

УДК 519.816

СИСТЕМНЫЙ ПОДХОД К РЕШЕНИЮ ЗАДАЧИ АВТОМАТИЗАЦИИ ИНДЕКСИРОВАНИЯ ПО УДК

Т.Н. Жук

Международный научно-учебный центр информационных технологий и систем

НАН и МОН Украины

e-mail: zh_tamara@yahoo.com

УДК (Универсальная десятичная классификация) является единственной международной универсальной системой, позволяющей достаточно детально раскрыть содержание справочно-информационных фондов и обеспечить быстрый поиск информации. Возможность единой систематизации информационных материалов делает систему УДК самой удобной в процессе обмена информацией между странами [1]. К важнейшим требованиям организации материала в систематических каталогах или картотеках относится соблюдение принципа единообразия: литература по какому-либо вопросу всегда должна относиться в один и тот же раздел каталога. Если этот принцип не будет соблюдаться, то, во-первых, каталог или картотека потеряют наиболее ценные поисковые качества – точность и полноту, во-вторых, поиск будет длительным. В издательствах, редакциях научно-технических журналов, в органах научно-технической информации, библиотеках используется обязательное индексирование всех публикаций по УДК, т.е. все информационные материалы в области естественных и технических наук издаются с индексами УДК. Поэтому задача автоматизации процесса индексирования по УДК является весьма актуальной.

Одной из главных отличительных особенностей УДК является иерархическое построение большинства разделов основной и вспомогательных таблиц по принципу детализации от общего к частному с использованием цифрового десятичного кода. Основной ряд УДК содержит классы от 0 до 9. Индексы УДК построены так, что каждая последующая цифра не меняет значения предыдущих, а лишь уточняет, обозначая более частное понятие. При детализации разделов УДК применяется стандартное обозначение тех или иных понятий, а также типовое и аналогичное подразделения:

- общие определители:
 - определители языка;
 - определители народов;
 - определители формы;
 - определители места;
 - определители времени;
 - общие определители с дефисом,
- специальные определители:
 - 1/.-9 определители с дефисом;
 - .01/.09 определители с точкой ноль;
 - '1/'9 определители с апострофом,
- типовые (скользящие) окончания,
- параллельное подразделение.

Для фиксации отношений, существующих между понятиями, отраженными в документах и создании правильного поискового образа документа применяются следующие знаки УДК:

- знак присоединения +
- знак распространения /
- знак отношения : (двоеточие)
- знак отношения :: (двойное двоеточие).

Исходными данными для создания системы автоматизации процесса индексирования по УДК являются название публикации и аннотация. Результатом должен быть индекс УДК (или множество индексов).

Целью создания системы автоматизации процесса индексирования по УДК является разработка системы, принимающей решение по соотносению исходных данных в соответствии с классификационными признаками к одному или нескольким классам рубрикатора и позволяющей облегчить задачу индексирования лицу, принимающему решение (ЛПР) по нахождению индекса УДК.

На основе структуры схемы классификации строятся правила индексирования документов по УДК. Таким образом, система автоматизации должна включать систему однозначного описания этих правил. При этом возникает ряд проблем.

Во-первых, поскольку УДК является единой интегральной системой, это влечет за собой множественность локализации понятий в таблицах УДК, т.е. повторение одного и того же понятия в различных разделах в зависимости от того, в каком аспекте это понятие рассматривается. При индексировании документа необходимо четко определять аспект, в котором рассматривается данный предмет, чтобы соответственно этому аспекту выбрать индекс в таблице УДК. Это означает, что недостаточно найти в таблице индекс данного понятия, надо также выяснить, к какой отрасли знания и в каком аспекте рассматривается это понятие. При этом возникает задача построения словаря слов-синонимов, т.е. множества признаков и соотношения их с множеством классов. Следовательно, система автоматизации индексирования по УДК на этом этапе должна включать в себя систему поддержки принятия решений (СППР).

Во-вторых, наличие в таблицах УДК различных индексов для аналогичных понятий является источником неоднозначного индексирования одного и того же документа, что значительно затрудняет однозначное определение индекса и, следовательно, создание системы автоматизации индексирования.

Задача автоматизации индексирования по УДК требует системного подхода к ее решению. Основной идеей системного подхода является нахождение всех возможных связей и отношений между элементами реализуемой системы. Суть системного подхода состоит в получении общей картины построения системы с учетом отдельных подсистем и взаимосвязей между ними [2]. При системном анализе необходимо учитывать роль лица, принимающего решение (ЛПР) и на каком этапе будет осуществляться взаимодействие системы с ЛПР. В [3] автор с точки зрения пользователя делит СППР на пассивные, активные и кооперативные. Пассивной СППР он называет систему, которая помогает процессу принятия решения, но не может вынести предложение, какое решение принять. Активная СППР делает предложение, какое решение следует выбрать. Кооперативная позволяет ЛПР изменять, пополнять или улучшать решения, предлагаемые системой, посылая затем эти изменения в систему для проверки. Система изменяет, пополняет или улучшает эти решения и посылает их опять пользователю. Процесс продолжается до получения согласованного решения.

Система автоматизации индексирования по УДК должна включать диалоговую систему для ввода исходных данных и взаимодействия с ЛПР, таблицы УДК, правила описания использования этих таблиц, словари ключевых слов по каждому из разделов классификации, СППР по нахождению индекса УДК.

На начальном этапе система автоматизации индексирования по УДК должна выполнять анализ исходных данных и создание списка ключевых слов. Затем осуществляется поиск в словаре, включая возможность сопоставления со словами-синонимами. В том случае, если слова являются омонимами, что затрудняет их соотношение с каким-либо классом, должен обеспечиваться диалог с ЛПР с целью уточнения данных. Следовательно, система принятия решений должна быть

кооперативной и осуществлять взаимодействие с ЛПР. На основе множества ключевых слов и правил индексирования по таблицам УДК необходимо сформировать множество альтернативных решений. СППР индексирования по УДК должна включать также критерий, по которому происходит отбор альтернатив [4]. В соответствии с критерием выполняется принятие решения или определяется возможное множество решений с указанием весов предлагаемых альтернатив. Результаты должны быть представлены лицу, принимающему решение.

Система автоматизации процесса индексирования по УДК представляет собой экспертную систему. В программной реализации диалогового компонента системы необходимо предусмотреть два входа в систему: в режиме пользователя и в режиме эксперта. В режиме пользователя предлагается ввести название материала и его аннотацию или набор ключевых слов. Возможен ввод формы текста, например: статья, учебник, справочник, патент и т.п. В базе знаний системы хранится набор правил, по которым осуществляется выбор индекса УДК. К примеру, индекс УДК для учебника формируется иначе, чем индекс для статьи. В соответствии с этими правилами осуществляется поиск по ключевым словам или их синонимам в таблицах УДК и формирование множества соответствующих индексов. Если множество ключевых слов не задано в явном виде, необходим анализ текста аннотации и формирование в рабочей области экспертной системы трех списков: нейтральных слов (например: разумеется, представлено, имеется, в, на, о, и, таким образом и т.д.), ключевых слов и неизвестных слов в данной системе. При вводе пользователем ключевых слов также формируется список неизвестных слов. В системе существует возможность ввода максимального процента неизвестных слов, при котором система выдаст решение. Если процент неизвестных слов превышает заданный процент, то пользователь получает рекомендацию обратиться в данном случае к эксперту. В режиме эксперта диалоговая система позволяет соотнести неизвестные слова к определенным словарям индексов УДК. Поиск в словарях ведется от нижнего уровня – наибольшей детализации индексов УДК к верхнему – к общим классам. В результате пользователь получает список возможных отсортированных по весам индексов УДК с объяснением их формирования.

Таким образом, применение системного подхода к решению задачи автоматизации индексирования по УДК позволяет смоделировать предметную область, представить постановку задачи более формализовано и обеспечить упрощенный поиск структуры проектируемой системы на этапе синтеза системы.

Литература

1. <http://www.gsnti-norms.ru/norms/common/doc.asp?2&/norms/udc/udcs.htm>
2. Основные сведения об УДК.
3. Моисеев Н.Н. Математические задачи системного анализа. М.: Наука, 1981.
4. Haettenschwiler P, Neues anwenderfreundliches Konzept der Entscheidungsunterstuetzung. Gutes Entscheiden in WirtSchaft, Politik und Gesellschaft. Zuerich: Hochschulverlag AG, 1999. – S. 189–208.
4. Алексеев А.В., Борисов А.Н., Вилюмс Э.Р., Слядзь Н.Н., Фомин С.А. Интеллектуальные системы принятия проектных решений. - Рига: Зинатне, 1997. - 320 с.