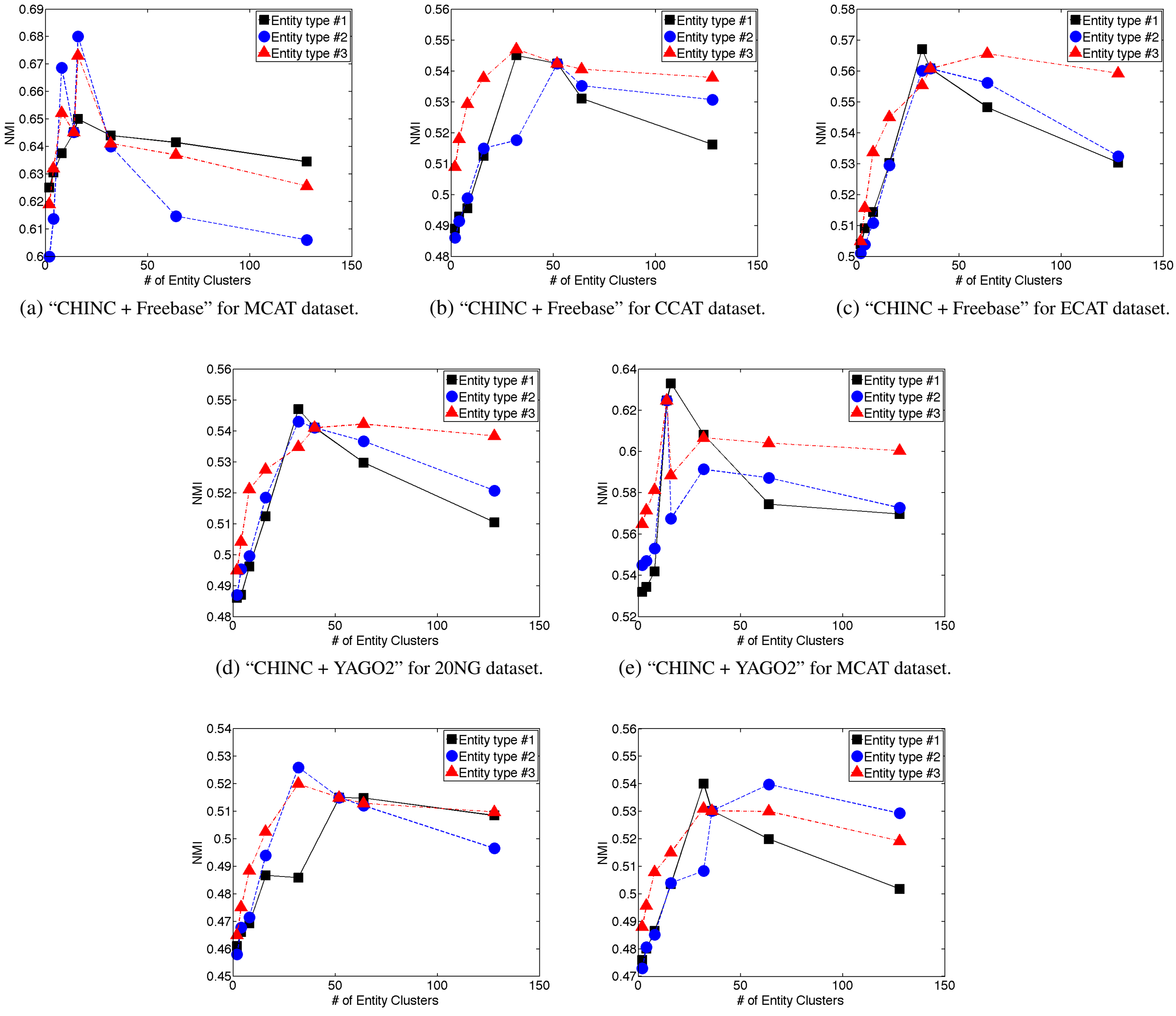
Supplementary Material for “Incorporating World Knowledge to Document Clustering via Heterogeneous Information Networks”

Chenguang Wang[†] , Yangqiu Song[‡] , Ahmed El-Kishky[‡] , Dan Roth[‡] , Ming Zhang[†] , Jiawei Han[‡]
[†]School of EECS, Peking University
[‡]Department of Computer Science, University of Illinois at Urbana-Champaign
wangchenguang@pku.edu.cn, {yqsong, elkishk2, danr, hanj}@illinois.edu, mzhang@net.pku.edu.cn

1. ADDITIONAL EXPERIMENTS



(a) “CHINC + Freebase” for MCAT dataset.

(b) “CHINC + Freebase” for CCAT dataset.

(c) “CHINC + Freebase” for ECAT dataset.

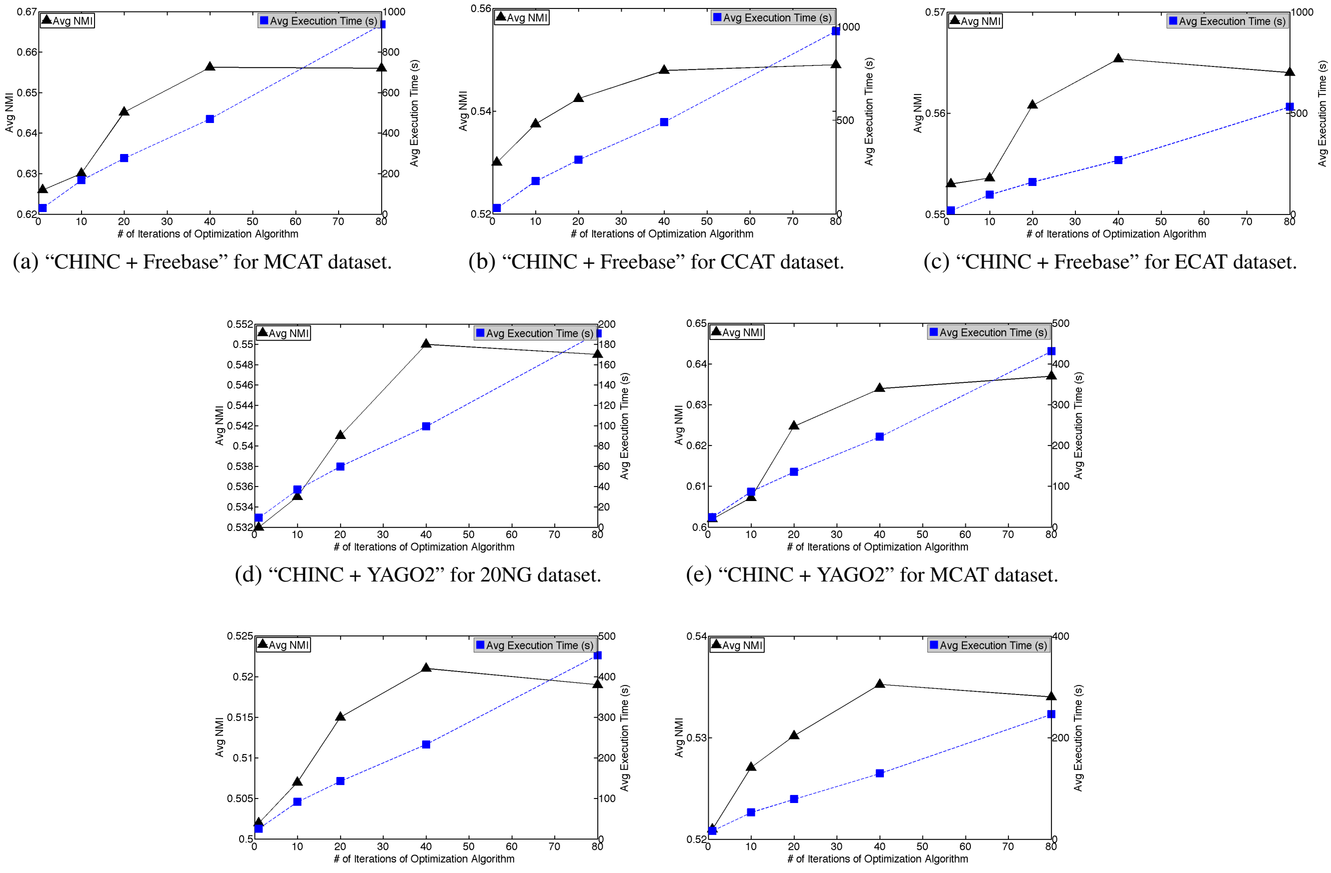
(d) “CHINC + YAGO2” for 20NG dataset.

(e) “CHINC + YAGO2” for MCAT dataset.

(f) “CHINC + YAGO2” for CCAT dataset.

(g) “CHINC + YAGO2” for ECAT dataset.

Figure 1: Effect of number of entity clusters of each entity type on document clustering on different dataset and world knowledge source combinations.



(a) “CHINC + Freebase” for MCAT dataset.

(b) “CHINC + Freebase” for CCAT dataset.

(c) “CHINC + Freebase” for ECAT dataset.

(d) “CHINC + YAGO2” for 20NG dataset.

(e) “CHINC + YAGO2” for MCAT dataset.

(f) “CHINC + YAGO2” for CCAT dataset.

(g) “CHINC + YAGO2” for ECAT dataset.

Figure 2: Analysis of # of iterations in alternating optimization algorithm on different dataset and world knowledge source combinations. Left y-axis: average NMI; Right y-axis: average execution time (s).