HierClasSArt: Knowledge-Aware Hierarchical Classification of Scholarly Articles

Mehwish Alam, Russa Biswas, Yiyi Chen, Danilo Dessì, Genet Asefa Gesese,
Fabian Hoppe, Harald Sack*
firstname.lastname@fiz-karlsruhe.de
FIZ-Karlsruhe, Leibniz Institute for Information Infrastructure, Karlsruhe Institute for Technology
Germany

ABSTRACT

A huge number of scholarly articles published every day in different domains makes it hard for the experts to organize and stay updated with the new research in a particular domain. This study gives an overview of a new approach, *HierClasSArt*, for knowledge aware hierarchical classification of the scholarly articles for mathematics into a predefined taxonomy. The method uses combination of neural networks and Knowledge Graphs for better document representation along with the meta-data information. This position paper further discusses the open problems about incorporation of new articles and evolving hierarchies in the pipeline. Mathematics domain has been used as a use-case.

CCS CONCEPTS

• **Computing methodologies → Knowledge representation and reasoning**; **Semantic networks**; **Information extraction**; **Topic modeling**.

KEYWORDS

Scholarly Data, Hierarchical Classification, Knowledge Graphs, Deep Learning

1 INTRODUCTION

Knowledge Graphs (KGs) have recently gained remarkable momentum for representing structured knowledge about a particular domain. Since its advent, the Linked Open Data (LOD) cloud has constantly been growing containing many KGs about many different domains such as artificial intelligence, biomedical domain, etc. There has been an increase in the studies generating KGs from

*All authors contributed equally to this work.

scholarly articles [3] since many new articles are published everyday which are either peer-reviewed or not (such as on arXiv). It is hard for the experts to organize and stay updated with the newly published articles which leads to the necessity of developing new methods for automatically organizing scientific articles. PubMed is one such platform which classifies the articles in the biological domain into pre-defined subjects, i.e., MeSH[1]. Same is the case with ACM Computing Classification System (ACCS)[2] for computer science articles and Mathematics Subject Classification (MSC) [8] for the articles from mathematical domain. However, these platforms are mostly based on the keywords defined in the articles from the taxonomy, i.e., without properly taking into consideration the content of the articles and other meta-data information.

This position paper discusses a proposal of an algorithm, namely *HierClasSArt*, which combines KGs with neural methods for the hierarchical classification of scholarly articles into predefined taxonomies. It leverages text, KGs (constructed from text), external cross-domain KGs such as Wikidata [19], meta-data and citation information. *HierClasSArt* first constructs KGs from the abstracts of the articles. It then generates latent representations of the documents by taking into account the text, entities, entity categories, along with the meta-data of the articles. These representations are then used for hierarchically classifying the documents. This would help in achieving more complete assignments of the articles to the taxonomy in case if scientists have missed or incorrectly labelled some classes from the taxonomy.

This paper further discusses the open problems originating from the real-world setting of Zentralblatt MATH (zbMATH)[3] by automatically assigning MSC classes to the documents. One of these problems includes the publication of new articles everyday. In such cases, the classifier has to be robust enough to take into account these constant updates. It was also observed that these taxonomies also are prone to evolution to reflect the current research, e.g., Mathematics Subject Classification (MSC) taxonomy, created in 1990, changes on regular intervals. These problems open opportunities for further research and discussion in the community.

The rest of this paper is structured as follows: Section 2 discusses the studies related to KG construction, scholarly data classification as well as studies using KG Embeddings for scholarly data. Section 3 provides detailed statistics and analysis of zbMATH while Section 4 gives a teaser of *HierClasSArt*. Section 5 discusses the open problems encountered and poses the questions for further discussion at the

---

[1] https://www.ncbi.nlm.nih.gov/mesh/
[2] https://dl.acm.org/ccs
[3] https://zbmath.org/

workshop. Finally, Section 6 summarizes the lessons learnt from this position paper.

2 RELATED WORK

The research community has been developing methods for several years for generating machine-understandable representations of scientific domain and is providing platforms for annotating scholarly articles [6], designing ontologies for representing scholarly data [15], and building a variety of large-scale KGs [3, 6]. For generating these resources, both manual and automatic approaches can be found in the literature. A recent effort of human-crafted KG is Open Research Knowledge Graph (ORKG)[4] [6] which facilitates human experts by providing the possibility to upload their scientific papers and provide machine-readable statements about them. On the other hand, automated approaches have been proposed and applied to build the Computer Science Ontology (CSO) [15] and the Artificial Intelligence KG (AI-KG) [3]. The CSO was built by using the Klink-2 algorithm [13] which uses a set of statistical indicators to estimate hierarchical, temporal, and similarity metrics to detect relationships between entities such as papers, authors, venues, and keywords. While CSO was built by using meta-data information of scientific papers, AI-KG was built by exploring the text of their title and abstracts using Natural Language Processing (NLP) [2] techniques. Although these methods are being explored for the Computer Science literature, there is still a gap for many other domains such as Mathematics. In [7], the authors analyzed the MSC 2010 based document classification of zbMATH by leveraging external knowledge from DBpedia to detect inconsistencies in the taxonomy.

Few of the efforts have also been directed towards using KG Embeddings on scholarly data. In [11], the authors propose a soft marginal TransE for performing KG Completion over the scholarly KGs. In another article [12], the authors propose a framework for embedding based recommendations on scholarly KG. More precisely, the study focuses on co-authorship link prediction using soft marginal TransE. In contrast to these approaches, the focus of this study is to give an overall view of a knowledge aware hierarchical classification approach using these embeddings over scholarly data. In [16], the authors train a logistic regression multi-class classifier on TF-IDF features to predict the 63 first-level classes of the MSC. A hierarchical multi-label classification of scholarly publications [10] has been proposed which considers also second and third level MSC classes with sufficient training data. This ML-KNN [24] inspired classifier is more resistant to noise and was tested on 240 third level classes. However, a drawback of supervised approaches is that they require sufficient training data for each class. Consequently, they cannot be applied to fine grained classes with only few examples. This problem is also emphasized by evolving taxonomies (cf. Section 5) which might include new classes with few or no examples. Thus, unsupervised models, which include external knowledge, have been proposed to tackle this issue. One of the most recent works in this scenario has been provided by [14] where the authors utilize word embeddings to calculate the similarity between computer science topics and the text of scientific papers to hierarchically classify them into the CSO. However, previous investigated

methods only rely on frequency and co-occurrence information about words, and do not consider any external knowledge source.

3 DATA STATISTICS

In this section, the statistics about MSC 2010 and MSC 2020 taxonomies as well as the zbMATH dataset are described.

3.1 Mathematics Subject Classification

Scholarly articles in mathematics are most commonly categorized by the MSC taxonomy [8]. This fine-grained taxonomy is organized into three hierarchical levels. The first level consists of 63 high-level classes covering all topics from *integral equations* to *mathematics education*. Each topic is further on structured into multiple sub-classes on the second and third level. The number of classes at each level is shown in Table 1. An alphanumerical code has been assigned to each of the classes in this hierarchy. For example, a MSC class '39A70' consists of the first level class '39 (Difference and functional equations)', the secondary class '39A (Difference equations)' and the third level class '39A70 (Difference operators)'. These codes, short class descriptions and general remarks about the labelling process are available online[5].

To keep up with the progress in the research within the mathematical domain, the MSC is subject to constant change. Therefore, it is revised by domain experts at regular intervals by adding and removing classes [4]. Table 1 shows the statistics of MSC 2010 and MSC 2020 generated from the publicly available MSC 2010 taxonomy [6] and MSC 2020 taxonomy [7] respectively. Naturally, the most changes occur on the third and second level.

Table 1: Number of classes at the different levels of MSC.

	Level 1	Level 2	Level 3
MSC 2010	63	591	5,607
MSC 2020	63	597	6,006

Another interesting property of the MSC taxonomy is the branching factor which is one of the key elements affecting the complexity of algorithms which leverage hierarchies such as hierarchical classification. As we can see in Table 2 the first and second level classes have on average around 9 sub-classes. However, the difference between median and mean, as well as the maximum indicate that there are some outlier classes with higher number of sub-classes.

Table 2: Branching factor at each category level of MSC.

	Level	Mean	Median	Max	Min
MSC 2010	1st	9.38	8	24	2
	2nd	9.38	7	38	1
MSC 2020	1st	9.48	9	24	2
	2nd	9.95	8	41	1

---

[4]https://www.orkg.org/orkg/

[5]https://zbmath.org/static/msc2020.pdf
[6]https://cran.r-project.org/web/classifications/MSC-2010.html
[7]https://www.zbmath.org/classification/?q=cc:*

3.2 zbMATH

zbMATH as a major reviewing service in the field of mathematics manages millions of mathematical papers. The metadata of papers managed by zbMATH are aggregated into a dataset, subsequently referred to as *zbMATH* dataset. This dataset consists of 4,171,550 articles, 1,855,774 unique authors and 3,169,168 articles with abstracts (76%). A brief overview of the length of the textual metadata (title, abstract) as well as the number of authors per article are given in Table 3.

Table 3: Number of tokens in titles and abstracts, number of authors per article.

	Mean	Median	Max	Min
Title	9.46	9	127	1
Abstract	120.10	100	9,634	1
Authors	1.82	2	130	1

In order to make the articles retrievable and support the reviewing process they are categorized into the MSC taxonomy. Each article can be labelled with multiple classes from all levels of the MSC. The zbMATH dataset contains 3,679,097 labelled articles (88%). The distribution of the classes over articles including all entailed labels based on the taxonomy at different levels is shown in Table 4.

Table 4: Class distribution over articles at different levels.

Class Level	Mean	Median	Max	Min
1st	115,931.11	53,560.5	679,914	1
2nd	13178.21	4731	216,836	1
3rd	1429.70	526.0	34,882	1

In zbMATH, each article can have more than one class and some classes occur together in different articles. The high frequency of the co-occurrence of classes implicates high attraction of intersecting interests. The most frequent classes indicate the subjects having the most intersecting interests in specific areas. Figure 1 shows a co-occurrence network of classes of the dataset. As the reader can see, one of the most salient classes (co-occur most frequently with other classes) is '68T05', which means that the topic *"Learning and adaptive systems in artificial intelligence"* is gaining a lot of attention in scientific research.

4 AN OVERVIEW OF HierClasSArt

This section gives a proposal of the pipeline of *HierClasSArt*, an approach for knowledge-aware hierarchical classification of the scholarly articles. As a first step, a KG from the content of the scholarly articles of *zbMATH* will be created, called as *Math-KG*. In the second step, entities will be linked to external KGs and entity categories will also be considered. Meta-data along with the citation information about scholarly articles will also be utilized. A knowledge aware document representation will then be generated for the articles by combining the vectors of the previously extracted information which can further be used for hierarchical classification.
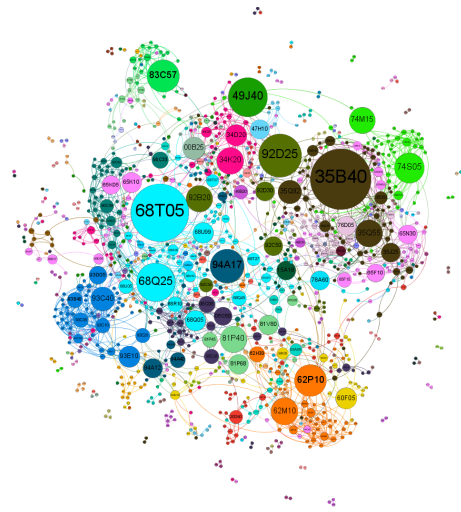

Figure 1: Co-occurrence network of classes in Articles. The same color of nodes refers to the same class at the first level.

4.1 Knowledge Graph Construction

The construction of Mathematical Knowledge Graph, named as *Math-KG*, starting from the abstracts of the articles in zbMATH will be a two step process: (i) extraction of entities from text, entity linking, and entity categorization, (ii) relation extraction.

4.1.1 Entity and Entity Category Extraction. Extracting entities from text requires the design of methodologies tuned on the peculiarities of the target domain. As such, two main directions will be explored: (i) open domain name entity recognition will be performed (e.g., using *Stanford Core NLP*[8]) and the results will be refined by exploiting existing vocabularies about mathematics, and (ii) design ML-based tools that recognize spans of the text as entities by exploring external knowledge about the domain; this will take inspiration from [21] where contextualized word embeddings were used within a Deep Learning model for recognizing entities of a predefined types. These entities will also be linked to Wikidata and also to Wikipedia-categories.

4.1.2 Relation Extraction. Relation extraction for building *Math-KG* is crucial for describing the semantics and support the classification approach. To this end, two directions will mainly be explored. A top-down approach which will include the definition of object properties to describe the relationships between entities with the help of domain experts. This approach might produce good quality outcomes but it requires human effort. A bottom-up approach focuses on NLP methods for identifying which elements commonly connect the entities (e.g., to describe the relationship between two entities a verb might often be used by domain experts). This may lead to noisy results and requires a post-processing step to identify significant elements. This can be solved by counting the frequency of the relations in the whole corpus of scholarly data.

---

[8]https://stanfordnlp.github.io/CoreNLP/ner.html

4.2 Exploiting Meta-data Information

Math-KG constructed in the previous step only contains the content of the abstracts of the scholarly articles. Additionally, meta-data of the scholarly articles provide a valuable information. This information includes, title, authors, venue, keywords, citation, etc. Combining this information with the KGs generated from the content will provide a better semantic representation of the mathematical research knowledge. For completing *Math-KG* with the meta-data information, a similar approach to AceKG [23] will be followed.

4.3 Knowledge Aware Representation and Classification of Scholarly Articles

Representation learning for scholarly articles can be achieved by leveraging the structured information from *Math-KG* and the textual information (article abstracts). Several approaches could be tried out to generate embeddings for *Math-KG*, i.e., translational models, semantic matching models, etc. [5]. In order to generate vector representations from text Doc2Vec [9] or BiLSTM [17] can be used. Moreover, the entities in *Math-KG* are being linked to external KGs such as Wikidata as well as Wikipedia-categories. These vector representations will be generated with the help of entity/entity category matrices. Finally, the vector representations of the KG of meta-data will be generated using TransE. All these vectors will be combined into one for generating knowledge aware representation for each document. This document representation can be further used in for any hierarchical classifier [18].

4.4 Utilizing the Citation Network of zbMATH

In case of classes in MSC where we have insufficient training data, the citations of the articles can be utilized. A citation network will be generated from the scholarly articles in zbMATH and the semantic similarity of the nodes in the network can be exploited with the help of network embeddings [1] to generate training data for insufficient classes. Moreover, the citation network can also help in improving a hierarchical classifier which classifies entities in the network into classes from the MSC with the assumption that : "Two papers p_1 and p_2 are classified under the same class or related classes in the MSC if and only if p_1 and p_2 share significantly enough number of papers as neighbours in the citation network".

5 OPEN PROBLEMS

This section discusses the open problems encountered while studying zbMATH which can lead to the development of a more sophisticated approach for the hierarchical classification of scholarly articles. These problems include constant addition of new articles as well as evolution of the taxonomy (more specifically, MSC taxonomy). The question here is how to capture these changes in this embedding based pipeline?

5.1 Adding New Articles

In case of the arrival of new articles, *Math-KG* should be updated automatically. This leads to the fact the KG Embeddings should also be prone to these updates instead of re-computation. Would these KG Embeddings help in dealing with the changes? How much computational power would it consume to take into account these updates? These are still open questions which are needed to be taken into account.

5.2 Evolving Hierarchies

The advancement of science and technology is continuously opening new research areas in various domains. The endless publication of huge number of research papers covering new topics leads to the problem of categorizing them in the existing classification model. Therefore, introduction of new classes in the hierarchy of classification model is necessary. For instance, in MSC 2020 new classes such as *Artificial neural networks and deep learning (68T07), Computational aspects of data analysis and big data (68T09), etc.* have been introduced as sub-classes of *Artificial Intelligence (68Txx)*. Moreover, the growth of research on the existing areas in a particular domain could also lead to further refinement of the topic, resulting in more fine grained classes in the hierarchical structure. Therefore, to maintain a proper classification structure for the research papers, changes in the hierarchy of the classification structure are inevitable. For example, the class in MSC 2010, *Relations with number theory and harmonic analysis (37A45)* is split up into two classes in MSC 2020 as *Relations between ergodic theory and number theory (37A44)* and *Relations between ergodic theory and harmonic analysis (37A46)*[9] (cf. Section 3).

In order to solve the problem of changes in the labels of the multi-label classification, Dynamic Multi-label Classification (DMC) approach has been introduced [22]. It is an incremental learning based approach to detect new classes and subsequently classify the text documents to the new classes. In this model, at first the text documents are classified to the already known classes followed by the detection of new classes which are then separated using Density based spatial clustering (DBSCAN). Finally, a new classifier is constructed for each new class which works collaboratively with the classifier for the known classes. To support this process, dynamic explicit representation modules [20] to track the changes into the MSC will be considered. However, the DMC model does not perform hierarchical classification. Future work will be conducted making this approach prone to evolution.

6 LESSONS LEARNT AND FUTURE WORK

This position paper indicates that there is still a research gap in leveraging KGs for the task of better organization and exploration of scholarly articles, specifically, in the task of hierarchical article classification in the domain of Mathematics. KGs generated from the content of the articles themselves would help to build a knowledge aware classifier. In addition to such KGs, citation networks also have significant advantage to help incorporate additional semantics to the classifier. It also has been noted that not enough attention has been given to the fact that new classes are constantly being added to existing taxonomies. However, it is important to properly handle such behaviour of taxonomies when building a classifier. The implementation of the proposed approach is part of further research where the *Math-KG*, the citation network and *HierClasSArt* classifier will be built and evaluated. Moreover, extensive experiments will be conducted using zbMATH and also use-cases from domains other than Mathematics.

---

[9]https://mathscinet.ams.org/msc/conv.html?from=2020

REFERENCES

[1] Peng Cui, Xiao Wang, Jian Pei, and Wenwu Zhu. 2019. A Survey on Network Embedding. *IEEE Trans. Knowl. Data Eng.* 31, 5 (2019), 833–852. https://doi.org/10.1109/TKDE.2018.2849727

[2] Danilo Dessì, Francesco Osborne, Diego Reforgiato Recupero, Davide Buscaldi, and Enrico Motta. 2021. Generating knowledge graphs by employing Natural Language Processing and Machine Learning techniques within the scholarly domain. *Future Gener. Comput. Syst.* 116 (2021), 253–264. https://doi.org/10.1016/j.future.2020.10.026

[3] Danilo Dessì, Francesco Osborne, Diego Reforgiato Recupero, Davide Buscaldi, Enrico Motta, and Harald Sack. 2020. AI-KG: An Automatically Generated Knowledge Graph of Artificial Intelligence. In *Proceedings of the 19th International Semantic Web Conference, Part II (Lecture Notes in Computer Science, Vol. 12507)*. Springer, 127–143.

[4] Edward Dunne and Klaus Hulek. 2020. Mathematics Subject Classification 2020. *EMS Newsletter* 3, 115 (2020), 5–6.

[5] Genet Asefa Gesese, Russa Biswas, Mehwish Alam, and Harald Sack. 2019. A Survey on Knowledge Graph Embeddings with Literals: Which model links better Literal-ly? *CoRR* abs/1910.12507 (2019). arXiv:1910.12507 http://arxiv.org/abs/1910.12507

[6] Mohamad Yaser Jaradeh, Allard Oelen, Kheir Eddine Farfar, Manuel Prinz, Jennifer D'Souza, Gábor Kismihók, Markus Stocker, and Sören Auer. 2019. Open research knowledge graph: next generation infrastructure for semantic scholarly knowledge. In *Proceedings of the 10th International Conference on Knowledge Capture.* ACM, 243–246.

[7] Maria Koutraki, O. Teschke, H. Sack, F. Müller, and Adam Bannister. 2017. Leveraging Mathematical Subject Information to Enhance Bibliometric Data. In *ESWC 2017.*

[8] Andreas Kühnemund. 2016. The role of applications within the reviewing service zbMATH. *PAMM* 16, 1 (2016), 961–962. https://doi.org/10.1002/pamm.201610459 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/pamm.201610459

[9] Quoc V. Le and Tomás Mikolov. 2014. Distributed Representations of Sentences and Documents. *CoRR* abs/1405.4053 (2014). arXiv:1405.4053 http://arxiv.org/abs/1405.4053

[10] Michał Łukasik, Tomasz Kuśmierczyk, Łukasz Bolikowski, and Hung Son Nguyen. 2013. Hierarchical, multi-label classification of scholarly publications: modifications of ML-KNN algorithm. In *Intelligent Tools for Building a Scientific Information Platform.* Springer, 343–363.

[11] Mojtaba Nayyeri, Sahar Vahdati, Jens Lehmann, and Hamed Shariat Yazdi. 2019. Soft Marginal TransE for Scholarly Knowledge Graph Completion. *CoRR* abs/1904.12211 (2019).

[12] Mojtaba Nayyeri, Sahar Vahdati, Xiaotian Zhou, Hamed Shariat Yazdi, and Jens Lehmann. 2020. Embedding-Based Recommendations on Scholarly Knowledge Graphs. In *Proceedings of the 17th International Conference, ESWC (Lecture Notes in Computer Science, Vol. 12123).* 255–270.

[13] Francesco Osborne and Enrico Motta. 2015. Klink-2: integrating multiple web sources to generate semantic topic networks. In *International Semantic Web Conference.* Springer, 408–424.

[14] Angelo Antonio Salatino, Francesco Osborne, Thiviyan Thanapalasingam, and Enrico Motta. 2019. The CSO Classifier: Ontology-Driven Detection of Research Topics in Scholarly Articles. In *Proceedings of Digital Libraries for Open Knowledge - 23rd International Conference on Theory and Practice of Digital Libraries, TPDL.*

[15] Angelo A Salatino, Thiviyan Thanapalasingam, Andrea Mannocci, Aliaksandr Birukou, Francesco Osborne, and Enrico Motta. 2020. The computer science ontology: A comprehensive automatically-generated taxonomy of research areas. *Data Intelligence* 2, 3 (2020), 379–416.

[16] Moritz Schubotz, Philipp Scharpf, Olaf Teschke, Andreas Kühnemund, Corinna Breitinger, and Bela Gipp. 2020. AutoMSC : Automatic Assignment of Mathematics Subject Classification Labels. In *Intelligent Computer Mathematics - 13th International Conference , CICM 2020, Bertinoro, Italy, July 26 -31, 2020, Proceedings (Lecture Notes in Computer Science, Vol. 12236)*, Christoph Benzmüller and Bruce R . Miller (Eds.). Springer, 237 – –250. https://doi.org/10.1007/978-3-030-53518-6_15

[17] M. Schuster and K. K. Paliwal. 1997. Bidirectional Recurrent Neural Networks. *IEEE Transactions on Signal Processing* 45, 11 (1997), 2673–2681. https://doi.org/10.1109/78.650093

[18] Roger Alan Stein, Patricia A Jaques, and Joao Francisco Valiati. 2019. An analysis of hierarchical text classification using word embeddings. *Information Sciences* 471 (2019), 216–232.

[19] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM* 57, 10 (Sept. 2014), 78–85. https://doi.org/10.1145/2629489

[20] Oleksandra Vsesviatska, Tabea Tietz, Fabian Hoppe, Mirjam Sprau, Nils Meyer, Danilo Dessi, and Harald Sack. 2021. ArDO: An Ontology to Describe the Dynamics of Multimedia Archival Records [to be published]. In *ACM, Symposium On Applied Computing.*

[21] David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, Relation, and Event Extraction with Contextualized Span Representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).* 5788–5793.

[22] Lun Wang, Wentao Xiao, and Shan Ye. 2019. Dynamic Multi-label Learning with Multiple New Labels. In *International Conference on Image and Graphics.* Springer, 421–431.

[23] Ruijie Wang, Yuchen Yan, Jialu Wang, Yuting Jia, Ye Zhang, Weinan Zhang, and Xinbing Wang. 2018. AceKG: A Large-Scale Knowledge Graph for Academic Data Mining *(CIKM '18)*. Association for Computing Machinery, New York, NY, USA, 1487–1490. https://doi.org/10.1145/3269206.3269252

[24] Min-Ling Zhang and Zhi-Hua Zhou. 2007. ML-KNN: A lazy learning approach to multi-label learning. *Pattern recognition* 40, 7 (2007), 2038–2048.