DOI 10.18699/MM-HPC-BBB-2018-14

Revealing the research institutes and their interactions: a case study of miRNA research

A. Firsov[1]*, I. Titov[2]
[1] *Novosibirsk State University, Novosibirsk, Russia*
[2] *Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia*
* *e-mail: artemijfirsov@mail.ru*

Key words: affiliation disambiguation, institution network, KOFER, K-Mer, miRNA

Motivation and Aim: A lot of digital libraries appeared with the growth of the Internet, thus, format of representation of many scientific articles changed. That way, we got a possibility to query articles metadata, gather some statistics, etc. This includes understanding the institutions' activity, their interactions, and other characteristics. However, to do that, one should identify affiliation in order to know in which articles the true underlying organization is mentioned. Issue of affiliation disambiguation is complex if you consider the dataset consisting of 2×10^7 articles, such as PubMed database. It becomes more complicated when you consider errors in affiliation made either by the author, or the editor. Moreover, sometimes institution name might be changed, or the affiliation from the papers metadata may have mixed institution names for different authors. E. g. if Author1 has "Institute of Cytology and Genetics, Novosibirsk, Russia" institution and Author2 has "Institute of Mathematics, Novosibirsk, Russia" institution, their resulting affiliation for paper might be "Institute of Cytology and Genetics, Institute of Mathematics, Novosibirsk, Russia". Moreover, affiliation can contain email, postal address and other artifacts.
Methods and Algorithms: In this work, we propose the method of the affiliation disambiguation based only on affiliations from papers metadata. The solution consists of 2 stages: preprocessing stage and clustering stage. At the preprocessing stage normalization and splitting of affiliation is performed. At the clustering stage the *DBSCAN* clustering is performed upon K-Mer features extracted from separated affiliations. Also, we proposed another clustering algorithm based on K-Mer Boolean feature vector sorting – KOFER. Parameters of the algorithm are trained on the Novosibirsk affiliation dataset consisting of 1000 samples.
Results: We show that *DBSCAN* method gives 0.81 v-measure score on the Novosibirsk affiliations dataset, while *KOFER* gives 0.9 v-measure score. We also present how affiliation grouping can be used to provide some statistics about institutional interactions, and provide institutions interaction network for Novosibirsk institutions and institutions in the miRNA science field gathered from PubMed database.
Conclusion: The results obtained show that institution from the miRNA conform network with *small-world* properties and that the proposed *KOFER* algorithm performs better than *DBSCAN* on the affiliations names data.

References

1. Titov I.I., Blinov A.A. (2014) Exploring the structure and evolution of the Novosibirsk biomedical co-authorship network. Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding. 18(4/2):939-944. (in Russian)
2. Fortunato S., Bergstrom C.T., Börner K., Evans J.A., Helbing D., Milojević S., Petersen A.M., Radicchi F., Sinatra R., Uzzi B., Vespignani A., Waltman L., Wang1 D., Barabási A.-L. Science of science. [Online] [Cited: 5 2, 2018.] http://science.sciencemag.org/content/359/6379/eaao0185.full.