Unsupervised Topical Organization of Documents using Corpus-based Text Analysis

Sarkis Sarkissian
School of Engineering,
Dept. of Elec. and Compt. Eng.
Lebanese American University
36 Byblos, Lebanon
sarkis.sarkissian01@lau.edu

Joe Tekli[†]
School of Engineering,
Dept. of Elec. and Compt. Eng.
Lebanese American University
36 Byblos, Lebanon
joe.tekli@lau.edu.lb

ABSTRACT

This study aims at automating the process of topical keyword organization of set of documents in an input text corpus. It is conducted in the context of a larger project to investigate efficient unsupervised learning techniques to automatically extract relevant classes and their keyword descriptions from a set of the United Nations (UN) documents, and use the latter to produce reference corpora allowing to classify future UN documents. We assume that the reference classes are unknown in advance, and thus suggest an unsupervised clustering approach which accepts as input a bunch of unstructured text documents, and produces as output groups of similar documents describing similar topics. The input document feature vectors are augmented with term co-occurrence and relatedness scores produced from a distributional thesaurus built on the same (or a related) corpus. The augmented feature vectors are then run through a hierarchical clustering process to identify groups of similar documents, which serve as candidates for topical organization and keyword extraction. Experiments on a manually labelled dataset of documents classified against the UN's Sustainable Development Goals (SDGs) confirm the quality and potential of the approach.

CCS CONCEPTS

Information systems - Information retrieval - Document representation • Information systems - Information retrieval - Retrieval tasks and goals - Clustering and classification • Information systems - Information retrieval - Retrieval tasks and goals – Information extraction.

KEYWORDS

Document clustering, Topical organization, Keyword extraction, Corpus statistics, Distributional thesaurus, Augmented TF-IDF.

ACM Reference format:

Sarkis Sarkissian and Joe Tekli. 2021. Unsupervised Topical Organization of Documents using Corpus based Text Analysis. In *Proceedings of the 13th International Conference on Management of Digital EcoSystems (MEDES'21)*. ACM, Hammamet, Tunisia.

[†]Corresponding author. The author is co-founder of the United Nations ESCWA Knowledge Hub (UNEKH), which framework provided the usecase for this project.

MEDES '21, November 1–3, 2021, Virtual Event, Tunisia

ACM ISBN 978-1-4503-8314-1/21/11…$15.00
https://doi.org/10.1145/3444757.3485078

1. Introduction

In recent years, proficiency at collecting socio-economic data has largely overcome the capacity for managing it. Various governmental institutions (e.g., ministries, agencies, and directorates) and non-governmental institutions (e.g., NGOs and UN offices) produce huge amounts of data which proper exploitation offers great challenges in terms both effectively and efficiently accessing and retrieving relevant information. This is especially true for governments in the MENA region where institutions handle large collections of data without a common or unified organization of the information. For example, a UN or government employee wishing to draft a report or policy for some socio-economic topic (e.g., energy management, sustainable garbage treatment, etc.) would need to search through a tremendous amount of unrelated and disconnected information scattered all around the place, hoping to find the needed information to draft her/his report/policy in due time, which is a tedious and difficult task. Classifying and organizing these documents manually requires a tremendous amount of effort from the end user. There is a need to automatically classify and organize the data to provide smart and efficient data access and mining services to the user.

This study aims at automating the process of generating a topical keyword organization of set of documents from an input text corpus. It is conducted in the context of a larger project to investigate efficient unsupervised learning techniques to automatically extract relevant classes from a set of the United Nations (UN) documents, and use the latter to produce reference corpora allowing to classify future UN documents. We assume that the reference classes are unknown in advance, and suggest an unsupervised clustering approach which accepts as input a bunch of unstructured text documents, and produces as output groups of similar documents describing similar topics. Our approach extracts the term frequency features from the input documents, and augments the latter with term co-occurrence scores produced from a distributional thesaurus built on the same (or a related) corpus. The augmented feature vectors are then run through a hierarchical clustering process to identify groups of similar documents, which serve as candidates for topical organization and keyword extraction at the following stage. Experiments on a manually labelled dataset of 158 documents classified against the UN's Sustainable Development Goals (SDGs) confirm the quality and potential of the approach.

The remainder of the paper is organized as follows. Section 2 briefly presents the related works. Section 3 provides background information on frequency-based text processing. Section 4 describes our unsupervised topical keyword organization approach. Section 5 presents the experimental results, before concluding in Section 6.

2. Related Works

2.1. Topical Exploration of Web Data

Providing techniques for Web data search, exploration, and visualization is gaining importance, where topical exploration and result organization become essential to allow easier and more effective access to the data. In [7], the authors introduce the concept of *inCloud* to transform the flat organization of the linked data into a special structure of clustered topics using dedicated data aggregation and abstraction techniques. An inCloud provides a high-level topical view of the data and consists of 3 main components: i) a circle box, representing a cluster, which is a group of data that focus on a specific topic, ii) a square box which representing a summary of the contents of the cluster, and iii) an arrow, which represents the relationship between clusters. The thicker the arrow is, the more connected the two clusters are. To build an inCloud, the first step is to perform similarity evaluation on the input linked data, and produce a similarity graph where similarity links are added to connect the data nodes together. The second step is to perform topical aggregation to identify a set of topical clusters within the connected similarity graph. Clusters are formed by going through the similarity graph and detecting data nodes that are highly interconnected. To do this, the authors in [10, 11] rely on the clique percolation method (CPM). CPM divides the graph into sub graphs, where each sub-graph has *k* connected nodes. Two sub graphs are defined as adjacent if they share *k*-1 nodes. In a subsequent study in [9], the authors introduced *inWalk*, an interactive system for the exploration of linked data based on the concept of inCloud from [7]. The goal of InWalk is to overcome the rigid Web interfaces of linked data repositories by providing topical high level data views built through similarity-based aggregation techniques.. In [8], the authors build on the concept of inClouds from [7] to define inClouds as entity-driven collections of Web resources aiming at providing information organization structures. An inCloud is used to collect information relevant for a given target entity. A target entity is a keyword-based specification of a topic of interest for the user. So all the resources found to be prominent to the target entity are properly arranged and presented to the user.

Note that Web data are usually designed for answering to a general-purpose informative need, and are characterized by a large number of features. Some of these features are related to the internal structure of the data repository and are useless for satisfying user queries (for example: the name of the user who inserted a data resource). Others are intended to provide an informative description of the real object described by the linked data resource (for example: person names, locations, professions). Hence, in the clustering approach, all the important features need to be used to determine the similarity between different data resources. To overcome the above limitation, the authors in [16] introduce a dimensional clustering approach capable of selecting the set of features to use for data clustering, which are then packaged into topical dimensions. This provides a description of the similarity value that generates each cluster. Therefore, resources with the same degree of similarity but with different sets of matching features are put in different clusters, resulting in more accurate and focused clustering results.

2.2. Topical Exploration across Data Streams

In many information repositories available over the Internet, data are accessible as a continuous stream of textual information. These data are dynamic, and the ability to deal with them as time passes is very important for the analysis of the continuous data flow [34]. As a result, users need to perform an exploratory analysis of the underlying data, driven by topics extracted from the data flow. In addition, the featured topics must be correctly located in the data flow timeline, to identify emergent topics and to study and understand topic evolution. In [10], textual data streams are represented and consumed as a continuous bootstrapping process, where each bootstrapping cycle works on an incoming document chunk that belongs to a fixed time window. Each incoming document is indexed to extract a representative keyword-set from its textual content to be used for bootstrapping. The solution consists in applying a bootstrapping cycle for each chunk of documents belonging to a time interval. For each document D_j, the acquired textual content is stored in a data stream repository along with the corresponding timestamp t_j. Each document is then associated with a keyword set extracted through the execution of a conventional text tokenization and normalization procedure. The bootstrapping cycle consists of the sequential execution of three tasks: i) document clustering (grouping the most similar documents together), ii) topic discovery (merging the most similar clusters together), and iii) topic assimilation (identifying the top keywords in every merged cluster). The bootstrapping cycle is based on the notion that similarity is first employed to compare documents, then for clusters, and finally for topics. Documents are represented as bags of keywords, and are compared using the Jaccard similarity index. Document clustering is performed using the HCf+ hierarchical clustering algorithm [16]. Topic discovery consists in applying a second round of clustering: merging similar clusters together to generate the corresponding topics. Clusters are also represented as bags-of-words and are compared using the Jaccard index. Topic assimilation consists in correctly linking newly-emerged topics (i.e., topics discovered in the current bootstrapping cycle) with topics discovered in the previous bootstrapping cycle. A new topic is linked with an existing one when it is recognized to be similar based on their keyword sets. The top keywords in the obtained clusters represent the corresponding topic descriptions.

While the above mentioned studies focus on linked data and Web data streams, our present study targets flat textual data and does not consider any data structure (no graph connections, attribute-value pairs, subject-predicate-object triplets, or time-based) that can be used in linking the data together or forming the seeds of thematic clusters. Our approach aims at producing a topical organization of the documents from scratch, considering only their flat textual contents and mining their augmented term co-occurrence relatedness scores.

3. Background on Text-based Processing

3.1. Document Representation and Term Weighting

Information retrieval (*IR*) is a branch of informatics concerned with the acquisition, organization, storage, search and selection of information [31]. The goal of *IR* is to efficiently identify and retrieve, from a data collection, information that is relevant w.r.t. (with respect to) the user's needs [5]. With conventional *IR*, documents and user queries usually consist of sets of keywords. Identifying documents that are relevant (similar) to a given query comes down to:

- Comparing the keywords of each document, in the document collection, to those of the query,
- Ranking the documents w.r.t. their keyword similarities with the query (document selection is undertaken using a similarity threshold, e.g., range queries [1] or KNN queries [30]).

In legacy IR solutions, a text document is usually represented as a bunch of keywords, which are commonly weighted in order to reflect their relative importance in the document at hand. The underlying idea is that terms that are of more importance in describing a given document are assigned a higher weight. As a weighting scheme, the standard *TF-IDF* (*Term Frequency – Inverse Document Frequency*) approach (and its variants) of the *vector space model* [26, 33] is usually used. In the standard vector space model[1], documents and queries are indexed in a similar manner, producing vectors in a space which dimensions represent, each, a distinct indexing unit t_i. An indexing unit usually stands for a single term, i.e., a keyword[2]. The coordinate of a given document D on dimension t_i, is noted $w_D(t_i)$ and stands for the weight of t_i in document D within a document collection. $w_D(t_i)$ is computed using a score of the *TF-IDF* family, taking into consideration both document and collection statistics. The relevance of a document D w.r.t. a query (or document) Q, designated as $Sim(D, Q)$, is evaluated using a measure of similarity between vectors such as the inner product, the cosine measure, the Jaccard index, the Dice coefficient, etc., [5, 21]. Note that while relevance in *IR* is a broad and imprecise notion, the abstract concept of *relevance* is generally concretized by the notion of *similarity* [29].

As for *TF-IDF*, different variations have been proposed in the literature [31-33]. In this study, we utilize the standard definition, consisting of two factors [31]:

- The *TF* (*Term Frequency*) factor which designates the number of times a term t_i occurs in document D (document statistics). The importance of a given term t_i in describing a document D increases with the frequent use of t_i in D.
- The *IDF* (*Inverse Document Frequency*) factor, emphasizing the fraction of documents that contain term t_i (collection statistics). Here, the importance of a given term t_i in describing a document D decreases with the frequent use of t_i in the document collection.

A common *TF-IDF* mathematical formulation [33] is as follows. $w_D(t_i) = TF \times IDF$:

- $TF = tf(t_i, D)$ is the number of times term t_i occurs in D
- $\text{IDF} = \log \frac{N}{df(t_i, D)}$ where N is the total number of documents in the document collection, and $df(t_i, D)$ is the number of documents containing term t_i

Using the legacy vector space model, the relevance (similarity) between documents is evaluated considering the documents' term weights and distributions. We aim to extend the latter using corpus-based term relatedness, by integrating term co-occurrence weights from a distributional thesaurus.

3.2. Distributional Thesaurus

A thesaurus is a type of dictionary that lists synonymous terms. For instance, the WordNet thesaurus [28] entry for term "education" includes synonyms "instruction", "teaching", "pedagogy", "didactics", and "educational activity". The latter can be of equal importance, or can be weighted and ranked following their relative importance in describing the target term. A thesaurus is traditionally generated manually by a group of people or linguistic experts (which is the case of WordNet). While descriptive for general-purpose scenarios, yet manually creating such thesauri for specific and dynamically changing application domains (which is the case for UN documents), requires a lot of time, effort, and human labor. In this context, distributional thesauri construction methods can be used, e.g., [6, 39], to allow mining the syntactic/lexical relatedness between terms in the documents. A distributional thesaurus is a thesaurus generated automatically from a given textual corpus (such as the Brown corpus [17], COCA [13], or the domain specific textual collection being mined), by finding words that co-occur together or that have similar contexts in the corpus. Generating related terms automatically can save time and human effort, albeit sacrificing accuracy. It is relatively cheap in terms of human effort (it does not require any manual labor) and can be updated by rebuilding the distributional thesaurus to reflect changes to the corpus at hand. Domain specific thesauri are capable of capturing specific and technical terminologies that are unambiguous to the domain at hand, and can dynamically adapt to the domain at hand, compared with statically generated general purpose dictionaries which are more difficult to adapt and change [27].

4. Unsupervised Topical Document Organization

The overall architecture of our proposal is shown in Figure 1. The system accepts as input a corpus of flat text documents, and produces as output a topical organization of the documents consisting of groups of similar documents associated with their most descriptive keywords/expressions. It consists of five main components: i) text preprocessing, ii) distributional thesaurus extraction, iii) feature vector representation, iv) document clustering, v) topic extraction. We describe each of the components in the following subsections.

4.1 Text Preprocessing

Many preprocessing tasks are conducted before the documents can be processed for feature extraction and clustering. First, this component performs data serialization which consists in extracting the raw text from the input pdf document, and transforms it into an in-memory data representation that can be processed by the application software. Second, it converts all words to their lowercase form, and removes all stop-words and punctuations from the text. Third, it performs stemming or lemmatization, following the user's preference: i) stemming converts all the words to their original syntactic forms (stems) using syntactic stemming rules[3]; ii) lemmatization transforming words into their original lexical forms using a lexical reference[4]. While lemmatization is generally more accurate than stemming, yet it requires significant additional processing time. Hence, the user can choose the former or the latter following her computation resources and needs.

[1] Note that various *IR* models, other than the vector space model, have been proposed in the literature, among which the Boolean model [25], the probabilistic model [18], the LSI (Latent Semantic Indexing) model [14], the DFR (Divergence From Randomness) model [3], etc. However, we restrict ourselves to the vector space model since it is the most commonly used, its performance being accredited in a broad variety of applications and scenarios (e.g., [22, 27]).

[2] A keyword can also consist of multiple words (phrase units).

[3] We adopt the Porter Stemmer [37] in our approach since it's one the most effective and commonly used in the literature.

[4] We adopt the WordNet lexical dictionary [28] to perform lemmatization, since it's one of the most commonly used machine-readable lexical knowledge bases in the literature.

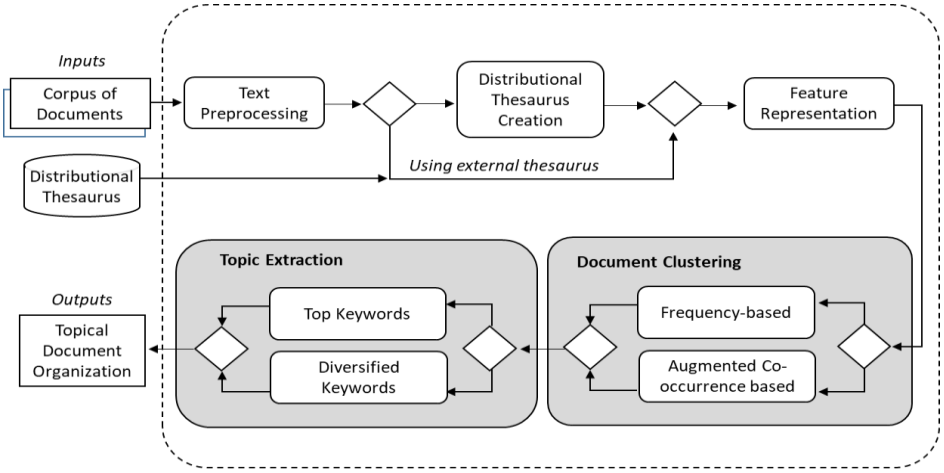

Figure 1: Simplified activity diagram describing our approach

4.2. Distributional Thesaurus Creation

The pseudo-code of our distributional thesaurus generation component is shown in Figure 2. It accepts as input: a text corpus *C*, as well as input parameters designating the co-occurrence *window size* and the number of *top-ranked terms* needed to identify related terms. For each term t_i in *C*, the algorithm creates a *relatedness vector* $RV(t_i)$ to store the co-occurrence frequencies of surrounding terms (lines 1-4). It identifies a window size consisting of the terms occurring to the left and right of the target term in the reference corpus, and adds all window term frequencies to the *relatedness vector* (lines 5-8). Once the vector has been obtained, we normalize vector scores w.r.t. overall maximum term co-occurrence frequency (line 9), and identify the top-ranked terms of the target term t_i, which are considered as the most related terms to t_i (lines 10-13). The output distributional thesaurus consists of the list of distinct terms from *C*, where every term t_i is associated a co-occurrence vector $\overrightarrow{V_{Occ}}$=< $occf(t_i, t_j)$, $occf(t_i, t_k)$, …> providing the co-occurrence frequencies of the top terms co-occurring with t_i in *C*.

```
Algorithm Distributional_Thesaurus_Generation
Input: C          // Text corpus
       c1, c2     // Input parameters: window size and top-ranked terms

Ouput: DT         // Distributional Thesaurus
Begin
DT = ∅                                                              1
For each term ti in C                                               2
{                                                                   3
    Create RV(ti) from C              // Relatedness vector for term ti  4
    For each term tj in window(ti, c1, C)                           5
    {                                                               6
        Add Freq(tj) to RV(ti)                                      7
    }                                                               8
    RV(ti) = RV(ti) / Max(RV(ti))     // Normalizing RV(ti) scores  9
    Ti = set of c2 top-ranked terms in RV(ti)                       10
    DT = DT ∪ (ti, Ti)                // Adding new DT entry        11
}                                                                   12
Return DT                                                           13
End
```

Figure 2: Pseudocode of distributional thesaurus generation

In our solution, the user can choose to generate the distributional thesaurus form the document corpus accepted as input for topic extraction, or can be generated based on an external corpus. The latter needs to be carefully chosen to describe the target documents at hand, since the effectiveness of a distributional thesaurus depends on the lexical coverage and expressiveness of its reference corpus. For instance, using a *sports* document corpus to identify the co-occurrence relationships between terms describing *medical events* will lead to a non-descriptive distributional thesaurus, thus negatively affecting document clustering and topic extraction accuracy.

4.3. Document Feature Representation

Our approach allows two document representations: i) term-frequency based, and ii) term co-occurrence based. The former represents documents as multi-dimensional feature vectors, where vector weights are computed using legacy TF-IDF term scoring techniques developed in IR (cf. Section 3.1). The latter augments the term feature vectors produced previously, with co-occurrence scores provided by the distributional thesaurus. Let t_i and t_j be two terms in the corpus *C*, let $occf(t_i, t_j)$ be the number of times t_i and t_j have co-occurred together in *C*, and $max\_occf(C) = \underset{\forall\, t_i, t_j \in C}{\text{Max}} \left(occf(t_i, t_j)\right)$ be the maximum number of co-occurrences in *C*. We compute the co-occurrence score of every term t_i in the document feature vector as the sum of: i) the term's initial frequency score (TF-IDF) and ii) the non-null co-occurrence scores of all terms in the document vector obtained from the distributional thesaurus. More formally, given document *D* with term frequency feature vector $\vec{V}$, and given $w_{\vec{V}}(t_i)$ the initial term frequency weight of t_i in *D*, the augmented co-occurrence weight of t_i, noted $w_{\vec{V}}(t_i)^*$ is computed as follows:

$$w_{\vec{V}}(t_i)^* = w_{\vec{V}}(t_i) + \alpha \sum_{\forall\, t_j \in D} \frac{occf(t_i, t_j)}{max\_occf(C)} \quad \textbf{(1)}$$

where α is a linear scaling factor allowing to assign more or less importance to the co-occurrence scores following user preference. The augmented co-occurrence weights allow to increase the descriptiveness of the terms in the document, following their co-occurrences with related terms from the reference corpus. Considering document *D* with term frequency vector $\vec{V}$ and the co-occurrence vectors from the extract distributional thesaurus in Table 1.a, the augmented term co-occurrence vector is shown in Table 1.b.

Table 1: Sample term frequency and augmented term co-occurrence vectors

a. Extract distributional thesaurus

$\overrightarrow{V_{Occ}}(t_1)$ = < (t_2, *occf*(t_1, t_2)=2), (t_3, *occf*(t_1, t_3)=1) >
$\overrightarrow{V_{Occ}}(t_2)$ = < (t_1, *occf*(t_2, t_1)=2) >
$\overrightarrow{V_{Occ}}(t_3)$ = < (t_1, *occf*(t_3, t_1)=1) >
$\overrightarrow{V_{Occ}}(t_4)$ = < (t_5, *occf*(t_4, t_5)=4) >
Max_occf(*C*)= 4

b. Sample term frequency and augmented term co-occurrence vectors, considering α=1

	t_1	t_2	t_3	t_4
$\vec{V}$	$w_{\vec{V}}(t_1) = 0.5$	$w_{\vec{V}}(t_2) = 0$	$w_{\vec{V}}(t_3) = 0$	$w_{\vec{V}}(t_4) = 0$
$\vec{V}*$	$w_{\vec{V}}(t_1) + \alpha \frac{occf(t_1, t_2)+occf(t_1, t_3)}{max\_occf(C)}$	$w_{\vec{V}}(t_2) + \alpha \frac{occf(t_2, t_1)}{max\_occf(C)}$	$w_{\vec{V}}(t_3) + \alpha \frac{occf(t_3, t_1)}{max\_occf(C)}$	$w_{\vec{V}}(t_4) = 0$
	$= 0.5 + \frac{2+1}{4} = 1.25$	$= 0 + \frac{2}{4} = 0.5$	$= 0 + \frac{1}{4} = 0.25$	

4.4 Document Clustering

We utilize the well known single link hierarchical clustering method [19, 20] although any form of clustering could be utilized. Given *n* documents, we construct a fully connected graph G with *n* vertices (documents) and $\frac{n \times (n-1)}{2}$ weighted edges. The weight of an edge corresponds to the similarity between the connected vertices. Consequently, the single link clusters for a similarity threshold $Thresh_{Sim}$ are identified by deleting all the edges with weights < $Thresh_{Sim}$. Therefore, the single link clusters will group together documents that have pair-wise similarity values greater or equal than $Thresh_{Sim}$. We utilize *cosine* as a similarity measure to compare document feature vectors, and the *silhouette score* as a stopping rule to choose the best hierarchical level to stop the clustering process [15]. Note that other vector similarity measures or stopping rules could be used from the literature, e.g., [2, 15].

4.5. Topic Extraction

Topic extraction in our approach consists in identifying the most important keywords describing the clusters generated previously. The first step consists in producing the feature vector representations describing every cluster. This is achieved by aggregating the term frequency (and augmented co-occurrence frequency) vectors of their constituent documents. Consequently, we consider two approaches to select the topical terms: i) top keywords, and ii) diversified keywords. The *top keywords* approach consists in representing every cluster by its top-*k* keywords, ranked following their term weights in the cluster feature vector. While straightforward, yet a main concern with the top keywords approach is that some subtopics whose keywords do not have high enough weights will not appear in the cluster's extracted topics. In addition, the top keywords might be very similar to each other and might lack diversity, since terms that commonly appear together are usually related in meaning. To address the latter problems, we introduce the *diversified keywords* approach, where we consider that an efficient topic extraction solution should be able to provide a global view of the clusters, identifying keywords that are both relevant and that cover diverse aspects of the cluster. This is based on the assumption that data clusters usually involve many aspects and target multiple sub-topics [2, 4]. By widening the pool of possible topics, one can increase the likelihood of the system providing the user with the needed information, thus increasing its effectiveness. To perform diversified keyword identification, we conduct a second round of clustering within the clusters obtained in the first clustering phase. We utilize the same single link hierarchical clustering algorithm used previously, with cosine as the similarity measure and silhouette score as the stopping rule. This results in sub-clustering every initial cluster into groups of most similar documents. The sub-clusters are then processed to produce their feature vectors and extract their top-*k* keywords. Consequently, the top keywords from every sub-cluster are combined and ranked following their weights, to form the diversified keyword list of the initial cluster, resulting in its topical representation.

5. Experimental Evaluation

5.1. Prototype Implementation

We have implemented our topical organization solution using the Python programming language, to test and evaluate its performance. We perform the data serialization using Python's *PDFToText* library. This is one of the high-performance libraries to parse pdf documents and convert them into string serializations. After parsing, we remove the stop-words and punctuations using Python's *NLTK* library. Later we convert the remaining text to lower-case and we stem each word using the *NLTK Porter Stemmer*. We utilize the WordNet library to perform lemmatization. We compute the term frequency vector for each document and store the results in a dataframe, where each row represents a document and each column represents a distinct term dimension. Each cell contains the term frequency weight of each term w.r.t. its document. We perform distributional thesaurus construction following our implementation of the pseudo-code in Figure 2. Consequently, we perform clustering and build the dendrograms using Python's *SciPy* hierarchical clustering library. We use the *Sklearn* cluster metrics library to determine the number of clusters which achieve the highest silhouette score, and which we utilize as a stopping rule in our approach.

Our implementation and experimental data are available online[1]. An executable version of our solution has been deployed in the UN-ESCWA's MANARA search engine[2].

[1] https://shorturl.at/afnJT
[2] https://manara.unescwa.org/home

5.2. Experimental Data

We conducted various experiments on a manually labelled dataset of 158 documents classified against the United Nations' 17 Sustainable Development Goals (SDGs)[1]. The labels are assigned by human experts and provide, for each document, the primary SDG describing it, and sometimes two or three more additional SDGs that are related to it. We consider both the primary SDG and the related SDGs in our evaluation. We divide the dataset into multiple subsets of equally sized classes to make sure that the reference data is not biased toward any specific SDG. The document subsets are described in Table 2.

Table 2: Characteristics of test documents

	# of docs	# of classes (primary SDGs)	# of docs per class	Avg. doc size (in KB)	Total subset size (in KB)
Subset 1	12	3	4	1,384	16,608
Subset 2	20	4	5	1,641	32,829
Subset 3	15	5	3	2,630	39,451
Subset 4	18	6	3	2,461.00	44,311
Subset 5	63	7	9	2,711	170,821

5.3. Experimental Metrics

5.3.1. Cluster Evaluation

Owing to their proficient usage of their traditional predecessors in IR evaluation, we make use of the *precision* (*PR*) and *recall* (*R*) metrics [12, 38] to evaluate the effectiveness of our clustering approach. For an extracted cluster C_i that corresponds to a given primary SDG_i:

- a_i is the number of documents in C_i that indeed correspond to SDG_i (correctly clustered documents).
- b_i is the number of documents in C_i that do not correspond to SDG_i (miss-clustered).
- c_i is the number of XML documents not in C_i, although they correspond to SDG_i (documents that should have been clustered in C_i).

Consequently, given n: the total number of generated clusters:

$$PR = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} a_i + \sum_{i=1}^{n} b_i} \in [0,1] \text{ and } R = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} a_i + \sum_{i=1}^{n} c_i} \in [0,1] \quad \textbf{(2)}$$

High *precision* denotes that the clustering task achieved high accuracy, grouping together documents that actually correspond to the SDGs mapped to the clusters. High *recall* means that very few documents are not in the appropriate cluster where they should have been. In addition, we evaluate *F-value*, which represents the harmonic mean of *precision* and *recall*:

$$F\text{-}value = \frac{2 \times PR \times R}{PR + R} \in [0,1] \quad \textbf{(3)}$$

High *precision* and *recall*, and thus high *F-value* indicates in our case excellent clustering quality, and thus excellent topical organization of the documents w.r.t. the SDGs.

5.3.2. Topic Keyword Evaluation

To evaluate the quality of the keywords extracted from the clusters as their topical representation, we first extract the *top keywords* and the *diversified keywords* describing every SDG, using the SDGs own textual description documents published online[3]. We use the latter as the ground-truth reference to compare with. Consequently, we compare each of our cluster keyword representations with the ground-truth SDG keyword representations, using the Jaccard similarity index as a commonly used metric to compare sets of textual tokens.

5.4. Experimental Results

5.4.1. Cluster Evaluation

We ran two cluster evaluation experiments. In Experiment 1, we evaluate clustering quality considering: i) the term frequency feature vector representation, and ii) the augmented co-occurrence feature vector representations. In Experiment 2, we evaluate clustering quality considering as reference for every cluster: i) the primary SDG only, and ii) the primary and the related SDGs together. In the first case, a document is considered to be clustered correctly if it matches the primary SDG only. In the second case, a document is considered to be clustered correctly if it matches the primary SDG or any of the related SDGs. We consider the primary SDG to be the only reference for every cluster in this experiment.

Experiment 1: Results in Table 3 show that augmented co-occurrence frequency vectors improve clustering quality in two of the five subsets, while it maintains the same quality levels for the remaining three subsets. This shows that using augmented co-occurrence scores can either maintain or improve topical organization quality, but does not seem to negatively affect quality. We are currently preparing an extended experimental evaluation using external corpora with varying linear scaling factors (α), to shed more light on the potential of this approach and its impact on topical organization quality.

Table 3: Experiment 1: *Precision*, *recall*, and *f-value* results averaged over all subsets of documents, considering term frequency versus augmented co-occurrence frequency feature vector representations (with α = 3, chosen empirically to emphasize impact of co-occurrence)

	Case1: Considering term frequency vectors			**Case 2:** Considering augmented co-occurrence frequency vectors		
	PR	Recall	F-Value	PR	Recall	F-Value
Subset 1	0.8	0.8	0.8	0.88	1	0.93
Subset 2	1	0.8	0.8889	1	0.8	0.8889
Subset 3	0.78	0.95	0.8566	0.78	0.95	0.8566
Subset 4	0.83	0.75	0.78	0.94	1	0.96
Subset 5	0.76	0.74	0.7499	0.76	0.74	0.7499

Experiment 2: Results in Table 4 show that using primary and related SDGs as cluster reference improves clustering quality in three of the five subsets, while it maintains the same quality levels for the remaining two subsets. This is expected since considering the related SDGs as relevant cluster references (in addition to the primary one) increases the chances of a cluster being mapped correctly to the SDGs, thus improving precision and recall accordingly. More importantly, results considering the primary SDG only are also very

[1] The SDGs are a collection of 17 global goals designed to be a "blueprint to achieve a better and more sustainable future for humanity". They were set in 2015 by the United Nations General Assembly and are intended to be achieved by the year 2030. They are available at: https://sdgs.un.org/goals

promising, reaching f-value levels comprised between 0.74 and 0.89, i.e., almost 90% accuracy in correctly organizing the documents following the relevant SDGs.

Table 4: Experiment 2: *Precision*, *recall*, and *f-value* results averaged over all subsets of documents, considering primary SDG only versus primary and related SDGs

	Case 1: Considering Primary SDG only			**Case 2:** Considering Primary and Related SDGs		
	PR	Recall	F-Value	PR	Recall	F-Value
Subset 1	0.8	0.8	0.8	0.9	0.8	0.8471
Subset 2	1	0.8	0.8889	1	0.8	0.8889
Subset 3	0.78	0.95	0.8566	1	0.95	0.9744
Subset 4	0.83	0.75	0.78	0.83	0.75	0.78
Subset 5	0.76	0.74	0.7499	0.9623	0.8361	0.8947

5.4.2. Topic Keyword Evaluation

As for topic keyword extraction, we evaluate the Jaccard similarity index between the ground-truth SDG keyword descriptions and the system generated cluster keyword descriptions. We compare two approaches: i) top keywords extraction and ii) diversified keyword extraction.

Table 5: Jaccard index results for keyword representations of system generated clusters, compared with ground-truth SDG keyword representations

	Mean Jaccard index for first *k* cluster keywords			
	k=10		**k=20**	
	Top Keywords	Diversified Keywords	Top Keywords	Diversified Keywords
Subset 1	0.5278	0.5293	0.5730	0.5736
Subset 2	0.7692	0.7292	0.7275	0.7303
Subset 3	0.7670	0.6548	0.7681	0.6833
Subset 4	0.4042	0.3247	0.4786	0.4783
Subset 5	0.6864	0.5454	0.5843	0.4961

	k=50		**k=100**	
	Top Keywords	Diversified Keywords	Top Keywords	Diversified Keywords
Subset 1	0.5788	0.5598	0.6203	0.5625
Subset 2	0.7576	0.7325	0.7915	0.7547
Subset 3	0.7519	0.7039	0.7061	0.6982
Subset 4	0.5255	0.5217	0.6126	0.5371
Subset 5	0.6538	0.5467	0.6485	0.5427

Results in Table 5 show that similarity between the ground-truth and the generated keyword descriptions, or both *top keyword* and *diversified keyword* approaches, varies from 0.5278 when considering the first 10 keywords, to 0.7915 (79.15%) when considering the first 100 keywords, such that similarity increases with the increase in number of keywords considered. This shows that our approach is capable of detecting the topical keyword descriptions of the input documents with accuracy levels varying between 52.78% and 79.15%. Results also show that the *diversified keywords* approach tends to perform better when considering a lesser number of keywords (10-and-20), and performs consistently worse than its *top keyword* counterpart when considering more keywords (50-and-100). This might indicate that diversification occurs seamlessly when many keywords are selected as representatives of their clusters (as a result of selecting more keywords which have decreasing weights in describing the cluster, and thus tend to be increasingly different from each other, i.e., more diversified). In other words, there might be a need to run a dedicated process to diversify the results when returning a small number of keyword representatives, but there is apparently no need for such a process when returning a larger number of keyword representatives. We are currently conducting more experiments on larger datasets to further confirm this observation.

6. Conclusion

This paper describes an approach for cluster-based topical document organization using corpus-based text analysis. We consider two methods for document clustering: the first uses legacy term-frequency feature vectors, while the second augments the frequency vectors with term co-occurrence scores generated from a distributional thesaurus. We also consider two methods for topical keyword extraction from the generated clusters: the first identifies the top keywords with the highest feature vector weights, while the second perform a second round of clustering within the previously generated clusters to produce more diversified keyword representations based on the produced sub-clusters. Experiments on a manually labelled dataset of 158 documents classified against the UN's Sustainable Development Goals (SDGs) confirm the quality and potential of the approach and its variant methods. We are currently conducting further experiments considering external corpora as references for the distributional thesaurus, and larger document datasets to further evaluate our approach. We are also investigating the interplay between keyword diversification and quality [23], which in general tends to be antinomic [24], i.e., the improvement of one of them usually results in a degradation of the other. Too much diversification may result in losing relevant keywords while increasing relevance only tends to provide many near duplicates [24]. We also aim to consider semantic-aware indexing capability [35-37], providing more opportunities toward knowledge-based topical extraction and organization. In this context, we aim to extend the approach by integrating a human tailored knowledge base such as the United Nations UNBIS thesaurus[1], and evaluate the quality of corpus-based versus knowledge-based feature vector augmentation, and their impact of topical keyword organization and extraction.

Acknowledgements

We would like to thank Dr. Tarcisio Alvarez, Chief, Programme Planning and Technical Cooperation Section (PPTCS), United Nations Economic and Social Commission for Western Asia (ESCWA), for contributing to the launching of this work, and for his support to complete the project.

REFERENCES

[1] Rakesh Agrawal, Christos Faloutsos, and Arun N. Swami, *Efficient Similarity Search in Sequence Databases*, 1993. International Conference on the Foundations of Data Organization and Algorithms (FODO), pp. 69-165

[2] Amir Ahmad and Shehroz Khan, 2019. *Survey of State-of-the-Art Mixed Data Clustering Algorithms.* IEEE Access. 7: 31883-31902.

[1] Available at: http://metadata.un.org/thesaurus/

[3] Gianni Amati and C. J. Van Rijsbergen, 2002. *Probabilistic models of information retrieval based on measuring the divergence from randomness.* ACM Transactions on Information Systems (TOIS). 20(4): 357-389.

[4] Bogdan Boteanu, Ionut Mironica, and Bogdan Ionescu, 2015. *Hierarchical Clustering Pseudo-Relevance Feedback for Social Image Search Result Diversification*. International Conference on Content-Based Multimedia Indexing (CBMI'15), pp. 1-6.

[5] Mohand Boughanem, 2006. *Introduction to Information Retrieval.* Proceedings of EARIA'06 (Ecole d'Automne en Recherche d'Information et Application), Ch. 1.

[6] Hiram Calvo, Alexander Gelbukh, and Adam Kilgarriff, 2005. *Distributional Thesaurus Versus WordNet: A Comparison of Backoff Techniques for Unsupervised PP Attachment.* International Conference on Computational Linguistics and NLP (CICLing) pp. 177-188.

[7] Silvana Castano, Alfio Ferrara, and Stefano Montanelli, 2011. *Thematic Exploration of Linked Data.* International Workshop on Very Large Data Search (VLDS), pp. 11-16.

[8] Silvana Castano, Alfio Ferrara, and Stefano Montanelli, 2012. *Structured Sata Clouding across Multiple Webs.* Information Systems. 37(4): 352-371.

[9] Silvana Castano, Alfio Ferrara, and Stefano Montanelli, 2014. *inWalk: Interactive and Thematic Walks inside the Web of Data*. International Conference on Extended DataBase Technology (EDBT'14), pp. 628-631.

[10] Silvana Castano, Alfio Ferrara, and Stefano Montanelli, 2017. *Exploratory Analysis of Textual Data Streams.* Future Generation Computer Systems. 68: 391-406.

[11] Silvana Castano, Alfio Ferrara, and Stefano Montanelli, 2018. *Topic Summary Views for Exploration of Large Scholarly Datasets.* Journal of Data Semantics. 7(3): 155-170.

[12] Theodore Dalamagas, Tao Cheng, Klaas-Jan Winkel, and Timos Sellis, 2006. *A Methodology for Clustering XML Documents by Structure.* Information Systems. 31(3):187-228.

[13] Mark Davies, *The Corpus of Contemporary American English as the first reliable monitor corpus of English.* Literary & Linguistic Computing, 2010. 25(4): 447-464.

[14] Scott Deerwester, Susan Dumais, and Thomas Landauer, 1990. *Indexing by Latent Semantic Analysis.* Journal of the American Society for Information Science. 41(6):391–407.

[15] Bernard Desgraupes, 2017. *Clustering Indices - Package clusterCrit for R.* University Paris Ouest, Lab Modal'X, 33 p.

[16] Alfio Ferrara, Lorenzo Genta, Stephano Montanelli, and Silvana Castano, 2015. *Dimensional Clustering of Linked Data: Techniques and Applications.* Transactions on Large Scale Data and Knowledge Centered Systems. 19: 55-86

[17] Nelson Francis and Henry Kucera, 1982. *Frequency Analysis of English Usage*. Houghton Mifflin, Boston.

[18] Norbert Fuhr, *Probabilistic Models in Information Retrieval*. 1992. The Computer Journal. 35 (3):243-255.

[19] J. C. Gower and G. J. S. Ross, 1969. *Minimum Spanning Trees and Single Linkage Cluster Analysis.* Applied Statistics, 18. pp. 54-64.

[20] Maria Halkidi, Yannis Batistakis, Michalis Vazirgiannis, 2001. *Clustering Algorithms and Validity Measures.* International Conference on Scientific and Statistical Database Management, 3-22.

[21] Ramzi Haraty R. and Mazen Hamdoun, 2002. *Iterative Querying in Web-based Database Applications.* ACM Symposium on Applied Computing (SAC), 458-462.

[22] Ramzi Haraty, Nashat Mansour, and Walid Daher, 2003. *An Arabic Auto-indexing System for Information Retrieval.* Applied Informatics, pp. 1221-1226.

[23] Bogdan Ionescu, Adrian Popescu, Mihai Lupu, Alexandru-Lucian Gînsca, Bogdan Boteanu, Henning Müller, 2015. *Div150Cred: A Social Image Retrieval Result Diversification with User Tagging Credibility Dataset.* ACM Multimedia Systems (MMSys), pp. 207-212.

[24] Bogdan Ionescu, Adrian Popescu, Anca-Livia Radu, Henning Müller, 2014. *Result Diversification in Social Image Retrieval: A Benchmarking Framework.* Multimedia Tools and Applications (MTAP), pp. 1–31.

[25] Joon Ho Lee, 1994. *Properties of Extended Boolean Models in Information Retrieval.* International ACM SIGIR Conference, Springer-Verlag, pp.182-190.

[26] Nashat Mansour, Ramzi A. Haraty, Walid Daher, Manal Houri, 2008. *An Auto-Indexing Method for Arabic Text.* Information Processing and Management journal, 44(4):1538-1545.

[27] Michael McGill, 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 400 p.

[28] George Miller, Christiane Fellbaum, 2007. *WordNet Then and Now*. Language Resources and Evaluation. 41(2): 209-214.

[29] J.C. van Rijsbergen, 1079. *Information Retrieval*. Butterworths, London, 208 p.

[30] Nick Roussopoulos, Stephen Kelley, Frédéic Vincent, 1995. *Nearest Neighbor Queries.* Proceedings of the ACM International Conference on Management of Data (SIGMOD), pp. 71-79.

[31] Gerard Salton, 1971. *The SMART Retrieval System.* Prentice Hall, N.J., 556 p.

[32] Gerard Salton and Chris Buckley, 1988. *Term-weighting Approaches in Automatic Text Retrieval.* Information Processing and Management. 24(5):513 -523.

[33] Gerard Salton and Michael Mcgill, 1983. *Introduction to Modern Information Retrieval.* McGraw-Hill, Tokio, 400 p.

[34] Jimmy Tekli, Bechara al Bouna, Youssef Bou Issa, Marc Kamradt, Ramzi A. Haraty, 2018. *(k, l)-Clustering for Transactional Data Streams Anonymization*. Information Security Practice and Experience. pp. 544-556.

[35] Richard Chbeir, Yi Luo, Joe Tekli, Kokou Yétongnon, Carlos Raymundo Ibañez, Agma J. M. Traina, Caetano Traina Jr., and Marc Al Assad, 2014. *SemIndex: Semantic-Aware Inverted Index.* Symposium on Advances in Databases and Information Systems (ADBIS), pp. 290-307.

[36] Joe Tekli, Richard Chbeir, Agma J. M. Traina, and Caetano Traina Jr., 2019. *SemIndex+: A Semantic Indexing Scheme for Structured, Unstructured, and Partly Structured Data*. Knowledge-Based Systems. 164: 378-403.

[37] Joe Tekli, Richard Chbeir, Agma J. M. Traina, Caetano Traina, Kokou Yétongnon, Carlos Raymundo Ibañez, Marc Al Assad, and Christian Kallas, 2018. *Full-fledged Semantic Indexing and Querying Model Designed for Seamless Integration in Legacy RDBMS*. Data and Knowledge Engineering, 117: 133-173.

[38] Joe Tekli, Richard Chbeir, and Kokou Yétongnon., *Structural Similarity Evaluation between XML Documents and DTDs.* Inter. Conf. on Web Information Systems Engineering (WISE), 2007, 196-211.

[39] Julie Weeds, David J. Weir, Diana McCarthy, 2004. *Characterising Measures of Lexical Distributional Similarity*. Int. Conf. on Comput. Linguistics (COLING), Article No. 1015.

[40] Peter Willett, 2006. *The Porter Stemming Algorithm: Then and Now*. Program. 40(3): 219-223.