Towards a Recommender System for the Choice of UDC Code for Mathematical Articles

Olga Nevzorova [1[0000-0001-8116-9446]] and Damir Almukhametov[1[0000-0002-4888-7937]]

[1] Kazan Federal University, Kremlevskaja str., 18, Kazan, 420008, Russia
[onevzoro, dnlanik]@gmail.com

Abstract. Authors of scientific papers in the field of mathematics usually use the universal decimal classification scheme to search for related articles. UDC is a hierarchical classification scheme that allows librarians and editors to specify one or more codes for publications. Typically, the classification code identifies a subject editor who is responsible for the review process for articles submitted to scientific journals. In this article, we will explore a new approach to assigning UDC code for mathematical work, based on the OntoMathPRO ontology.

This ontology is an applied ontology for the automatic processing of professional mathematical articles in Russian and English. An ontology defines concepts commonly used in mathematics, as well as an evolving and poorly established vocabulary extracted from contemporary scientific articles. OntoMathPRO covers a wide range of areas of mathematics such as number theory, set theory, algebra, analysis, geometry, computation theory, differential equations, numerical analysis, probability theory, and statistics. Each class has a textual explanation, Russian and English inscriptions, including synonyms.

We investigated a set of classification functions, which are presented as ontology concepts, and identified the most relevant ones for constructing code maps of some UDC codes in the field of mathematics. We found that the code maps of the considered UDC codes can be built on the basis of the selected features (method, equation, problem). The values of these features are determined using the OntoMathPRO ontology. The constructed code maps allow for successfully assigning the considered UDC codes for publications.

Keywords: Recommender system, classification of documents, OntoMathPro ontology, Universal Decimal Classification.

1 Introduction

Recommender systems are used in a variety of areas [1]. Recommender systems are classified as content, collaborative, knowledge-based, and hybrid [2]. Collaborative filtering approaches build a model from a user's past behavior. This model is used to predict items (or ratings for items) that the user may have an interest in. Content-based filtering approaches utilize pre-tagged characteristics of an item in order to

recommend additional items with similar properties. Current recommender systems typically combine one or more approaches into a hybrid system.

Knowledge-based recommender systems used in science and education are of particular interest. The classic tasks of such systems are searching related articles, building recommendations for the study of educational topics [3].

In this article, we consider recommender systems focused on publishing and the preparation of scientific publications [4]. Such systems form the digital infrastructure of electronic scientific journals, including a software platform that implements the main workflows for managing an electronic journal, and information systems that support basic and additional services, taking into account, in particular, the specifics of the subject area of this journal [5]. One of the important problems is the classification of articles submitted to the journal.

Classification of documents with the assignment of code-classifiers is a traditional way of systematizing and searching knowledge.

Classifiers are a type of metadata in scientific documents. There are various national and international universal classification systems. It is being widely used in Russia such the classification systems as the universal Library-Bibliographic Classification (LBC), the State Rubricator of Scientific and Technical Information (SRSTI), the Universal Decimal Classification (UDC).

The Universal Decimal Classification (UDC) (www.udc.org/index.html) underlies the systematization of knowledge presented in libraries, databases and other repositories of information. UDC is adopted for indexing scientific and technical documents in most countries of the world. In Russia, the UDC is a mandatory requisite for all book products and information on natural and technical sciences. At the end of 2019, this classifier contains about 126,441 codes. The classification is currently translated into more than 50 languages.

The classification codes selecting is associated with the analysis of the structure of the classifier tree and is quite time consuming. In this article, we consider the problem of automating the selection of the UDC classification code for a mathematical article based on a special resource – the OntoMathPro ontology for professional mathematics.

2 Related Work

The interest in the topic of scholarly text classification and recommendation has grown in recent years. Regarding the classification of scholarly texts according to the UDC [6], texts are classified by peers based on their keywords. Similarly, bibliographic metadata (title, description and subject tags) can be used to equip texts with Dewey decimal classification (DDC) to supplement bibliographic records of publications [7]. The spread of digital resources and their integration into the traditional library environment has created the need for an automated tool that organizes publications into library classification schemes.

A survey of methods, such as content-based, collaborative filtering, graph-based and hybrid methods can be found in the work of Bai et al. [8]. Analysis of the use of

recommendation-as-a-service for academia is presented in the study by Beel et al. [9]. In [10] a comprehensive summary of the state-of-the-art of deep learning based recommender systems is provided. The machine learning methods are used in scientific recommender system in various services [11,12]. In [11] the authors investigate the feasibility of automatically assigning a coarse-grained primary classification using the MSC scheme, by regarding the problem as a multiclass classification machine learning task. In [12], a machine learning model for the automatic classification of old digitized texts from the Slovenian digital library is discussed. The classification of the UDC of new scientific texts, assigned by human specialists, was used to build a classification model of the UDC of old digitized texts. This model uses various clustering algorithms. The authors argue that the best performing classifier was SVM using Tf-idf (CA 5 0.963). In contrast to these works, for the classification of mathematical articles, we use a different approach based on the OntoMathPro ontology of professional mathematics [13].

3 Ontology based Model for Recognition of UDC Code

3.1 The OntoMathPro Ontology

The OntoMathPRO ontology is an applied ontology for automatically processing professional mathematical articles in Russian and English. The ontology defines the concepts commonly used in mathematics. The OntoMathPRO ontology covers a wide range of fields of mathematics such as number theory, set theory, algebra, analysis, geometry, theory of computation, differential equations, numerical analysis, probability theory, and statistics. Each class has a textual explanation, Russian and English labels including synonyms. Terminological sources used in the development were classic textbooks, online resources such as Wikipedia and the Cambridge Mathematical Thesaurus, scientific articles from a scientific journal, such as the journal “Russian Mathematics (Iz. VUZ)”.

In the ontology, one could distinguish two taxonomies with respect to ISA-relationship – a hierarchy of fields of mathematics and a hierarchy of mathematical knowledge objects. The first one is rather conventional and close to the related part of the Universal Decimal Classification. The top level of the second taxonomy contains concepts of three types: i) basic metamathematical concepts, e.g. Set, Operator, Map, etc; ii) root elements of the concepts related to the particular fields of mathematics, e.g. Element of Probability Theory or Element of Numerical Analysis; iii) common scientific concepts: Problem, Method, Statement, Formula, etc. OntoMathPRO defines three types of object properties.

OntoMathPRO is developed in OWL-DL/RDFS languages. Numerically, OntoMathPRO contains 3 450 classes, 5 object properties, 3 630 subclass-of property instances, and 1 140 other property instances.

3.2 Main Approach

This article examined collections of 1356 mathematical articles published in the journal "Russian Mathematics (Iz. VUZ)" for 10 years (1999-2009). Each article has at least one UDC code. In the collection under consideration, the largest number of articles falls on the UDC code 517 ("Analysis"). 883 articles have this code.

The approach proposed in this article to the automatic recognition of the UDC code for a mathematical article is based on the use of the OntoMathPro ontology. As noted above, the ontology contains basic concepts such as a problem, system, theory, equation, formula, etc. The key idea of our approach is that the choice of the UDC code is determined by a certain set of classifying features that the author of the article uses. These features are represented in the ontology by basic mathematical concepts. And the task of our research was to select the most relevant features as ontological concepts that determine the choice of the UDC code. We asked the experts to answer the question, which features are decisive for them when choosing a UDC code for their scientific works, and we came to the conclusion that the most significant features are the method, problem and equation.

Therefore, in this article, we investigate the working hypothesis that the methods, problems and equations used will be the most relevant features to create a map of the UDC code in the "Mathematics" domain.

3.3 The Architecture of the Prototype for Assessing the Relevance of Classifying Features

The general infrastructure of the workflow can be divided into two main subprocesses, such as preparing subcollections with highlighted UDC codes (Fig. 1) and assessing the relevance of classifying features (Fig. 2).

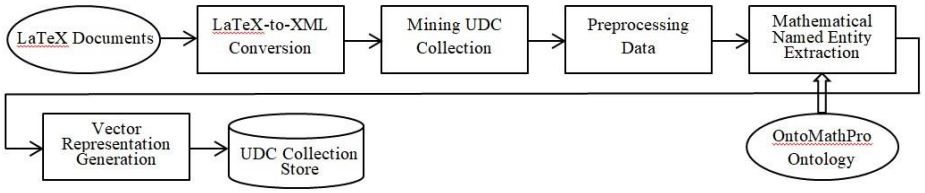

Fig. 1. The architecture of the prototype.

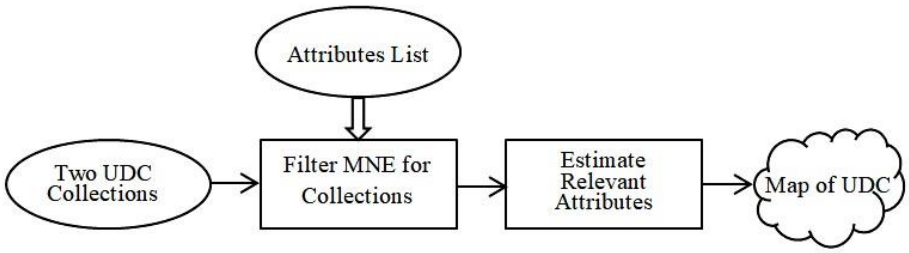

Fig. 2. The model of assessing the relevance of classifying features.

The collection preparation process includes five modules that can be combined into three subsystems as Format Conversion, Preprocessing and Semantic Annotation.

The Format Conversion subsystem provides conversion of a collection of mathematical articles into xml format. Next, the Preprocessing subsystem sorts articles by the specified UDC codes. At this stage, morphological analysis of the content of xml tags is performed using the pymorpthy2 library. The Semantic Annotation subsystem provides functionality for annotating articles in terms of a fixed set of subject areas of the OntoMathPro ontology. At this stage, all named entities recognized by the ontology are extracted from the text of the article, and a vector of the document is compiled based on the ontology dictionary.

The named entity recognition is implemented using fuzzy string comparison. The modified Levenshtein metric implemented by the fuzzywuzzy library is used as a comparison measure.

3.4 Assessing the Relevance of Features

The system for assessing the relevance of classification features is shown in Figure 2b. The Filter_MNE module receives on the input two collections with different the UDC codes and a list of classifying features. The result of the module's work is the formation of the code maps of UDC based on the selected classifying features. A UDC code map obtained with classifying feature is a set of feature values that are recognized in the corresponding subcollection of articles based on the OntoMathPro ontology. The Map_Estimate module compares these code maps of UDC obtained on these collections. At this stage, the general and specific terms of collection are determined. As a result, the module forms the code maps of UDC, which take into account the relevance of each term.

Experiments. We performed several experiments to test the working hypothesis on the most representative subcollection with the UDC code 517 (“Analysis”) in our collection of journal articles.

We carried out several experiments, pairwise comparing the constructed different code maps of UDC for different subcollections. The choice of UDC codes was based on the position of these codes in the UDC hierarchy (different first-level subtrees in the code tree with root vertice 517), relationship (descendants of one ancestor), and the size of subcollections.

The results of the experiments are presented in diagrams that show a number of common and UDC-specific terms.

Let us denote the complete code map of the classifying feature, built using the ontology, which includes all the values of this feature as *SF*, and the code map formed from the subcollection with a given UDC code, as the SF_{code}, for example, SF_{517}.

Thus, we determine the classifying feature and its code map (characteristic set of the feature) for the UDC code. Then we compare the code cards for two UDC codes and compute the relevance of the classifying feature (represented as a fuzzy linguistic variable with the values “weak”, “real”, “strong”). The relevance of the classifying feature for two code maps (*Rel_F (code1, code2)*) is calculated as:

$$Rel\_F(code1, code2) = \frac{SF_{code1} \cap SF_{code2}}{SF_{code1} \cup SF_{code2}}$$

If the Rel_F (code1, code2) value is in the [0..0.3] range, then we can talk about a strong difference in the UDC pair for this feature (the “strong” value).

If the value of Rel_F (code1, code2) is in the range [0.3..0.7], then the UDC pair is moderate distinguishable for this function (the "valid" value).

If the Rel_F (code1, code2) value is in the [0.7..1] range, then the UDC pair is poorly distinguishable for this feature (the “weak” value).

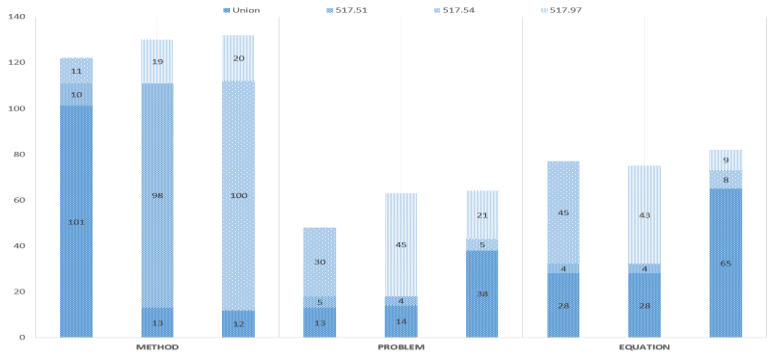

Fig. 3. The results of experiment 1.

Experiment 1. The first experiment involves UDCs of the same level and comparable sizes of subcollections (UDC 517.51 (89 articles) and 517.54 (87 articles)), as well as UDC 517.97 (75 articles) from another subdomain. We consider a method, an equation, and a problem as classifying features. The results of experiment 1 are shown in figure 3 and the interpretation of these results is in table 1.

Table 1. Assessing the classifying features in experiment 1.

	Method	**Problem**	**Equation**
517.51 & 517.54	weak	strong	valid
517.51 & 517.97	strong	strong	valid
517.54 & 517.97	strong	valid	weak

It can be seen that for UDC 517.51 and 517.54 the most relevant feature will be the methods, and for 517.97 – equations.

Experiment 2. In this experiment, we consider highly specialized UDCs: 517.956 (57 papers), 517.958 (59), 517.982(21) and 517.983(36). These UDC codes do not have a large number of representatives in the collection, but due to their high specialization, we believe that they should differ significantly in characteristics. The results of experiment 2 are shown in fig. 4 and the interpretation of these results is in table 2.

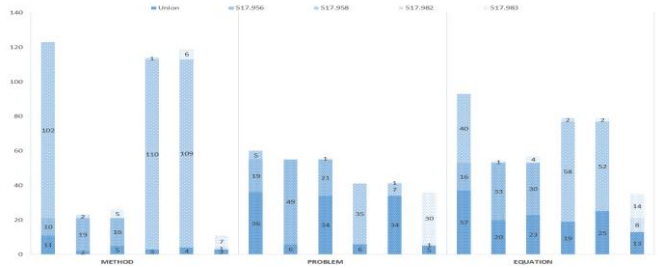

Fig. 4. The results of experiment 2.

Table 2. Assessing the classifying features in experiment 2.

	Method	**Problem**	**Equation**
517.956 & 517.958	strong	valid	valid
517.956 & 517.982	strong	strong	valid
517.956 & 517.983	strong	valid	valid
517.958 & 517.982	strong	strong	strong
517.958 & 517.983	strong	weak	strong
517.982 & 517.983	valid	strong	valid

The analysis of the above results shows that for UDC 517.956 the most relevant feature is the problem, for UDC 517.958 - methods, and for UDC 517.982 and 517.983 - equations.

Experiment 3. In this experiment, we investigated single-level UDCs of one parent node, which have the largest number of representatives in the collection: 517.92(129), 517.95(156) and 517.98(133).The results of experiment 3 are shown in figure 5 and the interpretation of these results is in table 3.

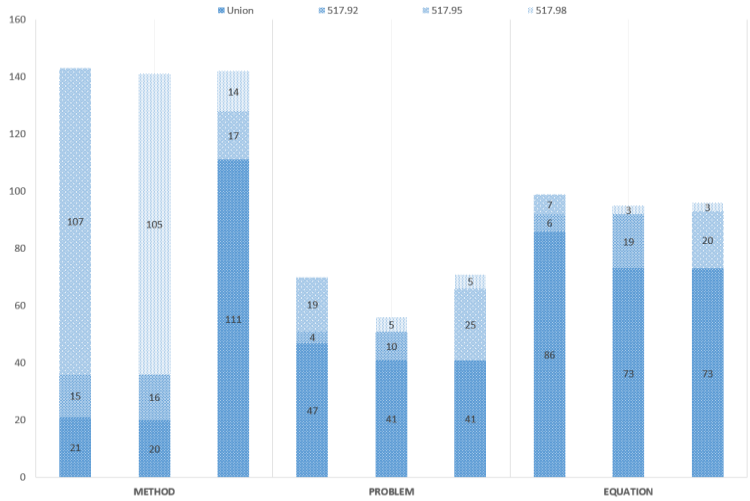

Fig. 5. The results of experiment 3.

Table 3. Assessing the classifying features in experiment 3.

	Method	**Problem**	**Equation**
517.92 & 517.95	strong	valid	weak
517.92 & 517.98	strong	weak	weak
517.95 & 517.98	weak	valid	weak

Analysis of the above results shows that methods and problems are the most relevant feature for explored UDC codes.

An important preliminary conclusion from the experiments carried out is the construction of code maps of the studied UDC codes based on the OntoMathPro ontology (see Table 4). The table contains values for each features (total and number of unique values).

Table 4. Digital code maps of the studied UDC codes.

	Method		**Problem**		**Equation**	
	All	Unique	All	Unique	All	Unique
517.51	111	10	18	5	32	4
517.54	112	11	43	4	73	8
517.97	32	19	59	21	71	9
517.956	21	10	55	19	53	16
517.958	113	102	41	5	57	40
517.982	4	1	6	0	21	1
517.983	10	5	35	1	27	2
517.92	36	15	51	4	92	6
517.95	128	17	66	19	93	7
517.98	125	14	46	5	76	3

4 Conclusion

The research carried out allows concluding that the combination of the selected features and their values can successfully classify collections by UDC codes. Some related groups of UDC codes can be classified according to only one feature, but with an increase in the degree of code relationship, the number of required classifying features increases. We also identified the most relevant features of the UDC code groups, by which we can classify them in the general UDC code tree.

The research carried out confirms our hypothesis that a group of mathematical UDC codes can be classified by the features such as “method”, “task” and “equation”.

Acknowledgement. The research was funded by RSF according to the project № 21-11-00105.

References

1. Jie Lu, Dianshuang Wu, Mingsong Mao, Wei Wang, Guangquan Zhang: Recommender system application developments: A survey. In: Decision Support Systems, vol. 74, pp.12-32 (2015).
2. Ricci F. (2014) Recommender Systems: Models and Techniques. In: Alhajj R., Rokne J. (eds) Encyclopedia of Social Network Analysis and Mining. Springer, New York, NY. https://doi.org/10.1007/978-1-4614-6170-8_88.
3. Liliana Shakirova, Marina Falileeva, Alexander Kirillovich, Evgeny Lipachev, Olga Nevzorova, Vladimir Nevzorov: Modeling and Evaluation of the Mathematical Educational Ontology. In: SSI–2019 Scientific Services & Internet Proceedings of the 21st Conference on Scientific Services & Internet (SSI–2019) Novorossiysk-Abrau, Russia, September 23–28, 2019. CEUR Workshop Proceedings, vol. 2543, CEUR-WS.org, pp. 305–319 (2020).
4. Elizarov A.M, Lipachev E.K.: Methods of processing large collections of scientific documents and the formation of digital mathematical library. In: CEUR Workshop Proceedings, vol. 2543, pp.354–360 (2020).
5. Elizarov A, Lipachev E.: Big math methods in Lobachevskii-DML digital library. In: CEUR Workshop Proceedings, vol.2523, pp.59-72 (2019).
6. Romanov, A.Y., Lomotin, K.E., Kozlova, E.S. and Kolesnichenko, A.L.: Research of neural networks application efficiency in automatic scientific articles classification according to UDC. In: International Siberian Conference on Control and Communications, SIBCON 2016 – Proceedings, pp. 7–11 (2016).
7. Khoo, M.J., Ahn, J.W., Binding, C., Jones, H.J., Lin, X., Massam, D. and Tudhope, D.: Augmenting Dublin core digital library metadata with Dewey decimal classification. In: Journal of Documentation, vol. 71, No. 5, pp. 976–998 (2015).
8. Bai, X., Wang, M., Lee, I., Yang, Z., Kong, X. and Xia, F.: Scientific paper Recommendation: a survey. In: IEEE Access, IEEE, vol. 7, pp. 9324–9339, doi: 10.1109/ACCESS.2018.2890388 (2019).
9. Beel, J., Aizawa, A., Breitinger, C. and Gipp, B.: Mr. DLib: recommendations-as-a-service (RaaS) for academia. In: Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL) (2017).
10. Shuai Zhang, Lina Yao, Aixin Sun, Yi Tay: Deep Learning Based Recommender System: A Survey and New Perspectives. In:ACM Computing Surveys, 52(1), pp. 1-38 (2019).
11. M. Schubotz et al.: AutoMSC: Automatic Assignment of Mathematics Subject Classification Labels. In: Proceedings of the 13th Conference on Intelligent Computer Mathematics (2020).
12. Matjaž Kragelj, Mirjana Kljajić Borštnar: Automatic classification of older electronic texts into the Universal Decimal Classification–UDC. In: Journal of Documentation, vol. 77, no. 3 (2021).
13. Olga A. Nevzorova, Nikita Zhiltsov, Alexander Kirillovich and Evgeny Lipachev: OntoMathPro Ontology: a Linked data hub for mathematics // 5th International Conference, KESW 2014, Kazan, Russia, September 29 – October 1, 2014. Proceedings. Series: Communications in Computer and Information Science, vol. 468, pp. 105–119. Springer (2014).