Fraunhofer
SCAI

Knowledge Discovery on Biomedical Literature: Validating and Quality Control on Cause and Effect Networks

J. Dörpinghaus*[1], A. Tom Kodamullil*[1] , S. Madan*

[1]Fraunhofer Institute for Scientific Computing and Algorithms SCAI, Sankt Augustin; *equal contributions

Outline

Today the biomedical field beside of in-vitro, assay experiments, clinical trials mostly relies on systems biology approaches such as integrative knowledge graphs to decipher mechanism of a disease, by considering system as a whole (holistic approach). In that, disease modeling and pathway databases plays an important role. Knowledge Graphs built using Biological Expression Language (BEL, see www.openbel.org) is widely applied in biomedical domain to convert unstructured textual knowledge into a computable form. The BEL statements that forms knowledge graphs are semantic triples that consist of named entities, functions and relationships (Fluck et al. 2013). We face several challenges while converting knowledge from literature into knowledge graphs. First challenge is dimension reduction, which is building the relevant literature corpora to build the knowledge graphs. It is hard to extract the relevant articles for a topic by an unaided human. Second challenge is the publication bias, meaning, biomedical research is biased towards certain well-known findings and it is obvious that you find more articles related to this well-known topic and relatively less number of articles representing novel findings.

Introduction

Here, we propose statistic measures based on document clustering (Dörpinghaus et al. 2017) to quantify completeness and coverage, to prove the quality of a knowledge graph by identifying the scope, to distinguish and prioritize well-known, novel and missing knowledge based on literature. We developed two methods: an internal criterion and an external criterion. The internal criterion helps to evaluate the model itself and to find the coverage and scope of the knowledge graph. The external criterion is to evaluate the network knowledge against all computable available scientific knowledge, for example in the entire MEDLINE. This does not cover all knowledge, but the digital data available.

While comparing this external criterion with internal criterion, we can define the completeness of the model. This will also quantify the missing knowledge in the network with respect to the data sources that can be added to the network. This is – once again – only a check against the digital data available.

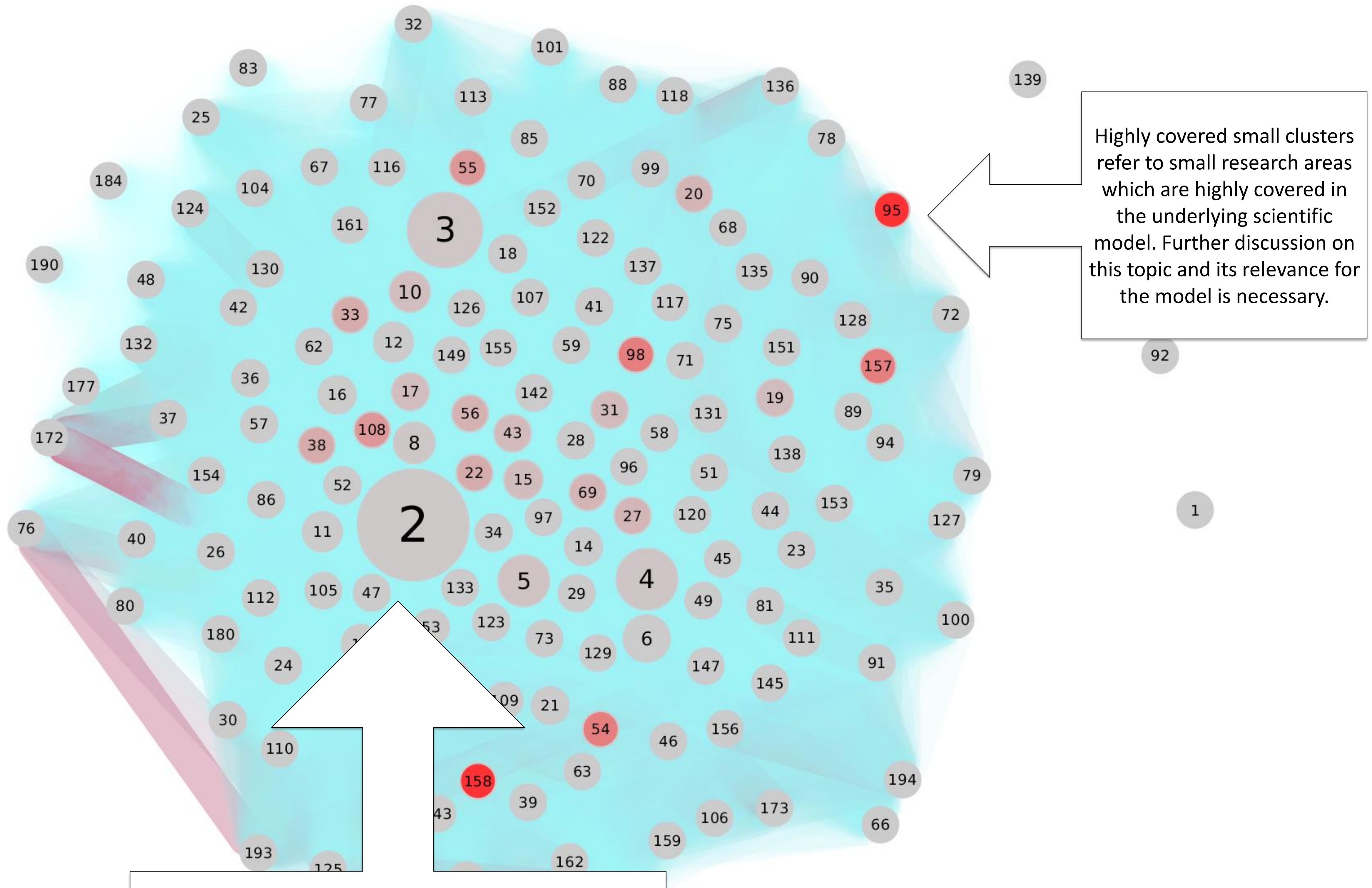
The different network attributes and properties obtained by the external criterion help to distinguishes which topic is overly represented. In addition it gives more information on ignoromes – the underrepresented novel findings

External Criterion

Several approaches for clustering textual data are known. For our application we need a discrete heuristic with flexible similarities without previous knowledge about what we want to see. In (Dörpinghaus et al. 2017) a novel soft-document clustering approach based on discrete algorithms was discussed.

Having a set of documents D and a subset R of documents contained in the model as well as n clusters $C_1,\ldots,C_n$ we can calculate the coverage of each cluster.

Figure 1 gives an example output of this method and some interpretation examples.



Highly covered small clusters refer to small research areas which are highly covered in the underlying scientific model. Further discussion on this topic and its relevance for the model is necessary.

Large clusters refer to greater research areas. Here we see a large cluster which is only sparsely covered in the model. This may have several reasons. For example there might not be more knowledge that needs to be represented in the model or which is relevant for the model. But this might be a good candidate to overcome the publication bias.

Fig. 1 – Visualization of the external criterion

The partition of the literature into clusters. The numbers identify the clusters. The size of a node is related to the number of documents included. The edges and their widths and color describe their weight. A darker blue edge has a greater weight.

In addition the red color indicates the coverage with the Alzheimer's model build by (Kodamullil et al 2015).

Internal Criterion

Finding ignorome can be done within the model itself. We can calculate the relative frequency of each statement witch helps to identify the most interesting statements in relation to those statements which are mentioned disproportionately high.

In addition we can find the most interesting documents containing the most significant statements by calculating the relative frequency of each statement. The documents obtained by this method may contain ignorome data. See the figures on the right for an illustration.

Discussion

In this paper we discussed some early work on the quality control or the validation on cause and effect networks generated from knowledge discovery and data mining methods on medical literature. These methods are quite general and could be applied to other applications as well.

We found that these methods are quite robust and give a valuable output and insights on generated models. Further research has to be done on the statistics for the internal criterion. What are significant values that should be examined? In addition we plan an integration into SCAIView software that makes it more easy to combine existing BEL-models with a corpus from literature. Here researches should get more easy feedback about ignorome and about important documents to consider.

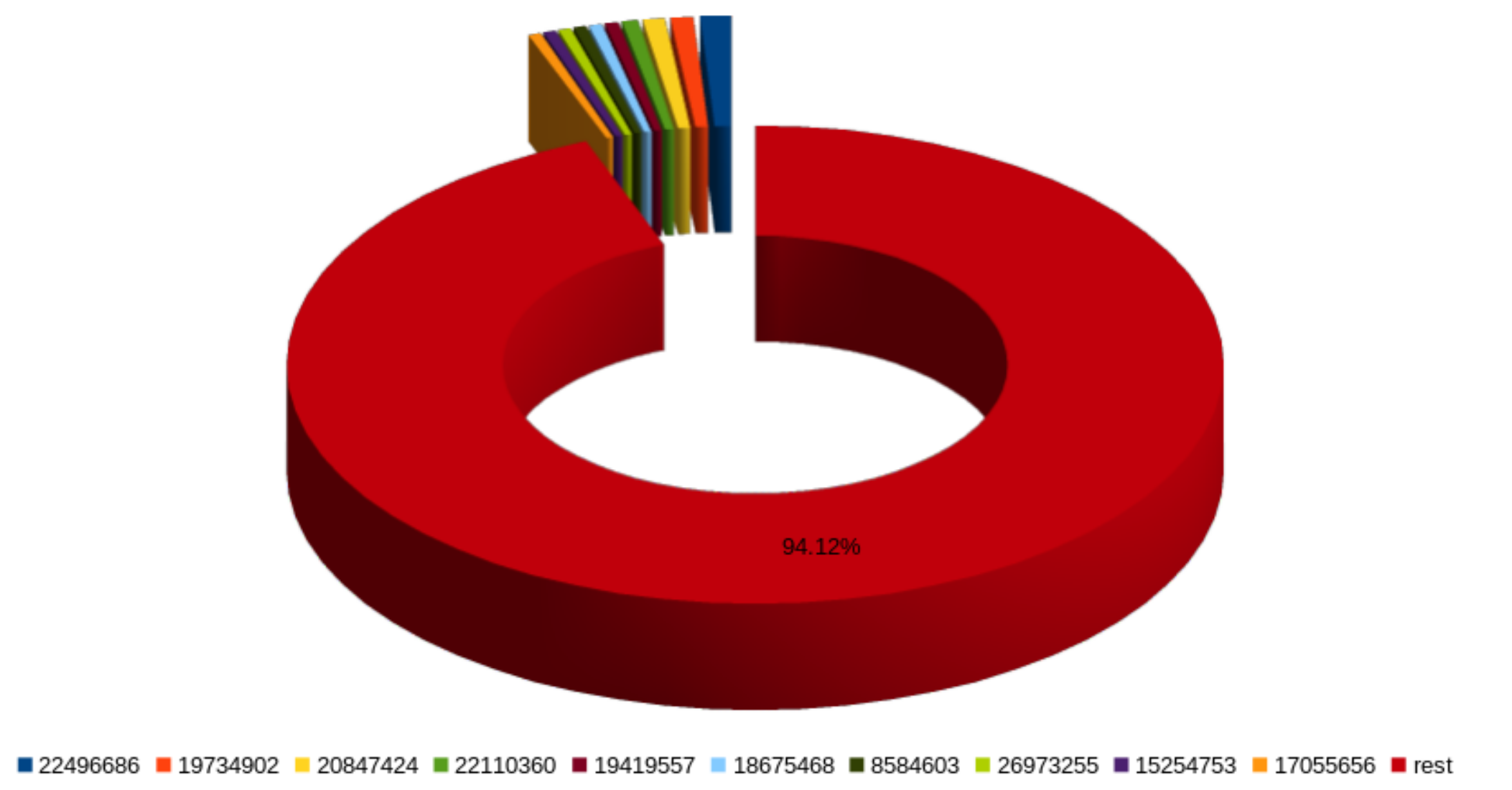


Fig. 2 – Visualization of the internal criterion: documents

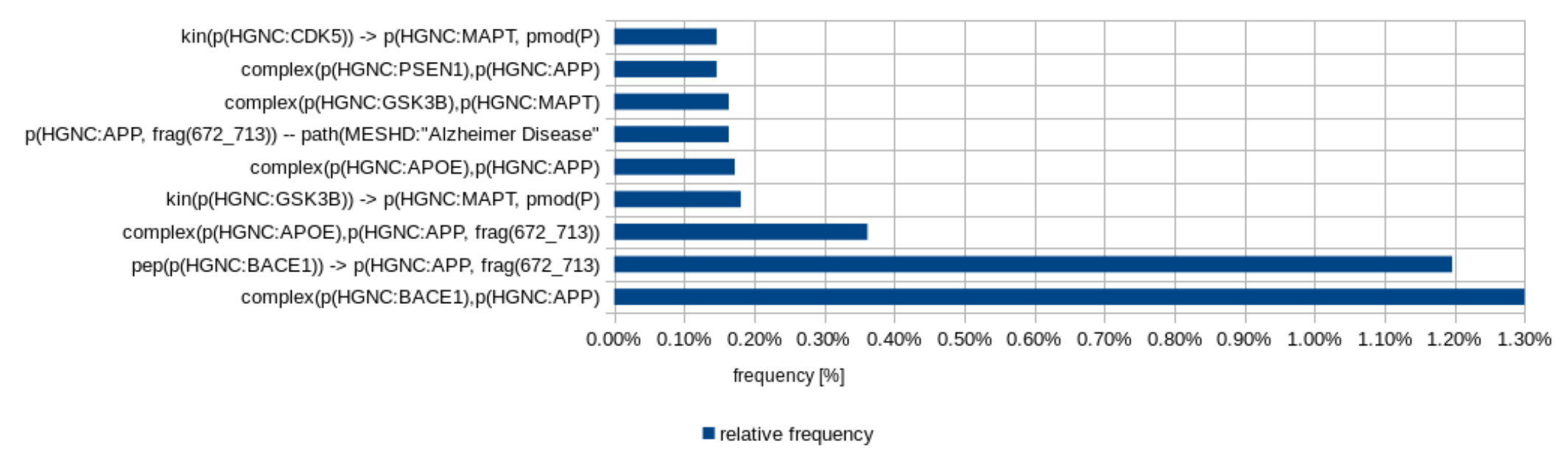
Information content of the 10 documents with the most statements. Here 10 documents contain 5.88% of all statements.

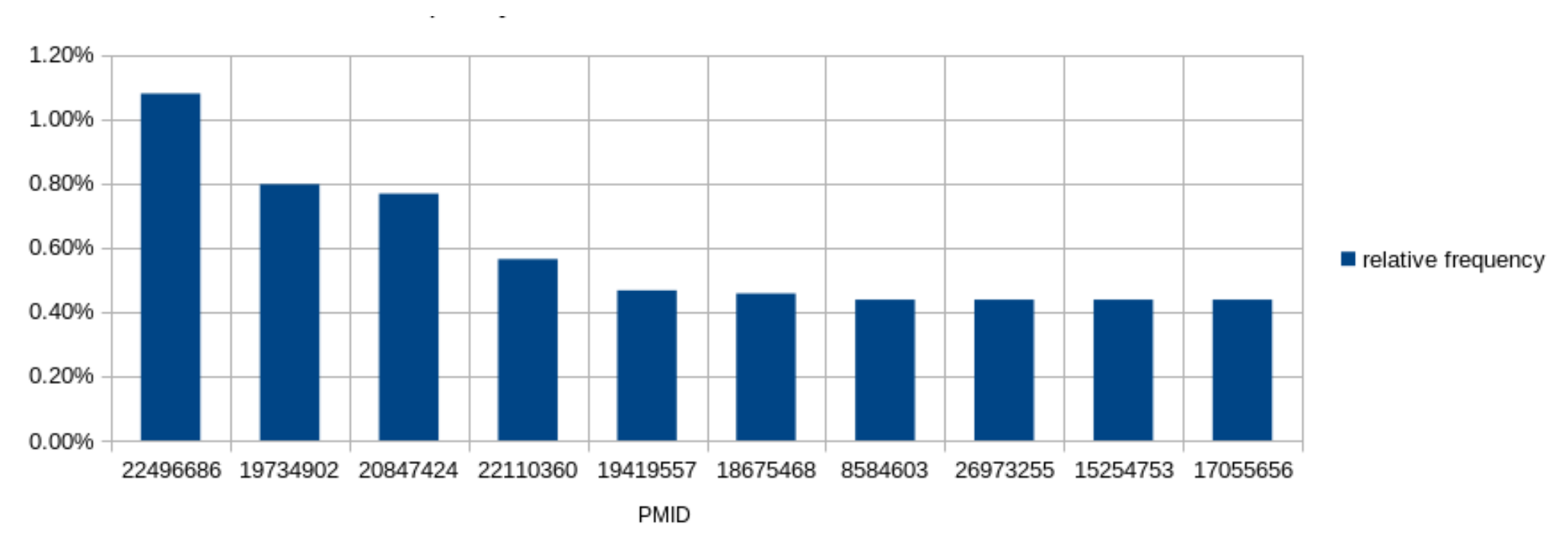
Fig. 3 – Visualization of the internal criterion: statements

Relative frequency of statements for the 10 documents with most statements.

Fig. 4 – Visualization of the internal criterion: statement relations

Relative frequency of 10 most frequent statements in relation to all annotated statements.





References

Fluck, J et al. BEL Networks derived from qualitative translations of BioNLP Shared Task annotations. *Proceeding of the 2013 Workshop on Biomedical Natural Language Processing*. 2013

Dörpinghaus, J, et al. "Document clustering using a graph covering with pseudostable sets." *Computer Science and Information Systems (FedCSIS), 2017 Federated Conference on*. IEEE, 2017.

COORDINATORS, NCBI Resource. Database resources of the national center for biotechnology information. *Nucleic acids research*, 2017, 45. Jg., Nr. Database issue, S. D12.

Kodamullil, Alpha Tom, et al. Computable cause-and-effect models of healthy and Alzheimer's disease states and their mechanistic differential analysis. *Alzheimer's & Dementia*, 2015, 11. Jg., Nr. 11, S. 1329-1339.

Contact

Jens Dörpinghaus jens.doerpinghaus@scai.fraunhofer.de
Alpha Tom Kodamullil alpha.tom.kodamullils@scai.fraunhofer.de
Sumit Madan sumit.madan@scai.fraunhofer.de

Fraunhofer Institute for Algorithms and Scientific Computing SCAI