Using Semantic Technologies to Construct Research Topicology

Gandhi Sunil, Peshave Akshay and Rao Raghavendra Shruti

Department of Computer Science and Engineering
University of Maryland, Baltimore County
Baltimore, Maryland 21250
{sunilga1, peshave1, rrao1}@umbc.edu

Abstract

Conference publications typically deal with research carried out in a focussed subject matter of any science. These publications, and in turn their authors, can be represented as clusters or well connected graphs based on shared context. In this work, using topics and knowledge areas as the fundamental context, we implement a logical linking of published artifacts, their authors and knowledge areas. This paves the way for a scalable and configurable knowledge linking and reasoning system which will find numerous applications in enterprise-wide information retrieval, entity relationships modelling, data visualisation etc.

1. Introduction

In a conference, people with expertise in specific research area publish their findings and report. Besides authors, there may be associations, topics, conference location and many other meta data linked to a publication. Names and associations are directly related to one topic, which means, that a topic can have multiple names and associations. Associations interconnect topics through directed edges. Two topics that represent persons might be connected through an association, which is a type "isFriendOf". As one can see, the topic modeling enables the definition and classification of knowledge structures by means of well defined constructs.

We propose to create a knowledge base 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 (PoC) we implement the logic and demonstrate the process on publications on the American Computer Society (ACM) forum using minimum manual intervention.
2. Related works

Currently relevant and popular technologies related to similar area of work are:

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**\(^2\): aggregates publications and conferences dealing with the Semantic Web knowledge area and generates triples for conference proceedings. Input is via RDFa, Excel or XML import. However, human intervention is required to input the imported 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 infer relationships amongst topics, authors and their publications.

4. **PubZone**\(^4\): is an open platform to discuss publications and rate the articles. It also provides bibliographic information for the articles selected.

5. **Facebook Graph Search**: It is a semantic search engine based on open graph protocol. Open graph protocol is based on RDFa standards. In Facebook graph search, people are connected to each other based on their social network attributes like photos they are tagged in, interests, pages liked etc.

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.

Our vision is of a scalable and robust system which can be deployed in a closed enterprise-wide environment to connect various entities based on intersecting domain or enterprise-specific contexts. We draw inspiration from Facebook’s graph search which profiles its users based on their activity to include similar connections in their knowledge base. The differentiator in our approach is our objective of an extensible and robust deployment of a similar framework within a closed enterprise scope with reduced time to release.

Research Topicology is our first attempt at a proof-of-concept(PoC) to this effect in the context of research work and publications thereof.
3. System Schematic:

A major portion of the PoC is data extraction and cleansing. We downloaded the citation web pages for a representative set of publications as well as the publication documents. These downloaded artifacts were then parsed by HTML/PDF parsers and extractors to mine meta-data of the publications such as the title, DOI, authors, topics and CCS classification. Figure 1 is a depiction of the the system schematic and data flow. Alchemy API brings a Natural Language Processing flavor to tag extraction.

4. Topic/Entity Extraction Modules:

We explored two approaches for topic generation: LDA [5] (Latent Dirichlet Allocation) and AlchemyAPI [6].

4.1 Latent Dirichlet Allocation :

It is a generative probabilistic model for collections of discrete data such as text corpora. It is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document.
4.2 Alchemy API:

We used Alchemy API to add a NLP flavor to the tags extraction. This API returns topics, authors, organization, city and other classes of information when a document is given as input. For this PoC we have utilized it to cleanse the document tags and keywords based on language and domain semantics which are inherent to the Alchemy engine.

The precision and effectiveness of topics modeled by LDA were not satisfactory. Also it models topics relative to a cluster of documents which is not desired in our work. Alchemy API works on NLP principles and hence can extract topics and keywords from a standalone document, augmenting the process with open linked data sources it is interfaced with.

5. Ontologies

The following ontologies are used for representing documents we selected from the Association of Computing Machinery website.

1. Dublin Core[7]: Dublin Core can be used to describe web resources like video, images, web pages, physical resources such as books and objects like artworks. Each publication is considered to belong to the Document class. We used following classes and properties from Dublin Core ontology:
   a. dcterms:Document: All publications belong to this class
   b. dcterms:identifier: Identifies each document by its DOI (Digital Object Identifier).
   c. dcterms:creator: author property of the document.
   d. dcterms:title: indicates the title of the document.
   e. dcterms:subject: Describes each topic mined for the document.
   f. dcterms:source: indicates the ACM CCS knowledge areas that the publication belongs to.
   g. dcterms:partOf: this is a generic property used for sioc:Community class to indicate any entities that are part of the community. Entities can be members of any class provisioned by any ontology.

2. FOAF[8]: FOAF (Friend of a friend) is a machine-readable ontology describing persons, their activities and their relations to other people and objects. We used following classes and properties from FOAF ontology:
   a. foaf:Person: all authors belong to this class
   b. foaf:interest: Describes a person’s interest. This is used to model the interest of a author in topics on which they have published papers.
   c. foaf:knows: used to model the inference that co-authors know each other.

3. SIOC[9]: The Semantically Linked Online Communities is an ontology for describing online forums and communities and items and contributors therein. We used following
classes and properties from the SIOC ontology:

a. sioc:Community: Each CCS knowledge area is considered to be a community of which publications mapped to it and authors contributing to it are part of.
b. sioc:topic: this property indicates the finer topics which this knowledge area deals with.

4. ACM 2012 Computing Classification System (CCS): This is a poly-hierarchical ontology that can be utilized in semantic web applications. We use ACM CCS classification system for categorizing documents. We also use this classification system to go from broader topic to narrower topic or from narrow topic to broader topic.

Figures 2, 3 and 4 depict the entity-relationship units. It consists of initial properties used to describe the extracted data-set and the inferred properties which are detailed in the subsequent section.

6. Inference Rules

Initial knowledge extracted from the publications and their meta-data is extended using inference logic. We have attempted to model the inference logic to provide a holistic perception of the topicology paradigm we have based this PoC on and the numerous perspectives and applications that may emerge thereafter.

Below is a gist of the inference logic we have put in place:

1. All publications belong the classes foaf:Document and sioc:Item.
2. All authors belong to the class foaf:Person: Following is rule which indicate that all the authors are foaf:person.
3. Co-authors know each other. The property foaf:knows is used symmetrically for this purpose.

4. Topics that a publication deals with are inferred to be the authors’ interests.
5. ACM CCS knowledge areas are sources for documents. Further each of these areas can be perceived as a knowledge community consisting of publications and contributors who are authors in that knowledge area.

![Fig 3: Generalised Topicology Under CCS Umbrella](image)

6. Each knowledge area (and the corresponding community) encompasses finer topics which can be inferred from dcterms:subject properties of all the documents in the community. The sioc:topic property is used here to differentiate semantics when discussing about topics related to a document and a community.

![Fig 4: Community Topicology](image)

A better understanding of the initial assertions about entities and the inferred facts about them can be obtained by perusing through excerpts of the data provided in Appendix A.

7. Conclusion:

We have been able to provide a PoC of the envisioned system through the Research Topicology project. The various components of the system from data extraction and cleansing through semantic assertions and inferences provide evidence of the holistic nature of this approach. The system is extensible and scalable due to the use of lightweight...
technologies. Interoperability is inherent since widely popular ontologies and taxonomies have been utilised.

8. Future work

Immense possibilities exist within this realm. The applications of such a framework are numerous and diverse. Key advancements to further the effectiveness of such a system are:

1. Encapsulation in a generic framework based on standard configurations. This will enable quick deployment in an organisation scope with standard interfaces to structure and semi-structured data sources. This would require building upon standard data interface protocols and providing APIs to configure components such as ontologies, inference logic etc.
2. Extending linkages to open linked data corpuses such as DBpedia and non-functional testing of querying frameworks for such data.
3. Evaluation of various topic modelling and entity extraction frameworks and introducing them as configurable plugins or agents for the system.

9. External Project Links:
Below are links to the various deliverables and components of the project.

- Initial Knowledge Base
- Knowledge Inference Logic
- Research Topicology Complete with Inferred Knowledge
- Knowledge Extraction and Cleansing Code Base
- Knowledge Codification Code Base
- ACM CCS SKOS Navigator
- Latent Dirichlet Allocation Trials

Bibliography:
APPENDIX A : Sample Research Topicology

Below is an excerpt from the initial topicology triples modeled using the publications and associated meta-data. The inferred properties follow which completes the research topicology as suggested in the above sections.

Excerpt 1: Sample Publication Assertions

Excerpt 2: Inferred knowledge for an author of publication in Excerpt 1
Excerpt 3: Community Topicology

"G.2.1" a :Community;
  <http://purl.org/dc/terms/hasPart> "Alex Samorodnitsky",
  "Asaf Levin",
  "Dana Moshkovitz",
  "Dana Ron",
  "Dorit S. Hochbaum",
  "Hiroshi Fujiwara",
  "Michael Elkin",
  "Muli Safra",
  "Noga Alon",
  "Omri Weinstein",
  "Ronitt Rubinfeld",
  "Shmuel Safra",
  "Susanne Albers",
  <http://dx.doi.org/10.1145/1103963.1103968>,
  <http://dx.doi.org/10.1145/1150334.1150336>,
  <http://dx.doi.org/10.1145/1290672.1290686>,
  <http://dx.doi.org/10.1145/2382559.2382562>,
  <http://dx.doi.org/10.1145/2438645.2438651>;
  :topic "algorithms",
  "almost k-wise independence",
  "almost shortest paths",
  "approximation algorithm",
  "biclustering",
  "combinatorial algorithms",
  "combinatorics",
  "competitive analysis",
  "derandomization",
  "dynamic programming",
  "flow time",
  "gene expression",
  "general",
  "generalized hashing",
  "graph algorithms",
  "graph theory",
  "group testing",
  "influence of a boolean function",
  "k-restriction",
  "offline algorithms",
  "online algorithms",
  "order preserving",
  "sequencing and scheduling",
  "set-cover",
  "spanners",
  "splitter",
  "sublinear query approximation algorithms",
  "symmetric chains",
  "theory",
  "variable-speed processor".