Using Semantic Technologies to Construct Research Topicology

Principal Investigators:
Gandhi, Sunil (sunilga1@umbc.edu)
Peshave Akshay (peshave1@umbc.edu)
Rao, Raghavendra (rrao1@umbc.edu)

Motivation

Conference publications typically deal with research in a focussed topic or knowledge area of any science. These publications, and in turn their authors, can be represented as clusters or well connected graphs based on their topic commonalities. Such logical representations find numerous applications in areas of information mining, pattern analysis, data visualisation etc.

Project Description

We propose the creation of triples representing these associations and associated inference logic using standard semantic web technologies and relevant ontologies. This will also enable a generalized partitioning of graphs based on standard scientific taxonomies such as the Computing Classification System (ACM-CCS), Mathematics Subject Classification (AMS-MSC) etc.
As a proof of concept we intend to implement the logic and demonstrate the process on publications on the American Computer Society (ACM) forum using minimum manual intervention.

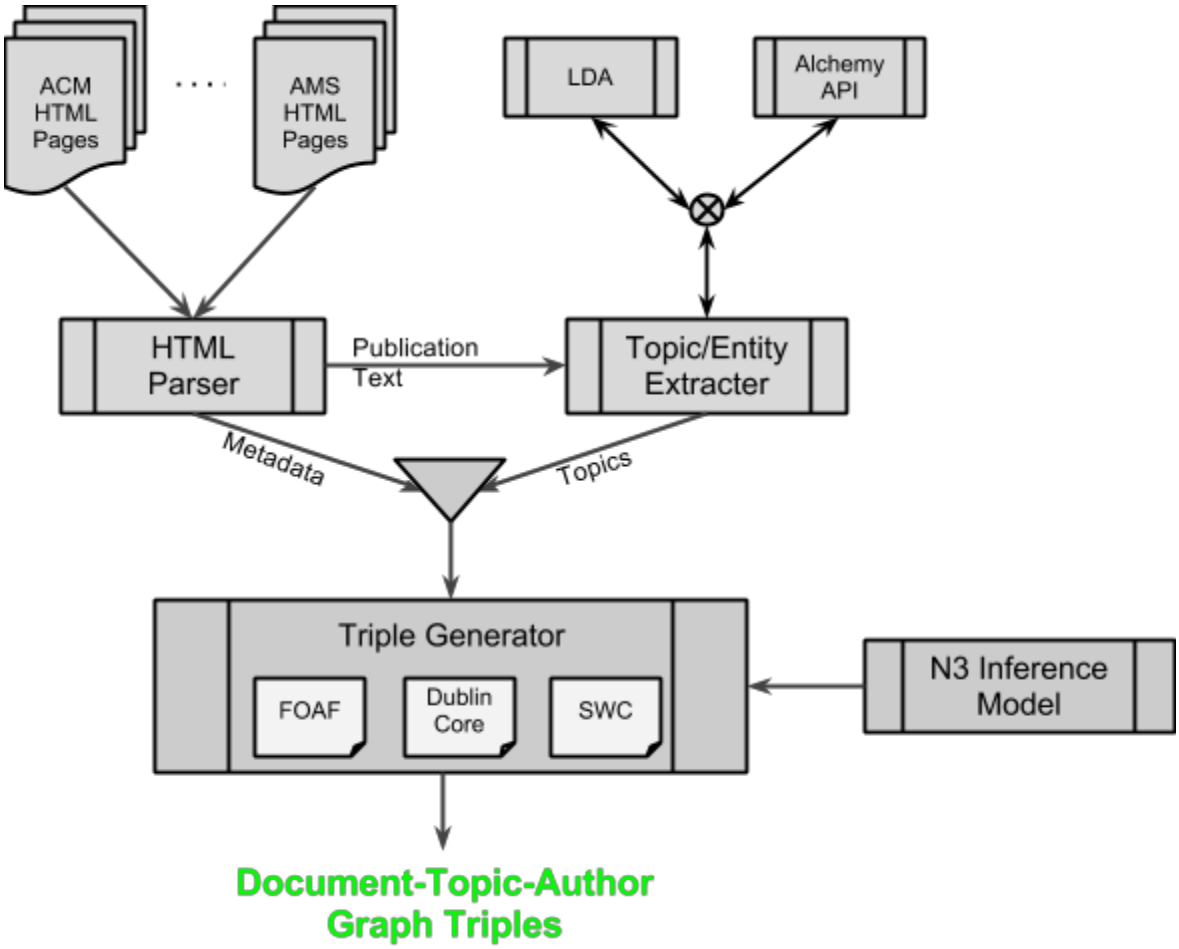
Related works

1. **Google Graph**[1]**:** This is a knowledge base used by google for performing search queries. It provides similar features as that of our proposed project, but for a larger domain. Although information like education, birth and related people for a person is easy to search, information about what topic a researcher is working on or bibliographic information of a particular paper that the person has authored is hard to find.
2. **Semantic Web Conference Corpus**[2]: aggregating publications to conferences dealing with the Semantic Web knowledge area only and generating triples for conference proceedings. Input is via RDFa, Excel or XML import. Human intervention is required to input the import file
3. **Google Scholar**[3]: is a simple tool to search for scholarly articles. Based on the search result, one can then explore related works, citations, authors, and publications. But It doesn't find infer about relationships amongst topics , authors and their publications.
4. **PubZone**[4]: is an open platform to discuss publications and rate articles. It also provides bibliographic information for the articles selected.

In most of the applications above, we find that the data is manually generated and imported in RDF, Excel or other structured formats. We believe commonalities in the structure of web sites

hosting these publication libraries can aid in reducing this human intervention and substantially automate the data import process. Also topic-centered inferences and graph generation is not seen extensively in the available solutions.

Proposed Approach



Data source : Publication/conference portals such as ACM, AMS etc.
Topic/Entity Extraction Modules: LDA[5] (Latent Dirichlet Allocation) libraries and/or AlchemyAPI [6]
Ontologies: Dublin Core[7], FOAF[8] and SWC[9]

Initial Bibliography:

[1] Google [Online]. Availabe: www.google.com

[2] Semantic Web Conference Corpus [Online]. Available: http://data.semanticweb.org/

[3] Google Scholar [Online]. Available: http://scholar.google.com/

[4] PubZone [Online]: Available: http://www.pubzone.org/index.do

[5] Latent Dirchlet Allocation [Online]: Availaible: http://www.cs.princeton.edu/~blei/papers/BleiNgJordan2003.pdf

[6] AlchemyAPI [Online]: Availaible: http://www.alchemyapi.com/

[7] Dublin Core Metadata Initiative [Online]: Availaible: http://dublincore.org/

[8] Friend-of-a-Friend Ontology [Online]: Availaible: http://xmlns.com/foaf/spec/

[9] Semantic Web Conference Ontology [Online]: Availaible: http://data.semanticweb.org/ns/swc/swc_2009-05-09.html