Информатика

Материалы международной научно-практической конференции «Наука, технологии и информация в библиотеках (LIBWAY-2018)»

УДК 025.45.05
ББК 78.364
DOI 10.20913/1815-3186-2018-4-106-110

ФОРМАЛЬНАЯ ГРАММАТИКА ИНДЕКСОВ УНИВЕРСАЛЬНОЙ ДЕСЯТИЧНОЙ КЛАССИФИКАЦИИ[1]

© В. Н. Белоозеров, А. В. Шапкин, 2018

Всероссийский институт научной и технической информации Российской академии наук, Москва, Россия; e-mail: nomoip@viniti.ru

Предложен алгоритм анализа и представления на естественном языке смысла сложных индексов Универсальной десятичной классификации (УДК). Алгоритм основан на формальном определении правильных индексов с помощью порождающей грамматики, задающей перечень структур, начиная от простых табличных кодов классов УДК, к которым последовательно добавляются отдельные символы, вспомогательные определители и самостоятельные индексы с обозначением отношений классов, соединяемых в сложном индексе.

Значения анализируемых сложных индексов выражаются наименованиями и примечаниями табличных классов, входящих в структуру анализируемого индекса. Описания классов сопровождаются логическими связками, основанными на функциях вспомогательных символов УДК и позволяющими составить представление о связи обозначаемых индексом понятий. Действие алгоритма изложено на примере анализа конкретного комбинированного индекса.

Предлагаемый алгоритм решает не только задачу визуализации значения сложных индексов, но и задачу выделения из него самостоятельных смысловых фрагментов, которые могут служить ключами для расширенного тематического поиска.

Ключевые слова: Универсальная десятичная классификация, индексы УДК, формальная грамматика индексов УДК, алгоритм перечисления индексов УДК, алгоритм расшифровки индексов УДК

Для цитирования: *Белоозеров В. Н., Шапкин А. В.* Формальная грамматика индексов Универсальной десятичной классификации // Библиосфера. 2018. № 4. С. 106–110. DOI: 10.20913/1815-3186-2018-4-106-110.

Indices formal grammar of the Universal Decimal Classification

V. N. Beloozerov, A. V. Shapkin

Russian Institute for Scientific and Technical Information Russian Academy of Sciences, Moscow, Russia; e-mail: nomoip@viniti.ru

The article proposes an algorithm for decoding and representation in natural language of the Universal Decimal Classifycation (UDC) complex class numbers. The algorithm is based on the formal definition of correct class numbers using a generative grammar, which sets the list of structures starting with simple table codes of UDC classes. Then separate integers, auxiliary and independent class numbers are sequentially attached to the codes with special symbols of relations of classes, which compose the complex class number.

The algorithm expresses the values of the analyzed complex indices by descriptions (names and notes) of the table classes included in the structure of the analyzed string. The class descriptions are accompanied with the logical connectors based on the functions of the auxiliary characters. They provide the idea on connection of concepts denoted in the class number.

The algorithm action is described evidently for the analysis of combined index 539.4.019: [535-15+537.8.029.6]. The proposed algorithm is applicable both to visualize the meaning of complex class numbers, and to ensure the completeness and accuracy of the documents retrieval by the UDC classes.

Keywords: Universal Decimal Classification, UDC class numbers, formal grammar of UDC class numbers, UDC class numbers enumeration algorithm, UDC class numbers decoding algorithm

Citation: *Beloozerov V. N., Shapkin A. V.* Indices formal grammar of the Universal Decimal Classification. *Bibliosphere. 2018. № 4. P. 106–110.* DOI: 10.20913/1815-3186-2018-4-106-110.

Работа с аналитико-синтетическими классификациями, каковыми являются Библиотечно-библиографическая классификация (ББК) и Универсальная десятичная классификация (УДК), требует определенной квалификации для составления и понимания классификационных индексов, которой

[1] Работа выполнена в рамках проекта РФФИ № 17-07-00153 «Исследование системы классификаторов по науке и технике и разработка механизма смысловой навигации и поиска знаний в информационных сетях».

авторы, издатели и библиотечные работники зачастую не имеют. Помочь делу могло бы программное обеспечение для автоматизации создания, проверки и расшифровки индексов. Что касается формирования индексов УДК, то в ГПНТБ России и ВИНИТИ РАН разработаны программы, позволяющие компоновать индекс на основе лексического поиска в классификационной таблице. Но задача распознавания смысла символьного индекса осложняется трудностью вычленения из него кодов табличных рубрик, значение которых было бы указано в классификационной таблице. Автоматическая разборка сложного комбинированного индекса на табличные коды решает задачу помощи в осознании смысла индексов и открывает возможность многоаспектного поиска информации, когда релевантность документа запросу устанавливается не только по полному совпадению индексов, но и по совпадению их отдельных компонентов, выражающих разные аспекты и элементы содержания документа.

Автоматический анализ индекса возможен, если индекс составлен в соответствии с четкими правилами. Хотя правила формирования индексов УДК имеются в каждом томе изданных таблиц [5] и в ряде руководств [3, 6, 10], они достаточно сложны и подчас трактуются индексаторами по-разному. Обеспечить однозначность понимания правил может принятие формального алгоритма построения индексов, основанного на системе «правил переписывания (rewriting rules)» (см. напр. [8, 9]), которые определяют форму сложных индексов на основе комбинирования простых. Формирование знаковых объектов по правилам переписывания, называемым также порождающей грамматикой (generative grammar), применяется во многих областях – от обработки текстов до музыкальной композиции [11]. Но для индексов УДК вариант такого алгоритма впервые предложен в работах В. Н. Белоозерова [1, 2].

Правильные индексы

Согласно этому алгоритму множество правильно образованных индексов УДК *U* строится (если отвлечься от разделительных точек, служащих для облегчения зрительного восприятия индекса) на основе «простых индексов», которыми могут быть все конечные цепочки десятичных цифр:

$$\{\text{Простые}\} = \{ц_1ц_2,...ц_n\} \subseteq \boldsymbol{U},$$

где $ц_i \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$, $n \in [1, n]$.

«Сложные индексы» образуются путем комбинирования «простых индексов»: {Сложные} = $\{[И_1]\times[И_2]\}$, где $И_1$ и $И_2$ – произвольные индексы, входящие в ***U*** (как простые, так и сложные), а знак × обозначает один из служебных символов комбинирования основных и вспомогательных классов по правилам УДК: плюс, двоеточие, дефис, апостроф, скобки, знак равенства, кавычки, астериск, сочетания точка-нуль, дефис-нуль и скобка-равно. Имеются также правила дополнения и упрощения индексов, на которых мы останавливаться не будем.

Приведем примеры индексов различной структуры.

- Простой индекс: ***538.945 Сверхпроводимость***
- Индекс, усложненный специальным определителем:

  537.8.029 Частотные диапазоны электромагнитных волн
  535-15 Инфракрасные лучи
- Индекс, усложненный общим определителем: **528.9(571.14) Картография Новосибирской области**
- Индекс, охватывающий тематику двух классов: **532+533 Механика текучих сред (жидкостей и газов)**
- Индекс тематики, входящей в два класса: **535:621.315.592 Оптика полупроводниковых материалов**
- Индекс, объединяющий два сложных индекса: **535-15+537.8.029.6 Физика и оптика СВЧ и инфракрасного излучения**
- Индекс третьей ступени усложнения: **539.4.019:[535-15+537.8.029.6] Влияние СВЧ и инфракрасного излучения на прочность материалов.**

Расшифровка индекса

Индексирование по УДК научных публикаций предназначено для того, чтобы впоследствии можно было бы найти соответствующие сведения по смыслу, зашифрованному в классификационном индексе. Однако нам не известны работы, ставящие перед собой задачу рассмотрения способов выявления смысла сложного индекса, который не сводится к табличному коду класса, отражает тематику публикации с указанием различных частных тем и аспектов. Некоторые указания на способы смыслового анализа индексов УДК можно найти в работе О. А. Антошковой и др. [4, с. 106–107].

Задача – показать, как на основе порождающей грамматики УДК строится алгоритм расшифровки смысла сложных индексов.

На сайте ВИНИТИ РАН действует программа расшифровки индексов УДК, которая отыскивает в анализируемом индексе фрагменты, соответствующие кодам табличных классов, выдает на экран их наименования [7]. Но информация о связях этих классов друг с другом теряется. Предлагаемый ниже алгоритм выдает информацию о табличных классах в сопровождении логических связок, основанных на функциях вспомогательных символов УДК.

Алгоритм будет выражать значения индексов наименованиями табличных классов, входящих в структуру анализируемого индекса. Поэтому система анализа должна содержать не только программу, реализующую алгоритм, но также рабочую классификационную таблицу УДК. При этом желательно использовать не универсальную таблицу официального эталона УДК, а именно свою рабочую таблицу, оптимизированную за счет исключения посторонней тематики и включения комбинированных индексов, представляющих классификационные решения, принятые в данном информационном органе.

Излагать алгоритм будем не путем декларирования и разъяснения операций, а путем показа сообщений, которые целесообразно выводить на выходную форму.

Очевидно, что перед началом обработки заданного индекса следует вывести его как заголовок последующих сообщений. В качестве материала для изложения алгоритма возьмем последний из приведенных выше примеров. Поэтому в выходную форму записываем такую строку:

«***Расшифровка индекса УДК 539.4.019:[535-15+537.8.029.6]***», а сам индекс записываем в буферный файл ***Б***, хранящий выражения, подлежащие непосредственному анализу.

Сам алгоритм начинается с выборки анализируемого выражения из буфера ***Б*** и поиска его в рабочей таблице УДК. Если он там нашелся, то в выходную форму выписывается (с новой строки) вся информация, относящаяся к этому индексу (наименование и примечания). Таким образом, если этот индекс имеется в рабочей таблице, то в выходной форме появляется строка «***Влияние СВЧ и инфракрасного излучения на прочность материалов***».

Если же анализируемое выражение отсутствует в рабочей таблице, то работа алгоритма разветвляется в зависимости от характера первого знака в индексе. Если первый знак – цифра, то алгоритм должен выделить простой индекс, с которого (как в нашем случае) начинается сложный индекс и который, скорее всего, имеется в рабочей таблице. Для этого нужно в цепочке символов найти служебный символ, делящий индекс на независимые части. Знак плюс соединяет наиболее независимые друг от друга классы. Этот знак, не включенный в квадратные скобки, ищем в цепочке символов в первую очередь, а предшествующую ему часть индекса разыскиваем в рабочей таблице классификации. Другие служебные символы будем отыскивать в индексе после обработки знака плюс в последовательности нижеприведенной таблицы обработки служебных символов. В буферном файле при этом остается правая часть индекса, начинающаяся с данного служебного символа.

В нашем примере знак плюс, не включенный в квадратные скобки, отсутствует, но мы обнаруживаем знак двоеточия и делим индекс по этому символу. Если бы в нашем примере не было двоеточия, то алгоритм отыскивал бы символы, следующие за двоеточием в таблице обработки служебных символов.

Таблица обработки служебных символов

Table of auxiliary symbols processing

Очередь обработки	Символ	Выводимый текст	Примечание
1	/	совместно	Переход в ветвь обработки диапазонов
2	+	а также (и/или)	
3	:	в аспекте, в сочетании с	
4	::	со свойством	
5	[	совместно	Обрабатываются символы до закрывающей скобки]
6	(=	народ	Обрабатываются символы до закрывающей скобки)
7	=	язык документа	
8	(	при этом имеем	Обрабатываются символы до закрывающей скобки)
9	«	время	Обрабатываются символы до закрывающих кавычек »
10	*	заимствованный код	

Каждый раз, когда выделенная из исходного индекса левая часть обнаруживается в классификационной таблице в качестве табличного индекса, она выводится на выходную форму вместе с наименованием и комментариями к этому индексу. В нашем примере это будет следующий текст (индекс присутствует в эталонных таблицах УДК):

539.4.019 Различные воздействия.

Этот текст уже кое-что говорит о смысле расшифровываемого индекса. Но поскольку в УДК зачастую (как в данном случае) описание класса предполагает знакомство с содержанием вышестоящего класса, выводим также и класс с индексом, в котором исключена последняя цифра (если последний знак есть цифра). Этот текст предваряем словами «*из класса*», напечатанными в отдельной строке. В нашем примере получается:

539.4.019 Различные воздействия.
из класса
539.4.01 Теория прочности. Сила сцепления молекул между собой. Различные воздействия на прочность.

Далее алгоритм переходит к анализу оставшейся в буфере правой части исходного индекса, которая начинается со служебного символа. В зависимости от этого символа в выходную форму записывается строка, текст которой указан в таблице служебных символов. В нашем примере это будет «в аспекте». Сама строка символов из буфера без начального

служебного символа поступает на поиск в классификационной таблице, и при обнаружении ее повторяется вышеуказанный вывод текста табличных индексов. Но если, как в нашем примере, поиск в таблице не удается, анализируемая цепочка символов подвергается снова описанной выше операции выделения начальной табличной части. В нашем примере наличие знака квадратной скобки приведет к записи в выходной форме строки «*совместно*» (или как-нибудь иначе, если кто придумает что-либо получше), а найденная в таблице левая часть выражения в скобках даст текст:

535-15 Инфракрасные лучи
из класса
535-1 Длинные волны. Инфракрасные лучи.

Знак плюс приведет к выдаче строки «*а также (и/или)*», а правая часть выражения в скобках даст текст:

537.8.029.6 Сверхвысокие частоты
из класса
537.8.029 Частотные диапазоны электромагнитных волн.

В случаях, когда цепочка символов не обнаруживается в рабочей таблице УДК и в ее составе нет разделяющих служебных символов, на печать выдается строка «***В таблице не обнаружено. Ошибка индекса***».

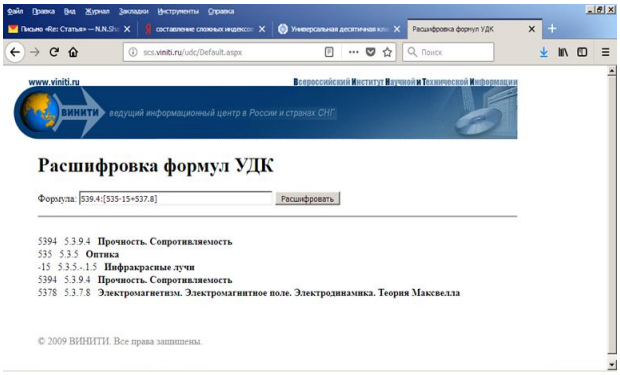
В целом выдача по нашему примеру имеет следующий вид:

Расшифровка индекса
УДК ***539.4.019:[535-15+537.8.029.6]***

539.4.019 Различные воздействия.
из класса
539.4.01 Теория прочности. Сила сцепления молекул между собой. Различные воздействия на прочность.
в аспекте
совместно
535-15 Инфракрасные лучи
из класса
535-1 Длинные волны. Инфракрасные лучи
а также (и/или)
537.8.029.6 Сверхвысокие частоты
из класса
537.8.029 Частотные диапазоны электромагнитных волн

По такой почленной расшифровке нетрудно сформулировать значение индекса на естественном языке: **«Влияние на прочность материалов сверхвысокочастотных и инфракрасных электромагнитных волн».**

В настоящее время в ВИНИТИ этот алгоритм реализован не в полной мере, но для сравнения ниже на рисунке приведен результат анализа менее сложного индекса, полученного из нашего примера путем удаления определителей с символом точка-нуль, так как опознавание их не реализовано в программе ВИНИТИ.



Пример расшифровки индекса в системе ВИНИТИ

An example of complex class number decoding by VINITI system

Заключение

Предлагаемый алгоритм решает не только задачу визуализации значения сложных индексов, но и задачу обеспечения полноты и точности поиска по индексам УДК. Он позволяет анализировать индекс УДК, выделяя из него самостоятельные смысловые фрагменты, которые могут служить ключами для тематического поиска в массивах документов, индексируемых классами УДК с разных точек зрения и с разной подробностью. Выявляемые в ходе анализа табличные индексы могут использоваться как расширители поискового образа. Для визуального восприятия смысла анализируемого индекса предусмотрен вывод на выходную форму наименований смысловых фрагментов с обозначением их логических связей.

В статье изложена только идея алгоритма. В реальной компьютерной программе должны быть предусмотрены многочисленные особые случаи и тонкости УДК, для чего требуется программирование достаточно высокого уровня, которое предполагается осуществить в ходе развития сервиса по расшифровке индексов УДК на сайте ВИНИТИ РАН.

Список источников

1. *Белоозеров В. Н.* Алгоритм построения индексов УДК // Перспективные направления исследований и критические технологии в классификационных системах : материалы науч.-практ. конф. (25–27 окт. 2017 г.). Москва, 2017. С. 36–39. URL: http://www.udcc.ru/MATERIALS/2017/CONFERENCE 2017.pdf (дата обращения: 13.09.2018).
2. *Белоозеров В. Н.* Правила алгоритмического порождения индексов УДК для тематической классификации информационных ресурсов // Научно-техническая информация. Серия. 2. Информационные процессы и системы. 2018. № 5. С. 32–38.
3. ГОСТ 7.90–2007 Система стандартов по информации, библиотечному и издательскому делу. Универсальная десятичная классификация. Структура, правила использования и индексирования. Москва : Стандартинформ, 2007. 22 с.
4. Индексирование фундаментальных научных направлений кодами информационных классификаций: Универсальная десятичная классификация / О. А. Антошкова, Т. С. Астахова, В. Н. Белоозеров [и др.]. Москва : ВИНИТИ РАН, 2010. 322 с.

5. УДК. Универсальная десятичная классификация : полное 4-е изд. Т. 1–10. ВИНИТИ РАН. Москва, 2000–2010.
6. Учебное пособие по универсальной десятичной классификации / ВИНИТИ РАН ; гл. ред. Ю. М. Арский. 3-е изд., испр. и доп. Москва, 2014. 185 с.
7. *Шапкин А. В.* Расшифровка формул УДК / ВИНИТИ РАН. URL: http://scs.viniti.ru/udc/Default.aspx (дата обращения: 13.09.2018).
8. *Baader F., Nipkow T.* Term rewriting and all that. Cambridge : Cambridge Univ. press, 1999. 316 p.
9. *Besem M., Klop J. W., De Vrije R.* [et al.]. Term rewriting systems («TeReSe»). Cambridge : Cambridge Univ. press, 2003. 884 p.
10. *McIlwaine I. C.* The universal decimal classification: guide to its use. Hague : UDC Consortium, 2000. 280 p.
11. *Rohrmeier M. A*. generative grammar approach to diatonic harmonic structure // Proceedings SMC'07, 4th Sound and Music Computing Conference (11–13 July, Lefcada, Greece). URL: https://ru.scribd.com/document/190067187/ (accessed 13.09.2018).

References

1. Beloozerov V. N. An algorithm for UDC class numbers construction. *Perspectivnye napravleniya issledovanii i kriticheskie tekhnologii v klassifikatsionnykh sistemakh : materialy nauch.-prakt. konf. (25–27 okt. 2017 g.).* Moscow, 2017, 36–39. URL: http://systemling.narod.ru/class/17-07-00153/materialy_seminara.pdf (accessed 13.09.2018). (In Russ.).
2. Beloozerov V. N. Rules of algorithmic generation of UDC class numbers for the information resources thematic classification. *Nauchno-tekhnicheskaya informatsiya. Seriya 2. Informatsionnye protsessy i sistemy*, 2018, 5, 32–38. (In Russ.).
3. *GOST 7.90–2007 Sistema standartov po informatsii, bibliotechnomu i izdatel'skomu delu. Universal'naya desyatichnaya klassifikatsiya. Struktura, pravila ispol'zovaniya i indeksirovaniya* [National Standard 7.90–2007. System of standards on information, librarianship and publishing business. Universal decimal classification. The structure, rules of indexing use]. Moscow, STANDARTINFORM, 2007. 22 p. (In Russ.).
4. Antoshkova O. A., Astakhova T. S., Beloozerov V. N. et al. *Indeksirovanie fundamental'nykh nauchnykh napravlenii kodami informatsionnykh klassifikatsii: Universal'naya desyatichnaya klassifikatsiya* [Indexing of fundamental scientific branches by codes of information classifications: Universal Decimal Classification]. Moscow, 2010. 322 p. (In Russ.).
5. Astakhova T. S. (ed.). *UDK. Universal'naya desyatichnaya klassifikatsiya : polnoe 4 izd. T. 1–10* [UDC. Universal Decimal Classification : full 4th ed. 1–10]. Moscow, 2000–2010. (In Russ.).
6. Arskii Yu. M. (ed.). *Uchebnoe posobie po Universal'noi desyatichnoi klassifikatsii* [Training manual on the Universal Decimal Classification]. 2nd ed. Moscow, 2014. 185 p. (In Russ.).
7. Shapkin A. V. Decoding of UDC formulas. *VINITI RAN.* URL: http://scs.viniti.ru/udc/Default. aspx (accessed 13.09.2018). (In Russ.).
8. Baader F., Nipkow T. Term rewriting and all that. Cambridge, Cambridge Univ. press, 1999. 316 p.
9. Besem M., Klop J. W., De Vrije R. [et al.] Term rewriting systems («TeReSe»). Cambridge, Cambridge Univ. press, 2003. 884 p.
10. McIlwaine I. C. The universal decimal classification: guide to its use. Hague, UDC Consortium, 2000. 280 p.
11. Rohrmeier M. A generative grammar approach to diatonic harmonic structure. *Proceedings SMC'07, 4th Sound and music computing conference (11–13 July, Lefcada, Greece).* URL: https://ru.scribd.com/document/190067187/ (accessed 13.09.2018).

Материал поступил в редакцию 09.10.2018 г.

Сведения об авторах: *Белоозеров Виктор Николаевич – кандидат филологических наук, доцент, ведущий научный сотрудник ВИНИТИ РАН, ORCID: 0000-0002-4200-1410,*

Шапкин Александр Владимирович – кандидат технических наук, заведующий отделом ВИНИТИ РАН, ORCID: 0000-0001-9714-4526