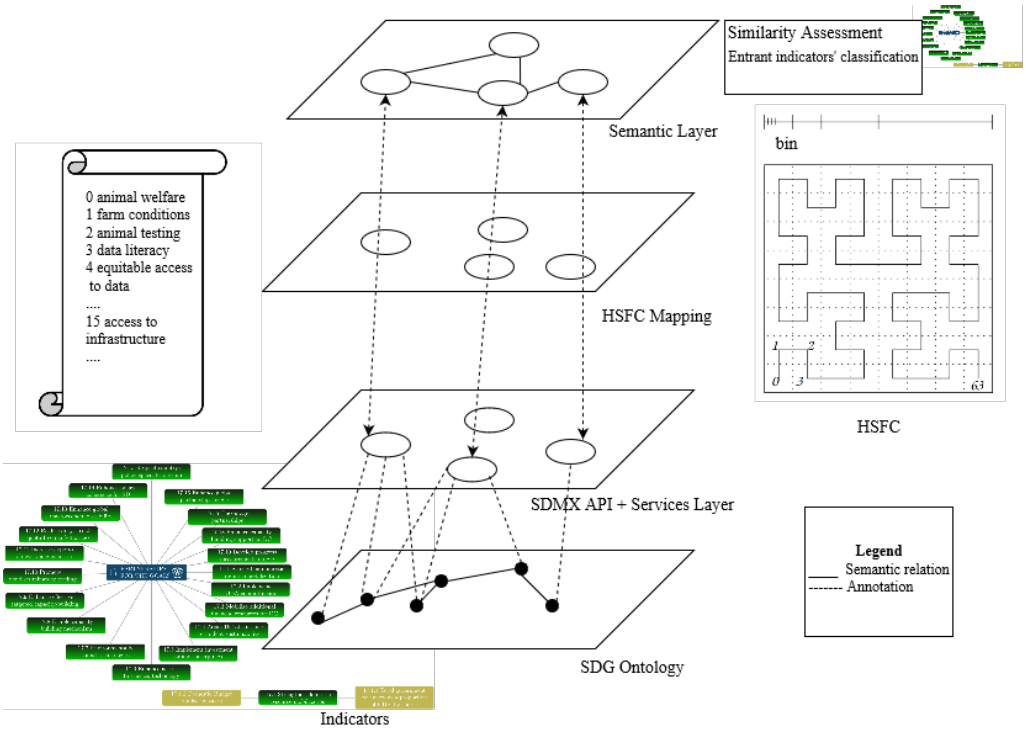

Figure 7: Suggested knowledge graph-based framework for efficient content similarity search of SDG data consisting of i)semantic representation of data, ii) a substrate of the network topology, where the indexing area is divided into semantically homogeneous areas through HSFCs, iii) SDMX-standardized code equivalents of data entries, iv) mapping to SDG ontology, when applicable.

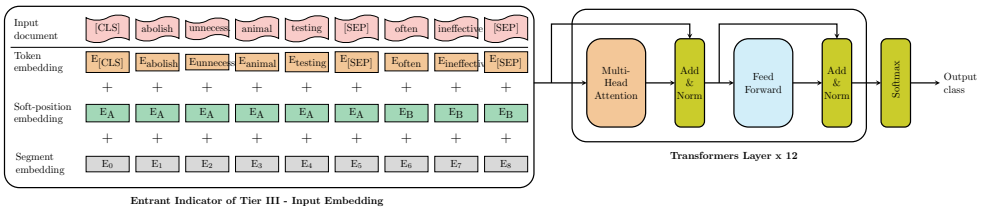

Figure 8: Transformer Architecture for Entrant indicator, implemented with 12 stacked transformer layers and consisting of token, position and segment embeddings.

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00206

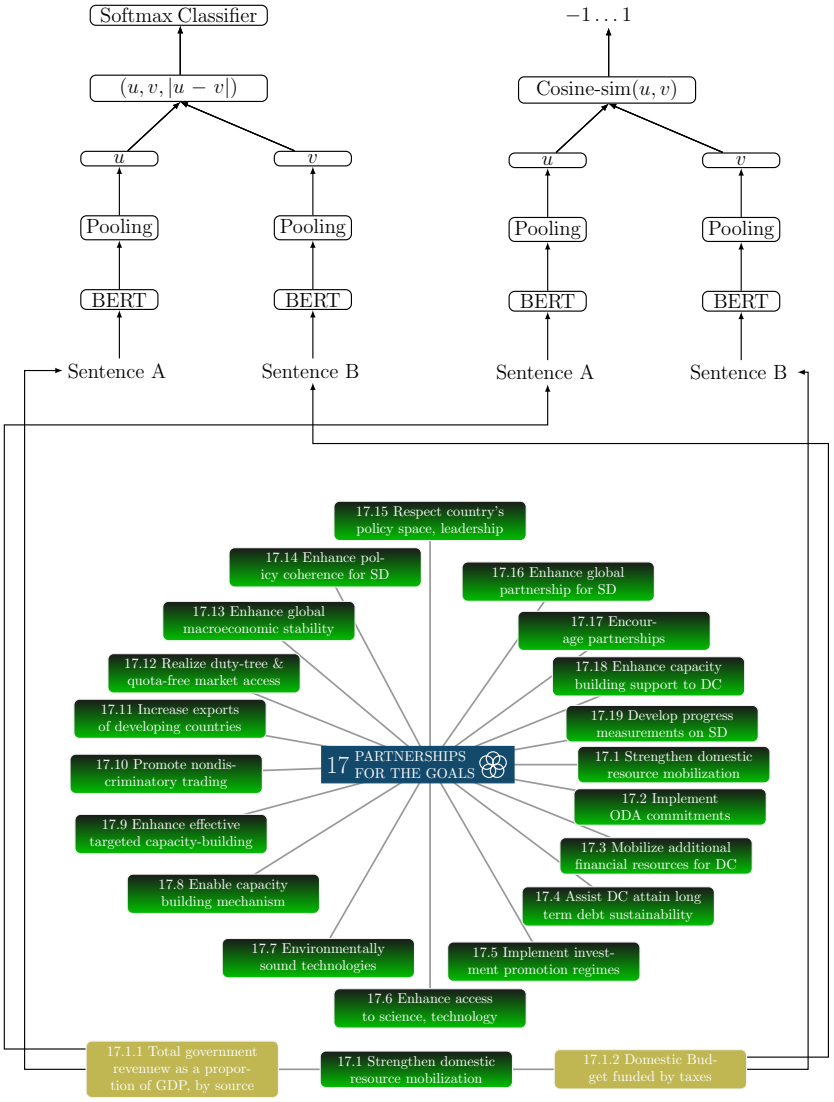

Figure 9: SDG Goal 17 consists of targets, such as Target 17.1: "Strengthen domestic resource mobilization" and entails indicators such as 17.1.2: "Total government revenue as a proportion of GDP, by source" [1]. Various indicators are fed into either Sentence-BERT (SBERT) [17] architecture with classification objective function, e.g., to tune the model, or SBERT architecture with the regression objective function, e.g., to compute indicator similarity scores.

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00206

Table 1: Notation Overview for knowledge graph-based framework for efficient content similarity search of Sustainable Development Goals data

Insertion of Indicators:	
N	grid size
$HSFC\_dims$	Hilbert dimension
$\tau, HSFC\_order$	Hilbert order
$indicator\_sentence$	indicator sentence
$indicator\_number$	number of indicators
$probe\_sentence$	probe sentence
μ	size of the reference set
bin	bin number
bin_size	bin size
T=(x,y) ∈ N	HSFC coordinates
M = ($indicator$, T= (x , y) ∈ N)	tuple
Similarity Assessment:	
$sentence_1$	ith $sentence$
w	number of words in the sentence
u, v	sentence embeddings
$W_t \in \mathrm{R}^{3z\times\omega}$	trainable weight
z	dimension of sentence embeddings
ω	number of labels
v_{ij}	vector representation of a word in sentence$_1$
Data Classification of Entrant Indicators:	
N_{DC}	individuals
Q	categorical variables
Z	indicator matrix
J	datasources
P	matrix of proportions
n	sample size
r, c	sums along the rows and along columns
$dist^2_{\chi^2}(ind_j, ind_{j'})$	chi-squared distances between two entrant-indicators
u_k	eigenvectors from eigenvalue decomposition of $PD_c^{-1}P^T D_r^{-1}$
λ_d	eigenvalue, variance of the points of indicator on d-dimension
D_r, D_c	the diagonal matrix of row and column masses respectively
Filtering of Similarity Search Results:	
δ_H	distance function for HSFC projection
$l_{\delta H}$	lower bound
k	maximum number of indicators in similarity search
q	similarity search query
R_H	rankings by means of lower bound
n_{Hq}	number of HSFC nodes with matching data for query q
Load Balancing:	
$physical\_node$	physical node
$virtual\_node$	virtual node
$virtual\_node\_number$	number of virtual nodes
$HSFC\_node\_number$	number of physical nodes
$load\_threshold$	load threshold
$load(HSFC\_node)$	load of HSFC node
$load(virtual\_node)$	load of virtual node

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00206

Table 2: Setup settings in an experimental distributed environment over a key-value store of SDG data

Parameter	***Value***
`Dataset`	2,21M. entries
`Number of servers`	3
`Number of virtual nodes per physical node`	1
`Queries`	SELECT for similarity search
`HSFC dimensions`	2
`HSFC order`	3
`Memcached server chunk size`	1MB
`Memcached server page size`	40

Table 3: Number of occurrences of uncategorised Tier III entrant indicators in datasources, namely nodes of the graph store where data associated with the indicators reside

Table	**Indicators**											
Datasource	***I1***	***I2***	***I3***	***I4***	***I5***	***I6***	***I7***	***I8***	***I9***	***I10***	***I11***	***I12***
`1`	16	16	16	16	17	17	16	16	16	16	16	16
`2`	15	18	18	14	16	16	15	16	17	16	17	15
`3`	16	17	17	15	16	16	14	17	18	16	17	16
`4`	14	18	16	14	16	18	16	16	17	17	18	16
`5`	16	16	16	16	16	17	16	16	15	16	16	16
`6`	15	17	15	15	15	17	15	16	17	16	18	15
`7`	16	17	17	15	15	16	15	18	17	17	17	15
`8`	14	18	17	14	15	16	16	16	18	16	18	15
`9`	16	16	15	16	17	18	16	15	16	16	16	16
`10`	16	16	15	16	16	17	16	16	15	17	16	16

Table 4: Tier III entrant indicators

Indicator	***Description***
`I1`	inclusive access to knowledge
`I2`	abolish unnecessary animal testing
`I3`	stop animal caging
`I4`	access to infrastructure
`I5`	cross border processing
`I6`	data erasure
`I7`	data portability
`I8`	data literacy
`I9`	improve animal welfare
`I10`	promote research
`I11`	improve farm conditions
`I12`	equitable access to knowledge

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00206

Author Biography

Dr. Irene Kilanioti works as Teaching Staff at NTUA. She received her B.Sc. and M.Sc. (Advanced Information Systems) from the Department of Informatics and Telecommunications, National Kapodistrian University of Athens (best student award). She received after evaluation one of the three scholarships of the Greek State Scholarships Foundation for a Ph.D. in informatics abroad (2012-2015). She received her Ph.D. (2016) entitled Improving Content Delivery with OSN-Awareness from the Department of Computer Science of the University of Cyprus. She has been a PostDoc researcher with the Ludwig-Maximilians-Universität in Munich (2018, 2019) and was accepted after evaluation at the LMU Mentoring Programme. Her research interests include Complex networks, Data mining and Big Data, Knowledge analysis, Content Delivery Optimization, Distributed computing and Adaptive educational software. She has publications in peer-reviewed conferences, journals and books and holds the best paper award in the 13th IEEE International Conference on Knowledge Graph 2022, Orlando, USA. She has been vice-coordinator for the use case Delivering Social Multimedia Content with Scalability, WG1, and trainer for the EU Cost action cHiPSet High-Performance Modeling and Simulation for Big Data Applications (2017-2019). She is an Editor for the journal Frontiers in Digital Education. She has worked as an informatics teacher in Greece and Germany and has served from seconded positions (National School of Public Administration, Information Society Office of the Greek Ministry of Education), as a software engineer (Vodafone, Enterprise Systems Development Department) and a teaching assistant (Design and Use of Database Systems, NKUA, Information-theoretic Data Mining, LMU). `https://www.ece.ntua.gr/en/staff/489`

Dr. George A. Papadopoulos holds the (tenured) rank of Full Professor in the Department of Computer Science, University of Cyprus. His research interests include Advanced Software Engineering, Ubiquitous Computing, Cloud Computing, Parallel and Distributed Programming Models, Technology Enhanced Learning, Medical Informatics, Assistive Technologies, Context Aware and Recommender Systems, and Internet Technologies. He has published over 150 papers as book chapters or in internationally refereed journals and conferences, he is a current or past member in the Editorial Board of 18 international journals and is serving or has served as a Chair or Steering or Program Committee member in more than 200 international conferences. Professor Papadopoulos is a recipient of a 1995 ERCIM-HCM scholarship award. He has been involved or is currently participating, as coordinator or partner, in more than 100 internationally and nationally funded projects (total budget for his participation close to 9.5 MEURO) and has been invited by the E.U. as an Expert Evaluator or Reviewer more than 50 times. Among his other activities, he is the Focal Point of Cyprus in COL's (`https://www.col.org/`) VUSSC (`https://vussc.col.org`) project. He is the Director of the Software Engineering and Internet Technologies (SEIT) Laboratory (`http://www.cs.ucy.ac.cy/seit`). More information can be found on his personal web site at: `http://www.cs.ucy.ac.cy/~george/`. His email is george@ucy.ac.cy.