УДК 519.7:007.52

ИДЕНТИФИКАЦИЯ ЗНАНИЙ В ЭЛЕКТРОННЫХ БИБЛИОТЕКАХ

О.В. КАНИЩЕВА

Рассмотрен метод компараторной идентификации как один из логических методов Data Mining для решения задач обработки текстов естественного языка в автоматизированных информационных библиотечных системах.

ВСТУПЛЕНИЕ. АКТУАЛЬНОСТЬ РАБОТЫ

Разработка и исследование электронных библиотек (ЭБ) — одно из актуальных направлений развития информационных систем в последние годы, привлекающее внимание специалистов различного профиля.

Специалисты в области библиотечного дела видят в ЭБ новые возможности для совершенствования автоматизированных библиотечных систем, превращения их в публичные ЭБ нового поколения с развитыми средствами представления разнообразных цифровых информационных ресурсов и доступа к ним, создаваемые с учетом необходимости интеграции издательских и библиотечных технологий.

Специалисты в области информационных систем рассматривают ЭБ как новый класс информационных систем, базирующихся на самых передовых достижениях информационных и телекоммуникационных технологий [1]. Разработки таких систем порождают разнообразные сложные теоретические и технологические проблемы, требующие отдельного исследования.

К числу наиболее острых технологических проблем развития ЭБ можно отнести следующие:

- Развитие методов представления информационных ресурсов ЭБ.
- Определение состава метаданных, независимых от применений и специфических для различных сфер приложения, разработка средств их представления.
- Развитие новых походов к каталогизации информационных ресурсов ЭБ.
- Разработка техники индексирования информационных ресурсов различной природы (текст, аудио, видео и т.п.), методов поиска и обнаружения релевантных ресурсов, а также принципов и средств их анализа.
- Интеграция неоднородных коллекций информационных ресурсов на логическом и семантическом уровнях.
- Разработка подходов к интеграции метаданных и методов их реализации.
- Создание функционально развитых пользовательских интерфейсов (многоязыковый доступ, визуализация данных, персонализация функций, поддержка семантического уровня общения пользователей с системой).
- Эффективное использование новых Веб-технологий, основанных на стандартах платформы XML.
- Исследование архитектурных аспектов ЭБ.
- Обеспечение безопасности информационных ресурсов ЭБ.

Однако для обычного пользователя наиболее актуальной остается проблема эффективного поиска, который даст на выходе релевантную, полезную информацию.

В настоящее время для анализа больших массивов информации на естественном языке практически во всех сферах деятельности человека, где накоплены большие объемы данных, используют технологии Data Mining и Text Mining.

Text Mining содержит новые методы для выполнения семантического анализа текстов, информационного поиска и управления. Синонимом понятия Text Mining является KDT (Knowledge Discovering in Text — поиск или обнаружение знаний в тексте).

В отличие от технологии Data Mining, которая предусматривает анализ упорядоченной в некие структуры информации, технология Text Mining анализирует большие и сверхбольшие массивы неструктурированной информации.

Программы, реализующие эту задачу, должны некоторым образом оперировать текстами на естественном языке и при этом «понимать» смысл анализируемого текста. Одним из методов Text Mining является метод сравнения, или метод компараторной идентификации.

Для формализации и хранения знаний в памяти интеллектуальной системы (ИС), а эта система также имеется и в ЭБ, рассматриваются задачи представления знаний. Для этого разрабатываются специальные модели, языки для описания и выделяются различные типы знаний, изучаются источники, из которых ИС может черпать знания, создаются процедуры и приемы, с помощью которых возможно приобретение знаний для ИС. Проблема представления знаний для ИС чрезвычайно актуальна, так как ИС — это система, функционирование которой опирается на информацию о проблемной области, хранящуюся в ее памяти.

В настоящее время из существующих моделей представления знаний наиболее популярны логические, сетевые, продукционные, фреймовые и формальные модели представления знаний [2]. В рассматриваемой задаче идентификации знаний в интеллектуальных системах предметной областью обычно называется множество предметов и процессов, которые составляют основу необходимой для решения задачи обработки информации.

Общеизвестно, что языки, предназначенные для описания предметных областей, называются языками представления знаний (ЯПЗ). Считается, что универсальным ЯПЗ является естественный язык. Однако использовать его в системах машинного представления знаний сложно, так как он тяжело поддается формализации из-за нерегулярности, полисемии, омонимии и т.д., а главное — из-за отсутствия формализации семантики естественного языка, которая имела бы достаточно эффективную операционную поддержку.

ИСПОЛЬЗОВАНИЕ МЕТОДА КОМПАРАТОРНОЙ ИДЕНТИФИКАЦИИ В ЗАДАЧАХ ПРЕДСТАВЛЕНИЯ ЗНАНИЙ

Классическая задача идентификации состоит в том, чтобы по входному x и выходному y сигналам объекта определить закон $y = F(x)$ преобразования. Такую идентификацию называют прямой, поскольку она осуществляется при непосредственном доступе к выходному сигналу. Однако в ряде случаев возникает необходимость в косвенной идентификации объекта, когда у исследователя нет прямого доступа к выходному сигналу. Многие задачи этого типа можно решать методом компараторной идентификации, который

позволяет излагать основные положения теории интеллекта дедуктивным способом, исходя исключительно из физически наблюдаемых фактов. Этот метод хорошо зарекомендовал себя при обработке лингвистических объектов различных уровней языка.

Компараторная идентификация используется для формального описания низших (периферических) механизмов интеллекта (восприятие, узнавание и понимание). Эти механизмы формируют физические реакции человека на внешние воздействия.

Обрабатываемые библиотечными системами объекты являются дискретными, конечными и детерминированными, что позволяет использовать при обработке объектов АИБС (автоматизированной информационной библиотечной системы) метод компараторной идентификации.

Для реализации метода компараторной идентификации необходим единый универсальный, хорошо разработанный математический аппарат, желательно, ориентированный и на моделирование всех уровней лингвистической обработки текстов документов. Опыт исследования закономерностей передачи информации на естественном языке, а именно с такой информацией мы имеем дело в библиотечных системах, показывает, что целесообразно пользоваться одним формальным аппаратом описания закономерностей передачи и интеллектуального преобразования информации [3]. Таким наиболее универсальным математическим языком, служащим для решения задач обработки текстовой информации, является алгебра конечных предикатов [4, 5].

Используя алгебру предикатов и предикатных операций, можно создать интегрированную модель представления знаний, основанную на традиционных моделях, а также на моделях представления знаний на естественном языке. Алгебра предикатов компенсирует необходимость в других ЯПЗ. При этом объекты и отношения во всех моделях представления знаний записываются в виде уравнений алгебры предикатов. Системы предикатных уравнений могут решаться с помощью универсального решателя, который представляет собой программу, написанную на некотором алгоритмическом языке высокого уровня. Кроме того, любое уравнение алгебры предикатов может быть представлено в виде переключательной цепи, что предоставляет возможность сконструировать процессор представления знаний из комбинаций таких цепей [6].

Исчисления высказываний (ИВ) и предикатов (ИП) гарантируют непротиворечивость вывода, алгоритмической разрешимости (для ИВ) и полуразрешимости (для ИП первого порядка).

Алгебра конечных предикатов (АКП) полностью характеризуется алфавитом A, состоящим из k символов $a_1, a_2, \ldots, a_k$ и алфавитом переменных B из n символов $x_1, x_2, \ldots, x_n$. Средствами АКП может быть описан любой n-местный k-ичный предикат $f(x_1, x_2, \ldots, x_n)$, заданный алфавитом A. Формулы АКП состоят из следующих символов: $a_1, a_2, \ldots, a_k$, переменных $x_1, x_2, \ldots, x_n$, знаков дизъюнкции $\vee$, конъюнкции $\wedge$, логических констант 0 и 1, называемых соответственно ложью и истиной.

Предикатом P, заданным на U^n, называется любая функция $\varepsilon = P(x_1, x_2, \ldots, x_n)$, отображающая множество U^n в множество $\sum$, где $\sum = \{0,1\}$.

Под универсумом элементов U^n будем понимать все возможные тексты документов полнотекстовой базы данных, вторичные документы (рефе-

рат, аннотация, библиография), ключевые понятия, дескрипторы, рубрики, подрубрики и т.д. Переменные $x_1, x_2, \ldots, x_n$ называются предметными, а их значения предметами. При $n=1$ предикат P является унарным, при $n=2$ — бинарным, при $n=3$ — тернарным. Если множество U конечно, как при моделировании библиотечных процессов, то и предикат P конечный. Предикаты, обозначаемые 1 и 0, называются тождественно истинными и тождественно ложными соответственно.

Множество всех n-арных предикатов, заданных на U^n, на котором определены операции дизъюнкции, конъюнкции и отрицания, называется алгеброй n-арных предикатов на U^n. При этом операции дизъюнкции, конъюнкции и отрицания являются базисными для алгебры предикатов, которая при любом значении n является разновидностью булевой алгебры, и в ней выполняются все ее основные тождества. Базисными предикатами для алгебры предикатов будут предикаты вида

$$x_i^a = \begin{cases} 1, \text{ если } x_i = a \quad (1 \le i \le n), \\ 0, \text{ если } x_i \ne a \quad (1 \le i \le n), \end{cases} \tag{1}$$

где $i = \{1, 2, \ldots, n\}$; a — любой элемент универсума. Предикат вида (1) называется предикатом узнавания предмета a по переменной x_i. Если универсум конечен и состоит из m элементов, всего имеется $m \times n$ различных базисных элементов. Алгебра предикатов полна в том смысле, что любой ее предикат можно представить в виде суперпозиции базисных операций, примененных к базисным элементам. На языке АКП могут быть описаны любые конечные отношения, поэтому другой математический аппарат, предназначенный для описания произвольных конечных отношений, в логическом смысле обязательно будет эквивалентен алгебре конечных предикатов.

ВЫВОДЫ

Рассмотренный метод описания знаний (метод компараторной идентификации) с использованием алгебры предикатов может дать в перспективе возможность единообразного представления знаний в электронных библиотеках в виде соответствующих уравнений. Любое такое уравнение можно реализовать аппаратно переключательной схемой. Используя переключательные цепи, можно конструировать технические средства обработки и хранения знаний как некоторый интеллектуальный процессор обработки знаний.

ЛИТЕРАТУРА

1. *Когаловский М.Р., Новиков Б.А.* Электронные библиотеки — новый класс информационных систем // Программирование. — 2000. — № 3. — С. 3–8.
2. *Искусственный* интеллект: В 3-х кн. — Модели и методы / Под ред. Д.А. Поспелова. — М.: Радио и связь,1990. — Кн. 2. — 304 с.
3. *Хайрова Н.Ф., Шаронова Н.В.* Автоматизированные информационные системы: задачи обработки информации. — Харьков: Нар. укр. акад., 2002. — 120 с.
4. *Шабанов-Кушнаренко Ю.П.* Теория интеллекта. Технические средства. — Харьков: Вища шк., 1986. — 136 с.
5. *Шабанов-Кушнаренко Ю.П., Шаронова Н.В.* Компараторная идентификация лингвистических объектов. — Киев: ІСДО, 1993. — 116 с.
6. *Шабанов-Кушнаренко Ю.П.* Теория интеллекта. Математические средства. — Харьков: Вища шк., 1984. — 144 с.

Поступила 30.05.2007