Realization of the semantic web on a large scale depends on semantic annotation of documents with a domain ontology that describes domain entities. Creation of ontologies is human dependent and maintenance of ontologies is very expensive. We propose a framework for automating the construction of lightweight ontologies for semantic annotations. Lightweight ontologies are typically defined as more hierarchical or classificatory in nature. More like taxonomies, but with greater connectedness, lightweight ontologies are often designed to represent subsumption or other relationships between concepts. They have not too many or not too complicated predicates (relationships). As relationships are added and the complexities of the world get further captured, ontologies migrate from the lightweight to the “heavyweight” end of the spectrum [3].

Lightweight ontologies are enough for annotating general domains, which is a basic foundation for semantic interpretation and disambiguation of domain entities and their intended context. We are developing a framework that consists of a set of modules that contribute in automating the construction of lightweight ontologies. We base most of our knowledge on the structure of the Wikipedia, which represents the hierarchical links between categories and links between pages, in addition to the context.

The main modules we focused on for this framework are: extracting domain concepts and terms, measuring relatedness between domain terms, defining boundaries of subdomains using concept clustering, and extracting relations, and defining named entities within each subdomain. We are usingDbpedia for defining named entities and linking named entities to Wikipedia definition.

**Methodology**

**Extraction of Domain Concepts and generation of domain boundaries**

Selection of seed concepts is done by a domain expert or by the help of an upper level topic map. Then we build the domain space by expanding the seed concepts by adding related nodes from the Wikipedia semantic network.

**Relatedness Measure**

Relatedness between extracted concepts is computed using our proposed algorithm which relies on Wikipedia link structure. We treat the Wikipedia as a semantic network. Our algorithm is based on the hypothesis that:

- The relatedness between two concepts is proportional to the number of shared nodes.
- The relatedness between two concepts is proportional to the number of shared parent nodes.
- The importance of the link between 2 articles depend on the section of the article where it is present.

**Clustering**

We build a relatedness matrix, which we use for creating the subdomain boundaries via clustering. We used Hierarchical clustering [6] because we don’t know the number of clusters (subdomains) in advance.

We developed a procedure to define the cutoff level in the produced dendrogram based on:

- Computing and plotting the inconsistency to choose the range of cut-off that will represent the clusters.
- We choose candidate heights with high inconsistency coefficient.
- Dunn’s Coefficient was computed and plotted for clusters produced by each candidate height, we choose the cutoff with highest Dunn’s index as the optimal static cut-off.
- Clusters with high intra-cluster distance and few number of concepts are considered outliers. Clusters with low intra-cluster (compared to other clusters) and high number of concepts - exceeding certain threshold - are re-clustered using K-medoid.

**Collection of domain relatedness**

An ontology is a formal, explicit specification of a shared conceptualization. We define an ontology as “closed world assumption” which means that the domain is a complete set and does not include any other classes, concepts, or properties. The core of a domain is the core set of concepts and relations that are considered essential to the domain. A domain ontology is a collection of domain relatedness in which concepts, relations and categories are defined.

**Related Annotation Models**

Annotation models are schemas for defining the annotation elements that will be used to annotate documents or concepts in domain. The primary goal of all developed annotation models is to provide a standard description mechanism for sharing Annotations between systems. Most known efforts are:

- Dublin Core: is a small set of vocabulary terms that can be used to describe web resources [3].
- Annota: is an RDF standard sponsored by the W3C to enhance document-based collaboration via shared document metadata based on tags, bookmarks, and other annotations [4].
- Schema.org: is an annotation schema widely supported by most search engines and users. It includes a structured definition for most named entities [5].

Why New Schema when there exists a lot?

Existence schemas lack granularity and uniqueness. The schema we propose inherit the definition of named entities from Schema.org and the document level annotations from Dublin core, extends that to add concept level annotations that define the knowledge domain and help in resolving context level ambiguity.

**Preliminary Results**

- One of the domains we developed our experiments on is the domain of “Information retrieval science”.
- We generated the seed concepts by exploring the main topics of this domain, we started with 7 core concepts .
- We generated the extended concept space which contained 516 concepts.
- Then we generated the weighted relatedness matrix that represents the concept-concept relation according to their structure in the Wikipedia.
- After clustering the domain concepts based on the generated relatedness matrix, we used our procedure for determining the best cutoff level (number of clusters). We found that the best Dunn’s coefficient was at 23 clusters with 3 clusters exceeding 37, which will be further clustered using K-medoid.

**Challenges**

We need to build an evaluation technique to measure the validation of:

- Wikipedia based relatedness measure.
- The Extracted domain concepts and the clusters extracted
- The Entity recognition and relation extraction.

In addition to the evaluation of the quality of the generated ontologies, the most effective method to evaluate ontologies is to compare it with ground truth which depends mainly on the existence of developed ontologies in the required.

**References**


