Quantifying the hierarchical adherence of modular documents

IOP Publishing J. Phys. Complex. 4 (2023) 045008 https://doi.org/10.1088/2632-072X/ad0a9b

Journal of Physics: Complexity

OPEN ACCESS

RECEIVED
8 May 2023

REVISED
31 October 2023

ACCEPTED FOR PUBLICATION
8 November 2023

PUBLISHED
23 November 2023

PAPER

Quantifying the hierarchical adherence of modular documents

Alexandre Benatti[1,*], **Ana C M Brito**[2], **Diego R Amancio**[2] and **Luciano da F Costa**[1]

[1] São Carlos Institute of Physics, Av. Trabalhador São-Carlense, 400, São Carlos, SP 13566-590, Brazil
[2] Department of Computer Science, Institute of Mathematics and Computer Science, University of São Paulo, São Carlos, SP, Brazil
[*] Author to whom any correspondence should be addressed.

E-mail: alexandre.benatti@usp.br

Keywords: hierarchical organization, coincidence similarity index, hierarchical adherence index, modular documents, knowledge networks

Abstract

Several natural and artificial structures are characterized by an intrinsic hierarchical organization. The present work describes a methodology for quantifying the degree of adherence between a given hierarchical template and a respective modular document (e.g. books or homepages with content organized into modules) organized as a respective content network. The original document, which in the case of the present work concerns Wikipedia pages, is transformed into a respective content network by first dividing the document into parts or modules. Then, the contents (words) of each pair of modules are compared in terms of the coincidence similarity index, yielding a respective weight. The adherence between the hierarchical template and the content network can then be measured by considering the coincidence similarity between the respective adjacency matrices, leading to the respective hierarchical adherence index. In order to provide additional information about this adherence, four specific indices are also proposed, quantifying the number of links between non-adjacent levels, links between nodes in the same level, converging links between adjacent levels, and missing links. The potential of the approach is illustrated respectively to model-theoretical networks as well as to real-world data obtained from Wikipedia. In addition to confirming the effectiveness of the suggested concepts and methods, the results suggest that real-world documents do not tend to substantially adhere to respective hierarchical templates.

1. Introduction

Several real-world and theoretical structures present an intrinsically *hierarchical* organization. For instance, the systematic description of concepts in a book, encyclopedia, or web page, involves several levels progressing from more general to more specific concepts (or vice-versa). For simplicity's sake, the present work focuses on the former type of structures.

A similar organization is often observed in taxonomic systems, such as the organization of living beings, molecules, etc. Actually, the own scientific method often leads to hierarchical models (e.g. [1]) and fractal models (e.g. [2]).

Hierarchical systems can be generally represented and modeled in terms of several modules organized as a respective *tree*, as illustrated in figure 1.

The tree in the example has $H = 3$ hierarchical levels indicated as $h = 1, 2, 3$. The number of modules (nodes) within each of these levels is represented as m_h so that, for this specific example we have $m_1 = 1, m_2 = 3$, and $m_3 = 5$.

Hierarchical structures can be expected to present a basic property, namely the existence of semantic overlap reflected in the content possibly shared by modules that are at adjacent or nearby hierarchies [1]. This is so because, at each successive level in a specifying hierarchy, it becomes necessary to reference concepts in modules at both previous and successive levels. Thus, in a sense, the modules within each hierarchical level can be understood as *bridges* with the preceding and succeeding levels. Though these overlaps may eventually extend along further away levels, the present work assumes only content shared between immediately adjacent levels.

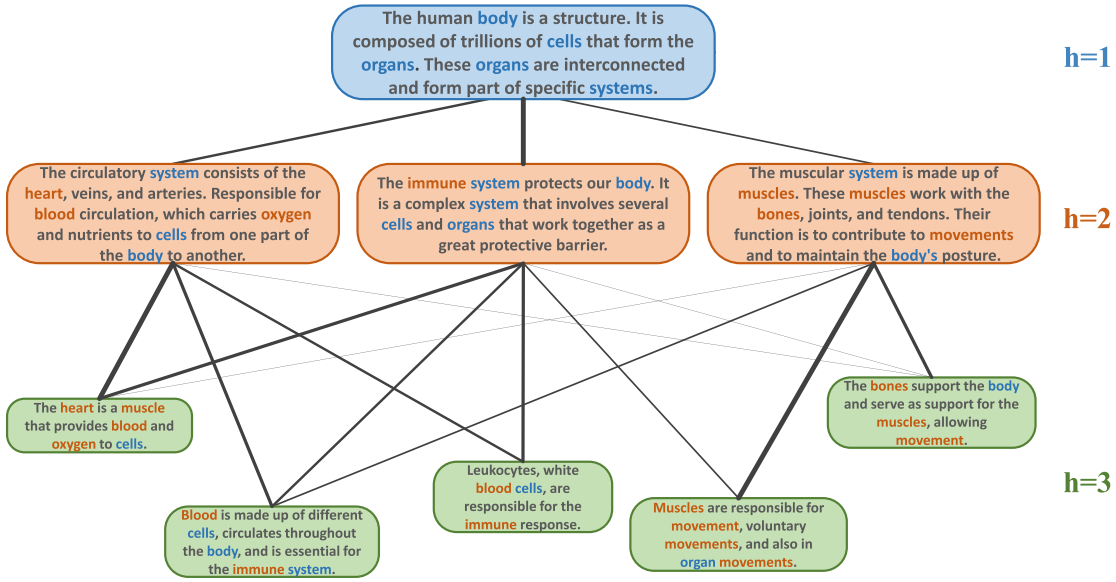

Figure 1. A case example of hierarchical structure considering, in a very simplified manner, properties of vertebrate living beings, including three hierarchical levels. The words shared between the intermediate level and the upper level are shown in blue, while the words common to the intermediate and lower level are presented in red.

Several interesting problems are defined respectively to the study of hierarchies. In the present work, we focus on the situation in which, given a *modular document*, i.e. a document organized into hierarchical modules (e.g. sections, subsections, etc), as well as a respective hierarchical template, one needs to quantify how much its hierarchical structure reflects the shared contents between adjacent modules. Though the original modular document does not need to have a hierarchical structure, full correspondence is required between the modules of the document and the template. More specifically, we aim at developing concepts and methods that can be applied in order to assign an overall index quantifying the degree of hierarchy in a given system.

These issues are interesting and important because hierarchical systems can differ in the levels of adherence between the modules content and the hierarchical levels, reflecting specific demands or characteristics. For instance, given a book organized into chapters, sections, subsections, etc it becomes interesting to have a quantification of how much the content of these modules adheres to the overall hierarchical template (e.g. respective systematic table of contents, with sections and subsections). Another possibility consists in having a set of independent pieces of knowledge (e.g. scientific articles or entries in an Encyclopedia), all of them related to some general topic, as well as a candidate hierarchical template. The adherence between these two structures can again be quantified to estimate the coherence between the content modules and the provided hierarchical template.

In order to supply additional information about this type of adherence, four specific indices are also proposed, quantifying the number of links between non-adjacent levels, links between nodes in the same level, converging links between adjacent levels, and missing links.

In summary, we approach the above problems by considering a *hierarchical template*, providing the reference, as well as a *document*, translated into a respective *content network*, whose content is to have its adherence to the template quantified in terms of an overall index.

Though a substantial number of approaches related to hierarchy have been reported in the literature (see section 2 for a review of some of the main related works), the specific problem of quantifying how much a document adheres to a given hierarchical structure seems to have received less attention. One possible related work has been reported in [3], which describes approaches for extracting hierarchies as well as for the quantification of the similarity between two hierarchies. The latter is done by taking into account the relative ratio (concerning one of the two provided hierarchies) of the number of parent-child pairs shared between the two hierarchies. The Jaccard index has also been used to quantify the similarity between ground truths and estimated hierarchies, such as in [4] aimed at inferring the hierarchy (classes) of binary documents (stripped binary pieces of code).

In the present work, we approach the interesting subject of quantifying the adherence between a modular document and a given hierarchical template by using the coincidence similarity index (e.g. [5, 6]) as a means not only for translating a modular document into a respective content network, but also for quantifying its

similarity between the hierarchical template while allowing not only for binary interconnections but also non-negative weighted links between the hierarchical modules. In addition, the coincidence similarity index has been shown to frequently implement more strict similarity comparisons than the cosine similarity and the Pearson correlation coefficient.

The potential of the suggested concepts and method is illustrated respectively in three main cases: (i) both the hierarchical template and the content network are model-theoretic; (ii) real-world hierarchical template and model theoretical content network; and (iii) real-world hierarchical template and content network. Several interesting results are presented and discussed, including the verification of the ability of the proposed method to quantify hierarchical adherence and the possibility that real-world documents do not tend to present substantial adherence to respective hierarchical templates.

This work is organized as follows. First, some of the main related works are briefly described. The section *Data and Basic Concepts and Methods* then presents the concepts and methods used in this work, including the adopted data, as well as the coincidence similarity index and methodology for translating data into networks. The methodological approach suggested in the current work is then presented in the section *Methodology*, which includes the overall flow diagram, the hierarchical adherence index, specific indices, as well as the approach adopted for synthesizing model-theoretical content networks. The results and discussion are presented next, followed by conclusions and suggestions for further related research.

2. Related works

This section presents a non-exhaustive review of previous approaches related to the current work.

Humans tend to frequently categorize objects and concepts into different groups. This process is related to our necessity to organize data and knowledge. Not surprisingly, data clustering algorithms have become a well-known researched field of study in Computer Science. This task can be generalized to every data type, from abstract to concrete objects, including transportation systems, social networks, as well as text data. When representing complex systems using networks (or graphs), nodes usually can be grouped into clusters by considering modular patterns or interconnection, known as communities [7]. Examples of text modeled as graphs include co-occurrence, mesoscopic, and enriched networks [8–12].

For text data, one way to organize concepts is by using *concept hierarchies* (e.g. [13]). This representation of knowledge consists of trees where the nodes correspond to concepts organized into levels of hierarchy. A class of text data usually arranged in a systematic way consists of textbooks. The extraction of concept hierarchies from textbooks constitutes the object of study in [14–16]. Both [14, 16] aim at extracting the hierarchical structure concepts of textbooks taking Wikipedia as a reference. A study of prerequisite relationships among intra- and inter- institutions graduation courses has been described in [15], yielding a type of concept hierarchy. Wikipedia was again used as a reference to construct a representative scheme of the concept space.

A hierarchical topical characterization of scientific subjects was proposed in [17]. The authors described a method to summarize a scientific field via the representation of citation networks [18]. Given a set of papers in a subfield, two papers were linked whenever they shared a citation link. This representation allowed a better understanding of the field since papers are highly clustered in subtopics, as revealed by network community analysis. The topics most relevant to each network community were generated via an approach similar to tf-idf, and the similarity of topics was then computed based on the average distance in the citation network of the papers comprising the words of interest. The hierarchy was then estimated by using an agglomerative clustering approach. The results observed for the field of complex networks revealed that both the clustering and hierarchy of topics are consistent and accurate.

The most frequent strategies used to validate the hierarchy identification methods can be divided into two main approaches: (i) a validation based on traditional metrics such as precision and recall [14, 15, 19, 20]; and (ii) a subjective qualitative validation consisting of an inspection of the case studies [16, 21, 22].

In [21], Liang *et al* predicted the prerequisite relation among concepts (if a concept is a prerequisite of another concept), and used precision and recall to quantify the quality of the proposed method. They also showed the case study of the concept 'deep learning', to inspect the set of concepts identified as prerequisites ('deep learning' is a prerequisite to 'feature learning' and 'artificial intelligence' a prerequisite to 'deep learning').

In [16], Wang *et al* extracted concepts using precision@n ($n = 1, 3, 5$) to compare their method with others available in the literature. To illustrate the concept hierarchies extracted by using their method, they showed the case study of 'computer network'. In [22], the task was developed in two parts concerning: (i) identifying concepts; and (ii) organizing them into a concept hierarchy. The results from the first part of their approach were quantified using precision, recall, and F-measure, and the results obtained in the second part

were inspected concerning two specific hierarchies, namely in the domains of tourism and biology. This inspection involved comparing the obtained hierarchy with a respective gold standard.

The article [23] describes developments that aim to simplify and enhance hierarchies by breaking respective cycles. To remove cycles from a given hierarchical structure, three criteria are employed: forward, backward, and greedy.

One approach to obtaining ground truths to be considered in validations consists in having experts to prepare them manually [19, 24, 25]. Book summaries have also been adopted as a reference to be taken as subsidy for validating the hierarchy of documents [14–16].

Similarity indices, such as the Jaccard index, have been utilized to estimate hierarchical relationships among entities or components within structures and systems. For example, [26] apply the Jaccard index to estimate the similarity between text attributes. They used the analytic hierarchy process (AHP) to assign respective weights. In [27], the authors described the application of similarity-based approaches, including new tree-based similarities, to study ontologies and hierarchical and graph relationships.

Hierarchical systems are used in many fields to classify and organize information in a logical and structured way. In linguistics, for example, hierarchies can be considered in order to better understand the relationships between sounds, words, and sentences [28]. In grammar, a sentence is typically composed of smaller units such as *noun phrases* and *verb phrases*, which are themselves composed of even smaller units such as *nouns, verbs, adjectives*, and so on [29]. Other areas of linguistics where the concept of hierarchy is important are phonetics, where the sounds of language can be organized into a hierarchy of phonemes, syllables, and semantics, where the meanings of words and sentences can be organized into a hierarchy of concepts, such as suggested by works on speech perception [30, 31].

In biology, hierarchies are commonly used to classify and organize living beings. In a hierarchical classification approach, organisms are grouped into successively larger categories, starting with individual species and moving up to broader groups like genera, families, orders, classes, phyla, and kingdoms [32]. This approach allows biologists to better understand the relationships between different organisms and to study their characteristics and behaviors at different levels of the hierarchy. In [33], the authors studied the evolution of hierarchical complexity in biological systems.

In neuroscience [34], hierarchies of brain function are used to understand how its different parts are involved in specific mental processes. Another study [35] shows that memory can be understood as a hierarchical structure in layers, where each layer represents a level of information processing, stored according to different time scales. The visual system can also be understood as being underlain by a hierarchical organization [36] that enables the brain to efficiently process and interpret visual information, as well as learn and adapt based on experience.

3. Data and basic concepts and methods

This section describes the adopted data and concepts and methods, including the coincidence similarity index and the respective method for translating datasets into networks.

3.1. Hierarchical templates and modular documents

Data pre-processing starts with the selection of a modular document to be analyzed. The documents can be of various types, such as web pages, books, or surveys. Though the original document may already have a hierarchical structure, e.g. identified as sections and subsections in a respective table of contents, it is also possible that the original document, though modular, has no hierarchy formally specified. In this case, the approach suggested in the present work can still be applied provided there is a one-to-one correspondence between the document modules and a provided hierarchical template.

Each module is determined by the structure of the document. For example, a survey document is structured hierarchically, with sections and subsections and each section will represent a module. In the case of a WWW document related to a general topic, the respectively linked pages can be considered modules. The present work considers this type of representation respectively to Wikipedia pages describing aspects of general interest about Brazil and the U.S.A. (e.g. geography, economy, demography, etc), as of 13 September 2022.

Having identified the modules from the original document, some steps must then be applied. We first remove the *stop words* (such as articles and prepositions), because they usually are not particularly specific (e.g. [37]) and can lead to highly connected nodes in respective network representations. In sequence, the text was tokenized into words. In other words, in this work, the texts are converted into linguistic units (*tokens*). Then, each of the identified modules is represented as a respective multiset containing the frequency of each token in that module.

The hierarchical template can be derived from the own document, but it is also possible that it has been provided independently, e.g. corresponding to an expected or candidate organization. The templates in the present work are of the former type, i.e. being from the original documents. In mathematical terms, we can describe this template as an adjacency matrix A_t corresponding to the hierarchical organization originally contained in the document (e.g. tables of contents and/or sections/subsections).

3.2. The coincidence similarity index

As described above, the modules extracted from the original document need to be interconnected according to their content similarity. In the present work, we employ the *coincidence similarity index* [5, 6, 38] for that finality, in order to perform more selective and sensitive similarity quantification.

The *Jaccard* index ($\mathcal{J}$) (e.g. [39–43]) is an approach that seeks to quantify the level of similarity between two non-empty sets A and B [5]. The basic *Jaccard index* can be expressed by the following equation:

$$\mathcal{J}(A,B) = \frac{|A \cap B|}{|A \cup B|}, \tag{1}$$

where A and B are any two non-empty sets and $||$ stands for the cardinality. We obtain $\mathcal{J}(A,B) = 1$ whenever $A = B$, and $\mathcal{J}(A,B) = 0$ in case of $|A \cap B| = \emptyset$ (where $\emptyset$ is the empty set).

The multiset version of the Jaccard index (e.g. [42]) between two non-negative and non-zero (with at least one element different from zero) multisets or vectors $\vec{x} = [x_1\ x_2 \ldots x_N]^T \neq \vec{0}$ and $\vec{y} = [y_1\ y_2 \ldots y_N]^T \neq \vec{0}$ can be expressed as:

$$\mathcal{J}(\vec{x},\vec{y}) = \frac{\sum_i \min\{x_i, y_i\}}{\sum_i \max\{x_i, y_i\}}. \tag{2}$$

The *Interiority* index ($\mathcal{I}$, also called overlap index [44]) quantifies how much a given set A is relatively interior to another set B [5]. Therefore, this index should return a minimum value when the two sets are disjoint ($A \cap B = \emptyset$). On the other hand, this index should return a maximum value when one set is contained in the other (i.e. $A \subset B$). The *Interiority* index can be expressed as follows:

$$\mathcal{I}(A,B) = \frac{|A \cap B|}{\min\{|A|, |B|\}}. \tag{3}$$

Both sets are assumed to be non-empty.

The multiset generalization of *Interiority* between two non-negative and non-zero multisets or vectors $\vec{x} \neq \vec{0}$ and $\vec{y} \neq \vec{0}$, can be expressed as:

$$\mathcal{I}(\vec{x},\vec{y}) = \frac{\sum_i \min\{x_i, y_i\}}{\min\{\sum_i x_i, \sum_i y_i\}}. \tag{4}$$

In order to integrate the complementary characteristics of the two previous indices, the *Coincidence similarity index* ($\mathcal{C}$) has been described in [5, 38, 45], corresponding to the product between the *Jaccard* (equations (1) and (2)) and *Interiority* (equation (3) and (4)) indexes, i.e.:

$$\mathcal{C}(\vec{x},\vec{y}) = \mathcal{J}(\vec{x},\vec{y})\ \mathcal{I}(\vec{x},\vec{y}). \tag{5}$$

As discussed in [5, 6], the coincidence index has been found to allow enhanced selectivity, sensitivity, and robustness as compared to alternative approaches such as the cosine similarity and the Pearson correlation coefficient. This index has been suggested as a way to implement strict similarity comparisons between non-null generic mathematical structures, including sets, multisets, vectors, functions, matrices, among other possibilities [45, 46].

3.3. Translating modules into content networks

Once the original document has been divided into its respective modules, and the text in these modules has been pre-processed as described in section 3.1, it becomes necessary to obtain a content network representing the similarity among the contents of the obtained modules, which is henceforth referred to as *content network*.

In this work, this is done by understanding each module as a respective multiset. Each multiset carries the information not only about the words contained in each module but also the frequency in which each word appears [47].

More specifically, similarly to the work [48], in our approach, each word in a module is taken as an element of the respectively associated multiset, while the number of repetitions of that word is understood as

its *multiplicity*. The quantification of the similarity of the content between each pair of modules can then be estimated by calculating the coincidence similarity index (equation (5)) between those two multisets (see section 3.2).

The similarity values between text modules can then be organized as a weight matrix W_c. In this way, a weighted similarity network (henceforth called the *content network*), is obtained with its nodes corresponding to the document modules, while the links indicate the similarity between the contents of the respective modules.

Now, the hierarchical organization of the obtained content network can be compared to the considered hierarchical template, yielding a respective overall index $\mathcal{H}$ (see equation (6)) quantifying the adherence between the two structures, as described in the following section.

4. Methodology

Figure 2 illustrates the diagram of the methodology described in the present work as a means of quantifying the level of adherence between a given hierarchical template and a specific respective document, represented as a content network.

The procedure described in this paper can be summarized as follows:

(i) *Document into modules*: The hierarchical template is assumed to be provided. In general, it can be derived from a respective index or table of content. However, it is also possible to consider a candidate hierarchical template that is expected to be compatible with some specific documents. The documents can refer to books, homepages, or surveys. The given document needs to be translated into a respective document (or content) network. In the present work, this is implemented by first dividing the document into parts or modules.

(ii) *From modules to content network*: Each of the obtained modules is understood as a node of the respective content network. The links between these nodes are then determined as having weights corresponding to the coincidence similarity between the content (words) of each pair of modules. This procedure results in the respective content network.

(iii) *Adhrence between hierarchical template and content network*: At this stage, the obtained content network is compared, via the coincidence similarity index, with a provided hierarchical template, both being represented in terms of the respective weight or adjacency matrices. This operation results in the respective hierarchical adherence index.

The operations involved in the above steps are presented in more detail in the subsequent sections.

Table 1 summarizes the main symbols and abbreviations related to the suggested method for quantification of the hierarchical adherence of documents.

4.1. The hierarchical adherence index

In the present work, we propose a *hierarchical adherence index* ($\mathcal{H}$) in equation (6) to quantify how much a given document adheres to a respectively supplied reference hierarchical scheme.

The hierarchical template is henceforth represented in terms of its respective weight matrix $W_t = [w_{i,j}]$, with each element $w_{i,j}$ being comprised in the interval $(0, 1)$. These weights can be provided *a priori*, corresponding to the expected shared contents between each pair of connected nodes. However, in the case no such weight values are available, it is possible to assume that $w_{i,j} = 0.5$, which is the approach henceforth adopted in this work.

Similarly, the content network can be represented in terms of its weights (each of them in the interval $(0, 1)$), yielding:

$$\mathcal{H} = \mathcal{C}\left(W_c, W_t\right), \tag{6}$$

with $0 \leqslant \mathcal{H} \leqslant 1$ in both cases. The higher this index, the greatest the adherence of the document content to the reference hierarchical scheme. The maximum adherence is observed when the two adjacency matrices are equal.

4.2. Specific adherence indicators

Given a hierarchical template and a content network, the above-presented methodology can be used to assign an overall hierarchical adherence index quantifying the coherence between the two given structures. However, in case an index smaller than one is obtained, it is not possible to know the reasons for the respective lack of adherence. In the present section, we provide four additional indices that can provide

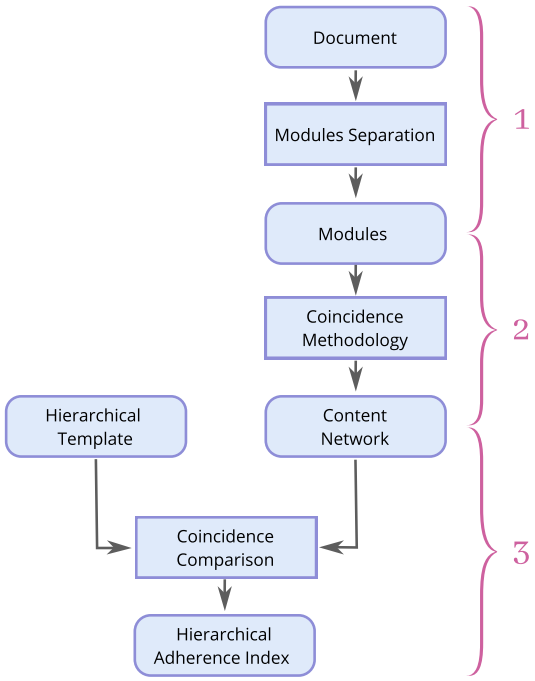

Figure 2. Diagram describing the main data and operations performed as part of the proposed methodology for quantifying hierarchical adherence between an original template and a network derived from a respective content network. See the main text for additional information.

Table 1. List of main symbols and abbreviations related to the adherence concept and methods described in the present work.

Index	Description
H	Number of hierarchical levels
h	Level index ($h = 1, \ldots H$)
m_h	Number of modules per level h
$n_{i,h}$	Number of elements in each module i in level h
N	Total number of possible textual elements
a_j	A given textual element among N possibilities
$r_{i,j,h}$	Number of occurrence of the textual element a_j in module i in level h
$\mathcal{H}$	Hierarchical Adherence Index

specific additional indication of four main reasons implying the overall hierarchical adherence to be smaller than its maximum value of one.

Figure 3 illustrates the main types of possible causes leading to a lack of hierarchical adherence, which include: (A) interconnections between non-adjacent levels; (B) interconnections between nodes in the same hierarchy; (C) converging connections between successive adjacent levels, and (D) missing interconnections. Each of these four problems can be quantified in terms of respective specific indices ϵ_A, ϵ_B, ϵ_C, and ϵ_D.

These indices are henceforth understood to be calculated from a threshold version of the obtained coincidence similarity network, therefore yielding a respective adjacency matrix A containing only 0 or 1 as entries. The four suggested indices are presented as follows:

- ϵ_A: can be obtained by considering all observed connections between nodes from non-adjacent levels, divided by the maximum number of possible such connections.
- ϵ_B: is defined as the number of interconnections between nodes in the same level, divided by the maximum possible number of this type of link.
- ϵ_C: is computed using the content network adjacency matrix A. More specifically, the index ϵ_C is calculated by dividing the number of unexpected links between adjacent levels by the maximum possible number of such incorrect connections.
- ϵ_D: is obtained by dividing the number of missing links, which is calculated from the matrix A and the hierarchical template, by the total number of expected links specified by the latter structure.

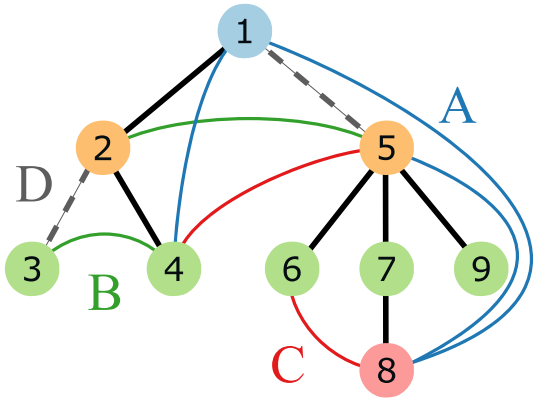

Figure 3. The four main types of effects contributing to the lack of hierarchical adherence: (A) interconnections between non-adjacent levels; (B) interconnections between nodes in the same hierarchy; (C) converging connections between successive adjacent levels, and (D) missing interconnections. Each of the specific effects can be quantified by respective indicators ϵ_A, ϵ_B, ϵ_C, and ϵ_D.

Table 2. The four specific indices obtained for the example network in figure 3.

Indices	Values
ϵ_A	37.50%
ϵ_B	18.18%
ϵ_C	22.22%
ϵ_D	25.00%

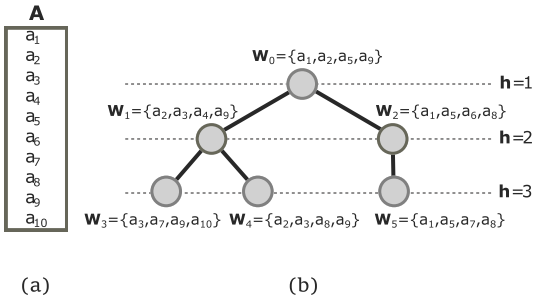

Figure 4. Example of generating a model-theoretic content network from a specific list of elements (a) and hierarchical template shown as the tree in (b). The resulting content network, shown in (b), has $n = 4$ elements within each of its modules.

In the case of the previous example (figure 3), we would have the specificity indices shown in table 2.

Observe that the above-described set of four specific indicators provides complementary information about specific aspects leading to the lack of hierarchical adherence. For instance, in the case of the above example, we would conclude that the main sources of adherence errors are interconnections between non-adjacent levels (37.5%), as well as missing links (25.0%).

4.3. Synthesizing hierarchical networks

In order to test and validate the proposed approach for quantifying the hierarchical adherence, in the present work we resource to three cases: (i) hierarchical and template content networks generated by a model-theoretic approach; and (ii) real-world templates and model-theoretical content networks, and (iii) real-world content networks about general aspects of countries (Brazil and U.S.A.) and respective templates obtained from the Wikipedia.

In the present section, we describe how the model-theoretical content networks have been obtained, based on [48].

The proposed approach for obtaining model-theoretic content networks starts with a given *hierarchical template*, represented as a tree. In the present work, the hierarchies adopted for the models correspond to the hierarchical templates of the real-world documents to be studied in section 5.3.

Figure 4 illustrates the structure of a model-theoretic content network respective to a specific hierarchical template containing $H = 3$ hierarchical levels, and $N = 10$ possible textual elements. Each node in this

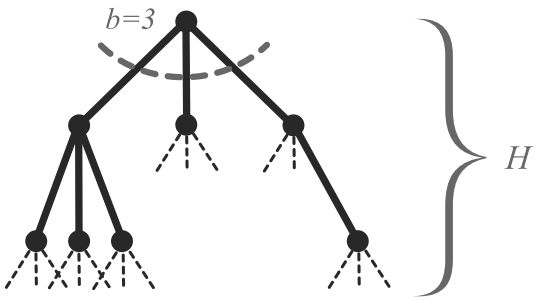

Figure 5. Example of a full tree (in this case with $b = 3$) to be considered as model-theoretical hierarchical templates in the present work.

network corresponds to a respective module, containing a set of words W_i, allowing for repetitions. In the case of this specific example, we have $n = 4$.

A set of all possible elements of a text $\mathbf{A}$ is specified as $\mathbf{A} = [a_1, a_2 \ldots a_N]$ (see figure 4(a)). The number of elements n in each of the modules, taking into account repetitions, also needs to be specified.

We define the parameter β (with $0 \leqslant \beta \leqslant 1$), which determines the mixture of elements between adjacent modules.

The elements in each module will be included while taking into account the provided hierarchical template. The first module ($\mathbf{W}_0$) will receive n (with $n < N$) aleatory elements randomly chosen from $\mathbf{A}$. The other modules ($\mathbf{W}_1$, $\mathbf{W}_2$, …) will receive n elements, being $(1 - \beta)n$ elements drawn from $\mathbf{A}$ and βn elements drawn from its parent, in other words, the predecessor adjacent module in the hierarchy.

The thus obtained model-theoretical content networks, as illustrated in figure 4, can therefore be used for testing, illustrating, and evaluating the approach for quantifying the hierarchical adherence between a provided hierarchical template and document networks.

In particular, we will consider the case of *full* hierarchies as illustrated in figure 5, containing H hierarchical levels and having each node connect to each of the b nodes at the following adjacent level. This particular example assumes $b = 3$.

5. Results and discussion

This section illustrates the application of the proposed concepts and methods respectively to the following three main approaches: (a) both hierarchical template and content networks are model-theoretical; (b) real-world hierarchical template and model-theoretical content networks; and (c) both hierarchical template and content networks are real-world.

5.1. Full model-theoretical case

In this case, we have both the hierarchical template and content network model-theoretical, obtained by using the methods described in section 4.3.

Two situations will be taken into account: one in which all the hierarchical levels have the same number of elements, and another in which this number increases up to the intermediate level, and then decreases, as illustrated in figure 6.

Figure 7 illustrates the hierarchical adherence indices, in terms of the mixture parameter β, obtained hierarchical template with $H = 3$ and $b = 3$, and 100 instances for each β of respective content networks with $N = 100$ and $n = 10$.

The obtained results indicate a plateau of high matching values (extending approximately from 0.3 to 0.8) between the hierarchical template and the respective content networks. In addition, the adherence tended to decrease markedly for $\beta < 0.3$.

Figure 8 illustrates the hierarchical adherence indices obtained for the same configuration as before, but using $H = 5$ hierarchical levels. Substantially smaller values of hierarchical adherence are obtained for this configuration, which is a consequence of the considered structures involving more levels, and therefore, an higher number of interconnected nodes.

Respectively to figure 7, it can also be observed that the hierarchical adherence curve shifted to the right-hand side of the plot, with the maximum correspondence between template and content networks being observed for larger values of β, implying more elements to be drawn from the parent of each module than from $\mathbf{A}$. The larger number of hierarchical levels in this example ($H = 5$) implies that more elements need to be inherited from the previous levels (parent) in order to preserve an adequate heterogeneity of

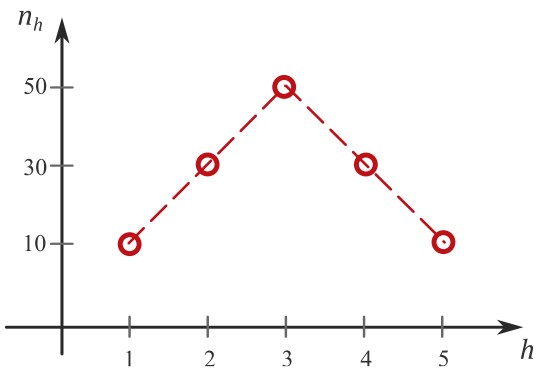

Figure 6. The number of elements per module as a function of the hierarchical level h, as considered for some of the experiments in this section.

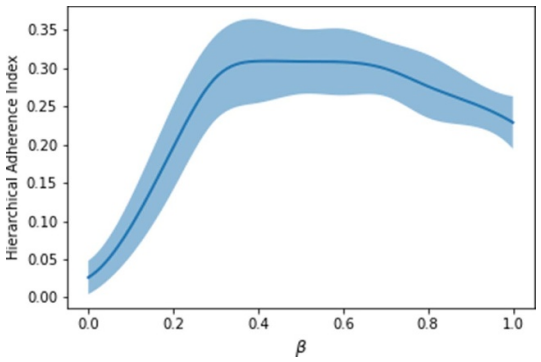

Figure 7. The hierarchical adherence indices (average ± standard deviation) obtained for hierarchical template with $H = 3$ and $b = 3$, and respective content networks with $N = 100$ and $n = 10$ for all modules.

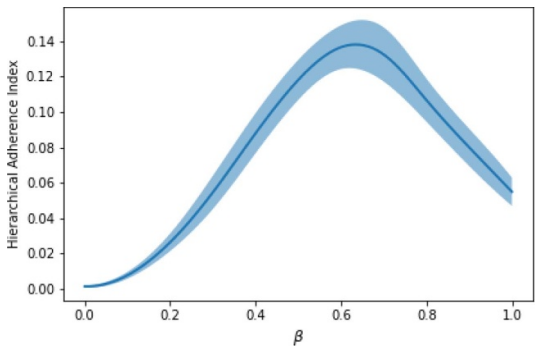

Figure 8. The hierarchical adherence indices (average ± standard deviation) were obtained for the hierarchical template with $H = 5$ and $b = 3$, and respective content networks with $N = 100$ and $n = 10$ for all modules.

symbols in the module at successive levels, required for more substantial content sharing and respective higher coincidence values.

We now proceed to model-theoretical content networks in which the number of elements per module varies with the respective hierarchical level (see figure 5). More specifically, we consider hierarchical template with $H = 5$ and $b = 3$, and content networks with $N = 300$ and $n_h = 10, 30, 50, 30, 10$. Figure 9 illustrates the therefore obtained hierarchical adherence indices.

The obtained result resembles that shown in figure 8, though being moderately narrower and with smaller error bars (standard deviations), possibly as a consequence of the relatively larger number of elements used in the intermediate levels.

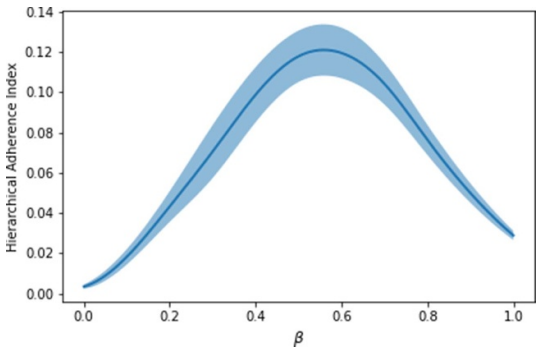

Figure 9. The hierarchical adherence indices (average ± standard deviation) obtained for hierarchical template with $H = 5$ and $b = 3$, and respective content networks with $N = 300$ and $n_h = 10, 30, 50, 30, 10$.

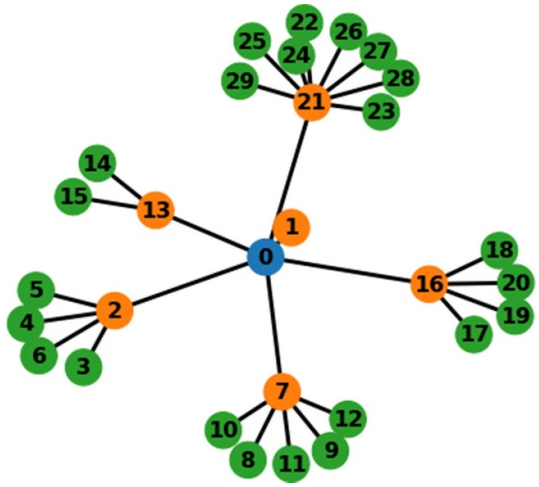

Figure 10. The hierarchical template for the illustration of the hierarchical index approach was applied to the hybrid case, as well as for the real-world Brazil Wikipedia content network. This hierarchy contains 29 nodes organized into 3 levels.

5.2. Real-world template and model-theoretical content case

Having studied the hierarchical adherence considering hierarchical templates and content networks being both model-theoretical, in the present section we describe experiments obtained respectively to *hybrid* approach, in which we use a real-world hierarchical template corresponding to that in the Brazil dataset described in section 5.3, but with modules containing sets of elements derived by using the respective model-theoretical approach described in section 4.3.

The hierarchical template was obtained from the Wikipedia homepage about several aspects of Brazil, which is shown in figure 10. This hierarchical template contains 29 nodes organized into 3 hierarchical levels.

Then, networks including the same number of nodes and levels as in the adopted template were generated while varying the β parameter between 0 and 1. The dictionary size was taken as $N = 2200$, and the number of words per module was $n = 190$. A total of 100 realizations were considered, yielding the average ± standard deviation values of the hierarchical adherence index.

The hierarchical adherence between each of these model-theoretic content networks and the respective template was then quantified by using the coincidence-based approach suggested in the present work. Figure 11 shows the respectively obtained results.

As can be observed in the average curve plot in figure 11, the coincidence similarity between the hierarchical template and the model-theoretical content networks tends to increase smoothly and steadily up to a peak taking place near $\beta = 0.4$, then decreasing also in a smooth fashion. The network corresponding to the obtained peak (b) can be verified to have a well-defined hierarchy that closely resembles that of the template hierarchy in figure 10.

The obtained results substantiate the ability of the coincidence similarity in identifying the maximum adherence between the given template and specific content networks.

In order to complement our analysis regarding the above template and model theoretical networks, we repeated the above experiment varying the parameters m and n.

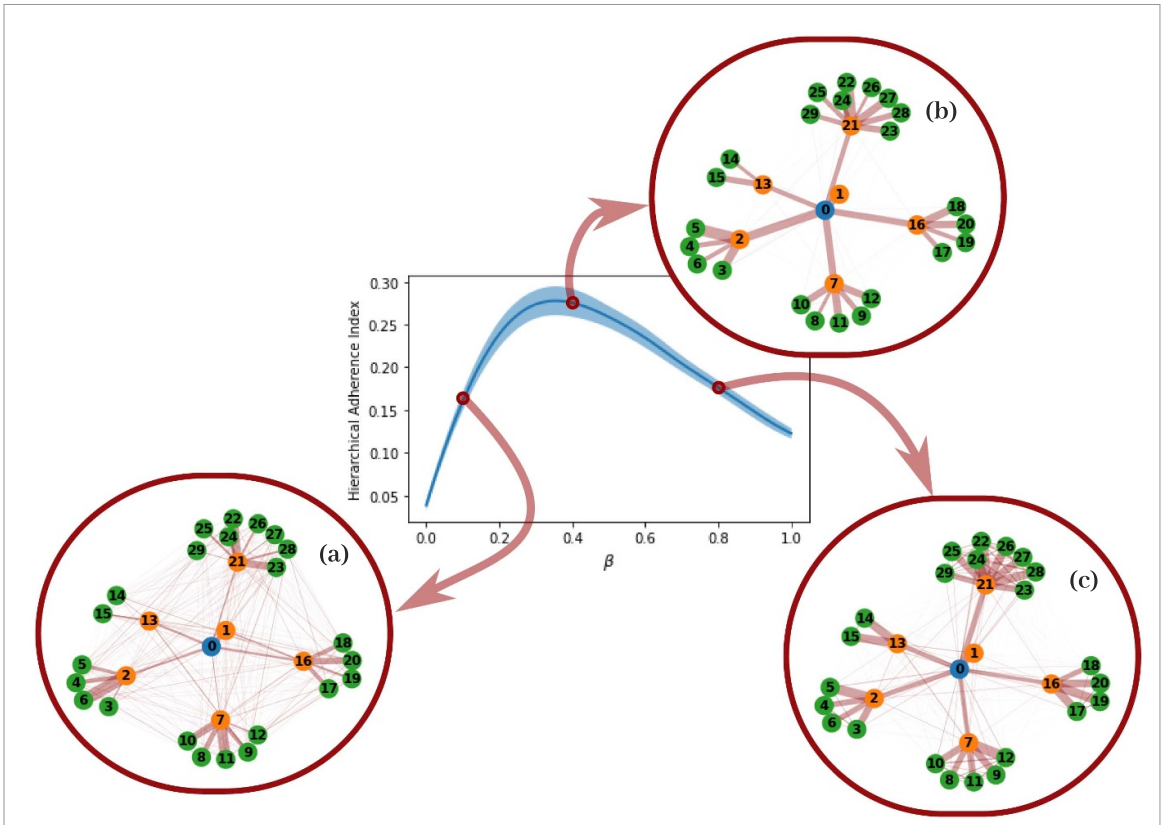

Figure 11. The quantification, in terms of the coincidence similarity, of the adherence of several model theoretical networks with the original template in figure 10 respectively to several model-theoretical content networks obtained for varying values of β. The curve shows the average $\pm$ standard deviations. Networks are illustrated respective to (a) $\beta = 0.1$, (b) $\beta = 0.4$, (c) $\beta = 0.8$. A well-defined peak can be observed near $\beta = 0.4$. Also shown are the error bars (blue) corresponding to the standard deviation of the coincidence similarities.

Figure 12 presents the average coincidence similarity values (hierarchical adherence index) obtained for $N = 50, 100$, and 200 for several values of β between 0 and 1. The error bars correspond to the standard deviation of the results.

The obtained results indicate that the larger the dictionary size, the better the adherence that tends to be observed between the template and the model theoretical networks. This follows from the fact that a larger dictionary allows the content of the nodes to become more complete and specific, contributing to enhanced mutual differentiation between nodes at the same and different hierarchical levels. Observe that the standard deviations also tend to increase with n.

Figure 13 presents the hierarchical adherence indices obtained by assuming $n = 5, 10$ and 20 while keeping $N = 100$, in terms of subsequent values of β taken between 0 and 1.

The results indicate, for $\beta > 0.6$, a moderate tendency of the hierarchical adherence index to increase with n, which is reasonable because a larger number of words per node tends to contribute to the respective specificity and differential of the modules and hierarchical levels.

To complement our theoretical analysis, we investigated how the similarity values estimated by the cosine, interiority, Jaccard, and coincidence indices change as the network leading to the maximum adherence in the above examples is progressively rewired (uniform probabilities). The results are shown in figure 14 for rewiring rates ranging from 0% to 100%. It can be readily observed that the cosine and interiority indices decay almost identically, presenting the smallest overall decreases as the rewiring rate increases. The Jaccard index resulted in an intermediate decreasing profile, while the coincidence index led to the fastest overall decrease, therefore corroborating its ability to implement more strict and selective comparisons.

5.3. Real-world case

The first real-world case-study we present concerns the hierarchical structure of the page of Brazil available on Wikipedia[3]. The template structure consists of 29 modules representing sections and subsections, for instance, 'History', 'Geography', 'Climate', and 'Tourism'. The hierarchical template of the connections

[3] https://en.wikipedia.org/wiki/Brazil.

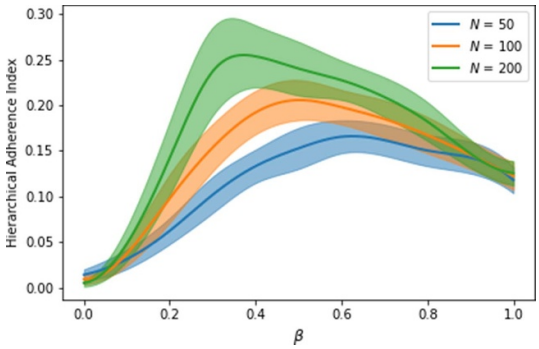

Figure 12. The average ± standard deviation of the hierarchical adherence index values obtained for the considered model theoretical case assuming $n = 10$ and $N = 50, 100$, and 200, estimated for several values of β. As could be expected, the hierarchical adherence decreases with n, achieving a peak at an intermediate value of β.

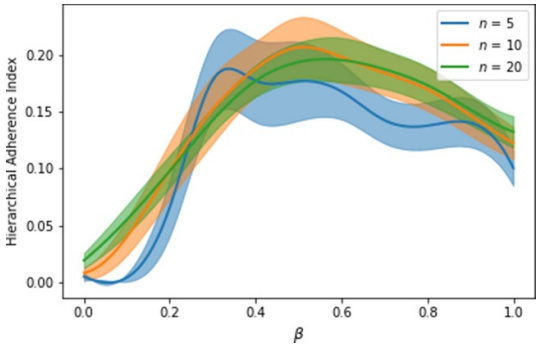

Figure 13. The average ± standard deviation of the hierarchical adherence index values obtained for the considered model theoretical case while keeping $N = 100$ and taking $n = 5, 10$ and 20, in terms of several values of β comprised between 0 and 1. Slightly higher values of the hierarchical adherence index have been obtained for $n = 20$.

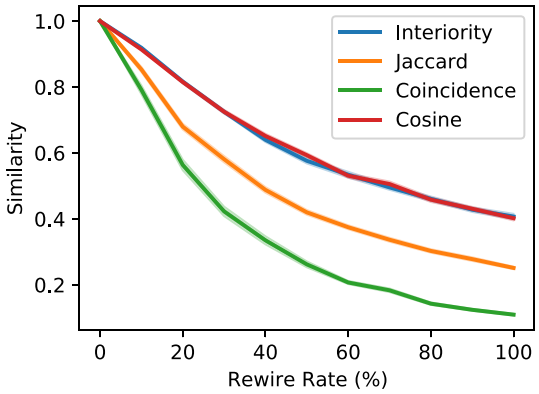

Figure 14. Comparison of the similarity values as obtained by applying the cosine, interiority, Jaccard, and coincidence indices to the comparison between the best matching network and the template in the above examples, as that network is progressively rewired. The coincidence index allowed the most strict and selective comparisons along the full range of rewirings.

between the modules is established according to their hierarchy presented in sections and subsections on Wikipedia, resulting in the tree shown in figure 10.

Each of the 29 modules was represented as respective nodes of a coincidence network, obtained by using the coincidence methodology described in section 3.2, with the weights of the links corresponding to the respective coincidence values between each pair of nodes. The thus obtained networks are henceforth called the *content network*. The respectively obtained result is shown in figure 15, including the identification of the respective hierarchical levels as indicated in the respective template in figure 10. More specifically, the hierarchical levels are identified successively (from higher to lower) by the colors blue, orange, and green. In addition, the width of the links indicates the obtained similarity values.

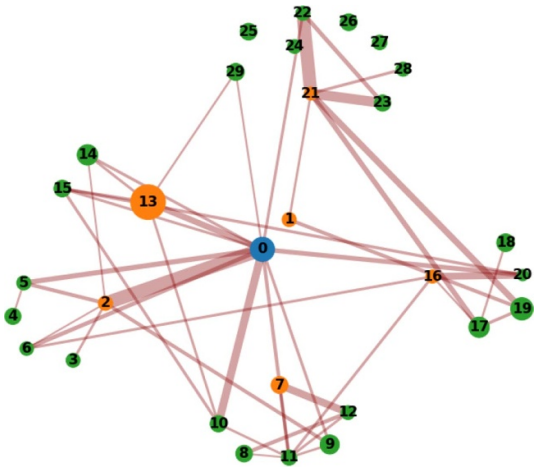

Figure 15. Coincidence similarity network obtained for the Wikipedia pages about Brazil, adopting $T = 0.362$. The colors indicate the hierarchical levels, according to the template hierarchy in figure 10. The width of the edges is proportional to the respective coincidence values. A moderate correspondence between the template and content network can be observed. The hierarchical adherence index for this case is $\mathcal{H} = 0.092$.

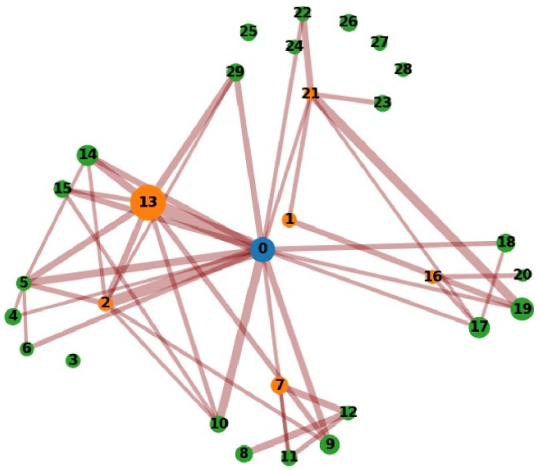

Figure 16. The content network of the Wikipedia pages about Brazil as resulting from the application of the cosine similarity, with $T = 0.497$. Observe that denser interconnections are obtained respectively to the network in figure 15, which are a consequence of the less strict comparison implemented by the cosine similarity index. The hierarchical adherence index for this case is $\mathcal{H} = 0.0679$.

For comparison purposes, figure 16 shows the similarity network obtained for the same dataset, but by adopting the cosine similarity instead of the coincidence index.

By comparing figures 15 and 16, it can be inferred that the content network obtained by using the coincidence similarity has less dense interconnections that also adhere more closely to the original template, as indicated by the respective hierarchical adherence index value of 0.092. At the same time, the content network obtained by using the cosine similarity is substantially denser than that obtained by the coincidence similarity, as a consequence of the less strict comparisons implemented by this measurement, being also less related to the original hierarchical template. Indeed, the hierarchical adherence index obtained in this case was equal to 0.0679.

The specific adherence indices of the coincidence similarity network obtained for the Wikipedia pages on Brazil, after thresholding by $T = 0.362$, are presented in table 3.

The obtained values of the specific indices reveal that the observed small adherence between the respective hierarchical template and content network is mostly accounted for by interconnections between non-adjacent levels (43.478%) and missed links (51.724%).

Figures 17(a) and (b) present the hierarchical template and content network resulting in the Wikipedia content about the U.S.A., which contains 21 modules representing, for instance, 'Government and politics', 'Economy', 'Science, technology, and energy', and 'Language'.

Table 3. The four specific indices obtained for the Brazil Wikipedia content network are shown in figure 17.

Indices	Values
ϵ_A	43.478%
ϵ_B	4.478%
ϵ_C	7.826%
ϵ_D	51.724%

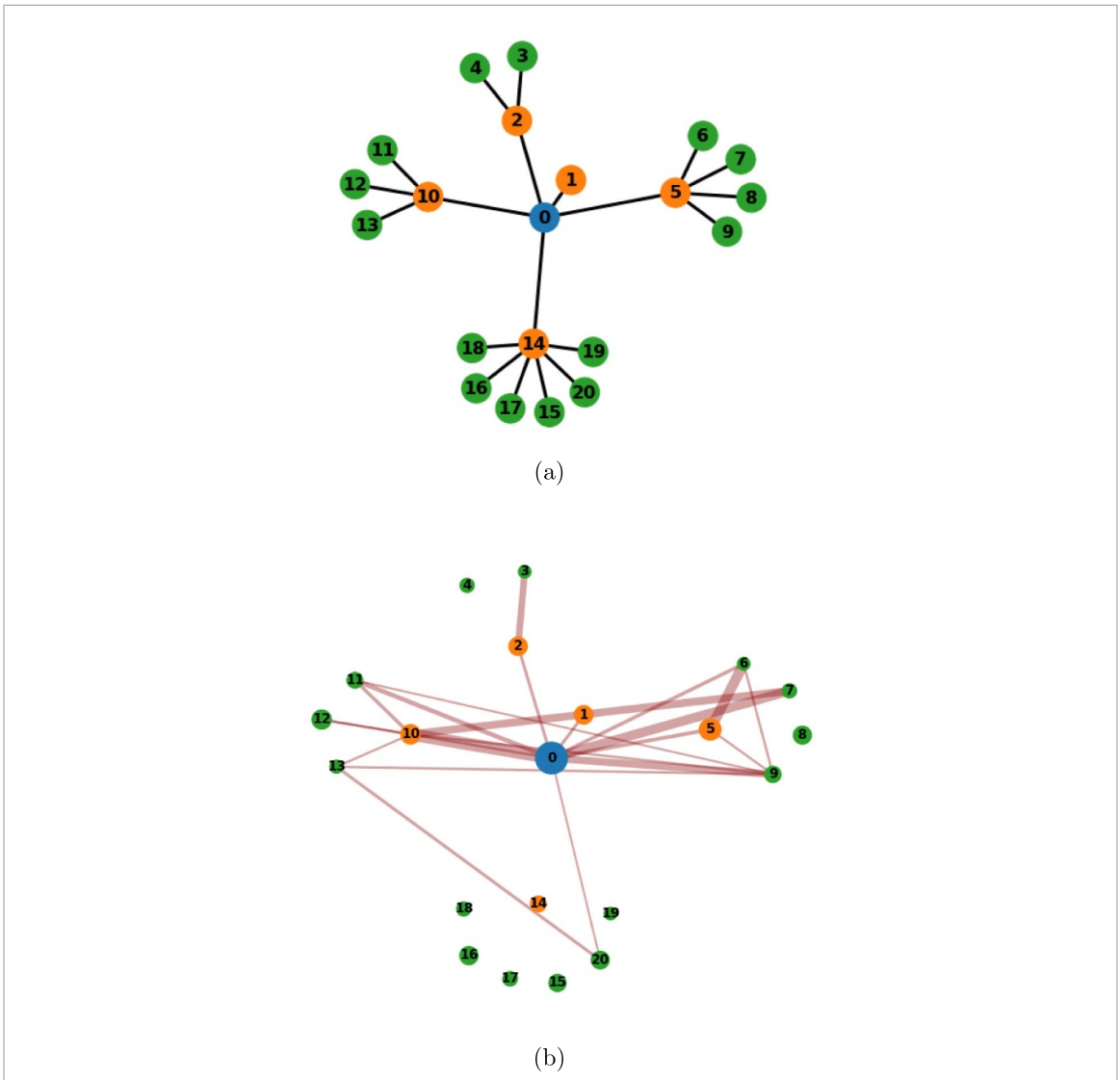

Figure 17. The hierarchical template (a) and content network were obtained from the Wikipedia pages on the U.S.A., using $T = 0.405$. The hierarchical adherence index for this case resulted equal to $\mathcal{H} = 0.1283$.

Table 4. The four specific indices obtained for the U.S.A. Wikipedia content network are shown in figure 15.

Indices	Values
ϵ_A	40.000%
ϵ_B	3.478%
ϵ_C	3.333%
ϵ_D	50.000%

The obtained specific indices of the content network obtained for the Wikipedia pages on the U.S.A., after thresholding by $T = 0.405$, are shown in table 4.

As with the Brazil content network, the obtained specific indices imply that the observed lack of adherence between the respective hierarchical template and content network is related to interconnections between non-adjacent levels (40.000%) and missed links (50.000%). It is interesting to observe that,

compared to the Brazil network, relatively smaller values of indices $\epsilon_A = 43.478\%$ and $\epsilon_D = 51.724\%$ were obtained in the case of the U.S.A. network, which is reflected in the respectively higher value of the respectively obtained overall hierarchical adherence index.

6. Conclusions

Several real-world and abstract systems and structures are inherently underlain by respective *hierarchical* organization, a property that has motivated substantial attention from the scientific community.

The present work has addressed a specific, but potentially important problem regarding the quantification of the adherence between a document (or structure), represented as a *content network*, and a given *hierarchical template*. This has been approached first in terms of an overall hierarchical adherence index based on the coincidence similarity, and then in terms of four specific indices related to respective problems of links between non-adjacent levels, links between modules at the same level, converging links implying in respective loops, as well as missing links. The proposed approach has been illustrated respectively to three main types of data: (i) model-theoretic; (ii) hybrid; and (iii) real-world.

In the first case, we had both the hierarchical templates and content networks to be derived in a respectively described model-theoretical approach. In particular, full trees were considered as hierarchical templates, while two types of content networks were adopted, being characterized by constant number of elements in all modules as well as a respective increasing/decreasing number of elements along the hierarchical levels. The concepts and methods described in the present work have been found to provide meaningful results in all respectively considered situations. More specifically, the hierarchical adherence index values obtained for successive values of the mixing parameters β were found to be characterized by a peak of adherence at an intermediate value of this parameter occurring at $\beta \approx 0.6$. In addition, the overall adherence was found to decrease substantially when proceeding from hierarchies with 3 to 5 hierarchical levels.

In the hybrid case, we adopted a real-world hierarchical template corresponding to the Wikipedia pages on aspects of Brazil, while the content networks were derived by using respectively described model-theoretical approaches. Again, a peak of similarity between the hierarchies was identified at intermediate values of the mixture parameter β. Additional experiments were also performed aimed at studying the effect of the parameters N and n on the overall adherence. It was found that higher values of N and n tend to lead to larger adherence values.

In order to illustrate the application of the proposed concepts and methods respectively to real-world data, we considered content networks obtained from the Wikipedia pages about Brazil and the U.S.A. The respective pages organization was taken as the hierarchical template. Many interesting results were observed, including possibly surprising small values of hierarchical adherence indices which, according to the respectively obtained four specific indices, were found to be mostly related to links between non-adjacent hierarchies and missing links. The adherence index obtained for the U.S.A. pages resulted about twice as large as that obtained for the Brazil pages, suggesting that the former data is moderately more complete and/or hierarchically organized. Interestingly, a substantially small number of links between modules at the same level and converging links were obtained, indicating that the resulting hierarchical structure is relatively robust regarding these two specific aspects.

The relatively small values of hierarchical adherence observed for the two real-world documents seem to suggest that this tendency could be found in other real-world structures and documents possibly as a consequence of the intrinsic sharing between the concepts between adjacent levels being implemented more at a *semantic, implicit* level, not necessarily being respectively reflected in the actual observed content of the modules.

Several are the possibilities for further research motivated by the reported concepts, methods, and results. For instance, it would be interesting to consider additional parameters that contribute to the hierarchical structures, including the number of levels, number of branches (b), number of elements per module, etc. The suggested approaches can also be used to investigate other real-world documents, including books and encyclopedias, among other data types. Of particular interest would be to check the possibility of semantic, implicit overlaps between the contents of adjacent modules not only in text networks but also in other types of data (e.g. [49]).

Data availability statement

No new data were created or analysed in this study.

Acknowledgments

Alexandre Benatti thanks Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001 (88882.328749/2019-01). Ana C M Brito acknowledges financial support from São Paulo Research Foundation (FAPESP Grant No. 2020/14817-2) and Capes-Brazil for sponsorship. Diego R Amancio acknowledges financial support from CNPq-Brazil (Grant No. 311074/2021-9) and FAPESP (20/06271-0). Luciano da F Costa thanks CNPq (Grant No. 307085/2018-0) and FAPESP (Grant No. 15/22308-2).

ORCID iDs

Alexandre Benatti https://orcid.org/0000-0002-7419-4712
Luciano da F Costa https://orcid.org/0000-0001-5203-4366

References

[1] da F Costa L 2021 An ample approach to data and modeling (available at: www.researchgate.net/publication/355056285_An_Ample_Approach_to_Data_and_Modeling)
[2] Liu J-B, Bao Y, Zheng W-T and Hayat S 2021 Network coherence analysis on a family of nested weighted n-polygon networks *Fractals* **29** 2150260
[3] Stadler T, Skylaki S, Kokkaliaris K D and Schroeder T 2018 On the statistical analysis of single cell lineage trees *J. Theor. Biol.* **439** 160–5
[4] Venkatesh S and Reps T 2014 Recovery of class hierarchies and compositionrelationships from machine code *23rd Int. Conf. on Compiler Construction (CC'14)*
[5] da F Costa L 2021 Further generalizations of the jaccard index (available at: www.researchgate.net/publication/355381945_Further_Generalizations_of_the_Jaccard_Index)
[6] da F Costa L 2021 Multiset neurons (available at: www.researchgate.net/publication/356042155_Common_Product_Neurons)
[7] Fortunato S 2010 Community detection in graphs *Phys. Rep.* **486** 75–174
[8] Cong J and Liu H 2014 Approaching human language with complex networks *Phys. Life Rev.* **11** 598–618
[9] de Arruda H F, Costa L and Amancio D R 2016 Using complex networks for text classification: discriminating informative and imaginative documents *Europhys. Lett.* **113** 28007
[10] Akimushkin C, Amancio D R, Oliveira O N Jr and Gao Z-K 2017 Text authorship identified using the dynamics of word co-occurrence networks *PLoS One* **12** e0170527
[11] de Arruda H F, Silva F N, Marinho V Q, Amancio D R and da F Costa L 2018 Representation of texts as complex networks: a mesoscopic approach *J. Complex Netw.* **6** 125–44
[12] Quispe L V C, Tohalino J A V and Amancio D R 2021 Using virtual edges to improve the discriminability of co-occurrence text networks *Physica* A **562** 125344
[13] Sanderson M and Croft B 1999 Deriving concept hierarchies from text *Proc. 22nd Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval* pp 206–13
[14] Wu Z, Li Z, Mitra P and Giles C L 2013 Can back-of-the-book indexes be automatically created? *Proc. 22nd ACM Int. Conf. on Information and Knowledge Management* pp 1745–50
[15] Yang Y, Liu H, Carbonell J and Ma W 2015 Concept graph learning from educational data *Proc. 8th ACM Int. Conf. on Web Search and Data Mining* pp 159–68
[16] Wang S, Liang C, Wu Z, Williams K, Pursel B, Brautigam B, Saul S, Williams H, Bowen K and Giles C L 2015 Concept hierarchy extraction from textbooks *Proc. 2015 ACM Symp. on Document Engineering* pp 147–56
[17] Silva F N, Amancio D R, Bardosova M, da F Costa L, Costa L Oliveira O N Jr 2016 Using network science and text analytics to produce surveys in a scientific topic *J. Inform.* **10** 487–502
[18] Amancio D R, Oliveira O N Jr and da F Costa L 2012 Three-feature model to reproduce the topology of citation networks and the effects from authors' visibility on their h-index *J. Inform.* **6** 427–34
[19] Anoop V, Asharaf S and Deepak P 2016 Unsupervised concept hierarchy learning: a topic modeling guided approach *Proc. Comput. Sci.* **89** 386–94
[20] Liu J-B, Bao Y and Zheng W-T 2022 Analyses of some structural properties on a class of hierarchical scale-free networks *Fractals* **30** 2250136
[21] Liang C, Wu Z, Huang W and Giles C L 2015 Measuring prerequisite relations among concepts *Proc. 2015 Conf. on Empirical Methods in Natural Language Processing* pp 1668–74
[22] Rios-Alvarado A B, Lopez-Arevalo I and Sosa-Sosa V J 2013 Learning concept hierarchies from textual resources for ontologies construction *Expert Syst. Appl.* **40** 5907–15
[23] Sun J, Ajwani D, Nicholson P K, Sala A and Parthasarathy S 2017 Breaking cycles in noisy hierarchies *Proc. 2017 ACM on Web Science Conf.* pp 151–60
[24] Zheng W, Fang H and Yao C 2012 Exploiting concept hierarchy for result diversification *Proc. 21st ACM Int. Conf. on Information and Knowledge Management* pp 1844–8
[25] Chen P, Lu Y, Zheng V W, Chen X and Yang B 2018 Knowedu: a system to construct knowledge graph for education *IEEE Access* **6** 31553–63
[26] Wang Y and Li Y H 2009 Deep web entity identification method based on improved jaccard coefficients *2009 Int. Conf. on Research Challenges in Computer Science* (IEEE) pp 112–5
[27] Blanchard E, Harzallah M and Kuntz P 2008 A generic framework for comparing semantic similarities on a subsumption hierarchy *18th European Conf. on Artificial Intelligence* vol 2008 pp 20–24
[28] Frank S L, Bod R and Christiansen M H 2012 How hierarchical is language use? *Proc. R. Soc.* B **279** 4522–31
[29] Frank S L and Bod R 2011 Insensitivity of the human sentence-processing system to hierarchical structure *Psychol. Sci.* **22** 829–34

[30] Crouzet O 2007 On segments and syllables in the sound structure of language: curve-based approaches to phonology and the auditory representation of speech *Math. Sci. Hum. Math. Soc. Sci.* **180** 57–71
[31] Patel A D 2003 Language, music, syntax and the brain *Nat. Neurosci.* **6** 674–81
[32] Whittaker R H 1969 New concepts of kingdoms of organisms: evolutionary relations are better represented by new classifications than by the traditional two kingdoms *Science* **163** 150–60
[33] Nehaniv C L and Rhodes J L 2000 The evolution and understanding of hierarchical complexity in biology from an algebraic perspective *Artif. Life* **6** 45–67
[34] Kiebel S J, Daunizeau J, Friston K J and Sporns O 2008 A hierarchy of time-scales and the brain *PLoS Comput. Biol.* **4** e1000209
[35] Hasson U, Chen J and Honey C J 2015 Hierarchical process memory: memory as an integral component of information processing *Trends Cogn. Sci.* **19** 304–13
[36] Hochstein S and Ahissar M 2002 View from the top: hierarchies and reverse hierarchies in the visual system *Neuron* **36** 791–804
[37] Bird S, Klein E and Loper E 2009 *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit* (O'Reilly Media, Inc.)
[38] da F Costa L 2022 Coincidence complex networks *J. Phys. Complex.* **3** 015012
[39] Jaccard P 1901 Étude comparative de la distribution florale dans une portion des alpes et des jura *Bull. Soc. Vaudoise Sci. Nat.* **37** 547–79
[40] Samanthula B K and Jiang W 1989 Secure multiset intersection cardinality and its application to jaccard coefficient *IEEE Trans. Dependable Secure Comput.* **13** 591–604
[41] Leydesdorff L 2008 On the normalization and visualization of author co-citation data: Salton's cosine versus the jaccard index *J. Am. Soc. Inf. Sci.* **59** 77–85
[42] Wikipedia 2021 Jaccard index (available at: https://en.wikipedia.org/wiki/Jaccard_index) (Accessed 10 October 2021)
[43] Schubert A and Telcs A 2014 A note on the jaccardized czekanowski similarity index *Scientometrics* **98** 1397–9
[44] Vijaymeena. M K and Kavitha K 2016 A survey on similarity measures in text mining *Mach. Learn. Appl.* **3** 19–28
[45] da F Costa L 2022 A brief guide to the coincidence similarity and its applications (available at: www.researchgate.net/publication/362713778_A_Brief_Guide_to_the_Coincidence_Similarity_and_its_Applications)
[46] da F Costa L 2021 On similarity (available at: www.researchgate.net/publication/355792673_On_Similarity)
[47] da F Costa L 2021 Multisets (available at: www.researchgate.net/publication/355437006_Multisets)
[48] Benatti A and da F Costa L 2022 Retrieving hierarchies (available at: www.researchgate.net/publication/359831098_Retrieving_Hierarchies)
[49] Amancio D R, Nunes M G V, Oliveira O N Jr and da F Costa L 2012 Using complex networks concepts to assess approaches for citations in scientific papers *Scientometrics* **91** 827–42