Mathematical World Knowledge Contained in the Multilingual Wikipedia Project

Dennis Tobias Halbach(✉)

University of Wuppertal, Wuppertal, Germany
dennis.halbach@uni-wuppertal.de

Abstract. The purpose of this project is to test and evaluate an approach for Formula Concept Discovery (FCD). FCD aims at retrieving a formula concept (in the form of a Wikidata item) together with its defining formula within documents, in this case 100 English Wikipedia articles. To correctly identify the defining formula of a Wikipedia article, this approach searches for shared formulae across Wikipedia articles available in different languages. The formula shared in the most languages is then assumed to be the defining formula. The results show that neither this approach alone nor a combination with an existing approach that considers the order of the formulae inside an article leads to satisfying results. It is thus concluded that the number of times a formula is shared across a Wikipedia article in different languages is not a good indicator to determine the defining formula with the current approach. Consequently, several ideas for further research are proposed which could improve the results.

1 Introduction

For many generations mathematical textbooks were the primary source of information for pupils and laypersons to acquire mathematical knowledge. However, since the beginning of the 21st century and the rise of collaborative online encyclopaediae such as Wikipedia, this situation is changing. Wikipedia can basically be seen as a digital book organized in classical articles with cross-references. This format is similar to printed textbooks and not designed to be machine-readable. Thus, the automated retrieval of properties (like a formula) describing a related topic is a non-trivial task. To assist this task, the Wikidata knowledge graph was established in 2012. Wikidata connects different language versions of Wikipedia and stores data related to Wikipedia as triples, linking a data item (via its unique identifier called a 'QID') to one or multiple properties and their respective value. One such property can be the so called *defining formula* of a Wikidata item, which can be stored in the Wikidata knowledge graph since 2016. For example the Wikidata item on Schwarz's theorem (Q1503239) connects Wikipedia articles in 15 languages on the topic. Here the formula

$$\frac{\partial}{\partial x}\left(\frac{\partial}{\partial y}f(x,y)\right) = \frac{\partial}{\partial y}\left(\frac{\partial}{\partial x}f(x,y)\right)$$

A. M. Bigatti et al. (Eds.): ICMS 2020, LNCS 12097, pp. 353–361, 2020.
https://doi.org/10.1007/978-3-030-52200-1_35

is the defining formula of this concept and thus included in most of the 15 articles, although partly in different mathematical notations (see Fig. 1). As the data format, Wikidata uses *Presentation MathML* as the exchange format and the LaTeX dialect *texvc* as the input format.

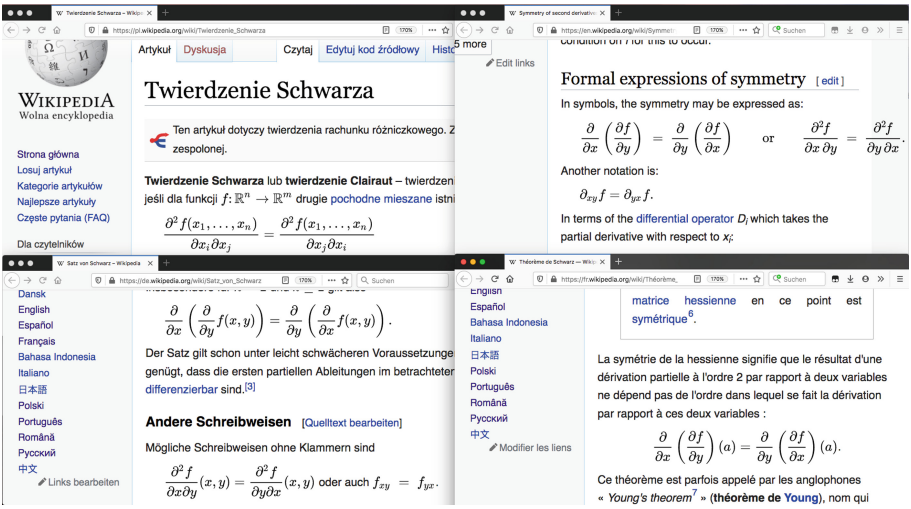

Fig. 1. The Wikipedia articles on Schwarz's theorem in the languages Polish, English, German and French. Accessed on 28th of March, 2020.

While some versions of Wikipedia like German and Portuguese include the exact form of the formula, the French and Spanish versions use a rather than x for the function argument and the English article lacks the function argument. Moreover, the Russian and Polish versions use numeric indices for the variables, i.e, x_1, x_2 instead of x, y. Still, judging from this one example, it seems possible to infer the defining formula from the reoccurrence of a formula across different Wikipedias, although advanced techniques might be needed, e.g. to recognize slightly different formulae as representing the same mathematical concept.

In this paper, we aim at improving the automatic extraction of defining formulae over an already existing approach from Schubotz et al. [2], who chose to extract the first formula included in the English Wikipedia article after a manual investigation showed that the first formula is often the most relevant one for that article [2] as it is frequently included in the introductory part of an article. Knowing the formula with the highest probability to be the defining formula can then be used to suggest formula edits to Wikidata editors.

2 Method

The Wikipedia articles used are obtained from a collection of Wikipedia article dumps[1] for all 309 official Wikipedia languages[2]. Specifically, we use the 100 articles defined as QIDs in the dataset from Schubotz et al. and their respective articles in other Wikipedia languages available. We infer the articles titles from the 100 QID using the MediaWiki API and use these titles to filter the dumps for all pages containing one of the titles in their title-tag. These pages will then be filtered to extract the formulae. Schubotz et al. consider a string a *formula* if it fulfills the following two conditions: Firstly, it has to be enclosed in a wikitext tag, namely 'math', 'ce', or 'chem'. Secondly, it needs to include (at least) one formula-indicator [1,4], in our case '=', '<', '>', '≤', '≥', '≈' and/or '≡' were used. These formula-indicators prevent that variables are recognized as formulae since a formula typically relates the definiens and the definiendum using formula-indicators.

After filtering all articles for formulae, the extracted strings[3] of all articles with the same QID are compared to determine the string shared by the most articles corresponding to each QID. The resulting 100 most common strings are then compared to a gold standard dataset derived from [2] to evaluate the results. This dataset has been built by randomly choosing 100 English Wikipedia articles, each containing at least one math-tag, and manually determining the correct defining formula for each article.[4] Thus, the gold standard consists of 100 defining formulae and their respective QID of the corresponding article(s). Instead of using the Latex notations for the 100 defining formulae provided by the gold standard dataset, we copied the current Latex notations from the dumps. This decreases the probability that a most common formula does not match an equivalent defining formula simply due to slightly different Latex notations: The notations were found to have changed since the publication of the dataset of [2], e.g. optional brackets were added in the formulae. Thus, this approach ensures better comparability of our results with the results from [2]: While they

[1] The downloaded Wikipedia data dump files were created on 2nd & 3rd of March, 2020, and are available on https://dumps.wikimedia.org/.

[2] https://meta.wikimedia.org/wiki/List_of_Wikipedias. Accessed on 6th of March, 2020.

[3] Note: While the word 'string' typically refers to a formula in this paper, it can also mean an empty string (if no formula exists in the article or the string shared most often across all articles is 'none').

[4] Note: While our definition of a 'formula' means a string containing a formula-indicator, the term 'defining formula' references an arbitrary, possibly empty string in the gold standard. This definition is in accordance with Wikidatas *defining formula* property, which does allow strings without a formula-indicator. Thus, it is obvious that our filtering approach cannot find the four defining formulae without a formula-indicator (e.g. $\pi \int_a^b [R(x)]^2 \, \mathrm{d}x$ is the defining formula of the article about Disc integration (Q3825524)).

manually confirmed if each extracted formula visually[5] matches the defining formula - an approach that does not depend on exactly matching Latex notations - we automatically check for matching strings. To ensure that we recognize most formulae that visually match their defining formula as a true positive, we check if they are similar: Two mathematical expressions are considered similar if they only differ due to whitespaces, irrelevant characters at the end (like a comma or dot that are part of the sentence surrounding the formula) or optional brackets around a sub- or superscript. These factors were found to be the cause for most different, despite visually matching formulae in a small manual investigation. We rectified the entry for 'plastic number' in the gold standard dataset by using $\rho = \sqrt[3]{\frac{9+\sqrt{69}}{18}} + \sqrt[3]{\frac{9-\sqrt{69}}{18}}$ instead of an empty string (no defining formula).

We classified a result as relevant if and only if its defining formula is not an empty string and is included in (at least) one of the articles of the corresponding QID. To make sure we correctly identify relevant results as such, a defining formula is considered 'included' in an article if (at least) one mathematical expression is similar to it. If a result is relevant and gets retrieved, i.e. the most common formula is the defining formula, it is counted as a true positive (TP). If a result is relevant, but the defining formula is not retrieved, this is classified as a false negative (FN). Non-relevant results are counted as true negatives (TN) if the most common string matches the defining formula, otherwise as false positives (FP). These definitions are in accordance with Schubotz et al. in order to ensure the comparability of the results.

We first investigate the results of our approach of counting the occurrences of formulae as well as a combined approach that also considers the order of the formulae in the articles. Afterwards, we inspect the findings of the combined approach with regard to the number of Wikipedia languages used. When filtering only a number of all 309 Wikipedia languages, we choose to filter the biggest language (English) as well as the biggest five and 20 Wikipedias, while excluding Cebuano and Waray-Waray, since both have a high number of bot-generated articles[6] and low number of community members (see footnote 2).[7] We determine the size of the Wikipedias by the number of articles in its respective language according to a list of all Wikipedias (see footnote 2). Afterwards we use our definition of similarity to determine the most common formula and investigate the number of true positives.

[5] Two mathematical expressions are considered visually matching if the expressions generated from the (possibly different) Latex notations look the same, e.g. x_i and $x\_\{i\}$ generate the same expression.

[6] https://stats.wikimedia.org/EN/BotActivityMatrixCreates.htm. Accessed on 22nd of March, 2020.

[7] As it turns out, this measure was unnecessary since neither language included an article corresponding to one of the 100 QIDs.

3 Evaluation

As a first, simple approach we filter one, five, 20, and all 309 Wikipedias while counting the number of articles a formula occurs in. If more than one formula is the most common one, the extracted formula is chosen randomly among them. The results show that while we do get better results by using more Wikipedias, the number of false results is always higher than 70 (mostly due to FN), irrespective of the number of Wikipedias used, and thus too high for this approach to reliably work. An investigation of the results when using 309 languages shows that more than half of the most common strings only occur in one or two languages, thus 54 most common strings have at least one other string with the same number of occurrences. Consequently, ∼80% of those are falsely identified — in comparison to ∼61% for the more common formulae. This shows an obvious problem in the data: A lot of strings only reoccur very rarely across articles, mostly because they occur in a similar mathematical form or depend on a different Latex notation to generate a visually equivalent formula. Before trying to solve this problem by recognizing similar formulae when determining the most common formula, we will focus on another point: Randomly choosing the extracted string among multiple most common strings is a simple but unsophisticated approach. Instead, we now use the order of the formulae as a measure in case two formulae have the same number of occurrences.

As it turns out, this allows us to easily reproduce the findings of Schubotz et al. when using English as the sole language to filter: Since we only count every formula in an article once, we essentially disregard the occurrences of the formulae when using only one language; instead only the order of the formulae will be taken into account, as is the case in [2].

The results in Table 1 reveal that we find nine TP less, while getting five FP and six FN more than Schubotz et al. The higher number of FN is in about four cases attributed to the fact that we — in contrast to Schubotz et al. — automatized the comparison of the extracted formulae with the defining formulae. As a consequence, four extracted formulae could not be identified as equal to their visually equivalent defining formula since they were not similar. The remaining five missing TP are probably attributable to the time-conditioned changes of the Wikipedia articles since the publication date of [2]: The gold standard depends on the defining formulae that were based on mathematical expressions of former Wikipedia sites. As such, today some Wikipedia articles only include a mathematically equivalent, but different formulation, which does not match our defining formula, e.g. $a = b$ and $b = a$. Thus, such a result is falsely recognized as 'non-relevant' and classified as a FP instead of TP.

Altogether, we can verify the findings of Schubotz et al. The investigation of the results revealed that an automated classification of results in 'relevant' and 'non-relevant' is not perfectly accomplishable with the current approach and a more sophisticated method is needed to determine if a formula matches its defining formula.

Table 1. Contingency table comparison of (a) our results and (b) the findings from Schubotz et al. when using one Wikipedia language (English)

(a)

	relevant	not relevant
retrieved	62 (TP)	22 (FP)
not retrieved	16 (FN)	0 (TN)

(b)

	relevant	not relevant
retrieved	71 (TP)	17 (FP)
not retrieved	10 (FN)	2 (TN)

Next, we take a look at the impact the combination of the approach of Schubotz et al. with our approach has on the results when using more than one language. The results in Fig. 2 show that we do get less TP as the number of languages increases and that we get the best results with one language. In other words, as the influence of the order of the formulae gets smaller and, consequently, as the influence of the reoccurrences of formulae gets bigger, the results worsen. This suggests that the order is significantly more important and thus, that the approach of using only the order of the formulae is most probably better than only choosing the most common formula. Note, however, that we cannot verify this: A direct comparison of both methods is not possible as the method of simply counting the reoccurrences always needs an accompanying measure in case multiple most common strings exist. While we could generate an arbitrary dataset such that in no case multiple most common strings exist — simply by excluding all QIDs whose article(s) contain more than one most common string — such a dataset would probably be biased: The number of most common strings might correlate with the number of formulae in the article and thus the length and quality of the article, consequently influencing the results. This was not further investigated.

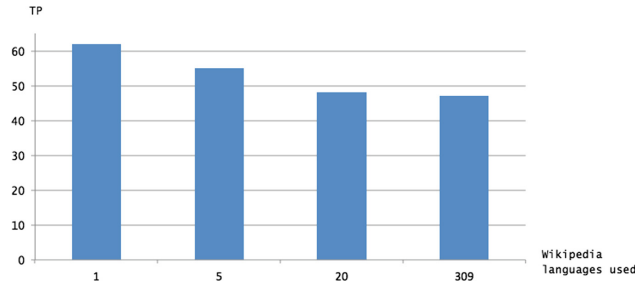

Fig. 2. Number of true positives (TP) depending on the number of Wikipedia languages used.

In the following, we examine the impact that checking for similarity has on the results when we check this not only when comparing a formula and a defining formula as before, but also when determining the most common formula. This allows us to recognize ∼13% of the formulae as similar to another formula found

and thus increases our number of occurrences per formula. Figure 3 shows that the number of TP negatively correlates with the number of languages used, as is the case in the last approach (see Fig. 2). Contrary to our initial expectations, the current approach could not improve our results compared to Fig. 2. The reason is most probably that the average number of articles containing the most common string increased from 4.0 to ∼4.7 (for 309 languages), thus the number of cases where only one most common string exists increased from 46 to 54. As a consequence, in eight fewer cases the order of the formulae is considered. This is another indicator that, as the impact of the occurrences of formulae on our results gets bigger, our results worsen.

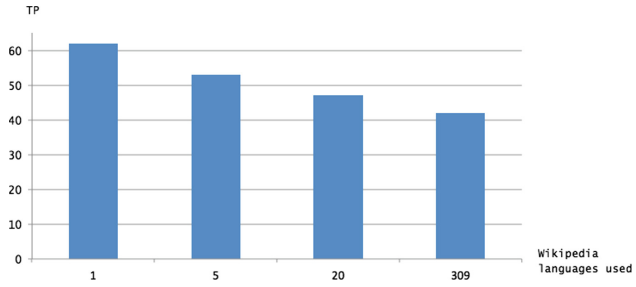

Fig. 3. Number of true positives (TP) depending on the number of Wikipedia languages used when checking for similarity in every comparison between strings.

4 Future Work

The current approach only takes the number of occurrences of a formula and the order of the formulae into consideration, which leaves out a lot of information like the quality of the article, the formula-indicator used in each formula or whether a formula is visually highlighted by placing it in a separate line. Thus, we propose a score-based system using all the information to determine the defining formula more accurately. The information should also include the number of occurrences of a formula, even though it might not improve the results as seen in this project. It is still believed that knowing how often a formula occurs across multiple articles is important information that can improve the detection-rate if used correctly. As the investigation shows, it cannot be the only information used in conjunction with the order of the formulae, although no advanced techniques like unification [3] were used to verify more similar formulae as actually being similar, which might better the results, although the results indicate otherwise. We suggest that in this proposed approach the occurrences of formulae should probably not be weighted heavily as this might negatively impact the results considering our findings.

To build the proposed score, it is necessary to find an optimal weighting of the different pieces of information. To do so, a bigger dataset is needed. We suggest to use Wikidata, which contains a manual assignment of the QIDs of more than 4,300 Wikipedia articles to their respective defining formula.

5 Conclusion

Our findings verify the results of Schubotz et al. who extracted the first formula of a Wikipedia article as an approach to obtain the defining formula related to an article. Nevertheless, it was not possible to achieve the same amount of true positives as Schubotz et al., most probably due to the lack of advanced techniques used to determine whether two formulae are equivalent.

Furthermore, our results were negatively impacted when considering the order of the formulae in their respective article together with the number of languages it occurs in. This suggests that the order of the formulae is a much more important indicator to determine the defining formulae than the number of its occurrences across multiple languages. Thus, reducing the influence the order has on the results in favor of the number of occurrences decreases the number of extracted defining formulae. This assumption is further supported by the fact that the number of true positives negatively correlates with the number of Wikipedia languages used, which in turn influences the number of languages a formula occurs in. Furthermore, when we determine the most common formula by regarding formulae as equal if they match our definition of being similar, the number of true positives further decreases. This is reasoned to be another indication that the number of occurrences of a string across articles is a bad factor for determining the most common formula. Consequently, other indicators are proposed that should be able to improve the current approach. It is worth including the number of occurrences across articles as one of the factors, as it cannot be said with certainty that the number of occurrences is an inherently bad indicator. It might be possible that much more sophisticated measures are needed to determine if two formulae are similar, though our findings suggest otherwise. Improving the results of the current approach will be a focus of future work.

Acknowledgments. The project is based on our contribution to the seminar 'Selected Topics in Data Science' from the Data and Knowledge Engineering group of the University of Wuppertal headed by Bela Gipp [ORCID: 0000 − 0001 − 6522 − 3019]. The author thanks his seminar-advisor Moritz Schubotz [ORCID: 0000−0001−7141−4997].

References

1. Schubotz, M., et al.: Evaluation of similarity-measure factors for formulae based on the NTCIR-11 math task. In: Kando, N., Joho, H., Kishida, K. (eds.) Evaluation of Similarity-Measure Factors for Formulae. Proceedings of the NTCIR National Institute of Informatics (NII) (2014)
2. Schubotz, M., et al.: Introducing MathQA – a math-aware question answering system. Inf. Discov. Deliv. **46**(4), 214–224 (2018). https://doi.org/10.1108/IDD-06-2018-0022

3. Sojka, P., Ruzicka, M., Novotný, V.: MIaS: math-aware retrieval in digital mathematical libraries. In: Cuzzocrea, A., et al. (eds.) Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, 22–26 October 2018, pp. 1923–1926. ACM (2018). https://doi.org/10.1145/3269206.3269233
4. Zhang, Q., Youssef, A.: An approach to math-similarity search. In: Watt, S.M., Davenport, J.H., Sexton, A.P., Sojka, P., Urban, J. (eds.) CICM 2014. LNCS (LNAI), vol. 8543, pp. 404–418. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-08434-3_29