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

Mario Cataldi
Università di Torino
Torino, Italy
cataldi@di.unito.it

K. Selçuk Candan
Arizona State University
Tempe, AZ 85283, USA
candan@asu.edu

Maria Luisa Sapino
Università di Torino
Torino, Italy
mlsapino@di.unito.it

ABSTRACT

Taxonomies 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 very dynamic content, they do not necessarily reflect the content knowledge. Thus, in this paper, 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 by user studies that show that the proposed algorithm is able to adapt the taxonomy in a new compact and understandable structure from a human point of view.

Categories and Subject Descriptors

H.5.4 [**Hypertext/Hypermedia**]: Navigation; I.7.3 [**Docu ment and Text Processing**]: Index Generation; H.3.3 [**Information Search and Retrieval**]: Information Filtering, Selection Process

General Terms

Algorithms, experimentation

Keywords

taxonomy adaptation, taxonomy segmentation, taxonomy summarization

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 properly reflect important details within the considered data collection.

In this paper we introduce a new method for distilling a taxonomical domain categorization from an existing one, *within the context of* a set of text documents that have to be represented and indexed by it. Thus, recognizing that the primary role of a taxonomy is to describe or narrate the natural relationships between concepts in a given domain to its users, we propose *A Narrative Interpretation for Taxonomy Adaptation* (ANITA), a novel distillation approach for adapting existing taxonomies to varying application contexts.

2. RELATED WORK

In the literature, many authors tried to automatically extract hierarchical categorizations from text corpora. [2] presents an overview about the many methodologies that have been proposed to automatically extract structured information from texts. In [10], Sanderson et al. present an unsupervised method to automatically derive from a set of documents a hierarchical organization of concepts, using cooccurrence information.

One of the most critical tasks is the definition of the semantical relationships among the retrieved concepts: [4] organized the extracted concepts by analyzing the syntactic dependencies of the terms in the considered text corpus. Many other methods rely on preliminary supervised operations to limit the noise in the retrieved concepts: in [1], the user sketches a preliminary ontology for a domain by selecting the vocabulary associated to the desired elements in the ontology (this phase is called lexicalisation). In [8], Ponzetto et al. investigate the problem of automatic knowledge acquisition from Wikipedia repositories.

Evaluation of the quality of automatically generated taxonomies is a very important and non-trivial task. In [11], authors determine the precision of the clustering algorithm by manually assigning a relevance judgment to the documents associated to the clusters. In [12], authors use the F-Score to evaluate the accuracy of the document associations (but the approach requires a ground truth, which is hard to determine in many cases). In [9] authors perform a user study to evaluate the qualities of the relationships between concepts and their children and parent concepts. In [6], the quality of the concepts is measured by evaluating their ability to find documents within the hierarchy.

CIKM'10, October 26–30, 2010, Toronto, Ontario, Canada.
Copyright 2010 ACM 978-1-4503-0099-5/10/10 ...$10.00.

3. NARRATIVE-DRIVEN TAXONOMY ADAPTATION PROCESS

Given an input taxonomy $H(C, E)$ (also called hierarchy in the paper) where $C = \{c_1, \ldots, c_n\}$ represents the set of n concept nodes (or categories) and E is the set of structural edges, our goal is to create an adapted taxonomy $H'(C', E')$, based on a given context defined by a corpus, D, of text documents. As described before, ANITA relies on a "narrative" interpretation of the input taxonomy to achieve this goal. Unlike the original taxonomy, which is hierarchical, the narrative is linear in structure; however, it is created in such a way that the structure of the narrative corresponds to the structure of the hierarchy. More specifically, the scope of each concept (represented as a sentence) is contextualized by those that precede and follow it, and this contextual scope corresponds to both the structural information (coming from the original structure) as well as the content of the considered corpus.

3.1 Step I: Narrative View of a Taxonomy

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 map concept-nodes of the input taxonomy into *concept-sentences*. What we refer to as concept-sentences are not natural language sentences, but vectors obtained by analyzing the structure of the given taxonomy and the related corpus of documents. Intuitively, these sentence-vectors can be thought of as being analogous to *keyword-vectors* commonly used in representing documents in IR systems.

Concept-sentences associate to each concept a coherent set of semantically related keywords, extracted from the associated text corpus. Thus, for each concept c_i in the considered hierarchy, we associate a sentence-vector $\vec{sv_{c_i}}$ as

$$\vec{sv_{c_i}} = \{w_{i,1}, w_{i,2}, w_{i,3} \cdots w_{i,v}\}$$

where v represents the total number of considered terms (the corpus vocabulary and labels in the taxonomy), and $w_{i,j}$ represents the semantical correlations between the j-th term and the i-th taxonomical concept.

Concept-sentences can be obtained in many different ways; [3], [4], [7] propose various approaches that leverage semantic similarities between concepts in a given context for obtaining such vectors. In this paper, we use the approach proposed in [3] to associate to each concept a keyword-vector, that tightly integrates terms extracted from text documents and labels of concepts obtained from the considered domain taxonomy. Thus, the resulting vectors reflect both the structural context (imposed by the taxonomy) and the documents' content (imposed by the corpus).

After the vector-based encoding of the *concept-sentences*, the creation of the narrative is completed by ordering these sentences (therefore the nodes in the original hierarchy) in an order representing the structure of the taxonomy. A 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 observation to create a narrative: the sentences associated to the nodes of the taxonomy can be read in different orders to obtain the narrative. In the evaluations presented in Section 4, we consider the pre-order reading, where each concept-sentence is immediately followed by its detailed description in terms of its specializations.

3.2 Step II: Segmentation of the Narrative

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

In the next step, we analyze this narrative 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, 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 this paper, in order to partition the narrative $\vec{sv}_1, \vec{sv}_2, \ldots, \vec{sv}_n$ into coherent segments, we seek partitions that correspond to similar internal coherence (defined in terms of the total amount of internal topic drift):

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

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

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 $\vec{sv}_i$ and $\vec{sv}_j$, 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 it holds that $drift_{i,j} < \lambda_{max}$, where $\lambda_{max} = \frac{drift_{1,n}}{k}$ is the *coherence threshold*, and k is the target size of the summarized taxonomy.

At the end of the process, we have a set of segments, or partitions, $P = \{P_1, P_2, \cdots, P_k\}$ that represent sequences of coherent narrative components. Note that, each partition is a sequence of concepts from the original taxonomy and defines a single concept in the revised taxonomy

3.3 Step III: Taxonomy Reconstruction

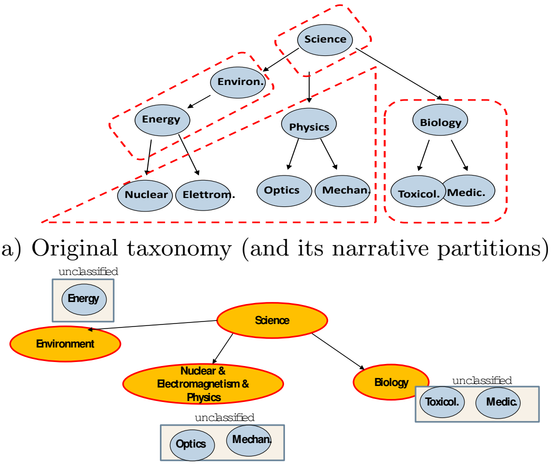
In order to construct the adapted taxonomy from the partitions created in the previous step, we need to re-attach the partitions in the form of a tree structure. Furthermore, for each partition, we need to pick a *label* that will be presented to the user and will describe the concepts in the partition.

3.3.1 Step IIIa: 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:

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

- Let us consider a pair, P_i and P_j, of partitions in P. The decision on whether (and how) the corresponding concepts



(a) Original taxonomy (and its narrative partitions).

(b) Re-constructed (or distilled) taxonomy.

Figure 1: Narrative-based adaptation of a taxonomy fragment: based on the structural constraints and the available NSF content (described in Section4) (a) the partitions are linked to each other. Finally, (b) each partition is labeled by selecting a representative label.

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 strength of the structural constraints implied by E in H, and we propose as our solution the adapted taxonomy which maximally preserves such constraints.

Let $e = \langle c_a, c_b \rangle$ be an edge in H that connects two different partitions P_i and P_j (i.e. $c_a \in P_i$, $c_b \in P_j$). The strength of the structural constraint e, $strength(e)$, (i.e., the strength of the structural constraints induced by e) is $1 + d_b$, being d_b the number of descendants of c_b in H that also belong to P_j.

Based on this, the decision of having the corresponding c'_i as the ancestor of c'_j is supported by the strength of the structural constraints associated to the edges in $E_{i,j}$.

Thus, the taxonomy H', is constructed by maximally preserving such constraints 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
   - $w_P(\langle P_i, P_j \rangle) = \sum_{e \in E_{i,j}} strength(e)$;
2. find a *maximum spanning tree* of G_P rooted at the partition P_{root},

For example, let us 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 three partitions have been attached to it by analyzing the constraints given by the structural original edges. Note that the distillation process can alter the structure of the hierarchy, since the relationships among concepts could change from one domain to another one. For example, in a political context, concepts "*nuclear*" and "*environment*" will may found to be strongly related, while in a context of a scientific taxonomy, the concept "*nuclear*" may be more rigorously related to the concept "*physics*" (in fact, as shown in Figure 1(b), considering the NSF awarded abstracts, "*nuclear*" has been connected to "*physics*"). Therefore, considering the knowledge expressed by the domain experts 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.

3.3.2 *Step IIIb: Partition Labeling*

In order to select a representative label for each partition we need to analyze the obtained partitions 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., $\forall c_j \in P_i \quad c_j$ is a descendant of 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. Intuitively, a concatenation implies that, in the given document context, these corresponding concepts are found to be not sufficiently distinguished from each other. On the other hand, any label that was in the original taxonomy, but is not included in the new taxonomy is found to be unnecessary in the new context. An example can be seen in Figure 1(b).

4. USER STUDY

In order to analyze the benefits of using an ANITA adapted categorization for text data indexing purposes, we conducted a user study and evaluate the feedback of 16 users when exploring a set of scientific abstracts from National Science Foundation[1] (~50K article abstracts describing NSF awards for basic research, with over ~30K unique keywords) using different taxonomies.

The users represent various range of ages, backgrounds, jobs and education level and they have intermediate web ability (they are not computer scientists or domain experts).

We presented to the users, three different taxonomies that indexed NSF documents: the original portion of DMOZ-extracted taxonomy, with 72 concepts (obtained considering the most relevant terms, in the considered domains, extracted from the corpora), its ANITA-based adaptation with 13 concepts (with k randomly set to 13) and the k-Means based adaptation (with same value of k) [2]. In order to avoid bias in the evaluation of the presented taxonomies, we presented the 3 taxonomies to the user in a random order.

Search Time and Interaction Counts.

Given a randomly selected concept label extracted from the original taxonomy (different for each partecipating user),

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

[2] For the k-Means clustering, the sentence-vector representation of the taxonomy nodes is used to support partitioning. Also, once the partitions are obtained, the same taxonomy re-construction and labeling strategies (described in Section 3.3) are used to stitch the taxonomy back.

Context: NSF Corpus		
	avg time (sec)	avg num. of interactions
Original (72 concepts)	23.5	5.1
ANITA (13 concepts)	9.7	2.3
k-Means (13 concepts)	11.0	2.9

Table 1: User Study: Average time and average number of interactions (clicks on the structure for expanding or collapsing nodes) per taxonomy, when the users explore the structure to retrieve documents related to a randomly selected concept.

we asked the users, for each presented taxonomy, to retrieve related documents by exploring the presented categorizations. Therefore, we analyze the time and the number of interactions (in terms of expansions/collapses of the presented nodes in the taxonomies) the user needs to reach satisfactory documents. As reported in Table 1, ANITA adapted taxonomy reports significant gains in terms of time (from an average of 23.5 seconds to an average of 9.7) and number of interactions (from 5.1 to 2.3) by reducing the number of nodes the user has to navigate through. On the other hand it is important to note that, even if k-Means adapted taxonomy presents the same number of nodes as ANITA, it is not able to guide the user as well as ANITA adapted taxonomies do; the user needs more time to find relevant documents (an average of 11.0 seconds) and more interactions to retrieve the appropriate contents (an average of 2.9 operations). Thus, we can state that ANITA is not only able to reduce the cardinality of the selected taxonomy, but also organizes the concepts in such a way to facilitate the retrieval operations.

Classification Accuracy.

Given a randomly selected article (different for each considered user), extracted from the considered NSF corpus of documents, we asked to the users, for each presented taxonomy, to select those nodes (if any) that would best represent the selected content. Then we compared these user associations with the ones automatically provided by the system [3], calculating the percentage of shared concepts associated. All the considered users provided, for each document, between two and three associated concepts per taxonomy. The results indicate that, for the original taxonomy, 67.7% of the user selected concepts were shared by the system. Similarly, the ANITA-based adapted taxonomy provides a 68.7% of shared concepts, indicating that the quality of the taxonomy is as good as original one despite containing much smaller number of concepts. On the other hand, the k-Means based adapted taxonomy does not perform well: only 37.4% of the user selected concepts had been effectively associated by the system to such nodes, highlighting the fact that a naive re-structuring process (such as k-Means) can cause a significant increase in terms of confusion and disorganization.

Subjective Questionnaire Measures.

After the study, each user also completed a brief questionnaire which included two questions ("Is the taxonomy easy to use?" and "Is the taxonomy sufficiently detailed?"); the users could quantify the responses using a 5-point scale ratings.

Context: NSF Corpus		
	easy to use	sufficiently detailed
Original (72 concepts)	4.1	3.8
ANITA (13 concepts)	4.1	3.6
k-Means (13 concepts)	3.3	2.6

Table 2: Subjective questions in the user study: for each question, each user has quantified her opinion by a 5-point scale rating.

As show in Table 2, the users reported that the ANITA adapted taxonomy was as "easy to use" as the original one (both 4.1) while the k-Means adapted taxonomy was significantly harder to use (3.3). Moreover, even if the number of presented nodes was dropped almost 80%, the users commented that, in terms of providing "sufficient details" (i.e., the number of alternatives), ANITA adapted taxonomy provides a good range of details, close to the original one (3.6 vs 3.8). We can summarize these results as follows: as initially supposed, the original taxonomies, developed by domain experts for broad coverage of documents, provide unnecessary details that can be removed without causing significant loss in terms of contextual knowledge. On the other hand, a general adaptation method such as k-Means, could introduce confusion and disorientation: the k-Means adapted taxonomy significantly reduces the "sufficiency" (only 2.6) and results in taxonomies that the users find harder to use (3.3 in terms of "easy to use").

5. CONCLUSIONS

In this paper, we introduced *A Narrative Interpretation of Taxonomies for their Adaptation* (ANITA) for re-structuring existing taxonomies to varying application contexts. The user studies validated the proposed technique from a human point of view.

6. REFERENCES

[1] C. Brewster, F. Ciravegna, and Y. Wilks. User-centred ontology learning for knowledge management. In *NLDB '02*, pages 203–207. Springer-Verlag, 2002.

[2] P. Buitelaar, P. Cimiano, and B. Magnini, editors. *Ontology Learning from Text: Methods, Evaluation and Applications.* IOS Press, 2005.

[3] M. Cataldi, C. Schifanella, K. S. Candan, M. L. Sapino, and L. Di Caro. Cosena: a context-based search and navigation system. In *MEDES '09*, pages 218–225. ACM, 2009.

[4] P. Cimiano, A. Hotho, and S. Staab. Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Intelligence Research*, 24:305–339, 2005.

[5] J. W. Kim and K. S. Candan. Cp/cv: concept similarity mining without frequency information from domain describing taxonomies. In *CIKM '06*, 2006.

[6] D. J. Lawrie and W. B. Croft. Generating hierarchical summaries for web searches. In *SIGIR '03*, pages 457–458, 2003.

[7] C. Muller, I. Gurevych, and M. Muhlhauser. Integrating semantic knowledge into text similarity and information retrieval. In *ICSC '07*, pages 257–264, 2007.

[8] S. P. Ponzetto and M. Strube. Deriving a large-scale taxonomy from wikipedia. In *AAAI*, pages 1440–1445, 2007.

[9] M. Sanderson. Word sense disambiguation and information retrieval. In *SIGIR '94*, pages 142–151. Springer-Verlag New York, Inc., 1994.

[10] M. Sanderson and B. Croft. Deriving concept hierarchies from text. In *SIGIR'99*, pages 206–213, 1999.

[11] O. Zamir and O. Etzioni. Web document clustering: a feasibility demonstration. In *SIGIR '98*, pages 46–54, 1998.

[12] Y. Zhao and G. Karypis. Evaluation of hierarchical clustering algorithms for document datasets. In *Data Mining and Knowledge Discovery*, pages 515–524. ACM Press, 2002.

[3] In order to obtain this classification, without loss of generality, we rely on the concept-vectors introduced in [5], by quantifying, as usual, the cosine similarity between the document keyword-vector (containing term frequencies) and the concept sentence-vector.