Learning Similarity Measures Based on Random Walks

[Extended Abstract]

William W. Cohen
Machine Learning Department
Carnegie Mellon University
Pittsburgh, Pennsylvania
wcohen@cs.cmu.edu

ABSTRACT

We describe a novel learnable proximity measure based on personalized PageRank (also known as "random walk with reset"). Instead of introducing one weight per edge label, as in most prior work, we introduce one weight for each edge label *sequence*. We show that this approach is advantageous for a number of real-world tasks, including querying graph databases, recommendation tasks, and inference in large, noisy knowledge bases.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

personalized PageRank, learning

1. OVERVIEW

The scientific literature can be represented as a graph of documents, terms, and meta-data, with edges corresponding to containment of a term in a document, authorship of a document by a person, and so on. One popular way of querying such a graph is via queries based on proximity measures, such as personalized PageRank, also known as Random Walk with Restart (RWR).

We describe a novel learnable proximity measure based on RWR. Instead of introducing one weight per edge label, as in most prior work, we introduce one weight for each edge label sequence. In this model proximity is defined by a weighted combination of simple ”random walk experts”, each corresponding to conducting a random walk constrained to follow a particular sequence of labeled edges.

Experiments on several tasks using graphs based on literature from two subdomains of biology show that the new learning method significantly outperforms the prior methods. We extend the method to support two additional types of experts to model intrinsic properties of entities: ”query-independent experts”, which generalize the PageRank measure, and ”popular entity experts” which allow rankings to be adjusted for particular entities that are especially important.

We also present experiments in which we use this approach to learn relationships in the ontology of NELL, a wide-coverage, large-scale information extraction system for web data. We show that these types of learnable ”proximity measures” are general enough to accurately model a significant number of real-world relations, and that they outperform an alternative technique that learns to model relations based on more traditional logical rules.

Acknowledgements

This is joint work with Ni Lao (now at Google Research) and Tom Mitchell. Parts of this work were supported by NIH under grant R01GM081293, by NSF under grant IIS0811562, by DARPA under awards FA8750-08-1-0009 and AF8750-09-C-0179, and by a gift from Google.

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.
ACM 978-1-4503-1156-4/12/10.