Analyzing Formula Concepts and Patterns to Improve Literature Exploration and Recommendation for STEM Documents

Philipp Scharpf
Department of Computer and Information Science
University of Konstanz, Germany
philipp.scharpf@uni-konstanz.de

ABSTRACT

In my dissertation, I will explore the following hypothesis: *Considering mathematical formulae using the open knowledge-base Wikidata improves content-based literature exploration and recommendation for STEM documents.*

CCS CONCEPTS

• **Information systems** → *Information retrieval*;

KEYWORDS

Literature Exploration and Recommendation, Semantic Similarity and Relatedness, Named Entity Recognition, Wikidata, Mathematical Information Retrieval, Ontologies and the Semantic Web

ACM Reference Format:
Philipp Scharpf. 2018. Analyzing Formula Concepts and Patterns to Improve Literature Exploration and Recommendation for STEM Documents. In *Proceedings of ACM/IEEE Joint Conference on Digital Libraries in 2018 (JCDL 2018).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Recommender systems (RS) are indispensable tools for filtering out relevant documents from a variety of publications, allowing researchers to discover relevant (similar or novel) content more quickly. This is particularly important for avoiding duplicate or sub-optimal research results that can arise if a researcher is not aware of already existing findings of others. Currently available content-based Recommender Systems rely on semantic similarity of textual elements or citations and do not take into account the mathematical concepts encoded in formulae, thus lacking an essential linkage for STEM documents.

JCDL 2018, June 2018, Fort Worth, TX

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM. . . $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

To improve the quality of content-based recommendations for STEM documents, I propose a method of discovery (definition) and recognition (identification) of formula concepts (statement) and formula patterns (sequence), which I will develop using example documents from mathematics and physics. For the evaluation of my hypothesis, I will semantically serialize a given STEM document by matching items (QIDs) from the open knowledge-base Wikidata to textual and mathematical concepts (named entity and formula recognition). This yields semantic markers (figure 1) that, like citations, do not depend on the specific language [12] and form of the document. Analyzing these *Wikidata markers* (WDM), I will be able to investigate four edge cases by including 1) only text occurrence, 2) text sequence, 3) text and formula occurrence and 4) text and formula occurrence and sequence in the document comparison. I will evaluate how stepwise including the individual cases improves the performance of the literature Recommender System *Mr. DLib*.

Concepts (occurrence)	Patterns (sequence)
1) text	2) text
3) text + formulae	4) text + formulae

Evaluation edge cases for Recommender Systems.

Furthermore, I will include an individual weighting option on specific WDM made by the user who will precedingly be enabled to assess the correctness of the WDM recognition in a feedback loop. Finally, the efficiency and effectiveness of my approach will be compared to other existing types of Recommender Systems, reviewed in [5].

Motivation

My personal motivation for a Ph.D. in information science was that during my research in theoretical physics I recognized the growing challenge of keeping track of all relevant publications, even if the subject area is very small and specific. Therefore, I consider it very important to use the available processing capabilities of computers to automatically assess the relevance and importance of newly published documents. Analysis of similarity of semantic content is indispensable for Recommender Systems for scientific documents. In the STEM subjects (Science, Technology, Mathematics and Engineering), however, it is necessary to include the essential mathematical content, which is mainly encoded in formulae. An understanding of the semantics is thus only possible if text and formulae are jointly processed. The research of my dissertation will focus on Recommender Systems for scientific publications and patents in the STEM disciplines.

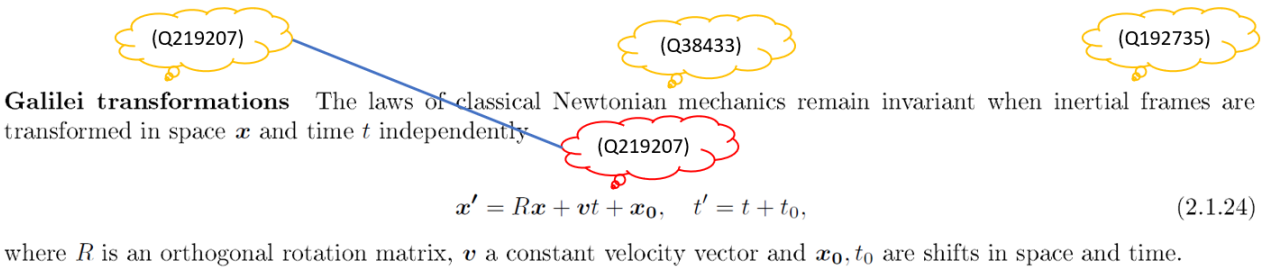

Figure 1: Example of *Wikidata markers* (WDM), annotating a short section of my master's thesis.

To facilitate the automated analysis of the semantic content by the computer in the STEM disciplines and thereby decisively improve Recommender Systems for STEM documents, I consider the study of the semantics of formulae as a crucial research topic, to which, in my dissertation, I would like to make a significant contribution. As a physicist, I see the great benefits of computer-aided analysis of formula concepts and patterns with a variety of additional applications for the STEM disciplines, i.e. Search Engines, Plagiarism Detection and Novelty Detection.

Problem and Vision

The starting point of my dissertation are the following shortcomings in the analysis of STEM documents: Currently available content-based Recommender Systems rely on semantic similarity of text elements or citations and do not take into account the mathematical concepts encoded in formulae, thus lacking essential linkage for STEM documents. To tackle this problem, I propose a new approach, the analysis of Formula Concepts and Patterns to build Formula Based Recommender Systems (FbRS) for STEM documents.
I understand here

- *Formula Concept*: Semantic content of a mathematical formula, represented by the name (if established) of the given equation or mathematical statement;
- *Formula Pattern*: Sequence of several mathematical formulae in a STEM document, which has a meaning (for example, the formula pattern represents a mathematical derivation or proof as a whole).

The following hypothesis will be reviewed: *The consideration of mathematical Formulae Concepts and Patterns improves content-based recommendations for STEM documents.*

2 RELATED WORK

At this point, the relevance of my approach to analyze formula concepts and patterns will be clarified by explaining, more detailed, the problems of currently available recommender systems, and the current state of research in Mathematical Information Retrieval (MathIR). The focus of my research will not be in recommender systems, but in MathIR - developing a method of Formula Concept Discovery and Recognition which will be described in section 3.

Recommender Systems

Recommender systems are an indispensable tool for filtering out relevant documents from a variety of published papers, allowing them to discover similar or novel content more quickly. This is particularly important for avoiding duplicate or suboptimal research results that can arise if a researcher does not learn about the already existing findings of others. Currently available recommender systems can be divided into two categories: The use of *content-based* or *user-based* methods for the calculation of similar, i.e., for the user relevant literature. In the *user-based* approach, recommendations are given on the basis of matching elements, *similar users* are interested in (*collaborate filtering*). Their calculation relies on ratings, downloads, behavioral and demographic characteristics, and shared interests. However, there are some problems that can become weaknesses of the system. Firstly, there may be a lack of diversity, for example when academic interests are very specialized, and there are not enough users with similar interests, which means that the system is not sufficiently trained. Secondly, it is often a major obstacle convincing scientists who are looking for suitable literature for their research to create an additional user account, just to generate recommendations that have to be superior to simple searches. In the simplest case, *content-based* methods are limited to text string matching, which is only effective as long as there are sufficient textual matches of the documents. In the case of STEM documents, the method is often inadequate, since the semantic information contained in formulae, tables and figures is thereby ignored. This may cause the system to find no similarity, although the various contents actually hold one that would be relevant to the user. There has been a lot of research on Recommender Systems for scientific papers. B. Gipp, J. Beel and C. Hentschel introduced *Scienstein: A Research Paper Recommender System* in 2009 [7, 8] and *Mr. DLib, a Machine-readable Digital Library* in 2011 [5, 6, 9, 10] with an integration into the reference manager JabRef done by S. Feyer et al. in 2017 [11]. I will collaborate with Jöran Beel (Trinity College Dublin) and test how the performance of the existing Recommender System *Mr. DLib* http://mr-dlib.org/ can be improved by the consideration of mathematical formulae.

Mathematical Information Retrieval (MathIR)

A large number of academic documents, almost all in the STEM disciplines, contain, in addition to citations, a crucial text-independent feature: mathematical formulae. Due to the links between text and formulae, automated access to capture the mathematical content of a STEM document is often very difficult. First of all, these parts of the document must be identified and labeled and, at best, directly contextualized by their relationships to the surrounding text and the section they are contained with. Since many documents are available in PDF format only and not as source code in LaTeX or annotated in meaning and context in the Mathematical Markup Language (MathML) [22], the conversion may already be a difficult task currently being researched [20]. Even if they have been successfully extracted, the study is particularly confronted with the problem of disambiguation, as there is only a limited number of characters (mainly Latin and Greek letters) representing a theoretically unlimited number of semantic content (various mathematical variables or physical quantities). A semantic enrichment is required to markup the formula parts analogous to the model of *Part-Of-Speech (POS) Tagging* from *Natural Language Processing (NLP)*. In a contribution to SIGIR 2016 [4], my group (M. Schubotz, N. Meuschke, B. Gipp et al.) proposed a new approach coining the term *Mathematical Language Processing (MLP)*, transferring NLP methods to math-specific processing and semantic enrichment of scientific texts on the level of mathematical concepts. For this purpose, the formulae are broken down into their structural components by means of the MathML language, i.e. they are tokenized (variables, operators, numbers etc.), typified (equation, term, definition, proof, etc.) and disambiguated (e.g. function f(a+b) vs. multiplication f ·(a+b)).

3 APPROACH

Research steps. Currently, I am working on comprehensive literature review of the state-of-the-art in Mathematical Information Retrieval (MathIR). For Recommender Systems, I refer to [5].

My research approach will be divided into the following working packages:

0. Selection of Data: Select suitable STEM documents from mathematics and physics containing content which I can judge.

1. Named Entity Recognition: Link text concept terms to Wikidata items using already available methods of named entity recognition.

2. Formula Concept Discovery: Retrieve mathematical formula concepts from the LaTeX source code of STEM documents or Wikipedia articles and create Wikidata items including names and characters of the occurring identifiers.

3. Formula Concept Recognition: Identify formulae in STEM documents or Wikipedia articles as existing Wikidata formula concept items.

4. Formula Pattern Discovery: Examine formula sequences in STEM documents to discover mathematical argumentation, especially proofs of mathematical theorems.

5. Formula Pattern Recognition: Identify mathematical argumentation (proofs of mathematical theorems) in STEM documents. Evaluate how the consideration of formula patterns improves the performance of the Recommender System.

6. Implementation in Recommender System: Use Mr. DLib http://mr-dlib.org/ to process the WDM annotated STEM LaTeX documents and evaluate the improvement of the formula inclusion in recommendation generation.

7. Evaluation: Comparison of different weighting schemes (text, formulae) and PD/RS approaches (collaborate-filtering etc.), evaluating the improvement of the Recommender System by formula consideration.

Selection of Data

Find suitable STEM documents. First of all, it will be necessary as preliminary work to find suitable documents for the respective work packages from various publicly available data servers. For the fundamental study of formula syntax and semantics, and the formula concepts and patterns, I will focus on examples from mathematics and physics retrieved from the ArXiV repository of electronic preprints (e-prints) [19]. An outstanding advantage of the ArXiV is that the LaTeX source code is provided for a large number of documents. As I develop methods to discover and recognize formula concepts and patterns, I will find experts to assess the performance of the Recommender System (preferably mathematicians, computer scientists, physicists and engineers). Having obtained a master's degree in physics, I consider myself suitable for finding instructive examples of formula concepts (occurrences) and formula patterns (order) to investigate my research questions. In addition, I will make use of the contacts of my group (Moritz Schubotz), especially Howard Cohl, Michael Kohlhase, Akiko Aizawa and Richard Zanibbi to exchange datasets and ideas and discuss issues.

Named Entity Recognition

Having chosen a suitable subset of STEM documents from the ArXiV, I can begin with the major task of annotating the LaTeX code. I will start with the textual named entities, i.e. linking text concept terms to Wikidata items. For this I can build on and exploit already existing research and code, e.g., [13–15].

Formula Concept Retrieval

Observation. My planned in-depth research on Formulae Concept Retrieval is motivated by the following observation: Consider for example different equations of physics (e.g., Klein-Gordon equation, Planck's law of radiation, Schroedinger's equation, Einstein's field equations, Lorentz force, etc.). It is striking that a particular formula appears in a variety of different but equivalent representations - in numerous publications, it is at least slightly different from the rest.

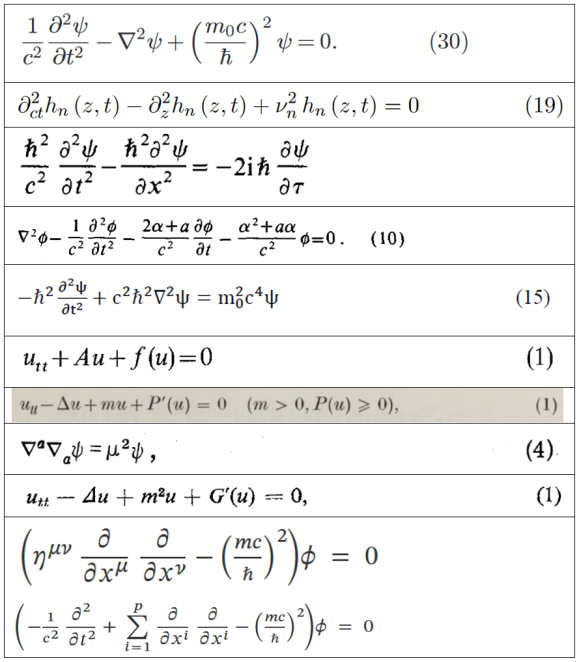

Figure 2: Various representations of the Klein-Gordon equation from arbitrarily selected sources.

As an illustration, Figure 2 shows some representations of the Klein-Gordon equation that seem very different at first glance but actually represent the same mathematical concept. Moreover, many of these formulae are not machine-readable, much less machine-interpretable.

Solution. I am striving to develop a method that will be able to map all of these formulae to the same concept, that is in this case *Klein-Gordon equation*[1].

For this purpose, a given equation must be broken down into its structural components, the identifiers which have to be recognized individually and assembled to identify the target concept.

Figure 3 and 4 each show a first own suggestion for a scheme and process of a Formula Concept Discovery (FCD) and Formula Concept Recognition (FCR) - as a starting point of my research.

The goal of *Formula Concept Discovery* (FCD) is to develop an understanding of the definition of a mathematical concept by scanning Wikipedia articles or STEM documents from the ArXiV database for recurrences of a particular defining formula for a mathematical concept. Often this is simply the article's first occurring formula, which however cannot be reliably generalized, so far leaving a human judgment to be unavoidable. Thus, my research here focuses on finding intelligent ways for automatic retrieval of the defining formula. Once a defining formula is discovered, a concept item can be created (if not already existing) in Wikidata,

[1]In particular, linkable to the Wikidata entry: https://www.wikidata.org/wiki/Q868967 (July 25, 2018).

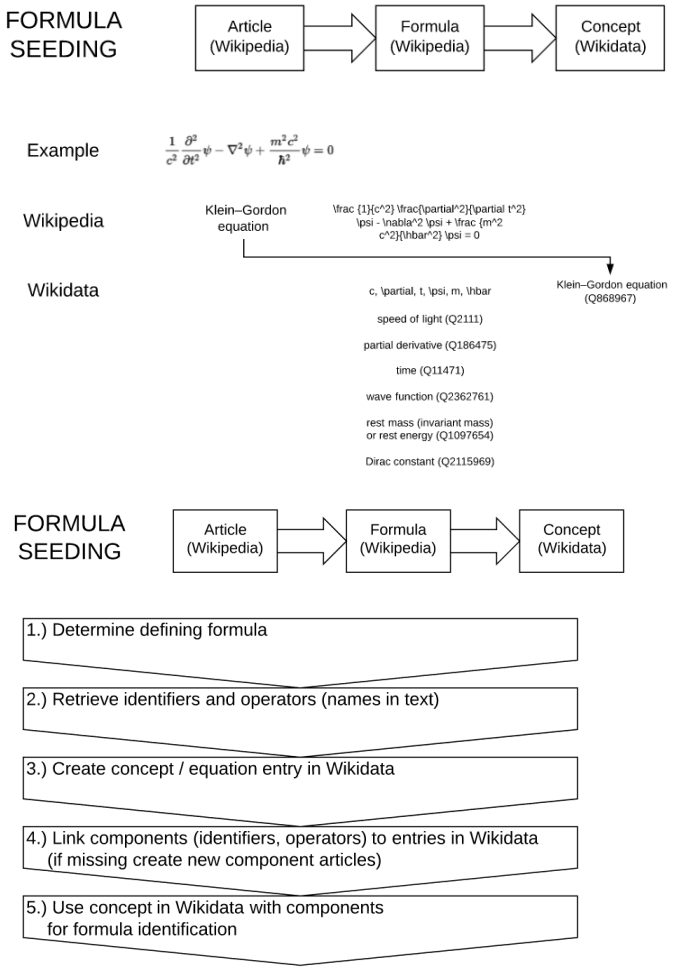

Figure 3: My suggestion for a scheme (above) and a process (below) of the Formula Concept Discovery (FCD) for seeding a formula into the database *Wikidata*.

whereby the individual components of the formula (identifiers, operators) must, in turn, be linked to known or new Wikidata items. The goal of *Formula Concept Recognition* (FCR) is then to identify a formula that occurs in a LaTeX STEM document by comparing its semantic components (i.e., the constituting identifiers) with potential contents of Wikidata items. I suggest a percentage recognition measure based on the number of matching components. This will often deviate from 100 % due to equivalence transformations where the identifiers change or include others (additional syntax level). My research is therefore here concerned with the optimization of the recognition process, for which possible changes in presentation should be analyzed to finally ensure a largely representation-independent identification of formula concepts.

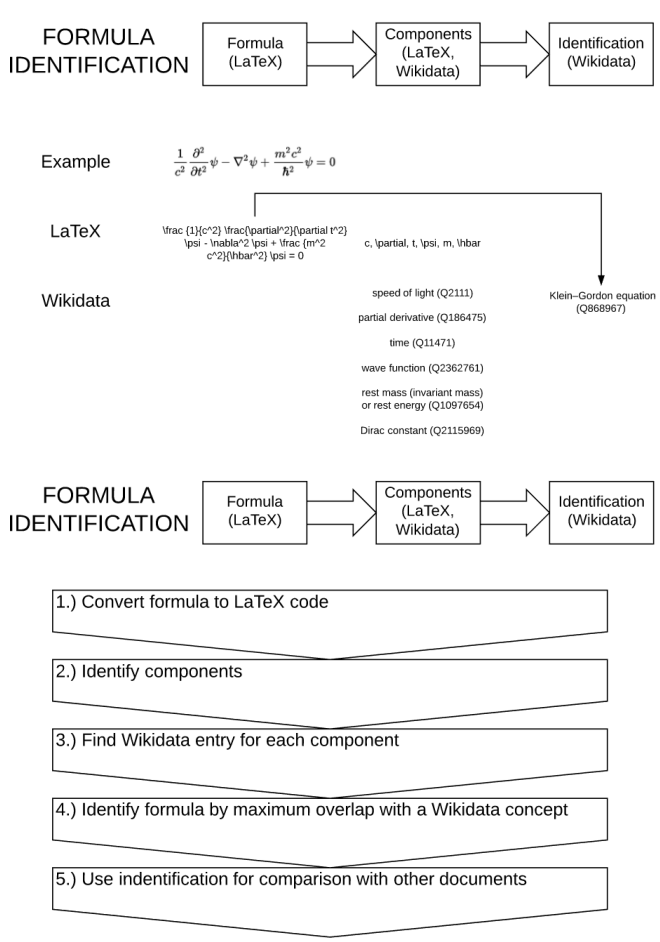

Figure 4: My suggestion for a scheme (above) and a process (below) of the Formula Concept Recognition (FCR) for retrieving a formula from the database *Wikidata*.

Research goals.

- Development and optimization of Formula Concept Discovery for Wikipedia articles and STEM documents from the ArXiV.
- Seeding a scalable large number to Wikidata with suitable annotation of the formula parts (identifiers).
- Development and optimization of Formula Concept Recognition techniques that match a given formula to a Wikidata concept item, independently from its specific representation.

Formula Concept Discovery (FCD)

As stated before, Formula Concept Discovery aims to retrieve mathematical formula concepts from ArXiV documents or Wikipedia articles and create Wikidata items including names and characters of the occurring identifiers. Before being able to seed a formula concept into Wikidata, preprocessing of the ArXiV formula has to be done.

Analysis of formula syntax. If the LaTeX source code of a STEM document is available, an analysis of the formula syntax has to be performed as preparation for all further examinations. First, there is a differentiation between mathematical and non-mathematical content, so the formulae are separated and extracted from the text. Subsequently, the complex mathematical expressions are tokenized, i.e., the formulae are decomposed into their components (identifiers, operators, numbers, etc.). A rough structural analysis (Part-Of-Math-tagging) should be carried out in advance: A formula consists of different terms, which must be clearly distinguished from each other. The Klein-Gordon equation from quantum physics, again used as an instructive example,

$$\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\psi - \nabla^2\psi + \frac{m^2c^2}{\hbar^2}\psi = 0$$

contains a term $\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\psi$ with a double time derivative, one with a double space derivative $\nabla^2\psi$ as well one with a constant prefactor $\frac{m^2c^2}{\hbar^2}\psi$. The first term can then be further decomposed into its characters (tokens), that is, the denominator c for the speed of light, the operator $\frac{\partial}{t}$ with an exponent (number) 2 and the identifier ψ for the physical (quantum) wave function. The structure decomposition is important for the subsequent semantic analysis of the components for Formula Concept Discovery and Recognition.

To analyze the formula syntax and semantics, M. Schubotz et al. have done important preparatory work, especially [1, 2] distinguishing within the functionality of the MathML markup-language between *presentation*, *content* and *semantics* of a formula.

Analysis of formula semantics. To grasp the meaning of a formula and its components, a typification of the mathematical statement is required in advance. It has to be determined whether it is, e.g., a definition of a variable, a theorem or part of a mathematical proof? While in the first case, a pure analysis of the formula concept is sufficient, in the second case an additional analysis of the formula patterns occurring in the entire proof is needed to capture the formula semantics. At this point, we are confronted with the problem of the identifier ambiguity, which requires disambiguation with the help of the partial clarifications available in the text that is a combination of mathematical and linguistic tokens. A single identifier faces a theoretically unlimited number of possible meanings. For example, E in physics often refers to both an energy and an electric field, generally mathematically an expected value, and so on. Schubotz, Kraemer, Meuschke et al. recently [3] showed that a new machine learning approach combining previously successful methods was able to increase the precision of the identifier definition extraction from 17.85% to 48.60%. The recall also increased from 22.85% to 28.06%. While still being far from satisfactory, the approach is promising with great development potential, and the following steps in my work will benefit significantly from an improvement.

Analysis of formula concepts. This is where the main challenge of my work begins. Once the structure of a formula has been determined, broken down into its components, and the meaning of each identifier is known after a disambiguation using the explanations in the text, a possible assignment to an abstract mathematical concept can be explored. For this purpose, a database is needed, with the help of which the components can be completely or at least partially identified. As already stated, I will use the free knowledge database Wikidata [23]. The reason for my choice is that Wikidata is free, open and can be read and edited by humans and machines. Before a successful Formula Concept Recognition can be performed by *fetching* from the database, a previous Formula Concept Discovery as *seeding* into the database is required.

Formula Clustering. As an essential part of Formula Concept Discovery, I will cluster formulae to have candidates for mathematical concepts that can be created as Wikidata items. This will serve as a preparation for a final global formula-based clustering of STEM documents as a whole to provide Wikidata item labeled documents. The annotated (WDM labeled) documents will be the basis for the evaluation of the improvement of recommendation generation by considering formulae. I will use k-means algorithms to cluster formulae by their syntax and semantics respectively (feature vectors containing information about the parse tree and constituent names). If a formula is discovered numerous times (above a defined threshold) in different documents (possibly in slightly different notation), it is considered to be an important mathematical concept that is frequently used. The retrieved formula concepts will be subsequently seeded into Wikidata, i.e., numerous items with the name of the respective concept and its defining formula will be created. I will compare and discuss several possible seeding methods (different possibilities to include the formula identifier parts).

Formula Concept Recognition (FCR)

Identify formulae in LaTeX documents or Wikipedia articles as Wikidata formula concept items.

My goal here is to introduce a measure of similarity that allows assigning a formula to a mathematical concept (equation) if it exceeds a defined threshold. A first rough approach would be

$$\text{MS (matching score)} = \#\text{ recognized elements} \,/\, \#\text{ total elements} \quad .$$

To successfully identify a single element, for example, the Laplace operator $\nabla^2 = \Delta$, it must be assigned to the corresponding concept in Wikidata. The Laplace operator, for example, can be retrieved at https://www.wikidata.org/wiki/Q203484 (July 25, 2018), so as request QID *Q203484*. My formula recognition project is designed to motivate active users of Wikidata to gradually build a hierarchical structure of the formula elements, assign elements to all available formulae (property *has part*) and create new items for formulae concepts directly including the parts. I will compare and discuss several possible formula recognition methods (e.g., simple TeX string search vs. parts identification, recognition by identifier name, symbol and value). I will represent formulae in a *bag-of-words* vector for the occurring identifiers and operators respectively where also the order of occurrence can be considered.

Formula Pattern Discovery

Formula Pattern Discovery aims to examine formula sequences in STEM documents to discover mathematical argumentation, especially proofs of mathematical theorems. Once one or several formulae have been typed as a (or at least part of a) mathematical proof, it is necessary to analyze the position and order of the individual formulae within the proof. This makes it theoretically possible to identify a proof, i.e., to assign the formula pattern to a formula proof concept. In my research I will examine if and how the consideration of the formula patterns (formula concept sequence) improves the recommendation generation.

Formula Pattern Recognition

Formula Pattern Recognition aims to identify mathematical argumentation (proofs of mathematical theorems) in STEM documents. Subsequently, I will evaluate how the consideration of formula patterns (include sequence of formulae in recommendation generation) improves the performance of the Recommender System.

Implementation in Recommender System

Having a sufficient number of annotated STEM ArXiV documents (around 100.000), I can collaborate with Jöran Beel (Trinity College Dublin) and use the existing Recommender System *Mr. DLib* http://mr-dlib.org/, to process the WDM annotated documents and evaluate the improvement of its performance.

4 EVALUATION

As stated at the beginning, the goal of my Ph.D. thesis is to review the following hypothesis that I have prepared: *Considering mathematical formulae improves content-based literature exploration and recommendation for STEM documents.* An evaluation of my hypothesis should be based on Information Retrieval (IR) State of the Art statistical quality criteria for the binary classification of documents that may need to be adapted to the specific needs of my thesis in the course of my research. I define the classification for the Recommender System as *recommendation helpful* = positive test result, *recommendation not helpful* = negative test result, while checking whether a recommendation is helpful is done by experts. This is used to evaluate the correct and false classification rates, i.e. *precision*, *recall*, as well as the combined harmonic mean *F-value*. Subsequently, I compare the results of the recommendation generation, in which the semantic similarity was determined purely on the basis of text and citations, to one with additional consideration of formula concepts and patterns to assess my above-stated hypothesis.

Beforehand, the underlying methods of Formula Concept Discovery and Recognition will be evaluated respectively.

5 PRELIMINARY RESEARCH

Starting in November 2017, I already made some advancements towards the essential Formula Concept Discovery by helping M. Schubotz and A. Greiner-Petter to construct a MathML Gold Standard of Wikidata annotated formulae. The benchmark *MathMLben* was used to evaluate the performance of conversion tools from LaTeX code via MathML to a selection of Computer Algebra Systems (CAS). The paper *Improving the Representation and Conversion of Mathematical Formulae by Considering their Textual Context* was submitted to and accepted by the JCDL 2018. Currently, I am working with M. Schubotz on a publication of the first mathematical Question Answering System using Wikidata. To have enough items with a defining formula that can be queried, we first had to import data from Wikipedia and seed it into Wikidata. Our research is a starting point for the development of effective methods to automatically seed Wikidata with mathematical formulae from Wikipedia articles or ArXiV documents. The paper *Introducing MathQA - a Math-Aware Question Answering System* will be submitted to the CICM 2018.

6 CONCLUSION AND OUTLOOK

The aim of my dissertation is to improve Recommender Systems for STEM documents by considering an analysis of the occurring formula concepts and patterns that are an essential part of mathematical literature.

Having developed effective methods for Formula Concept Discovery and Recognition to automatically annotate STEM documents, numerous applications are promising. First of all, it will be a decisive step towards a semantification of STEM literature for the Semantic Web [16]. A semantic annotation by Wikidata markers will facilitate grasping the gist of a document by identifying a list of concepts or prospectively even statements in the form of RDF triples [24]. The next step is then to construct a Wikidata OWL ontology from a given STEM document (= Ontology Learning [17]) and visualize the semantic relations of the occurring Wikidata items. This can be used to compare the extracted text ontology (claim) with a current Wikidata ontology (state) of a given knowledge topic to verify the alleged semantic relations and assess the novelty of their insights. Elizarov et al. recently proposed a mathematical knowledge management technology *OntoMath* to create a World Digital Mathematical Library in OWL form from ArXiV documents [18]. Being able to automatically or at least semi-automatically extract (Wikidata) ontology statements is definitely a challenging research goal that is worth pursuing for the next decades.

ACKNOWLEDGMENTS

The author would like to thank Bela Gipp and Moritz Schubotz for their support.

REFERENCES

Books

[1] M. Schubotz, *Augmenting mathematical formulae for more effective querying & efficient presentation*. Verlag epubli, 2017. https://www.amazon.de/Augmenting-Mathematical-Effective-Efficient-Presentation/dp/3745062086/ref=sr_1_1?ie=UTF8&qid=1512131844&sr=8-1&keywords=moritz+schubotz (July 25, 2018)

Publications

[2] M. Schubotz, N. Meuschke, T. Hepp. H. S. Cohl, B. Gipp. *VMEXT: A Visualization Tool for Mathematical Expression Trees*. In Proceedings Conference on Intelligent Computer Mathematics (CICM), 2017.

[3] M. Schubotz, L. Kraemer, N. Meuschke, F. Hamborg, and B. Gipp. *Evaluating and Improving the Extraction of Mathematical Identifier Definitions*. In Proceedings of the 8th International Conference of the CLEF Association (CLEF), 2017.

[4] M. Schubotz, N. Meuschke, B. Gipp et al. *Semantification of Identifiers in Mathematics for Better Math Information Retrieval*. In Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. Full Paper, 2016.

[5] J. Beel, B. Gipp, S. Langer, and C. Breitinger. *Research-paper recommender systems: a literature survey*. International Journal on Digital Libraries, pp. 1-34, 2015.

[6] J. Beel, S. Langer, M. Genzmehr, B. Gipp, C. Breitinger, and A. Nuernberger. *Research Paper Recommender System Evaluation: A Quantitative Literature Survey*. In Proceedings of the Workshop on Reproducibility and Replication in Recommender Systems Evaluation (RepSys) at the ACM Recommender System Conference (RecSys), 2013.

[7] B. Gipp, J. Beel, and C. Hentschel. *Scienstein: A Research Paper Recommender System*. In Proceedings of the International Conference on Emerging Trends in Computing (ICETiC'09), Virudhunagar, India, 2009.

[8] B. Gipp and J. Beel. *Identifying Related Documents For Research Paper Recommender By CPA And COA*. In Proceedings of The World Congress on Engineering and Computer Science 2009, Berkeley, USA, 2009.

[9] J. Beel et al. *Introducing Mr. DLib, a Machine-readable Digital Library*. In Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries. ACM, 2011.

[10] J. Beel et al. *Mr. DLib: Recommendations-as-a-Service (RaaS) for Academia*. Digital Libraries (JCDL), 2017 ACM/IEEE Joint Conference on. IEEE, 2017.

[11] S. Feyer et al. *Integration of the Scientific Recommender System Mr. DLib into the Reference Manager JabRef*. European Conference on Information Retrieval. Springer, Cham, 2017.

[12] B. Gipp and J. Beel. *Citation Based Plagiarism Detection – A New Approach to Identify Plagiarized Work Language Independently*. Proceedings of the 21st ACM Conference on Hypertext and Hypermedia (HT'10). New York, NY, USA, 2010.

[13] J. Geiß, A. Spitz, and M. Gertz. *NECKAr: A Named Entity Classifier for Wikidata*. International Conference of the German Society for Computational Linguistics and Language Technology. Springer, Cham, 2017.

[14] A. Spitz et al. *State of the Union: A Data Consumer's Perspective on Wikidata and Its Properties for the Classification and Resolution of Entities*. Wiki@ICWSM, 2016.

[15] P. Taufer. *Named Entity Recognition and Linking*. Faculty of Mathematics and Physics, Charles University, Prague Czechia. Master's Thesis, 2017.

[16] T. Berners-Lee, J. Hendler, and O. Lassila. *The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities*. Scientific American 284 (5), 2001.

[17] A. Maedche and S. Staab. *Learning ontologies for the semantic web*. In Proceedings of the Second International Conference on Semantic Web-Volume 40. CEUR-WS.org, 2001.

[18] . A. Elizarov et al. *Digital ecosystem ontomath: Mathematical knowledge analytics and management*. International Conference on Data Analytics and Management in Data Intensive Domains. Springer, Cham, 2016.

Web addresses

[19] ArXiV - e-Print archive, https://arxiv.org/ (July 25, 2018).
[20] InftyReader - OCR Tool zur Formelübersetzung, http://www.inftyreader.org/.
[21] Scikit-learn - Module for Python, http://scikit-learn.org/stable/modules/outlier_detection.html (July 25, 2018).
[22] Mathematical Markup Language (MathML) - W3C Recommendation, https://www.w3.org/TR/MathML3/ (July 25, 2018).
[23] Wikidata - Free and open knowledge base, www.wikidata.org (July 25, 2018).
[24] Resource Description Framework (RDF): Concepts and Abstract Syntax. W3C Recommendation 2004. https://www.w3.org/TR/rdf-concepts/ (July 25, 2018)