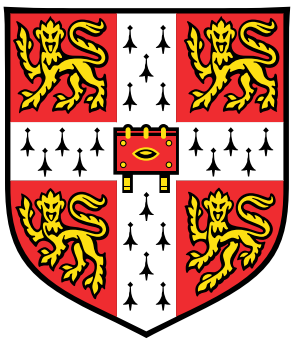
Social Semantic Network Analysis using Tourism Ontologies

Using the CUED template



Praveen Koirala

Department of Computer Science and IT Engineering

Pokhara University

This dissertation is submitted for the degree of

Master in Computer Information System

Nepal College of Information Technology

February 2015

Acknowledgements

I would never have been able to finish my dissertation without the guidance of my committee members, help from friends, and support from my family.

I would like to express my deepest gratitude to my advisor, Asst. Professor Suresh Pokhrel, for his excellent guidance, caring and patience for doing research. I would like to thank Asst. Professor Mohan Timilsina, who let me experience the research of Social Network Analysis in the field and practical issues beyond the textbooks, patiently corrected my writing. I would also like to thank Prof. Dr. Sashidhar Ram Joshi, Ass. Professor Saroj Shakya, Asst. Professor Sanjeeb Panday, Asst. Professor Kumar Pudasaini, Asst. Professor Basanta Joshi for guiding my research and helping me to develop my background in Semantic Web and Data Mining.

I would like to thank Mr. Ganesh Joshi, who as a good friend, was always willing to help and give his best suggestions. I would also like to thank my parents, and my brother. My research would not have been possible without their helps. They were always supporting me and encouraging me with their best wishes.

Abstract

Ontologies enhances the semantics by providing richer relationships between the terms of a vocabulary. Ontologies are usually expressed in a logic-based language, so that detailed and meaningful distinctions can be made among the classes, properties, and relations. Web Ontology Language (OWL) is a standard ontology description language, built on the Resource Description Framework(RDF). A key argument for modeling knowledge in ontologies is the easy re-use and re-engineering of the knowledge. However, beside consistency checking (e.g Pellet, HermiT, FacT++, RacerPro), current ontology engineering tools provide only basic functionalities for analyzing ontologies. Since ontologies can be considered as graphs(labeled, directed), graph analysis techniques are a suitable answer for this need. There are only very few social network analysis(SNA) applications up to now, and virtually none for analyzing the structure of ontologies. In my thesis I would like to propose the usage of existing state-of-the-art network statistics(e.g to measure Centrality,Cohesion) and algorithm (CPM algorithm) in a domain of ontologies to understand their structure, extract the hidden communities of ontologies and visualize them in a tourism sector.
Keywords: Ontology, OWL, RDF, Social Network Analysis, Semantic Web, Tourism

Table of contents

List of figures

Chapter 1

Introduction

1.1 Background

Social Web sounds like a pleonasm: user interactions and social networks are among the cornerstones of the Web. Human participation and freeform contributions are at the core of most popular web sites, creating shared spaces where people can freely gather, interact, and explicitly connect. From these usages, online communities of interests spontaneously emerge with roles and life cycles that are inherent in their members' interactions and involvements. Guided by common interests and goals, these communities publish, filter and organize directories of references in their domains at an impressive scale with very agile responses to changes. Now, we have access to an evergrowing long tail of information and knowledge.

The main problem is no more to collect and publish resources but mainly to structure and mash them in a way that matters to people and to their communities. Consequently, intelligent agents are crawling web resources, mining and indexing them in order to provide added-value services and extended information to web users. Interested by the audience driven through such activities, content providers make explicit and available their public data through the form of API or mark-ups in their pages. The activity of these agents is made easier and easier by the growing adoption of Semantic Web technologies to capture, publish and access data with standard machine-readable formats and protocols. In particular, we are witnessing the outburst of standard semantic mark-ups inside HTML pages, thanks to their consideration by biggest web actors (e.g. Google, Facebook, Yahoo). This exponential growth of readily available semantic data foster the deployment of more and more intelligent software that consume these linked and structured data to personalize, enrich and multiply user experiences (e.g. web augmented reality).

Intranets of organizations are progressively reproducing various web evolutions and web

based social applications are progressively deployed inside companies. For instance, Wikis are used to foster collaborative editions and knowledge capture, and social networking services to increase and ease sharing between employees. Intranet users are now able to partially adapt the flow of information inside the company to their daily tasks and evolving needs. However, social web applications inside intranets are more often disconnected, and corporate information is still more structured according to the organization chart rather than to how people use it. Beyond the reluctance related to emerging and auto-organized information, data that are produced by these applications lack the semantics and interoperability to be mashed and integrated in the intranet structure. The adoption of Semantic Web technologies could greatly benefit such social intranet by turning its information into structured data and connect it. Once semantically revealed, structured and connected, social data can in turn be exploited to develop functionalities that will structure information according to the need and the use of intranet users.

Several researches have been conducted to develop this social semantic perspective of web based applications, and we now dispose of standards to capture, to represent and to interlink socially produced and structured metadata. However, this important step toward applications that easily collect, mash and publish data, puts users and companies in front of a huge amount of social signals that need to be filtered and organized to avoid hindering their initial benefits. In particular, socially issued metadata embed an emergent structure that is inherent in user relations, interactions, and affiliations. Revealing this social structure would enable its exploitation to help filtering and organizing this huge amount of data.

This thesis investigates methods for identifying the social structure emerging from the semantic representation of online social activities. Building on top of Semantic Web technologies and classical graph theory, we propose a novel approach to take benefits of both models and conduct a semantic social network analysis. We will see how to semantically represent, link and access online social networks, how to enable classical operators of social network analysis to consider the semantics of these networks, and how these semantics could be exploited to enhance community detection.

This report is organized as follow:

Chapter 2 reviews the literature and definitions of the basic notions related to social network analysis and online social networks. Chapter 3 presents how Semantic Web technologies enable us to structure, link and exchange social networking data across web sites to conduct a semantic social network analysis and detect communities. Chapter 4 presents the Ongoing & Future works.

1.2 Problem statement

The semantic web community has introduced many, independently created, ontologies. These ontologies cover real-world domains, but are created and structured by humans. Current ontology engineering tools (e.g RDF2Go,Protg,OntoEdit) provide only basic functionalities for analyzing ontologies. Since ontologies can be considered as labeled and directed graphs, graph analysis techniques are a promising tool to understand their structure. This problem can be overcome using Semantic Network Analysis with graph representation of these ontologies.

1.3 Research Objective

The objective of this research is to

- to use novel graph algorithm(OntoGraf,OWLviz using protege tool) to construct the ontology graph
- to run an analysis to explore the ontology communities and,
- to identify concept groups which are hidden in the graph.

1.4 Significance of the study

The Resource Description Language (RDF) specifies a way to state information in terms of statements[4] which is a triple consisting of a subject, object and a predicate. The RDF Schema (RDFS) allows us to define an RDF based vocabulary. The Web Ontology Language (OWL) is the ontology language for the semantic web . The semantic web vocabularies are represented as graphs. In RDF, the predicate represents an edge between the subject and object nodes. RDFS also allows for relationship between predicates, preventing them from being viewed as pure edges from a network analysis perspective since such relationships implies that there are edges between predicates. The semantic modeling of tourism information enables intelligent tourism information systems to provide personalized services. An intelligent tourism information system includes ontology-driven subject domain and repository of tourism information. It is adaptive to users needs (e.g. user requires. Information management tasks are annotated in terms of subject domain. So that we can build intelligent recommendations services in tourism support systems using ontologies. A suite of tourism ontologies can be developed and engaged to enable a prototypical e-tourism system with

various knowledge-based recommendation capabilities. A usability evaluation of the system yields encouraging results as a demonstration of the viability based on the relationship among ontologies.

Chapter 2

Litrature Review

2.1 Semantic Web

2.1.1 The Semantic Web Initiative

In 1992 Tim Berners-Lee created the World Wide Web Consortium (W3C) with the goal to develop,extend, and standardize the Web. W3C research eventually led to the conceptual development of the so called Semantic Web, that is described by Berners-Lee et al. (2001) as an extension of the current Web in which information is given well defined meaning, better enabling computers and people to work in cooperation. Van Harmelen et al. (2000) describe the Semantic Web as a range of standards, modeling languages and tool development initiatives aimed at annotating Web pages with well defined metadata, so that intelligent agents can reason more effectively about services offered at particular sites.

2.1.2 Ontologies

An ontology is an agreed vocabulary that provides a set of well-founded constructs to build meaningful higher level knowledge for specifying the semantics of terminology systems in a well defined and unambiguous manner Bettina Hoser [2]. In other word, An ontology is a conceptualization of a domain into machine readable format. For a particular domain, an ontology represents a richer language for providing complex constraints on the types of resources and their properties.

Ontologies can be used to increase communication both between humans and computers. The three major uses of ontologies are:

- To assist in communication between humans.

- achieve interpretability and communication among software systems.
- improve the design and the quality of software systems.

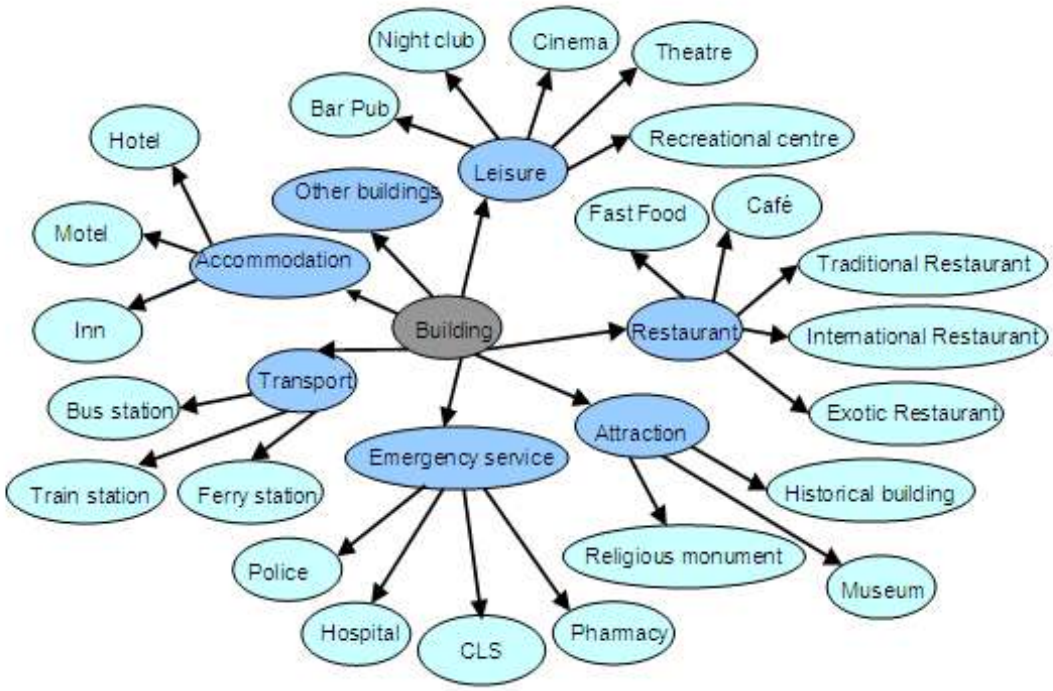


Fig. 2.1 Ontology: Domain,Classes and property

2.1.3 Ontology Web Language (OWL)

The OWL language, which was created by the W3C Web Ontology (WebOnt) Working group derive from DAML+OIL. Like DAML+OIL, OWL builds on RDF and RDF Schema and adds more vocabulary for describing properties and classes: among others, relations between classes (e.g. disjointness), cardinality (e.g. "exactly one"), equality, richer typing of properties, characteristics of properties (e.g. symmetry), and enumerated classes Guillaume Ereteo [6].

2.1.4 RDF: a Standard Resource Description Framework

RDF (Manola and Miller 2004), which stands for Resource Description Framework, is a data model and syntax specification for representing information about Web resources. An RDF Model is a set of statements, each consisting of a triple (i.e. subject, predicate, object). RDF

statements can either be represented as a graph or in an XML format known as RDF/XML serialization.

RDF enables us to make assertions and describe resources with triples (subject, predicate, object) that can be viewed as "the subject, verb and object of an elementary sentence", "a natural way to describe the vast majority of the data processed by machines" [Berners-Lee 2001]:

- The subject represents the described resource.
- The predicate represents the property used to describe the resource.
- The object represents the value of the property for the described resource.

2.1.5 SPARQL: Protocol and RDF Query Language for Querying and Accessing Data

Once provided with a semantic dimension, we can query this graph with SPARQL (SPARQL Protocol and RDF Query Language). SPARQL is an RDF query language and data access protocol. It defines a query language to query triples, different protocols to send queries and their results across the web, and a result format to exchange these results Olivier Corby [9]. The queries are composed of four blocks:

- PREFIX "to declare the schemas used in the query", this clause is optional.
- A clause to determine the type of query and identify the values to be returned. We have three types of clauses:
  - SELECT: "returns all, or a subset of, the variables bound in a query pattern match".
  - CONSTRUCT: "returns an RDF graph constructed by substituting variables in a set of triple templates".
  - ASK "returns a boolean indicating whether a query pattern matches or not".
  - DESCRIBE: "returns an RDF graph that describes the resources found".
- FROM clause "to identify the data source to query". This clause is optional and the default graph is queried when it is not used.
- WHERE clause, "a conjunction of triples" that defines "the triple/graph pattern to be matched against the triples/graphs of RDF"

The following example proposes to return all the agents and their first names (and only the persons that have a first name):

```
PREFIX foaf: < http://xmlns.com/foaf/0.1/>
SELECT ?person ?name
WHERE {
?person rdf:type foaf:Agent
?person foaf:firstName ?name
}
```

2.2 Semantic Analysis of Social Networks

The social data that emerge in online social applications embed rich social links,between their users, that have to be revealed and reified in order to be mined and exploited. In particular, these applications enable their users to connect, interact and develop interest affiliations between each other, which enable us to build and mine the resulting social networks. However the structures of these social networks are complex to represent, due to the multiplicity of context, roles and identities, and to their distribution across applications Guillaume Ereteo [6]. Each user of an application represents a person, in a particular role and a given context that constitute a fragment of its identity. Consequently, a person develops different social links, across several applications,which are contextualized by the different fragments of its identity. An effective mining of the resulting global social network should consider such specificities and require thus an adequate representation.

Semantic Web technologies answer the problematic of exchanging, mashing and querying data across applications. Based on these technologies, we need to reuse existing models and develop new ones, if necessary, to smartly represent people, user profiles and their different social links for revealing the online social networks they form. Once represented in a uniform structure, these social networks can then be mined for extracting the metrics that will be used for managing social data. Bizer and Berners-Lee [3]

Social Network Analysis is particularly well suited for understanding and determining the global structure of a social network, the distribution of actors and activities, and the strategic positions and actors. The result of the network analysis can be exploited for leveraging the social experience of collaborative tools. On one hand, we can better organize and filter social data in every step of the business intelligence process. During the searching, monitoring, collecting and handling steps, the presentation of social data to the users should consider the insight of a network analysis as well for classification purposes as for information quality indicators. During the disseminating step, the network analysis metrics

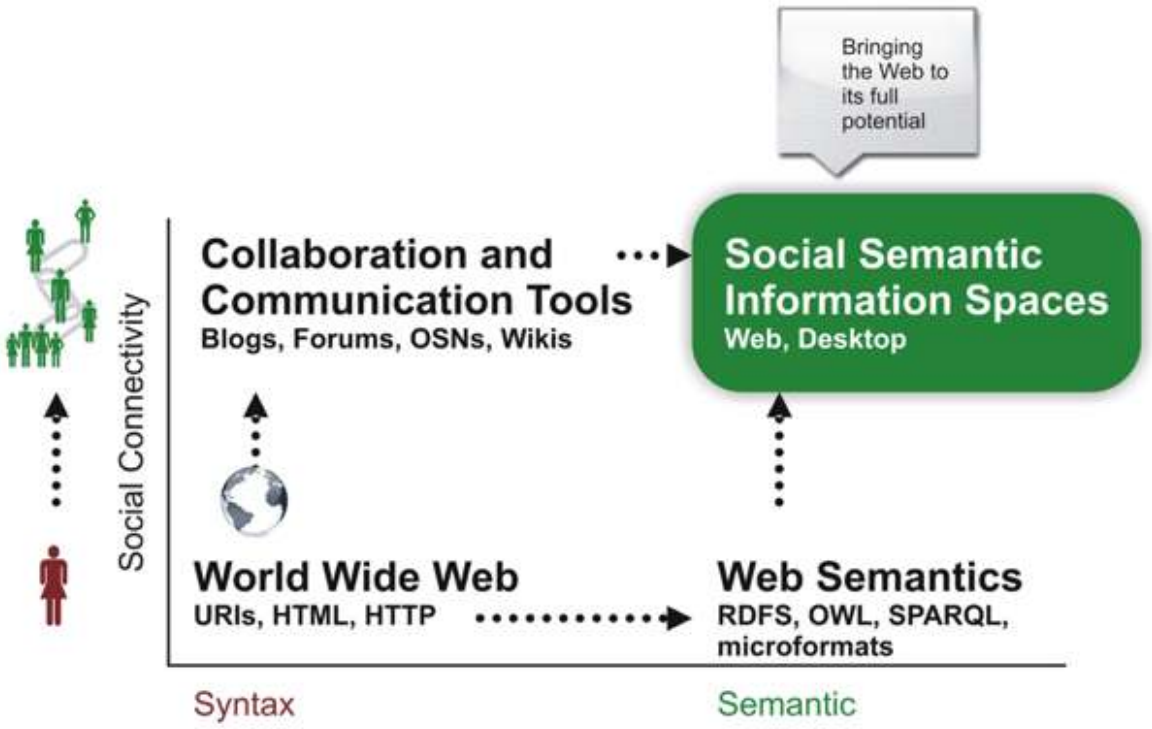


Fig. 2.2 Social Semantic Web: Syntax and Semantics

will help propagating the produced information toward the relevant part of the network and connect it to the targeted communities. On the other hand, the analysis can be used for strengthening the network structure in order to increase its efficiency, both locally and globally. At a local scale, the analysis can be used to assist applications' users in maintaining their relationships and developing relevant new ones that could best serve their efficiency.At a global scale the analysis could be used to stimulate new connections that would empower the whole network efficiency, such as bridging disconnected communities that would benefit from collaborating. Alexander Maedchea [1]

However, social network analysis' algorithms are only based on the linking structure of the network and do not exploit the semantics that are embedded in such typed representations from which they could highly benefit. The richness and the specificities of online social networks offer many perspectives for conducting more accurate analysis. In particular, the online artifacts that mediate interactions and develop affiliations provide social networks with the purposes of the creation, the maintenance or the disappearance of social links. In addition, the semantic structuring of the vocabulary generated by the users provides social networks with the knowledge that is produced, maintained and shared by their members to support their exchanges. Social network analysis is now provided with

these multidimensional representations that include not only the linking structure but also the shared knowledge of the social network.

Our Research objective is to leverage social network analysis metrics for handling the semantic representations of social networks, which is a necessary step for fully achieving the social evolution of the business intelligence.

2.3 Semantic Web Framework for Social Network analysis

The Figure 2.3 presents the stack of tools to conduct a semantic social network analysis. The goal of this stack is to provide a framework that enables us to consider not only the network structure embedded in social data, but also the schemas that are used to structure, link and exchange these data. This stack is composed of

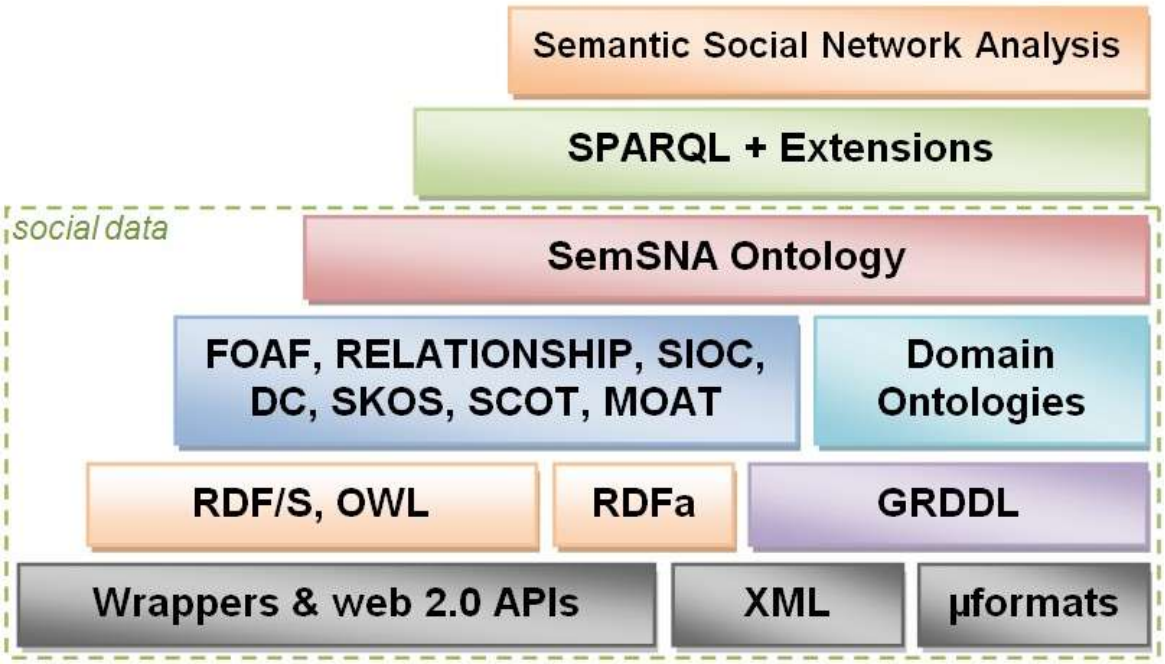


Fig. 2.3 Abstraction Stack for Semantic Social Network Analysis

1. tools for building, representing and exchanging social graphs and
2. tools for extracting social network analysis metrics and leveraging social graphs with their characteristics.

2.4 State of the Art on Social Network Analysis and Its Application on the Web

Social Network Analysis (SNA) provides graph algorithms to characterize the structure of social networks, strategic positions in these networks, specific sub-networks and decompositions of people and activities. This domain has raised lots of interests and the outburst of social data on the web has led to the collection of the biggest social networks ever Guillaume Ereteo [7].

In this Chapter, we will review

1. the traditional methods and models that are used to build and represent social networks,
2. the different metrics and algorithms of social network analysis, and
3. the applications of social network analysis to online social networks.

2.4.1 Representing Social Networks

A social network is made of actors that are linked by social relations. Social actors can be people, organizations, or groups of actors. A wide range of social relations exists between actors; we can group these relations in three categories:

- *explicit and declared relations* between humans
- *interactions* between actors.
- *affiliation* between actors.

Explicit relations include all the relations we can define between persons (e.g. parent,sibling, cousin, friendship, love, simple acquaintance, co-worker, etc.), between persons and organizations (e.g. member, employee, etc.), and inside organizations (e.g. owner,manage, etc.).

Interactions represent all the exchanges that could be observed between actors such as a discussion, a collaboration, a meeting or any action that involves at least two actors. Some interactions actively implicate all the concerned actors, like a synchronous discussion. Others are initiated by some actors and target other actors, like a single message that has a sender and a recipient Bettina Hoser [2].

Affiliations correspond to any similarity between actors that links them, like, for instance, sharing the same attributes, the same interests, the same activities, the same objects, or the same organizations.

2.4.2 Social Network Representation

In order to graphically visualize social networks, in 1930s, Moreno systematized the first representations of social networks: the sociograms [Moreno 1933]. Sociograms consist in representing people by points and relationships by lines connecting points.These representations were also named 'web' due to their spider web aspect, this is an interesting unintentional coincidence of history. As little innovative as it may appears today, this type of visualisations offered to quickly detect some network features that are highlighted by specific visual patterns. As an example, Moreno introduced the concept of "star" for designing people having the most connections in a social network,due to the star shape formed by a point and its numerous connected lines. Sociograms were the first step for further involvement of mathematicians in social network analysis.

Definitions

Node: basic unit of a network that represents a resource, also called a vertex. In a social network we talk about actors or agents.

Edge: a connection between two nodes. We also use the terms arcs or links.

Hyperedge: an edge than connects more than two nodes.

Directed edge: an edge used in only one direction, from its source node to its end node. In opposition, an undirected edge can be used in both directions and does not distinguish its extremity nodes.

Weighted edge: an edge with an assigned a value, called a weight, to represent the importance of this edge.

Labelled edge: an edge with a term used to label the relation.

Graph: a graph is defined by a set of nodes and a set of edges.

Hypergraph: an hypergraph is defined by a set of nodes and a set of hyperedges [Berge 1985]

Directed graph: a directed graph is defined by a set of nodes and and a set of directed edges.

Weighted graph: a weighted graph is defined by a set of nodes and a set of weighted edges.

Labelled graph: a labelled graph is defined by a set of nodes and a set of labelled edges.

Representing Graphs With Matrices

Matrix: A matrix is a rectangular table of values, in which each cell is noted a_{ij} with i and j that are the row and the column of the cell.

Matrices are popular mathematical objects for handling graphs. Usually, when modeling a graph with a matrix, the rows and the columns of a matrix represent the nodes of the graph, and the value in the cell a_{ij} in the matrix represents an edge (or an absence of edge) between the corresponding nodes v_i and v_j. Generally, two types of matrices are used for representing social networks:

1. adjacency matrices which rows and columns represent *the same sets of nodes*, and
2. incidence matrices which rows and columns represent *different sets of nodes*.

Adjacency Matrix: An adjacency matrix is a squared matrix in which rows and columns are labelled by the same list of nodes (the i^{th} row and the i^{th} column represent the same node).

Incidence Matrix: An incidence matrix have two types of nodes (e.g. authors and papers) with rows representing one type and columns the other one.

Degree Matrix: The degree matrix of a graph is a diagonal matrix with the degree of the graph's nodes on the diagonals

2.5 Social Network Analysis

SNA tries to understand and exploit the key features of social networks in order to manage their life cycle and predict their evolution. Much research has been conducted on SNA using graph theory Freeman [4]. Among important results is the identification of sociometric features that characterize a network. SNA metrics can be decomposed into two categories;

1. some provide information about the position of actors and how they communicate and
2. others give information about the global structure of the social network.

The *Centrality* highlights the most important actors and the strategic positions of the network. The main question of centrality is to define what makes an actor more central than another one Tim Finin [10]. Different criteria have been considered to define the centrality, and the chosen criteria enable to obtain different information about the position of actors. The three main definitions of centrality are resumed by [Freeman 1979]: the degree centrality, the betweenness centrality and the closeness centrality.

Degree Centrality: Degree centrality gives us a measure of how well connected a node is in a graph. The degree of a resource is the number of resources adjacent to it. The Degree centrality considers nodes with the highest degrees (number of adjacent edges) as the most central.

Betweenness centrality: The *betweenness centrality* considers nodes that are more often on shortest path between other nodes as the most central. Betweenness centrality gives the normalized shortest path between nodes that pass through the given node. The betweenness centrality focuses on the capacity of a node to be an intermediary between any two other nodes.

Closeness centrality: The *closeness centrality* considers as most central the nodes that have the smallest average length of the roads (sequence of relationships) linking an actor to others. The closeness centrality of a resource represents its capacity to join (and to be reached by) any resource in a network. It reveals the ability of a node to quickly connect with all the other actors of the network.

2.6 Protege

Protege is an ontology and knowledge base editor produced by Stanford University. Protege is a tool that enables the constriction of domain ontologies,customized data entry forms to enter data. Protege allows the definition of classes, class hierarchies,variables,variable value restrictions,and the relationships between classes and the properties of these relationships.Protege is free and can be downloaded from http://Protege.stanford.edu. Protege comes with visualization packages such as OntoGraf and OWLviz; all of these help the user visualize ontologies with the help of diagrams. The main strong point of protege is that it support at the same time tool builders, knowledge engineers and domain specialists. This is the main difference with existing tools, which are typically targeted at the knowledge engineer and lack flexibility for meta-modeling. This latter features makes it easier to adopt protege too new requirements and/or changes in the model structure.

Protege 5.0 and GraphViz supports two ontology viewing tools namely OWLViz and OntoGraf. OWLViz: shows only classes hierarchy but doesn't show relationships and properties. OntoGraf shows classes with hierarchy and relationships, but does not print the relationship names on the graph. The tool is configurable or flexible. Taxonomy is a central part of most conceptual models. Properly structured taxonomies gives substantial order to elements of a model, it is useful in presenting limited views of a model for human interpretation, and plays critical role in reuse and integration tasks. Properly constructing taxonomy providing insights that have been focused on the semantics of the taxonomic relationship (also called is-a,class inclusion.), on different kinds of relations (generalization, specialization, subset hierarchy) according to the constraints involved in multiple taxonomic relationships (covering, partition, etc.), on the taxonomic relationship in the more general framework of data abstractions, or on structural similarities between descriptions.

2.7 Litrature Review

Bettina Hoser [2] performed social network analysis on the SUMO (Suggested Upper Merged Ontology) and SWRC (Semantic Web for Research Communities) ontologies. They found that social network analysis provide useful insights into the structure of ontologies. They found the need to preprocess ontologies to a simpler structure prior to the social network analysis. In this paper the authors explored the use of centrality analysis on the ontologies.

They specifically identified betweenness centrality and eigenvector centrality for these two ontologies.They consider the betweenness centrality useful in identifying the core concepts in the ontology. Stuckenschmidt analyzed ontologies and used relative strengths to determine if ontology needs to be partitioned. In the paper the author represented the ontology as a proportional strength network where the weight of the relationship is determine by the inverse of the degree of the node. The partitions were then determined by applying minimal cut algorithm on the graph.

G. Coskum [5] used social network analysis on ontologies to identify concept groups. In this paper the authors investigate nine different representation of ontology as a graph. The three basic representation being a plain RDF graph structure, a graph where the predicates are also represented as nodes and a third where only the classes are represented as nodes. Each of these representations had two extensions, one where the literals were ignored and another where the RDF, RDFS, OWL and XML Schema nodes were ignored.

Social network analysis has also been used for the development of ontologies Mika [8].

Chapter 3

Methodology

3.1 Preprocessing the Ontologies

In order to generate the Tourism Ontology datasets, we have collected three different OWL ontologies from three different source. As Social Network Analysis techniques work on graphs, we have first transform the ontology into a suitable graph.

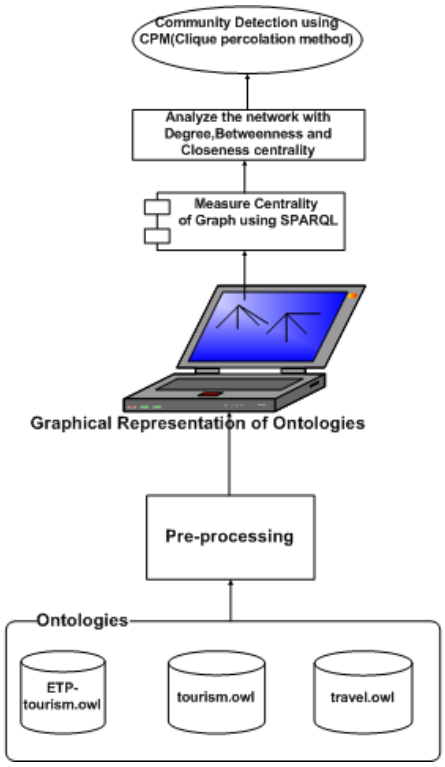


Fig. 3.1 Methodology and System Architecture Process Flow

3.2 Graphical Representation of Ontologies

A social network indicates the ways in which nodes (e.g. individual or organization) are connected through various relationships to each other. The social network analysis, in this context, is a way of processing and interpreting the nodes and relationships to realize mutual interests and connection between different groups and individuals in a community. Several studies have been carried out, mostly focusing on graph theory and statistical methods, to analyze relationships and connections in a social network. The Semantic Web technologies support social network analysis by providing explicit representation of the social network information.

The logical reasoning could be outlined based on stated triples. Each triple T is defined as a 3-tuple T = (D; P;R), where D represents the domain of the triple; P represents the property connecting the concepts D and R; and R represents the range variable of the triple. Each triple can be considered as a link that connects two neighborhood conceptual nodes, and the list of triples together forms an integrated ontology graph.

Other spatial and temporal similarities between entities could also be considered in a semantic association discovery.

3.3 Measure Centrality with SPARQL

In this section, we present how we query the RDF social graph using path extensions in SPARQL to compute centralities and identify strategic positions using the three definitions of Freeman. The path filtering features are used to parameterize the analysis: in order to deal with the diversity of social interactions that are captured online, we parameterize SNA features extraction with parameterized SPARQL queries and focus on relevant RDF subgraph.

Most forms of SPARQL queries contain a set of triple patterns called a basic graph pattern. Triple patterns are like RDF triples except that each of the subject, predicate and object may be a variable. A basic graph pattern matches a subgraph of the RDF data when RDF terms from that subgraph may be substituted for the variables. Hence executing SPARQL queries generally involves graph pattern matching.

SPARQL graph patterns that involves pathes through the RDF graph convert to subject-object joins in the SQL and patterns that involves multiple attributes about the same entity involves subject-subject joins in the SQL.

Centrality highlights the most important actors and the strategic positions of the network.

3.3.1 Degree Centrality

Degree centrality is a measure of the centrality of a node in a network and is defined as the number of edges (including self-loops) that lead into or out of the node. Degree centralities therefore lie between and inclusive, where is the number of vertices in a graph, and identify nodes in the network by their influence on other nodes in their immediate neighborhood.

Degree Centrality will give high centralities to vertices that have high vertex degrees. The vertex degree for a vertex is the number of edges incident to v. For a directed graph, the in-degree is the number of incoming edges and the out-degree is the number of outgoing edges.For an undirected graph, in-degree and out-degree coincide.DegreeCentrality works with undirected graphs, directed graphs, multigraphs, and mixed graphs.

The n-degree of a resource in a RDF graph is the number of paths of length n or less starting from or ending to this resource in other words, this is the number of sequences of n properties having this resource at an end.

We notate the parameterized degree: $deg_{<type,length>}(y)$ and extract it with the following query:

```
select ?y count(?x) as ?degree where{
{?x $path ?y
filter(match($path, star(param[type])))
filter(pathLength($path) <= param[length]) }
UNION
{?y $path ?x
filter(match($path, star(param[type])))
filter(pathLength($path) <= param[length]) }
} group by ?y
```

3.3.2 Betweenness Centrality

Betweenness Centrality returns a list of nonnegative machine numbers that approximate particular centrality measures of the vertices of a graph. Betweenness Centralities lie between and inclusive, where is the number of vertices in a graph. A betweenness centrality is a measure of the centrality of a node in a network based on the number of shortest paths that pass through it. Betweenness Centrality therefore identifies nodes in the network that are crucial for information flow.

It will give high centralities to vertices that are on many shortest paths of other vertex pairs and it works with undirected graphs, directed graphs, multigraphs, and mixed graphs.

According to Freeman, the betweenness of k for a couple of resource (i, j) is the probability for k to be on a shortest path from i to j:

$$b_{ij}(k) = \frac{g_{ij}(k)}{g_{ij}} \tag{3.1}$$

with g_{ij} the number of shortest paths between i and j and $g_{ij}(k)$ the number of shortest paths between i and j going through k. Then the betweenness centrality of k in the whole network is the sum of its betweenness for all possible couples of resources:

$$C_b(k) = \sum_{i}^{n} \sum_{j=1}^{n} b_{ij}(k) \tag{3.2}$$

We compute $g_{count<type>}$(from,to) the parameterized number of shortest paths between from and to, using the following query:

```
select ?from ?to count($path) as ?nbPaths
where{
?from $path ?to
filter(match($path, star(param[type]), 'sa'))
} group by ?from ?to
```

We compute $g_{count<type>}$(b,from,to) the parameterized number of shortest paths between from and to going though b using the following query:

```
select ?from ?to ?b count($path) as ?nbPaths
where{
?from $path ?to
graph $path{?b param[type] ?j}
filter(match($path, star(param[type]), 'sa'))
filter(?from != ?b)
optional { ?from param[type]::?p ?to }
filter(!bound(?p))
} group by ?from ?to ?b
```

We consequently defined the parameterized betweenness:

$$B_{<type>}(b, from, to) = \frac{g_{count<type>}(b, from, to)}{g_{count<type>}(from, to)} \tag{3.3}$$

Finally, compute $C_{b<type>}$(b) the parameterized betweenness centrality, by summing them:

Let $C_{b<type>}(b) \leftarrow 0$
For each pair <?from, ?to> connected by a shortest path going through b :
$C_{b<type>}(b)+=B_{<type>}(b,from,to)$

3.3.3 Closeness Centrality

Closeness Centrality returns a list of nonnegative machine numbers that approximate particular centrality measures of the vertices of a graph. Closeness centralities lie between 0 and 1 inclusive. A closeness centrality is a measure of the centrality of a node in a network based on the mean length of all shortest paths from that node to every other reachable node in the network. Closeness Centrality therefore identifies nodes in the network that are crucial for the quick spread of information.

It will give high centralities to vertices that are at a short average distance to every other reachable vertex and works with undirected graphs, directed graphs, weighted graphs, multigraphs, and mixed graphs.

The closeness centrality of a node is the inverse sum of its shortest distances to each other resource with n the number of nodes and d(k,i) the length of a shortest path from k to i the closeness centrality is:

$$C_c(k) = \sum_{i=1}^{n} d(k,i) \tag{3.4}$$

A shortest path in an RDF graph corresponds to a minimum number of triples that connect two resources. We note $C_{c<type>}(y)$ the parameterized closeness centrality and we extract it with the following query:

```
select ?y ?to pathLength($path) as ?length
sum(?length) as ?centrality where {
?y $path ?to
filter(match($path, star(param[type]), 's'))
} group by ?y
```

3.4 Community Detection

We Perform community detection on the ontologies to understand its structure. If there are islands within the network the nodes are not connected to the rest that would indicate that the ontology lacks cohesion. We can use WCC (weakly Connected Components) to help identify if the ontology contains any unrelated island of concepts.

We analyze the graph using the Clique Percolation Method (CPM) to detect closely related topics/concepts the ontology.

Modeling of social networks can be aided by ontologies out of several reasons. First ontologies are commonly deployed for the specification and explication of concepts and relationships related to a given domain. Social networks have the same purpose but with the focus on social relations and entities, hence domain ontologies related to social entities and relations can be designed and deployed.

Second, through reasoning and inference ontologies do not allow the modeling of contradictory or inconsistent information. Modeling social networks via ontologies ensure the validity of the information encoded.

Third ontologies, together with the inference mechanism, enable information gain through deploying rules to infer new information. Inference mechanism can be facilitated over ontology based social networks to come up with new relations and concepts out of the already existing ones between the social entities i.e. people, organizations and events, locations.

CFinder as a software tool for finding and visualizing overlapping dense groups of nodes in networks, based on the Clique Percolation Method (CPM) of Palla et. al. CFinder was recently applied to the quantitative description of the evolution of social groups: Palla et. al.,. It offers a fast and efficient method for clustering data represented by large graphs, such as genetic or social networks and microarray data. CFinder is also very efficient for locating the cliques of large sparse graphs.

3.4.1 The outline of the community finding algorithm

- The k-clique community finding algorithm built in CFinder first extracts all complete subgraphs of the network that are not parts of larger complete subgraphs.These maximal complete subgraphs are simply called cliques, (and the difference between k-cliques an cliques is that k-cliques can be subsets of larger complete subgraphs).

- Once the cliques are located, the clique-clique overlap matrix is prepared. In this symmetric matrix each row (and column) represents a clique and the matrix elements are equal to the number of common nodes between the corresponding two cliques, and the diagonal entries are equal to the size of the clique.

- The k-clique-communities for a given value of k are equivalent to such connected clique components in which the neighboring cliques are linked to each other by at least k-1 common nodes. These components can be found by erasing every off-diagonal entry smaller than k-1 and every diagonal element smaller than k in the matrix, replacing the remaining elements by one, and then carrying out a component analysis

of this matrix. The resulting separate components are equivalent to the different k-clique-communities.

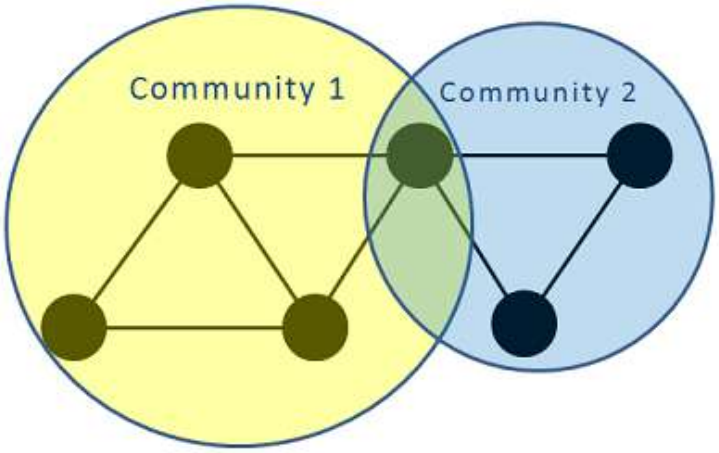


Fig. 3.2 Illustration of Clique Percolation Method: Community Structure k=3

Chapter 4

Data Set,Experiment and Results

4.1 Data Sources

Structured data is preferred (i.e. RDF or OWL) to analyze the Network and their relation to identify the hidden communities with in the network. Semi-structured or parsable unstructured data (i.e. XML) can be transformed to structured data using xPath or XSLT. Data with rich metadata and relations is preferred. For example, for a "Computer Scientist" class, the source also provides "address", "country" attributes as well as some relations with other classes such as "Research Area", "Publication", "Organization". The data set should have rich relations and large amount of instances which are highly connected.

The raw Ontology datasets have been collected from three semistructured sources.

1. tourism.owl: This ontology is collected from *code.google.com ontology repository*
2. ETP-tourism.owl: This ontology is collected from *Department of Computer Science - University of Quebec at Montreal*
3. travel.owl: This ontology is collected from *Protege Community Wiki*

The table below summarizes the number of statements,subjects,objects and properties in the above ontologies. The number of statements in OWL ontologies is usually an order of magnitude higher than those in RDFS Schemas.

4.1.1 Basic Network Properties of Ontologies

The table below summarizes some of the computed basic properties of the graph representation of the selected ontologies.

Table 4.1 Property comparison of three different otologies

Metrics			
Ontology Properties	ETP-tourism.owl	tourism.owl	travel.owl
Axiom	907	80	145
Logical axiom count	533	52	93
Class count	194	11	35
object property count	41	10	6
data property count	46	7	4
individual count	19	0	14
Class Axioms			
Ontology Properties	ETP-tourism.owl	tourism.owl	travel.owl
Subclass of axioms count	216	13	30
Equivalent classes axioms count	2	3	7
Disjoint classes axioms count	93	0	10
GCI(General Concept Inclusion) count	0	0	0
Hidden GCI count	2	0	0
Object Property axioms			
Ontology Properties	ETP-tourism.owl	tourism.owl	travel.owl
Sub object property of axioms count	4	9	0
Equivalent object properties axioms count	0	0	0
Inverse object properties axioms count	5	1	1
functional object property axioms count	7	2	0
Data Property axioms			
Ontology Properties	ETP-tourism.owl	tourism.owl	travel.owl
subdataproperty of axioms count	0	6	0
functional data property axioms count	19	0	4
data property range axioms count	46	5	4

Table 4.2 Degree and betweenness centrality of concepts and relations

S.No.	Label	d_o	d_i	b_c
1	Attraction	3	4	0.63
2	CulturalHeritage	2	3	0.31
3	ThemePark	2	4	0.24
4	ZooAndAquariam	2	1	0.03
5	NaturalHeritage	2	1	0.52
6	ReligiousHeritage	1	2	0.04
7	Sport	5	4	0.27
8	Gulf	2	3	0.39
9	Jogging	1	2	0.54
10	Skiling	1	3	0.08
11	Byking	1	1	0.
12	Swimming	1	2	0.04
13	Surfing	2	2	0.27
14	Adventure	2	3	0.41
15	Hunting	1	2	0.54
16	Hiking	2	3	0.61
17	Fishing	1	2	0.
18	DesertExeursion	2	3	0.44
19	Safari	2	3	0.4
20	BoatExeursion	1	2	0.6
21	Recreation	3	2	0.
22	Dancing	2	3	0.01
23	SightSeeing	3	4	0.05
24	Sunbath	2	1	0.04
25	Shopping	2	2	0.02
26	Training	1	2	0.
27	MusicLessons	3	2	0.6
27	LanguageCourse	3	2	0.5
28	Craftsmanship	2	1	0.01
29	FoodAndDrinks	2	1	0.05
30	Restaurant	3	3	0.04
31	Cafe	1	2	0.3
32	Pub	3	2	0.017
33	Club	3	2	0.05
34	Bar	2	1	0.01
35	Tavern	3	1	0.05
36	Accommodation	2	4	0.04
37	Cottage	1	2	0.08
38	Chalet	1	2	0.03
39	Hotel	3	4	0.06
40	Hostel	2	3	0.07
41	BreadAndBreakfast	2	1	0.05
42	GusetRoom	1	3	0.04
43	RestaurantRoom	1	3	0.04
44	Festival	2	3	0.14
45	Ceremony	3	4	0.054

Table 4.3 statements,subjects,objects and properties of Ontologies

Ontology	Properties	Objects	Statement	Subject
tourism.owl	72	136	425	95
ETP-tourism.owl	48	112	242	72
travel.owl	32	94	186	56

Table 4.4 Properties of Graph Representation of Ontologies

Ontology	Nodes	Edges	Avg. Degree
tourism.owl	54	326	14.2
ETP-tourism.owl	28	224	15.6
travel.owl	16	188	13.4

4.2 Graphical Representation of Ontologies

The Graphical Representation of Ontologies using Protege OntoGraf and OntoViz with AT&T's highly sophisticated Graphviz visualization software are shown in figure below.

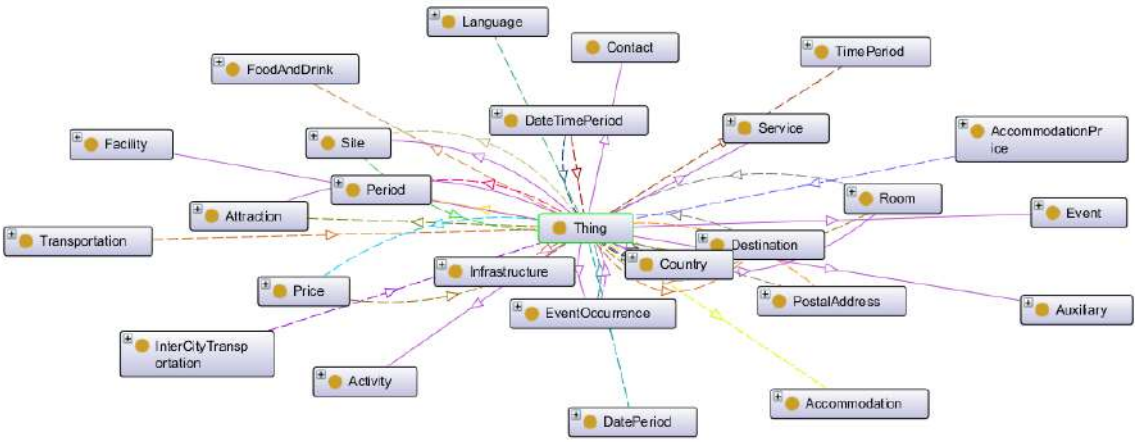


Fig. 4.1 Illustration of OntoGraph for EPT-toursim Ontology

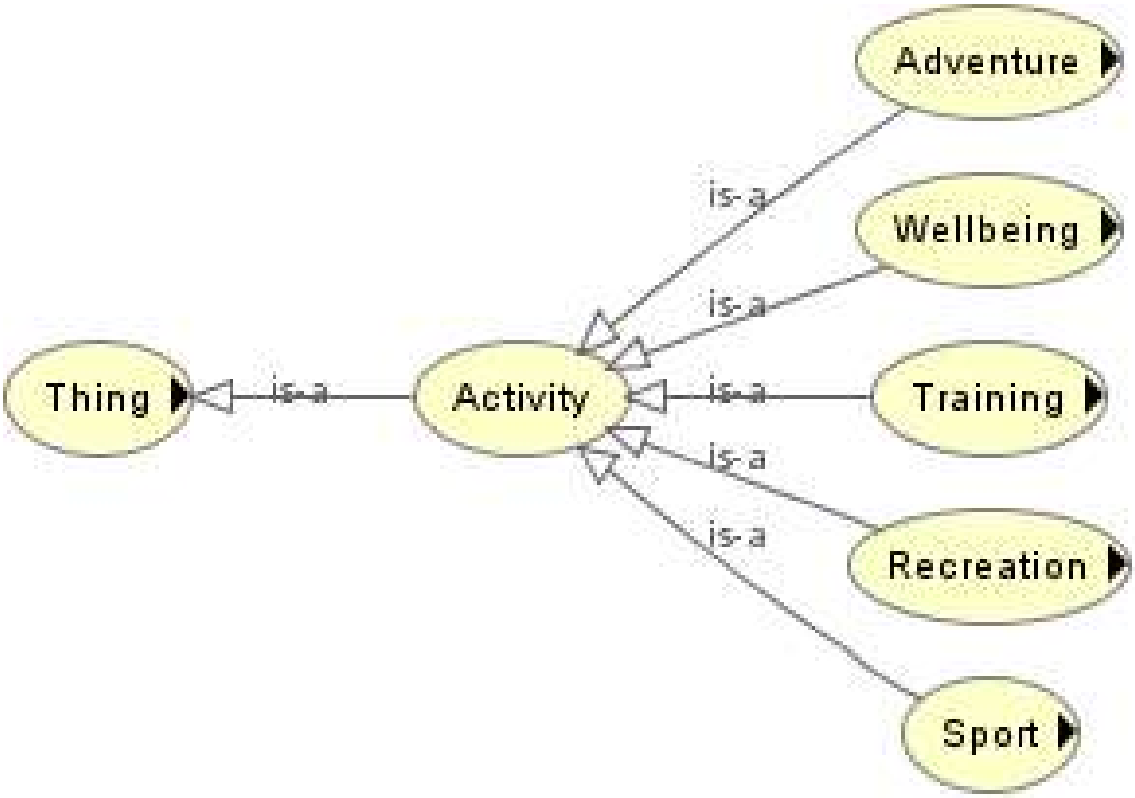


Fig. 4.2 Taxonomical relationship of Activity class with its subclass

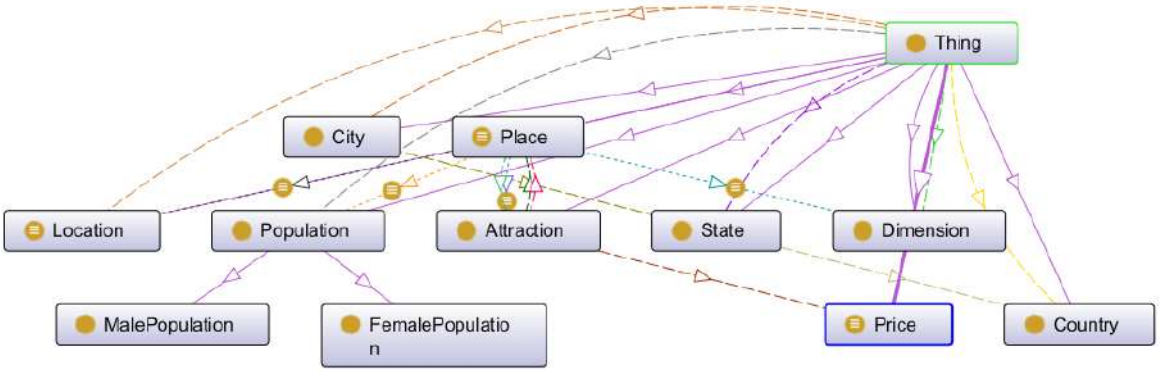


Fig. 4.3 Taxonomical relationship of class with its subclass of Tourism Ontology

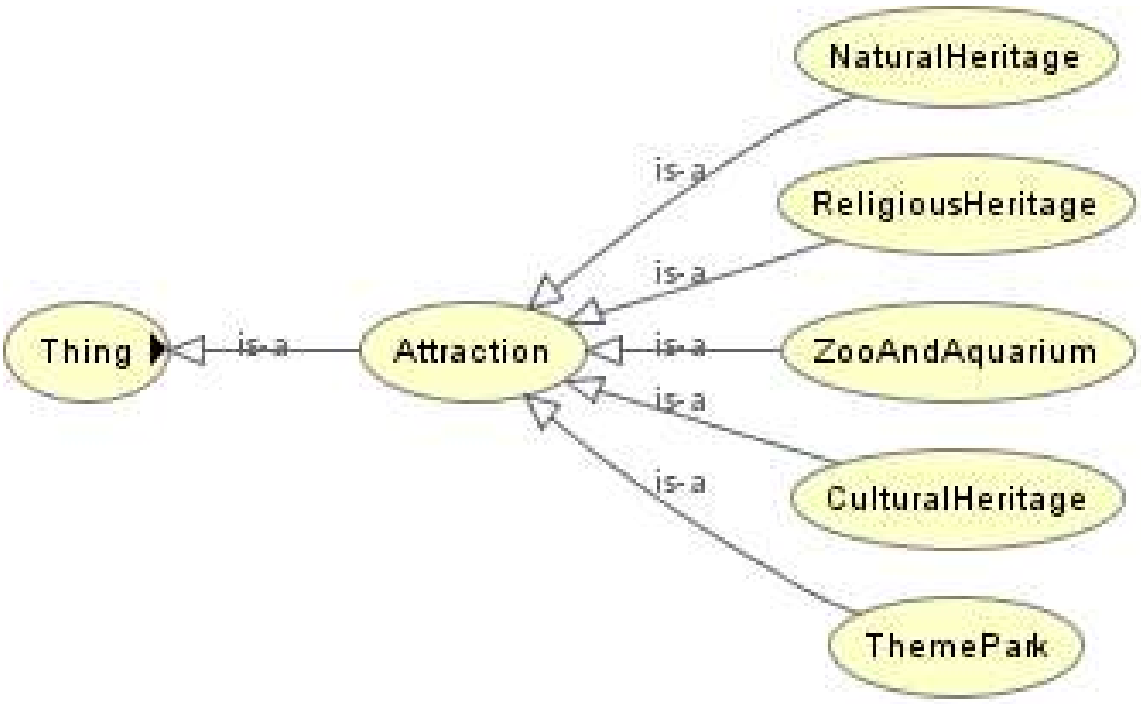


Fig. 4.4 OWLViz of Attraction Class Asserted Model showing is-a relation with subclass

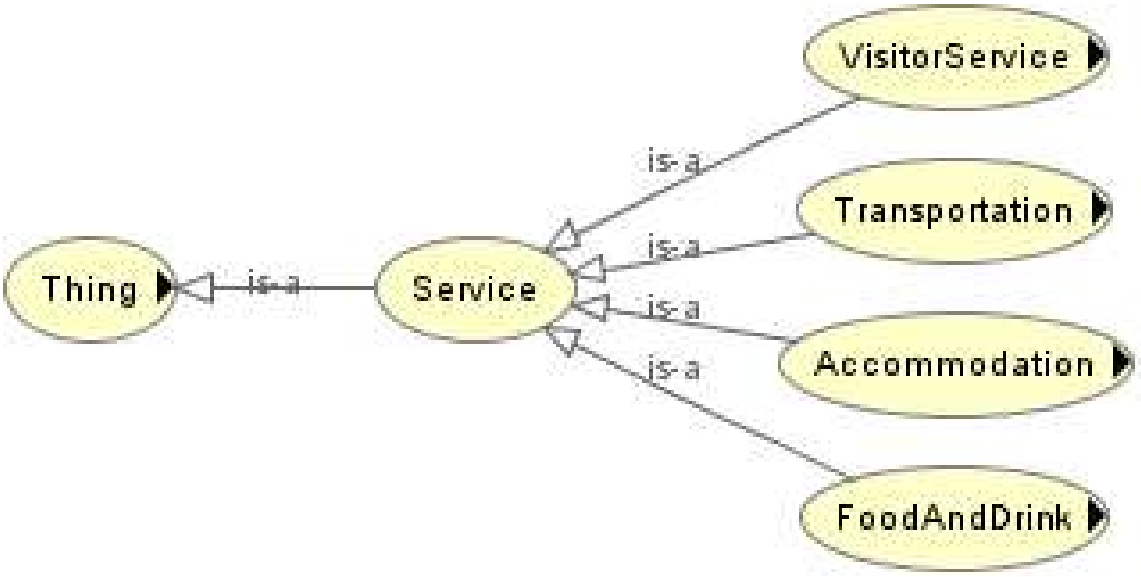


Fig. 4.5 OWLViz of Service Class Asserted Model showing is-a relation with subclass

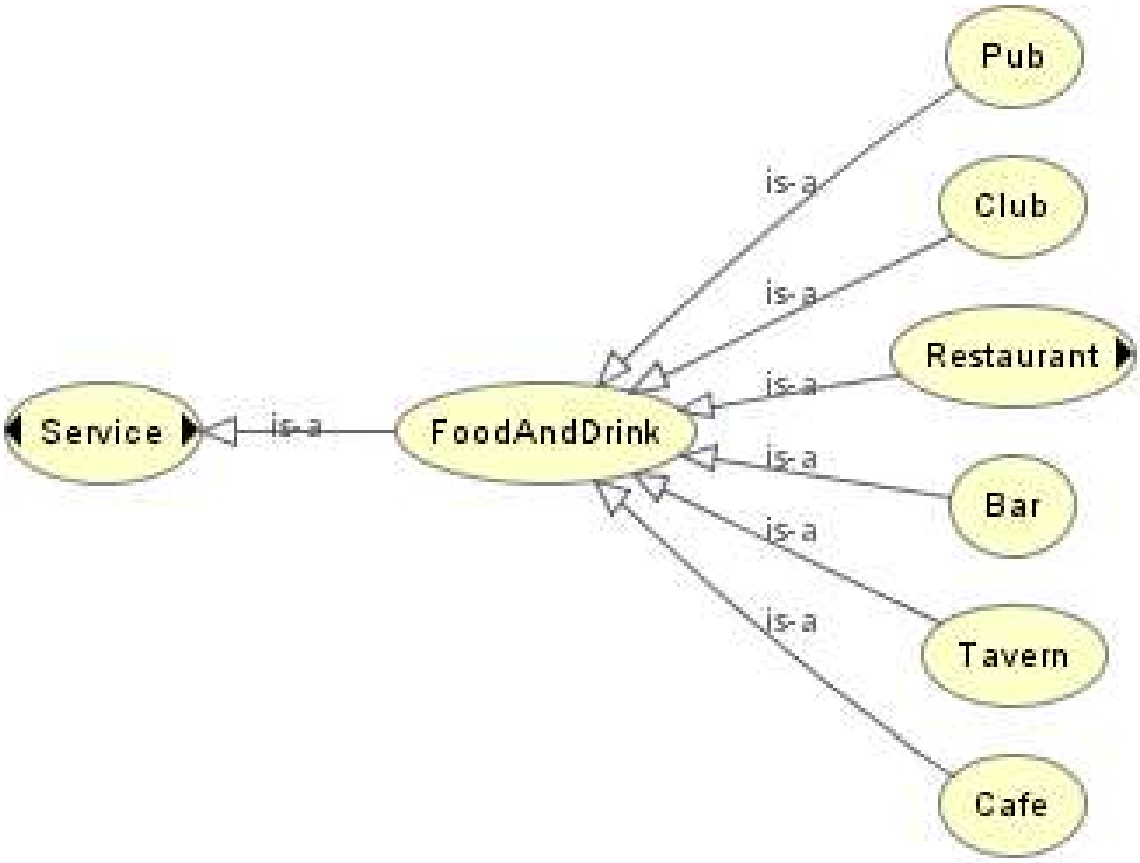


Fig. 4.6 Ontology Visualization of Food and Drink Class Asserted Model showing is-a relation with subclass

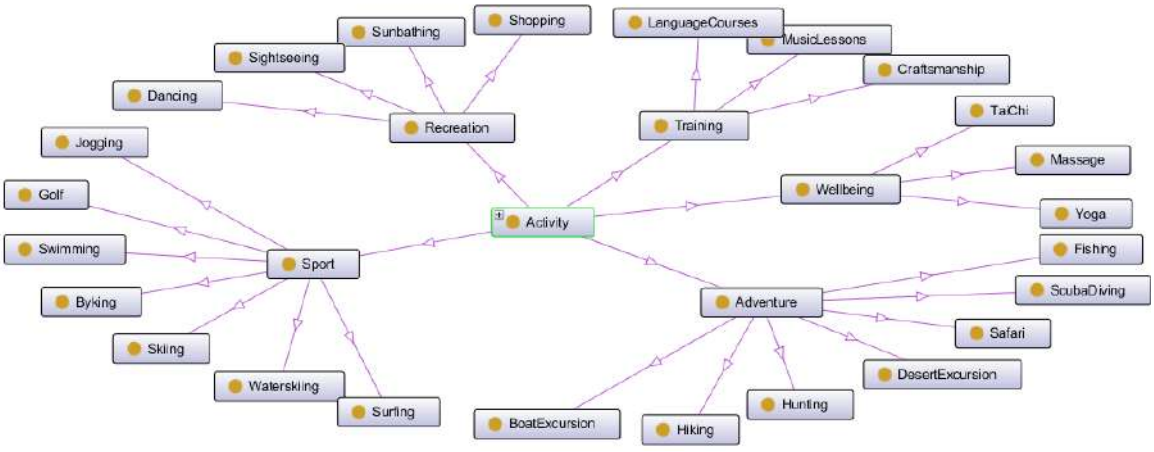


Fig. 4.7 Taxonomical relationship of Activity class with its subclass

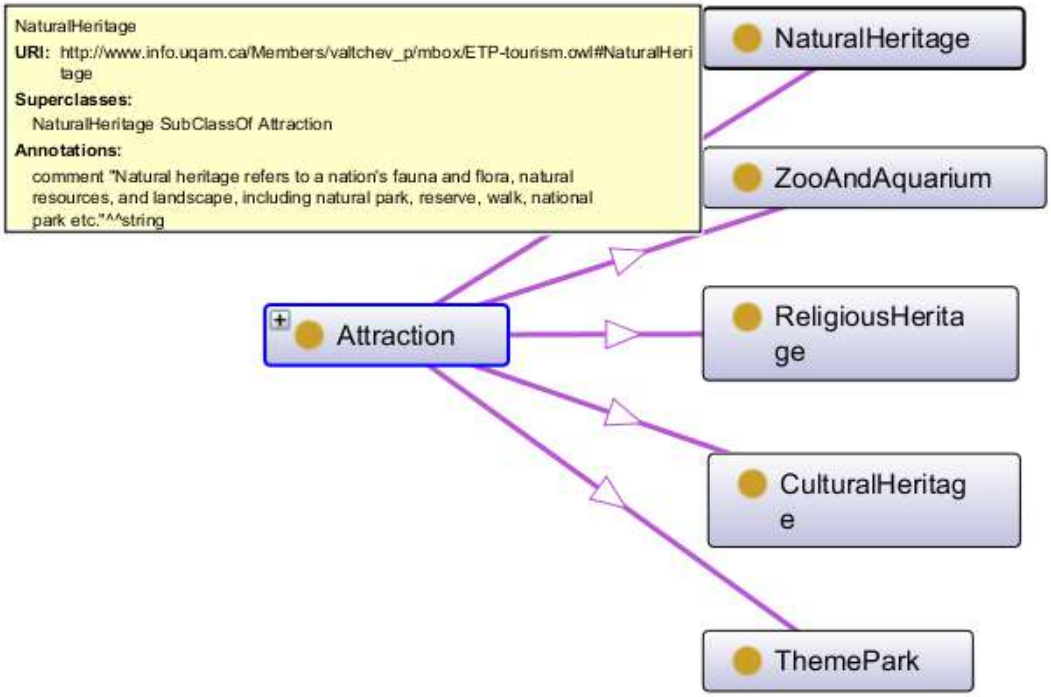


Fig. 4.8 Taxonomical relationship of Attraction class with its subclass

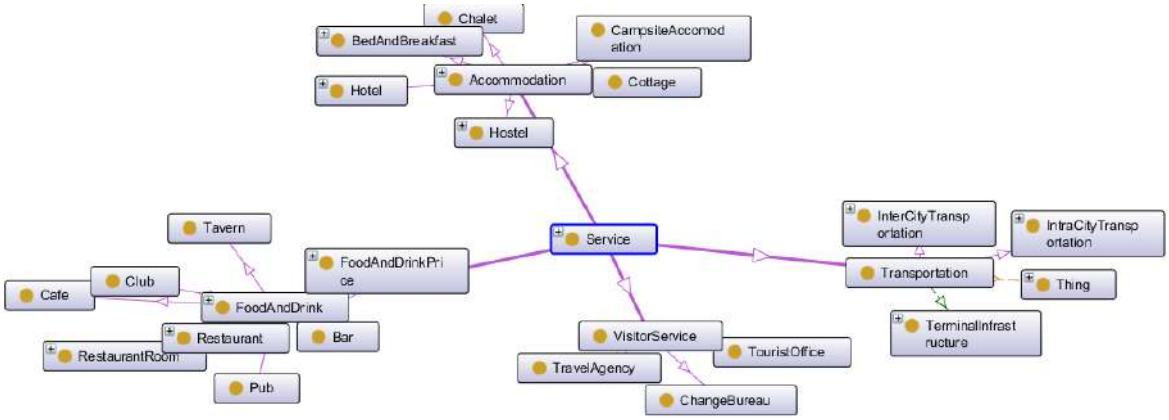


Fig. 4.9 Taxonomical relationship of Service class with its subclass

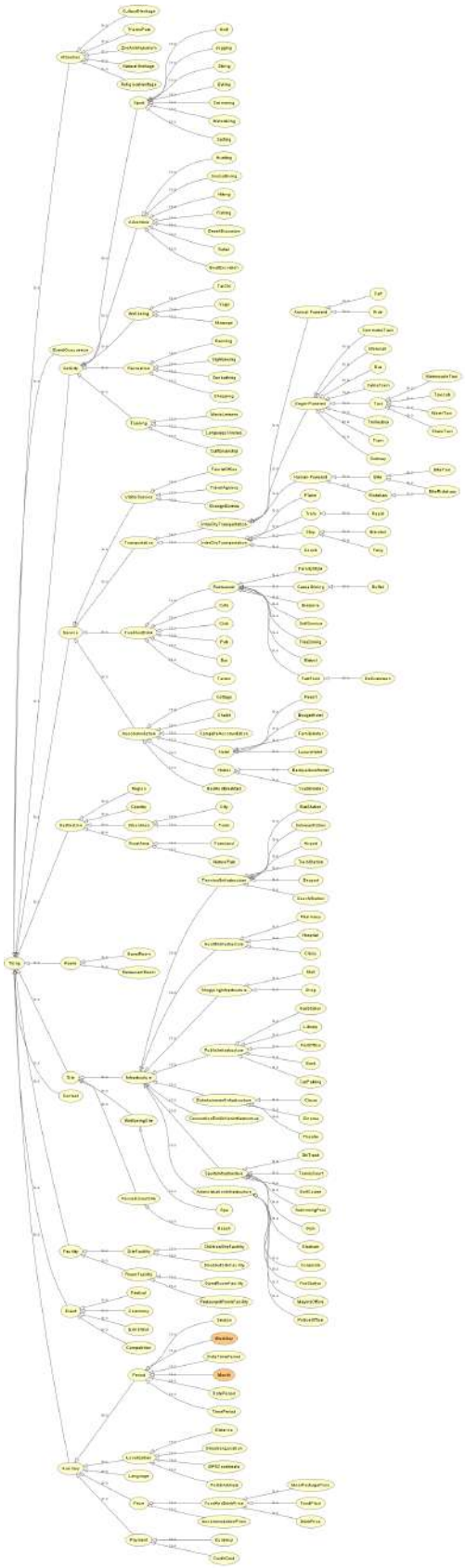


Fig. 4.10 ETP-toursim Ontology Visualization with its class and subclasses

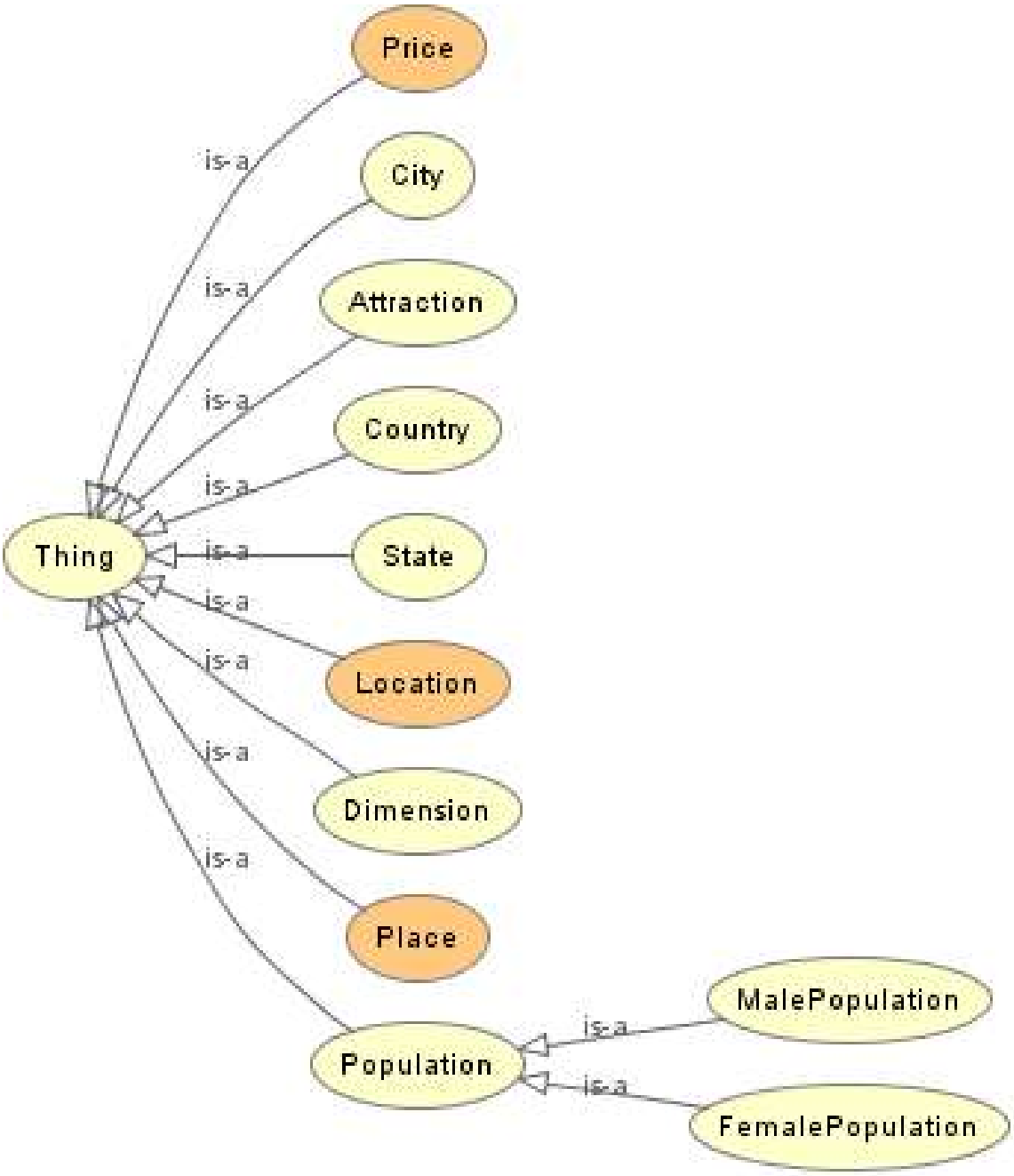


Fig. 4.11 Illustration of OntoGraph for toursim Ontology

Fig. 4.12 Illustration of OntoGraph for travel Ontology

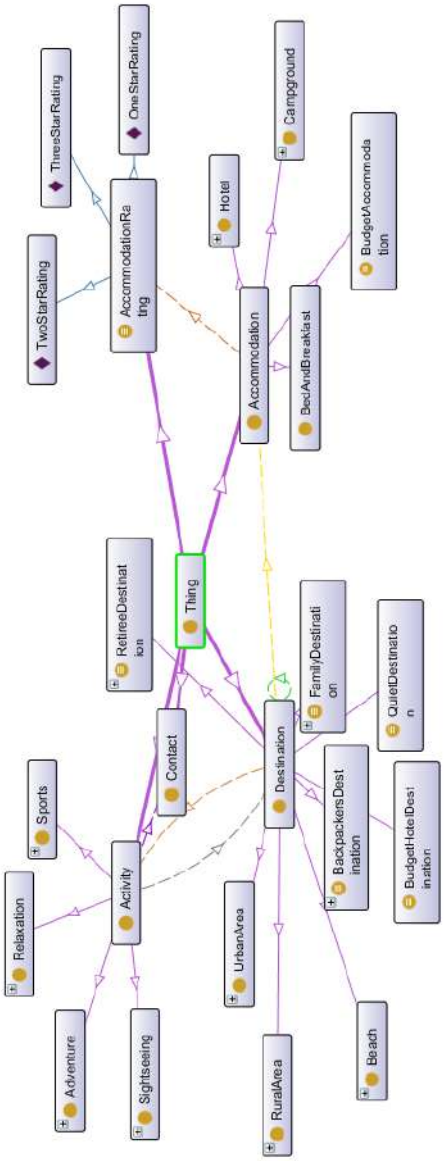


Fig. 4.13 Illustration of OntoGraph for Travel Ontology

4.3 Network Analysis

The data collected through Ontologies allowed a graph of network to be drawn. After that, the main characteristics were deduced with the analysis starting with the categorization of the static properties and then proceeding with the study of its dynamics and evolution. All the values computed: order (number of nodes) and size (number of links) and a random distribution of the links. This procedure allows a better understanding of the significance and of the physical interpretation of the quantities involved.

The calculations were performed by using several software packages widely employed in network analysis, and with the help of programs specifically developed for the purpose of the present study.

- SPSS: for statistical data analysis and charting (SPSS, 2004)

Fig. 4.14 Indegrees, Outdegrees and Betweenness Centralities of the nodes extracted from the EPT-Tourism.owl Ontology

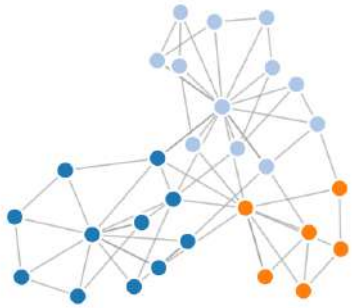


Fig. 4.15 Indegrees, Outdegrees and Betweenness Centralities of the nodes extracted from the Tourism.owl Ontology

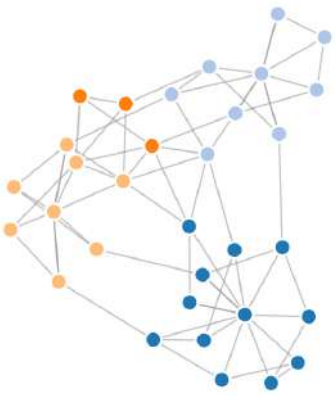


Fig. 4.16 Indegrees, Outdegrees and Betweenness Centralities of the nodes extracted from the Travel.owl Ontology

4.4 Community Detection

We analyze the graph using the Clique Percolation Method (CPM) to detect closely related topics/concepts the ontology. The Clique Percolation Method (CPM) was very effective for grouping together closely related concepts. The communities in ontologies identified using CPM are shown in figure below.

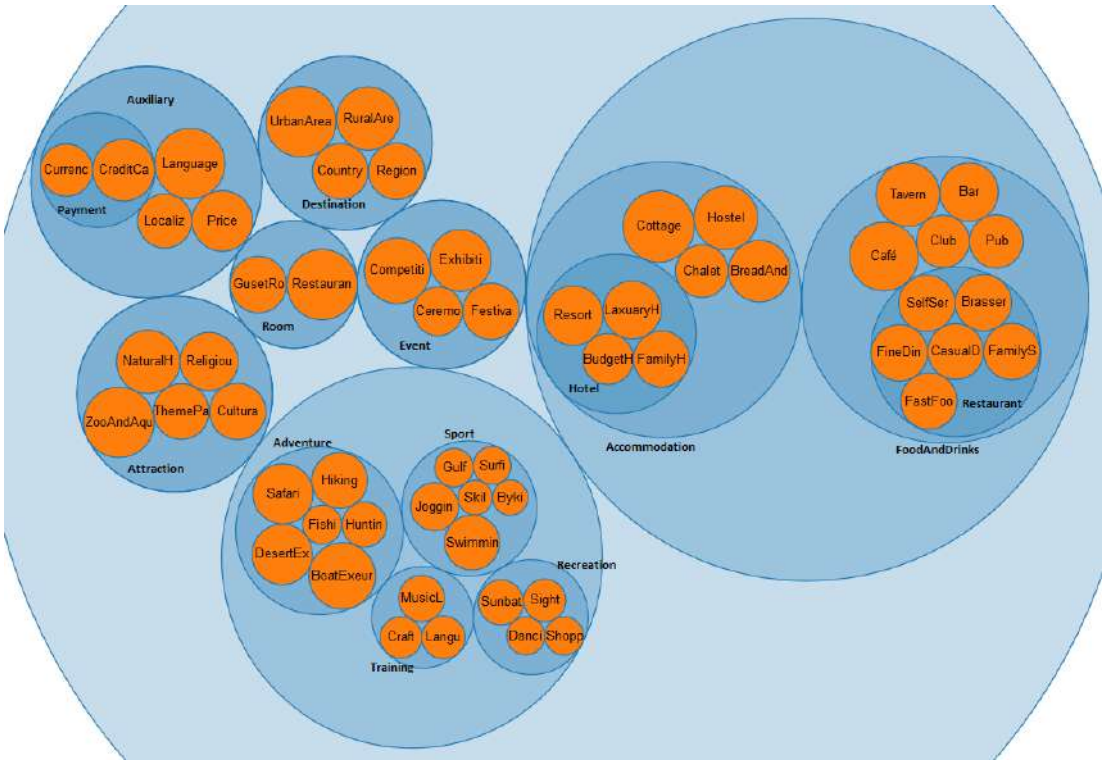


Fig. 4.17 Community Detection from the EPT-Tourism.owl Ontology

Fig. 4.18 Community Detection from the Tourism.owl Ontology

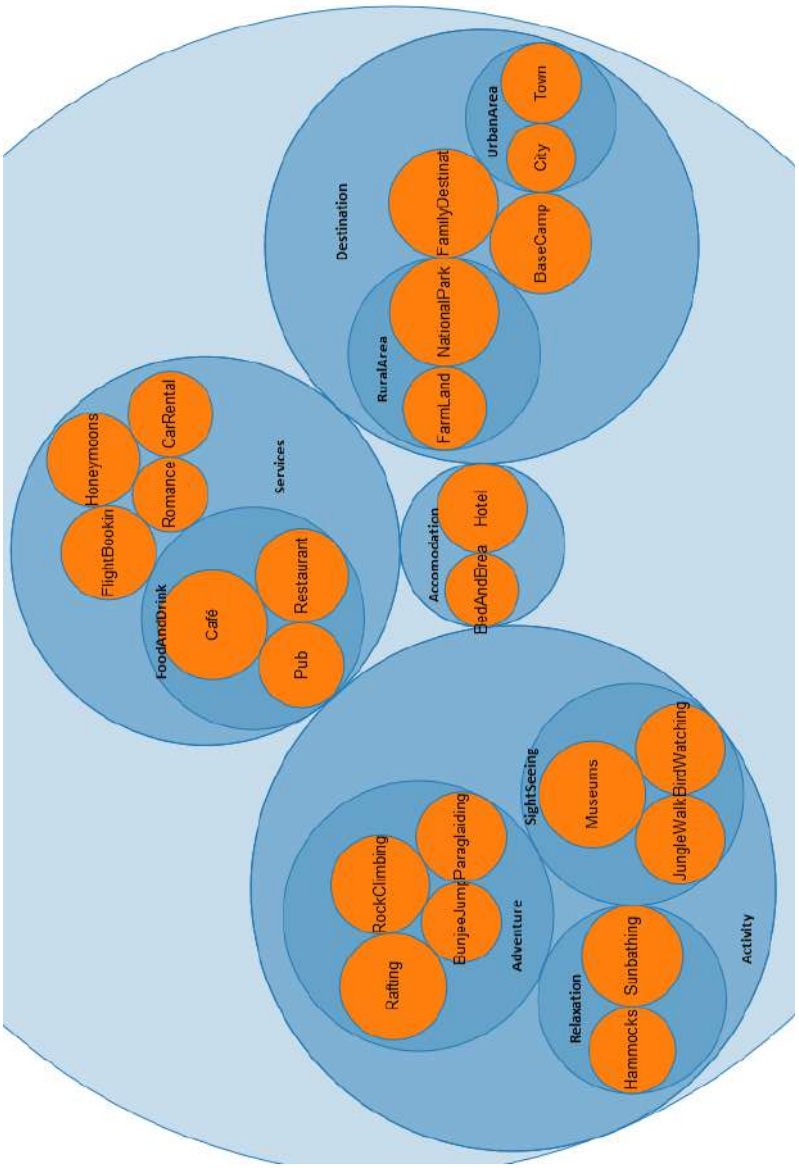


Fig. 4.19 Community Detection from the Travel.owl Ontology

Chapter 5

Conclusion and Future works

5.1 Conclusion

Social Network Analysis provides a promising set of tools for analyzing ontologies and Semantic Web applications, providing deep insights into the structure of ontologies and knowledge bases. In particular, we have seen that the analysis of a given ontology can be done very thoroughly at different levels of granularity. The gained insights may help to design or redesign ontologies in such a way as to find redundancies or holes that should be mended. The analysis is also of use for selecting the right ontology for reuse.

There are multiple ways in which the ontology can be represented as a graph.For the purpose of this analysis,the predicate was represented as a node in the graph representation.Network analysis of the ontologies generated useful results. Since the ontologies have specific focus areas, they have higher average degrees. The basic network properties are useful in comparing ontologies in terms of complexity and level of detail.

Betweenness centrality was useful in identifying the central concept in the ontologies. It was also used to identify the core constructs in RDF, RDFS and OWL. It is interesting to note that, even though the ontologies were developed independently, they follow the power-law and their scaling factor are similar.

The Clique Percolation Method (CPM) was very effective for grouping together closely related concepts.

5.2 Future works

Future analysis could consider different representation of the graph for the ontology and identify the best representation for the analysis. We could also investigate using the declared

vs. the inferred model for the analysis. We could also analyze the strength of relationship between the nodes to perform community detection.

References

[1] Alexander Maedchea, S. S. (2011). Applying semantic web technologies for tourism information systems. *Research Center for Information Technologies at the University of Karlsruhe Research Group Knowledge Management (WIM),Germany*.

[2] Bettina Hoser, Andreas Hotho, C. S. (2006). Semantic network analysis of ontologies. *3rd European Semantic Web Conference*, 1(4).

[3] Bizer, C., H. T. and Berners-Lee (2009). Linked data - the story so far. *International Journal on Semantic Web and Information Systems*, 5:1–22.

[4] Freeman, L. (2009). Centrality in social networks: Conceptual clarfication. *Social Network Elsevier*, pages 215–239.

[5] G. Coskum, M. Rothe, K. T. A. P. (2011). Applying community detection algorithms on ontologies for identifying concept groups. *Modular Ontologies IOS Press*, 17.

[6] Guillaume Ereteo, Freddy Limpens, F. G. O. C. M. B. (2011). Semantic social network analysis: A concrete case. *Handbook of Research on Methods and Techniques for Studying Virtual Communities: Paradigms and Phenomena, IGI Global*, hal-00562056>:122–156.

[7] Guillaume Ereteo, Michel Buffa, F. G. P. G. M. L. P. S. (2008). A state of the art on social network analysis and its applications on a semantic web. *Social Data on the Web*, 405:34–52.

[8] Mika, P. (2007). Ontologies are us: A unified model of social networks and semantics. *Elsevier, The Journal of Web Semantics*, 5:5–11.

[9] Olivier Corby, Leila Kefi-Khelif, H. C. F. G. K. K. (feb. 2009). Querying the semantic web of data using sparql,rdf and xml. *INRIA Rapport de recherche*, pages 1–22.

[10] Tim Finin, Li Ding, L. Z. A. J. (2009). Social networking on the semantic web. *University of Maryland University of Maryland*, pages 418–434.