Автоматическое выделение гипертекстовых переходов в текстах документов

Губин М.В. (max@kodeks.ru)
Меркулов А.И. (u_andrewshi@pisem.net)
Информационная компания «Кодекс»

Современные информационные полнотекстовые системы - электронные библиотеки, правовые информационные системы, хранилища систем документооборота, практически всегда поддерживают возможность создания гипертекста, однако разметка ссылок требует больших трудозатрат. В статье рассматривается разработка, настройка и оценка качества работы системы автоматического выделения ссылок в нормативных, технических и экономических документах. В таких документах потенциальные связи выделены определенными языковыми конструкциями, как правило, содержащими атрибуты документа. Например: «согласно ст.15 Конституции РФ», «удовлетворяет требования ГОСТ 12.307-94», «действуя на основании Устава» и т.д. В основе используемого алгоритма выделения ссылок использовался синтаксический анализатор на базе конечного автомата, который описывается в специальном настроечном файле системы. Анализатор, сканируя текст, выделяет варианты атрибутов документов и мест в документе, куда должен осуществляться переход. Далее эти варианты проверяются путем поиска документа по атрибутам, и, с помощью определенных эвристик, отбирается наиболее вероятный вариант. Для удобства настройки поиск и отбор варианта программировался на специальном языке сценариев и находился в том же файле настроек, что и описание анализатора. В статье описываются особенности реализации и проблемы, с которыми пришлось встретиться разработчикам при адаптации к различным коллекциям нормативных и нормативно-технических документов. Качество работы системы проверялось на нескольких тысячах реальных документов путем ручной проверки экспертами. Точность работы системы не хуже 75%, то есть не менее 3/4 ссылок было правильно размечено автоматически. Производительность системы составляет несколько документов в секунду. В статье приведены подробные данные об оценке системы. В статье описываются предполагаемые направления развития и совершенствования системы.

Современные информационные полнотекстовые системы - электронные библиотеки, правовые информационные системы, хранилища систем документооборота, практически всегда поддерживают возможность создания гипертекста. Но разметка ссылок в текстах требует больших трудозатрат. В статье рассматривается разработка, настройка и оценка качества работы системы автоматического выделения ссылок в нормативных, технических и экономических документах. В таких документах потенциальные связи выделены определенными языковыми конструкциями, как правило, содержащими атрибуты документа. Например: «согласно ст.15 Конституции РФ», «удовлетворяет

требования ГОСТ 12.307-94», «действуя на основании Устава» и т.д. Это делает возможным создание автоматизированной системы, которая выделяет такие конструкции и размечает ссылки. Данная статья посвящена описанию подобной системы.

Постановка задачи

Сформулируем проблему, которую нам необходимо решить:
Имеется коллекция документов, в нашем случае рассматривались коллекции нормативных (федеральное законодательство России) и нормативно-технических документов (СНиПы и ГОСТы). Из практики использования нормативных справочных систем, систем технической документации и хранилищ систем документооборота можно выделить следующие атрибуты документа:

1. Название документа.
2. Вид документа (закон, приказ, ГОСТ, СНИП и т.д.).
3. Номер документа. (Идентификационный номер, присвоенный автором или при регистрации документа).
4. Дата принятия, регистрации или подписания документа.

В текстах документов содержаться ссылки на другие документы с указанием их атрибутов. Задача состоит в:

1. Выделение фрагментов текстов, содержащих ссылки на другие документы;
2. Выделение из фрагмента поисковых атрибутов и поиск по ним документов;
3. Поиск рядом с выделенным фрагментом указаний на место в документе, куда идет ссылка (глава, статья и т.д.) и установка гиперссылки на это место в документе. Если указание на место не найдено, то связь устанавливается на начало документа.

Известные существующие решения

Из описанных в литературе систем, которые решают подобные задачи, нам известна только система Syntalex[1]. Судя по описанию, авторы использовали подход аналогичный нашему, но это система предназначена только для обработки англоязычных текстов в области права, кроме того, авторы не приводят никаких данных о качестве работы системы и ее производительности. Первые публикации об Syntalex появились в 2000 году, в то время как первые версии нашей системы были практически внедрены в 1997.
Наиболее близким по подходам из известных российских систем является продукт «RCO Pattern Extractor библиотека выделения объектов в тексте», разработанный фирмой «Гарант-Парк-Интернет»[2]. Однако авторы этого продукта, насколько нам известно, не использовали ее с целью выделения ссылок и не проводили оценки качества выделения каких-либо объектов в тексте с помощью своей библиотеки.

Описание использованного метода

После анализа текстов, предназначенных для обработки, стало очевидно, что в большинстве случаев подобные фрагменты можно было выделить, основываясь на пунктуации и стоящих рядом ключевых словах – названия заключены в кавычки, перед номером документа, как правило, предшествует символы N или №.
Для выделения ссылок использовался простой анализатор текста, выделяющий последовательности объектов, описываемые шаблоном. Виды выделяемых объектов приведены в «Таблица 1. Выделяемые из текста объекты».

Таблица 1. Выделяемые из текста объекты

Объект	**Значение**
"пример"	Такому объект соответствуют любые однокоренные слова к слову пример. Т.е. - "примеров", "примеры", "примерный" и т.д. в любом падеже.
"пример!"	Имеется в виду точное соответствие, т.е. слово "пример".
"=1:10"	Подстрока-число, принадлежащее указанному интервалу, в данном случае - от 1 до 10.
"#*"	Любой номер документа (цифро-буквенная последовательность), например, № 52-ФЗ.
"#52-ФЗ"	Конкретный номер документа, в данном случае - № 52-ФЗ.
"$выражение"	Подстрока, соответствующая регулярному выражению выражение.
"*D"	Подстрока, представляющая собой любую дату, указанную в последовательности число-месяц-год, например: 12 января 97 года 5 марта 1995 10.08.2003 01.03.03
"*20021201"	Подстрока, представляющая собой конкретную дату, в данном случае - 1 декабря 2002 года.
"*"	Любое слово.
"~Q"	Любая фраза, заключенная в кавычки, например, название документа. При этом, в случае вложенных кавычек, в соответствие берется фраза целиком - по границам внешних кавычек: "О банкротстве" "О внесении изменений в закон "О бакротстве"

Объекты типа «Номер» или «Дата» несколько избыточны, т.к. их можно задать через объекты других типов, но они введены отдельно т.к. они часто встречаются и, кроме того, их выделение уже было реализовано для реализации других задач обработки текстов.
Примером простейшего шаблона, используемого системой для выделения гиперссылок:

"от!" "*D" "#*"

Данному шаблону соответствует фрагментам текста: «от 12.02.94 N 137» или «от 21 февраля 1991 года № 123-ПГ».
Для задания более сложных фрагментов можно использовать группирующие скобки и оператор ИЛИ, например такой шаблон:

"статья" "=1:137" "Конституции! " (("Российской!" "Федерации!")|"РФ"|"России!")

соответствует следующим фрагментам текста: «статья 5 Конституции Российской Федерации» или «статьи 9 Конституции России Федерации» или «статьей 10 Конституции РФ».
Как видно, описание шаблонов не предполагает рекурсии и даже включения шаблона в шаблон. Это было сделано специально, по следующим причинам:

- Введение этих возможностей значительно усложняет написание и отладку шаблонов. Если необходимо большое количество однотипных шаблонов, то они генерируются статически внешней утилитой.
- Отсутствие этих средств позволяет значительно ускорить обработку большого количества шаблонов.

Шаблоны описываются в виде текстовых строк в специальном настроечном файле в формате XML. При создании шаблонов использовалась ручная обработка и генерация шаблонов с помощью специальной утилиты с последующим ручным редактированием. Каждому шаблону соответствует своя функция-обработчик, при обнаружении шаблона она вызывается, и, в качестве параметров, ей передаются найденные в тексте значения объектов шаблона. Обычно логика работы функций-обработчиков достаточно проста – сформировать запрос по полученным данным и в случае, если он вернул один документ сформировать ссылку и уведомить систему о том, что для данного фрагмента текстов дополнительной обработки не нужно. Однако, в некоторых случаях требуются более сложные действия, например, в случае использования атрибута названия, при неудачном поиске функция пытается произвести поиск не учитывая название и потом анализирует найденные документы. Если у одного из них название достаточно близко к заданному, то данный документ считается подходящим, иначе происходит переход к следующему шаблону. Кроме этого, функция обработчик имеет доступ ко всем атрибутам обрабатываемого документа и массиву уже выделенных ссылок, что позволяет обрабатывать ссылки, например, на статьи упоминавшихся выше документов, например фрагмент «согласно 3 статье указанного Постановления».

Шаблоны в настроечном файле упорядочены специальным образом, т.к. если для данного фрагмента текста подходит несколько шаблонов, то первым вызывается обработчик для шаблона, указанного в файле ранее. Например, учитывающий наименование документа шаблон размещен в файле ранее, чем учитывающий только номер и дату, тем самым обеспечивается разметка ссылок максимально охватывающих в тексте упоминание документа.

Система может включать несколько настроечных файлов шаблонов, что позволяет организовать их модульную структуру и не обрабатывать не нужные для данного набора документов. Например, при анализе актов регионального Московского законодательства используются блоки шаблонов для федерального законодательства и специальный массив шаблонов московского законодательства. Отдельный модуль шаблонов позволяет выделять внутренние ссылки между частями документа.

Количество шаблонов используемых, например, для обработки российского законодательства составляет около 150, а общее количество шаблонов во всех созданных у нас сейчас файлах более 1000.

Оценка качества

Разработка данной системы велась для ведения информационно-справочной базы данных с очень высокими требованиями по качеству обработки материалов. При этом каждый документ, после автоматического выделения гиперссылок, проверялся операторами вручную, при этом все установленные ссылки тестировались и, при необходимости, исправлялись и устанавливались пропущенные.

Для проверки качества работы системы было написано специальное программное обеспечение, которое производило повторную расстановку связей в документе, проверенным оператором, и сравнение ссылок. При этом вычислялось количество ссылок, которые автомат установил правильно и количество пропущенных ссылок. В «Таблица 2. Оценка качества работы» приведены результаты тестирования на документах российского законодательства и нормативно-технических документах.

Таблица 2. Оценка качества работы

Выборка	**%**	**Найдено**	**Всего**

«Военная служба» 7 федеральных законов	82	201	246
Законы о коммерческих организациях 5 федеральных законов	97	525	541
Законы о пенсионном обеспечении 9 федеральных законов	85	462	547
Законы о рынке ценных бумаг 4 федеральных закона	96	174	181
Законы о труде 11 федеральных законов	83	169	203
Кодексы РФ 25 документов	81	5970	5252
Санитарные номера и правила	88	804	910

Данный метод оценки вносит некоторую ошибку, связанную с тем, что кроме ссылок описанного типа, операторы проставляли так же дополнительные ссылки. Например, многие статьи кодексов Российской Федерации содержат отсылки вида «определяется Федеральным Законом», без указания конкретных атрибутов закона, т.к. на момент принятия кодекса такой закон еще не был разработан. Однако оператор, являясь юристом-специалистом в данной области, может указать ссылку на закон, что, конечно, не может сделать система в рамках описанного алгоритма, т.к. атрибутов документа не достаточно. Именно это объясняет плохие результаты по выборкам «Кодексы» и «Законы о труде». В случае, если ссылка содержит упоминания атрибутов документов, то вероятность распознавания близка к 99%.

Производительность системы составляет единицы документов в секунду на компьютере Pentium II 400 Mhz и определяется в основном скоростью обработки запросов на поиск документов по данным атрибутам.

Выводы и перспективы

Использованный подход позволяет с высокой степенью достоверности выделять ссылки на другие документы и правильно размечать их в текстах нормативных и деловых документов.

Система используется для разметки ссылок у нас с 1997 года, кроме этого она внедрена в ряде сторонних организаций, в том числе в Научном Центре Правовой Информации Минюста РФ.

Была предпринята попытка использования данного алгоритма для обработки запросов пользователей к поисковой системе. Но предварительная оценка показала, что реально пользователи очень часто формулируют запросы не так, как документы упоминаются в тексте, поэтому выделение фрагментов не срабатывало, и качество поиска заметно не изменилось. Однако для обработки поисковых запросов был разработан свой набор поисковых шаблонов.

Система постоянно совершенствуется – разрабатываются новые библиотеки шаблонов для различных предметных областей.

Литература

1. Needle J, 'The Automatic Linking of Legal Citations', 2000 (3) The Journal of Information, Law and Technology' (JILT). http://elj.warwick.ac.uk/jilt/00-3/needle.html
2. "RCO Pattern Extractor библиотека выделения объектов в тексте" http://www.rco.ru/product.asp?ob_no=17