Вопросы разработки естественно-языкового интерфейса к правовым БД.

Обзор правовых документов.

Одной из важнейших составляющих успешного функционирования государства является законодательная система. При этом важно не только качество самих законов и гарантия их исполнения, но и знание гражданами страны этих законов, способность ими правильно пользоваться. Строго говоря, одно без другого невозможно. На практике в современном государстве система законов всегда очень сложна. Одним из путей преодоления этого противоречия является создание электронных правовых баз, которые могли бы обеспечивать удобный поиск нужных законодательных документов. В идеале от таких систем требуется не только обеспечивать поиск, но и давать несложные консультации, т.е. поддерживать диалог по тематике найденных документов.

На сегодняшний день существует несколько популярных электронных баз законодательных документов: Консультант, Гарант, Кодекс. Наиболее крупные из них содержат более 1 миллиона документов. Они поддерживают достаточно развитые возможности поиска. Но принципиально – это поиск только двух видов: на основе ключевых слов и на основе предопределенных рубрик. Все остальные виды поиска (например, интеллектуальный поиск в системе «Кодекс» или поиск по ситуациям в системе «Гарант») являются лишь их вариациями и комбинациями. Эти виды поиска могут быть успешны, только, если пользователю хорошо известна как структура самих документов, так и принципы рубрикации. В остальных случаях необходима возможность искать документы по смыслу или типу тех ситуаций, которые они регулируют. Причем один и тот же смысл может выражаться достаточно разными фразами. Для реализации этих возможностей нет необходимости восстанавливать полную семантику связного текста. Достаточно уметь выделять основные типы ситуаций и отношения между ними. Процесс поиска таких ситуаций не может быть полностью автоматическим. Он должен направляться диалогом системы с пользователем, в результате которого уточняются искомые ситуации и отсеиваются лишние. В результате пользователь должен получать не только набор законодательных документов, но и некоторую их трактовку в применении к интересующим его ситуациям. Такая трактовка ни в коей мере не заменит консультацию с профессиональным юристом, но позволит хотя бы в общих чертах ориентироваться в проблеме.

Рассмотрим основные виды анализа, необходимые для реализации вышеупомянутых возможностей.

1. Предварительный анализ документа. На этом этапе распознаются специальные символы и элементы разметки документа, которые не являются обычными словами, сокращениями, числами или датами.
2. Анализ структуры документа. Особенностью правовых документов в сравнении с произвольными текстами является достаточно жесткая и формальная структура текста. Значительная часть семантики определяется именно той структурной частью, к которой относится фрагмент текста. На основе правил, описывающих общую структуру документа, могут распознаваться элементы структуры (такие как название, дата утверждения, подписи, ссылки на другие документы и т.д.). В процессе анализа

используются результаты предварительного анализа, а также элементы морфологического и синтаксического анализа.

3. Морфологический и синтаксический анализ, в результате которого определяется структура предложений документа. Для этих целей предполагается использовать уже существующую систему, рассчитанную на работу с общеязыковыми текстами (см.[1]).
4. Семантический анализ, в результате которого всем значимым словам и словосочетаниям присваиваются семантические классы. После того как присвоены семантические классы, возможен поиск с учетом семантики понятий. Этот этап также основывается на системе, описанной в [1] и [2] Здесь требуется доработка семантического словаря, пополнение его понятиями, специфическими для юридической тематики, уточнение существующих определений.
5. Ситуационный или прагматический анализ. Позволяет выявлять основные типы ситуаций вне зависимости от способа их выражения в тексте документа. Эта задача значительно упрощается тем обстоятельством, что законодательные документы по самой своей сути предназначены для описания стандартных ситуаций и отношений. Степень формализованности как самих ситуаций, так и способов их выражения, значительно выше, чем в произвольных текстах. Здесь необходима разработка прагматического словаря или словаря «типичных ситуаций».

Остановимся подробнее на первых двух видах анализа.

В общем случае, кроме текста правового документа в документе есть еще и дополнительная информация, которая важна для организации хранения и поиска документов и может быть значима для их правильного использования. Например, дата выпуска документа, считается датой, с которой содержимое документа становится истинным с точки зрения закона, за исключением тех случаев, когда в тексте явно не указана дата введения документа в силу. Другие документы могут ссылаться на искомый документ по заголовку к тексту и его регистрационному номеру.

Задачами анализа структуры документа являются:

1. Выделение реквизитов документа необходимых для распознавания ссылок на него (заголовок к тексту, регистрационный номер).
2. Выделение реквизитов несущих дополнительную информацию (дата документа, название типа)
3. Выделение текста документа для передачи его семантическому анализатору.

Все законодательные документы можно разделить на две большие группы. К одной группе относятся базовые законодательные документы. В рамках Российского государства - это свод законов, который представлен статьями. Однако при выявлении тех административных случаев, когда существующих статьей не достаточно для урегулирования сложившейся ситуации, органы управления выпускают дополнительные организационно-распорядительные документы (постановления, распоряжения, приказы и т.д.).

Анализ структуры документов

Как уже говорилось, мы имеем дело с документами базовыми (статьи) и дополнительными организационно-распорядительными (постановления, распоряжения, приказы и т.д.). Задачей анализа структуры законодательного документа будем считать выделение содержимого документа и получение других его реквизитов. Для выявления реквизитов документа рассмотрим подробнее форматы законодательных документов.

статья.

Статьи в зависимости от тематики группируются по кодексам. За кодексом закрепляется дата выпуска, которая так же является датой вступления в силу его статей. Другими словами дата вступления в силу статьи определяется датой выпуска кодекса. Кодекс так же имеет название, которое является уникальным.

Таким образом, статья имеет два дополнительных реквизита:

- Дата вступления в силу;
- Название кодекса, в котором содержится статья.

Рассмотрим теперь какие данные можно извлечь непосредственно из документа статьи.

пример статьи из Гражданского Кодекса Российской Федерации:

Статья 9. Осуществление гражданских прав

1. Граждане и юридические лица по своему усмотрению осуществляют принадлежащие им гражданские права.
2. Отказ граждан и юридических лиц от осуществления принадлежащих им прав не влечет прекращения этих прав, за исключением случаев, предусмотренных законом.

Как видно из примера кроме текста статья еще имеет порядковый номер в кодексе и название. Для выделения этой информации из документа нам достаточно знать, что такие документы начинаются со слова *Статья,* затем идет ее номер затем название, которое заканчивается пустой строкой. Все остальное содержимое статьи считается текстом.

Таким образом, статья имеет следующие реквизиты:

- Название кодекса, в котором она опубликована;
- Дата выпуска (соответствует дате выпуска кодекса);
- Номер в кодексе;
- Название статьи;
- Содержимое статьи.

Организационно-распорядительные документы.

Дополнительные организационно-распорядительные документы есть принадлежность выпускающих правовых организаций и не имеют группировки по тематикам. Таким образом, рассматриваемые документы не имеют реквизитов объявленных вне документа. Формат такого документа описывается в стандарте унифицированной системы организационно-распорядительной документации “ТРЕБОВАНИЯ К ОФОРМЛЕНИЮ ДОКУМЕНТОВ”

ГОСТ Р 6.30-97 (от 31 июля 1997 г.) с изменениями (от 21 января 2000) введенными в действие с 1 апреля 2000 года.

Такой документ в соответствии с указанным стандартом делится на три части

1. Заголовочная часть документа;
2. Основная часть документа;
3. Оформляющая часть документа.

Указанный стандарт охватывает более широкий спектр документов, чем рассматриваемый в данной публикации. Поэтому в дальнейшем будем рассматривать лишь те части документа, которые входят в состав организационно-распорядительного документа.

Заголовочная часть документа

В соответствии с вышеуказанным стандартом организационно-распорядительные документы федеральных органов государственной власти, органов государственной власти субъектов Российской Федерации начинаются с названий тех правовых организаций и их подразделений, которыми был выпущен.

В случае если таковая организация одна, то за ее наименованием следует наименование типа документа, а затем дата этого документа и регистрационный номер, а затем заголовок к тексту.

Например:

Администрация Санкт-Петербурга
КОМИТЕТ ЭКОНОМИЧЕСКОГО РАЗВИТИЯ,
ПРОМЫШЛЕННОЙ ПОЛИТИКИ И ТОРГОВЛИ

ПРИКАЗ

от 20 мая 2002 года N 98-п

Об утверждении Временного положения о порядке расчета
и предоставления субсидий на работы и услуги по содержанию и
текущему ремонту жилищного фонда жилищно-строительных и жилищных
кооперативов, товариществ собственников жилья и иных объединений
собственников недвижимости в жилищной сфере, расположенного
на территории Санкт-Петербурга

...

В случае если же авторами документа являются несколько организаций, то непосредственно за наименованием каждой организации следует дата и регистрационный номер, а затем уже наименование типа документа и заголовок к тексту.

Например:

Администрация Санкт-Петербурга
КОМИТЕТ ПО СОДЕРЖАНИЮ ЖИЛИЩНОГО ФОНДА

от 26 декабря 2001 года N 198

КОМИТЕТ ФИНАНСОВ

от 26 декабря 2001 года N 99-р

РАСПОРЯЖЕНИЕ

О введении билета для студентов и учащихся 9-11 классов
на основе смарт-карт

Датой документа является дата его подписания или утверждения, для протокола - дата заседания (принятия решения), для акта - дата события. Если авторами документа являются несколько организаций, то датой документа является наиболее поздняя дата подписания.

Регистрационный номер документа состоит из его порядкового номера, который можно дополнять по усмотрению организации индексом дела по номенклатуре дел, информацией о корреспонденте, исполнителях и другой. Регистрационный номер документа, составленного совместно двумя и более организациями, состоит из регистрационных номеров документа каждой из этих организаций, проставляемых через косую черту в порядке указания авторов в документе.

Заголовок к тексту содержит краткое изложение основного смысла составляемого документа. Заголовок должен грамматически согласовываться с название вида документа, отвечая на вопрос "о чем?"

Например:

Приказ
О создании аттестационной комиссии

Основная часть документа

Стандарт ГОСТ Р 6.30-97 накладывает некоторые правила на содержимое текста документа, которые в последствии могут помочь при построении семантического дерева.

Если текст содержит несколько решений, выводов и т.д., то в соответсвии со стандартом он должен быть разбит на разделы, подразделы, пункты, которые нумеруют арабскими цифрами.

В распорядительных документах (приказ, распоряжение и т.д.) организаций, действующих на принципах единоначалия, а также документах, адресованных руководству организации, изложение текста идет от первого лица единственного числа ("приказываю", "предлагаю", "прошу").

В распорядительных документах коллегиальных органов текст излагают от третьего лица единственного числа ("постановляет", "решил").

В совместных распорядительных документах текст излагают от первого лица множественного числа ("приказываем", "решили").

Оформляющая часть документа.

Оформляющая часть документа в первую очередь содержит реквизит “подпись” и печать. Так как электронная копия документа в текстовом формате не содержит печать то этот реквизит выходит из рассмотрения.

В состав реквизита "Подпись" входят: наименование должности лица, подписавшего документ (полное, если документ оформлен не на бланке документа, и сокращенное - на документе, оформленном на бланке); личная подпись; расшифровка подписи (инициалы, фамилия),

При подписании документа несколькими должностными лицами их подписи располагают одну под другой в последовательности, соответствующей занимаемой должности

В документах, составленных комиссией, указывают не должности лиц, подписывающих документ, а их обязанности в составе комиссии в соответствии с распределением.

Например:
Председатель комиссии Н.В.Куликов
Член комиссии К.М.Артемьева

Таким образом, кроме содержимого организационно-распорядительные имеет следующие реквизиты:

- Данных от тех правовых организаций и их подразделений, которыми был выпущен (название организации, дата , регистрационный номер);
- Дата документа (является наиболее поздняя дата подписания дате среди дат закрепленными за организациями авторами);
- Регистрационный номер (состоит из регистрационных номеров документа каждой из этих организаций, проставляемых через косую черту в порядке указания авторов в документе);
- Заголовок к тексту.
- Содержимое документа.
- Подписи должностных лиц (названия должностных, инициалы и фамилии должностных лиц).

О реализации механизма анализа структуры документов.

Из описания форматов законодательных документов можно сказать, что анализ статьей есть более простая задача, чем анализ организационно-распорядительных документов. Однако в любом случае это можно с уверенностью сказать, что задача подготовки законодательного документа имеет решение. Теперь остается вопрос о выборе средства для реализации такого анализа.

Знания только о порядке частей документа не всегда достаточно при анализе организационно-распорядительных документов. Например, как определить, что текущий блок документа представляет собой часть содержимого документа или же представляет собой реквизит подпись, если предыдущий блок представлял собой часть содержимого.

Анализируя многие реквизиты документов можно найти общие или схожие части в рамках приведенного примера можно сказать, что реквизит подпись начинается с названия должности (множество названий должностей имеющих право подписывать организационно-

распорядительные документы есть число ограниченное) и заканчивается инициалами и фамилией должностного лица.

При рассмотрении других реквизитов документа можно так же найти характерные особенности. Например, реквизит дата и регистрационный номер имеют декларированный стандартом формат, заголовок к тексту начинается с “О “ или “Об “. Количество наименований типов документов в рамках нашей задачи есть число конечное (распоряжение, приказ и т.д.).

Таким образом, задача построения механизма подготовки сводиться к построению таблицы возможных переходов реквизитов при последовательном проходе текстовых блоков документа и в рамках разрешения конфликтных ситуаций механизм идентификации реквизита. Последняя задача очень просто решается при использовании шаблонов регулярных выражений.

Использование шаблонов регулярных выражений.

Регулярные выражения или шаблоны впервые в реализации появились в системах Юникс. Выражения и синтаксис регулярных выражений заимствован из свободно распространяемых процедур V8 Генри Спенсера (Henry Spencer). Затем механизм регулярных выражений с расширенным синтаксисом появился в языке Перл.

Рассмотрим детально как можно квалифицировать различные реквизиты документа с помошью регулярных выражений. Для начала разобьем документ на блоки разделителем которых в исходном документе будет наличие пустой строки.

При выявлении того или иного реквизита будем так же использовать знание о порядке реквизитов. Таким образом будем считать что рассматриваемый блок документа является соответсвующим реквизитом если:

- Удовлетворяет соответствующему шаблону регулярных выражений.
- Является корректным в соответвии со структурой документа (на основе знаний о предшествующем реквизите).

В рамках регулярных выражений реквизит заголовка документа можно выразить следующим шаблоном

*(О|Об)[ ].**

а подпись

(Президент|Губернатор|Председатель|Глава|Член)[ ]..[А-Я][.][А-Я][.][А-Я].**

Номер статьи и ее название определяется шаблоном

*Статья (\n|\n\n|\n\n\n).**

Дата документа и регистрационный номер:

от (((\d|\d\d)
(января|февраля|марта|апреля|мая|июня|июля|августа|сентября|октября|ноября|декабря)
\d\d\d\d)|(\d|\d\d[.]\d\d[.]\d\d\d\d)|(\d\d\d\d[.]\d\d\[.]d\d)) года N.*

Название типа документа:
РАСПОРЯЖЕНИЕ|ПРИКАЗ|ПОСТАНОВЛЕНИЕ|РАСПОРЯЖЕНИЕ ...

Так как текстом может быть любая комбинация символов, то представление текста следует выразить как

.*

Для выявления данных правовой организации (автора), шаблон можно оставить таким же как и у текста документа . В данном случае совпадение шаблонов допустимо потому что в соответсвии со структурой документа декларированной в стандартах [4] и [5] эти реквизиты никогда не могут быть расположены друг за дркгом.

На базе описанных знаний был разработан програмный модуль на языке программирования Java. Даный модуль реализовывает преобразовывание электронного документи из текстового формата в XML формат с выделением реквизитов документа в отдельные элементы для выполнения коректных морфологических и синтаксических анализов в последствии.

Литература

1. Тузов В.А. Компьютерная семантика русского языка // ДИАЛОГ'2001, Труды межд. семинара. М., 2001.
2. Каневский Е.А., Тузов В.А. НЕКОТОРЫЕ ВОПРОСЫ ПОПОЛНЕНИЯ СЕМАНТИЧЕСКОГО СЛОВАРЯ ТЕРМИНАМИ ПРЕДМЕТНОЙ ОБЛАСТИ // ДИАЛОГ'2002, Труды межд. семинара. М., 2002.
3. Электронные публикации Кодексов http://znai-zakon.narod.ru/kodeks.htm
4. Унифицированная система организационно-распорядительной документации ГОСТР 6.30-97 //Принят и введен в действие постановлением Госстандарта РФ от 31 июля 1997 г. # 273 с 1 июля 1998 г
5. ИЗМЕНЕНИЕ N1 ГОСТ Р 6. 30 - 97//Изменение подготовлено Федеральной архивной службой России и утверждено постановлением Госстандарта России "О принятии и введении в действие изменения ГОСТ Р 6. 30-97" от 21.01.2000 N 9-ст с датой введения в действие с 1 апреля 2000 года.
6. Владимир Маслов. Введение в Perl (раздел "Шаблоны регулярных выражений") http://www.utl.ru/pub/DOCS/PERL/perl.htm