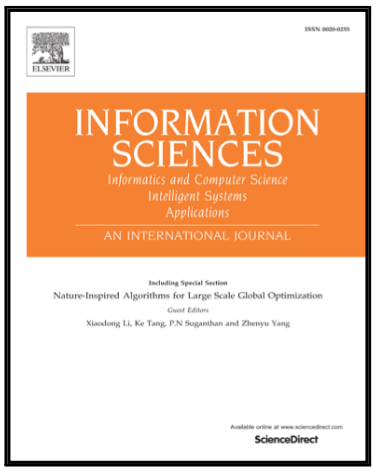

A Novel Approach for Entity Resolution in Scientific Documents Using Context Graphs

Changqin Huang, Jia Zhu, Xiaodi Huang, Min Yang, Gabriel Fung, Qintai Hu

PII: S0020-0255(17)31147-7
DOI: 10.1016/j.ins.2017.12.024
Reference: INS 13318

To appear in: *Information Sciences*

Received date: 1 November 2017
Revised date: 15 December 2017
Accepted date: 20 December 2017

Please cite this article as: Changqin Huang, Jia Zhu, Xiaodi Huang, Min Yang, Gabriel Fung, Qintai Hu, A Novel Approach for Entity Resolution in Scientific Documents Using Context Graphs, *Information Sciences* (2017), doi: 10.1016/j.ins.2017.12.024

A Novel Approach for Entity Resolution in Scientific Documents Using Context Graphs

Changqin Huang[a,b,*], Jia Zhu[b], Xiaodi Huang[c], Min Yang[d], Gabriel Fung[e], Qintai Hu[a,b]

[a] *School of Information Technology in Education, South China Normal University, Guangzhou, China*
[b] *Guangdong Engineering Research Center for Smart Learning, South China Normal University, Guangzhou, China*
[c] *School of Computing and Mathematics, Charles Sturt University, Albury, Australia*
[d] *Department of Computer Science, The University of Hong Kong, Hong Kong, China*
[e] *Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong, China*

Abstract

Entity resolution refers to disambiguating and resolving entities in structured and unstructured data. Developments of effective resolution algorithms are significant for processing scientific documents, particularly for biomedical literature. Specifically, name ambiguity among biomedical entities is a primary task that needs to be solved in the knowledge extraction process. In this paper, we present a novel approach to disambiguating gene/protein names by using context graphs. A set of abstracts of documents is used to build the context graphs through disclosing the indirect co-occurrence relationships among words. Feature vectors of the graphs can be constructed according to information gain (IG) on the word set. To evaluate the IG values, we propose a new metrics that integrates the word frequency (WF), dispersion degree (DD) and concentration degree (CD). Finally, entity resolution is performed by applying a support vector machine (SVM). Compared to existing approaches, the proposed method is capable of discovering latent information from the context of entity names, rather than using some statistical information such as the number of occurrences of words. Based on the results from comprehensive experiments over two benchmark datasets, we conclude that our proposed method, compared to several existing solutions, for resolving

*Corresponding author
Email address: cqhuang@scnu.edu.cn (Changqin Huang)