16. Гиляревский Р. С., Родионов И. И. и др. Информатика как наука об информации: Информационный, документальный, технологический, экономический, социальный и организационный аспекты / Гиляревский Р. С., Родионов И. И., Залаев Г. З., Цветкова В. А., Барышева О. В., Калин А. А.; под ред. Гиляревского Р. С ; авт.-сост. Цветкова В. А.- М.: ФАИР-ПРЕСС, 2006.— 592 с.

17. Гиляревский Р. С. Статус и перспективы технической документации в России // Международный форум по информации.— 2005. Т. 30.— № 2.— С. 4.

18. Арский Ю. М., Черный А. И. Информационные ресурсы для устойчивого развития общества // Международный форум по информации.— 2003. Т. 28.— № 4. С. 3-9.

19. Антопольский А. Б., Майстрович Т. В. Электронные библиотеки: основные принципы создания. Научно-методическое пособие.- М.: Изд-во "ЛИБЕРЕЯ", 2006.

20. Межгосударственный стандарт. Система стандартов по информации, библиотечному и издательскому делу. Содержание записи. Издание официальное Бз2-2001/12. Межгосударственный совет по стандартизации, метрологии и сертификации. — Минск, 2001.

21. Кедровский О. В. Информационные ресурсы и информационная политика // НТИ. Сер. 1. 1988. № 7.-- С. 2 4.

22. Маркусова В. А., Вилсон К. Базы данных как источник информации о международном сотрудничестве: модели российско-скандинавского сотрудничества // НТИ. Сер. 1. 2002. № 12. С. 27 34.

23. Маркусова В. А., Кречмер Х., Янец М. Модели сотрудничества российских исследователей с учеными Германии и Индии // Международный форум по информации.— 2002.— Т. 27.— № 2.— С. 25 28.

24. Персел Г., Херсон П. Национальная служба технической информации и распространения информационной литературы / Перевод Хахновича А. Н. М.: ВЦП, 1990.— С. 30.

25. Родионов И. И., Цветкова В. А. Конвергенция услуг связи и продуктов // НТИ. Сер. 1.-- 2004.- N. 8. С. 10 12.

26. Савин А Н. Мотивация управления в информационной сфере как решающий фактор повышения эффективности работы информационных служб и организаций // Всероссийский научно-практический семинар руководителей и специалистов Объединения "Росинформресурс" / Тез. докл. - Адлер, - 3-6 октября 2000. - С. 3.

27. Савин А. Н. О некоторых проблемах стимулирования труда работников организаций информационной сферы в условиях оптимизации хозяйственного механизма / Тез. докл. Всероссийской научно-практической конференции "Проблемы организации использования результатов научно-технической деятельности в интересах экономического и социального развития регионов Российской Федерации" 11 13 апреля 2000. Ярославль. С. 79.

28. www.http://ecatalog.csti.yar.ru

29. Левинский Л. С., Савин А. Н. Объединение "Росинформресурс" единый информационно-технологический комплекс // Информационные ресурсы России. 2006. N. 2 С. 9 11.

Материал поступил в редакцию 08.10.07

УДК 002.1:004.91

В. С. Егоров, Т. Н. Чернобровская

Автоматизированное определение тематики публикаций при библиографической обработке первоисточников

Рассматриваются различные способы автоматической разметки, применяемые в технологическом цикле обработки входного потока научно-технической литературы для подготовки информационных продуктов ВИНИТИ РАН.

Определение тематики публикаций одна из ключевых задач в процессе обработки научно-технической литературы (НТЛ). От точности ее решения зависит качество подготовки информационных продуктов ВИНИТИ РАН и обслуживания потребителей. При глубоком индексировании первоисточников конечный пользователь (потребитель информации) не будет испытывать затруднений в поиске необходимых ему сведений, несмотря на постоянно увеличивающиеся объемы НТЛ. При хорошем знании Рубрикатора ВИНИТИ РАН легко можно собрать информацию по малознакомой тематике.

Выявление тематики первоисточника многоэтапный процесс, который начинается с грубой оценки тематической направленности публикации (тематическая разметка на входе) и заканчивается точным индексированием по Рубрикатору (специалистами в отделах научной информации проставляется соответствующая рубрика). На разных этапах обработки НТЛ для определения тематики доступны различные сведения о публикации, и это является предметом сложных процессов аналитико-синтетической переработки в любом информационном центре. Специалист, выполняющий тематическую разметку, должен обладать энциклопедическими знаниями, достаточно глубоко ориентироваться не только в одной научной отрасли, но и во

всех, соприкасающихся с ней. На первичном этапе разметки необходимо обладать знанием всех отраслей науки и техники. К сожалению, такого рода специалистов не готовит ни одно учебное заведение и процесс вхождения в специальность происходит в ходе практической работы.

ВИНИТИ РАН, имея значительный потенциал в создании автоматизированных технологий обработки НТЛ, уже давно ведет исследования в области автоматизации процессов разметки. Особенно эта проблематика была популярна в 70-х, начале 80-х годов XX в., т. е. в начале массового внедрения компьютерной техники. Работы носили теоретический характер с небольшими модельными экспериментами. Полномасштабного внедрения автоматической разметки первоисточников в технологический процесс не было по крайней мере по двум причинам: отсутствие на входе Института потока НТЛ в электронной форме и слабость компьютерной техники. Сейчас в условиях появления значительного потока литературы в электронной форме, создания эффективной системы автоматизации процессов обработки входного потока ВИНИТИ РАН (система АС "ВХОД"), резкого сокращения штатов внимание к этой проблеме приобретает практический интерес. Институту необходимы решения, дающие реальную экономию трудовых ресурсов и сокращение сроков обработки НТЛ Такие работы были начаты в 2000 г и осуществляются в двух направлениях·

автоматическая классификация, основанная на элементах искусственного интеллекта;

автоматическая классификация, основанная на использовании классификационных меток, проставленных в экземплярах первоисточников при их издании сторонними организациями, реализующая сложившиеся в институте схемы обработки НТЛ и использующая элементы библиографического описания литературы (БО).

Безусловно, первое направление является более трудным и перспективным, однако ожидать реального практического внедрения достаточно сложно, хотя одна из подобных систем уже находится в опытно-промышленной эксплуатации. Настоящая статья посвящена описанию результатов, внедренных в промышленную эксплуатацию по второму направлению. Эта технология является частью Автоматизированной системы комплектования и регистрации входного потока (АСКР) [1, 2], которая эксплуатируется в ВИНИТИ РАН с 2000 г. и позволяет создавать механизмы автоматической тематической разметки на основе библиографических сведений в обрабатываемых изданиях.

Из всех видов НТЛ были выделены те, которые имеют в исходном документе классификационные рубрики или специальную терминологию или объединены общностью замысла, тематики, целевым или читательским назначением, выходящие под общим названием и в однотипном типографском оформлении.

Результатом визуальной оценки первоисточников стало выявление в БО тех элементов данных (ЭД), которые можно использовать для определения тематики, т. е. выявление тематически нагруженных ЭД:

Код вида издания [3]
Классификационные индексы
Организации, участвующие в подготовке издания:
Коллективный автор публикации
Место выполнения опубликованной работы
Место защиты диссертаций
Издательство

При наличии признака тематической разметки одновременно в нескольких ЭД, определение тематики осуществляется строго в указанной выше последовательности.

ЭД, участвующие в процедуре автоматической разметки (профилирования), снабжены специальным признаком, значение которого соответствует штампу Отдела научной информации (ОНИ) по соответствующей тематике. Согласно принятой в ВИНИТИ РАН технологии одноразового реферирования существует понятие 1-го ОНИ, в реферативных изданиях которого отражается реферат на первоисточник, и понятие 2-го и последующих ОНИ, которые заимствуют данный реферат.

Рассмотрим технологию автоматической разметки с использованием разных ЭД.

1. Профилирование по коду вида издания

Код вида издания (поле 626) – обязательный элемент БО. В автоматическую разметку включены следующие виды изданий.

Картографические виды изданий (626=*КГИ*). Автоматическая разметка **всех** картографических изданий осуществляется с помощью специального маршрута, который автоматически проставляется им при регистрации. В качестве 1-го ОНИ этому виду изданий проставляется штамп ОНИ по географии, а следующими определены ОНИ по геологии, геофизике (астрономии) и охране окружающий среды [4].

Издания книжного типа, а именно книжные серии (626=*ИКТ.СЕР*). Выпуски книжных серий составляют заметную часть входного потока изданий книжного типа. Традиционная технология библиографической обработки таких изданий предусматривает заполнение элемента данных "сведения о серии" (название серии, коллективный автор и издательство). При этом каждый раз значение этих полей вводилось заново, независимо от ранее поступивших выпусков той же серии. Такая технология приводила к разночтению описания книжной серии и, как следствие, не позволяла группировать выпуски одной серии, а для номерных серий не было возможности следить за их "лакунами".

Между тем описание серии является довольно стабильным. При наличии соответствующего аппарата накопления и хранения сведений о книжных сериях можно, во-первых, существенно сократить трудозатраты на ввод библиографического описания выпуска и, во-вторых, — использовать общие данные для определения тематической направленности выпусков одной серии. С этой целью в состав описания книжной серии был включен специальный элемент "Признак автоматической разметки" поле kind_razm и поле "Код профильности" (штамп ОНИ). Значение поля kind_razm=1 означает, что данное издание должно быть направлено на реферирование в указанный ОНИ. Около 80% всех книжных серий размечаются автоматически.

Периодические издания (626=*СИ*). На основе статистической обработки периодических изданий были выявлены так называемые "монаразметочные издания" такие, которые отражаются

в выпусках Реферативного журнала только одного ОНИ на протяжении 3-х лет. Аппарат автоматической разметки данного вида издания аналогичен аппарату автоматической разметки книжных серий.

2. Профилирование по классификационным индексам

Автоматическая разметка по классификационным индексам осуществляется для авторефератов диссертаций и государственных и отраслевых стандартов. Авторефераты отечественных диссертаций в качестве обязательного элемента БО содержат специфический элемент "Номер специальности" по классификации ВАК, а стандарты имеют индекс "Общероссийского классификатора стандартов". Для автоматической разметки этих видов изданий были созданы специальные словари "Классификатор ВАК" и "Классификатор НТД" [5]. Большая часть этих словарей имеет прямо выраженную тематику, которая проставлена в специальные поля "Штамп ОНИ1" (однозначные) и "Штамп ОНИ2" (многозначные). Наполнение указанных полей осуществляется строго по словарю SUBJECT (Тематика (коды, штампы)). Коды конкретных ОНИ1 и ОНИ2 проставляются только в том случае, если практически все выпуски издания по этой рубрике отражаются в изданиях данных ОНИ. Например, для специальностей ВАК 25.00.21 "Теоретические основы проектирования горно-технических систем" и для кода НТД 75.020 "Добыча и переработка нефти и природного газа" автоматически все авторефераты и нормативные документы по этим рубрикам будут направлены в ОНИ по геологии.

Для изданий, у которых рубрика однозначно не определяет ОНИ, в код ОНИ1 проставляется признак "ручной разметки" ТО! Например, для рубрики VAK 05.02.01 "Материаловедение (по отраслям)" осуществляется традиционная ручная разметка.

Авторефераты диссертаций, тематика которых не соответствует Рубрикатору ВИНИТИ РАН (например, "Литературоведение"), автоматически исключаются из дальнейшей обработки и им автоматически присваивается маршрут "Научные фонды" (НФ).

Если в БО документа была допущена полиграфическая ошибка в индексе классификации, то система автоматической разметки определяет его как несуществующий. Для таких случаев в словари был введен "псевдокод", например 00.00.00 в словаре "Классификатор ВАК" и 00 в словаре "Классификатор стандартов", на их основании такой документ передается на участок традиционной ручной разметки.

Сегодня автоматической разметке подвергаются практически 95% всего потока авторефератов и 75% нормативных документов.

3. Профилирование организаций

Профилирование организаций распространяется на издательства, коллективных авторов, центры-депозитарии, места защиты и выполнения работы, организаторов мероприятий, так как есть основания для них предполагать специализацию по тематике.

Профилирование организаций основывается на статистической обработке входного потока документов, обрабатываемых в ВИНИТИ РАН за последние четыре года, и на принципе представительности выборки — количества наименований публикаций, которые подготовлены к изданию данной организацией.

На основе анализа статистических данных, снятых с таблицы analit.ORG_SHTAMP [2], были выявлены организации, которые готовят информационную продукцию по одному устойчивому тематическому профилю. Устойчивые тематические профили были занесены в описания объектов-организаций.

Как в случае с сериальными изданиями и книжными сериями в состав описания объектов-организаций был включен специальный элемент "Признак автоматической разметки" поле kind_razm. Значение поля kind_razm=1 означает, что для издания, подготовленного данной организацией, может быть применена процедура автоматической разметки, и для него выбирается соответствующий код профильности. Это в первую очередь место защиты иностранных диссертаций, у которых в этом поле указан факультет.

СПИСОК ЛИТЕРАТУРЫ

1. Шапкин А. В. Автоматизированная система комплектования и регистрации входного потока ВИНИТИ. Ч. 1. // НТИ. Сер. 1. 2005. № 3. С. 8 19

2. Шапкин А. В. Автоматизированная система комплектования и регистрации входного потока ВИНИТИ. Ч. 2. // НТИ. Сер. 1. 2005. № 4. С. 16 31

3. Дивильковская Т. Ю., Козачук М. В., Чернобровская Т. П. Классификация изданий книжного типа в БД ВИНИТИ // Сборник трудов Международной конференции Информационное общество. Интеллектуальная обработка информации. Информационные технологии. (ИТИ-2002), 16 18 октября 2002. ВИНИТИ. М., 2002. С. 126.

4. Чернобровская Т. П., Денисова Л. А., Пономаренко Т. П. Обработка картографических изданий в АСКР ВИНИТИ. Там же. С. 376.

5. Чернобровская Т. Н, Блат К. Д., Маркова Л. И. Словарная база АС "Вход". Там же. С. 377.

Материал поступил в редакцию 12.11.07