Knowledge-Based Systems 23 (2010) 844–855

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Grammar-based geodesics in semantic networks

Marko A. Rodriguez [a,*], Jennifer H. Watkins [b]

[a] T-7, Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, NM 87545, United States
[b] International and Applied Technology, Los Alamos National Laboratory, Los Alamos, NM 87545, United States

ARTICLE INFO

Article history:
Received 6 March 2009
Received in revised form 23 May 2010
Accepted 24 May 2010
Available online 8 June 2010

Keywords:
Geodesics
Network analysis
Centrality
Ranking
RDF
Semantic Web

ABSTRACT

A geodesic is the shortest path between two vertices in a connected network. The geodesic is the kernel of various network metrics including radius, diameter, eccentricity, closeness, and betweenness. These metrics are the foundation of much network research and thus, have been studied extensively in the domain of single-relational networks (both in their directed and undirected forms). However, geodesics for single-relational networks do not translate directly to multi-relational, or semantic networks, where vertices are connected to one another by any number of edge labels. Here, a more sophisticated method for calculating a geodesic is necessary. This article presents a technique for calculating geodesics in semantic networks with a focus on semantic networks represented according to the Resource Description Framework (RDF). In this framework, a discrete "walker" utilizes an abstract path description called a grammar to determine which paths to include in its geodesic calculation. The grammar-based model forms a general framework for studying geodesic metrics in semantic networks.

Published by Elsevier B.V.

1. Introduction

The study of networks (i.e. graph theory) is the study of the relationship between vertices (i.e. nodes) as defined by the edges (i.e. arcs) connecting them. In path analysis, a path metric function maps an ordered vertex pair into a real number, where that real number is the length of the path connecting to the two vertices. Metrics that utilize the shortest path between two vertices in their calculation are called geodesic metrics. The geodesic metrics that will be reviewed in this article are shortest path, eccentricity [1], radius, diameter, betweenness centrality [2], and closeness centrality [3].

If G^1 is a single-relational network, then $G^1 = (V,E)$, where $V = \{i,\ldots,j\}$, is the set of vertices and $E \subseteq (V \times V)$ is a subset of the product of V. In a single-relational network all the edges have a single, homogenous meaning. Because an edge in a single-relational network is an element of the product of V, it does not have the ability to represent the type of relationships that exist between the two vertices it connects. An edge can only denote that there is a relationship. Without a distinguishing label, all edges in such networks have a single meaning. Thus, they are called single-relational networks.[1] While a single-relational network supports the representation of a homogeneous set of relationships, a semantic network supports the representation of a heterogeneous set of relationships. For instance, in a single-relational network it is possible to represent humans connected to one another by friendship edges; in a semantic network, it is possible to represent humans connected to one another by friendship, kinship, collaboration, communication, etc. relationships.

A semantic network denoted G^n can be defined as a set of single-relational networks such that $G^n = (V, \mathbb{E})$, where $\mathbb{E} = \{E_0, E_1, \ldots, E_n\}$ and for any $E_k \in \mathbb{E},\ E_k \subseteq (V \times V)$ [5]. The meaning of a relationship in G^n is determined by its set $E_k \in \mathbb{E}$. Perhaps a more convenient semantic network representation and the one to be used throughout the remainder of this article is that of the triple list where $G^n \subseteq (V \times \Omega \times V)$ and Ω is a set of edge labels. A single edge in this representation is denoted by a triple $\tau = \langle i,\omega,j \rangle$, where vertex i is connected to vertex j by the edge label ω.

In some cases, it is possible to isolate sub-networks of a semantic network and represent the isolated network in an unlabeled form. Unlabeled geodesic metrics can be used to compute on the isolated component. However, in many cases, the complexity of the path description does not support an unlabeled representation. These scenarios require "semantically aware" geodesic metrics that respect a semantic network's ontology (i.e. the vertex classes and edge types) [6]. A semantic network is not simply a directed labeled network. It is a high-level representation of complex objects and their relationship to one another according to ontological constraints. There exist various algorithms to study semantically typed paths in a network [7–11]. Such algorithms assume only a path between two vertices and do not investigate other features of the intervening vertices. The benefit of the grammar-based

* Corresponding author.
E-mail address: marko@markorodriguez.com (M.A. Rodriguez).

[1] It is noted that bipartite networks allow for more than one edge meaning to be inferred because V is the union of two disjoint vertex sets. Thus, edges from set $A \subset V$ to set $B \subset V$ (such that $A \cap B = \emptyset$) can have a different meaning than the edges from B to A. Also, theoretically, it is possible to represent edge labels as a topological feature of the graph structure [4]. In other words, there exists an injective function (though not surjective) from the set of semantic networks to the set of single-relational networks that preserves the meaning of the edge labels.

0950-7051/$ - see front matter Published by Elsevier B.V.
doi:10.1016/j.knosys.2010.05.009

geodesic model presented in this article is that complex paths can be represented to make use of path "bookkeeping." Such bookkeeping investigates intervening vertices even though they may not be included in the final path solution. For example, it may be important to determine a set of "friendship" paths between two human vertices, where every intervening human works for a particular organization and has a particular position in that organization. While a set of friendship paths is the result of the function, the path detours to determine employer and position are not. The technique for doing this is the primary contribution of this article.

A secondary contribution is the unification of the grammar-based model proposed here with the grammar-based model proposed in [12] for calculating stationary probability distributions in a subset of the full semantic network (e.g. eigenvector centrality [13] and PageRank [14]). With the grammar-based model, a single framework exists that ports many of the popular single-relational network analysis algorithms to the semantic network domain. Moreover, an algebra for mapping semantic networks to single-relational networks has been presented in [15] and can be used to meaningfully execute standard single-relational network analysis algorithms on distortions of the original semantic network. The Semantic Web community does not often employee the standard suite of network analysis algorithms. This is perhaps due to the fact that the Semantic Web is generally seen as a knowledge-base grounded in description logics rather than graph-or network-theory. When the Semantic Web community adopts a network interpretation, it can benefit from the extensive body of work found in the network analysis literature. For example, recommendation [16], ranking [17], and decision making [6] are a few of the types of Semantic Web applications that can benefit from a network perspective. In other words, graph/network theoretic techniques can be used to yield innovative solutions on the Semantic Web.

The first half of this article will define a popular set of geodesic metrics for single-relational networks. It will become apparent from these definitions, that the more advanced geodesics rely on the shortest path metric. The second half of the article will present the grammar-based model for calculating a meaningful shortest path in a semantic network. The other geodesics follow from this definition.

2. Geodesics in single-relational networks

This section will review a collection of popular geodesic metrics used to characterize a path, a vertex, and a network. The following list enumerates these metrics and identifies whether they are path, vertex, or network metrics:

- in- and out-degree: vertex metric,
- shortest path: path metric,
- eccentricity: vertex metric,
- radius: network metric,
- diameter: network metric,
- closeness: vertex metric,
- betweenness: vertex metric.

It is worth noting that besides in- and out-degree, all the metrics mentioned utilize a path function $\rho: V \times V \to Q$ to determine the set of paths between any two vertices in V, where Q is a set of paths. The premise of this article is that once a path function is defined for a semantic network, then all of the other metrics are directly derived from it. In the semantic network path function, $\rho: V \times V \times \Psi \to Q$ returns the number of paths between two vertices according to a user-defined grammar Ψ.

Before discussing the grammar-based geodesic model for semantic networks, this section will review the geodesic metrics in the domain of single-relational networks.

2.1. In- and out-degree

The simplest structural metric for a vertex is the vertex's degree. While this is not a geodesic metric, it is presented as the concept will become necessary in the later section regarding semantic networks.

For directed networks, any vertex $i \in V$ has both an in-degree and an out-degree. The set of edges in E that have i as either its in- or out-edge is denoted $\Gamma^-: V \to E$ and $\Gamma^+: V \to E$, respectively. If

$$\Gamma^-(i) = \{(x,y) | (x,y) \in E \wedge y = i\}$$

and

$$\Gamma^+(i) = \{(x,y) | (x,y) \in E \wedge x = i\}$$

then, $\Gamma^-(i)$ is the subset of edges in E incoming to i and $\Gamma^+(i)$ is the subset of edges outgoing from i. The cardinality of the sets is the in- and out-degree of the vertex, denoted $|\Gamma^-(i)|$ and $|\Gamma^+(i)|$, respectively.

2.2. Shortest path

The shortest path metric is the foundation for all other geodesic metrics. This metric is defined for any two vertices $i,j \in V$ such that the sink vertex j is reachable from the source vertex i in G^1 [18]. If j is unreachable from i, the shortest path between i and j is undefined. The shortest path between any two vertices i and j in an unweighted network is the smallest of the set of all paths between i and j. If $\rho: V \times V \to Q$ is a function that takes two vertices and returns a set of paths Q where for any $q \in Q$, $q = (i, \ldots, j)$, then the shortest path between i and j is the $min(\bigcup_{q \in Q} |q| - 1)$, where min returns the smallest value of its domain. The shortest path function is denoted $s : V \times V \to \mathbb{N}$ with the function rule

$$s(i,j) = min\left(\bigcup_{q \in \rho(i,j)} |q| - 1\right).$$

It is important to subtract 1 from the path length since a path is defined as the set of edges traversed, not the set of vertices traversed. Thus, for the path $q = (a,b,c,d)$, the $|q|$ is 4, but the path length is 3.

Note that ρ returns the set of all paths between i and j. Of course, with the potential for loops, this function could return a $|Q| = \infty$. Therefore, in many cases, it is important to not consider all paths, but just those paths that have the same cardinality as the shortest path currently found and thus are shortest paths themselves. It is noted that all the remaining geodesic metrics require only the shortest path between i and j.

2.3. Eccentricity, radius, and diameter

The radius and diameter of a network require the determination of the eccentricity of every vertex in V. The eccentricity metric requires the calculation of $|V| - 1$ shortest path calculations of a particular vertex [1]. The eccentricity of a vertex i is the largest shortest path between i and all other vertices in V such that the eccentricity function $e : V \to \mathbb{N}$ has the rule

$$e(i) = max\left(\bigcup_{j \in V} s(i,j)\right),$$

where max returns the largest value of its domain.

The radius of the network is the minimum eccentricity of all vertices in V [19]. The function $r : G \to \mathbb{N}$ has the rule

$$r(G^1) = min\left(\bigcup_{i \in V} e(i)\right).$$

Finally, the diameter of a network is the maximum eccentricity of the vertices in V [19]. The function $d : G \to \mathbb{N}$ has the rule