УДК 025.43:002.6

Тезаурусы по физике и электронике для навигации по цифровому пространству знаний

Шабурова Н.Н.,
кандидат педагогических наук, заведующая научной библиотекой, Институт физики полупроводников им. А. В. Ржанова Сибирского отделения Российской академии наук, Новосибирск, Россия, shaburova@isp.nsc.ru
Белоозеров В.Н.,
кандидат филологических наук, ведущий научный сотрудник, Всероссийский институт научной и технической информации Российской академии наук, Москва, Россия, systemling@narod.ru

Аннотация. Описаны принципы и опыт построения отраслевых тезаурусов, предназначенных для представления предметной области в рамках формальной онтологии единого пространства знаний. Лексика тезаурусов содержит термины, реально используемые для тематического индексирования документов — ключевые слова и классификационные рубрики. Система тезаурусных отношений установлена полуавтоматической процедурой и включает: эквивалентность (синонимию), родовидовую иерархию, ассоциацию и дефинитивную связь.
Ключевые слова: онтология, тезаурус, ключевые слова, классификационные системы, единое пространство знаний, родовидовые связи, дефинитивная связь.

Thesauri in Physics and Electronics for Navigating the Digital Knowledge Space

Shaburova N. N.,
Candidate of Pedagogical Sciences, Head of the scientific library, Rzhanov Institute of Semiconductor Physics, Siberian branch of the Russian Academy of Sciences, Novosibirsk, Russia, shaburova@isp.nsc.ru

Beloozerov V. N.,
Candidate of Philological Sciences, Leading Researcher, All-Russia Institute for Scientific and Technological Information of the Russian Academy of Sciences, Moscow, Russia, systemling@narod.ru,

Abstract. The article describes the principles and experience of building brach thesauri designed to represent subject areas within the formal ontology of the unified knowledge space. Ttrms jf the thesauri include the words that are actually used for thematic indexing of documents — keywords and classification categories. The system of the thesaurus relations is established by a semi-automatic procedure and includes: equivalence (synonymy), generic hierarchy, association, and definitive relationship.

Keywords: ontology, thesaurus, keywords, classification systems, unified knowledge space, generic relationships, definitive relationships.

DOI: 10.31432/1994-2443-2020-15-3-81-85

Цитирование публикации: Шабурова Н. Н., Белоозеров В. Н. Тезаурусы по физике и электронике для навигации по цифровому пространству знаний // Информация и инновации. 2020, Т. 15, № 3. с. 81–85. DOI: 10.31432/1994-2443-2020-15-3-81-85
Citation: Shaburova N. N., Beloozerov V. N. Thesauri in physics and electronics for navigating the digital knowledge space // Information and Innovations 2020, Т. 15, № 3. p. 81–85. DOI: 10.31432/1994-2443-2020-15-3-81-85

Создание единого пространства знаний состоит в объединении источников знания общими средствами поиска и извлечения знаний из источников. Универсальным средством для этого является естественный язык, который и служит людям в этом качестве в условиях,

когда источники знания представлены в воспринимаемой человеком форме и в обозримом количестве. Развитие компьютерных средств хранения знаний требует использования компьютерных языков поиска и интерпретации компьютерных данных. Перспективы компьютерного анализа смысла полного текста остаются проблематичными. Поэтому целью нашей работы стало исследование возможности и путей использования метаданных публикаций, которые сами уже являются продуктом интеллектуальной интерпретации документа с целью облегчить поиск содержащихся в публикации знаний. Такими метаданными являются ключевые слова и классификационные индексы, которыми снабжаются практически все документы в информационном пространстве науки и техники. Этим наш подход отличается от направления развития современных поисковых машин, которые ориентированы на «свободную лексику» и игнорируют труд индексаторов научных публикаций, результаты которого уже вложены в имеющиеся информационные ресурсы.

Необходимость искать информацию среди разобщённых источников ставит задачу согласования средств, с помощью которых систематизированы сведения в разнородных ресурсах. Однако сейчас отсутствует единство и универсальность этих средств описания тематики. Ключевые слова отнюдь не всегда позволяют определить тематику работы. Библиотеки пользуются для этого классификационными системами областей знания, но 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 смысловых связей рубрик используемых классификационных систем при привязке к ним ключевых слов, индексирующих документы данной рубрики и также связанных сетью смысловых отношений.

Иными словами, для реализации в среде связанных компьютерных ресурсов единого пространства знаний по определённой научной области требуется создать классификационно-тезаурусную сеть смысловых связей терминов и тематических рубрик, которые были использованы при создании информационных ресурсов данной области.

Для физики и электроники мы начали осуществлять эту идею с разработки в 2009 г. первой версии “Тезауруса тематических рубрик по физике полупроводников” (ТТРФПП) [1], который включал рубрики ряда классификаций, в том числе УДК, ББК и ГРНТИ. Тезаурус, содержащий около 1600 дескрипторов, был выложен на сайте научной библиотеки ИФП СО РАН и депонирован в ВИНИТИ [2]. В тезаурусе дескрипторам приписаны коды пяти классификаций и обычные тезаурусные связи.

В дальнейшем тематика физики полупроводников была дополнена терминами смежных областей физики, электроники и нанотехнологий, а модель тезауруса была предложена для построения онтологий других предметных областей [3]. Теперь в рамках работ по проекту РФФИ № 20-07-103[1]) 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 ключевых слов ТЕРМИН, в которой реализована семантическая сеть смысловых связей терминов, описана в статьях [4, 5, 6]. В рамках этой концепции разработаны тезаурусы «Физика» и «Электроника» для областей знания, определяемых разделами ГРНТИ: *29 Физика* и *47 Электроника. Радиотехника*. В основе тезаурусов лежат ключевые слова, указанные индексаторами ВИНИТИ и БЕН РАН как наиболее важные для каждой подрубрики данных разделов ГРНТИ. Каждое ключевое слово сопровождалось определением соответствующего понятия, почерпнутым, как правило, из авторитетных словарей и руководств. Словник был дополнен терминами, выделенными из наименований рубрик, входящих в данные разделы ГРНТИ, а также из связанных по смыслу рубрик других классификаций — ББК, УДК, PACS, Международной патентной классификации (в русском переводе). Лексика других рассмотренных классификаций

[1] Грант РФФИ: 20-07-00103 «Разработка методологии навигации и поиска знаний в гетерогенной сетевой среде на основе универсального интеллектуального конвертера метаданных»

(Scopus, Web of Science, OECD Fields of Science, Номенклатура ВАК) вполне укладывается в пределы собранного массива терминов. Термины из классификаций также были снабжены определениями.

Объём лексики обоих тезаурусов составляет около 1200 словарных единиц (примерно 700 по физике и 500 по электронике).

Выбор терминов из используемых классификаций и назначенных документам ключевых слов обеспечивает привязку нашей системы к реальным признакам описания тематики документов, имеющихся в пространстве поиска.

При установлении смысловых связей терминов мы исходили из того, что тезаурусы будут составлять компонент формальной онтологии, в которой они должны отражать родовидовые связи объектов, прагматические отношения которых должна отражать система ассоциативных связей разного рода, которые в свою очередь могут составлять отдельный тезаурус связей с открытым составом членов. В силу этого мы устанавливали иерархическую связь «выше — ниже» только между понятиями одной категории, которые соотносятся как род и вид, т. е. множество денотатов одного термина является подмножеством другого. Остальные случаи смысловой связи обозначались ассоциативным отношением терминов. Этим тезаурусы «Физика» и «Электроника» базы данных ТЕРМИН отличаются от обычных информационно-поисковых тезаурусов (включая наш ТТРФПП), в которых основанием для иерархической связи является соотношение не денотатов понятий, а стоящих за терминами множеств документов. То же самое справедливо и для отношения эквивалентности терминов. Термины эквивалентны тогда и только тогда, когда они обозначают тождественные множества объектов. Отношение тождества терминов означает именно тождество денотатов, а не тождество массивов документов про них. 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 на тему «полнота системы аксиом» релевантны запросу «неполнота системы аксиом», поскольку критерии этих явлений совпадают, и указание этого критерия равно определяет и то, и другое понятие. Но в онтологическом смысле эти понятия исключают друг друга и не могут рассматриваться как эквивалентные или пересекающиеся.

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

Базовый массив связей терминов в системе ТЕРМИН 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» осуществлялась интеллектуальным рассмотрением вручную. При этом существенную долю дефинитивных связей (примерно четверть) пришлось исключить как установленные из-за формального совпадения слов с семантически не связанными смыслом (например, когда термины были приписаны к понятию «время» из-за того, что в их определениях встретилось выражение *«в настоящее время»*).

Но и среди оставленных действительных связей не все целесообразно использовать при обычном документном поиске. Так, например, понятия «электричество» и «аккумулятор» явно связаны, и это может быть отражено в их определениях. Но использовать документы об аккумуляторах как релевантные для поиска документов об электричестве (и наоборот) вряд ли целесообразно в общем случае. Однако если в поисковой системе будут реализованы специфические модальности поиска *«средства накопления»* или *«устройства, основанные на…»*, то связь этих 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 объектов онтологической реальности.

Таким образом, в словарях «Физика» и «Электроника» базы данных ТЕРМИН устанавливаются следующие виды связей терминов:

- 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.

В настоящее время тезаурус «Физика» содержит 714 терминов, тезаурус «Электроника» — 488 терминов, а в целом база данных ТЕРМИН состоит из 63 тематических словарей, в которые введено 12090 терминов с определениями (12880). Термины связаны сетью 298381 отношений. База данных постепенно пополняется новыми понятиями и связями по мере выявления перспективных объектов исследования в данной области знания. Опыт практической работы в базе данных с тезаурусом «Электроника» обобщён в [7]. В настоящее время проводятся работы по объединению базы данных ТЕРМИН с Системой классификационных схем ВИНИТИ (См. доклад А. В. Шапкина и др. в наст. издании).

Доклад подготовлен в рамках работ по проекту РФФИ 20-07-00103 «Разработка методологии навигации и поиска знаний в гетерогенной сетевой среде на основе универсального интеллектуального конвертера метаданных». Авторы выражают благодарность коллегам по проекту за сотрудничество и предоставленные материалы.

ЛИТЕРАТУРА

1. Белоозеров В. Н., Шабурова Н. Н. Сопоставительный тезаурус классификационных систем по физике полупроводников // *Информационное обеспечение науки: новые технологии: Сборник научных трудов* / Н. Е. Калёнов (ред.). С. 311–322. — М.: Научный Мир, 2009.

2. Белоозеров В. Н., Шабурова Н. Н. Тезаурус тематических рубрик по физике полупроводников // *Депонировано в ВИНИТИ* 2013-12-24, № 379-В2013.

3. Белоозеров В. Н., Шабурова Н. Н. Тезаурус библиографических классификаций как модель интеграции информационных ресурсов // *Международная конф. 27-28 окт. 2011, [Москва] «Информационное общество: Состояние и тенденции межгосударственного обмена научно-технической информацией в СНГ».* — С. 8–9. — М.: ВИНИТИ, 2011.

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

5. Antopol'skii A. B. [et al.]. The development of a semantic network of keywords based on definitive relationships // *Scientific and Technical Information Processing*. 44(4). — P. 261–265. — 2017.

6. Якшин М. М., Калёнов Н. Е. Классификаторы: создание базы данных терминологических словарей // *Информационное обеспечение науки: новые технологии: Сб. научных трудов*. — Москва: БЕН РАН, 2915. — С. 137-146.

7. Shaburova N. N. Operational experience in DB "TERMIN" // *Journal of Information Science Theory and Practice.* — 2019. — V. 7 (3). — P. 21-30.

REFERENCES

1. Beloozerov V. N., SHaburova N. N. Sopostavitel'nyj tezaurus klassifikacionnyh sistem po fizike poluprovodnikov // Informacionnoe obespechenie nauki: novye tekhnologii: Sbornik nauchnyh trudov / N. E. Kalyonov (red.). S. 311–322. — M.: Nauchnyj Mir, 2009.

2. Beloozerov V. N., SHaburova N. N. Tezaurus tematicheskih rubrik po fizike poluprovodnikov // Deponirovano v VINITI 2013-12-24, № 379-V2013.

3. Beloozerov V.N., SHaburova N.N. Tezaurus bibliograficheskih klassifikacij kak model' integracii informacionnyh resursov // Mezhdunarodnaya konf. 27-28 okt. 2011, [Moskva] «Informacionnoe obshchestvo: Sostoyanie i tendencii mezhgosudarstvennogo obmena nauchno-tekhnicheskoj informaciej v SNG». — C. 8–9. — M.: VINITI, 2011.

4. Antopol'skii A. B. [i dr.]. Razrabotka ontologii informacionnogo prostranstva znanij na osnove definitivnyh svyazej // Nauchno-tekhnicheskaya informaciya. Seriya 1. Organizaciya i metodika informacionnoj raboty, № 11. — C. 19–24. — 2017.

5. Antopol'skii A. B. [et al.]. The development of a semantic network of keywords based on definitive relationships // Scientific and Technical Information Processing. 44(4). — P. 261–265. — 2017.

6. YAkshin M. M., Kalyonov N. E. Klassifikatory: sozdanie bazy dannyh terminologicheskih slovarej // Informacionnoe obespechenie nauki: novye tekhnologii: Sb. nauchnyh trudov. — Moskva: BEN RAN, 2915. — S. 137-146.

7. Shaburova N.N. Operational experience in DB "TERMIN" // Journal of Information Science Theory and Practice. — 2019. — V. 7 (3). — P. 21-30.