MAY 12, 2023

THE UN SUSTAINABLE DEVELOPMENT GOALS AND THE UNIVERISTY SYSTEM

INFORMATICS FOR SUSTAINABLE SYSTEMS

MATT HODGES
CEE:5310

1 Introduction

Universities play influential roles in society, both from a cultural perspective and as discoverers and disseminators of knowledge. As such, they are often tasked with tackling the biggest, most intractable problems in society, the ones that planners and policymakers call wicked problems. Similarly, the United Nations' Sustainable Development Goals (SDG) address wicked problems, intending to provide guidance on how these societal challenges can be overcome. Though there may seem an alignment of purposes here, it is not always clear that the two institutions work toward the same end. The expectation is *not* that universities would be perfectly aligned with the UN SDGs, but as such influential components of society, to what degree should we place the responsibility of sustainability on the shoulders of universities? Measuring how closely the research produced by universities aligns with the SDGs might be a good place to start to assess the contribution they make in addressing sustainability issues.

Salvia and Brandli (2020) had a similar line of inquiry but specific to one SDG, namely Goal 7 which deals with affordable and clean energy. Through a systematic literature review of research articles available through the *ScienceDirect* and *Web of Science* databases, they determined, "Regarding the role of universities, it is important to note that this review points out that not only energy efficiency and sustainability could be improved, but that more attention should be paid to other areas as well…It indicates that universities could engage more [in broader areas of sustainable development]" (Salvia and Brandli, 2020). Their work suggests that by systematically reviewing the scientific literature the university system's contribution to sustainable development *can* be measured. This project aims to extend their work, updating their analysis to include more recent years of published work (their analysis included articles from 2007 through 2017) and extending it beyond the single SDG analyzed. It will use the same two databases of scientific literature, shifting the publication date range to 2012 through 2022, and will use natural language processing (NLP) techniques to analyze articles as they relate to a broader set of the SDGs, specifically goals 6, 7, 9, 11, 12, 13, 14, and 15.

Incorporating dashboarding tools to enhance the usability of this analysis, this project will allow policymakers, researchers, and students to identify research hotspots where universities are producing work that addresses a specific SDG of their interest. For instance, an applicant to graduate programs across the US may want to know where universities produce research on sustainable cities and communities (SDG 11). The dashboard tool will allow them to specify the SDG and targets that they want to study and will identify on a map the concentration of published work on their selected areas, along with summary statistics of the count, type, and timeframe of the work produced. For the potential graduate student, this may help them decide to which schools to apply, where to the researcher and policymaker it will help them zero in on not only research produced in their area of interest but also potential partners with whom to work to address their sustainable development area(s) of concern. For the purposes of this project, the geographic scope will be limited to the United States, although future analysis could easily incorporate locations outside the US with geographic information of the other universities listed in the research articles database.

2 Methods

2.1 Coding Structure

The web application portion of this project is built around a combination of hypertext processor (PHP), JavaScript (JS), cascading style sheets (CSS), and hypertext markup language (HTML) coding scripts. The partitioning of the coding in these formats makes for a more robust, dynamic system that allows for ease of user interface and system updates when new data is added or additional functionality is required. Data is being pulled from the server by the PHP script, which then is referenced in the JS to build functionality into the web application, and then flows through to the HTML coding to be displayed on the user webpage. By separating the system like this, troubleshooting can be identified and isolated more easily and updates can be much more targeted and require less repetition through the process. Figure 1 shows the output for the *getallaffiliates.php* script that pulls information from the database on affiliate ID, name, city, state, target, target ID, latitude, and longitude for any given article in the dataset. The

[[191649, 'Hofstra University', 'Hempstead', 'NY', '6.b', 55, 40.71595900, -73.60078000], [228778, 'The University of Texas at Austin', 'Austin', 'TX', '13.a', 115, 30.28282500, -97.73827300], [166027, 'Harvard University', 'Cambridge', 'MA', '11.2', 92, 42.37447100, -71.11831300], [147767, 'Northwestern University', 'Evanston', 'IL', '7.1', 56, 42.05035600, -87.67985800], [110583, 'California State University-Long Beach', 'Long Beach', 'CA', '6.6', 53, 33.78281800, -118.11204000], [228778, 'The University of Texas at Austin', 'Austin', 'TX', '11.2', 92, 30.28282500, -97.73827300], [228778, 'The University of Texas

Figure 1. Output from *getallaffiliates.php*

results of this output are limited to a certain range of possible targets, as identified as a limiting factor for the purposes of this study in the introduction. Figure 2 includes the SQL query that generates Figure 1's output, in addition to two other queries for the *getaffiliatedates.php* and the *getaffiliatecounts.php*, which are used for time series information and to aggregate the articles for a given affiliate and target. Finally, Figure 3 demonstrates the relational schema that

```
$query = "SELECT * FROM articles JOIN sdgtargets ON articles.targetid = sdgtargets.
    targetid JOIN universities ON articles.affiliateid = universities.univid  WHERE
    sdgtargets.target ilike '6.%' OR sdgtargets.target ilike '7.%' OR sdgtargets.
    target ilike '9.%' OR sdgtargets.target ilike '11.%' OR sdgtargets.target ilike
    '12.%' OR sdgtargets.target ilike '13.%' OR sdgtargets.target ilike '14.%' OR
    sdgtargets.target ilike '15.%'";

$query = "SELECT date, target, count(target) AS count FROM articles JOIN sdgtargets
    ON articles.targetid = sdgtargets.targetid JOIN universities ON articles.
    affiliateid = universities.univid WHERE sdgtargets.target ilike '6.%' OR
    sdgtargets.target ilike '7.%' OR sdgtargets.target ilike '9.%' OR sdgtargets.
    target ilike '11.%' OR sdgtargets.target ilike '12.%' OR sdgtargets.target
    ilike '13.%' OR sdgtargets.target ilike '14.%' OR sdgtargets.target ilike '15.%'
     GROUP BY date, target ORDER BY date";

$query = "SELECT name, target, count(target) AS count FROM articles JOIN sdgtargets
    ON articles.targetid = sdgtargets.targetid JOIN universities ON articles.
    affiliateid = universities.univid WHERE sdgtargets.target ilike '6.%' OR
    sdgtargets.target ilike '7.%' OR sdgtargets.target ilike '9.%' OR sdgtargets.
    target ilike '11.%' OR sdgtargets.target ilike '12.%' OR sdgtargets.target
    ilike '13.%' OR sdgtargets.target ilike '14.%' OR sdgtargets.target ilike '15.%'
     GROUP BY name, target ORDER BY name";
```

Figure 2. SQL queries for PHP scripts

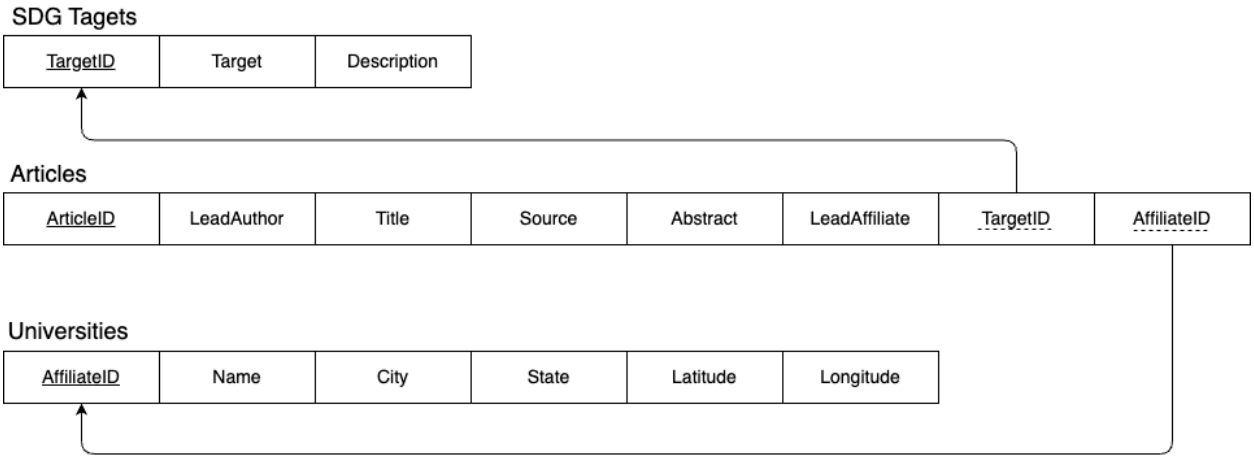


Figure 3. Relational schema of the Articles, Universities, and SDG Targets tables

underpins the three tables from the database upon which these queries are built. The dataset is drawn together around the articles table, with TargetID and AffiliateID serving as foreign keys in the Articles table. The data can be joined in the SQL queries upon this basis, bringing together the analysis that lets universities be associated with the SDGs.

Following the throughline from PHP to JS to HTML, the JavaScript code creates a series of functions that define variables, pull data from URLs (the PHP web services), identify certain elements of the interface (e.g., Google map), and creates IF-Statements and/or For-Loops that set conditions or iterate through the data to make the user experience dynamic. For instance, Figure 4 shows the beginning of the *createmarkers* function, that creates the markers for the map by cycling through a lengthy IF-Statement to identify which color-coded icon should be used for the marker (as it sits now, there are 70 conditions it must run through to find the correct icon; there is likely a more efficient to way to set this up, but it was not discovered before the project deadline). Another function worth highlighting are the map parameters defined by the *createmap* function. Figure 5 outlines that the map is centered on a point variable defined as 38 latitude and -103 longitude, slightly askew from where the center of the US would be should the map take up the whole screen because there are buttons to the left-hand side of the interface that requires the landmass of the US to be right of the screen. The map is also zoomed to 4.25, which sets it perfectly in the frame.

2.2 Keyword, Clustering, and Similarity Analyses

In the description of their methodology of article classification, Salvia and Brandli do a good job of defining the four categories of classification, but they do not explain at all *how* they classified the articles as they did. Without assuming too much, it is probably most easily implied that they “handpicked” the articles for each category by reading them. For this study, full texts of the articles were not acquired, so classification analysis using NLP techniques will be performed on the abstract of each article using three methods: keyword, clustering, and similarity analyses. Prior to running any of these analyses, some preprocessing of the data was required, including removing unnecessary attributes, standardizing attribute titles, types,

```
function createmarkers(affiliates) {
    var marker;
    var markericon;
    for (var i = 0; i < affiliates.length; i++) {
        point = {lat: affiliates[i][6], lng: affiliates[i][7]};

        if (affiliates[i][5]==48) {
            markericon = '6.png';
        } else if (affiliates[i][5]==49) {
            markericon = '6.png';
        } else if (affiliates[i][5]==50) {
            markericon = '6.png';
        } else if (affiliates[i][5]==51) {
            markericon = '6.png';
        } else if (affiliates[i][5]==52) {
            markericon = '6.png';
        } else if (affiliates[i][5]==53) {
            markericon = '6.png';
        } else if (affiliates[i][5]==54) {
            markericon = '6.png';
        } else if (affiliates[i][5]==55) {
            markericon = '6.png';
        } else if (affiliates[i][5]==56) {
            markericon = '7.png';
        } else if (affiliates[i][5]==57) {
            markericon = '7.png';
```

Figure 4. Beginning of IF-Statement for *createmarkers* function

```
function createmap() {
    var point = {lat: 38, lng: -103 };
    var mapbox = document.getElementById('map');
    map = new google.maps.Map(
        mapbox,
        {
            center: point,
            streetViewControl: false,
            scaleControl: true,
            fullscreenControl: false,
            zoom: 4.25
        }
    );
    infobox = new google.maps.InfoWindow();
    getallaffiliates();
}
```

Figure 5. *createmap* function that defines the map's parameters

formats, and sizes, merging the data from each database, deleting records with NULL values, and deleting duplicate records.

Keyword analysis will be applied to the collections of the abstracts as a whole. The abstracts will be individually tokenized, and parts-of-speech tags were applied to the tokenized words. The words will then be collected in an array that lists the word and the count of how many times the word appears throughout the abstracts. Stopwords—common words like "the", "and", or "in"—will be removed from these collections, both a standardized list of stopwords and a localized list identified through the first collections output. The final process applied to the keywords was network analysis, which finds connections between keywords that occurred in the same abstract (example outputs of this and the other analyses in the appendix).

Using Term Frequency-Inverse Document Frequency (TF-IDF) and *k*-means clustering, multiple collections of the abstracts will be identified for analysis as a potential way to classify the abstracts. TD-IDF is a method of counting terms in a document then assigning a weight to those terms based on their importance within that document, in this case a single abstract is considered a document. Each term and document will then be placed in a vector, or table, where the terms are in the attribute position and the documents are in the records position. Within this vector, a certain number of observations are produced that can be plotted on x and y coordinates. Finally, the unsupervised algorithmic *k*-means clustering technique will calculate based on the specified number of clusters desired (in this case, likely the specific number of targets being analyzed) and will then label each abstract with a number based on the predictive modeling of *k*-means clustering.

Two forms of similarity analysis will be performed on the abstracts: Jaccard similarity coefficient and cosine similarity. The former is the simpler equation, dividing the size of the intersection of a set by the size of the union of the set. Applied to texts, all matching words are the intersection, and the union is the total set of words. The larger the intersection, the larger the similarity between texts. Cosine similarity uses much more complicated math. Basically, it uses TF-IDF again to create vector observations of the texts that can then be plotted three dimensionally. Measuring from zero to the three-dimensional point of the observation of two texts, a cosine angle can be determined, and the smaller the cosine angle the more similar the two texts are to each other. Each method returns a ratio between 0 and 1, with scores closer to 1 having a higher similarity. For this study, the textual descriptions of the SDG targets will be compared using Jaccard and cosine methods to each of the abstracts to return a similarity score for each method. For each of those methods, the highest score across the comparisons will then used to place each abstract to one of the SDG target categories.

3 Results & Discussions

3.1 Data Life Cycle

Collect – Research article data will be collected from the *ScienceDirect* and *Web of Science* databases, accessible through login credentials to the University of Iowa Libraries databases system. Each of these databases has a unique user interface, offering advanced search features that allow for more targeted searches for fields like Author, Title, Keywords, and Abstracts. Additional data will be acquired on both the SDGs and on geographic data on universities across the United States. For the former, it is just the descriptive text of the goals, targets, and indicators, which can be compiled from the United Nations website. For the latter, the US Department of Education has downloadable shapefiles on geographic information on postsecondary institutions across the United States.

Assure – Data preprocessing will primarily consist of removing redundant and unnecessary features and removing records with missing information. The geographic information in the article databases will be challenging to standardize. *Addresses* is an attribute in the database, but an alternative approach may be more expeditious. The

data contains a feature called *Affiliations* that lists the higher education institutions that produced the research, which can be joined with the US Department of Education dataset to produce the geographic information necessary to pinpoint on a map. Other steps will likely be required to normalize and standardize the data.

Describe – The types and sizes of features in the dataset are fairly simple, with the heft of the analysis being conducted on the text attributes that describe the SDGs and the text attributes of the research article abstracts, which will be assessed for comparability. Other feature types include dates for publication month and year and latitude and longitude for geolocation information on universities.

Preserve – Concerns of data preservation are minimum for this project, however, the *Web of Science* database can only be accessed through University of Iowa institutional login credentials, so the information is not readily available to the public. As such, there may be concerns about what data can be shared through the dashboard. The university, SDG, and author information is publicly available data, so there should be no concerns with share it in the dashboard.

Discover, Integrate, and Analyze – Some of the early discoveries about the data are mentioned above, and some more preprocessing is anticipated to be necessary, requiring the data to flow back through some of the earlier stages of the data cycle. Integration steps will include identifying the best way to join the article information with the geographic information on US postsecondary educational institutions (also highlighted above), as well as standardizing the information from the two separate article databases, which in their raw form do not completely match each other. Natural language processing techniques will be deployed to compare the text of the SDGs, targets, and indicators to the article abstracts, with the intent of identifying research produced in each area of the SDGs. This part of the analysis will need to be tinkered with as the project moves along, ensuring an appropriate comparison threshold is established to secure accurate analysis.

3.2 Discussion

The main difference between this project and Salvia and Brandli's (2020) work will be where they started their analysis with categories of university contribution to sustainable development determined from another author's work (Filho, 2011), this study will not approach the scientific literature with predetermined categories but rather will use NLP techniques to attempt to identify research conducted on the various SDGs. It is unclear from their writing how they assigned each article to their categories, so it might be that they employed NLP techniques as well. Instead of using those tools to categorize articles that were already identified as related to SDG 7, however, this study will use NLP to find the relevant research on SDGs out of a broader pool of articles that may or may not be relevant to research on the SDGs. In other words, we will be applying the NLP tools at a step further back in the analysis process. Salvia and Brandli (2020) distilled the description of SDG 7's goal, targets, and indicators down to two strings that they then used to search for articles on *ScienceDirect* and *Web of Science*. Where this approach is likely more beneficial when analyzing a single SDG, we are hoping that our

approach will be more useful when considering a larger number of the SDGs and their targets and indicators (about half of them).

Currently, that larger pool of research articles has been sourced only from *Web of Science*, which has an easier user-interface that yields more manageable data than *ScienceDirect*. So far, 75,818 article abstracts have been collected, with publication dates ranging from 2018 through 2022 and with the designation of "Highly Cited Papers". This narrowed the list of articles to be analyzed from 11,745,333, which itself was narrowed from a larger pool of document types for those years of 17,213,922, both of which would be an unwieldy set of documents to analyze. The proof-of-concept, so to speak, of this type of analysis will be performed on the 75,818 article abstracts, with further analysis being executed on a larger set of articles should the models prove useful. As such, the *ScienceDirect* database will likely be left out of the analysis for the time being (barring identifying a more efficient way to access their articles' abstracts), as well as the broader date range mentioned in early portions of this proposal (that range of 2012 through 2022 produced double the total documents and about 2.5 times as many highly cited papers as the smaller set of years).

https://s-l112.engr.uiowa.edu/inf/inf_student06/project/home/index.html

References

Filho, W.L., (2011). About the Role of Universities and Their Contribution to Sustainable Development. *Higher Education Policy*, 24:427-438.

Gupta, S., (2018, May 15). Overview of Text Similarity Metrics in Python. *Medium*. https://towardsdatascience.com/overview-of-text-similarity-metrics-3397c4601f50

Salvia, A.L., & Brandli, L.L. (2020). Energy Sustainability at Universities and Its Contribution to SDG 7: A Systematic Literature Review. In: , *et. al.* Universities as Living Labs for Sustainable Development. World Sustainability Series. Springer, Cham.

Appendix