

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[‡]

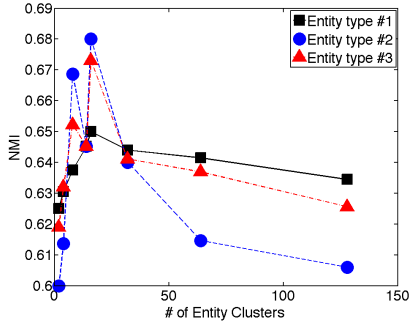
[†]School of EECS, Peking University

[‡]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

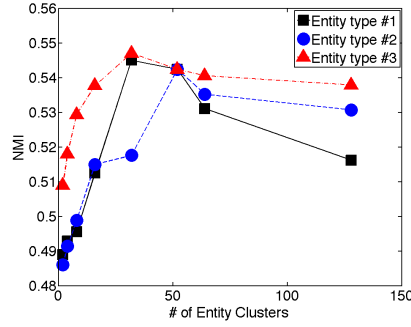
1. ADDITIONAL EXPERIMENTS

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

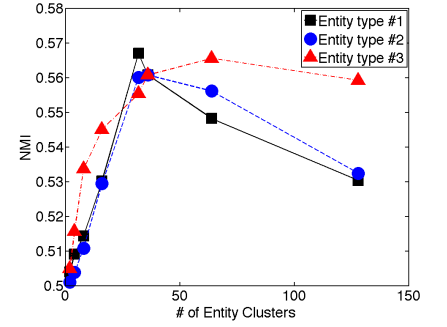
Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...\$15.00.



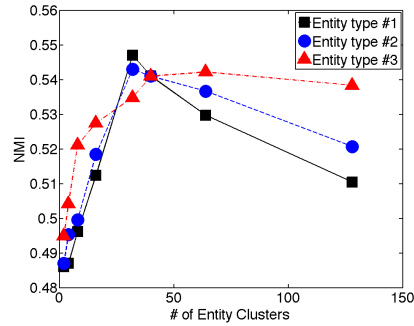
(a) "CHINC + Freebase" for MCAT dataset.



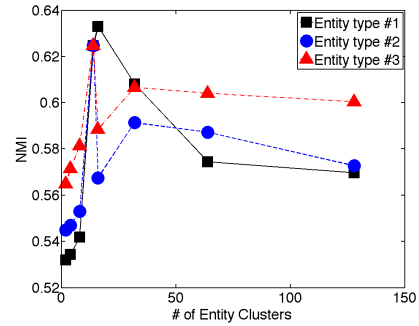
(b) "CHINC + Freebase" for CCAT dataset.



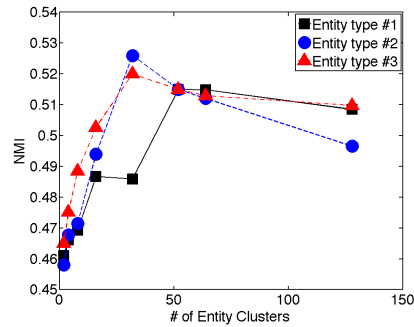
(c) "CHINC + Freebase" for ECAT dataset.



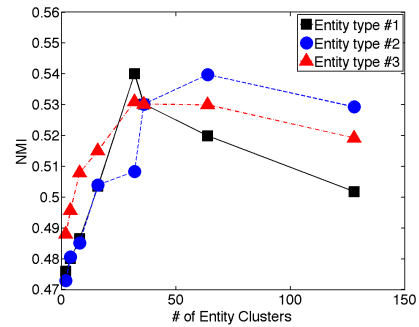
(d) "CHINC + YAGO2" for 20NG dataset.



(e) "CHINC + YAGO2" for MCAT dataset.

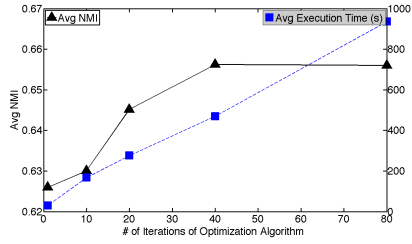


(f) "CHINC + YAGO2" for CCAT dataset.

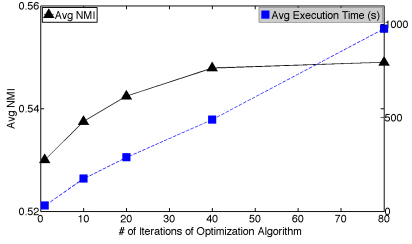


(g) "CHINC + YAGO2" for ECAT dataset.

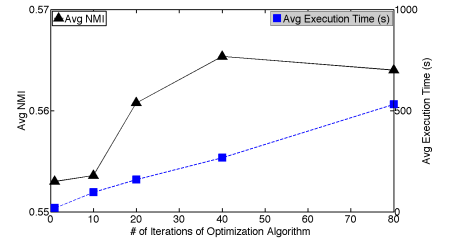
Figure 1: Effect of number of entity clusters of each entity type on document clustering on different dataset and world knowledge source combinations.



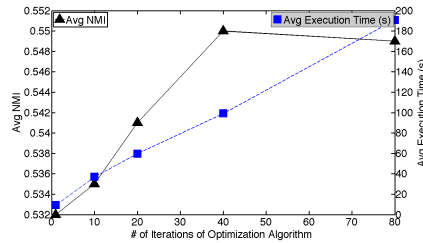
(a) “CHINC + Freebase” for MCAT dataset.



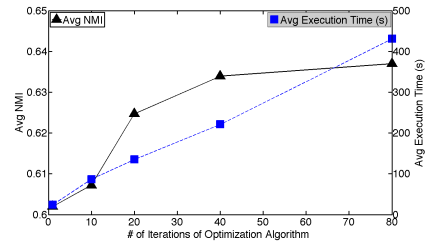
(b) “CHINC + Freebase” for CCAT dataset.



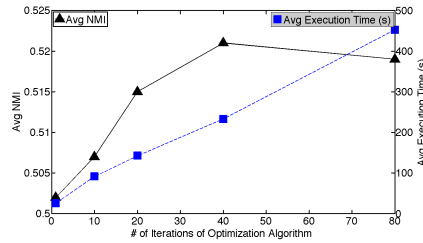
(c) “CHINC + Freebase” for ECAT dataset.



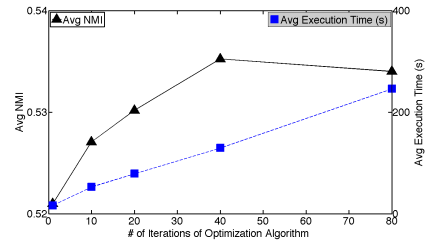
(d) “CHINC + YAGO2” for 20NG dataset.



(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).