Участие ИПС «Кодекс» в семинаре РОМИП 2004

© Губин М.В.

ИК «Кодекс»
max@gubin.spb.ru

Аннотация

Эта статья описывает опыт участия информационно-поисковой системы «Кодекс» в семинаре РОМИП в 2004 году.
В статье кратко описывается история создания коллекции, представленной нашей фирмой в этом году для использования в заданиях семинара.
Описаны основные изменения, которые были внесены в систему в этом году и отличия варианта системы, на котором выполнялись задания семинара, от коммерческой версии. В статье перечислены цели и задачи, которые мы ставили перед собой, принимая участие в семинаре, и что из запланированного удалось выполнить.
Статья содержит краткий анализ полученных результатов и наши предложения по заданиям семинара на 2005 год.

Введение

Мы участвуем в семинаре РОМИП уже во второй раз. Данная инициатива нам кажется очень полезной для развития средств поиска в разрабатываемых нами систем. Пользуясь материалами семинара, общаясь со своими коллегами из других организаций, мы смогли, как нам кажется, лучше понять какие направления развития системы являются более перспективными, получили новый опыт и методики.

В 2004 в РОМИП в заданиях двух дорожек использовались материалы, подготовленные нашей фирмой. Поэтому, в этой статье мы хотим кратко рассказать историю возникновения и основные принципы формирования этих материалов. Кроме этого, в ней будет

рассказано об основных изменениях, которые были внесены в систему по сравнению с прошлым семинаром и отличия тестовой системы, участвовавшей в семинаре от коммерческой версии.

К сожалению, в этом году не все прошло настолько гладко, как хотелось бы. В тестовом программном обеспечении, когда уже были получены результаты, была обнаружена ошибка, что не позволило продемонстрировать настолько высокие результаты, насколько мы ожидали.

Коллекция

В этом году на семинар нашей фирмой была представлена коллекция нормативно-правовых документов. Она содержит основные правовые документы законодательства России, изданные федеральными органами власти, на состояние начала 2004 года. Все документы были представлены последними версиями, старые версии не включались. Из текстов документов были максимально удалены все комментарии и другая вспомогательная информация. Это было сделано чтобы, с одной стороны, сделать коллекцию максимально «нейтральной», не привязанной к данному производителю, с другой стороны, на эти материалы накладывались существенные ограничения по авторским правам.

Небольшая историческая справка: Данная коллекция ведется с 1990 года, однако она содержит и документы изданные ранее, но действующие сейчас. Первоначально эта база данных была создана по заказу Исполкома Ленсовета, и в нее вносились в основном Ленинградские документы. В 1992 году проект стал коммерческим. Сейчас наша фирма и ее партнеры ведет около 300 документарных баз данных разных направлений. Законодательство России было выделено в отдельную базу в 1994 году. В настоящее время ежедневно в базу вводят около 300 документов и актуализируют тексты до 50 документов. Базу данных России ведет около 60 специалистов – юристов, операторов, корректоров, машинисток.

В семинаре с использованием этой коллекции было выполнено две дорожки – legal ad hoc и классификация. Первая представляла собой задание поиска по документам с использованием реальных запросов пользователей, взятых из протоколов некоторых правовых Интернет сайтов. Вторая дорожка – классификация документов. В качестве рубрик использовалось подмножество рубрик созданного нашей фирмой классификатора правовых документов.

Немного об этом классификаторе. В первых версиях системы (до 1993 года) не было средств поиска по текстам, и классификатор был основным средством поиска по содержанию документов. В основу

первых версий классификатора легли некоторые разделы УДК (Универсального Десятичного классификатора) и классическое юридическое разбиение правовой информации по видам. Каждый документ коллекции мог тогда, так же как сейчас, иметь произвольное количество рубрик. Ориентация на классификатор, как поисковое средство, привело к тому, что в нем появилось множество рубрик, которые удовлетворяли конкретным запросам пользователей. Для этого же некоторые темы, например связанные с налогами, «поднимались» на более высокие уровни иерархии классификатора.

С появлением в системе мощных средств контекстного поиска роль классификатора как поискового средства постепенно снижалось. По нашим данным, сейчас менее 5% пользователей при формировании запросов указывают тематики классификатора. По нашему мнению это связано со следующим:

1. Мощные средства поиска по тексту позволяют искать информацию, не обращаясь к классификатору;
2. Даже для пользователя хорошо знакомого с классификатором сложно в поисковой форме выбрать наиболее подходящую рубрику в большом их списке;
3. Пользователи, имеющие опыт работы с Интернет поисковыми системами предпочитают стандартные средства поиска по тексту, как наиболее привычные.

В связи с этим, классификатор в настоящее время используется как рубрикатор - средство навигации в правой базе. К каждой рубрике написан комментарий, который поддерживается юристами в актуальном состоянии. Кроме этого, классификатор используется для нашей внутренней работы над документами.

Для РОМИП использовался второй уровень иерархии классификатора. Все интересующиеся могли получить полный классификатор, чтобы иметь представление о взаимном положении в нем тематик.

В этом году мы не участвовали в дорожке классификации. Это связано с тем, что используемые у нас алгоритмы автоматической классификации не подходят для участия в семинаре. В задании необходимо было классифицировать документы на основании относительно небольшой обучающей выборки с использованием текстов документов. В нашей практической работе мы имеем большое количество уже классифицированных документов, и задачу классифицировать один новый добавляемый документ. Кроме этого, документ имеет намного больше атрибутов, чем в коллекции РОМИП – вид документа, принимающий орган и т.д. Ну и

последнее, крайне важное отличие, состоит в том, что результаты классификации всегда проверяются экспертами юристами, что обусловлено высокими требованиями по качеству обработки. При этом эксперту намного проще удалять «лишние» тематики, чем добавлять новые, поэтому наши алгоритмы ориентированы на установку избыточного числа тематик. Так как переделка наших технологий потребовала бы слишком больших изменений, мы не участвовали в дорожке классификации в этом году.

Тестовая система

В этом году использовалась доработанная версия системы. Основной целью изменений было внесение в алгоритм взвешивания документов учета взаимного положения слов.

Взвешивание документов с учетом взаимного положения слов.

Ряд исследований [1], в том числе и выполненных нами [2], показывают, что качество поиска может быть улучшено за счет учета взаимного положения слов в документах. Для этого к весу документа добавляется дополнительная составляющая, величина которой определяется взаимным положением слов в текстах документов. Эта величина определяется с помощью специальной функции взвешивания, которая учитывает координаты слов в документе.

При выборе этой функции мы использовали следующие эвристические предположения:

1. Вес документа тем выше, чем выше вес входящих в него фрагментов, содержащих слова запроса. Так как практически во всех функциях взвешивания, отдельных терминов, вес документа представляет собой сумму весов терминов, то здесь используется такое же предположение.
2. Вес каждого фрагмента тем выше, чем больше слов запроса он содержит.
3. Чем ближе в фрагменте располагаются слова запроса, тем больший вес он имеет. Данное предположение является следствием лингвистического наблюдения о том, что связанные между собой слова в тексте располагаются ближе друг к другу.
4. Для того, чтобы алгоритм взвешивания выше оценивал устойчивые словосочетания, фрагменты, которые чаще

встречаются в коллекции, должны получать больший вес.

Для выделения фрагментов используется алгоритм, основанный на «скользящем окне» (sliding window) длиной N. Где N – некоторая константа. В нашей системе мы выбирали ее значение, линейно зависящее от числа слов в запросе. При этом из документа последовательно выделяются фрагменты с 1 по N слово, затем со 2 по N+1, и так пока не будет достигнут конец документа. Каждый выделенный фрагмент обрабатывается по следующему алгоритму:

1. Если фрагмент содержит менее двух слов запроса, то система переходит к следующему фрагменту.
2. Если фрагмент содержит два и более слова запроса, то для данного фрагмента вычисляется вес. В системе реализовано 3 различных функции вычисления веса фрагмента:
    a. Константа. Фрагмент получает вес Wcnst вне зависимости от расстояния между словами.
    b. Линейная функция. Фрагмент получает вес, вычисленный по формуле:
    $W = W_{max} - k*l_{av}$, где
    l_{av} – среднее расстояние между словами запроса в фрагменте,
    k, W_{max} – некоторые константы.
    Линейная функция позволяет учесть то, что фрагменты, в которых слова ближе друг к другу имеют больший вес.
    c. Квадратичная функция. Фрагмент получает вес, вычисленный по формуле:
    $W = W_{max} - k*l_{av}^2$, где
    l_{av} – среднее расстояние между словами запроса в фрагменте,
    k, W_{max} – некоторые константы.
    Данная функция быстрее убывает с ростом расстояния, чем линейная.
3. Просматривается массив ранее отобранных фрагментов. Система хранит массив ранее отобранных фрагментов, при этом каждый элемент массива уникально идентифицируется по тем словам запроса, которые встретились в данном фрагменте. В каждом элементе хранится список идентификаторов документов. Кроме идентификатора, в каждом элементе списка хранится сумма весов фрагментов для данного документа.

После просмотра всех документов анализируется массив отобранных фрагментов. При этом для каждого элемента массива обрабатывается список идентификаторов документов. Вес каждого фрагмента умножается на логарифм длины списка. Тем самым обеспечивается больший вес устойчивых словосочетаний. Полученное значение прибавляется к весу документа с данным идентификатором.

При реализации данного алгоритма, для ускорения обработки запросов, система не просматривает документы, а использует инвертированный файл, в котором сохранены позиции слов внутри документов. При этом алгоритм вычисления веса сканирует не текст документа, а пост листы вхождения слов запроса.

Отличия тестовой версии системы от коммерческой

Задания РОМИП были представлены в специальном формате, отличающимся от обычно используемого нами. При этом представлении документ содержит только текст с элементами форматирования и гиперссылками. Никаких других атрибутов документ не имеет. Поэтому, при выполнении заданий РОМИП был внесен ряд упрощений в алгоритм поиска по сравнению с коммерческой версией:

1. В наших базах данных имеется рубрикатор документов, который использовался в задаче классификации РОМИП. Коммерческая версия системы производит поиск не только по документам, но и по тематическому рубрикатору – по названию рубрик и имеющимся у них комментариям. Далее система увеличивает вес документов, которые имеют отобранные поиском рубрики. Т.к. коллекция РОМИП не имела размеченных рубрик, и для рубрик не передавались комментарии, то данная функция была отключена.
2. Система имеет достаточно сложный алгоритм анализа пользовательского запроса. Используя набор стандартных шаблонов, из запросов выделяются атрибуты документов и специально обрабатываются при поиске. Например, если пользователь ввел запрос “Федеральный закон от 24.03.2001 N 33-ФЗ”, система выделит из запроса дату и номер, отбирая документы с соответствующими атрибутами. Так как другие участники РОМИП, очевидно, не имели таких специальных алгоритмов, ориентированных на

предметную область, мы не использовали их при выполнении заданий.

Выполение заданий РОМИП

Цели участия и выполненные задания

В этом году нами были представлены результаты пяти «прогонов»:

1. Коллекция web, прогон с функцией взвешивания не учитывающей взаимное положение слов.
2. Коллекция web, прогон с функцией взвешивания учитывающей взаимное положение слов с линейным взвешиванием фрагментов.
3. Коллекция legal, прогон с функцией взвешивания не учитывающей взаимное положение слов и не учитывающей гипертекстовые связи между документами.
4. Коллекция legal, прогон с функцией взвешивания учитывающей взаимное положение слов с линейным взвешиванием фрагментов и не учитывающая гипертекстовые связи между документами.
5. Коллекция legal, прогон с функцией взвешивания учитывающей взаимное положение слов с линейным взвешиванием фрагментов и учитывающая гипертекстовые связи между документами.

Представляя такой набор прогонов, мы хотели достигнуть следующих целей:

1. Определить как на качество поиска влияет учет взаимного положения слов и гиперссылок. Для web коллекции, по нашему мнению, проверять влияние гиперссылок не имело смысла, т.к. использовалась прошлогодняя коллекция, в которой большинство размеченных ссылок либо являются навигационными, либо ведут вне коллекции. В этом смысле коллекция правовых документов была намного более интересна, т.к. ссылки всегда указывали на документы коллекции.
2. Сравнить насколько совпадают оценки качества поиска по методикам РОМИП и по используемым нами внутренним методикам на одной коллекции и сходным запросам.

К сожалению, поставленные цели не были достигнуты. Получив результаты, мы были удивлены незначительным улучшением

результатов у алгоритмов, учитывающих взаимное положение слов. Анализ системы показал, что при построении индекса по тексту коллекций в формате РОМИП, для слова вместо номера позиции указывалось смещение слова от начала файла. Эта ошибка привела к тому, что алгоритм работал не совсем правильно и показал результаты значительно хуже ожидаемых.

Web коллекция

Данная коллекция является для нас не обычной, и здесь не ожидалось высоких результатов. Если сравнивать результаты нашей системы с результатами других систем, то мы опять выступили «середнячком».

Несмотря на ошибку в программном обеспечении, согласно результатам получилось, что учет взаимного положения слов улучшил точность поиска. Увеличение, для различных методик оценок релевантности, составило от 12 % до 227%. Это достаточно очевидный результат, т.к. замена положений слов на их смещения привело к тому, что система использовала «скользящее окно» очень маленькой длины. Для некоторых видов оценки релевантности полнота уменьшилась на 19%. Однако, в ряде случаев, наблюдается и увеличение полноты на 40%. Зависимость полнота/точность выглядит вполне разумной. Наибольший рост точности наблюдается у тех методик оценки, где полнота уменьшилась и наоборот.

Значительный разброс результатов в зависимости от методики определения релевантности, по нашему мнению, говорит о большой «зашумленности» коллекции и запросов, что делает крайне сложным определения уровня релевантности для асессора.

Коллекция правовых документов

Очевидно, что на «своей» коллекции мы надеялись достигнуть хороших результатов. Несмотря на ошибку реализации мы, если сравнить показания точности и полноты, достигли лучших показателей практически для всех методик оценки качества поиска.

Обнаружив ошибку в программном обеспечении, мы проверили, насколько данная ошибка ухудшает результаты поиска. По нашим данным, это должно привести к значительному ухудшению качества – до 50% по полноте на некоторых запросах. Как ни странно, по результатам РОМИП у нас получилось, что введение учета взаимного положения слов незначительно (на 5-10%) ухудшило точность и незначительно (2%) увеличило полноту. В качестве гипотезы, объясняющей этот результат, можно предположить, что улучшение полноты укладывается в погрешность метода измерения.

Введение учета гипертекстовых ссылок для наших внутренних методик оценки давало значительное улучшение точности – до 75%, при неизменной полноте. Однако, т.к. в данном прогоне использовалась функция взвешивания с учетом взаимного положения слов, которая имела ошибки реализации, то заметного улучшения точности мы не получили. Другой причиной расхождения результатов может быть то, что по нашей внутренней методике мы использовали очень маленькую глубину пула (10 документов) и очень высокие требования к определению релевантности. Поэтому изменения в начале списка документов, отсортированных по весу, давали значительные изменения в наших оценках, в то же время, скорее всего, слабо влияли на изменение оценок по методике РОМИП.

Так же мы успели провести к моменту публикации этой статьи анализ запросов, по которым наша система не вернула релевантных документов, согласно методике РОМИП при строгой оценке релевантности (relevant minus). Нас интересовали причины неудачи при обработке этих запросов. Вот результаты этого анализа:

Запрос «**лишение премии»**. Прежде всего, данный запрос плохо обрабатывается, т.к. подобное словосочетание является неформально-бытовым, и редко встречается в правовых документах. В коммерческой версии системы данный запрос отрабатывается заметно лучше, за счет использования системой рубрики «Трудовой распорядок», комментарий к которой содержит эти слова, возвращая такие документы, как «Трудовой кодекс».

Запрос «**Комментарий к закону об инвестиционной деятельности»**. Опять же запрос является «сложным» для системы, т.к. во-первых, в коллекции РОМИП не было комментариев к нормативным актам, во-вторых, такого закона в России нет. В данном случае опять же в коммерческой системе результат поиска лучше за счет использования рубрик, т.к. за счет рубрики «Инвестиционный фонд», в комментарии которой содержит много слов из запроса, в результат попадают законы «Об инвестиционных фондах» и «О рынке ценных бумаг».

Таким образом, заметное улучшение качества поиска можно было получить за счет использование информации о дополнительных атрибутах документа.

Выводы и планы

К сожалению, досадная ошибка в программном обеспечении не позволила нам полностью выполнить намеченные цели. Однако

полученные материалы нам позволят провести эту оценку в дальнейшем.

Большое различие результатов, полученных по нашим методикам и по методикам РОМИП, говорит о том, что нам требуются дополнительные проверки наших методик. Мы планируем провести их в течении следующего года, пересчитав наши метрики согласно методикам РОМИП.

Предложения по проведению семинара в 2005 году.

Организация семинара

По нашему мнению, организация семинара в этом году была на достаточно высоком уровне. Трудно высказать какие-либо предложения, кроме, возможно, более оперативного освещения хода семинара на сайте http://romip.narod.ru. Замечательно, что в этом году было разослан подробный документ с описанием методик, жаль только что эта рассылка была произведена очень поздно.

Предложения по новым дорожкам.

Интерес к новым дорожкам, конечно, определяется теми практическими проблемами, которых нам приходится решать. В настоящий момент для нас интересны следующие задания:

1. Ad hoc дорожка по очень большой коллекции. Имеется ввиду классический поиск на больших (терабайты) коллекциях документов. При этом нас интересует, насколько масштабирование влияет на качество поиска. Практически это может быть значительно расширенная Web коллекция, с тем же набором запросов, что и обычная web дорожка.
2. Поиск с выдачей фрагментов документов (passage retrieval). Данная дорожка является вариантом поиска по legal коллекции со стандартным набором запросов, где в качестве результатов возвращаются фиксированное количество фрагментов текстов документов коллекции заданной длины. Далее для этих фрагментов оцениваются точность и полнота по методикам РОМИП. Данная дорожка важна, т.к. в настоящее время, при постоянном росте размеров документов в наших базах данных, пользователи уже не удовлетворены выдачей в качестве результатов ссылки на документ, требуется более точная «детализация» результатов поиска.

Литература

[1] Proceedings of the TREC-5 conference, National Institute of Standards and Technology, Washington DC, USA, November 1996.
[2] Максим Губин. Исследование качества информационного поиска с использованием пар слов. In *RCDL 2003*, p/ 186-191, 2004

The Kodeks Information System at RIRES 2004

Maxim Gubin

We present the Kodeks Information System evaluation at RIRES 2004. The main aim of our participation was to test our new algorithms and to compare RIRES and our internal evaluation methods. The article describes algorithms used by the system and analyzes achieved results.