Scalable Relational Learning for Large Heterogeneous Networks

Ryan A. Rossi and Rong Zhou
Palo Alto Research Center
{rrossi, rzhou}@parc.com

Abstract

Relational models for heterogeneous network data are becoming increasingly important for many real-world applications. However, existing relational learning approaches are not parallel, have scalability issues, and thus unable to handle large heterogeneous network data. In this paper, we propose Parallel Collective Matrix Factorization (PCMF) that serves as a fast and flexible framework for joint modeling of large heterogeneous networks. The PCMF learning algorithm solves for a single parameter given the others, leading to a *parallel scheme* that is fast, flexible, and general for a variety of relational learning tasks and heterogeneous data types. The proposed approach is carefully designed to be (a) efficient for large heterogeneous networks (linear in the total number of observations from the set of input matrices), (b) flexible as many components are interchangeable and easily adaptable, and (c) effective for a variety of applications as well as for different types of data. The experiments demonstrate the scalability, flexibility, and effectiveness of PCMF. For instance, we show that PCMF outperforms a recent state-of-the-art parallel approach in runtime, scalability, and prediction quality. Finally, the effectiveness of PCMF is shown on a number of relational learning tasks such as serving predictions in a real-time streaming fashion.

TABLE I. DATASET STATISTICS

graph datasets	**(semantics) node & edge types**	m	n	$\lvert\Omega\rvert$	max d_{out}	max d_{in}	$\bar{d}_{\text{out}}$	$\bar{d}_{\text{in}}$	weight range	μ_{out}	μ_{in}	ρ	$\lvert\Omega^{\text{test}}\rvert$
amazon	*users*-rate-*items*	2.1M	1.2M	4.7M	9.8k	2.5k	2.18	3.79	1-5	3.57	3.77	10^{-6}	1.2M
dating	*users*-rate-*users*	135k	169k	16M	24k	32k	121.83	97.7	1-10	5.93	6.03	<0.01	868k
eachmovie	*users*-rate-*movies*	1.6k	74k	2.2M	26k	1.2k	1.4k	30.22	1-6	3.99	3.45	0.02	562k
epinions	*users*-rate-*items*	120k	756k	13M	159k	1.2k	111.18	17.73	1-5	4.64	4.31	<0.01	137k
	users-trust-*users*	120k	120k	607k	1.8k	3.2k	5.03	5.03	1	0.47	0.35	10^{-5}	
flickr	*users*-friend-*users*	1.9M	1.9M	18M	21k	13k	9.72	9.72	1	0.68	0.8	10^{-6}	
	users-join-*group*	1.7M	104k	8.5M	2.2k	35k	4.94	82.45	1	0.23	1	10^{-5}	4.5M
lastfm	*users*-listento-*songs*	992	1.1M	15M	146k	14k	15k	14.13	1	1	0.93	0.01	
	users-listento-*band*	992	174k	19M	183k	115k	19k	110.01	1	1	1	0.11	3.8M
livejournal	*users*-friend-*users*	5.2M	5.2M	39M	12k	928	7.56	7.56	1	0.55	0.95	10^{-6}	
	users-join-*groups*	3.2M	7.5M	112M	300	1.1M	35.08	15	1	1	1	10^{-6}	9.8M
movielens10M	*users*-rate-*movies*	72k	65k	9.3M	7.3k	29k	129.97	142.8	0.5-5	3.42	0.52	<0.01	699k
stackoverflow	*users*-favorite-*posts*	545k	97k	1.0M	4.0k	4.9k	1.91	10.78	1	0.86	0.92	10^{-5}	260k
yelp	*users*-rate-*businesses*	230k	12k	225k	575	824	0.98	19.53	1-5	0.73	3.66	10^{-5}	4.6k
youtube	*users*-friend-*users*	1.2M	1.2M	4.0M	23k	20k	3.42	3.42	1	0.46	0.86	10^{-6}	
	users-join-*groups*	664k	30k	293k	1.0k	7.6k	0.44	9.75	1	0.14	1	10^{-5}	989k