Query Expansion Method Application for Searching in Mathematical Subject Domains

Olga Ataeva[0000-0003-0367-5575], Vladimir Serebryakov[0000-0003-1423-621X]
and Natalia Tuchkova[0000-0001-6518-5817]

Dorodnicyn Computing Center FRC CSC of RAS, Vavilov str., 40, 11933, Moscow, Russia
oli@ultimeta.ru, serebr@ultimeta.ru, natalia_tuchkova@mail.ru

Abstract. The article focuses on the problem of information retrieval in the mathematical scientific articles**.** Issues related to scientific publication "in the scientific information environment" are considered. The possibilities of expanding the search query in the presence of a subject domain thesaurus are discussed. The formulation of an information request in specific subject domains is not always possible for an insufficiently competent user. The query expansion process is of particular importance for subject domains, where the search is based on special terminology. Automatic query expansion can serve as a necessary tool with which the user can get a pertinent search result. Mathematical texts are differed by the presence of a specific structure and the use of formulas. Formulas can often reflect the main results of research and therefore it is important to be able to use them in an advanced query. Modern programming tools allow you to search by formulas and use them for indexing. It is proposed to include formulas in mathematical thesauruses and in this way to create a context of formulas for their effective inclusion in search queries. The role of the context defined by the relations of thesaurus concepts is both to refine the query and to increase the scale of the sample on request. Examples are implemented for subject areas of ordinary differential equations, equations of mixed type, special functions of mathematical physics, etc.

Keywords: Comparison of Scientific Texts, Semantic Search, Thesaurus for the Ontology of Knowledge, Information Query using the Thesaurus, LibMeta

1 Introduction

Many scientific publications are devoted to the problem of expanding queries. It is known that the unreasonable addition of text in a query leads to additional information noise, and only the use of logically related terms can lead to a refinement of the query and an improvement in the search result. The query expansion tools allow you to refine the query using hints to the user, narrowing the search field using thesaurus descriptors, and use the existing relationships of terms (synonyms, abbreviations, etc.), increasing the search field and thereby receiving additional information noise. These two processes are in contradiction, but in the end lead to a pertinent result, that is, satisfying the informational interest of the user. Developments in this direction have been underway for a long time, and many information systems allow the expansion of the request. In [1], results are given that indicate that using synonyms from the WordNet database that are not related to context does not improve the quality of the information request. And only using the technology of prescribing manual "semantic" links allows you to expand the query to a useful information field, but, of course, that way it will not be possible to cover any significant number of links. As a result, there is a need to formulate the problem of automatic accounting for semantic relations, which is possible in the presence of a thesaurus corresponding to the subject domain. The particular difficulty of the process of searching for scientific information is clarifying and expanding the information request. The basis for such search is the use of special terminology and the connections defined by the domain logic. The hierarchical presentation of scientific data is also difficult when the problem of establishing associative relationships between concepts appears [2]. Using the example of problems of mathematical physics and related fields, it is proposed to show how expanding a query by adding mathematical symbols and formulas from the thesaurus can improve search results.

2 About the Query Expansion Methods

The extension of the information query involves reformulating the original query in order to improve the search result. This process is directly related to the understanding of the subject of search both from the side of the user (level of competence in a certain subject area), and from the side of the information retrieval system (it means availability of information and functional tools of expanding and refining the query).

It was noted in [4] that the history of studies of the expansion of an information query can be traced from 1965, when a formalized description of the relevance of search query results based on a vector feedback model (known as Rocchio algorithm). Earlier research into the "weight estimates" of related and non-related terms when expanding the query belongs to Spark Jones [6] and van Rijsbergen [7]. The idea of Relevance Feedback (RF) is to engage the user in the search process in order to improve the final list of results. In particular, the user informs the system of the relevance of the documents in the initial list of search results. The Rocchio algorithm is a classic algorithm for implementing the RF method. It adds a relevance feedback model to the vector space model [8]. The automatic generation of thesauruses was discussed by Qui and Frei [9] and Schütze [10]. The use of local and global methods for expanding queries has been investigated by Croft et al. [11].

These works became the foundation for further research in the field of expanding the information request for text documents, there were not enough tools for processing symbolic information. Currently, software has been developed that allows the use of formulas and symbolic expressions in databases, which opens up the possibility of using formula entries to expand a search query. The most famous example of such an information resource is Zentralblatt MATH [12].

This study proposes an approach based on taking into account relations from related domains, based on associative relations of terms and thesaurus formulas. Earlier, a technology was proposed for filling the addressee thesaurus and thesaurus of ordinary differential equations and mixed type equations [13, 14]. This technology is for the informed (competent) users.

If the user is not sufficiently familiar with the subject area, then any information can be a replenishment of the addressee (recipient's) thesaurus. A search query extension for a such user can be a useful hint in the search process.

It is known that the same mathematical expressions are found in the description of various phenomena. This fact is important to use a pair: “formula + term” to search in a specific subject area. Moreover, the formula, of course, can be associated with various terms that make up the semantic context of various subject domains. This allows us to consider options for expanding the information request for the ontology of one subject domain and ontologies of various subject domains.

2.1 Query Expansion "within the subject domain"

The option of expanding the information request is being considered, as a result of which it is expected to achieve greater coverage of related data in a certain subject domain. Such a situation arises if it is necessary immediately as a result of one request to obtain a variety of information on a specific topic. The recipient may be satisfied with the search result or at the next stage of the search try to refine the request.

A simple example is a search in scientific texts by the name of the author, and then a selection by keyword or other known data. In particular, mathematicians are often more comfortable communicating in the language of formulas. To search for a mathematical result, a specialist must be given the opportunity to expand the query with formulas. It is proposed to use specialized LibMeta thesaurus [3], where along with definitions in the natural language there are symbolic expressions, formulas. This approach allows you to refine the query within the subject domain, using a mathematical entry in TeX notation, regardless of the language of the source. For example, the well-known equation for the Gellerstedt problem is logically connected with a number of items, Fig. 1.

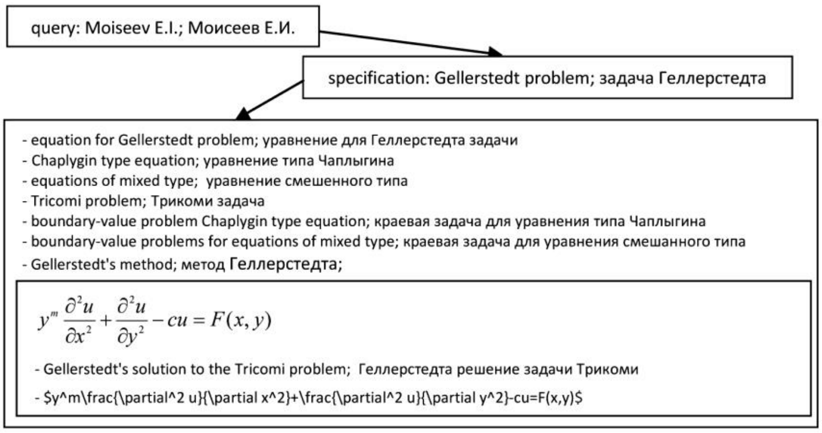

query: Moiseev E.I.; Моисеев Е.И.
specification: Gellerstedt problem; задача Геллерстедта
- equation for Gellerstedt problem; уравнение для Геллерстедта задачи
- Chaplygin type equation; уравнение типа Чаплыгина
- equations of mixed type; уравнение смешенного типа
- Tricomi problem; Трикоми задача
- boundary-value problem Chaplygin type equation; краевая задача для уравнения типа Чаплыгина
- boundary-value problems for equations of mixed type; краевая задача для уравнения смешанного типа
- Gellerstedt's method; метод Геллерстедта;
$y^m \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} - cu = F(x, y)$
- Gellerstedt's solution to the Tricomi problem; Геллерстедта решение задачи Трикоми
- $y^m\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}-cu=F(x,y)$

Fig. 1. Scheme for using a symbolic expression in a search fields.

On a request for an author: “Moiseev E.I.” (Моисеев Е.И., in Russian) and a refinement of the “Gellerstedt problem” (задача Геллерстедта, in Russian), a whole list of related terms can be obtained. Among other things, there is a term with a formula expression, since it is stored in the corresponding thesaurus article relating to equations of a mixed type. It is this hint in the form of a formula that will allow you to make a choice when searching inside the subject area for “equations of mathematical physics of a mixed type”. You can then use this expression as an extension of the search query and obtain other search results related to this equation.

2.2 Query Expansion for “related subject domains”

It is known that the same phenomena encountered in the natural sciences can be modeled in various fields of knowledge, and identical (similar) symbolic expressions can be used. For example, the “wave equation” is used to model various technical processes. The record of the wave equation is almost the same everywhere. The structure of the formula, i.e. its record, in different cases of use can be the same, only interpretations of the input and output data are different, except for the natural time parameter or input data.

For example: $u_{tt} = a^2 u_{xx}$, and in TeX notation: $u_{tt}=a^2u_{xx}$ – the equation of free transverse vibrations of a string, although it can also be found in other applications. On the example of Fig. 2 shows the increase in the search fields for scientific publications due to the terminology of related fields and the reformulation of the query.

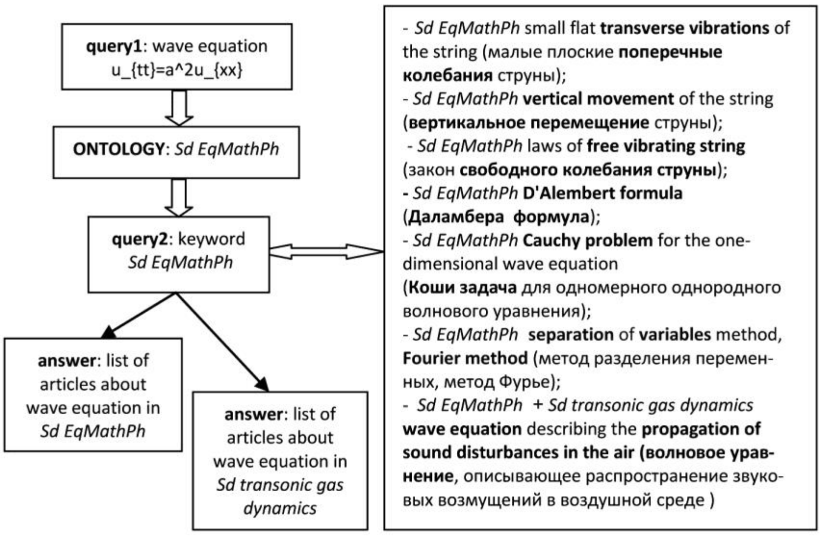

query1: wave equation
u_{tt}=a^2u_{xx}
ONTOLOGY: *Sd EqMathPh*
query2: keyword
Sd EqMathPh
answer: list of articles about wave equation in *Sd EqMathPh*
answer: list of articles about wave equation in *Sd transonic gas dynamics*
- *Sd EqMathPh* small flat **transverse vibrations** of the string (малые плоские **поперечные колебания** струны);
- *Sd EqMathPh* **vertical movement** of the string (**вертикальное перемещение** струны);
- *Sd EqMathPh* laws of **free vibrating string** (закон **свободного колебания струны**);
- *Sd EqMathPh* **D'Alembert formula** (**Даламбера формула**);
- *Sd EqMathPh* **Cauchy problem** for the one-dimensional wave equation (**Коши задача** для одномерного однородного волнового уравнения);
- *Sd EqMathPh* **separation** of **variables** method, **Fourier method** (метод разделения переменных, метод Фурье);
- *Sd EqMathPh* + *Sd transonic gas dynamics* **wave equation** describing the **propagation of sound disturbances in the air (волновое уравнение**, описывающее распространение звуковых возмущений в воздушной среде)

Fig. 2. Query extension scheme for related subject domain, where: Sd is the “subject domain”, EqMathPh is the “equations of mathematical physics”.

According to the relationships of the thesaurus of equations of mathematical physics, one can easily proceed to search from articles arrays in one subject domain to another. In the example in Fig. 2 shows how, from the query on the topic “wave equation”, the transition to the formulation on the topic “transonic gas dynamics” is carried out, where equations from an adjacent subject domain are considered. Another example: the “Tricomi equation” from the section “equations of mixed type” also has numerous applications, from the description of the problems of “magneto hydrodynamic flows” to the problems of “transonic gas dynamics” from one more general subject area of “equations of mathematical physics”. An unlimited number of these examples can be found, since the equations of mathematical physics, as a subject area, appeared for modeling physical and technical processes, i.e., it has many applications and related fields. Their informational images in search engines can be covered, thanks to the capabilities of query expansion.

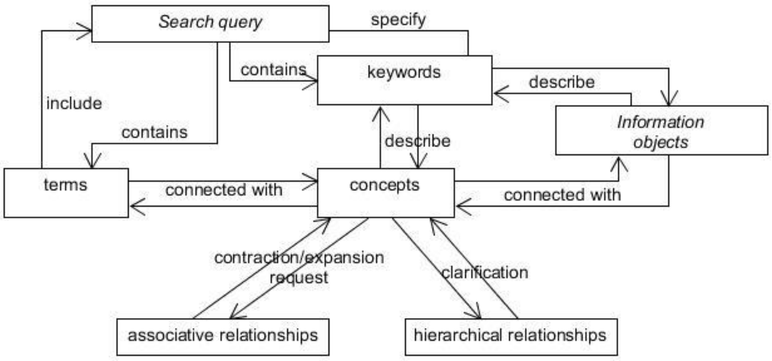


Fig. 3. Scheme of qualifying queries based on the concepts of the thesaurus and their relationships.

In Fig. 5 schematically shows the ways of forming clarifying queries using the concepts of the thesaurus and its main connections in the LibMeta system. Moreover, any path from the “Search Query” to the “Information Object” may be sufficient to obtain the necessary response to the query. Keywords can be used not only in the conventional sense, but, as described above, they can be, for example, formulas.

3 Benefits of Data Integration to Expand Query in Mathematical Subject Domains

At first glance, the advantages are obvious: the greater the scope of the request, the more information as a result the user of the integrated information system will receive. However, it is also known that the expansion of the query leads to an increase in information noise, which in no way can be attributed to the advantages of the search. The combination of these two features should take on some "optimal" values so that the search query extension service constitutes a useful property of the information system.

The optimal properties of the integrated system, ensuring the effective expansion of the information request, are realized thanks to the following features:

- special data structure;
- functional properties;
- the possibility of "tuning" to the user's subject domain.

3.1 Structure of Mathematical Data in LibMeta

Ontology describes the resources of the subject domain and their relationship. For each subject LibMeta's domains, the set of resources may differ both in format and in the set of resources themselves.

LibMeta uses the semantic content of a specific domain and concepts common to any domain to describe the library. A set of concepts is proposed that formulate a description of the content of the library, universal enough to include a specific subject domain in the system. This approach allows implementing data integration tools within the library, adaptable to the conditions of any subject domain, taking into account its specificity. Thanks to this, one of the main problems of integrating data from various sources is solved, namely, the reconciliation of heterogeneous digital information.

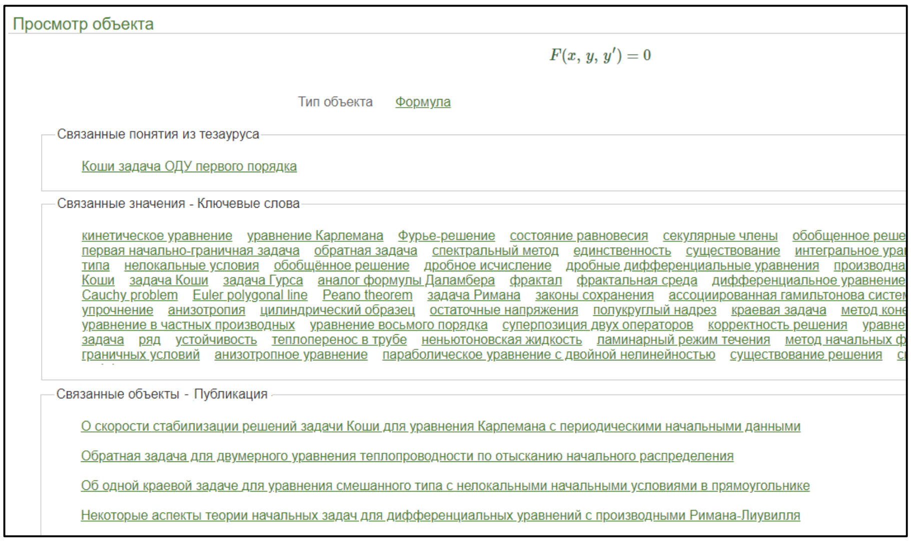
The concepts of ontology in the LibMeta system can be divided according to the functional purposes as following:

- the description of the content of the subject domain;
- the formation of a thesaurus of any subject domain;
- the description of thematic collections,
- the description of the task of integrating library content with data from external sources.

Between groups of concepts defined semantically significant relationships.

To support work with mathematical texts in the system, the concept of Formula was introduced, which allows you to store the original line of the formula from the source of origin. The formula string can be in the format Content MathML, Presentation MathML, LaTeX. If necessary, the number of types of presentation of the formula in various notations can be expanded.

The concept of a Formula is connected with relations with information objects that make up the content of a semantic library and with the concepts of a thesaurus. Thus, we can always build a network of relations of the formula with the concepts of the thesaurus and with various information objects of the system. Each formula can be supplemented with keywords, classifier codes, etc. Keywords can be put down by an expert of the system or added automatically, coming along with the formula from its source, as well as supplemented with keywords related objects. In Fig. 4 shows a set of relationships for a formula extracted from the concept of “Cauchy first-order ODE problem”, automatically constructed from thesaurus links.



Просмотр объекта

$F(x, y, y') = 0$

Тип объекта Формула

Связанные понятия из тезауруса

Коши задача ОДУ первого порядка

Связанные значения - Ключевые слова

кинетическое уравнение уравнение Карлемана Фурье-решение состояние равновесия секулярные члены обобщенное реше
первая начально-граничная задача обратная задача спектральный метод единственность существование интегральное ура
типа нелокальные условия обобщённое решение дробное исчисление дробные дифференциальные уравнения производна
Коши задача Коши задача Гурса аналог формулы Даламбера фрактал фрактальная среда дифференциальное уравнение
Cauchy problem Euler polygonal line Peano theorem задача Римана законы сохранения ассоциированная гамильтонова систем
упрочнение анизотропия цилиндрический образец остаточные напряжения полукруглый надрез краевая задача метод коне
уравнение в частных производных уравнение восьмого порядка суперпозиция двух операторов корректность решения уравне
задача ряд устойчивость теплоперенос в трубе неньютоновская жидкость ламинарный режим течения метод начальных ф
граничных условий анизотропное уравнение параболическое уравнение с двойной нелинейностью существование решения с

Связанные объекты - Публикация

О скорости стабилизации решений задачи Коши для уравнения Карлемана с периодическими начальными данными

Обратная задача для двумерного уравнения теплопроводности по отысканию начального распределения

Об одной краевой задаче для уравнения смешанного типа с нелокальными начальными условиями в прямоугольнике

Некоторые аспекты теории начальных задач для дифференциальных уравнений с производными Римана-Лиувилля

Fig. 4. Formula's semantic relationships.

3.2 Configuring a User Subject Domain in LibMeta

Consider the user’s mathematical subject domain. In the LibMeta structure, these are lists of publications on a certain topic, secondary documents of publications, and formulas as well.

The adaptation of this data of the subject area is interpreted as the “tuning” of the sources, in which several main stages (i-iii) can be distinguished.

i. Connecting a data source. Each data source is characterized by a unique URL and some set of parameters necessary for accessing the data. A preliminary analysis of the information available from the source is carried out; in particular, the types of its resources and their properties involved in the integration are determined. The result of this first step is the determination of that part of the source circuit from which data will be extracted.

ii. Defining LibMeta library resource types corresponding to source resource types. For each source resource defined by its schema extracted at this stage, the LibMeta library resource is mapped. The result of this step is the establishment of a link between the library resource and the source resource. Relationships are established using the appropriate operation, which declares that there are resource instances corresponding to the same real-world object. Based on certain (identified) relationships, the next step is the mapping of attributes.

iii. For each LibMeta resource, the mapping of attributes to the corresponding properties of the data source resource is determined. First of all, a mapping is constructed for identifying attributes that are mandatory, then for the rest. For each such pair, the type of connection and the set of operations are determined.

Thanks to this construction of mappings, we obtain a set of rules by which one can represent each object found in the source within the framework of the concepts of the LibMeta library. Next, save the metadata of objects in local storage at the request of the user, or simply save the connection between the found object in the source and the object in the library.

3.3 Data Integration of Different Subject Domains

A formal model of the process of integrating data from various subject areas can be represented as follows.

Based on the basic concepts of LibMeta, the content model of the *G* library is a few sets:

- the set of resources $R = \{r_j\}$,

- the set of attributes $A = \{a_i\}$,

- the set of attributes $N(r)\subset A$,, i.e., $r_j(a_1,.., a_n)$, $a_n\in N(r)$, defined for each resource.

Each set of attributes includes identifying attributes, $I(r)\subset N(r)\subset A$, used to uniquely identify the information objects of this resource.

Formally, the integration subsystem I_T is represented by a triple $<G,\{S_i\},\{M_i\}>$, where *G* is a predefined content model consisting of a set of resources *R* and their descriptions in the form of a set of attributes *N (r)*, *Si* is an *i*-th source connected to the system, *Mi* is the display of the *i*-th source, $1\leq i\leq n$, where *n* is the number of data sources.

Using a data source can occur in two scenarios:

- in the mode of affixing relations with objects available in the library;

- in the attribute search mode by the data source within the specified by the presentation (information image).

At the same time, data on objects from sources can be saved in two ways:

- “linking” – this method is identical to establishing a “see also” connection and means that at one end there is more complete and extensive information on the resource;

- “identification” – this method is identical in meaning to establishing a connection "the same as" and means that at one end contains exactly the same quality of information object, as with the other.

Due to the flexibility of the library content model, a scenario is possible for creating additional types of resources for plug-in sources, the information from which can be used as the values of some attributes of the main resources.

The library resource scheme *G*, both the data source *S* and the content, can be represented in the form of a graph (Fig. 5), which includes objects and relationships.

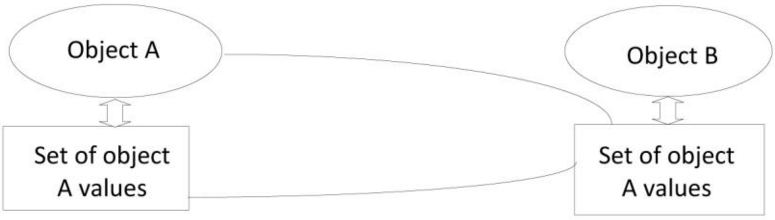


Fig. 5. The scheme of the examples of the formation of clarifying queries based on the concepts of the thesaurus and their relationships.

Integration of data of various origin allows you to implement the pre-property of query expansion when searching for mathematical texts. For example, if similar formulas are found in publications, then expanding the query through formulas will help to increase the search coverage and obtain a more complete amount of relevant information.

In Fig. 6, an example of a user request “Riccati equation” is given. This concept is found both in the thesaurus in the subject field "ordinary differential equations" (ODE), and in the mathematical encyclopedia. Information from both sources was presented to the user in an integrated form so that he could select clarifying information. As can be seen from the diagram, for accessing Russian-language objects (2) and (3) as a result, there are several ways for them to relate to the concepts of the thesaurus. Without taking into account relations with formulas that act as independent information objects, inclusion of object (1) in the result set is not possible.

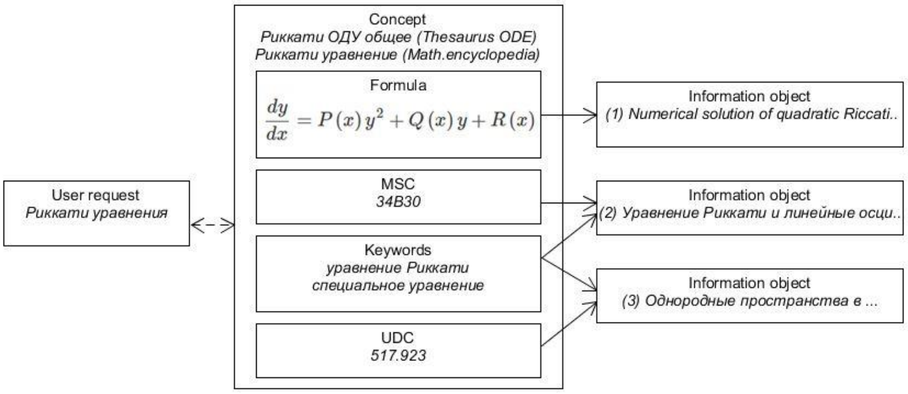

Concept
Риккати ОДУ общее (Thesaurus ODE)
Риккати уравнение (Math.encyclopedia)
Formula
$\frac{dy}{dx} = P(x)\,y^2 + Q(x)\,y + R(x)$
Information object
(1) Numerical solution of quadratic Riccati..
User request
Риккати уравнения
MSC
34B30
Information object
(2) Уравнение Риккати и линейные осци..
Keywords
уравнение Риккати
специальное уравнение
Information object
(3) Однородные пространства в ...
UDC
517.923

Fig. 6. The diagram of the ways of forming clarifying queries based on the concepts of various thesauruses and their relationships.

4 Conclusions and Outlook

The ontological approach to presenting information in a digital environment allows the use of subject area semantics for search query extensions. New technologies for data recognition and analysis make it possible to include symbolic information and formulas in articles of subject thesauri, thereby expanding the possibilities of their use for indexing and searching publications. This allows a new perspective on the representation of mathematical knowledge in digital space. Taking into account the associative relations of terms and formulas in the thesaurus allows you to search not only in related fields, but in digital arrays of various fields of knowledge, without increasing search noise. These conclusions are quite expected, and the problems discussed in the work are relevant from the point of view of combining ontologies of individual areas of knowledge without increasing the search time.

In the given examples, the data of mathematical subject areas are used as typical for query expansion due to formulas in adjacent domains, which, of course, does not limit the query extension to other subject areas integrated in LibMeta. The LibMeta project implements links with any sources that meet the requirements of LOD, as well as content, and testing links with a linguistic database and mathematical encyclopedia. Research in this direction is the subject of further work.

This work was supported in part by the Russian Foundation for Basic Research, projects #18-00-00297komfi, #20-07-00324.

References

1. Voorhees, E.M.: Query expansion using lexical-semantic relations. In SIGIR 94. ACM 1994., pp. 61–69 (1994).
2. Golden, P., Shaw, R., Buckland, M.: Decentralized coordination of controlled vocabularies. In: Proceedings of the American Society for Information Science and Technology. Annual Meeting, Seattle, WA, USA. (2014), https://doi.org/10.1002/meet.2014.14505101146 77th ASIS&T.
3. Serebryakov, V. A., Ataeva, O. M.: Ontology of the digital semantic library LibMeta. Informatics and Applications. 12(1), pp. 2-10 (2018).
4. Vechtomova, O.: Query Expansion for Information Retrieval. In: LIU L., ÖZSU M.T. (eds.) Encyclopedia of Database Systems. Springer, Boston, MA. (2009), https://doi.org/10.1007/978-0-387-39940-9_947.
5. Salton, G.: The SMART retrieval system (Chapter 14). Prentice-Hall, Englewood Cliffs NJ. (Reprinted from Rocchio J.J. (1965). Relevance feedback in information retrieval. In Scientific Report ISR-9, Harvard University) (1971).
6. Spärck Jones, K.: Automatic keyword classification for information retrieval. Butterworths, London (1971).
7. van Rijsbergen, C.J.: A theoretical basis for the use of co-occurrence data in information retrieval. J. Doc. 33(2), pp. 106–119 (1977).
8. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval, Cambridge University Press, Cambridge (2008).
9. Qui, Y., Frei, H.: Concept based query expansion. SIGIR '93 Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information

retrieval, pp. 160–169. Pittsburgh, Pennsylvania, NY, USA. (1993) https://doi.org/10.1145/160688.160713.

10. Schütze, H.: Automatic Word Sense Discrimination // Computational Linguistics, Special Issue on Word Sense Disambiguation. 24(1), pp. 97–123 (1998), https://www.aclweb.org/anthology/J98-1004.pdf, last accessed 25.11.2019.
11. Larkey, L.S., Croft, W.B.: Combining classifiers in text categorization. SIGIR '96 Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 289–297 Zurich, Switzerland, (1996). https://doi.org/110.1145/243199.243276.
12. Zentralblatt MATH https://zbmath.org, last accessed 25.11.2019.
13. Muromskij, A.A., Tuchkova, N.P.: Ob ontologii adresata v matematicheskoj predmetnoj oblasti. Russian Digital Libraries Journal, 21(6), pp. 506–533 (2018).
14. Moiseev, E.I., Muromskij, A.A., Tuchkova, N.P.: O tezauruse predmetnoj oblasti smeshannye uravneniya matematicheskoj fiziki. CEUR Workshop Proceedings. 2260, pp. 395–405 (2018) https://doi.org/10.20948/abrau-2018-43.
15. Ataeva, O.M., Serebryakov,V.A., Tuchkova, N.P.: Podhody k organizacii matematicheskih znanij pri formirovanii predmetnyh tezaurusov razlichnyh razdelov matematiki. CEUR Workshop Proceedings. 2260, pp. 42–54 (2018). https://doi.org/10.20948/abrau-2018-66.
16. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data – The Story So Far. International Journal on Semantic Web and Information Systems. 5(3) (2009) URL: https://eprints.soton.ac.uk/271285/1/bizer-heath-berners-lee-ijswis-linked-data.pdf, last accessed 25.11.2019. https://doi.org/10.4018/jswis.2009081901.
17. Moiseev, E.I., Lihomanenko, T.N.: Sobstvennye funkcii zadachi Trikomi s naklonnoj liniej izmeneniya tipa // Differencial'nye uravneniya. 52(10), pp. 1375–1382 (2016).
18. Vinogradov, I.M.: Matematicheskaya entsiklopediya [Mathematical Encyclopedia]. Sovetskaya entsiklopediya Publ., Moscow (1979).