УДК 025.4
DOI: 10.15827/2311-6749.18.4.3

СЕТЬ КЛАССИФИКАЦИОННЫХ СИСТЕМ ВИНИТИ РАН

В.Н. Белоозеров, к.филол.н., ведущий научный сотрудник, systemling@narod.ru;
А.В. Шапкин, к.т.н., начальник управления; Ю.Н. Щуко, к.геогр.н., врио директора
(ВИНИТИ РАН, г. Москва, 125190, Россия)

Описываются история, методика создания и текущее состояние базы данных ВИНИТИ РАН, содержащей основные используемые в информационной практике классификационные системы, между рубриками которых установлены смысловые связи. Эта сеть связей предлагается для использования при навигации и поиске тематической информации в пространстве разнородных информационных ресурсов.

Ключевые слова: *классификационные системы, смысловые связи, поиск информации, сопоставление классификаций, разнородные информационные ресурсы, навигация по информационным сетям.*

Система информационно-поисковых языков, установившаяся в библиотечно-информационном пространстве нашей страны к 80-м годам прошлого века, основывалась на трех классификационных слонах: национальная *Библиотечно-библиографическая классификация* (ББК) применялась в универсальных и массовых библиотеках, международная *Универсальная десятичная классификация* (УДК) – в научно-технических библиотеках и информационных центрах, *Государственный рубрикатор научно-технической информации* (ГРНТИ) – в автоматизированных процессах и электронных БД. Классификационные средства поиска дополнялись отраслевыми и специализированными информационно-поисковыми тезаурусами для поиска по терминам, выражающим потребность клиента в информационном обслуживании. В те годы была надежда, что разработка достаточно мощных единых средств поиска решит все задачи информационного обеспечения науки. Однако в 90-х годах с отказом от методов централизованного управления экономикой стало очевидно, что реальностью является неуправляемый рост разнообразных источников информации и средств доступа к ним. Положение усугубилось тем, что развитие информационных технологий, появление Интернета показали, что эффективность поиска информации дешевле и быстрее достигается простым увеличением объемов памяти и быстродействия компьютеров, чем развитием интеллекта поисковых машин. Поэтому прекратились разработки единых классификаторов и тезаурусов, которые должны были бы реализовывать интеллект информационного поиска. Однако потребность работать с разнообразными языковыми средствами привела к созданию ряда репозиториев тезаурусов, классификаций и других систем организации знаний. Коллекция классификационных систем ВИНИТИ РАН начала создаваться в начале 2000-х годов на основе УДК и ГРНТИ, ответственность за поддержание которых легла на институт. Эти работы были описаны в ряде статей [1, 2] и в монографии [3].

В связи с дальнейшим расширением Интернета и развитием информационных технологий была поставлена задача создания семантического веба, то есть средств осмысленной навигации по сетевым ресурсам в соответствии с их фактическим содержанием и поиска в них необходимого знания, отвечающего информационной потребности. Для решения этой задачи требуется научить системы поиска информации пониманию текстов и других данных, предоставляемых информационными ресурсами. Основным методом информационного поиска в последние годы стал поиск по свободной лексике (лексический поиск), на котором основаны распространенные поисковые машины (Яндекс, Google). Однако такой поиск дает низкие характеристики полноты и точности, в частности, потому, что он не учитывает семантические связи понятий. При этом игнорируется интеллектуальный вклад разработчиков информационно-поисковых языков и индексаторов научной информации, заложенный в большинстве ресурсов научно-технических знаний. Этот вклад состоит в индексах классификационных систем и ключевых словах, которые присвоены практически каждому документу в сфере научно-технической информации. Эти связи, заложенные в классификациях и тезаурусах, при поиске могли бы помочь точно отобрать документы, содержащие требуемую информацию, и найти те документы, в которых тема выражена другими языковыми средствами.

Трудность, однако, состоит в том, что в разных источниках тематика выражена различающимися языковыми средствами. Например, в сети библиотек РАН по естественным наукам часть библиотек индексируют документы кодами УДК, часть – кодами ББК, а часть имеет фонды, индексированные выборочно либо тем, либо другим способом. Единый для всего электронного пространства научных знаний язык ГРНТИ недостаточно точен для поиска детальных сведений, а большинство пользователей предпочитают формулировать свою потребность терминами научных публикаций, а не классификационными индексами, о существовании которых они могут вообще не подозревать. С другой стороны, поиск информации в документах определенного вида даст наилучший результат при использовании специализированной классификации, рассчитанной на данный вид документов. Так, поиск патентных данных сле-

дует вести по *Международной патентной классификации* (МПК), поиск стандартов – по *Международной классификации стандартов* (МКС), поиск решений математических проблем – по господствующей в этой области знания *Mathematical Subject Classification* (MSC). Поиск данных о научном использовании публикаций следует вести по классификаторам мировых библиографических систем Scopus и Web of Science.

Таким образом, очевидно, что пространство научно-технической информации в настоящее время предстает как конгломерат в определенной степени изолированных частей, в соответствии с теми средствами, которые позволяют осуществлять навигацию (поиск) по тематике источников знания. Интеграция пространства может быть осуществлена интеграцией (объединением) средств тематической навигации – используемых классификаторов тематического описания документов. Объединение классификаторов возможно путем установления смысловых соответствий классификационных рубрик. За рубежом такая работа проводится довольно активно. В Базельском регистре тезаурусов, онтологий и классификаций [4] насчитывается 15 конкордансов классификационных систем, главным образом – сопоставление проблемных классификаторов с *Десятичной классификацией Дьюи* (ДКД). Пример интеграции в одном инструменте различных средств доступа к содержанию ресурсов можно видеть в успешно действующей в Национальной библиотеке по медицине США Единой медицинской лингвистической системе, включающей Метатезаурус [5], объединивший лексику около двухсот классификаторов, справочников, словарей и номенклатур по частным медицинским проблемам. В Германии запущен проект COLI-CONC [6] создания сетевых программных модулей, объединяющих единым доступом различные системы организации знаний (онтологии, тезаурусы, классификаторы) для эффективного управления и их практического использования. Это говорит об актуальности и своевременности предпринятой в ВИНИТИ РАН работе по созданию сети сопоставления классификационных систем, используемых в российской информационной практике [7].

На первом этапе работ методом интеллектуального анализа были получены сопоставления рубрик ГРНТИ (около 8 тыс. позиций) следующим одиннадцати классификаторам: УДК (около 130 тыс. позиций), ББК (на верхнем уровне), МПК (фрагментарно), Номенклатура специальностей научных работников ВАК (501 специальность), классификаторы РИНЦ, Scopus, Web of Science, РФФИ, РНФ, РГНФ, ФАНО [8].

Сопоставляемые классификаторные рубрики отмечались тремя показателями соответствия – эквивалентность тематики, включение тематики одной рубрики в другую и наличие значимого пересечения тематик. Эти отношения позволяют контролировать при навигации по рубрикам характер возможного отклонения тематики от исходного поискового запроса. В таблице 1 показаны обозначения этих отношений.

Таблица 1

Обозначение и значение отношений рубрик

Обозначение	Чтение	Значение
А = Б	А эквивалентна Б	Тематика рубрик А и Б совпадает
А > Б	А включает Б	Рубрика А включает тематику Б
А < Б	А входит в Б	Тематика А входит в рубрику Б
А >< Б	А пересекается с Б	Тематика рубрик А и Б имеет значимые общие части

На втором этапе работ, который продолжается в настоящее время, поставлена задача установления прямых соответствий всех классификаций со всеми. Преодолеть неисчислимость количества возникающих при этом комбинаций помогла программа транзитивного установления соответствий [9]. Ее смысл заключается в следующем: если рубрика А из одного классификатора и рубрика Б из другого связаны с одной и той же рубрикой Г из ГРНТИ, то в большинстве случаев можно констатировать определенную непосредственную связь рубрик А и Б (см. табл. 2).

Таблица 2

Установление транзитивных связей

Исходные связи	Г = Б	Г > Б	Г < Б	Г >< Б
А = Г	**А = Б**	**А > Б**	**А < Б**	**А >< Б**
А > Г	**А > Б**	**А > Б**	**А >< Б**	**А >< Б**
А < Г	**А < Б**	**?**	**А < Б**	**?**
А >< Г	**А >< Б**	**?**	**А >< Б**	**?**

В результате получен основной массив связей (20 4649 соответствий) без затраты ручного интеллектуального труда. Однако результаты автоматических операций всегда требуют интеллектуальной верификации. Сплошной просмотр выявил чрезвычайно мало ложных связей, буквально единицы процента. Гораздо больше случаев, когда специалист может уточнить вид связи. Такие уточнения пока делались только эпизодически, по отдельным тематическим направлениям, но и без этого имеющиеся автоматические связи позволяют переходить от одной классификации к другой по смыслу, но с более размытой оценкой соответствия тематик. Еще больше случаев, когда транзитивный поиск соответствий не срабатывает: между рубриками А и Б разных классификаций объективно есть четкое смысловое соответствие (по мнению специалиста), но в ГРНТИ не найдена такая рубрика Г, которая была бы связана одновременно и с А, и с Б. По оценке авторов, таких случаев, по крайней мере, столько же, сколько установлено автоматически. Поиском этих дополнительных связей в настоящее время занимаются наши специалисты. Но уже имеющиеся связи соединяют рассматриваемые классификации довольно густой сетью, позволяющей переходить по связям от одной классификации к любой другой, сохраняя тематическую преемственность.

На предстоящем третьем этапе работы предполагается увеличить семантическую силу созданной сети классификаций путем включения в нее полного Рубрикатора ВИНИТИ, дающего возможность углубляться в детали естественных и технических наук до десятого уровня дробности, что позволяет обозначать классификационным индексом отдельные проблемы, обсуждаемые исследователями. А система связей классификаций позволит отыскивать им аналоги в посторонних информационных ресурсах, систематизированных другими классификациями знания, выявляющими иные аспекты знаний. Круг классификаций будет также расширен за счет включения некоторых актуальных отраслевых и специализированных классификаторов, таких как международная математическая классификация, классификации патентов и стандартов. При этом основное внимание будет уделено тематике естественных и технических наук в соответствие со специализацией ВИНИТИ РАН.

Необходимым дополнением сети классификационных систем должны служить термины областей знания, на языке которых формулируют свои потребности рядовые пользователи информации. В определенной степени эти термины представлены ключевыми словами, назначенными каждому документу либо авторами, либо индексаторами. БД ВИНИТИ позволяет отобрать наиболее представительные ключевые слова для каждой рубрики классификаторов, с тем чтобы дать возможность перехода от запроса пользователя к сети классификационных рубрик, которые обеспечат полноту поиска по тематике точно указанной наименованием рубрики. Задача сопряжения классификаций с ключевыми словами будет решаться при дальнейшем развитии системы.

В настоящее время сеть классификаций характеризуется следующими объемными показателями. Всего в БД записано более 20 классификаторов разных объема и тематики, которые содержат 241 936 рубрик, описывающих некоторую научную область исследований. На 12 классификаций из этого набора наложена сеть более чем 200 000 смысловых связей, число которых растет по мере интеллектуальной обработки массива.

Сеть классификационных систем предполагается использовать при диспетчеризации запросов по различным БД, а также в составе поисковых машин Интернета для реализации идеи семантического веба. Надежда на развитие средств извлечения смысла из полных текстов далека от осуществления, поэтому осмысленный поиск сведений в сети должен использовать тот интеллектуальный вклад, который сделали разработчики классификаций и индексаторы информационных систем для визуализации этого смысла с помощью классификационных и лексических метаданных информационных ресурсов.

Литература

1. Шапкин А.В. Практические вопросы построения системы классификационных схем // НТИ: Инф. процессы и системы. 2006. № 6. С. 1–14.

2. Шапкин П.А. Применение технологий ASP, NET и Semantic Web для обеспечения доступа к системе классификационных схем // НТИ: Инф. процессы и системы. 2006. № 5. С. 20–25.

3. Гиляревский Р.С., Шапкин А.В., Белоозеров В.Н. Рубрикатор как инструмент информационной навигации. СПб: Профессия, 2008. 352 с.

4. BARTOC.org: Basel Register of Thesauri, Ontologies and Classifications. Basel University Library, Switzerland. URL: https://bartoc.org/en/content/about (дата обращения: 21.09.2018).

5. Metathesaurus. U.S. National Library of Medicine. URL: https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/index.html (дата обращения: 21.09.2018).

6. COLI-CONC: Development of an infrastructure to facilitate management and exchange of concordances between library knowledge organization systems. Verbundzentrale des GBV (VZG). URL: https://coli-conc.gbv.de/ (дата обращения: 21.09.2018).

7. Арский Ю.М. [и др.] Сопоставление ГРНТИ с другими классификационными системами с целью совершенствования системы тематической кодификации НИР, НИОКР гражданского назначения. Формирование системы соответствий между различными классификаторами в сфере научно-технической информации. М.: Изд-во ВИНИТИ РАН, 2014.

8. Антошкова О.А., Белоозеров В.Н., Дмитриева Е.Ю. Разработка базовых соответствий между ГРНТИ и другими классификационными системами // Информационное обеспечение науки: новые технологии: сб. тр.; [под ред. Н.Е. Калёнова, В.А. Цветковой]. М.: Изд-во БЕН РАН, 2015. С.137–146.

9. Белоозеров В.Н., Шабурова Н.Н. Метод сопоставления библиографических классификаций на основе соответствий с ГРНТИ (на примере УДК и ББК) // НТИ: Инф. процессы и системы. 2016. № 10. С. 13–24.