

TaxaMiner: An Experimentation Framework for Automated Taxonomy Bootstrapping

V Kashyap^{1*}, C. Ramakrishnan², C. Thomas², D. Bassu³, T. C. Rindflesch¹ and A. Sheth²

¹Clinical Informatics R&D, Partners HealthCare System, 93 Worcester St, Wellesley, MA 02481

²LSDIS Lab, Department of CS, University of Georgia, 415 GSRC, Athens, GA 30602

³Applied Research, Telcordia Technologies, 445 South Street, Morristown, NJ 07960

vkashyap1@partners.org

Abstract

Hierarchical taxonomies and thesauri are frequently used by content management systems for indexing, search and categorization. They are also being viewed as rudimentary ontologies for the emerging Semantic Web infrastructure. However, to date, development of taxonomies and thesauri are human intensive processes, requiring huge resources in terms of cost and time. It is critical that approaches to reduce human effort and resource commitments be investigated. Towards this end, we present an experimentation framework for automated taxonomy construction from a large corpus of documents. Our approach involves: (a) generation of a document cluster hierarchy; (b) extraction of a taxonomy from this hierarchy; and (c) assignment of labels to nodes in this taxonomy. We draw upon a suite of clustering and NLP techniques and identify parameters which form the basis of an experimentation framework. We also propose metrics to measure taxonomy quality and evaluate the impacts of these parameters on these quality metrics. The MEDLINE® database is used as the document corpus and the MeSH thesaurus as the gold standard. Insights from these experiments are presented and discussed.

1. Introduction

A large portion of information content on the web is in the form of unstructured text documents. Documents are increasingly being annotated with metadata descriptions and being stored in back-end relational databases mapped to structured schemata describing the content in a domain specific manner. However, machines today understand very little of available web content. In fact, most of the annotations are in the form of tags that describe structure, formatting or presentation information. Approaches for annotation have primarily been manual [2],[3], though

there have been some attempts at exploring semi-automatic approaches for metadata annotation [4],[5]. As observed in these efforts, two resources necessary for realizing the semantic web are: (a) large scale availability of domain specific taxonomies; and (b) large scale availability of annotations or metadata descriptions created by using terms, concepts or relationships provided by these taxonomies.

Taxonomies and concept trees also facilitate users in forming effective queries through browsing and navigating information [6]. These artefacts organize documents into navigable structures and (general to specific) concepts to assist users in finding relevant information. Searches within a category of concepts typically produce more relevant results than un-scoped searches. Fully automatic approaches for taxonomy construction have lead to unsatisfactory results. On the other hand, taxonomies built and maintained by human experts, for e.g., Yahoo!, Open Directory Project and MeSH [7] taxonomies, require huge resources in terms of cost and time.

In this paper, we propose an experimental framework for taxonomy bootstrapping aimed at minimizing (not eliminating) human involvement. Measures to evaluate the quality of the taxonomies generated as a part of the framework are proposed. These measures in conjunction with the framework will enable us to identify the various algorithm parameters that need to be configured to optimize the quality of the taxonomy. We will also present insights and conclusions based on experimentation with a real world document corpus, MEDLINE® and MeSH as the Gold Standard.

The paper is organized as follows. In Section 2, we review relevant work, focusing on the attempts made by other researchers to address (parts of) this problem. The experimentation framework for taxonomy generation is described in detail in Section 3. The various components of the framework are discussed in detail in Sections 4-9. Experiments and evaluations are presented in Section 10. Section 11 discusses the conclusions and future work.

* Work on this project was initiated when this author was at the National Library of Medicine.

2. Related Work

Approaches for semi-automatic taxonomy generation typically utilize a combination of:

- Supervised machine learning approaches that require a collection of training examples.
- NLP approaches for generating taxonomic concepts and relationships between them; and
- Clustering and data mining approaches that facilitate search, categorization and visualization of data.

The concept forming system COBWEB [8] has been used to perform incremental conceptual clustering on structured instances of concepts extracted from the web [9]. An approach that uses training examples consisting of structured concept instances is presented in [10]. A classification taxonomy based on a set of structured rules is proposed in [11]. Naïve Bayesian approaches for classification have been presented in [20].

Empirical and corpus-based NLP methods to build domain specific lexicons have been proposed in [12] and used in [13]. Approaches presented in [14] apply shallow parsing, tagging and chunking, along with statistical techniques to extract terminologies or enhance existing ontologies. In [16], a thesaurus is built by performing clustering according to a similarity measure after retrieving triples from a parsed corpus. Linguistic structures such as verbs, appositives, nominal modifications and lexico-syntactic patterns have been used to identify is-a relationships in text [17],[18]. Salient words and phrases extracted from the documents are organized hierarchically using subsumption type co-occurrences in [19].

Effectively mining relevant information from a large volume of unstructured documents has received considerable attention in recent years [22]. Document clustering has been used for browsing large document collections in [23], using a “scatter/gather” methodology. Clustering of Web documents to organize search results has been proposed in [24]. Physicists have used clustering to find the spatial grouping of stars into galaxies [26]. An approach that pre-processes documents by applying background knowledge in order to improve the clustering results is proposed in [27].

Frameworks and hybrid approaches, combining the above techniques are presented in [28],[6][29]. A complementary approach that uses the structure and content of web pages on the Web to generate ontologies is presented in [30]. Hybrid approaches have also been used to automate semantic annotation, a closely related task, examples of which are the SemTag [4] and OntoMate – Annotizer systems [31], and the Semagix content management platform [5].

In this paper, we present a comprehensive framework consisting of the following novel features:

- Combination of clustering, NLP and various customized techniques for taxonomy generation.

- Identification of statistical parameters computed during the clustering process that characterizes the notion of *differentiation* in the taxonomic structure.
- Techniques for *extraction of a taxonomy* from the cluster hierarchy that results from our clustering process.
- Techniques for automatic generation and refinement of labels for nodes in the final taxonomy.
- Metrics for evaluating the quality of the generated taxonomy. The impacts of various components of the experimentation framework on the quality of the taxonomy.
- Initial validation of our approach using a real world data set, the MEDLINE® database and real world taxonomy, the MeSH thesaurus.

The taxonomy generation framework is discussed next.

3. The Taxonomy Generation Framework

The components of a framework for generating taxonomic/thesauri structures from textual documents are illustrated in **Figure 1**.

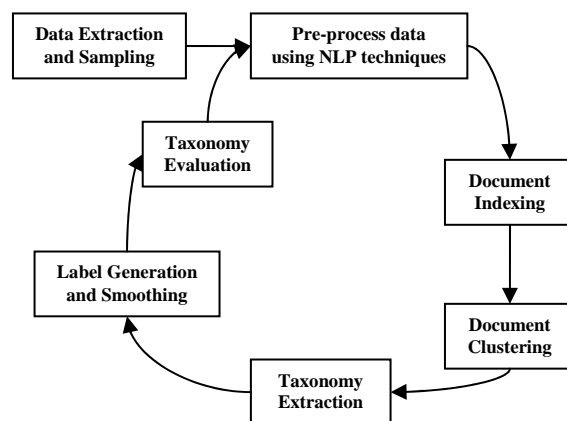


Figure 1: The Taxonomy Generation Framework

Data Extraction and Sampling A *gold standard taxonomy* (MeSH [7]) is chosen and relevant abstracts are sampled from the MEDLINE® bibliographic database. We chose the sub-tree under the concept *Neoplasms* consisting of 649 concepts as our *gold standard taxonomy* for experiments in this paper. Multiple data sets of different sizes are sampled using techniques such as uniform or density biased sampling, based on the underlying distribution of the documents *wrt* the concepts in the taxonomy.

NLP techniques for Pre-processing NLP techniques such as part of speech tagging and chunk parsing are used to extract noun phrases from the abstracts.. These phrases may be simple (1-2 words long), macro (2-3 words long) or mega (3-5 words long). Not pre-processing the documents is also an option. We compare these two options in our results.

Document Indexing The abstracts (documents) are mapped to a vector space, the dimensions of which could either be words or extracted phrases. Singular value decomposition (SVD) [32] may be applied and latent eigenvectors may be used as dimensions.

Document Clustering A bisecting K-Means strategy [33] or a Principal Direction Divisive Partitioning (PDD) [39] may be used for document clustering. Interesting variations in the clustering process might be: different cluster quality measures used to guide the clustering process, the type of distance metric (Euclidean *vs.* cosine), and term *vs.* document clustering. In our current experiments, we have adopted the bisection K means strategy with the cosine distance metric.

Taxonomy Extraction The hierarchy generated by the above process is an artefact of the clustering process and is not a taxonomy. The notion of differentiation is captured by the difference in the “cohesiveness” (defined later) between successive layers of the taxonomy. The taxonomy designer suggests a list of cohesiveness levels (which can be *tuned* to reflect a user’s perspective), based on which the taxonomy extraction algorithm extracts a subset of nodes from the clustering hierarchy and identifies the taxonomic structure.

Label assignment and smoothing A set of potential labels, based on the cluster centroids are assigned to the nodes in the taxonomy. Various techniques such as propagation of labels to parent nodes, use of thesauri such as WordNet or the UMLS Metathesaurus can be used refine these labels. Lexico-syntactic patterns [17] can be used to identify potential *subClassOf* relationships. In our current work, we have not employed artefacts such as thesauri for label assignment and smoothing.

Taxonomy Evaluation Finally, the generated taxonomy is evaluated *wrt* the gold standard taxonomy using a variety of measures, that measure content-based similarity (i.e., overlap between the labels extracted) and the structural similarity (i.e., consistency of parent-child relationships) between the two taxonomies.

The dimensions of our experimental framework are:

1. Sampling:
 - a. Uniform sampling
 - b. Density biased sampling
2. Natural Language Processing
 - a. No Tagging/Chunking
 - b. Noun Phrases: (i) Simple, (ii) Macro, (iii) Mega
 - c. Verb Phrases
3. Indexing:
 - a. Term-based dimensions: Word-based *vs.* Phrase based
 - b. SVD eigenvector-based dimensions: Word-based *vs.* Phrase-based
4. Clustering
 - a. Bisecting K Means *vs.* PDDP
 - b. Document *vs.* term based clustering
5. Distance Measures
 - a. Euclidean

- b. Cosine
6. Cluster Quality Measures:
 - a. Internal Measures:
 - (i) Pair wise distance,
 - (ii) Distance from Centroid
 - b. External Measures
7. K-Means Number of Iterations
8. Label assignment:
 - a. Threshold: (Value of Top K)
 - b. Use of Noun Phrase Matching
 - c. Use of Taxonomic Label Propagation
9. Use of Thesauri: Yes/No
10. Use of Lexico-Syntactic Patterns: Yes/No

4. Sampling the Data Set

A subset of the MEDLINE® bibliographic database satisfying the following conditions is extracted: (a) the MEDLINE® citation should be annotated by one of the 649 concepts present in the gold taxonomy, i.e. the MeSH sub-tree under the concept *Neoplasms*; (b) the concepts that annotate the citation should be identified as “preferred”; and (c) the citation should have a non-empty abstract.

MeSH, which is used as the gold standard in our experiments, while not a taxonomy in the formal sense from a knowledge representation viewpoint, is however on the most widely used organizations of concepts in the biomedical field. It has been created by domain experts and is used to index over 14 million MEDLINE® citations. These features have influenced us in our choice of the MeSH as the gold standard taxonomy and the MEDLINE® as the experimental data set.

Uniform random sampling is frequently used in practice and also frequently criticized because it will miss small clusters. Many natural phenomena are known to follow Zipf’s distribution and the inability of uniform sampling to find small clusters is of practical concern. In the context of our approach, sampling is likely to be biased in such a way as to produce a taxonomy containing concepts which appear only in a large number of MEDLINE® citations. Hence, we adopt the approach of density biased sampling as proposed in [34] where we probabilistically under-sample dense regions, i.e., concepts that appear as annotations of a large number of MEDLINE® citations; and over-sample sparse regions.

Density biased sampling relies on the a priori approximate grouping of data points in the sample. It then samples points from these groups whilst ensuring that dense regions are under-sampled and sparse regions over-sampled. The advantage we have in our experiment is that we know exactly what these groups are a priori. This enables us to greatly simplify the sampling process in our experiments. In the absence of such a priori knowledge one may use the more complex algorithms suggested in [34]. As discussed in [34], the data sets sampled have the following characteristics:

- Given a MeSH concept, documents are selected with a uniform probability. The probability function is:

$$f(\text{Concept}_i) = \frac{\alpha}{\sqrt{\text{size}(\text{Concept}_i)}}$$

- The sample is density preserving and biased by group size, where $\text{size}(\text{Concept}_i)$ is the number of citations in which Concept_i appears as a preferred annotation.
- For a given sample size M , the value of α is given by:

$$\alpha = \frac{M}{\sum_{i=1}^{\text{Nodes}} \sqrt{\text{size}(\text{Concept}_i)}}$$

Where Nodes = Number of MeSH terms in the *gold standard taxonomy*, 649 in our experiments.

5. NLP based pre-processing of the Data Set

The PhraseX system, developed at the National Library of Medicine, extracts noun phrases by referring to the syntactic structure provided by the SPECIALIST minimal commitment parser, which relies on the SPECIALIST Lexicon as well as the Xerox stochastic tagger [35]. The output of PhraseX contains simple noun phrases. The authors in [36] refer to these phrases as "core noun phrase," that is, a noun phrase with no modification to the right of the head.

The SPECIALIST parser is based on the notion of barrier words [37] which indicate boundaries between phrases. After lexical look-up and resolution of category label ambiguity by the tagger, complementizers, conjunctions, modals, prepositions, and verbs are marked as boundaries. Subsequently, boundaries are considered to open a new phrase (and close the preceding phrase). Any phrase containing a noun is considered to be a (simple) noun phrase, and in such a phrase, the right-most noun is labelled as the head; all other items (other than determiners) are labelled as modifiers. An example of the output from the SPECIALIST parser is given in (2) for the input in (1).

(1) Kupffer cells from halothane-exposed guinea pigs carry trifluoroacetylated protein adducts.

(2)

```
[mod([lexmatch(['Kupffer']),
inputmatch(['Kupffer']),tag(noun)]),
head([lexmatch([cells]),
inputmatch([cells]),tag(noun)])],
[prep([lexmatch([from]),
inputmatch([from]),tag(pre)]),
mod([lexmatch([halothane]),
inputmatch([halothane]),tag(noun)],
punc([inputmatch([-])]),
mod([lexmatch([exposed]),
inputmatch([exposed]),tag(adj)]),
head([lexmatch(['guinea pigs']),
inputmatch([guinea,pigs]),
tag(noun)])],
[verb([lexmatch([carry]),inputmatch([carry]),
tag(verb)])],
[mod([lexmatch([trifluoroacetylated]),
inputmatch([trifluoroacetylated]),
```

```
tag(adj)]),
mod([lexmatch([protein]),
inputmatch([protein]),tag(noun)]),
head([lexmatch([adducts]),
inputmatch([adducts]),tag(noun)]),
punc([inputmatch(['.'])])]]]
```

The underspecified structure produced by the SPECIALIST parser serves as the basis for the extraction of noun phrase strings. In addition to the simple noun phrase (labeled as "simp" in output), PhraseX identifies two additional structures. One of these is the complex noun phrase in which a head is followed by contiguous prepositional phrases to its right ("macro"). The first preposition in this structure can be anything, but all the rest must be "of". The second structure is not a canonical syntactic phenomenon, but may be important for information processing. Such a phrase includes all the content words that occur in a sentence either to the left or the right of a finite verb ("mega"). Examples of these strings as extracted from the syntactic structure in (2) are given in (3).

(3)

```
00000000|simp|kupffer cells
00000000|simp|halothane exposed guinea pigs
00000000|simp|trifluoroacetylated protein
adducts
00000000|macro|kupffer cells from halothane
exposed guinea pigs
00000000|mega|kupffer cells from halothane
exposed guinea pigs
00000000|mega|trifluoroacetylated protein
adducts
```

In the experiments described in this paper, we will focus on simple noun phrases.

6. Document Indexing

There are two prevalent methods related to indexing a set of documents using a vector-space representation:

- Terms are used as dimensions of the underlying vector space as in the SMART Indexing and Retrieval Engine [38].
- The Latent Semantic Indexing approach [32], where a *Singular Value Decomposition (SVD)* analysis identifies the underlying eigenvectors. These are used as dimensions of a common "latent" space in which both term and document vectors can be represented.

In the experiments described in this paper, we use the SMART Indexing and Retrieval Engine.

7. Document Clustering

In this set of experiments, we employ a bisecting K-means strategy [33] for document clustering. A hierarchical cluster tree is generated. Consider a set of document vectors $\mathbf{D} = \{d_1, \dots, d_M\}$ in the Euclidean space \mathbf{R}^N . Let the centroid of the set be denoted by:

$$m(\mathbf{D}) = \frac{1}{M} \sum_{i=1}^M d_i$$

The intra-cluster (or intra-set) cohesiveness is defined as:

$$c(\mathbf{D}) = \frac{1}{M} \sum_{i=1}^M \cos(d_i, m(\mathbf{D}))$$

Let $\{\pi_i\}_{i=1}^k$ be a partition of \mathbf{D} with the corresponding centroids $m_1 = m(\pi_1), \dots, m_k = m(\pi_k)$. The parent/child relationships are established as follows:

$$\forall i, 1 \leq i \leq k, \text{child}(\mathbf{D}) = \pi_i \text{ and } \text{parent}(\pi_i) = \mathbf{D}$$

The quality of the partition increases if the intra-cluster cohesiveness increases. Thus the quality \mathcal{Q} of the partition $\{\pi_i\}_{i=1}^k$ is given by:

$$\mathcal{Q}(\{\pi_i\}_{i=1}^k) = \frac{1}{k} \sum_{i=1}^k c(\pi_i)$$

We start with the set of all the documents as the initial cluster. Let C_1, \dots, C_i be the set of clusters at i^{th} iteration. We choose a cluster S using a selection rule and apply k-means clustering with $k=2$ to give $(i+1)$ clusters. We check to determine if there is significant improvement in the partition quality. In case there is, we run k-means on all the $(i+1)$ clusters to stabilize the clusters at this level. Changes in the clusters are noted and the above process is repeated until a significant increase in the quality measure is not seen. The algorithm is presented below.

1. Start with a single cluster \mathbf{D} at level = 1.
2. At tree level = L ,
 - a. Select a cluster $\pi_{j,L}$ from the partition $\{\pi_{i,L}\}_{i=1}^k$ which has the lowest value for $c(\pi_{j,L})$
 - b. Run k-means clustering on $\{\pi_{j,L}\}$ with $k = 2$ to obtain a new partition with $k+1$ clusters $\{\pi_{i,L+1}\}_{i=1}^{k+1}$. This includes the clusters $\{\pi_{j,L+1}, \pi_{k+1,L+1}\}$ generated from cluster $\pi_{j,L}$.
3. Check if $\mathcal{Q}(\{\pi_{i,L+1}\}_{i=1}^{k+1})$ is significantly greater than $\mathcal{Q}(\{\pi_{i,L}\}_{i=1}^k)$
4. If there are significant gains,
 - a. Copy the centroids to initialize a new partition at level $L+1$, i.e., $m_i = m(\pi_{i,L+1})$
 - b. Establish the following relationships:
 - i. $\text{child}(\pi_{j,L}) = \pi_{j,L+1}$
 - ii. $\text{child}(\pi_{j,L}) = \pi_{k+1,L+1}$
 - iii. $\text{child}(\pi_{i,L}) = \pi_{i,L+1}$ for other clusters.
 - c. Run $k+1$ means clustering on $\{\pi_{i,L+1}\}_{i=1}^{k+1}$ to stabilize the clusters at level $L+1$
 - d. Goto step 2.
5. Stop.

The clustering strategy involves design choices geared towards taxonomy generation, discussed below:

- **Document vs term clustering:** Document clustering is preferred over term clustering, as in most real data sets there are more terms than documents, giving the clustering algorithm a greater discerning power to differentiate clusters.
- **Cluster Selection:** In the bisecting K-means strategy, selecting the next cluster to split is a critical choice. We select a cluster with the least cohesiveness. Since we seek to generate a *differentiated* taxonomy, it is better strategy to choose

clusters with low cohesiveness that offer a higher likelihood of a cleaner differentiation at the lower levels.

- **Clustering Termination Condition:** The algorithm computes a partitioning quality measure and tracks changes in quality at each successive partitioning. If the quality of the partitions doesn't increase significantly or decreases, we terminate the partitioning process for a given cluster. Another pragmatic condition for terminating the partitioning process is when the size of a cluster drops below a certain threshold, say 50 documents.
- **Avoiding Local Extrema:** This is a crucial problem encountered in the bisecting K-means approach. We adopt two strategies to avoid the clustering process from getting caught in a local extrema:
 - The clustering process is initiated by generating random seed centroid and then performing K-means iterations till convergence is reached or a maximum number of iterations have been performed (Step 2b of the algorithm). We repeat this process multiple times, each time with a seed generated from a different part of the vector space and chose the best partitioning (based on the quality measure discussed) across all the K-means runs.
 - At each stage, we stabilize the clustering by initializing the centroids from the 2-means step and performing a K-means run at each stage (Step 4c of the algorithm).

It should be noted that the hierarchical cluster tree is an artefact of the clustering algorithm and is **not** the generated taxonomy. Some parameters that will be useful in extracting the final taxonomy are: (a) the intra-cluster cohesiveness $c(\pi_i)$; the centroid vector $m(\pi_i)$, and the various parent child relationships generated.

8. Taxonomy Extraction

The notion of differentiation is captured by the difference in the *cluster cohesiveness* between successive layers of the hierarchical cluster tree. The taxonomy creator or user is expected to suggest a set of cohesiveness levels which correspond to differentiation between the various layers of the taxonomy. In the course of our experimentation, it was observed that the successive values of cohesiveness down a cluster hierarchy are *monotonically increasing* in value.

Given a set of cohesiveness parameters, the taxonomy extraction algorithm extracts a subset of nodes from the cluster hierarchy and identifies the taxonomic structure (**Figure 2**). The input to this algorithm is a cluster hierarchy (CH) with the computed cohesiveness measure $c(\pi_i)$ and a set of thresholds: $\mu_1 \geq \dots \geq \mu_N$ and the output is an extracted taxonomy (T).

A set of paths belonging to a tree T is denoted by $\text{paths}(\mathbf{T}) = \{p_1, \dots, p_M\}$ and contains the paths originating from the root of the tree and ending at the leaf nodes of

the tree. The paths corresponding to the cluster hierarchy \mathbf{CH} in **Figure 2** are:

$paths(\mathbf{CH}) = \{“DSSS..”, “DHH_4 H_4..”, “DHKLH_3..”, …\}$.

Each node in \mathbf{CH} corresponds to a cluster of documents. A set of selected nodes corresponding to a cohesiveness threshold μ_j is denoted by $selectedNodes(\mu_j)$ and identifies clusters π_j s.t. $c(\pi_j)$ is closest to μ_j . The selected nodes as illustrated in **Figure 2** are:

$$selectedNodes(\mu_1) = \{S, H_4, K\}$$

$$selectedNodes(\mu_2) = \{H_3, H_1, H_2\}$$

The algorithm for taxonomy extraction is as follows:

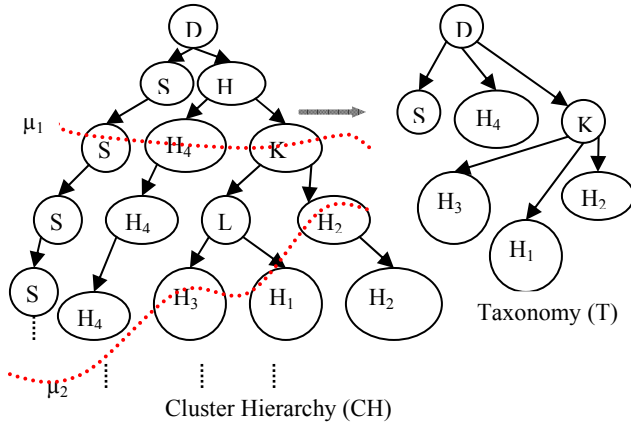


Figure 2: Taxonomy from Hierarchical Clusters

1. For each path p_i in $paths(\mathbf{CH})$ do
 - a. For $j = 1$ to N do
 - i. Find nodes A and B in p_i s.t. $c(A) \leq \mu_j \leq c(B)$
 - ii. If $(\mu_j - c(A)) \leq (c(B) - \mu_j)$
Insert A in $selectedNodes(\mu_j)$
Else, Insert B in $selectedNodes(\mu_j)$
2. Collapse \mathbf{CH} : For $i = 1$ to N do
 - a. For each Node A in $selectedNodes(\mu_i)$ do
 - i. If $i > 1$,
Find ancestor(A) in $selectedNodes(\mu_{i-1})$
 - ii. If $i=1$, ancestor(A) = root(\mathbf{CH})
 - iii. Delete all nodes from on the path from A to ancestor(A)
 - iv. Establish ancestor(A) as the parent of A in the extracted taxonomy \mathbf{T}
3. End Extract Taxonomy

The basis of the taxonomy extraction process discussed above is the hypothesis: nodes at lower levels in the taxonomy capture categories/concepts that are narrower than categories/concepts captured in nodes at higher levels. The rationale behind this is that document clusters at higher levels of the hierarchy are a **superset** of document clusters below them in the same sub-tree. From the IR literature, a concept C_1 is assumed to be narrower than a concept C_2 , if a query comprising of C_1 returns a subset of the documents returned by a query comprising of C_2 . Assuming that each document cluster represents an implicit concept provides a rationale for the hypothesis mentioned above.

The taxonomy creator or user is expected to suggest a set of cohesiveness levels which correspond to differentiation between the various layers of the taxonomy. In general, this will be an iterative process involving display of the raw clustering and labeling results to the user. This will give him/her a better idea of how to set up the cohesiveness levels to produce the desired taxonomy. The levels of cohesiveness are thus parameters which can be varied to better “tune” a taxonomy that corresponds to the taxonomy creator’s perspective of the information domain. The process of interaction between the taxonomy creator and the TaxaMiner system and “tuning” of the parameters are beyond the scope of this paper.

9. Taxonomy Node Labeling

After the taxonomy nodes have been extracted from the cluster hierarchy tree, the following steps are performed:

- For each node in the extracted taxonomy, a set of potential labels are assigned
- These sets of labels are then pruned and refined using noun phrases extracted from the dataset and taxonomic label propagation.

Labels are assigned to each node, based on the node’s centroid vector. In the case of SMART [38], we simply choose the top K weighted values of the centroid vector and determine the terms which contribute to the top K terms. In the case of the LSI [32], terms and documents are represented in the same “latent” space. This will enable us to compute the (Euclidean or cosine) distance between the centroid vector and the term vectors. We anticipate that experiments using LSI will greatly improve the quality (and reduce the number) of candidate labels generated for each node in the final taxonomy,

Given a cluster node π_i , we define the $labels(\pi_i)$ to contain the labels assigned to the cluster in the taxonomy tree.

$$childLabels(\pi) = \bigcup_{A \in children(\pi)} labels(A)$$

$$parentLabels(\pi) = labels(parent(\pi))$$

$$taxLabels(\mathbf{T}) = \bigcup_{A \in \mathbf{T}} labels(A)$$

The same labels can appear in multiple nodes of the extracted taxonomy. One approach of refining the labels of the taxonomy involves using noun phrases, extracted from the data set. Individual words can them be combined into potential phrases in the lexicon, reducing the number of labels. Another approach, referred to as *taxonomic propagation*, involves propagation of labels across different levels of the taxonomy.. Some heuristics are for label propagation are:

- **Propagate to Child:** If a label appears both in the parent and one or few of it’s children, the label will be propagated to the child and removed from the parent. A parent node in a taxonomy is a

generalization of its children. Hence the parent should not have a label that only one or few of its children have.

- **Propagate to Parent:** If a label has been assigned to all the children of a node, the label will be propagated to the parent and removed from all the children nodes at which it appears. If every child of a node in a taxonomy has a label that the node itself has, having that label in the parent node suffices to convey the fact that children of this node also talk about the concept that the label represents.

In our experiments described in this paper we use noun phrase replacement and the label propagation techniques described above to prune the labels. The algorithm for label propagation is as follows:

```

1. Start with the Root ( $\mathbf{T}$ )
2. For each cluster node  $\pi_i$  at level L do
  a. For cluster node  $\pi_j \in \text{children}(\pi_i)$  do
    i. If  $\Delta = \text{labels}(\pi_i) \cap \text{labels}(\pi_j) \neq \phi$ 
    ii.  $\text{labels}(\pi_i) = \text{labels}(\pi_i) - \Delta$ 
3. End Propagate to Children
4. Start with cluster nodes in leaves ( $\mathbf{T}$ )
5. For each cluster node  $\pi_i$  at level L do
  a. If  $\Delta = \text{labels}(\pi_i) \cap \text{childLabels}(\pi_i) \neq \phi$ 
  b.  $\text{labels}(\pi_i) = \text{labels}(\pi_i) + \Delta$ 
  c. For  $\pi_j \in \text{children}(\pi_i)$  do
    i.  $\text{labels}(\pi_j) = \text{labels}(\pi_j) - \Delta$ 
6. End Propagate to Parent
7. End Label Propagation

```

10. Experiments and Evaluation

We now discuss metrics used to evaluate the quality of the taxonomy generated by our algorithms. Experiments investigating the impact of the various factors on the quality of the taxonomy generated are also presented.

10.1 Taxonomy Quality Metrics

We propose to separate the content and structural aspects of a taxonomy, in an attempt to discover trade-offs and dependencies that might exist between the two. This will enable us to determine which of the steps in our process contributes to an increase in the quality of the taxonomy generated. Towards this end, we propose simple and pragmatic metrics to evaluate the generated taxonomy *wrt* a gold standard taxonomy. There are two classes of metrics: those that measure the quality of the content (labels); and those that measure the structure of the generated taxonomy. A discussion of the various quality metrics is presented below:

- **Content Quality Metric (CQM):** This measures the overlap in the labels present in the generated Taxonomy, T_{gen} and the gold standard taxonomy T_{gold} . There are two variants of this metric:
 - **CQM-P:** This measures the precision, i.e., the percentage of labels in T_{gen} that appear in T_{gold}

$$CQM-P = \frac{|\text{taxLabels}(T_{gen}) \cap \text{taxLabels}(T_{gold})|}{|\text{taxLabels}(T_{gen})|}$$

- **CQM-R:** This measures the recall, i.e., the percentage of labels in T_{gold} that appear in T_{gen}

$$CQM-R = \frac{|\text{taxLabels}(T_{gen}) \cap \text{taxLabels}(T_{gold})|}{|\text{taxLabels}(T_{gold})|}$$

- **Structural Quality Metric (SQM):** This measures the structural validity of the labels, i.e., when two labels appear in a parent child relationship in T_{gold} , they should appear in a consistent relationship (parent-child or ancestor-descendant) in T_{gen} or vice versa. Based on the above discussion, let:

$$pcLinks(\mathbf{T}) = \{ \langle a, b \rangle \mid a \text{ is parent of } b \text{ in } \mathbf{T} \}$$

$$adLinks(\mathbf{T}) = \{ \langle a, b \rangle \mid a \text{ is ancestor of } b \text{ in } \mathbf{T} \}$$

$$adLinks(\mathbf{T}) \supseteq pcLinks(\mathbf{T})$$

As above, there are two variants of the SQM metric:

- **SQM-P:** This measures the precision, i.e., the percentage of parent-child relationships in T_{gen} that appear consistently in T_{gold} .

$$SQM-P = \frac{|\text{pcLinks}(T_{gen}) \cap \text{adLinks}(T_{gold})|}{|\text{pcLinks}(T_{gen})|}$$

- **SQM-R:** This measures the recall, i.e., the percentage of parent-child relationships in T_{gold} that appear consistently in T_{gen} .

$$SQM-R = \frac{|\text{pcLinks}(T_{gold}) \cap \text{adLinks}(T_{gen})|}{|\text{pcLinks}(T_{gold})|}$$

The above measures are scaled appropriately. It is quite likely, especially for smaller data sets, that the number of concepts generated in T_{gen} is likely to be less than the number of concepts in T_{gold} . This will have an impact on the recall related quality measures and the respective denominators in CQM-R and SQM-R will be scaled appropriately to reflect this.

10.2 Experimental Results

We present an initial set of experiments evaluating the impact of the following on the quality of the taxonomies generated.

- The effect of varying the size of the data sets.
- The effect of varying the number of labels extracted.
- The effect of pre-processing the document set using limited NLP techniques (Noun Phrase Extraction).

The content (CQM-R, CQM-P) and structural quality (SQM-R, SQM-P) measures defined in the previous section will be used with the following caveats:

- In the current set of experiments, we have generated only 50 levels of the clustering hierarchy. This will have an impact on the recall related quality measures (CQM-R, SQM-R) and we scale those measures

appropriately to reflect this. In the future, we plan to generate more levels in further experiments which will lead to better results.

- A subject matter expert is required to set the threshold levels for taxonomy extraction, i.e., the μ values discussed in Section 8. In our current experiments we have assigned μ values automatically based on the minimum and maximum values of cohesiveness. We believe that the involvement of an expert would significantly improve the quality measures.
- It may be noted that our techniques will not be able to generate labels in the taxonomy that do not appear in the text of the MEDLINE® abstracts.

The gold standard taxonomy and an example of a learned taxonomy are illustrated in the **Appendices A.1 and A.2** at the end of this paper. We begin with a set of experiments involving multiple data sets that have been pre-processed using NLP techniques. Taxonomy content quality measures are computed for each of the taxonomies for different values of K (the size of the label sets extracted at each cluster node in the taxonomy) and different data set sizes.

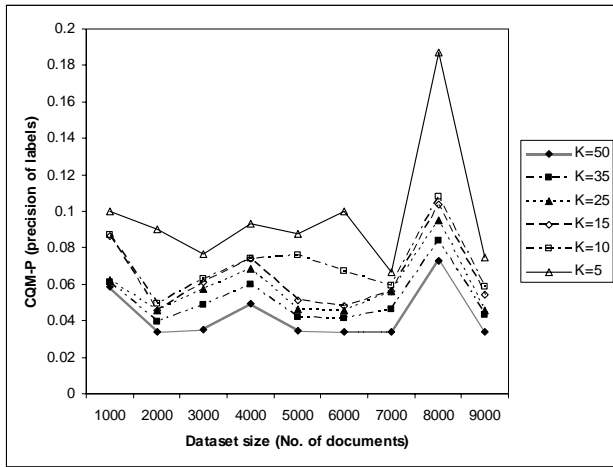


Figure 3: Content Quality Metric (Precision)

Figure 3 above, illustrates the impact of using datasets of different sizes on CQM-P. Some interesting trends that may be observed are:

- Increasing the data set size does not necessarily increase the values of CQM-P. In fact, we notice a trend that suggests that CQM-P peaks for a certain value of the data set size and then deteriorates for larger data sets. We observe this behaviour across all values of K (the number of labels extracted).
- Extracting a lesser number of labels for each cluster node (the value of K) gives better results for CQM-P

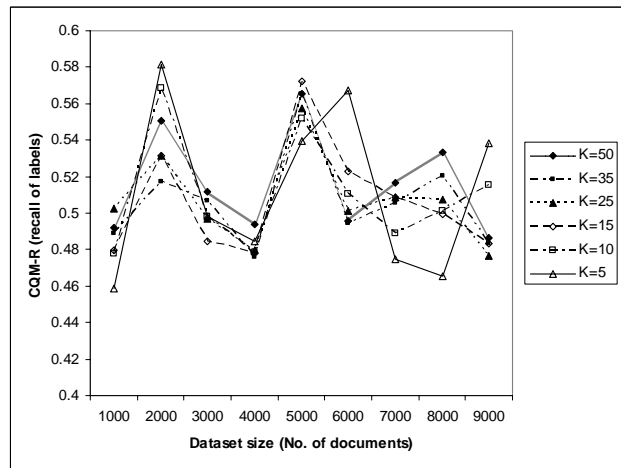


Figure 4: Content Quality Metric (Recall)

Figure 4 above, illustrates the impact of using datasets of different sizes on CQM-R. An interesting observation that may be from the above is that the value of CQM-R stays in a narrow band (0.5 – 0.6) and appears to be relatively independent of the size of the dataset and the number of labels extracted.

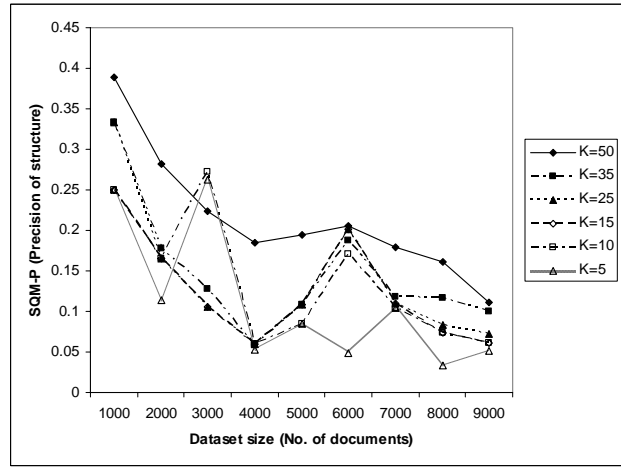


Figure 5: Structural Quality Metric (Precision)

Figure 5 above illustrates the impact of using datasets of different sizes on SQM-P. Some interesting trends that might be observed are:

- In contrast to the content quality metric CQM-P, increasing the value of K (the number of extracted labels), gives better values of SQM-P.
- More interestingly, the values of SQM-P show a downward trend *wrt* dataset size, i.e., increasing the size of the dataset, results in a decrease in SQM-P.

Figure 6 below, illustrates the impact of using datasets of different sizes on SQM-R. Some interesting trends that might be observed are:

- In a manner similar to SQM-P, SQM-R shows a downward trend on increasing the dataset size, i.e., as the data set size increases, SQM-R tends to decrease.

- Also, similar to SQM-P again, extracting a larger number of labels (value of K), tends to increase SQM-R, notwithstanding a few “crossings”

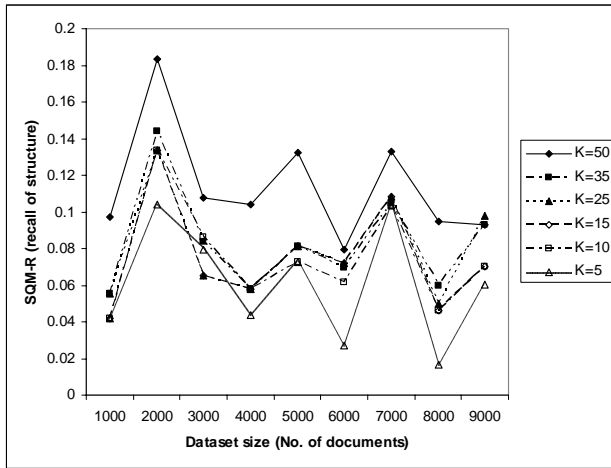


Figure 6: Structural Quality Metric (Recall)

In the next set of experiments, we investigate the impact of pre-processing the document set using limited NLP techniques, such as noun phrase extraction.

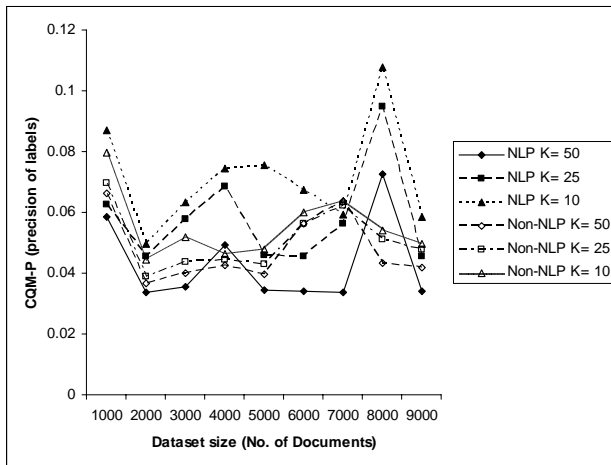


Figure 7: NLP vs. Non-NLP for CQM-P

Figure 7 above compares the impact of using NLP techniques vis-à-vis not using them on the precision-based content quality measure. We observe that for each value of K (the number of labels extracted), the values for CQM-P are consistently better for the NLP case in comparison to the non-NLP case.

Figure 8 below compares the impact of using NLP techniques vis-à-vis not using them on the recall-based content quality measure. We observe that for each value of K (the number of labels extracted), the values for CQM-R are consistently better for the non-NLP case in comparison to the NLP case. This is an interesting trend in complete contrast to the trend observed in the case of CQM-P.

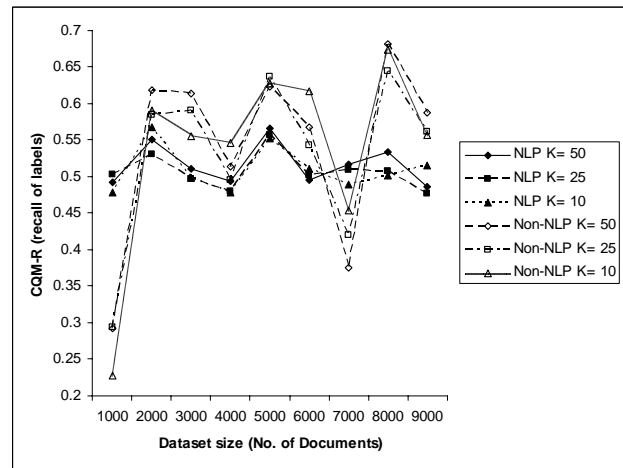


Figure 8: NLP vs. non-NLP for CQM-R

A comparison of SQM-P and SQM-R for the NLP and non-NLP cases does not reveal any conclusive trend or insight. As illustrated in **Figure 9** and **10**, the curves representing the NLP and non-NLP cases cross each other. A deeper investigation is required to conclude whether NLP is useful in this context or not.

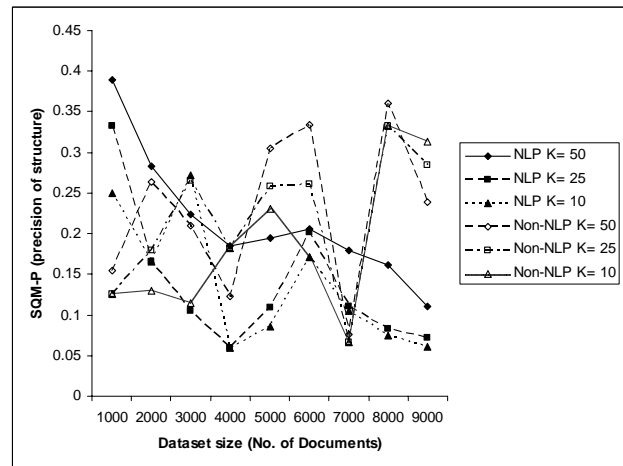


Figure 9: NLP vs. non-NLP for SQM-P

10.3 Discussions and Insights

The experiments discussed above are a component of extensive ongoing work in evaluating a suite of taxonomy generation techniques. They have provided us with some interesting insights, which indicate further areas of research and investigation.

We have observed differing trends and behaviours across the content (CQM-R, CQM-P) and structural (SQM-R, SQM-P) metrics when we varied the data set sizes and the number of labels extracted. Assuming that in

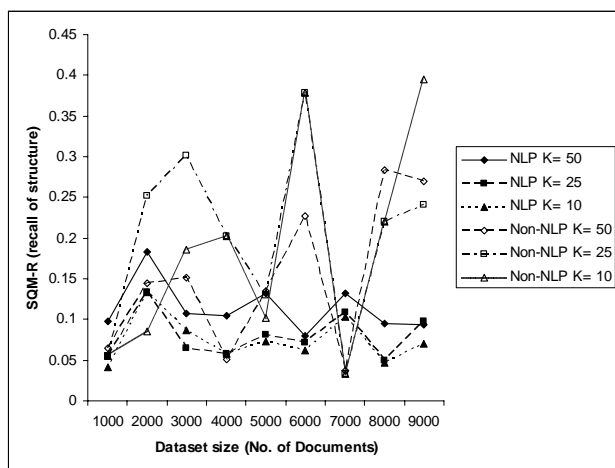


Figure 10: NLP vs. non-NLP for SQM-R

general, the goal of a taxonomy designer will be to optimize both the content and structural quality of a taxonomy, the trends observed in the previous section point to an optimal size of the data set and an optimal number of labels to be extracted at each node. We reiterated some of the observations made in the previous section which suggest this:

- There is a trend for CQM-P to peak for a particular size of the data set and this trend is visible across all values of K (number of labels extracted).
- Extracting a lesser number of labels (value of K) tends to increase the values of CQM-P.
- The values of SQM-P on the other hand (in contrast to CQM-P) tend to improve for a higher number of labels extracted (value of K).
- Also, in contrast to CQM-P, the values of SQM-P decrease with an increase in the size of the dataset.
- The values of SQM-R, in a manner similar to SQM-P also tend to increase with an increase when the dataset size decreases.

The low values of the various quality measures (except CQM-R) suggest that a deeper investigation is needed to obtain better results. We expect meaningful user input in the form of judiciously chosen cohesiveness thresholds (μ values) to alleviate the problem by identifying the correct level of differentiation and alignment. In our current experiments, we have implemented heuristics that identify these μ values automatically. These heuristics need to be further enhanced, in conjunction with reference taxonomies of the domain to automatically recommend a range of μ values to the user for taxonomy extraction.

For certain applications like information filtering and metadata annotation, the content quality measure might have more importance as opposed to the structural quality measure. We need to investigate a composite measure that gives different weights to the content and structural components and configure the taxonomy generation algorithm appropriately.

Finally, pre-processing documents using NLP techniques such as noun phrase extraction, gives better values of CQM-P but poorer values of CQM-R. One plausible explanation is the fact that NLP pre-processing removes from consideration certain terms and labels that might actually correspond to labels in the Gold Taxonomy. Obviously, the negative impact of NLP pre-processing on CQM-R will have to be compensated by re-configuring the algorithm to change the size of the data set or to increase the number of labels extracted at each node. A deeper investigation into this phenomenon would enable us to develop hybrid clustering and NLP approaches to optimize a combination of content and structural taxonomy quality.

11. Conclusions and Future Work

The main contributions of this paper are:

- Intuitive taxonomy quality metrics that differentiate between the content and structural aspects of a taxonomy and measures them in an explicit manner
- A comprehensive experimental framework that combines statistical clustering and NLP techniques for taxonomy generation. We have identified various parameters that provide means to optimize the quality of a taxonomy. Some novel features of our techniques are:
 - 1) Exploitation of the statistics generated during the clustering process to extract a more meaningful taxonomy.
 - 2) Identification of statistical parameters that characterize the notion of “differentiation” in the taxonomic structure.
 - 3) Techniques for automatic generation and refinement of labels for creating the final taxonomy.
- A set of experiments based on a real world data set, viz., the MEDLINE® collection, and a real world taxonomy, viz., the MeSH taxonomy, which has given us good initial results and interesting insights into this difficult problem.

A major insight suggests that a generated taxonomy consists of intrinsic information content and that analyzing larger data sets and extracting more labels do not necessarily guarantee good results. Human involvement, though minimal is crucial to the process of creating good quality taxonomies. The notion of taxonomy quality is multi-dimensional, involving content-based, structural aspects and may be viewed from correctness (reflected in precision) and completeness (reflected in recall) perspectives. Thus, the quality of a taxonomy needs to combine all these aspects in an application specific manner. An optimal strategy based on such a metric involves a joint optimization of various parameters.

Whereas, we have addressed the issues related to quality metrics and the experimentation framework in a

significant manner, we are not yet willing to declare victory *wrt* to the results obtained in our experiments. Initial results, which are based on completely automatic approaches, not involving domain experts are promising, even though, viewed in isolation, they may appear to be on the low side. However, it may be noted that there are intrinsic constraints associated with this problem, one of which is the fact that labels that do not appear in the text of the documents will not be generated by our system. However, similar labels are generated by our system, though not counted as a "match" in our metrics, resulting in under-estimation of the various quality metrics.

Given the above discussion, we have made a promising start and are investigating techniques to further improve the quality of the taxonomies generated. This work is an ongoing collaboration between researchers at the National Library of Medicine, LSDIS Lab at the University of Georgia and Applied Research at Telcordia Technologies. Some issues that we are investigating are:

- Algorithmic techniques for improving the structural quality of the generated taxonomies.
- Understand and leverage the human expert, especially in the context of identifying the levels of differentiation in the taxonomy that corresponds to his/her perspective of the application or domain. Combined quality metrics that better reflect the needs of the user.
- Investigation of the notion of an optimal set of parameters for generating a taxonomy. For example, processing a bigger data set can be avoided if we know that the resulting improvement in the taxonomy quality will be negligible.
- Investigation of NLP and other techniques [17] to further refine the taxonomies generated into richer ontologies.

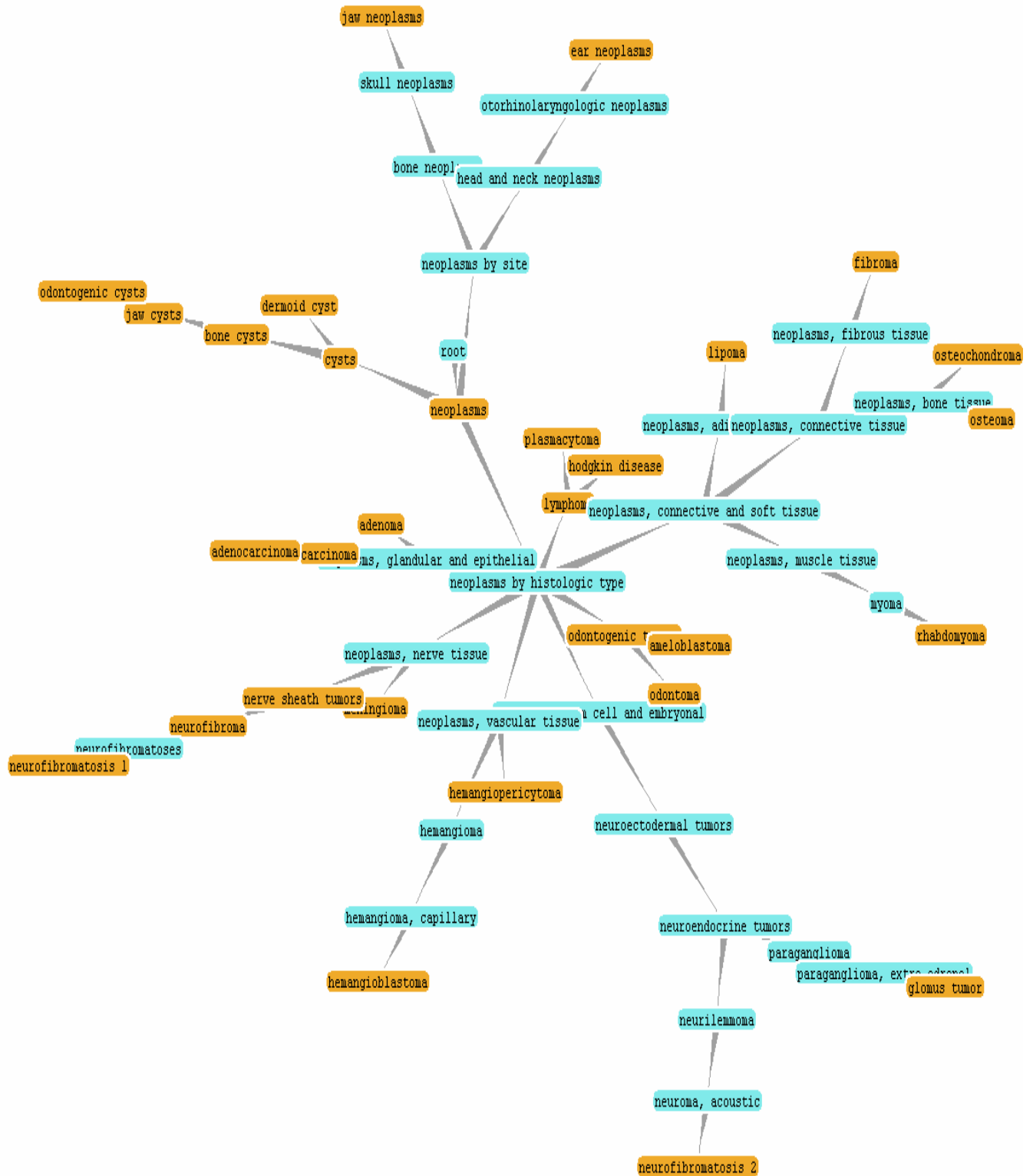
We believe that pragmatic issues as enumerated above are crucial for automating the generation of taxonomies in a scalable and feasible manner and that we have taken a very important first step in this direction.

12. References

- [1] T. Berners Lee, J. Hendler and O. Lassila. The Semantic Web. *Scientific American*, May 2001.
- [2] J. Kahan, M-R. Koivunen, E. Prud'Hommeaux and R. Swick. Annotea: An open RDF Infrastructure for shared annotations. *Proceedings of the 10th International WWW Conference (WWW 2002)*, Hong Kong, May 2001
- [3] S. Handschuh, S. Staab and R. Volz. On Deep Annotation. *Proceedings of the 12th International WWW Conference (WWW 2003)*, Budapest, Hungary. May 2003.
- [4] S. Dill et. al. SemTag and SemSeeker: Bootstrapping the Semantic Web via automated semantic annotation. *Proceedings of the 12th International WWW Conference (WWW 2003)*, Budapest, Hungary, May 2003.
- [5] B. Hammond, A. Sheth, and K. Kochut. Semantic Enhancement Engine: A Modular Document Enhancement Platform for Semantic Applications over Heterogeneous Content. In *Real World Semantic Web Applications*, V. Kashyap and L. Shklar, Eds., IOS Press.
- [6] C Y Chung, R. Lieu, J. Liu, A. Luk, J. Mao and P. Raghavan. Thematic Mapping – From Unstructured Documents to Taxonomies. *Proceedings of the 11th International Conference on Information and Knowledge Management (CIKM 2002)*, McLean, VA, November 2002.
- [7] MeSH. Medical Subject Headings. Bethesda (MD): National Library of Medicine, 2003. <http://www.nlm.nih.gov/mesh/meshhome.html>
- [8] D. H. Fisher. Knowledge Acquisition via incremental conceptual clustering. *Machine Learning 2:139-172*, 1987
- [9] P. Clerkin, P. Cunningham and C. Hayes. Ontology Discovery for the Semantic Web using Hierarchical Clustering. *Proceedings of the Semantic Web Mining Workshop co-located with ECML/PKDD 2001*, Freiburg, Germany, September 2001
- [10] W. W. Cohen and H. Hirsh. Learning the CLASSIC Description Logic: Theoretical and Experimental Results. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Fourth International Conference*, 1994.
- [11] H. Suryanto and P. Compton: Learning Classification taxonomies from a classification knowledge based system. In *Proceedings of Workshop on Ontology Learning at ECAI-2000*, 2000.
- [12] E. Riloff and J. Shepherd. A corpus-based approach for building semantic lexicons. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP-97)*, Providence, RI, 1997
- [13] S. Dill et. al. SemTag and SemSeeker: Bootstrapping the Semantic Web via automated semantic annotation. *Proceedings of the 12th International WWW Conference (WWW 2003)*, Budapest, Hungary, May 2003.
- [14] B. Daille. Study and implementation of combined techniques for automatic extraction of terminology. In P. Resnick and J. Klavans, editors, *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, MIT Press, Cambridge, MA, 1996

- [15] M. Missikoff, P. Velardi and P. Fabriani. Text Mining Techniques to automatically enrich a Domain Ontology. *Applied Intelligence* 18, 323-340, 2003.
- [16] D. Lin. Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL-98*, 1998.
- [17] Hearst, M.: *Automatic acquisition of hyponyms from large text corpora*. In Proceedings of the 14th International Conference on Computational Linguistics. Nantes, France, 1992.
- [18] M. Fiszman, T. C. Rindfleisch and H. Kilicoglu. Integrating a Hypernymic Preposition Interpreter into a Semantic Processor for Biomedical Texts. In *Proceedings of the AMIA Annual Symposium on Medical Informatics*, 2003.
- [19] M. Sanderson and B. Croft. Deriving Concept Hierarchies from Text. *International Conference on Research and Development in Information Retrieval (SIGIR 1999)*, 1999.
- [20] S. Chakrabarti, B. Dom, R. Agrawal and P. Raghavan. Using taxonomy, discriminants, and signatures to navigate in text databases. In *VLDB*, Athens, Greece, September 1997.
- [21] A. McCallum and K. Nigam. A comparison of event models for naïve Bayes text classification. In *AAAI/ICML-98 Workshop on Learning for Text Categorization*, pages 41-48. AAAI Press, 1998.
- [22] S. Chakrabarti. Data Mining for Hypertext: A Tutorial Survey. *ACM SIGKDD Explorations*, 1(2):1-11, 2000.
- [23] D. R. Cutting, D. R. Karger, J. O. Pedersen and J. W. Tukey. Scatter/Gather: A cluster-based approach to browsing large document collections. In *Annual International Conference on Research and Development on Information Retrieval*, Denmark, 1992.
- [24] O. Zamir and O. Etzioni. Web Document Clustering: A Feasibility Demonstration. In *Proceedings of ACM SIGIR Conference*, 1998.
- [25] C. Buckley, M. Mitra, J. Walz and C. Cardie. Using clustering and superconcepts within SMART: TREC 6. In *Sixth Text Retrieval Conference (TREC-6)*, Gaithersburg, MD, November 1997.
- [26] J. Kepner, X. Fan, N. Buhcall, J. Gunn, R. Lupton and G. Xu. An Automated Cluster Finder: The Adaptive Matched Filter. *The Astrophysics Journal*, 517, 1999.
- [27] A. Hotho, S. Staab and A. Maedche. Ontology-based Text Clustering. In *Proceedings of the IJCAI 2001 Workshop on Text Learning: Beyond Supervision*, Seattle, USA, 2001.
- [28] A. Maedche and S. Staab. Ontology learning for the Semantic Web. *IEEE Intelligent Systems*, 16, 2001.
- [29] M. Reinberger, P. Spyns, W. Daelemans and R. Meersman. Mining for Lexons: Applying unsupervised learning methods to create ontology bases. In *Proceedings of the International Conference on Ontologies, Databases and Applications of Semantics (ODBASE)*, November 2003.
- [30] H. Davulcu, S. Vadrevu and S. Nagarajan. OntoMiner: Bootstrapping and Populating Ontologies from Domain Specific Websites. *Proceedings of the First International Workshop on Semantic Web and Databases (SWDB 2003)*, Berlin, September 2003.
- [31] S. Handschuh, S. Staab and R. Volz. On Deep Annotation. *Proceedings of the 12th International WWW Conference (WWW 2003)*, Budapest, Hungary. May 2003.
- [32] Deerwester, S., Dumais, S. T., Landauer, T. K., Furnas, G. W. and Harshman, R. A., "Indexing by latent semantic analysis. *Journal of the Society for Information Science*, 41(6), 391-407, 1990.
- [33] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In *KDD Workshop on Text Mining*, 2000.
- [34] C. R. Palmer and C. Faloutsos. Density Biased Sampling: An Improved Method for Data Mining and Clustering. In Proceedings of ACM SIGMOD International Conference on Management of Data, May 2000
- [35] D. R. Cutting, J. Kupiec, J. O. Pedersen and P. Sibun. A practical part-of-speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, 1992.
- [36] R. Weischedel, M. Meteor, R. Schwartz, L. Ramshaw and J. Palmucci. Coping with ambiguity and unknown words through probabilistic models. *Computational Linguistics*, 19(2), 1993.
- [37] K. W. F. Tersmette, A. F. Scott, G. W. Moore and R. E. Miller. Barrier word method for detecting molecular biology multiple word terms. *Proceedings of the 12th Annual Symposium on Computer Applications in Medical Care*, 1988
- [38] G. Salton. The SMART Retrieval System – Experiments in Automatic Document Retrieval, Prentice Hall, 1971.
- [39] Boley, D.L. Principal Direction Divisive Partitioning, *Data Mining and Knowledge Discovery, Volume 2(4)*, 1998.

Appendix A.1 The Gold Taxonomy



Appendix A.2 The Generated Taxonomy

