Система классификационных схем для индексирования документов в области физики

System of Classification Schemes for Indexing Documents on Physics

Система класифікаційних схем для індексування документів у галузі фізики

В. Н. Белозеров
Всероссийский институт научной и технической информации РАН, Москва, Россия

Н. Н. Шабурова
Институт физики полупроводников СО РАН, Новосибирск, Россия

Viktor Belozerov
All-Russian Institute of Scientific and Technical Information of the Russian Academy of Sciences, Moscow, Russia

Nataliya Shaburova
Institute of Semiconductor Physics, Siberian Division of the Russian Academy of Sciences, Novosibirsk, Russia

В. М. Бєлозьоров
Всеросійський інститут наукової та технічної інформації РАН, Москва, Росія

Н. М. Шабурова
Інститут фізики напівпровідників СВ РАН, Новосибірськ, Росія

Сопоставление классификационных систем, используемых для тематического индексирования в области научной и технической информации, показало целесообразность разработки тезауруса, дескрипторами которого выступают наименования и понятия, соответствующие содержанию классификационных рубрик из различных классификаций. Такой тезаурус построен для области физики полупроводников и смежных проблем. Тезаурус оказался полезен для облегчения процесса индексирования публикаций и для повышения эффективности поиска данных. Приведены основные параметры тезауруса.

Comparison of classification systems used for subject indexing in the sphere of scientific and technical information showed practicability of developing thesaurus, in which names and notions corresponding to the content of classification sections of different classifications act as descriptors. Such thesaurus was developed for semiconductor physics and relating spheres. It turned out to be useful for facilitating the process of indexing publications and for increasing data search efficiency. Main characteristics of the thesaurus are represented.

Зіставлення класифікаційних схем, що використовуються для тематичного індексування в галузі наукової та технічної інформації, показало доцільність розробки тезауруса, дескрипторами якого виступають наймення та поняття, які відповідають змісту класифікаційних рубрик із різних класифікацій. Такий тезаурус побудовано для галузі фізики напівпровідників і суміжних проблем. Тезаурус виявився корисним для полегшення процесу індексування публікацій та для підвищення ефективності пошуку даних. Наведено основні параметри тезауруса.

По различным историческим причинам на сегодняшний день не существует единой классификационной системы тематического кодирования научной информации, их множество: Государственный рубрикатор научно-технической информации – ГРНТИ (обязателен для автоматизированных информационных систем), Универсальная десятичная классификация – УДК (используется в

московских библиотеках – БЕН РАН, ГПНТБ России; в централизованной библиотечной системе УрО РАН), Библиотечно-библиографическая классификация – ББК (применяется в централизованной библиотечной системе СО РАН), Рубрикатор ВИНИТИ (разработан ВИНИТИ для систематизации реферативной информации), Optical Classification and Indexing Scheme – OCIS, Physics and Astronomy Classification Scheme – PACS (используются многими международными издателями, в России – при издании англоязычных версий отечественных журналов). Такой разброс может затруднять взаимодействие фондов разных регионов: в некоторых случаях установить соответствия между различными классификациями удается только на самом общем уровне, при котором поиск становится бессодержательным. Также изначальное индексирование научных результатов в зависимости от требований одной публикующей организации может при последующем их поиске стать причиной потери необходимой информации из-за формулировки запроса в кодировке иной издающей организации.

С другой стороны, каждая из классификаций базируется на собственных подходах к характеристике объекта и выделяет в теме свои аспекты. Например, в ББК выделен специальный класс для описания свойств полупроводников и явлений, связанных с ними. В УДК соответствующая тематика рассыпана по разделам изучаемых явлений и их применений в технике. PACS, так же как и УДК, не имеет специального раздела физики полупроводников, эта тематика обозначена конкретными подрубриками тех разделов физики, для которых свойства полупроводников представляют существенный интерес. Наиболее подробно физика полупроводников разработана в Рубрикаторе ВИНИТИ, где ей посвящено более 200 рубрик (против 51 рубрики в PACS и 36 – в ББК). Разработанность тематики исключает необходимость использовать сочетание рубрик для обозначения вопросов, возникающих на стыках различных направлений исследования, они обычно уже отражены соответствующей подрубрикой данного раздела. Но в своей существенной части классификационные подразделения Рубрикатора выделены по отличным от ББК основаниям. Так, в последнем имеется несколько классов изучения структуры полупроводников, а в Рубрикаторе вопросы структуры рассматриваются в разных рубриках в связи с другими аспектами. С другой стороны, в ББК нет классов для общего рассмотрения кинетических эффектов, коллективных процессов, неоднородных систем и других вопросов, выделенных в Рубрикаторе на переднем плане.

Установить точное соответствие терминов всегда возможно на уровне конкретных понятий, несмотря на то, что в сравниваемых классификациях они отнесены к рубрикам с различным общим содержанием. Более того, именно совмещение индексов различных классификаций характеризует предмет исследования с разных точек зрения, что открывает возможность более точного поиска данных. Для отражения сложных смысловых отношений между пятью указанными выше классификационными системами в библиотеке ИФП СО РАН разработан информационно-поисковый тезаурус тематических рубрик по физике полупроводников[1], как технологический инструмент одновременного применения средств тезаурусного описания и классификационного индексирования на основе лексики классов по тематике одной области знания. В настоящее время ведётся работа по наполнению тезауруса понятиями смежных областей – физики наноструктур и спинтроники. Форма представления и методика разработки в целом соответствуют стандартам ИСО 2788 и ГОСТ 7.25-2001[2]. Согласно информационной теории и данным зарубежной практики[3] этот инструмент может быть языком-посредником и связующим звеном в сети взаимодействующих информационных ресурсов, которые обмениваются данными на основе тематического описания документов и информационных запросов.

Тезаурус предназначен для решения следующих задач:

---

[1] Тезаурус зарегистрирован Аналитической службой ведения информационных языков Государственной системы научно-технической информации (ГСНТИ) ВИНИТИ РАН - № 132.09, включен в фонд языковых средств ГСНТИ и рекомендован для использования в информационных органах.

[2] ГОСТ 7.25-2001 СИБИД. Тезаурус информационно-поисковый одноязычный. Правила разработки, структура, состав и форма представления. – М., 2002.

3 См. например U.S. National Library of Medicine. Fact sheet:. UMLS Metathesaurus / National Institutes of Health, 28 March 2006. [Электронный ресурс]. – Доступ: http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html (дата обращения 2009-05-07)

1. содержательное индексирование документов и запросов в интеллектуальном и автоматическом режиме,

2. поиск информации в базах данных, систематизированных по одной из классификационных систем, включённых в тезаурус,

3. поиск информации в фондах документов, заиндексированных ключевыми словами,

4. интеллектуальный поиск информации по полному тексту,

5. передача поисковых образов из одной информационной системы в другую,

6. интеграция разнородных информационных ресурсов в единую сеть с общими средствами для тематического и предметного доступа к данным.

Тезаурус используется для индексирования и поиска документов в научной библиотеке Института физики полупроводников СО РАН и размещён на сайте института.

Решение поставленных задач с помощью тезауруса достигается следующими путями.

1) Индексирование документов (или запросов).

Индексатор определяет тематику документа (как правило, по заглавию). Затем находит в тезаурусе дескрипторы, наиболее близкие по значению к формулировке темы документа. Если эти дескрипторы отражают тематику с достаточной точностью, они принимаются как поисковый образ данного документа. Классификационные коды, указанные в тезаурусе при дескрипторах, используются далее как индексы документа в соответствующих системах классификации. Если при дескрипторе не указан код желаемой классификации, то индексатор может использовать коды, указанные при вышестоящих и (или) ассоциативных дескрипторах, указанных в тезаурусной статье ссылками «выше» и «ассоциация». В некоторых случаях также можно использовать нижестоящие дескрипторы, указанные ссылками «ниже». Если действует программное обеспечение автоматического индексирования, выбирается из текста заранее определённое число дескрипторов, имеющих наибольший информационный вес в тексте.

2) Поиск информации в базе данных по одной из классификационных систем, включённых в тезаурус.

В тезаурусе отыскиваются дескрипторы, наиболее полно соответствующие тексту (смыслу запроса) путём интеллектуального анализа или автоматического индексирования. В качестве поискового предписания в базу данных предъявляются коды соответствующей классификации, указанные при найденных дескрипторах, а затем коды при ассоциативных дескрипторах и по всем цепям видовых (нижестоящих) дескрипторов.

3) Поиск информации в базах данных с доступом по ключевым словам.

В тезаурусе отыскиваются дескрипторы, наиболее близкие по смыслу к формулировке запроса. Найденные дескрипторы образуют поисковый образ запроса, который может быть дополнен нижестоящими дескрипторами из всех цепей родовидовых и партитивных ссылок «ниже».

4) Поиск информации в ресурсах с пословным индексированием (как в поисковых машинах Интернета).

Интеллектуальность информационного поиска может быть повышена добавлением к поисковому образу запроса терминов, связанных в тезаурусе любым типом отношений.

5) Сбор документов из разных источников.

Если информационный ресурс собирает документы из разных источников**,** надлежащие классификационные индексы могут быть выявлены по тезаурусу на основе тех индексов, которые имеются в исходных источниках, или с помощью процедуры, описанной выше в пунктах 1) и 2).

6) Интеграция разнородных информационных ресурсов в единую сеть с общими средствами тематического доступа.

Такая интеграция будет достигнута, если поисковый образ запроса формулировать сразу на языке всех классификационных систем, включённых в тезаурус (УДК, ББК, PACS и др.), на языке ключевых слов и на языке пословного индексирования. При этом наиболее точную и полную информацию даёт обращение именно к тому ресурсу сети, в классификационной системе которого обнаруживается по тезаурусу класс, точно соответствующий запросу.

Таким образом, тезаурус включает всего статей – 1023, всего элементов дескрипторных статей (индексов и связей) – 4167, определений – 163. Для выявления опосредованных связей классификационных систем дескрипторы заиндексированы по правилам многоаспектного индексирования, предусмотренным методиками применения ББК и УДК. Соответствующие комбинированные

индексы помечены астериском: *. Комбинированные коды УДК, использованные в рабочей таблице БЕН РАН, помечены знаком процентов %. Дескрипторы, значение которых не раскрывается формой термина, снабжены определениями. Основные параметры тезауруса указаны в таблицах 1–3.

Таблица 1

Наличие дескрипторов и их источники

PACS	164	Предметный указатель
ББК	343	Таблица ИФП
УДК	177	Выборка из эталона
% УДК	42	Рабочая таблица БЕН РАН
ГРНТИ	56	Раздел физики твердого тела
ВИНИТИ	242	Раздел физики полупроводников
Всего	1023	

Таблица 2

Наличие и количество комбинированных индексов

*УДК	308
*ББК	120

Таблица 3

Наличие и количество ссылок

С: (синоним) и См (смотри)	75
В: (выше) и Н: (ниже)	1300
А: (ассоциация)	138
Всего	2888

Применимость тезауруса для индексирования текущих научных публикаций показана в таблице 4 на примере поиска классификационных индексов для ряда докладов на недавно состоявшейся международной школе по физике полупроводников в Екатеринбурге (10-15 февраля 2010 г.). Индексы определялись по выше указанным схемам на основании формального совпадения слов заглавий докладов и дескрипторов тезауруса.

Таблица 4

Индексирование докладов по физике полупроводников[4]

Доклад (обозначение, автор, заглавие)	**Дескрипторы, найденные в тезаурусе**	**Классификационные индексы**
NM-03 G.V.Lashkarev	цинк и его соединения – полупроводниковые свойства	ББК В379.2
Zinc oxide as semiconductor material of nonrealized possibilities		ГРНТИ 29.19.31
		ВИНИТИ 291.19.31
	квантовая теория полупроводников	ББК В379.13
NM-06 V.V.Kabanov Magnetic quan-tum oscillations in doped anti-ferromagnetic semiconductors		УДК 530.145:621.315.592
		ГРНТИ 29.05.15
		ВИНИТИ 291.05.15
	легирование полупроводников	PACS: 61.72.–y
		ББК В379.1
	антиферромагнетизм полупроводников	ББК В379.233.4
		УДК 537.611.45:621.315.592
		ГРНТИ 29.19.43
		ВИНИТИ 291.19.43
	цинк и его соединения – полупроводниковые свойства	ББК В379.2
NM-10 M.Godlewski		УДК 538.9:546.47:621.315.592
Zinc oxide for photovoltaic and optoelec-tronic applications		
	фотовольтаический эф-фект в полупроводниках	ББК В379.231.4
	оптоэлектронные полупроводниковые детекторы	PACS: 85.60.–q

[4] Если даны альтернативные индексы одной классификации, то выбран один наиболее адекватный, либо составлен комбинированный индекс из обоих вариантов.

Продолжение таблицы

NM-11 S.A.Dvoretsky Control and growth of HgTe quantum wells	квантовые ямы в полупроводниках	ВИНИТИ 291.19.31.46.21
	теллур и его соединения – полупроводниковые свойства	ББК 379.2
		УДК 538.9:546.24:621.315.592
NM-14 A.S.Moskvin	электронная структура полупроводников	ББК В379.1
Electron structure of hole cen-ters in CuO2 planes of cuprates		
	дырки в полупроводниках	ББК В379.13
		УДК 538.913:544.022.373-027.21:612.315.592
NM-20 L.K.Orlov	полупроводниковые гетероструктуры	ББК 379.2
3C-SiC/SiGeC/Si heterocom-positions: physical properties, application prospects		УДК 538.975.5:621.315.592 025.25
		ГРНТИ 29.19.31
		ВИНИТИ 291.19.31