# MATHEMATICAL KNOWLEDGE ONTOLOGIES AND RECOMMENDER SYSTEMS FOR COLLECTIONS OF DOCUMENTS IN PHYSICS AND MATHEMATICS

A.M. Elizarov, A.V. Kirillovich, E.K. Lipachev, A.B. Zhizhchenko, N.G. Zhil'tsov

One of the methods for increasing the efficiency of work with Web content is the use of services offered by various systems of recommendation preparation. Such services are incorporated into popular on-line libraries, such as the Google Scholar retrieval system, the Scopus abstract database, the Mendeley control system of bibliographic information, the electronic library eLIBRARY.ru, and other systems. There exist two main types of recommender systems, content-oriented systems and social systems (of collaborative filtration); see, e.g., [1] – [4]. The former are based on a representation of users' preferences obtained by analyzing the contents of recommendation elements. Systems of the second type model preferences by estimating the proximity of user profiles. In what follows, by a recommender system we understand an information system which

- forms a conceptual representation of an electronic collection (e.g., MathNet) on the basis of ontologies of knowledge domains;
- calculates the measure of thematic proximity between documents by using this representation;
- outputs a list of documents reflecting the information interests of the user.

In comparison with information retrieval systems, recommender systems are most useful when the user experiences difficulties with formulating an effective retrieval request. Such difficulties arise most frequently in dealing with scientific content.

For preparing high-quality recommendations, a user profile constructed only on the basis of the history of documents browsed by the user in the collection being analyzed is insufficient; it is necessary to take into account models of knowledge domains, in particular, by using ontologies, and the user scenario. Users of a recommender system may be

- authors of publications, who are most interested in close settings and solution methods of stated problems in the document collection;
- referees, who estimate the novelty and importance of a particular scientific document in comparison with other existing documents on the same topic;
- scientists searching for materials related to their studies and readers interested in documents useful in understanding a topic and explaining the necessary basis notions.

For example, in analyzing the text of the same paper, the author of a planned new publication seeks a material useful for preparing a survey of related papers, a referee is interested in materials with definitions of specific terms and related papers needed to

estimate the novelty of the content, and a post-graduate student searches for classical papers with definitions of basis notions and fundamental statements.

At present, ontologies have become widespread as a tool for representing knowledge in a subject area, in particular, for recommender systems [5], by means of semantic relations. Such relations ensure, in particular, the improvement of search completeness and precision, as well as multilingualism. However, one of the obstacles in the propagation of ontological approaches in recommender systems is the laboriousness of the development of models for each knowledge domain separately. The application of the ontological approach in recommender systems for physico-mathematical content has become possible after the appearance of physico-mathematical knowledge ontologies [6] – [10]. In this paper, we describe a recommender system based on the mathematical knowledge ontologies OntoMathPRO and Mocassin and tested on the unique collection MathNet of publications in physics and mathematics [11, 12].

*Mathematical knowledge ontologies. Mocassin* is an ontology of the logical structure of mathematical documents developed by these authors for automatically analyzing mathematical publications in the LaTeXformat. This ontology formally (in the OWL language) describes the semantics of structural elements of mathematical documents (e.g., theorems, lemmas, proofs, definitions, etc.) expressed in the form of classes and properties. In addition, this ontology contains axioms of cardinality and transitivity.

OntoMath$^{PRO}$ [6] – [8] contains about 3500 concepts, which are organized into two hierarchies: mathematical objects and domains of mathematics. In this ontology, there are defined the antisymmetric relations

- "subclass" → "class" ("manifold" → "set");
- "defined by means of" ("atlas" → "manifold"),

and three types of symmetric relations:

- "associative relation" ("differentiable manifold" → "flow");
- "problem" → "solution method" ("system of linear equations" → "Gaussian elimination");
- "area of mathematics" → "mathematical object" ("algebraic topology" → "homology group").

Below, we describe a model of the application of the ontology structure to represent documents. The ontology can be represented in the form of a directed weighted graph $G = \langle T, E \rangle$, whose vertices are the elements of the set of concepts $T$ and edges are the elements of the set of relations $E$. To an antisymmetric relation, there corresponds one directed edge, and to a symmetric relation, two oppositely directed edges. Each edge is assigned a weight $w_r$ depending on the type of the relation $r$ and the scenario of the system operation.

We introduce the following measure of proximity between terms $s$ and $d$:

$$A[s,d] = \begin{cases} 1, & s = d, \\ 0, & \text{не существует пути между s и d}, \\ 1/dist(s,d) & \text{в остальных случаях}, \end{cases}$$

where $dist(s,d)$ is the shortest path length between $s$ and $d$.

We propose a method which processes a collection of mathematical documents and forms a list of related papers for each document. The method consists of the following steps. At the first step, from each document key words are extracted; these are mentions of terms described in the OntoMathPRO ontology and their mathematical notations. At the second step, the logical structure of the document is analyzed and its fragments are semantically marked on the basis of the Mocassin ontology [9]. At the third step, a vector representation of the document is constructed, which takes into account its terminological composition, position of terms in the logical structure, and connections between terms in the OntoMathPRO ontological graph. After that, the proximity measure of the constructed vectors of document in the collection is calculated and recommendations are formed.

Extraction of logical structure. At this step, the logical structure of the document is extracted on the basis of the Mocassin ontology. Fragments of the document are annotated as instances of ontology classes. The information on the classes to which fragments of the document belong is then used for weighting general terms in order to calculate the proximity measure between documents. Each concept of the Mocassin ontology is assigned a weight determining its significance in the document structure (the maximum weight is assigned to the title).

Extraction of key words. At this step, mathematical terms are extracted from the text of the document. A special feature of this step is that terms are often mentioned in symbolic form, as elements of mathematical notation. We use the OntoMathPRO ontology as the base of terms and Textocat API (a cloud text analytics platform developed with the participation of these authors) as an analysis tool. The extraction of terms is based on standard approaches to resolving the lexical multivaluedness of named entities [13] and consists of two main steps: extraction of name groups as phrase-candidates for being linked with ontology terms and the application of a binary classifier making a decision on linking phrase-candidates with ontology terms of specified confidence measure. Then, the extracted notions are tied to variables, which represent copies of these notions in mathematical formulas [14]. Thereafter, all mentions of a variable are regarded as mentions of the corresponding term.

As a result, we obtain a set $M_p$ of mentions of ontology terms in a document $p$, in which each element is a tuple $m = \langle t, l \rangle$, where $t$ is an ontology term and $l$ is the position of this term in the document.

Construction of a publication vector. At this step, a vector representation of the document in the terminological space is formed. As the terminological base we use the terms of the OntoMathPRO ontology. We order the set $T$ as follows.

The document $p$ is represented as a vector in which every component is the weight of the corresponding term in the given document:

$$v(p) = (weight(t_1), weight(t_2), \ldots, weight(t_n)), n = |T|,$$

where $weight(t)$ is the weight of a term $t$ in the document. Each mention increases the weight of the corresponding term; the increase in the weight of a term depends on the position of the term in the logical structure of the document. Moreover, each mention of a term increases the weight of all terms connected with it according to the proximity measure. Thus, each component of the vector is calculated by

$$weight(t^*) = idf(t^*) \cdot \sum_{\langle t,l \rangle \in M_p} \beta_l \cdot A[t, t^*],$$

where $idf(t) = \log \dfrac{N}{N_t}$, $N$ is the number of documents in the collection, $N_t$ is the number of documents containing the term $t$, and $\beta_l$ is the weight of the term determined by the position $l$ of this term in the logical structure of the document. The use of $idf$ makes it possible to reduce the influence of common terms (such as "number" or "method"), which weakly reflect the specifics of a particular document.

On the basis of the document vector, the recommender system forms a data set containing a list of related documents in the collection and an extended list of key words selected according to their weights in the publication vector. Since each key word corresponds to an ontology term, the system forms recommendations containing references to various definitions of the term, the position of the term in the ontology hierarchy, and a list of documents in the collection which contain this term.

Formation of a list of related papers. At this step, for each document in the collection, a set of related documents is constructed. As a measure of proximity between documents we use the well-known cosine proximity measure of their vectors. The recommended documents are those with proximity measure larger than a preset threshold value.

Substantiation of the formed recommendation. An explanation of a list of recommendations is a list of marks reflecting the most important, from the point of view of the model, attributes $weight(t^*)$ taken into account in calculating the proximity measure. Examples of marks are "related terms" and "similar problem".

Thus, the created recommender system is characterized by the following features:

- it takes into account the professional profile of a particular user and of other users interested in a given topic;
- it forms different recommendations for different scenarios of work with the system (referee, user being introduced in the topic, etc.);
- it assigns different weights to different concepts; thus, for a scientific review, concepts denoting areas of mathematics are more important than those characterizing mathematical objects, while for a beginning researcher, survey papers containing notions from different areas of mathematics and references to original works are important;
- it supposes further integration with MathNet and other scientific collections;
- it allows the ontology to be supplemented with new concepts, including notions from Mathematical Encyclopaedia.

This work was supported by the Russian Foundation for Basic Research, project nos. 15-07-08522 and 15-47-02472.

## Список литературы

[1] *Ricci F., Rokach L., Shapira B., Kantor P.B. (Eds.)* Recommender Systems Handbook. Springer-Verlag New York. 2010.

[2] *Bobadilla J., Ortega F., Hernando A., Gutierrez A.* Recommender systems survey. Knowledge-Based Systems. 2013. V. 46. P. 109-–132.

[3] *Lampropoulos A.S., Tsihrintzis G.A.* Machine Learning Paradigms. Applications in Recommender Systems. Springer International Publishing Switzerland, 2015. 125 p.

[4] *Verbert K., Manouselis N., Ochoa X., Wolpers M., Drachsler H., Bosnic I., Duval E.* Context-aware recommender systems for learning: a survey and future challenges // IEEE Transactions on Learning Technologies. 2012. V. 5. No 4. P. 318–335.

[5] *Middleton S.E., De Roure D., Shadbolt N.R.* Ontology-Based Recommender Systems // In Staab S., Studer R. (Eds.) Handbook on Ontologies. Springer-Verlag. Berlin. Heidelberg. 2009. P. 779–796.

[6] *Elizarov A.M., Lipachev E.K., Nevzorova O.A., Solovyev V.D.* Methods and Means for Semantic Structuring of Electronic Mathematical Documents // Doklady Mathematics. 2014. Vol. 90. No. 1. P. 521-524. C. 642–645.

[7] *Elizarov A.M., Kirillovich A.V., Lipachev E.K., Nevzorova O.A., Solovyev V.D., Zhiltsov N.G.* Mathematical knowledge representation: semantic models and formalisms // Lobachevskii J. of Mathematics. 2014. V. 35, No 4. P. 347–353.

[8] *Nevzorova O., Zhiltsov N., Kirillovich A., Lipachev E.* OntoMath$^{PRO}$ ontology: a linked data hub for mathematics //In Knowledge Engineering and the Semantic Web. Springer International Publishing. Communications in Computer and Information Science. V. 468. P. 105-119, 2014.

[9] *Solovyev V., Zhiltsov N.* // Proc. of the International Conference on Web Intelligence, Mining and Semantics. ACM, 2011.

[10] *Aberer K., Boyarsky A., Cudré-Mauroux P., Demartini G., Ruchayskiy O.* ScienceWISE: a Web-based Interactive Semantic Platform for scientific collaboration // 10th International Semantic Web Conference (ISWC 2011-Demo), Bonn, 2011.

[11] *Zhizhchenko A.B., Izaak A.D.* The information system Math-Net.Ru. Application of contemporary technologies in the scientific work of mathematicians // Russian Math. Surveys. 2007. 62:5. P. 943–966.

[12] *Chebukov D., Izaak A., Misurina O., Pupyrev Yu., Zhizhchenko A.* Math-Net.Ru as a digital archive of the Russian mathematical knowledge from the XIX century to today. Lecture Notes in Computer Science. 2013. **7961**. P. 344–348.

[13] *Shen W., Wang J., Han J.* Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions // IEEE Transactions on Knowledge and Data Engineering. 2015. V. 27. Issue 2. P. 443–460.

[14] *Nevzorova O., Zhiltsov N., Zaikin D., Zhibrik O., Kirillovich A., Nevzorov V., Birialtsev E.* Bringing Math to LOD: A Semantic Publishing Platform Prototype for Scientific Collections in Mathematics. // In The Semantic Web–ISWC. Springer Berlin Heidelberg. 2013. P. 379-394.