

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

A narrative interpretation of taxonomies for their adaptation to text collections. ANITA (ANITA) is a system for the automatic adaptation of taxonomies to text collections. It is based on a set of rules that allow the system to automatically generate a narrative interpretation of a taxonomy. The system is able to handle taxonomies of arbitrary complexity and to generate a narrative interpretation that is suitable for use in text collections. The system is implemented in Java and is available as a free software package.

Categories and Subject Descriptors

F.4 Hypertext/Hypermedia; I.2 Document and Text Processing; I.3 Information Search and Retrieval

General Terms

Algorithms; Theory of computation

Keywords

taxonomy; adaptation; text collections; ANITA

1. INTRODUCTION

The purpose of this paper is to present ANITA, a system for the automatic adaptation of taxonomies to text collections. ANITA is based on a set of rules that allow the system to automatically generate a narrative interpretation of a taxonomy. The system is able to handle taxonomies of arbitrary complexity and to generate a narrative interpretation that is suitable for use in text collections. The system is implemented in Java and is available as a free software package.

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.

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

t a g a t o a a a s s t o t a a a d e o t s -
c t u a t v g a s t o r t t c a h y t s t s e t a s v
t o t a o a t a t g o a t o t a t s t s a s s
t o t t s i a t t a a a d t a c o n y s a l s l a v
s e p o s c a c e a g e o p e a t s a o s e
n y a a g t o o v t u n y o t a t t a s w t t
e o s a t a e o t o
i t s a w t o l a w n t o o s t g
a t a c o n y a o n y a t a t g o a t o r o n y a i s t g o
w i t h i n t h e c o n t e x t o f a s t o r t t c a h y t s t a t a t o s
s t a s t a s t l s g e g t a t t
n y a v o c p a t a e o n y s t o s a s o a a t t a t
u a a t o s s o t w e a t s a g o n y a t o t s
l s s w o e s A N a r r a t i v e I n t e r p r e t a t i o n f o r T a x o n o m y
A d a p t a t i o n (A N I T A) a o s t a t o a c a e c a a t
g i s t g a t o n y s t o a v g a a t o e o t s

2. RELATED WORK

i t t a t u n y a a l t o s t t o a l t o n a t a v
t a t a a a a t a t o a t o s e o n t t t o c a 2
s t s a o w a e l t n y a n t o c e s t a t
a s o e s t o a l t o n a t a v t a t s t l t u
o n y a t o n y t s i f o a s o t a s t a
u s u s n t o t o a l t o n a t a v r o n y a s t o
c a h y t s a a a a e g a a t o c p e a t s l s g e o
c u a c o n a t o
o c t n e s t a t a t a s s s t t o c o r t s -
n y a t a a t o s s a n o g t t e o e t s
e g a t t a t a t o t s o v a a v g t s v t a e
t a v c t n t o s v o n y a v s u s o -
a t o s t o n t t o s t t e o a t s i
t l s s t a s a n y a v o t o e v o a o n y a s v
s a t g t c e a u a v a s s o a t o t s n y t s
t o t o e v (t s a s s e a a a s a t o) i 8
o t t o t a s t r a t t e n y o a l t o n a t a v e w
g a a u s t o r o n y l a t o s t o s
E a l t o c o r t l a t v o n a l t o n a t a v g a t a t a -
o n y s s a v n o t a t a o t a t a s i i p
a l t o s t n y t a s o c o r t e l s t g a g o t n y
s v n y a l a v a s s e g a a a j u g n t t o t c a l l
n y t s a s s o a t t o t a l s t s i i 2 a l t o s l s t
i - e o t o a l a t t a a u a v c o r t c a h y t a s s o a -
t o s e l t a c a u s a g e u t t w e s
a t o t n y n y a v a s s) i l a l t o s o n y
a l s s t u v t o a l a t t l a t s c o r t a t o s s
s t w e o a t s a t a a a t a o a t s
i 6 t l a t v c o r t e o a t s s n y a s u s v a l a t g
t a d t v t o c a h y t s w t t a e v

3. NARRATIVE-DRIVEN TAXONOMY ADAPTATION PROCESS

Given a text $T \in H(C, E)$ (a sequence of words t_1, \dots, t_n) and a set of clusters $\mathcal{C} = \{c_1, \dots, c_n\}$ (represented as $\{c_1, \dots, c_n\}$), we adapt the taxonomy by iteratively refining the clusters based on the narrative structure. The process involves identifying key concepts and their relationships within the text, such as the example of ANITA's story.

3.1 Step I: Narrative View of a Taxonomy

As a narrative T is processed, we extract *concept-sentences* and *keyword-vectors* from the text. These are used to associate each cluster c_i with a set of keywords $\{w_{i,1}, w_{i,2}, \dots, w_{i,v}\}$.

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

We then calculate the *drift* between clusters c_i and c_j based on their associated keyword vectors. The drift is defined as the cosine distance between the vectors.

The drift between clusters c_i and c_j is calculated as $drift_{i,j} = 1 - \cos(sv_{c_i}, sv_{c_j})$. This metric helps in identifying clusters that are too similar and should be merged or separated.

At the end of the process, we have a refined taxonomy where clusters are better aligned with the narrative content. This involves adjusting the clusters based on the identified concept-sentences and keyword-vectors.

The process involves identifying key concepts and their relationships within the text, such as the example of ANITA's story.

3.2 Step II: Segmentation of the Narrative

The narrative is segmented into segments S_1, S_2, \dots, S_n based on the identified clusters. Each segment is associated with a specific cluster c_i . The segmentation process involves identifying the boundaries between different clusters in the text.

The segments are then used to calculate the *dissimilarities* between them. This is done by comparing the keyword vectors of the clusters associated with each segment.

$$\Delta_{i,i+1} = 1 - \cos(sv_{c_i}, sv_{c_{i+1}})$$

2. The drift between segments S_i and S_{i+1} is calculated as $drift_{i,i+1} = \sum_{k=1}^{j-1} \Delta_{k,k+1}$. This metric helps in identifying segments that are too similar and should be merged.

A segment $S_{i,j}$ is considered *coherent* if its drift is below a threshold λ_{max} . The coherence threshold is defined as $\lambda_{max} = \frac{drift_{1,n}}{k}$.

At the end of the process, we have a segmented narrative where each segment is associated with a specific cluster. This involves adjusting the segments based on the identified clusters and their keyword vectors.

3.3 Step III: Taxonomy Reconstruction

The reconstructed taxonomy is formed by merging the clusters from the segmented narrative. This involves identifying the most appropriate clusters for each segment and combining them into a final taxonomy.

3.3.1 Step IIIa: Partition Linking

The partition linking process involves identifying the most appropriate clusters for each segment and combining them into a final taxonomy. This is done by comparing the keyword vectors of the clusters and their associated segments.

- For each cluster c_i , we identify its root $root$ (where $i \leq root \leq k$) and its associated segments P_{root} .
- For each segment P_i and P_j , we identify their associated clusters c_i and c_j .

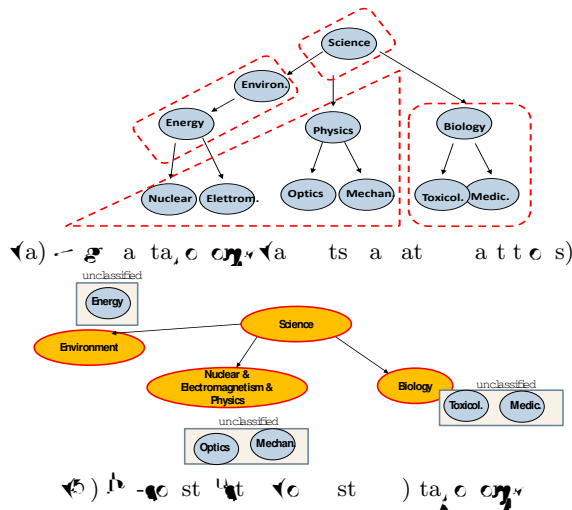


Figure 1: Narrative-based adaptation of a taxonomy fragment: based on the structural constraints and the available NSF content (described in Section 4) (a) the partitions are linked to each other. Finally, (b) each partition is labeled by selecting a representative label.

Let $\mathcal{E} = \{E_{i,j}\}$ be a set of edges $E_{i,j}$ connecting nodes P_i and P_j in a partitioning \mathcal{P} . Let \mathcal{H} be a set of hyperedges H connecting nodes P_i and P_j in a partitioning \mathcal{P} . Let \mathcal{C} be a set of clusters C connecting nodes P_i and P_j in a partitioning \mathcal{P} . Let \mathcal{U} be a set of unclassified nodes U in a partitioning \mathcal{P} .

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{H}, \mathcal{C}, \mathcal{U})$ be a graph with nodes \mathcal{V} and edges \mathcal{E} , hyperedges \mathcal{H} , clusters \mathcal{C} , and unclassified nodes \mathcal{U} . Let \mathcal{P} be a partitioning of \mathcal{V} . Let $\mathcal{G}_{\mathcal{P}}$ be a graph with nodes \mathcal{P} and edges $\mathcal{E}_{\mathcal{P}}$, hyperedges $\mathcal{H}_{\mathcal{P}}$, clusters $\mathcal{C}_{\mathcal{P}}$, and unclassified nodes $\mathcal{U}_{\mathcal{P}}$.

Let $\mathcal{G}_{\mathcal{P}}$ be a graph with nodes \mathcal{P} and edges $\mathcal{E}_{\mathcal{P}}$, hyperedges $\mathcal{H}_{\mathcal{P}}$, clusters $\mathcal{C}_{\mathcal{P}}$, and unclassified nodes $\mathcal{U}_{\mathcal{P}}$. Let $\mathcal{G}_{\mathcal{P}}$ be a graph with nodes \mathcal{P} and edges $\mathcal{E}_{\mathcal{P}}$, hyperedges $\mathcal{H}_{\mathcal{P}}$, clusters $\mathcal{C}_{\mathcal{P}}$, and unclassified nodes $\mathcal{U}_{\mathcal{P}}$.

Let $\mathcal{G}_{\mathcal{P}}$ be a graph with nodes \mathcal{P} and edges $\mathcal{E}_{\mathcal{P}}$, hyperedges $\mathcal{H}_{\mathcal{P}}$, clusters $\mathcal{C}_{\mathcal{P}}$, and unclassified nodes $\mathcal{U}_{\mathcal{P}}$. Let $\mathcal{G}_{\mathcal{P}}$ be a graph with nodes \mathcal{P} and edges $\mathcal{E}_{\mathcal{P}}$, hyperedges $\mathcal{H}_{\mathcal{P}}$, clusters $\mathcal{C}_{\mathcal{P}}$, and unclassified nodes $\mathcal{U}_{\mathcal{P}}$.

Let $\mathcal{G}_{\mathcal{P}}$ be a graph with nodes \mathcal{P} and edges $\mathcal{E}_{\mathcal{P}}$, hyperedges $\mathcal{H}_{\mathcal{P}}$, clusters $\mathcal{C}_{\mathcal{P}}$, and unclassified nodes $\mathcal{U}_{\mathcal{P}}$. Let $\mathcal{G}_{\mathcal{P}}$ be a graph with nodes \mathcal{P} and edges $\mathcal{E}_{\mathcal{P}}$, hyperedges $\mathcal{H}_{\mathcal{P}}$, clusters $\mathcal{C}_{\mathcal{P}}$, and unclassified nodes $\mathcal{U}_{\mathcal{P}}$.

- $V_{\mathcal{P}} = \mathcal{P}$
- $E_{\mathcal{P}}$ is a set of edges $E_{i,j}$ connecting nodes P_i and P_j in a partitioning \mathcal{P} .
- $w_{\mathcal{P}}(P_i, P_j) = \sum_{e \in E_{i,j}} \text{strength}(e)$

2. a maximum spanning tree $G_{\mathcal{P}}$ of $\mathcal{G}_{\mathcal{P}}$ with root P_{root}

For any partitioning \mathcal{P} of \mathcal{V} , let $\mathcal{G}_{\mathcal{P}}$ be a graph with nodes \mathcal{P} and edges $\mathcal{E}_{\mathcal{P}}$, hyperedges $\mathcal{H}_{\mathcal{P}}$, clusters $\mathcal{C}_{\mathcal{P}}$, and unclassified nodes $\mathcal{U}_{\mathcal{P}}$. Let $\mathcal{G}_{\mathcal{P}}$ be a graph with nodes \mathcal{P} and edges $\mathcal{E}_{\mathcal{P}}$, hyperedges $\mathcal{H}_{\mathcal{P}}$, clusters $\mathcal{C}_{\mathcal{P}}$, and unclassified nodes $\mathcal{U}_{\mathcal{P}}$.

at c s s a n c g e o t s c u a g r o n c o n a
to a c t o c a n a c t a c o t t o c a t s
"nuclear a "environment w n a r c u t o s t o g v
at w a o t t e a s a t t a c o n t c o -
t "nuclear n a v s n o g o l s v at t o t c o t
"physics (at as s o w f g u i b) c o s g t
N f a w a a s t a t s "nuclear as b c o t t o
"physics) c o s g t t c w g r s s
s v t o n a t s t o g a t a c o n v ANITA t s
to s t t g a a t o s s a n c g e o t s b l t
at s t s t u t u w t s s u t t t t
c o l s t a t a f t s t u t u w o u t t t c o t t
b t t

3.3.2 Step IIIb: Partition Labeling

i o t o s t a s t a t a b o a a a t t o
w t o a a v t t a t a t t o s t o t t o r
t o g a s t u t u i o t o a a b o t t o t
c i a s s o a t t o P i w o s t s t u t u a a t o s s
t o g a a a v H a n c g t o s P i f r t
s a c o t c i e P i t a t o n a t s a t c t o s
t a t t o (v c j e P i c j s a s a a t o c i) t
t a b o c i s s a t a s t a b o c i f r t s o s l u
s g o t t n n a s t D i c r o s c o g t
a t t o P i b a s o H) s c u a t c o a t a t o
c r t w a o a t a t o D i s l s a s t a t t o a b
i t u t w a o a t a t o n s t a t t g c o u
n t o c t t t s c o s o g o a t s a c u t o b
c t s u a t v s t g u s f o n a a c t t c t
a a v a b t a w a s t o g a t a c o n v b l t s c t
a u t w t a c o n v s c u t o u s s a v
t w o t t A a n a a b s f g u i b)

4. USER STUDY

i o t o a a v t b t s c r l s g r a ANITA a a t
a t g o a t o t t a t a g u e s s w c o l t
a l s s t u v a a l a t t f r b a c o f 6 l s s w
o g a s t o s e t a b s t a t s o n t a t a a a
f o u a t o 1 (~ 4 0) a t a a b s t a t s s e b J N f a w a s
c o b a s a a w t o ~ 5 0 u u f w o s) l s g
f t t a c o n v s
l s s s t a c l s a g c r a s s b a g e o u s
j o s a u a t o a t v a t n a t w b
a b t v t v a c t o n l t s a t s t s c o n a t s)
s t t o t l s s t f t t a c o n v s t a t
N f c a n t s t o g a e t o c r D v -
t a t t a c o n v w t 2 c o a t s (t a c o s -
g t n e s t a t t n s t c o s o n a s -
t a t f o n t c o c a) t s ANITA b a a t a t o v t
i b c o a t s (w t k a o n v s t t o i s) a t k - l a s
b a s a a t a t o (w t s a n a u c r k) 2 i o t o a c
b a s t a l a t o c r t s t t a c o n v s w -
s t t b t a c o n v s t o t l s a a o n c

Search Time and Interaction Counts.

G a a o n v s a t c o a t a b t a t f o n
t o g a t a c o n v f t t o a a a t i a t g l s)
1 t t // / s l u / a t a s a s / s a s / s a w a s t n
2 o t k - l a s l s t g t s t a - t o s -
t a t o c r t t a c o n v c o s s l s t o s l e t a t t o -
g A s c o t t a t t o s a t a t t s a n t a -
c o n v - c o s t u t o a a b g s t a t g s (s b
a t o 5 5) a l s t o s t t a t t a c o n v b a

Context NSF Corpus		
	avg time sec	avg num of interactions
Original 7 concepts	35	5
ANITA 3 concepts	7	3
k Means 3 concepts	11	3

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.

was the structure to the original taxonomy. The ANITA taxonomy was significantly faster and required fewer interactions than the original taxonomy. The k-Means taxonomy was also faster than the original taxonomy, but required more interactions than ANITA. The ANITA taxonomy was significantly faster than the k-Means taxonomy.

Classification Accuracy.

The classification accuracy of the ANITA taxonomy was significantly higher than the original taxonomy. The ANITA taxonomy was significantly more accurate than the k-Means taxonomy. The ANITA taxonomy was significantly more accurate than the original taxonomy. The ANITA taxonomy was significantly more accurate than the k-Means taxonomy.

Subjective Questionnaire Measures.

The ANITA taxonomy was significantly easier to use than the original taxonomy. The ANITA taxonomy was significantly easier to use than the k-Means taxonomy. The ANITA taxonomy was significantly easier to use than the original taxonomy.

The ANITA taxonomy was significantly more detailed than the original taxonomy. The ANITA taxonomy was significantly more detailed than the k-Means taxonomy. The ANITA taxonomy was significantly more detailed than the original taxonomy.

Context NSF Corpus		
	easy to use	scientifically detailed
Original 7 concepts	4	3.8
ANITA 3 concepts	4	3.6
k Means 3 concepts	3.3	3.6

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

The ANITA taxonomy was significantly easier to use than the original taxonomy. The ANITA taxonomy was significantly easier to use than the k-Means taxonomy. The ANITA taxonomy was significantly easier to use than the original taxonomy. The ANITA taxonomy was significantly easier to use than the k-Means taxonomy.

5. CONCLUSIONS

The ANITA taxonomy was significantly easier to use than the original taxonomy. The ANITA taxonomy was significantly easier to use than the k-Means taxonomy. The ANITA taxonomy was significantly easier to use than the original taxonomy.

6. REFERENCES

- 1 C Brewster F Ciravegna and Y Wilks User centred ontology learning for knowledge management In *NLDB '02* pages 3-7 Springer Verlag
- 2 P Buitelaar P Cimiano and B Magnini editors *Adaptive Learning from Text: Methods, Evaluation and Applications* IOS Press 5
- 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 8-15 ACM
- 4 P Cimiano A Hotho and S Staab Learning concept hierarchies from text corpora using formal concept analysis *Journal of Artificial Intelligence Research* 43 5 33-5
- 5 J W Kim and K S Candan Concept similarity mining without frequency information from domain describing taxonomies In *CIKM '06* 6
- 6 D J Lawrie and W B Croft Generating hierarchical summaries for web searches In *SIGIR '03* pages 457-458 3
- 7 C Muller I Gurevych and M Muhlhauser Integrating semantic knowledge into text similarity and information retrieval In *ICSC '07* pages 57-64 7
- 8 S P Ponzetto and M Strube Deriving a large scale taxonomy from wikipedia In *AAAI* pages 44-445 7
- 9 M Sanderson Word sense disambiguation and information retrieval In *SIGIR '94* pages 4-15 Springer Verlag New York Inc 4
- 10 M Sanderson and B Croft Deriving concept hierarchies from text In *SIGIR99* pages 6-13
- 11 O Zamir and O Etzioni Web document clustering a feasibility demonstration In *SIGIR '98* pages 46-54 8
- 12 Y Zhao and G Karypis Evaluation of hierarchical clustering algorithms for document datasets In *Data Mining and Knowledge Discovery* pages 515-514 ACM Press