УДК 681.3

Е. С. Подвальный,
С. А. Рыков

ПОСТРОЕНИЕ ИНТЕЛЛЕКТУАЛЬНОГО ПОИСКА ДАННЫХ В ЮРИДИЧЕСКИХ СПРАВОЧНЫХ ИНФОРМАЦИОННЫХ СИСТЕМАХ

Рассматриваются вопросы создания интерфейса интеллектуального поиска данных в юридических справочных информационных системах.

Ключевые слова: *правовые справочные информационные системы; правоприменение; программный продукт; интеллектуальный поиск; технологии доступа к данным; программный интерфейс; драйвер; СУБД; словари; нейронные сети.*

UDK 681.3

E. S. Podvalny,
S. A. Rykov

CREATION OF AN INTELLECTUAL SEARCH OF THE DATA IN LEGAL HELP INFORMATION SYSTEMS

Examines questions of creating the interface of intellectual data search in legal reference information systems.

Key words: *legal reference information systems; enforcement; software product; the intellectual search; data access technology; software interface; driver; database; dictionaries; neural networks.*

Правоприменение является одной из наиболее распространенных форм юридической деятельности, которая обеспечивает реализацию правовых норм на практике и от которой во многом зависит нормальное функционирования общества, государственных структур и правовой системы в целом. Несмотря на то, что внешне правоприменение, прежде всего, связано с нормативными аспектами, теорией права, логикой и структурой правовых норм, даже философией права, этот вид деятельности всегда связывался и с решением многих технических проблем. Правоприменительная деятельность не может быть эффективной, если, например, нормативные акты плохо распространяются среди их потребителей, если знание закона является прерогативой только узкого круга лиц и т.д.

Существующая в настоящее время ситуация с правовой базой нашего общества создает множество проблем для юристов-

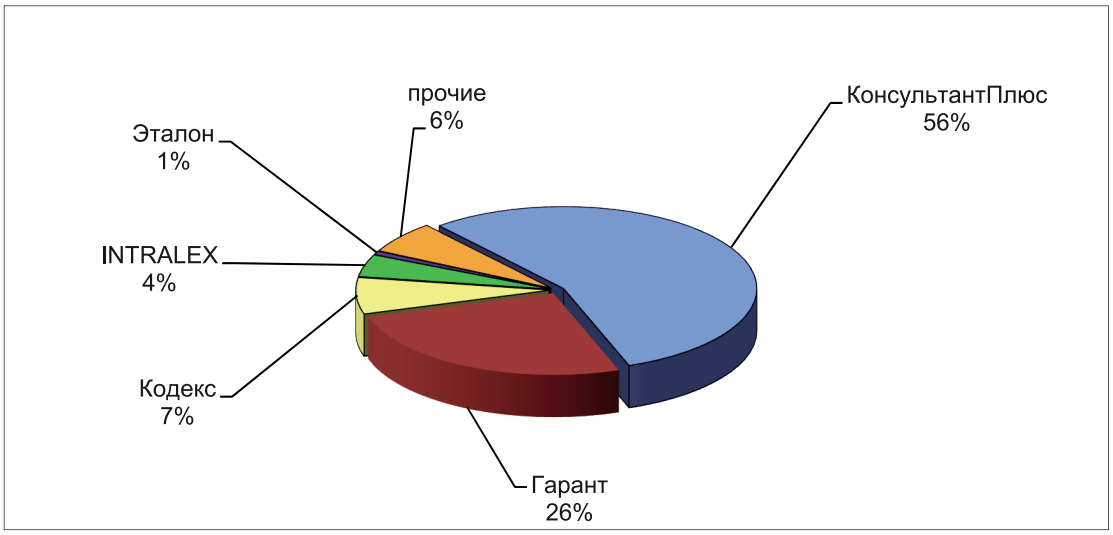


Рис. 1. Соотношение используемых СПС в России по результатам 2011 года

практиков. Прошли времена относительной неизменности законодательной базы, и сегодня поиск соответствующих редакций каких-либо правовых документов является вопросом номер один для всех практикующих юристов – ведь ежедневно принимаются изменения и поправки к законам, указам и т.п. И если при этом федеральное законодательство, в основном, публикуется в специализированных изданиях, то зачастую поиск необходимых ведомственных подзаконных актов сильно усложнен.

Наиболее известными и распространенными в нашей стране можно считать справочные системы «Консультант Плюс» - более 60.000 пользователей, и «Гарант» - свыше 25.000 пользователей. На текущее время распределение используемых СПС на рынке представлено на рисунке 1.

Всего в России насчитывается около 10 производителей различных компьютерных правовых систем, но остальные СПС (кроме упомянутых), по разным причинам, менее распространены и не оказывают заметного влияния на ситуацию на рынке СПС. Это, конечно, совершенно не умаляет их достоинств и ни в коей мере не свидетельствует о технологическом несовершенстве. Однако в качестве исследуемой справочно-правовой системы для построения программного продукта интеллектуального поиска данных используем СПС «Консультант Плюс».

При разработке программного продукта альтернативного интеллектуального поиска в справочной правовой системе «Консультант Плюс» необходимо использовать технологии доступа к данным. Рассматривая ODBC, JDBC, DAO, RDO, OLE DB, ADO как технологии доступа к базе данных справочно-правовой системы, выбор был остановлен на ODBC по причине возможности ее расширения и дополнения собственными драйверами и взаимодействием с любой средой разработки программного обеспечения. Стандартными средствами рассматриваемых технологий доступа к данным, возможность подключения “напрямую” к базе знаний отсутствуют, поэтому необходимо разрабатывать собственный драйвер для ODBC к СПС “Консультант Плюс”.

ODBC (Open Database Connectivity) — широко распространенный программный интерфейс фирмы Microsoft, удовлетворяющий стандартам ANSI и ISO для интерфейсов обращений к базам данных (Call Level Interface, CLI). Архитектура средств реализации ODBC-технологии приведена на рисунке 2.

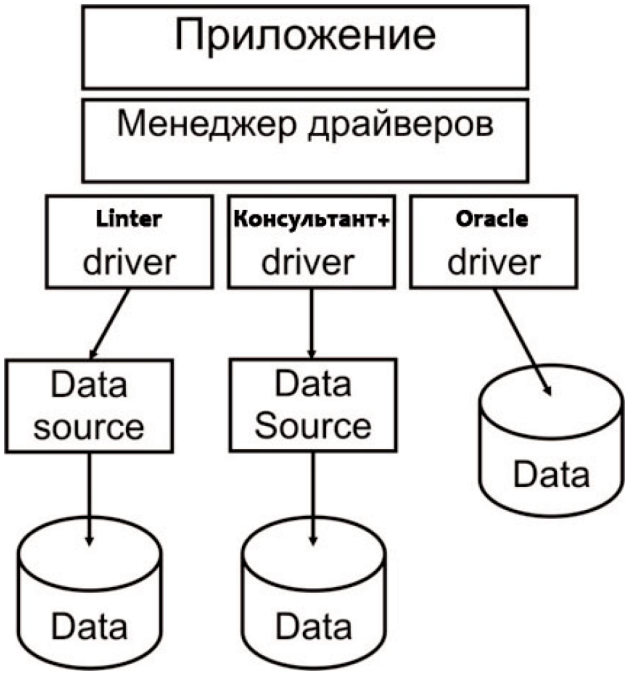


Рис. 2 – Архитектура средств реализации ODBC-технологии

ODBC-драйвер представляет собой динамически загружаемую библиотеку (DLL), которую клиентское приложение загружает в свое адресное пространство и использует для доступа к источнику данных. Для каждой, используемой СУБД, необходим собственный ODBC-драйвер, так как ODBC-драйверы используют функции клиентских API, разные для различных СУБД. При построении драйвера к справочно-правовой системе необходимо проанализировать структуру СУБД и определить основные и индекс файлы.

Конечная справочно-правовая система «Консультант Плюс» содержит множество различных исполнимых и информационных файлов. При разработке программного обеспечения интеллектуального информационного поиска необходимо проанализировать структуру файлов в каталоге BASE в котором содержатся следующие компоненты СУБД: COMMON – общие для комплекта многобазовые словари; DOCS, KRS, MLAW, PAP, PBI –некоторые информационные базы СПС, содержащие файлы, расширения которых означают: *.DT1 – названия документов, *.DT2 – тексты документов, *.DT3

– формы Word, Excel, *.DT4 – гипертекстовые ссылки, *.DT5 – графические объекты и служебная информация, *.DTI – служебная информация для формирования файлов USR. Индексные словари NX формируются на основе соответствующих файлов с расширением *.DT*: NX1 – названий документов, NX2 – текстов документов, NX3 - формы Word, Excel, NX4 - гипертекстовые ссылки, NX5 - графические объекты и служебная информация.

Индексные словари необходимы для оперативного поиска данных по заголовкам документов в условиях работы стандартного приложения «Консультант Плюс», которые являются инверсией данных основной БД. В связи с необходимостью построения более углубленного, но не менее оперативного поиска, в рамках разрабатываемого программного продукта, необходимо реализовать собственный индексатор таблиц базы данных. В результате анализа структуры файлов данных СПС сделан вывод о необходимости построении драйвера ODBC для таблиц данных, схожих по структуре с DBF, формата "*.DT?".

Модуль индексирования (IndexNRN) разрабатываемого интеллектуального интерактивного диалога в условиях использования справочно-правовой системы «Консультант плюс» будет состоять из трех вспомогательных программ (роботов):

Spider (паук) – программа, предназначенная для формирования данных основных разделов (DT1, DT2) данных СУБД и извлечения всех внутренних ссылок на дополнительные разделы. В момент работы данного приложения в ответ робот получает текстовый поток, содержащий служебную информацию и непосредственно сами документы.

Crawler («путешествующий» паук) – программа, которая автоматически проходит по всем внутренним ссылкам (DT3, DT4, DT5), найденным Spider. Выделяет все ссылки, присутствующие в документе. Его задача - определить, куда дальше должен идти паук, основываясь на ссылках или исходя из заранее заданного списка адресов. Crawler, следуя по найденным ссылкам, осуществляет поиск новых документов, еще неизвестных поисковой системе.

IndexNRN (робот- индексатор) - программа, которая анализирует результаты данных сформированные программами Spider и Crawler. Индексатор разбирает документ на составные части и анализирует их, применяя собственные лексические и морфологические алгоритмы. Анализу подвергаются различные элементы страницы, такие как текст, заголовки, ссылки структурные и стилевые особенности, специальные служебные теги и т.д.

Таким образом, модуль индексирования позволяет обходить по ссылкам заданное множество ресурсов, скачивать встречающиеся документы, извлекать ссылки на новые документы и результатами производить обучение нейронной сети Кохонена интеллектуального диалога справочно-правовой системы.

Интеллектуальный анализ данных, основанный на нейронной сети Кохонена, состоит из трех основных этапов: подготовки данных индексатором, извлечения правил (структуры ссылок), оценки правил, графически представленных на рисунке 3. Процесс индексирования и подготовки данных формирует их таким образом, чтобы сделать пригодными для конкретных методов интеллектуального анализа.

Анализ данных, основанный на нейронной сети, работает только с числовыми данными, из чего следует, что необходимо преобразовывать символьные данные в числовые. В разрабатываемом приложе-

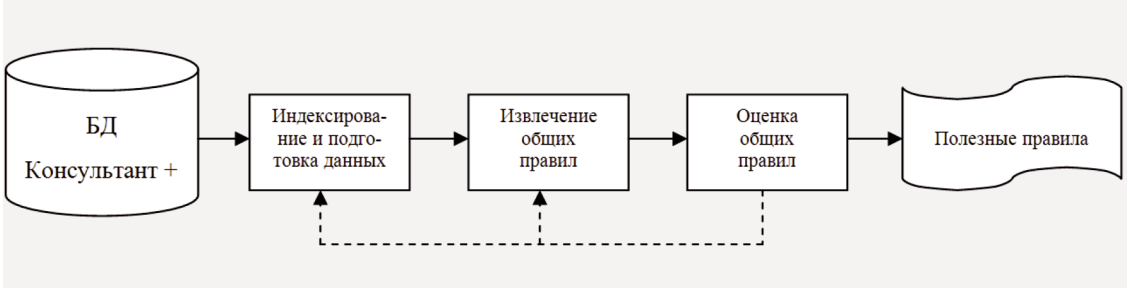


Рис. 3 – Процесс анализа данных, основанный на нейронной сети Кохонена

нии интеллектуального поиска в СПС применяется подход преобразования строковой интерпретации данных в числовое с использованием хэш-функций для создания уникальных числовых данных, соответствующих каждой строке. Несмотря на то, что в реляционной базе данных есть множество типов данных, все они в основном могут быть приведены к символьным, дискретным числовым и непрерывным числовым данным, то есть три логических типа данных.

Преимуществом использования нейронной сети Кохонена в области классификации, кластерного анализа, прогнозирования, сжатия данных следующее: устойчивость к зашумленным данным, неуправляемое обучение, быстрое обучение, возможность визуализации, возможность упрощения многомерной структуры.

Результатом использования справочно-правовой системы, технологии доступа к данным ODBC, построения нового драйвера ODBC, индексатора таблиц базы знаний СПС и нейронной сети Кохонена является конечное программное приложение, реализованное в среде построения ПП C#, обеспечивающее интеллектуальный поиск данных в «Консультант Плюс», представленный в виде интерактивного диалога с пользователем.

ЛИТЕРАТУРА

1. Козлов Е. Б., Метелкин А. В., Хорошевский В. Ф. Мультиагентная система поиска информации в Интернет // Труды седьмой национальной конференции по искусственному интеллекту с международным участием КИИ'2000. –– М.: Физматлит, 2000, С. 840–850.

2. Куршев Е. П., Осипов Г. С., Рябков О. В., Самбу Е. И., Соловьева Н. В., Трофимов И. В. Интеллектуальная метапоисковая система // Труды международного семинара Диалог'2002. Компьютерная лингвистика и интеллектуальные технологии.. –– М.: Наука, 2002, С. 320–330.

3. Культин Н. Основы программирования в Delphi - Спб.: «БХВ-Петербург», 2009 г. – 640 с.

4. Пахомов Б., Самоучитель C/C++ и C++ Builder 2007 - Спб.: «БХВ-Петербург», 2008. — 670 с.

5. Кормалев Д. А., Куршев Е. П., Осипов Г. С., Сулейманова Е. А., Трофимов И. В.: Препринт // Методы поиска и анализа информации. Автоматическое извлечение данных. –– Переславль-Залесский, ИПС РАН, 2003.

6. Хомоненко А., Ададуров С. Работа с базами данных в C++. / А. Хомоненко, С. Ададуров – СПб.:BHV, 2005. – 496 с.