Leveraging Mathematical Subject Information to Enhance Bibliometric Data

Maria Koutraki[1], Olaf Teschke[2], Harald Sack[1], Fabian Müller[2], and Adam Bannister[2]

[1] FIZ Karlsruhe – Leibniz Institute for Information Infrastructure
Karlsruhe Institute of Technology, Institute AIFB, Germany
[2] FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Berlin, Germany
firstname.lastname@fiz-karlsruhe.de

Abstract. The field of mathematics is known to be especially challenging from a bibliometric point of view. Its bibliographic metrics are especially sensitive to distortions and are heavily influenced by the subject and its popularity. Therefore, quantitative methods are prone to misrepresentations, and need to take subject information into account. In this paper we investigate how the mathematical bibliography of the abstracting and reviewing service Zentralblatt MATH (zbMATH) could further benefit from the inclusion of mathematical subject information MSC2010. Furthermore, the mappings of MSC2010 to Linked Open Data resources have been upgraded and extended to also benefit from semantic information provided by DBpedia.

Keywords: scientometrics, bibliometrics, linked data, mathematics

1 Introduction

The field of mathematics is known to be especially challenging from a bibliometric point of view. The application of general bibliometric methods have led to generally non-satisfactory outcomes, resulting in a broad rejection of such statistical measures in the mathematical community [1]. Several specifics of mathematical publications come into effect here: first of all, since mathematics is a relatively small area in terms of number of documents and references, metrics are especially sensitive to distortions [3]. This leads, e.g., to the situation that a journal impact factor is currently uncorrelated with its scientific quality [3,18]. A second factor is the unusual longevity of mathematical research [5]. With a citation half-life well beyond the period of ten years, standard measures fail to count the most significant impact quantities.

A third effect is the very diverse nature of mathematics. Being the language of modern exact science, mathematical research is interspersed with basically all scientific subjects, which heavily influences the publication behavior in all possible aspects such as availability, peer review policies, publication delay, or coauthor networks [13]. Consequently, quantitative measures vary vastly even

when restricted to mathematical research alone. Publication and citation frequencies in particular are heavily influenced by the subject. A known effect is the perceived topical bias in generalist math journals [6]: several mathematical areas, which contribute largely to the overall quantitative publication (and citation) numbers, as e.g. (mathematical) computer science, statistics, mathematical economics, or mathematical physics, contribute only marginally to generalist top tier mathematical journals. As shown in [10], this effect prevails for all generalist math journals independent of specifics like region, editorial board, or publisher.

Therefore, quantitative methods (bibliometric analysis, identification of trends or hot research topics, etc.) are prone to misrepresentations, and need to take subject information into account. An adequate starting point would be to employ the Mathematical Subject Classification (MSC2010).

The Mathematics Subject Classification (MSC)[3] is a classification scheme introduced in 1970 and maintained by Mathematical Reviews and zbMATH. The MSC has been revised every decade to adapt to the development of mathematics. Its current version MSC2010 was published in 2009. Traditionally a hierarchical system, its suitability to reflect the underlying connections between mathematical subjects is limited, though the recent versions include some attempts like cross-references to improve on this. For bibliometric studies, it would be highly desirable to derive further information on the similarity of MSC classes.

As MSC2010 has already been represented as Linked Data and mapped to the DBpedia[4] knowledge base as well as to the ACM Computing Classification System[5], the mapped information sources can be deployed as complementary information for further bibliometric and scientometric analysis.

The statistical analysis in this paper is based on the MSC assignments to mathematical publications in the zbMATH database. zbMATH (formerly Zentralblatt MATH) [6], is the world's most comprehensive and longest-running abstracting and reviewing service in pure and applied mathematics. Produced by FIZ Karlsruhe – Leibniz Institute for Information Infrastructure (FIZ Karlsruhe), it is edited by the European Mathematical Society (EMS), FIZ Karlsruhe, and the Heidelberg Academy of Sciences and Humanities, and distributed by Springer.

Earlier work [16] applied NLP methods to the zbMATH corpus, and obtained the following overlap of top-level MSC classes (cf. fig. 1). Darker colors indicate a higher similarity of subjects, hence the overall picture suggests that the classification is far from being relatively homogeneous, but rather contains many intrinsic relations which can be further studied by similarity analysis.

In this paper we make the following contributions:

(i) The already existing MSC2010 mapping to DBpedia has been corrected, upgraded to the most recent version, and enriched with additional subject mappings via the SKOS vocabulary [11].

[3] http://msc2010.org/
[4] http://dbpedia.org/
[5] http://www.acm.org/about/class/
[6] https://zbmath.org/

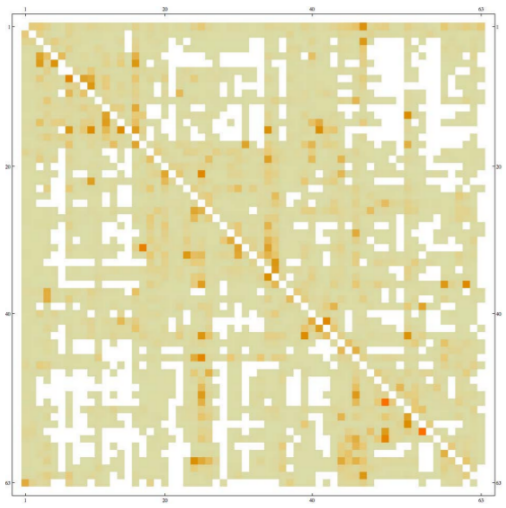

Fig. 1. Top-level MSC overlap after [16].

(ii) Statistical and semantic measures have been applied to compute the similarity of the MSC categories.
(iii) Inconsistencies and other issues have been detected in MSC2010 that should be addressed in the subsequent version MSC2020.

The paper is structured as follows: Sec. 2 presents the underlying zbMATH data sources as well as a brief outline of related work. In Sec. 3, the statistical analysis of the MSC2010 subject classification based on the bibliographic data of zbMATH is presented including the linking of MSC2010 to DBpedia as well as the semantic similarity computation based on DBpedia. Sec. 4 discusses the achieved results and Sec. 5 concludes the paper with a brief summary and an outlook on future work.

2 Related Work

The zbMATH database contains more than 3.5 million bibliographic entries with reviews or abstracts currently drawn from more than 3,000 journals and serials, and 170,000 books. Almost 10 million matched references provide the links of the citation network. Reviews are written by more than 10,000 international experts, and the entries are classified according to the MSC scheme (MSC 2010). The coverage starts in the 18th century and is complete from 1868 to the present by the integration of the "Jahrbuch über die Fortschritte der Mathematik" database. zbMATH is a subscription service but also allows non-subscribers to ask queries and access the freely accessible zbMATH author profile pages.

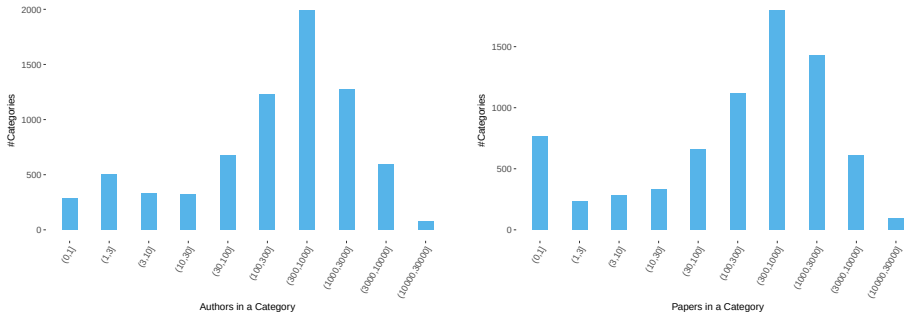

Fig. 2. Authors distribution (left) and papers distribution (right) across the different categories.

The Mathematical Subject Classification MSC2010 is organized as three-level classification tree with 63 first-level nodes, over 400 second-level nodes and more than 5,000 leaf nodes. MSC2010 is used by zbMATH to provide subject information for more than 3 million research articles, chapters, and proceedings papers, which are already indexed using this schema. MSC is designed for indexing resources at the granularity of articles or conference proceeding papers, i.e. it exposes article topics in a general way, but does not include specific theorems, functions, or sequences proven, which are discussed in a paper [14]. Documents from 1970–2009 classified by earlier MSC versions have been mapped to the recent MSC2010 by conversion tables. Fig. 2(left) shows the distribution of the zbMATH papers in the different MSC categories while fig. 2(right) shows the distribution of the authors of the papers to the MSC categories. Fig. 3 illustrates the highly diverse citation frequency (i.e., average citations to a paper indexed in zbMATH in a given subject) for the top level MSC categories, which differs by a factor greater than 10. This reinforces the necessity to take subject information into account for bibliometric studies in mathematics.

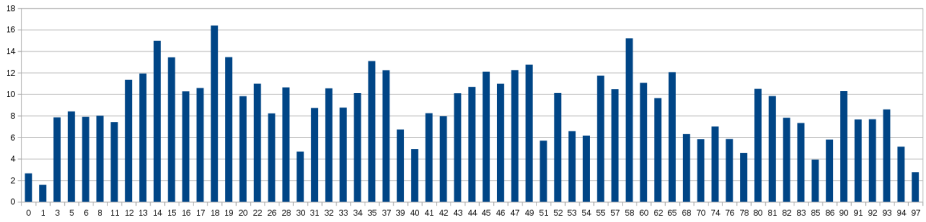

Fig. 3. Citation frequency for top level MSC categories.

MSC taxonomy has become part of the Web of Data being represented via SKOS (Simple Knowledge Organisation System, [12]) vocabulary and RDF (Re-

source Description Framework, [9]) and mapped to identical concepts in DBpedia as well as ACM Computing Classification System via `owl:sameAs` links [8]. In [7], Hu et al. describe a similar effort where data and metadata from the Semantic Web Journal are exposed as linked data, using a SPARQL endpoint. In this work, instead of the MSC classification, the authors use and extend the Bibliographic ontology[7].

One of the subtasks in this work is to reveal the relations between the MSC2010 categories. To this purpose, we use the `owl:sameAs` links of the MSC categories to the corresponding ones in DBpedia and compute the semantic similarity of the second to use as an extra factor to the general analysis of the MSC2010 taxonomy. Several methods have been proposed to compute semantic similarity in ontologies as [17] and [2]. I this work we decided to use the proposed approach in [15] since it is focused on computing the semantic similarity among the DBpedia resources.

3 Statistical and Semantic Analysis

For this work, already existing mappings of MSC2010 categories and DBpedia entities by Lange et al. [8] had to be corrected and synchronized with the current version of DBpedia[8]. In the course of this process, new mappings between MSC2010 categories and DBpedia entities have also been created (cf. fig. 4). The following adjustments have been made:

- The original mappings contain links to many so-called redirect pages, i.e. URIs (Uniform Resource Identifiers) that do not directly identify a DBpedia entity. DBpedia entities correspond to Wikipedia pages. Thus, DBpedia redirect URIs correspond to Wikipedia redirect pages. These serve the purpose of linking a Wikipedia page to alternative spellings, misspellings, or synonyms of the title of the underlying subject. DBpedia redirect URIs had to be replaced by their original DBpedia entity they are redirecting to.
- Since DBpedia is based on Wikipedia snapshots, which are published only a few times per year, entities represented in DBpedia are subject to changes [4]. As e.g. several names of Wikipedia categories or YAGO categories for mathematical subjects corresponding to MSC2010 categories have been changed and thereby had to be substituted by its successors whenever possible.
- MSC2010 categories are arranged as a taxonomy. Subordinate categories are linked to their superordinate categories via `skos:narrower` and vice versa via `skos:broader`. Thereby, DBpedia entities linked via `owl:sameAs` can be considered identical to their corresponding MSC2010 categories, and new `skos:narrower` and `skos:broader` relations can be created for the superordinate or subordinate MSC2010 categories and their corresponding DBpedia entities. Likewise, the MSC2010 categories linked by `skos:seeAlso` to other

[7] http://bibliontology.com/
[8] DBpedia version 2016-04, http://wiki.dbpedia.org/dbpedia-version-2016-04

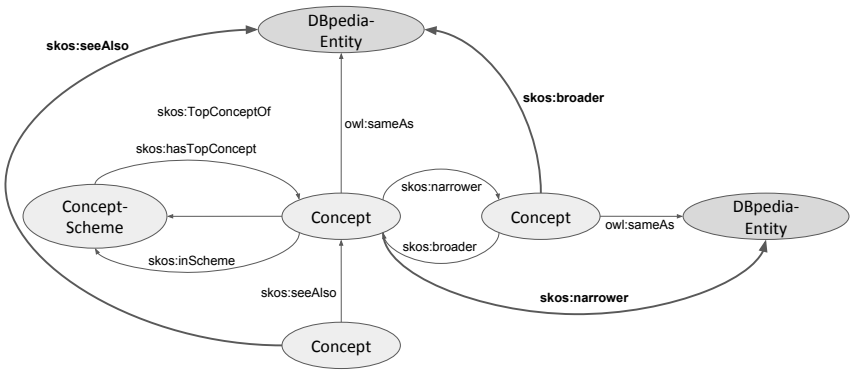

Fig. 4. SKOS based mappings and relations for MSC2010 concepts with DBpedia entities.

MSC2010 categories can also be linked to their corresponding DBpedia entities via the same property. Fig. 4 depicts the newly created mappings with thick arrows and property names in bold face.

In the original MSC2010 DBpedia mapping, 970 MSC2010 subjects are linked to 2,690 DBpedia entities via `owl:sameAs`. 5,245 MSC2010 subjects remained without a mapping to DBpedia. Via the new mappings mentioned above, 1,525 MSC2010 subjects could be mapped to DBpedia entities via `skos:narrower`, an additional 234 MSC2010 subjects via `skos:broader`, as well as 283 MSC2010 subjects via `skos:seeAlso`[9].

To better understand the similarity of the MSC2010 categories based on the number of co-occurring papers, the available mappings of MSC2010 to DBpedia have been used to compute the semantic similarity between the mapped DBpedia entities via the semantic similarity measure proposed by [15]. The semantic similarity of two DBpedia entities is based on taking into account the similarity of the properties of these resources as well as satisfying the fundamental axioms for similarity measures such as "equal self-similarity", "symmetry" and "minimality". For each pair of MSC2010 categories, the semantic similarity of corresponding DBpedia entities linked via `owl:sameAs` has been computed as well and compared to the similarity of MSC2010 categories based on the papers they are assigned to.

To compute the similarity of the MSC2010 categories we use the Jaccard similarity measure (see Equation 1). In our setting, for each MSC2010 category pair (cat_a, cat_b), the similarity is translated as the number of papers assigned to

[9] MSC2010 mapping is available at http://bit.ly/MSC2010mapping-ntriples

both categories cat_a and cat_b, normalized by the total number of papers assigned to cat_a or to cat_b.

$$\text{Jaccard}(cat_a, cat_b) := \frac{cat_a \cap cat_b}{cat_a \cup cat_b} \tag{1}$$

We compute the Jaccard similarity between categories w.r.t the zbMATH collection of documents from the mathematical domain. Each paper from zbMATH is assigned to one or more MSC2010 categories.

Furthermore, apart from the Jaccard coefficient presented in Equation 1, we compute the *asymmetric* Jaccard between two categories as in Equation 2.

In other words, the number of papers assigned to both categories cat_a and cat_b normalized by the number of papers assigned to cat_a. The asymmetric Jaccard is a useful indicator for the discovery of subclass relationships between the categories based on the papers assigned to them.

$$\text{asymmetric\_Jaccard}(cat_a \rightarrow cat_b) := \frac{cat_a \cap cat_b}{cat_a} \tag{2}$$

Since both measures, Jaccard and semantic similarity, are not directly comparable due to their different scaling behavior, we measure their correlation and see how the different measures capture the similarities between the different categories. We measure the correlation via *Pearson's rank correlation*.

Table 1 shows the achieved correlation results for symmetric Jaccard as well as for asymmetric Jaccard with the semantic similarities, based on DBpedia entities. For the experiment the top 10,000 similar MSC2010 categories for symmetric Jaccard as well as for asymmetric Jaccard have been taken into account. Of these categories only 937 for symmetric and 672 for asymmetric ($cat_a \rightarrow cat_b$) could be mapped to DBpedia via `owl:sameAs`. The results are discussed in the subsequent section.

similarity measure	r
semantic – Jaccard(cat_a, cat_b)	0.34
semantic – asymmetric_Jaccard($cat_a \rightarrow cat_b$)	-0.04

Table 1. Correlation coefficient comparing Jaccard based similarities for MSC2010 categories based on papers per category and semantic similarity of MSC2010 categories based on DBpedia entities corresponding to MSC2010 classes.

4 Discussion of the Achieved Results

In this section, we discuss our findings related to (i) the statistical and semantic similarity measures applied to the MSC taxonomy and (ii) the issues in the MSC2010 that can be addressed in the next 2020 version.

4.1 Similarity measures comparison

Table 1 presents the correlation between the two similarity measures, the *semantic* and the Jaccard similarity. In the first row, we show the correlation between the semantic similarity and the symmetric Jaccard coefficient. We find moderate correlation between the two measures with a score of $r = .34$. As far as it concerns the last row of the table, the correlation for the semantic similarity and the asymmetric Jaccard is weak. The explanation for this lies in the different nature of the two measures. The asymmetric Jaccard is a measure that is suitable for revealing subsumption relationships in contrast to the semantic similarity. The solution to this will be to compute both ways asymmetric Jaccard for each pair and consider them as similar if both ways asymmetric Jaccard is high. That is what the symmetric Jaccard is designed for.

4.2 Issues in MSC2010

Structural issues. One of the contributions of this work is to be able to discover and suggest part of the MSC2010 taxonomy that should be changed or improved in the next revision of the MSC, the MSC2020. A first finding that came out of the semantic analysis of the corpus was that subtrees in the taxonomy in which the first- or second-level category bears exactly the same name as the child category in the next level (second or third), the additional level of hierarchy generally does not contribute any distinguishing value. Furthermore, in the majority of the cases in the third level there is only one more category labeled as *"None of the above, but in this section"*. Worth to be mentioned here is that the vast majority of the papers related to those first-level categories from the zbMATH data are directly associated to the second-level category with the same label. Only very few of them are associated to the second-level category with label *"None of the above, but in this section"* and none are directly associated to the first-level category.

To this end, a suggestion for the next MSC version would be to revise those branches in the taxonomy and possibly merge them into one category per subtree. Table 2 presents a non-exhaustive list of categories that fall under the described exceptional case.

Wrong `owl:sameAs` *links.* Another interesting observation we made concerns the `owl:sameAs` links that exist between the MSC2010 and DBpedia. Many of the existing `owl:sameAs` relations link the MSC categories to the redirect resources of DBpedia instead of linking them to the correct resources themselves. Moreover, there are even erroneous links between the two schemas, for examples the category `65A05` with label *Tables* is linked (among others) to the `dbr:tablets`[10]. Therefore, together with the new revision of the MSC2020 another effort can be made in improving and extending the existing `owl:sameAs` links.

[10] http://dbpedia.org/resource/Tablets

Supercategory	Subcategory	label
13Gxx	13G05	*Integral domains*
14Txx	14T05	*Tropical geometry*
22Cxx	22C05	*Compact groups*
45Bxx	45B05	*Fredholm integral equations*
45Dxx	45D05	*Volterra integral equations*
45Pxx	45P05	*Integral operators*
45Qxx	45Q05	*Inverse problems*
62Qxx	62Q05	*Statistical tables*
65Axx	65A05	*Tables*
70Cxx	70C20	*Statics*
83Axx	83A05	*Special relativity*
83Fxx	83F05	*Cosmology*
85-XX	85Axx	*Astronomy and astrophysics*

Table 2. Cases that should be revised in the next MSC version. Categories and subcategories that share the same label and papers.

5 Conclusion and Outlook on Future Work

In this paper we have shown that beyond the three-level hierarchical structure of the MSC2010 taxonomy, intrinsic similarities can be measured in several alternative ways. These indicate options to enhance the taxonomy structure towards a more detailed ontology. The updated and partially corrected DBpedia linking provides additional information. A first useful result is the detection of a systematic flaw in the current MSC2010 scheme concerning the *"None of the above, but in this section"* leafs. A more detailed future analysis will be aimed to suggest more adaptations. Trend mining of publications since 2010 will indicate new research areas, which currently are not covered sufficiently, and a further discussion of similarities is likely to provide structural insights supporting the ongoing task of MSC2020 revision.

References

1. Adler, R., Ewing, J., Taylor, P.: Citation statistics. A report from the International Mathematical Union (IMU) in cooperation with the International Council of Industrial and Applied Mathematics (ICIAM) and the Institute of Mathematical Statistics (IMS). Stat. Sci. 24(1), 1–14 (2009)
2. Alsubait, T., Parsia, B., Sattler, U.: Measuring similarity in ontologies: A new family of measures. In: EKAW. pp. 13–25. Lecture Notes in Computer Science (2014)
3. Arnold, D.N., Fowler, K.K.: Nefarious numbers. Notices Am. Math. Soc. 58(3), 434–437 (2011)
4. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: Dbpedia - a crystallization point for the web of data. Web Semant. 7(3), 154–165 (Sep 2009), http://dx.doi.org/10.1016/j.websem.2009.07.002

5. Bouche, T., Teschke, O., Wojciechowski, K.: Time lag in mathematical references. Eur. Math. Soc. Newsl. 86, 54–55 (2012)
6. Grcar, J.F.: Topical bias in generalist mathematics journals. Notices Am. Math. Soc. 57(11), 1421–1429 (2010)
7. Hu, Y., Janowicz, K., McKenzie, G., Sengupta, K., Hitzler, P.: A linked-data-driven and semantically-enabled journal portal for scientometrics. In: International Semantic Web Conference (2). pp. 114–129 (2013)
8. Lange, C., Ion, P., Dimou, A., Bratsas, C., Corneli, J., Sperber, W., Kohlhase, M., Antoniou, I.: Reimplementing the mathematics subject classification (msc) as a linked open dataset. In: Proceedings of the 11th International Conference on Intelligent Computer Mathematics. pp. 458–462. CICM'12, Springer-Verlag, Berlin, Heidelberg (2012), http://dx.doi.org/10.1007/978-3-642-31374-5_36
9. Manola, F., Miller, E.: RDF primer. W3C recommendation, W3C (Feb 2004), http://www.w3.org/TR/2004/REC-rdf-primer-20040210/
10. Mihaljević-Brandt, H., Teschke, O.: Journal profiles and beyond: what makes a mathematics journal "general"? Eur. Math. Soc. Newsl. 91, 55–56 (2014)
11. Miles, A., Bechhofer, S.: Skos simple knowledge organization system reference. w3c recommendation 18 august 2009. (2009), http://www.w3.org/TR/2009/REC-skos-reference-20090818/
12. Miles, A., Bechhofer, S.: SKOS simple knowledge organization system reference. Tech. rep., W3C (2009), http://www.w3.org/TR/skos-reference
13. Müller, F., Teschke, O.: Progress of self-archiving within the DML corpus, with a view toward community dynamics. In: Intelligent computer mathematics. 9th international conference, CICM 2016, Bialystok, Poland, July 25–29, 2016. Proceedings, pp. 63–74. Cham: Springer (2016)
14. National Research Council: Developing a 21st Century Global Library for Mathematics Research. The National Academies Press, Washington, D.C (2014), https://arxiv.org/pdf/1404.1905.pdf
15. Piao, G., Ara, S.s., Breslin, J.G.: Computing the Semantic Similarity of Resources in DBpedia for Recommendation Purposes, pp. 185–200. Springer International Publishing, Cham (2016), http://dx.doi.org/10.1007/978-3-319-31676-5_13
16. Schöneberg, U., Sperber, W.: POS tagging and its applications for mathematics. Text analysis in mathematics. In: Intelligent computer mathematics. International conference, CICM 2014, Coimbra, Portugal, July 7–11, 2014. Proceedings, pp. 213–223. Berlin: Springer (2014)
17. Staab, S.: Ontologies and similarity. In: Case-Based Reasoning Research and Development - 19th International Conference on Case-Based Reasoning, ICCBR 2011, London, UK, September 12-15, 2011. Proceedings. pp. 11–16 (2011)
18. Teschke, O.: Negligible numbers. Eur. Math. Soc. Newsl. 82, 54–55 (2011)