

Abstract

Relational or semi-structured data is naturally represented in a graph schema, where nodes denote entities and directed typed edges represent the relations between them. Such graphs are heterogeneous, in the sense that they include different types of nodes and different types of links. For example, we represent personal information as a graph, in which email messages, meeting entries, persons, text and a timeline are inter-connected via relations derived from textual and structural information residing in a personal workstation.

A question of interest is how to determine the nature of relationship between entities in the graph. We suggest a framework, where random graph walks (e.g., Personalized PageRank) are applied to derive an extended measure of entity similarity. In previous works, special graphs have been designed in several domains to optimize performance for pre-defined tasks. In this thesis, we make the following arguments: (a) structured and semi-structured data can be naturally represented as a graph; (b) Given a general graph, multiple tasks can be processed in terms of entity similarity using the same underlying graph. We claim that graph walks can provide good performance for arbitrary queries in this framework, and that the graph walk derived similarity measure can be further adapted per task with learning.

In the thesis, we consider several learning approaches. First, we evaluate a method that tunes the set of graph weights defined per edge type in the graph, such that the probability flow in the graph is directed towards relevant nodes; we also propose re-ranking as an alternative and complementary learning method, using features that capture "global" properties of the graph walk. Finally, we suggest a path constrained graph walk variant, in which the graph walk process is guided by high-level knowledge about meaningful edge sequences; that is, we allow the probability flow in the graph to be conditioned on the history of the walk.

The framework is evaluated for two different domains and a variety of tasks. In the personal information management domain, we show how seemingly different tasks like person name disambiguation, threading and alias finding are addressed uniformly using the adaptive graph-walk based similarity measure. The second domain evaluated is parsed text, where we represent a corpus of parsed text as a graph, and induce inter-word similarity measures for tasks like coordinate term extraction. Finally, based on our empirical results, we discuss general design and scalability considerations in applying the framework.