DOI 10.32460/ion_nt-2018-0009
УДК 025.4::[53+621.38]
ББК 78.653.4:78.364:22.3:32

О разработке классификационно-тезаурусной онтологии для предметной области физики и радиоэлектроники

Белоозеров В. Н. (ВИНИТИ РАН, Москва)
Шабурова Н. Н. (Научная библиотека ИФП СО РАН, Новосибирск)

Описывается комплекс работ по моделированию онтологии информационного пространства области знания «физика и радиоэлектроника» на основе сопоставления тематических рубрик библиографических классификационных систем в табличном и тезаурусном формате. Ставится задача создания единой семантической сети терминов индексирования, которая могла бы обеспечить доступ к содержанию разнородных информационных ресурсов, систематизированных по различным схемам классификации и (или) предметного индексирования.

Ключевые слова: классификационные системы, смысловые связи, сопоставление классификаций, тезаурус тематических рубрик, онтология информационного пространства, физика, электроника

The article describes the complex of works on modeling of ontology of information space of physics and electronics field of knowledge on the basis of comparison of thematic headings of bibliographic classification systems in tabular and thesaurus formats. The task is to create a single semantic network of indexing terms, which could provide access to the content of heterogeneous information resources, systematized according to different schemes of classification and(or) subject indexing.

Keywords: classification systems, semantic links, mapping of classifications, thesaurus of thematic headings, ontology of information space, physics, electronics

Введение

Необходимость искать информацию в пространстве разобщённых источников ставит задачу согласования средств, с помощью которых систематизированы сведения в разнородных ресурсах. Эта задача вхо-

дит в число актуальных направлений работ по созданию «семантического вэба» [11, с. 239–243]. Современные поисковые машины ориентированы на поиск по «свободной лексике», т. е. используют «мешочную грамматику слов» [12] и игнорируют вложенный в информационных ресурсах труд индексирования классификационными системами и ключевыми словами. Поскольку перспективы автоматического извлечения смысла из текста весьма туманны, реализация семантического веба нереальна без использования представления содержания документов средствами классификаций и ключевых слов. Однако, в настоящее время отсутствует единство средств описания тематики. Часть библиотек использует Универсальную десятичную классификацию (УДК), часть – Библиотечно-библиографическую классификацию (ББК), патенты эффективнее искать по патентной классификации, материалы по физике – с помощью Системы классификации по физике и астрономии (Physics and Astronomy Classification Scheme – PACS) и т. д. Эта ситуация может быть преодолена установлением сети связей рубрик используемых классификационных систем при привязке к ним ключевых слов, индексирующих документы данной рубрики. Для физики и электроники мы начали осуществлять эту идею с 2008 г. [3].

Тезаурус тематических рубрик

Сопоставление нескольких классификаций в области физики полупроводников привело к идее тезауруса классификационных рубрик, в котором дескрипторами были бы сами рубрики, а также термины из наименований рубрик нескольких классификаций [5], включая УДК, ББК и Государственный рубрикатор научно-технической информации (ГРНТИ).

Такой тезаурус – Тезаурус тематических рубрик по физике полупроводников (ТТРФПП), содержащий около 2000 дескрипторов, разработан, выложен на сайте научной библиотеки ИФП СО РАН и депонирован в ВИНИТИ [7]. В Приложении 1 показаны образцы статей тезауруса. Каждому дескриптору приписаны коды пяти классификаций и даны обычные тезаурусные связи дескрипторов.

В дальнейшем тематика физики полупроводников была дополнена терминами смежных областей физики, электроники и нанотехнологий, а модель тезауруса предложена для построения онтологий других предметных областей [6].

В рамках работ по двум проектам РФФИ № 17-07-00153 и РГНФ № 17-03-12013 реализуется именно эта модель как широкая система классификаций и ключевых слов для различных областей знания.

Постановка задачи и начальный этап работ по созданию базы данных «Термин», в которой реализована сеть смысловых связей ключевых слов, реально использованных для индексирования научных публикаций, описаны в статьях [1, 9]. Создаваемая семантическая сеть состоит из 63 тезаурусов с перекрёстными связями по тематике всех основных разделов ГРНТИ. Сеть включает в частности тезаурусы «Физика» и «Электроника», соответствующие разделам ГРНТИ 29 и 47 соответственно. Однако, если разработка ТТРФПП шла «снизу», от мелких деталей выделенной тематики, зафиксированных в подробном Рубрикаторе ВИНИТИ, то новые проекты идут «сверху», от лексики верхних разделов классификаций. Поэтому объём лексики всего раздела физики и электроники пока не превосходит 1200 словарных единиц (примерно 700 по физике и 500 по электронике). Целесообразно поставить вопрос о массовом автоматическом вводе лексики Тезауруса в систему «Термин». Пока же материал ТТРФПП использовался только в режиме ручного установления связей терминов. При этом следует отметить, что если в ТТРФПП термины происходят исключительно из классификационных таблиц, то в основе лексики проекта РФФИ лежат ключевые слова, выделенные индексаторами как наиболее важные для описания содержания документов по данной области знания. Такой выбор терминов обеспечивает привязку нашей системы к реальным метаданным описания тематики документов.

Таблицы соответствия классификаций

Одной из главных задач разработки онтологии информационного пространства является задача тематической навигации среди ресурсов, систематизированных разными классификациями. Для этой задачи сопоставление классификаций удобнее производить не в тезаурусном, а в табличном формате, когда таблица показывает непосредственно связи от одного классификатора к другому. В статьях [2, 10] описана методика создания Сети классификационных систем ВИНИТИ путём полуавтоматического установления прямых смысловых связей между всеми классификациями, для которых в базе данных ВИНИТИ указано

соответствие рубрикам ГРНТИ. Методика предусматривает ручное редактирование и пополнение автоматически установленных связей. В экспериментальном режиме такая работа проводится для нескольких выделенных разделов ГРНТИ – физика, электроника, информатика, экономика и стандартизация. Для наиболее важных в библиотечной практике классификаций – ББК, УДК и ГРНТИ поученные таблицы соответствий по разделам физики и электроники депонированы в ВИНИТИ [4]. В Приложении 2 к статье приведено начало таблицы УДК-ББК: слева – классы УДК, справа – ББК в двух вариантах (буквенном и цифровом). Посередине – символ вида смысловой связи рубрик.

Характерной особенностью таблицы является изобилие соответствий между комбинированными индексами, не представленными в эталонных таблицах. Такие индексы присваиваются документам в ходе каталогизации для обозначения их тематики, если она захватывает значения различных классов из эталонных таблиц. В настоящее время Сеть классификационных систем ВИНИТИ расширила свою функциональность и теперь способна обрабатывать и комбинированные индексы. Это позволяет загрузить туда наши таблицы.

Особенности тезаурусов базы данных «Термин»

Возвращаясь к тезаурусному представлению соответствий, нужно отметить, что установление смысловых связей между терминами физики и электроники в базе данных «Термин» БЕН РАН не предполагает автоматического объединения с прежним Тезаурусом тематических рубрик по полупроводникам (ТТРПП) и его дальнейшим развитием. Если ТТРПП был ориентирован только на документный поиск, то теперь мы, следуя тенденциям современной информатики, ставим задачу построения в перспективе универсальной онтологии предметной области, которая будет пригодна для решения также задач фактографического поиска, управления, извлечения знаний и других задач искусственного интеллекта. Поэтому мы переосмыслили значения применяемых тезаурусных отношений. Они понимаются не в поисковом, а в онтологическом смысле – как пересечения не классов релевантных документов, а как пересечения классов описываемых в документах реалий. Поэтому отношение тождества терминов означает именно тождество денотатов, а не тождество массивов документов про них. Отношение «выше-ниже»

рассматривается как вхождение класса денотатов одного термина в класс денотатов другого. Это – более строгая интерпретация отношений, чем принято для документного поиска. Если между двумя понятиями А и Б установлено онтологическое отношение А=Б или А>Б, то это отношение будет справедливым и в поисковом смысле, но не наоборот.

Различие между двумя интерпретациями отношений можно пояснить следующим примером. В поисковом смысле часто антонимы можно рассматривать как эквиваленты. Например, все статьи о «неустойчивости плазмы» релевантны запросу об «устойчивости плазмы», поскольку критерии этих явлений совпадают, и их определения равно соответствуют и тому, и другому понятию. Но в онтологическом смысле эти понятия исключают друг друга и не могут рассматриваться как эквивалентные или пересекающиеся.

Что же касается «ассоциативного» отношения А×Б, то в онтологическом смысле оно понимается как наличие у денотатов общих атрибутов, что вполне сходится с его пониманием в поисковом смысле как пересечение массивов релевантных документов.

Базовый массив связей терминов в системе БЕН РАН был получен автоматически на основе «дефинитивного» поиска соответствий (употребление одного термина в составе дефиниции другого). Такой поиск не даёт возможности квалифицировать найденную связь по категориям видов тезаурусных отношений. Уточнение вида автоматически установленной связи по категориям «совпадение – вхождение – пересечение» осуществлялось интеллектуальным рассмотрением вручную. При этом существенную долю дефинитивных связей (примерно четверть) пришлось исключить как установленные из-за формального совпадения слов с семантически не связанными значениями (например, когда термины приписаны к понятию «*время*» из-за того, что в их определениях встретилось выражение *«в то время как»*).

Но и среди оставленных действительных связей не все целесообразно использовать при обычном документальном поиске. Например, понятия «автомобиль» и «бензин» явно связаны, и это может быть отражено в их определениях. Но использовать документы о бензине как релевантные для поиска документов об автомобилях (и наоборот) вряд ли целесообразно в общем случае. Однако если в поисковой системе будет реализован специфические модальности поиска *«источник энергии для»* или *«применяется в»*, то связь таких терминов будет

востребована. Поэтому мы такие связи не ликвидировали, а оставляли как особую категорию «слабых пересечений» в качестве кандидатов на установление специфических режимов поиска, учитывающих прагматических отношения объе4ктов онтологической реальности.

Таким образом, в словарях «Физика» и «Электроника» базы данных «Термин» БЕН РАН устанавливаются следующие виды связей терминов, показанные на таблице ниже:

Таблица 1. Смысловые связи сопоставляемых рубрик

Знак	Названия	Значение
А = Б	совпадает, равно, тождественно	Термины А и Б обозначают тождественные множества реалий (синонимы)
А>>Б	больше, шире, включает	Термин А обозначает множество реалий, в которое включено множество реалий, обозначаемых термином Б, при чём объёмы этих множеств соизмеримы.
А<<Б	меньше, уже, входит в	Термин А обозначает множество реалий, включённое во множество реалий, обозначаемых термином Б, при чём объёмы этих множеств соизмеримы..
А><Б	пересекается с	Множества реалий, обозначаемые терминами А и Б пересекаются в существенной части.
А – Б	дефинитивно связаны	Реалии, обозначаемые терминами А и Б, связаны прагматическими связями, но их множества, вероятно, не пересекаются, относясь к различным онтологическим категориям.

Этот набор типов связи отличает наши словари от других словарей в БД «Термин». Связи дескрипторов в словарях «Экономика», «Языкознание», «Информатика» и «Стандартизация» следуют традиционному пониманию связей в информационно-поисковых тезаурусах. Это различие целесообразно на этапе разработки экспериментальной версии инструмента построения информационных онтологий: оно позволяет провести сравнительное исследование эффективности различных моделей и выбрать оптимальную для промышленной реализации.

Ниже на рисунке (рис. 1) показана главная страница тезауруса «Физика» в базе данных «Термин», на которой открывается доступ к редактированию его содержимого. Включая добавление и уничтожение терминов

(711 единиц) определений, связей, индексов ББК, УДК и ГРНТИ. Об опыте работы с базой данных «Термин» на материале тезауруса «Электроника» рассказано в докладе на конференции LIBWAY [8].

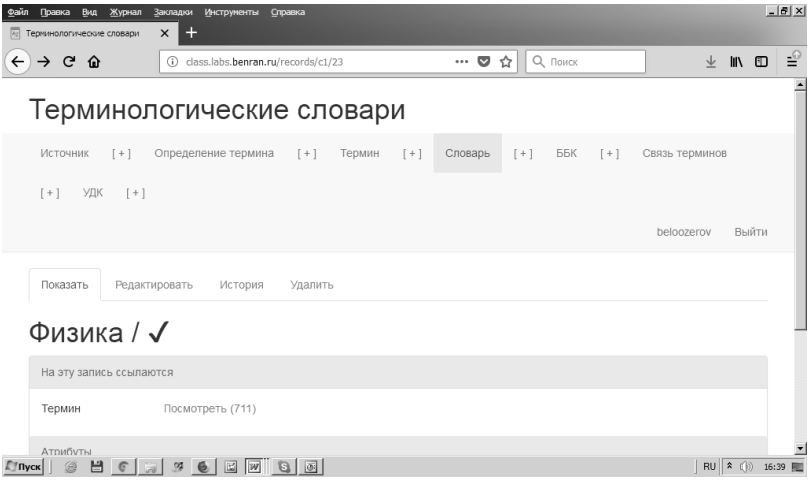

Рис. 1

Заключение

В настоящее время для обширной тематической области, включающей физику и радиоэлектронику, созданы семантические сети терминов, являющихся входными точками тематического поиска в различных информационных источниках, наделённых различными системами доступа – классификационными и предметными индексами. Сети терминов представлены в тезаурусном и в табличном вариантах, каждый из которых объединяет смысловыми связями наборы классификационных систем, не совпадающие друг с другом и различающиеся по деталям методики. На текущем этапе работ стоит задача объединения тезаурусного и табличного подходов к сопоставлению классификационных систем и построение тезаурусно-классификационной онтологии на единой материально-технической платформе (в БЕН и/или в ВИНИТИ). При этом необходимо решить следующие задачи:

- Пополнение БД «Термин» лексикой и связями тезауруса ИФП СО РАН.
- Редактирование связей Тезауруса ИФП СО РАН в соответствии с онтологическим подходом к их интерпретации.

- Введение в БД Термин» полных таблиц ББК.
- Установление связей ББК с актуальными международными классификациями (WoS, Scopus).
- Пополнение лексики ключевыми словами наиболее важных (наиболее цитируемых) работ.
- Обеспечение автоматического управления соответствием взаимообратных связей.
- Обеспечение автоматического управления иерархией (исключение дальних связей при наличии ближних).
- Решение вопроса о соответствии связей эквивалентности и ссылок «см.».

Литература

1. Антопольский А. Б. Разработка онтологии информационного пространства знаний на основе дефинитивных связей / А. Б. Антопольский [и др.] // Научно-техническая информация. Сер. 1. Организация и методика информационной работы. 2017. № 11. С. 19–24.

2. Антошкова О. А. Методика построения онтологии информационных ресурсов в виде сети библиографических классификаций / О. А. Антошкова [и др.] // Научно-техническая информация. Сер. 1. Организация и методика информационной работы. 2017. № 11. С. 25–30.

3. Белоозеров В. Н. Классификационные системы как средство поиска информации по физике полупроводников / В. Н. Белоозеров, Н. Н. Шабурова // Библиосфера. 2008. № 3. С. 34–42.

4. Белоозеров В. Н. Сопоставительные таблицы классификационных систем УДК и ББК по тематике «Радиоэлектроника и физика твёрдых тел» / В. Н. Белоозеров, Н. Н. Шабурова, О. В. Смирнова. Депонировано ВИНИТИ РАН 24.01.2018. № 9-В2018. 38 с.

5. Белоозеров В. Н. Сопоставительный тезаурус классификационных систем по физике полупроводников / В. Н. Белоозеров, Н. Н. Шабурова // Информационное обеспечение науки: новые технологии : сб. науч. тр. Москва : Научный Мир, 2009. С. 311–322.

6. Белоозеров В. Н. Тезаурус библиографических классификаций как модель интеграции информационных ресурсов / В. Н. Белоозеров, Н. Н. Шабурова // Информационное общество: состояние и тенденции межгосударственного обмена научно-технической информацией в

СНГ : междунар. конф. (27–28 окт. 2011 г., Москва). Москва : ВИНИТИ, 2011. С. 8–9.

7. Белоозеров В. Н. Тезаурус тематических рубрик по физике полупроводников / В. Н. Белоозеров, Н. Н. Шабурова. Депонировано в ВИНИТИ РАН 24.12.2013. № 379-В2013.

8. Шабурова Н. Н. Опыт работы с БД «TERMIN» / Н. Н. Шабурова // Наука, технологии и информация в библиотеках (LIBWAY-2018) : междунар. науч.-практ. конф. (12-15 сент. 2018 г., Новосибирск). Режим доступа: https://www.libway.ru/program/Saeffd4 c7045cd0be895b01a.

9. Antopol'skii A. B. The Development of a Semantic Network of Keywords Based on Definitive Relationships / A. B. Antopol'skii [et al.] // Scientific and Technical Information Processing. 2017. Vol. 44, N 4. P. 261–265.

10. Antoshkova O. A. On a Method for Constructing an Ontology of Scientific and Technical Information as a Network of Bibliographic Classifications / O. A. Antoshkova [et al.] // Scientific and Technical Information Processing. 2017. Vol. 44, N 4. P. 266–272.

11. Bast H. Semantic Search on Text and Knowledge Bases / H. Bast, B. Buchhold, E. Haussmann // Foundations and Trends in Information Retrieval. 2016. Vol. 10, N 2–3. P. 119–271. DOI: 10.1561/1500000032.

12. Li H. Semantic Matching in Search / H. Li, J. Xu // Foundations and Trends in Information Retrieval. 2013. Vol. 7, N 5. P. 343–469. DOI: 10.1561/1500000035.

Приложение 1

Образцы статей Тезауруса тематических рубрик по физике полупроводников

антисегнетоэлектрики = *диэлектрики, не являющиеся сегнетоэлектриками, но обладающие определенной спецификой электрических свойств. Основной признак антисегнетоэлектрика - наличие структурного фазового перехода, сопровождающегося значительной аномалией диэлектрической проницаемости*
*ББК В379.331.5:331.7
ВИНИТИ 291.19.35
ГРНТИ 29.19.35
УДК 537.226.4
В: сегнетоэлектричество

антиферромагнетизм полупроводников = *одно из магнитных состояний вещества, отличающееся тем, что элементарные (атомные) магнитики соседних частиц вещества ориентированы навстречу друг другу (антипараллельно), и поэтому намагниченность тела в целом очень мала. Этим антиферромагнетизм отличается от ферромагнетизма, при котором одинаковая ориентация элементарных магнитиков приводит к высокой намагниченности тела.*
ББК В379.233.4
*УДК 537.611.45:537.622.5:621.315.59
%УДК 537.311.322.04:537.611.45
В: антиферромагнетизм – теория
В: антиферромагнетики и слабый ферромагнетизм
В: магнитные свойства полупроводников

антиферромагнетизм – теория
ББК В373.35
УДК 537.611.45.01
В: магнетизм – теория
Н: антиферромагнетизм полупроводников

Антиферромагнетики
См: антиферромагнитные материалы

антиферромагнетики и слабый ферромагнетизм
ББК В373.34/35
ВИНИТИ 291.19.43
ГРНТИ 29.19.43
УДК 548:537.611.45/.46
В: магнитные свойства твёрдых тел
Н: антиферромагнетизм полупроводников
Н: антиферромагнитные материалы

Продолжение приложения 1

антиферромагнитные материалы
УДК 537.622.5
С: *антиферромагнетики*
В: антиферромагнетики и слабый ферромагнетизм
В: магнитные материалы
В: магнитные свойства материалов

арсенид галлия /полупроводниковые свойства/
*ББК Г125.315:Г123.31-2:з843.3
*УДК 546.681’19:621.315.59
В: галлий и его соединения – полупроводниковые свойства
Н: полупроводниковые соединения

Приложение 2

Фрагменты таблицы соответствия УДК - ББК

УДК		Связь	ББК		
Индекс	Содер-жание		Индексы		Содержание
1	2	3	4	5	6
004.33	Блоки памяти	>	з971.31-045	32.971.31-045	Аналоговые запоминающие устройства
		>	з971.32-044	32.971.32-044	Запоминающие устройства [*цифровые*]
		>	з971.32-046.4	32.971.32-046.4	Внешние запоминающие устройства
		><	з871.9	32.871.9	Другие виды записи и воспроизведения звука
		><	з871	32.871	Запись и воспроизведение звука
		><	з871.2	32.871.2	Механическая запись и воспроизведение звука
004.35.085	*Оптические накопители*	<	з871.4	32.871.4	Оптическая и магнитооптическая запись и воспроизведение звука
		=	з971.32-046.41	32.971.32-046.41	Оптические и магнитооптические накопители
		><	з971.31-045	32.971.31-045	Аналоговые запоминающие устройства
		><	з971.32-044	32.971.32-044	Запоминающие устройства [*цифровые*]

… … …

Продолжение приложения 2

534.84	Акустика помещений ...	><	з872	32.872	Озвучивание и звукоусиление
534.85	Запись и воспроизведение звука	<	з871	32.871	Запись и воспроизведение звука
		>	з871.3	32.871.3	Магнитная запись и воспроизведение звука
		>	з871.9	32.871.9	Другие виды записи и воспроизведения звука
		><	з971.31-045	32.971.31-045	Аналоговые запоминающие устройства
		><	з971.32-044	32.971.32-044	Запоминающие устройства [*цифровые*]
		><	з971.32-046.4	32.971.32-046.4	Внешние запоминающие устройства
		><	з871.2	32.871.2	Механическая запись и воспроизведение звука
534.852	Магнитная запись	<	з871.3	32.871.3	Магнитная запись и воспроизведение звука
		>	з971.32-046.41	32.971.32-046.41	Магнитные накопители
		><	з971.31-045	32.971.31-045	Аналоговые запоминающие устройства
		><	з971.32-044	32.971.32-044	Запоминающие устройства [*цифровые*]
534.853	Фотографическая запись	<	з871.4	32.871.4	Оптическая и магнитооптическая запись и воспроизведение звука
		>	з971.31-045	32.971.31-045	Аналоговые запоминающие устройства
		><	з971.32-044	32.971.32-044	Запоминающие устройства [*цифровые*]
		><	з971.32-046.41	32.971.32-046.41	Оптические и магнитооптические накопители