IRIS AperTO

UNIVERSITÀ DEGLI STUDI DI TORINO

AperTO - Archivio Istituzionale Open Access dell'Università di Torino

Reducing metadata complexity for faster table summarization

This is a pre print version of the following article:

Original Citation:

Availability:

This version is available http://hdl.handle.net/2318/73787 since 2021-04-29T21:32:48Z

Publisher:

ACM International Conference Proceeding Series

Terms of use:

Open Access

Anyone can freely access the full text of works made available as "Open Access". Works made available under a Creative Commons license can be used according to the terms and conditions of said license. Use of all other works requires consent of the right holder (author or publisher) if not exempted from copyright protection by the applicable law.

(Article begins on next page)

31 July 2024

ANITA: A Narrative Interpretation of Taxonomies for their Adaptation to Text Collections

ABSTRACT

Taxonomies and concept hierarchies embody formalized knowledge and define aggregations between concepts/categories in a given domain, facilitating the organization of the data and making the contents easily accessible to the users. Since taxonomies have significant roles in the data annotation, search and navigation, they are often carefully engineered. However, especially in dynamically evolving domains such as news, they do not necessarily reflect the content knowledge. Moreover, when the user's interests are highly focused, available taxonomies –which are often designed for broad coverage of concepts in a given application domain– may fail to reflect details within the users foci of interest. Thus, in this paper, we ask and answer, in the positive, the following question: "*is there a feasible approach to efficiently and effectively adapt a given taxonomy to a usage context defined by a corpus of documents?*". In particular, we recognize that the primary role of a taxonomy is to describe or *narrate* the natural relationships between concepts in a given document corpus. Therefore, a corpus-aware adaptation of a taxonomy should essentially *distill* the structure of the existing taxonomy by appropriately segmenting and, if needed, summarizing this narrative relative to the content of the corpus. Based on this key observation, we propose *A Narrative Interpretation of Taxonomies for their Adaptation* (`ANITA`) for re-structuring existing taxonomies to varying application contexts and we evaluate the proposed scheme using different text collections.

1. INTRODUCTION

While there are many strategies for organizing text documents, hierarchical categorization –usually implemented through a pre-determined taxonomical structure– is often the preferred choice. In a taxonomy-based information organization, each category in the hierarchy can index text documents that are relevant to it, facilitating the user in the navigation and access to the available contents. Unfortunately, given a set of text documents, it is not easy to find the appropriate categorization that best describes the contents. In fact the available taxonomies are usually designed for broad coverage of concepts in a considered domain, failing to reflect important details (within the users foci of interest) expressed by the considered data set. Indeed, especially in dynamically evolving domains, the available taxonomies could not necessarily reflect the content knowledge.

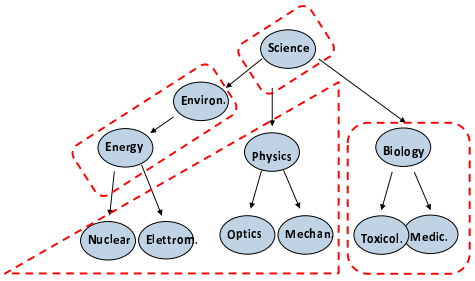
In this paper we introduce a new method for distilling a taxonomical domain categorization from an existing one, based on a given set of text documents that have to be represented and indexed by it.

For this purpose, we recognize that the primary role of a taxonomy is to describe or *narrate* the natural relationships between concepts in a given domain to its users. Therefore, a contextually relevant adaptation of a taxonomy should essentially distill and manipulate the structure of the existing taxonomy by appropriately segmenting and, if needed, summarizing this narrative relative to the documents in a given corpus. Based on this key observation, we propose *A Narrative Interpretation for Taxonomy Adaptation* (`ANITA`), a novel taxonomy distillation approach for adapting existing taxonomies to varying application contexts. The specific contributions of this paper are as follows:

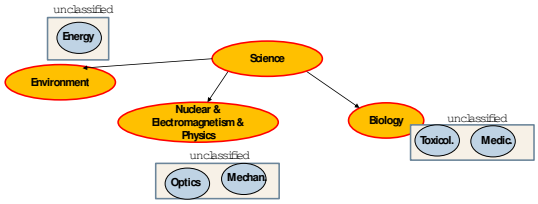
- *A narrative view of taxonomies:* we view a taxonomy as a discourse that is describing the relationships between concepts/categories in a given domain. Thus, as described in Section 3.1, we transform a given taxonomy into a *narrative.*
- *Corpus-driven reinterpretation of the narrative:* Given a context defined by the considered corpus of text documents, this narrative is re-interpreted and re-structured (Section 3.2) based on a statistical analysis of terms and structural analysis of the taxonomy.
- *Segmentation of the narrative*: This narrative is then analyzed and segmented based on a narrative-development analysis, highlighting where the narrative significantly drifts from one concept-topic to another (Section 3.3).
- *Re-construction (or distillation) of an adapted taxonomy based on the segmentation results:* The resulting narrative segments (each describing a group of concepts/categories that collectively act as a single topic) are re-organized into a hierarchical structure, linking each concept-segment to others that are structurally related to it (Section 3.4).

The result of the above process is a contextually-relevant *adapted taxonomy*, where details are highlighted where they

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Copyright 200X ACM X-XXXXX-XX-X/XX/XX ...$10.00.



(a) Original taxonomy (and its narrative partitions)



(b) Re-constructed (or distilled) taxonomy

Figure 1: Adaptation of a DMOZ taxonomy fragment about "science": (a) the original taxonomy is partitioned through narrative-based interpretation, based on available content (NSF data set described in Section 4) and (b) the partitions are used to create a concise taxonomy reflecting the content.

matter and suppressed where they do not support the current context. Suppressed categories are attached to the relevant category of the adapted taxonomy as *un-classified concepts* (Figure 1). In Section 4, we evaluate the proposed scheme using text collections.

2. RELATED WORK

In order to adapt/summarize hierarchical structures to represent the available contents, various hierarchical clustering methods have been proposed. There are two major approaches for hierarchical clustering: agglomerative clustering and divisive clustering. In [1], the centroids of each class are used as the initial seeds and then a projected clustering method is applied to build the hierarchy. In [12] a linear discriminant projection is applied to the data first and then the hierarchical clustering method UPGMA [7] is exploited to generate a binary tree. [15] applies a divisive hierarchical clustering: authors generate a taxonomy with each node associated with a list of the categories. [5] associates word distribution conditioned on classes to each node: the method uses a variance of the EM algorithm to cluster nodes. Similarly, [19] presents a method in which concepts are probabilistically modelled. The probabilistic classes are organized in hierarchies by relying on the KL divergence measure between the probability distributions associated to the concepts.

A successful approach for organizing web query results based on available web structure is topic distillation [11]. The basic idea in topic-distillation is to consider the structure of the Web and propagate scores between pages in a way to organize topic spaces in terms of smaller sets of authoritative pages. Moreover, given a query, other methods propagate the term frequency values between neighboring pages [18] or the relevance score itself between web pages connected with hyperlinks [20]. In a text environment, a concept taxonomy can be also used to flexibly describe a user/group's interests with varying granularity. [21] addresses the problem of how to adapt a topic taxonomy in order to reflect the change of a group's interest to achieve dynamic group profiling.

Summarization of a text stream relies on the analysis of the evolution of the arguments expressed by the sequence of sentences. In document summarization, summary sentences are typically arranged in the same order that they were found in the full document, although [8] reports that human summarizers do sometimes change the original order. The task of ordering sentences to obtain a meaningful narrative in a way that reflects a given context has also been extensively investigated in the text and discourse generation literature [13, 14, 6].

3. NARRATIVE-DRIVEN TAXONOMY ADAPTATION PROCESS

Given a taxonomy H (also called hierarchy in the paper) with n nodes (concepts or categories in the paper), our goal is to create an adapted taxonomy H', based on a given context defined by a corpus of text documents. As described before, `ANITA` consists of four steps. In this section, we present the details of each of these steps.

3.1 Step I: Narrative View of a Taxonomy

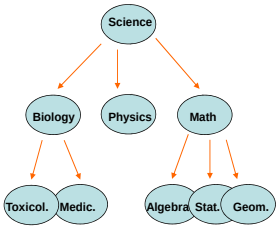
As we mentioned in the introduction, the given taxonomy H is transformed into a *narrative* which captures the structural relationships of concepts/categories in the taxonomy. Note that, whereas a taxonomy is a hierarchy of concept nodes, a *narrative* is a sequence of sentences. Therefore, in order to create a narrative corresponding to the taxonomy, we need to achieve two goals: (a) we need to map concept-nodes of the input taxonomy into "*concept-sentences*" and (b) we need to pick an appropriate ordering of these concept-sentences.

3.1.1 *Step Ia. Mapping from the Concept-Nodes to "Concept-Sentences"*

What we refer to as "*concept-sentences*" above is not natural language sentences, but concept-vectors obtained by analyzing the hierarchical structure of the given taxonomy. Intuitively, these concept-vectors can be thought of as being analogous to *keyword vectors* commonly used in representing documents in IR systems.

Given a taxonomy, we associate a concept-vector to each concept in the underlying hierarchy using the CP/CV structural encoding scheme proposed in [10]. Given a taxonomy $H(C, E)$ with $n = |C|$ concepts, [10] represents each node in the hierarchy as an n-dimensional vector of related concepts. More specifically, authors introduce a concept propagation (CP) scheme, which relies on the structural relationships between concepts implied by the hierarchy, to annotate each concept-node with a concept-vector (CV). Each vector encodes the *structural* relationship between this node and all the other nodes in the hierarchy.

CP/CV is a spreading activation [2] like algorithm, where before the propagation process starts, the concept-vector corresponding to c_i is initialized by setting the weight corresponding to the concept c_i to 1 and all others to 0. Then, the vectors are propagated (as in spreading activation) between parent/child concepts, but taking into account the structural distance between them. This propagation process is repeated until all concepts are informed of all the others.



(a)

	science	math	physics	biology	algebra	geometry	statistics	toxicology	medicine
$\vec{cv}_{science}$	0.450	0.169	0.141	0.158	0.018	0.018	0.018	0.021	0.021
$\vec{cv}_{math}$	0.052	0.469	0.006	0.006	0.156	0.156	0.156	0.0003	0.0003
$\vec{cv}_{physics}$	0.100	0.012	0.873	0.012	0.0006	0.0006	0.0006	0.0007	0.0007
$\vec{cv}_{biology}$	0.057	0.007	0.007	0.520	0.0003	0.0003	0.0003	0.204	0.204
$\vec{cv}_{algebra}$	0.004	0.100	0.0002	0.0002	0.872	0.012	0.012	0	0
$\vec{cv}_{geometry}$	0.004	0.100	0.0002	0.0002	0.012	0.872	0.012	0	0
$\vec{cv}_{statistics}$	0.004	0.100	0.0002	0.0002	0.012	0.012	0.872	0	0
$\vec{cv}_{toxicology}$	0.006	0.0003	0.0003	0.165	0	0	0	0.806	0.023
$\vec{cv}_{medicine}$	0.006	0.0003	0.0003	0.165	0	0	0	0.023	0.806

(b)

Figure 2: (a) A portion of taxonomy about *science* and (b) the corresponding concept vectors obtained by the CP/CV process. In this paper, we treat each row as a "*concept-sentence*" where the weights reflect the contribution of the node- (or column-labels) to the description of concept represented by each row

Consider, for example, the taxonomy fragment (containing nine concept nodes) presented in Figure 2(a). CP/CV maps each concept into a 9-dimensional vector (Figure 2(b)). For example, the root is represented by the vector

$$\langle 0.450, 0.169, 0.141, 0.158, 0.018, 0.018, 0.018, 0.021, 0.021 \rangle,$$

in which the first component (the one associated to the tag "science"), dominates over the others that contribute to the definition of the concepts. The second, third and fourth components reflect the weight of "math", "physics" and "biology" respectively in the semantic characterization of "science", while the remaining components represent the weights of the three descendants of "math" and of the two descendants of "biology". [10] showed that the semantic similarities of the concepts can be computed using the cosine similarities of the concept vectors and such that similarity measurements are quite in line with the human judgments of similarities. In a sense, each concept-vector can be thought of as a sentence which describes the given concept in terms of its relationships to the other concepts in the hierarchy. Notice that, all concept-sentences are formed by the same terms (the labels of the concepts), but the weights associated to each term are different from concept-sentence to concept sentence.

3.1.2 *Step Ib. Ordering the "Concept-Sentences" into a Narrative*

After the vector-based encoding of the *concept-sentences*, the next step in narrative creation is ordering these sentences (therefore the nodes in the original hierarchy) in an order representing the structure of the taxonomy.

Ancestor-Descendent Ordering.

In this paper we consider and evaluate three different narrative orders: the pre-order traversal, the parenthetical traversal and the post-order traversal of the taxonomy.

- *Pre-order Traversal of the Taxonomy.* A hierarchy (especially a concept hierarchy) is structured in a way that the most general concept is used as the root of the hierarchy and the most specific ones are the leaves. In a sense, each node provides more specialized knowledge within the context defined by all its ancestors. We leverage this aspect by defining a narrative in which the sentences associated to the nodes of the taxonomy are read in pre-order; i.e., each concept sentence is immediately followed by its detailed description in terms of its specializations.

- *Post-order Traversal of the Taxonomy.* This traversal of the tree generates a narrative in which the different concepts are presented bottom-up: after presenting the most specific concepts, their super-concept is narrated. Any super-concept presented after the narration of its children can be seen as summarizing the description of its sub-concepts.

- *Parenthetical Traversal of the Taxonomy.* Intuitively, the parenthetical traversal is analogous to a narrative where each passage is presented with an *introduction* and goes in *details* until a general *conclusion*. In *parenthetical traversal* of the tree, each parent node is visited twice, representing the general introduction to the argument that the children specialize and their conclusion.

Distance-Preserving Sibling Ordering.

While pre-order, parenthetical, and post-order traversal of the tree help us decide in which order ancestors and descendants are to be considered, they do not help us choose the order in which the siblings in the hierarchy are to be concatenated, in the narrative. Let us consider a node c_0 with m children $\{c_1, c_2 \cdots c_m\}$. Our primary goal is to ensure that the siblings are ordered in a way that preserves the similarities – or dissimilarities – among these m concepts (as well as their parent c_0). For this purpose, we first compute the cosine dissimilarity matrix M based on the concept-vectors corresponding to all $m+1$ concepts (the parent and the m children); in other words,

$$M[i][j] = 1 - cos(\vec{cv}_{c_i}, \vec{cv}_{c_j})$$

We then use a distance-preserving embedding technique to map these concepts onto a one-dimensional ordering. In particular, without loss of generality, we use multi-dimensional scaling (MDS [22]), to embed the concepts onto a 1-dimensional order. MDS works as follows: given as inputs (1) a set of N objects, (2) a matrix of size $N \times N$ containing pairwise distance values and (3) the desired dimensionality k, MDS tries to map each object into a point in the k-dimensional space in such a way that a stress value, defined as

$$stress = \sqrt{\frac{\sum_{i,j}(d'_{i,j} - d_{i,j})^2}{\sum_{i,j} d^2_{i,j}}},$$

where $d_{i,j}$ is the actual distance between two objects o_i and o_j and $d'_{i,j}$ is the distance between the corresponding points in the resulting k-dimensional space, is minimized. Therefore, by providing as input $N = m+1$ input concepts and $k = 1$ target dimension, the resulting order of concepts would preserve the semantic ordering between the concepts as best as possible.

Notice that, due to the special nature of the node c_0 (it is the parent), we need to make a minor modification in the MDS algorithm: in particular, we constrain the stress minimization process in a way that forces the position of c_0 at the beginning of the list. This way, the resulting order of the children concepts will reflect the concept similarities with respect to the position of the parent concept in the narrative.

One difficulty with narrative ordering approach is that in many cases the hierarchy itself is not sufficiently informative to order the siblings. For example, if we look at the concept vectors (Figure 2(b)) related to the children of the concept node "*biology*" ("*medicine*" and "*toxicology*"), we can see that the hierarchy does not impose a true order between these two siblings. To cope with this, we leverage the information provided by the text corpus to order the siblings most effectively and distinguish among them in the considered corpus.

3.2 Step II: Re-interpretation of the Narrative based on the Context

The first step in re-interpreting the narrative based on the context defined by the corpus of text documents is to find which documents are associated to which concepts in the taxonomy. Thus, the corpus D, with $p = |D|$ text documents, is analyzed and a representative keyword vector is generated for each document. As usual, the keyword extraction includes a preliminary phase of stop word elimination and stemming. The weight associated to each stemmed term is computed in the augmented normalized term frequency form [17]. For the t^{th} document, a document vector $\vec{dv}_t = \{w_{t,1}, w_{t,2}, ..., w_{i,v}\}$ is defined, where v is the size of the vocabulary in D, $1 \leq t \leq p$, and $w_{t,u}$ is the normalized term frequency of the u^{th} vocabulary term in the t^{th} document. Then, for each taxonomy (if we need to consider more than a single categorization), the classification process takes as input the set of CP/CV vectors associated to hierarchy elements (already computed for constructing the narrative) and the set of document vectors associated to the corpus; therefore, as in most IR systems, it associates each document to the few best matching concepts.

Next we search for the most contextually relevant keywords corresponding to each concept. More specifically, we compute the degree of matching between the given concept c_i and a keyword k_j, which occurs in the associated documents, by treating the set of the associated documents A_{c_i} as positive evidence of relationship between k_j and c_i within the given context. Similarly, we treat the documents containing the same keyword, but not belonging to the set A_{c_i}, as negative evidence against the relationship between c_i and k_j. Intuitively, this is analogous to treating (a) the concept vector corresponding to the concept c_i as a query and (b) the set of associated documents as relevance feedback on the results of this query. Thus, given a concept c_i and the corresponding set of associated documents, A_{c_i}, we identify a weight $w_{i,j}$ for each keyword k_j using a probabilistic feedback approach [16].

At the end of this process, we merge both information (concept vector and relevant keywords) in a new *extended vector* $\vec{ev}_{c_i}$, which describes the same concept in terms of both other concepts as well as contextually relevant keywords, thanks to their relevant keywords components. Consequently, two sibling concepts c_i and c_j that are not properly differentiated within the taxonomy can now be differentiated based on the context defined by the text corpus. The sibling order imposed by the distance-preserving ordering (such as MDS) thus will be computed on these extended vectors, to properly reflect the usage context.

At this point the narrative is a sequence of sentences, each including the information coming from the structural knowledge (hierarchy) and the context knowledge (corpus of documents), defining a global discourse that covers all the topics addressed by the content, according to the knowledge expressed by the taxonomy.

3.3 Step 3: Segmentation of the Narrative

In this step, we analyze the narrative obtained in the previous step to identify segments (or partitions) that are highly correlated. The idea is that if, in the given corpus, two concepts are highly correlated, they may not need two separate nodes in the adapted taxonomy. In contrast, we can think that, if there is a significant difference between two portions of the narrative, then these two portions (or segments) do necessitate different concepts in the resulting taxonomy.

In the literature, there are various techniques for segmenting a narrative into coherent units [4, 3, 9]. Textile [4, 3] and Vectile [9] algorithms, for example, plot similarity scores (based on lexical co-occurrence and distribution analysis) of neighboring portions of the text. The dips (i.e., local minima) in the resulting similarity curve correspond to regions of the text where there is significant change in the content. Therefore, these dips are identified as text segment boundaries.

In this paper, in order to partition the narrative $\vec{ev}_1, \vec{ev}_2, \ldots, \vec{ev}_n$ into coherent segments, we use a similar strategy. However, instead of searching for local minima of similarities, we look for portions of the narrative where the change is above a threshold:

1. Given the narrative (i.e., ordered sequence of extended vectors), we first compare each pair of neighboring vectors, $\vec{ev}_i$ and $\vec{ev}_{i+1}$ $(1 \leq i \leq n-1)$ by computing their *dissimilarities*:

$$\Delta_{i,i+1} = 1 - cos(\vec{ev}_i, \vec{ev}_{i+1})$$

2. The sequence of vectors is then analyzed for *topic drifting*. We say that a topic drift occurs for a given segment of the narrative when the degree of change between its starting and ending points is above a given threshold. If $Seg_{i,j}$ denotes a segment from the vector $\vec{ev}_i$ and $\vec{ev}_j$, the corresponding degree of drift is defined as $drift_{i,j} = \sum_{k=i}^{j-1} \Delta_{k,k+1}$.

   A segment $S_{i,j}$ is said to be *coherent* if holds that $drift_{i,j} < \lambda_{max}$, where λ_{max} is the *coherence threshold*.

The narrative is segmented in such a way that each segment is *maximally* coherent (i.e. no other segment containing it would be coherent). Note that the value of the threshold λ_{max} will determine the number of resulting segments. If $\lambda_{max} \approx 0$, then the resulting segments will be as many as the original nodes. On the other hand, if $\lambda_{max} \approx 1$, we will obtain a few, large segments. Since the number of segments will determine the size of the adapted taxonomy, we pick the value of λ_{max} based on the target taxonomy size, k.

In particular, the value of λ_{max} is iteratively adjusted using binary search starting from 0.5, until the desired partition cardinality is reached[1].

At the end of the process, we obtain a set of segments, or partitions, $P = \{P_1, P_2, \cdots, P_k\}$ that represents sequences of coherent narrative components. Intuitively, each partition P_i defines a sequence of concepts.

Notice that, the parenthetical traversal introduces each parent concept twice; in this case, if a parent node is associated to two different partitions, it is removed from the partition whose drift value (with respect to neighbour nodes in the sequence) is higher.

3.4 Step 4: Taxonomy Distillation from the Partitions

In order to construct the adapted taxonomy from the partitions created in the previous step, we need to re-attach the partitions in the form of a tree structure. Furthermore, for each partition, we need to pick a *label* that will be presented to the user and will describe the concepts in the partition.

3.4.1 Partition Linking

The adapted taxonomy, $H'(C', E')$ with $C' = \{c'_1, \ldots, c'_k\}$ (where each node c'_i represents the partition P_i) should preserve the original structure of $H(C, E)$ as much as possible. Thus,

- The root of H' is c_{root} $(1 \leq root \leq k)$ such that the corresponding partition P_{root} contains the root concept of H.

- Let us consider a pair, P_i and P_j, of partitions in P. The decision on whether (and how) the corresponding concept c'_i and c'_j should be connected is based on the following analysis. Let $E_{i,j}$ be the set of edges in E linking any concept in P_i to any concept in P_j. Similarly, let $E_{j,i}$ be the set of edges in E linking any concept in P_j to any concept in P_i. With the goal of preserving to the best the structure of H, we measure the cost of violating the structural constraints implied by E in H, and we propose as our solution the adapted taxonomy which minimizes such cost. The cost is defined by cases.

  - The cost of having the corresponding c'_i as the ancestor of c'_j is the cost of the violations of the constraints associated to the edges in $E_{j,i}$.
  - Similarly, the cost of having c'_j as the ancestor of c'_i is the cost of the violations of the constraints associated to the edges in $E_{i,j}$.
  - The cost of non directly connecting c'_i and c'_j is the cost of the violations of the constraints associated to the edges in $E_{i,j} \cup E_{j,i}$.

  Let $e = \langle c_i, c_j \rangle$ be an edge in H that connects two different partitions P_i and P_j. The cost of breaking e, $cost(e)$, (i.e., the cost of the violation of the structural constraints induced by e) is $1+d_j$, being d_j the number of descendants of c_j in the H that also belong to P_j.

Thus, the taxonomy H', minimizing the errors due to structural constraint violations is constructed as follows:

1. create a complete weighted directed graph, $G_P(V_P, E_P, w_P)$, of partitions, where
   - $V_P = P$,
   - E_P is the set of edges between all pairs of partitions, and
   - $w_P(\langle P_i, P_j \rangle) = \sum_{e \in E_{j,i}} cost(e)$;
2. find a *maximum spanning tree* of G_P rooted at the partition P_{root},

For example, let us re-consider the taxonomy fragment and its partitions shown in Figure 1(a). In the adapted hierarchy (Figure 1(b)), `ANITA` picks as root the partition containing the root node ("science"). Then, the remaining two partitions will be attached to it by analyzing the constraints given by the broken edges. Note that the distillation process can alter the structure of the hierarchy, since the relationship among concepts could change from one domain to another one. For example, in "terrorism"-related news articles, two geographical concepts as "USA" and "Afghanistan" will result strongly related, while in an "economical" context they can be very far from each other. Therefore, considering the knowledge expressed by the domain expert in the original taxonomy, ANITA tries to preserve the original relationships among concepts, but alters the structure when there is sufficient evidence in the corpus that a different structure would reflect the content better.

3.4.2 Partition Labeling

In order to select a representative label for each partition we need to analyze the obtained partition in the context of the original structure. In order to pick a label for the node c'_i associated to P_i, we consider the structural relationships in the original hierarchy H among the nodes in P_i. If there is a concept $c_i \in P_i$ that dominates all the other nodes in the partition (i.e., $\forall c_j \in P_i \;\; c_j \preceq c_i$), then the label of c_i is selected as the label of c'_i. If there is no such single node, then the minimal set D_i of nodes covering the partition P_i (based on H) is found and the concatenation of the concept labels in D_i, is used as the partition label. An example can be seen in Figure 1(b).

4. EVALUATION

In our experiments, we used two different data sets: a corpus of news articles from New York Time data set [2] (containing 64,000 text entries with over 100,000 unique keywords) and a set of scientific abstracts from National Science Foundation[3] (containing 49,000 article abstracts describing NSF awards for basic research, with over 30,000 unique keywords).

For each data set, we considered a corresponding domain taxonomy extracted from the *DMOZ* categorization[4]. Specifically, we considered a taxonomy of science (with 72 nodes) which we contextualized in the domain of the NSF abstracts, and geographical taxonomy (181 nodes), against which we classified the articles from the New York Times.

[1]Note that there can be cases in which the specific target value cannot be reached within a predetermined number of steps; in such a case we use the λ_{max} value that leads to a partition set closest to the targeted size.

[2]http://archive.ics.uci.edu/ml/datasets/Bag+of+Words

[3]http://kdd.ics.uci.edu/databases/nsfabs/nsfawards.html

[4]accessible at the link *http://www.dmoz.org/*

Note that, to experiment with different taxonomies for the same domain, we selected different subsets of these original taxonomies by randomly removing some of their nodes. Specifically, we created a total of 18 distinct taxonomies for each domain, obtained by removing anywhere between 10% to 60% of the concepts the original DMOZ taxonomies. The results reported in this section for each domain are averages of the results for all the taxonomies.

4.1 Effectiveness Measures

In order to better understand the behavior of ANITA under different settings and to compare its performance to other algorithms on a concrete basis, we quantify the quality of the adapted taxonomies using the following 4 measures:

1. *Domain coverage*: An important role of taxonomies in many applications is to help provide search and access to text documents. Therefore, it is essential that they properly reflect the content of the corpus. Given a corpus of documents D and a taxonomy $H(C, E)$, the coverage of D by H is the percentage of documents in D that can be associated to at least one concept in C using the process described in Section 3.2. The higher the coverage, the more effective the taxonomy in indexing the documents in D.

   Let $A_{c_i} \subseteq D$ be the set of documents associated to the concept $c_i \in C$. We define the domain coverage measure as

   $$cover(H, D) = \frac{\left|\bigcup_{c_i \in C} A_{c_i}\right|}{|D|}.$$

   Note that the labels of some nodes in the adapted taxonomy may consists of the concatenation of multiple labels from the original taxonomy. In the case of a concatenated label, a document is considered a match for the corresponding node as long as the keyword vector of the document matches the concept-vectors of at least one of the constituent labels.

2. *Redundancy*: Note that it would be trivial to increase the domain coverage simply by concatenating more and mode labels. This would not result in a desirable taxonomy. Therefore, it is important to quantify other properties, such as the degree of discrimination of the nodes of the hierarchy, along with domain coverage.

   The redundancy measure, defined as

   $$redundancy(H, D) = \frac{|overlap(D, H)|}{\left|\bigcup_{c_i \in C} A_{c_i}\right|},$$

   where $overlap(D, H)$ returns the set of documents in D associated to *at least* two concepts in H. Thus, this formula quantifies the discrimination power of the concepts in the resulting taxonomy, i.e, the degree of overlapping in the sets of documents associated to different concepts. The lower the redundancy, the higher the discrimination power, and thus the more effective the taxonomy in helping search and access text documents.

3. *Coherence*: it measures the effectiveness of a taxonomy with respect to the homogeneity of the documents corresponding to each individual node into the hierarchy.

Context: NSF Corpus				
	cover.	redund.	coher.	Ltl
Pre-Order (dist. pres.)	0,123	0,551	0,106	1,724
Parenth. (dist. pres.)	0,128	0,510	0,097	1,681
Post-Order (dist. pres.)	0,128	0,530	0,088	1,702
Pre-Order (no dist.pres.)	0,128	0,725	0,046	1,423
Parenth. (no dist.pres.)	0,125	0,729	0,049	1,402
Post-Order (no dist.pres.)	0,128	0,736	0,053	1,463

Table 1: Impact of different narrative orders

Context: NYTimes Corpus				
	cover.	redund.	coher.	Ltl
Pre-Order (dist. pres.)	0,752	0,634	0,082	2,289
Parenth. (dist. pres.)	0,759	0,573	0,081	2,2044
Post-Order (dist. pres.)	0,755	0,612	0,088	2,277
Pre-Order (no dist.pres)	0,755	0,792	0,063	2,0634
Parenth. (no dist.pres.)	0,758	0,789	0,068	1,966
Post-Order (no dist.pres.)	0,756	0,792	0,060	1,809

Table 2: Impact of different narrative orders

   The higher the average intra-class mutual similarities among documents, the more coherent the nodes of the taxonomy:

   $$coherence(H, D) = \frac{\sum_{c_i \in H} \left(\frac{\sum_{d_k, d_h \in A_{c_i}} cos(\vec{d_k}, \vec{d_h})}{|A_{c_i}| \cdot |A_{c_i}|} \right)}{|C|}.$$

4. *Label term-length*: Finally, the *label term-length* measure simply reports the average number of labels in the original taxonomy included in the labels of the adapted hierarchy.

   Given an initial taxonomy $H(C, E)$ and its adapted version $H'(C', E')$, let $length(label_{c'_i}, H, H') = l$ iff $label(c'_i) = label(c_{j1}), \ldots, label(c_{jl})$, with $c_{j1}, \ldots, c_{jl} \in C$. Then, label term-length is defined as

   $$ltl(H, H') = \frac{\sum_{c'_i \in H'} length(label_{c'_i}, H, H')}{|C'|}.$$

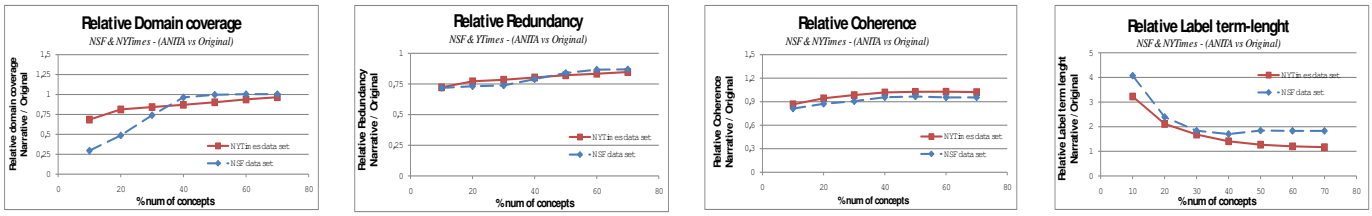
In the rest of the paper, we present experiment results that rely on these four measures.

4.2 Impact of the Narrative Orders

Tables 1 and 2 present the values of the effectiveness measures for the three proposed narrative orderings, with and without distance preserving sibling ordering. The values presented in the table are averages of the performance results for five different target taxonomy sizes (from 10% to 50% of the original number concepts).

From these two tables, we observe that sibling ordering results in slightly higher label term-length. This behavior is due to the fact that the ordering of siblings is likely to lead to longer sequences of similar siblings, which will be concatenated if the sequence does not contain the parent. However, it is important to note that this lengthening of the labels does not result in any increase in the redundancy or drop in the coherence of the resulting taxonomies. In all cases, the versions with sibling ordering has significantly smaller redundancy than the corresponding versions with the random ordering of siblings. Moreover distance preserving sibling ordering also provides significantly higher coherences. The differences in terms of their domain coverages are negligible.

When we consider the different traversal strategies, we observe that, for both data sets, parenthetical traversal generally provides lower redundancies. In terms of coherence, however, there is no clear winner among the three traversal



(a) (b) (c) (d)

Figure 3: (a) Domain coverage, (b) redundancy, (c) coherence, and (d) label term-length ratio ($\frac{\texttt{ANITA}}{Original}$) curves. The two curves on each of the charts correspond to the NSF and NY Times data sets.

strategies. In general, parenthetical traversal leads to lower label term-length. Parenthetical traversal also provides high coverage, especially when distance preserving sibling ordering is used. Therefore, in the rest of the section, we only consider the parenthetical traversal with distance preserving sibling ordering.

4.3 Comparison wrt. the Original Taxonomy

In this subsection, we quantify how much difference in coverage, redundancy, and coherence with respect to the original taxonomy occurs for varying target taxonomy sizes. Figure 3 shows the ratios between the considered effectiveness measures on the adapted and the original taxonomies (the two curves in these charts correspond to the NSF and NY Times data sets). Figure 3(a) shows that, for both data sets, the relative domain coverage is very close to 1.0 for adaptations with $\geq 30\%$ of the nodes; this means that the adapted taxonomy can index the same amount of contents as the original taxonomy. As expected, the coverage drops when the size of the adapted taxonomy is pushed further down, even though `ANITA` increases the label length to compensate for this drop. Note that, despite this increase in the label lengths, `ANITA` is still able to lower the redundancy in the taxonomy, while preserving the coherence of the adapted taxonomies even when the compression rates are lowered down to 10% range. Finally, note that the similarities between the NSF and NY Times curves on these charts highlight that the performance of `ANITA` (especially in terms of redundancy and coherence) is largely independent of the data set.

4.4 ANITA vs. other Concept Clustering Methods

In Figure 4 we compare the impacts of narrative-based partitioning against k-Means clustering, with k also being equal to the target taxonomy size requested from `ANITA`. In both cases, extended vector representation of the taxonomy nodes are used to support partitioning. Also, in both cases, once the partitions are obtained, the same taxonomy re-construction and labeling strategies (described in Section 3.4) are used to stitch the taxonomy back.

In this experiments, we considered target taxonomy sizes between 10% and 70%. Each point in these charts denotes the performance measure obtained in a single experiment using k-Means (y-axis) vs. `ANITA` (x-axis) on the same input. These results show that while the two clustering approaches show similar behaviors in terms of domain coverage, coherence, and label length (Figures 4(a),(c), and (d)), `ANITA` provides significant gains in terms of lowering the amount of redundancy (Figure 4(b)) in the taxonomy. This is consistent with the key design goals of `ANITA`; i.e., creating compact taxonomies that provide high category differentiation (to support effective navigation), while not losing in terms of domain coverage or coherence.

Context: NSF+NYTimes Corpora				
	Ltl	cover.	redund.	coher.
ANITA	1.840	0,266	0,737	0,075
EM	4.434	0,121	0,794	0,045
X-Means	5.314	0,169	0,775	0,056
H-EM	6.340	0,232	0,797	0,067

Table 3: `ANITA` (without a target taxonomy size) vs. EM, X-Means, and Hierarchical-EM

We also compared narrative-based partitioning approach with non-parametric clustering methods which do not require the number of target clusters as input. For these experiments, we run `ANITA` without providing the target taxonomy size input. Instead, we stopped the segmentation process described in Section 3.3 after its first step, with the initial drifting threshold λ_{max} set to 0.5. Table 3 compares the results obtained by `ANITA` against results obtained by clustering algorithms, such as EM, $X-Means$, and $Hierarchical-EM$ ($Hierarchical-EM$ method applies EM clustering strategy to each sibling group). As these results show, for this λ_{max} value, `ANITA` provides better results in terms of all parameters against these alternative clustering strategies.

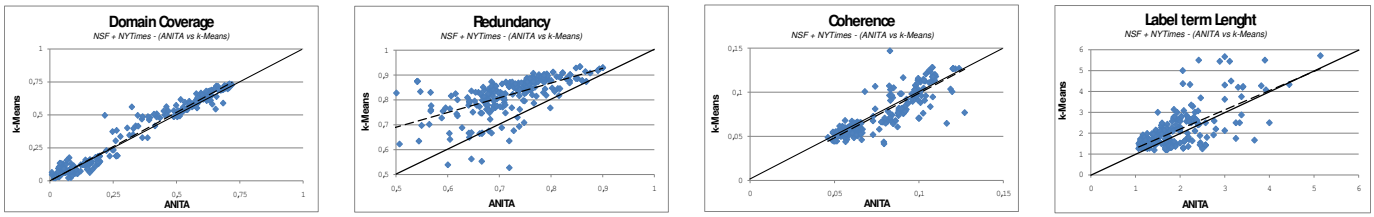
4.5 Impact of the Corpus Context

Lastly, we evaluate the impact of the using the corpus context, represented by the extended vectors introduced in Section 3.2. In particular, we compare the application of `ANITA` on these extended vectors with a version of `ANITA` where the narrative structure is created on the CP/CV vectors (Section 3.1.1) which reflect only the structure of the taxonomy and do not take into account the corpus in any way. This corresponds to omission of the *contextual re-interpretation* step described in Section 3.2.

The four charts in Figure 5 plots the performance ratio, $\frac{\texttt{ANITAwithcontext}}{\texttt{ANITAwithoutcontext}}$ for both NSF and NY Times data and for different target taxonomy sizes.

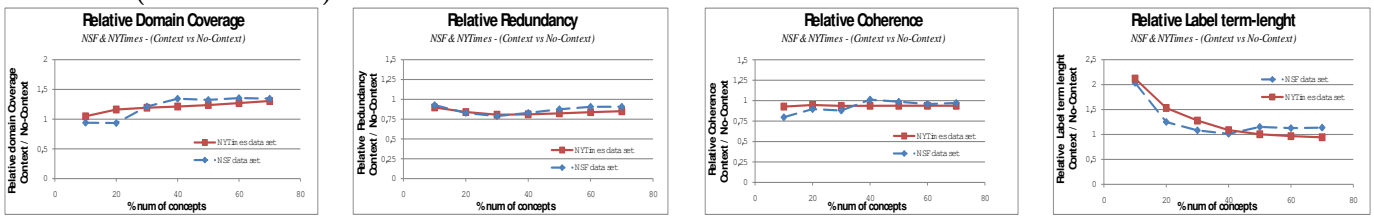
For both data sets, the use of corpus context based reinterpretation of the narrative promotes improved domain coverage and lower redundancy (Figures 5(a) and (b)). In terms of the lengths of the term labels, especially for very low target taxonomy sizes, the ratio is close to 2.0 for both data sets, indicating that the use of context results in longer descriptors. This, however, does not result in a reduction in coherence as shown in Figure 5(c): for both data sets, the coherence ratio is close 1.0, indicating that the use of context corpus does not impact the coherence of the nodes with respect to each other at all.

Finally, once again, the almost identical behaviors of `ANITA` with respect to these two very different data sets (and



(a) (b) (c) (d)

Figure 4: (a) Domain coverage, (b) redundancy, (c) coherence and (d) label term-length for ANITA and k-Means (both data sets)



(a) (b) (c) (d)

Figure 5: (a) Domain coverage, (b) redundancy, (c) coherence, and (d) label term-length ratio ($\frac{\texttt{ANITAwithContext}}{\texttt{ANITAwithoutContext}}$) curves. The two curves on each of the charts correspond to the NSF and NY Times data sets.

corresponding taxonomies) provide strong evidence that the results presented in this section are not data set specific.

5. CONCLUSIONS

In this paper, we introduced a novel narrative interpretation of a taxonomy, where it is viewed as a discourse describing the relationships among concepts/categories in a given domain, for re-structuring existing taxonomies to varying application contexts. The experimental results showed that the proposed *A Narrative Interpretation of Taxonomies for their Adaptation* (`ANITA`) technique provides significant benefits in terms of reducing the redundancies in the taxonomies, while improving their domain coverages relative to the given corpora of documents.

6. REFERENCES

[1] C. C. Aggarwal, S. C. Gates, and P. S. Yu. On the merits of building categorization systems by supervised clustering. In *KDD-99*, pages 352–356.

[2] H. Chen and T. Ng. An algorithmic approach to concept exploration in a large knowledge network (automatic thesaurus consultation): symbolic branch-and-bound search vs. connectionist hopfield net activation. *J. Am. Soc. Inf. Sci.*, 46(5):348–369, 1995.

[3] M. H. Computer and M. A. Hearst. Texttiling: A quantitative approach to discourse segmentation. Technical report, 1993.

[4] M. A. Hearst. Texttiling: segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.*, 23(1):33–64, March 1997.

[5] T. Hofmann. The cluster-abstraction model: Unsupervised learning of topic hierarchies from text data. In *In IJCAI*, pages 682–687, 1999.

[6] E. H. Hovy. Automated discourse generation using discourse structure relations. *Artif. Intell.*, 63(1-2):341–385, 1993.

[7] A. K. Jain and R. C. Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

[8] H. Jing. Summary generation through intelligent cutting and pasting of the input document, 1999.

[9] S. Kaufmann. Cohesion and collocation: Using context vectors in text segmentation. In *ACL*, 1999.

[10] J. W. Kim and K. S. Candan. Cp/cv: concept similarity mining without frequency information from domain describing taxonomies. In *CIKM '06*.

[11] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999.

[12] T. Li, S. Zhu, and M. Ogihara. Hierarchical document classification using automatically generated hierarchy. *J. Intell. Inf. Syst.*, 29(2):211–230, 2007.

[13] K. R. McKeown. *Text generation: using discourse strategies and focus constraints to generate natural language text*. Cambridge University Press, 1985.

[14] J. D. Moore and C. L. Paris. Planning text for advisory dialogues. In *Association for Comp. Ling.*, pages 203–211, 1989.

[15] K. Punera, S. Rajan, and J. Ghosh. Automatically learning document taxonomies for hierarchical classification. In *WWW '05*. ACM, 2005.

[16] I. Ruthven and M. Lalmas. A survey on the use of relevance feedback for information access systems. *Knowl. Eng. Rev.*, 18(2):95–145, 2003.

[17] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. In *Information Processing and Management*, pages 513–523, 1988.

[18] J. Savoy. Ranking schemes in hybrid boolean systems: a new approach. *J. Am. Soc. Inf. Sci.*, 48(3), 1997.

[19] E. Segal, D. Koller, and D. Ormoneit. Probabilistic abstraction hierarchies. In *In Advances in Neural Information Processing Systems 14*. MIT Press, 2001.

[20] A. Shakery and C. Zhai. Relevance propagation for topic distillation uiuc trec 2003 web track experiments. In *TREC*, pages 673–677, 2003.

[21] L. Tang, H. Liu, J. Zhang, N. Agarwal, and J. J. Salerno. Topic taxonomy adaptation for group profiling. *ACM TKDD*, 1(4):1–28, 2008.

[22] W. S. Torgerson. *Theory and methods of scaling*. R.E. Krieger Pub. Co.