Участие ИПС «Кодекс» в семинаре РОМИП 2005

© Губин М.В.

ИК «Кодекс»
max@gubin.spb.ru

Аннотация

Данная статья является отчетом об экспериментах, проведенных в рамках семинара РОМИП 2005. В этом семинаре мы уделили особое внимание поиску по смешанной коллекции и контекстно-зависимому аннотированию текста. Рассматриваются использованные подходы и анализируются основные полученные результаты.

1. Введение

ИК Кодекс участвует в РОМИП уже третий раз. Данная конференция позволяет провести независимое объективное исследование разрабатываемых нами подходов к анализу текстов.
В этом году мы участвовал и в следующих дорожках:

1. поиск по Веб-коллекции;
2. поиск по коллекции нормативно-правовых актов;
3. поиск по смешанной коллекции;
4. контекстно-зависимое аннотирование документов.

В первых двух дорожках мы участвовали и в прошлых конференциях и в этом году. Мы не планировали продемонстрировать каких-то новых технологий и подходов. Поэтому в данной статье мы не рассматриваем участие в этих экспериментах, а опишем участие в дорожках поиска по смешанной коллекции и формированию контекстно-зависимых аннотаций документов.

2. Поиск по смешанной коллекции

Мы являлись одним из инициаторов данной дорожки. Мы хотели проанализировать изменение характеристик информационного поиска для случая, когда коллекция содержит разнородные документы. Подобная ситуация является достаточно распространенной на практике, например, в случае библиотеки электронных книг это могут быть книги в различных форматах или написанные в разные периоды времени. Для информационно-правовой системы, это, например, нормативные документы, книги и учебники по праву, словари, энциклопедии и т.д. Для системы локального поиска - файлы, почтовые сообщения, сообщения интернет-пейджера. Для глобальной поисковой системы в Интернет - сайты, новостийные ленты, сообщения конференций, блоги и т.д

2.1 Использованные подходы

Нами были проведены два эксперимента с использованием двух подходов поиска по смешанной коллекции:

1. «Смешанный подход». При этом подходе смешанная коллекция рассматривались как один массив документов. Статистики терминов запросов рассчитывались для всей коллекции, после чего осуществлялся классический поиск с использованием алгоритма TF*IDF. При реализации данного подхода мы не строили общие индексы по объединенной коллекции, а при поиске производили объединение постлистов непосредственно при выполнении запроса.

2. «Раздельный подход». При данном подходе поиск по алгоритму TF*IDF производился отдельно для каждой коллекции (web-коллекции и коллекции нормативных документов), а результат поиска объединялся с использованием весов, присвоенных каждой из коллекций.

Вес для каждой из коллекций при «раздельном» подходе получался с помощью следующих характеристик коллекций:

1. Статистика распределения терминов запроса в коллекции. Вводится следующая величина для каждой из коллекций:

   $$F_{ri} = F_i / F \text{ , где}$$

   F - общая частота терминов в двух коллекциях;

   F_i - частота термина в коллекции i.

2. Количество документов, которые система вернула по запросу из каждой коллекции. Вводится следующая величина для каждой из коллекций:

$$R_{ri} = S_{ri}/S_i, \text{ где}$$

S_{ri} - количество документов, отобранных как релевантные из данной коллекции;

S_i - размер i-ой коллекции в документах.

При «раздельном подходе» использовался следующий алгоритм:

1. Производится поиск по каждой из коллекций;
2. Получается оценка веса, каждой из коллекций, на основании величин F_{ri}, R_{ri}. Использовалась следующая простейшая формула, где Σ – сумма по всем терминам запроса:

$$W_i = R_{ri} \Sigma F_{ri}$$

3. Формируется объединенный список. При этом используется следующий алгоритм:
    - Формируется массив, который заполняется вычисленными на предыдущем этапе весами.
    - Выбирается элемент массива с наибольшим значением и в результирующий список перемещается первый документ из результата поиска соответствующего массива документов.
    - Ко всем элементам массива весов, кроме большего, прибавляется вес соответствующего массива документов.
    - Цикл отбора повторяется , пока не будут изъяты все документы из всех списков.

Более подробный анализ алгоритмов приведен в статье [1]

2.2 Экспериментальные результаты

Использование «раздельного» подхода к поиску позволило улучшить качество поиска, хотя и не столь значительно, как мы предполагали. Стандартный 11 точечный график релевантность-полнота для случая сильных требований релевантности приведен на рисункеРисунок 1.

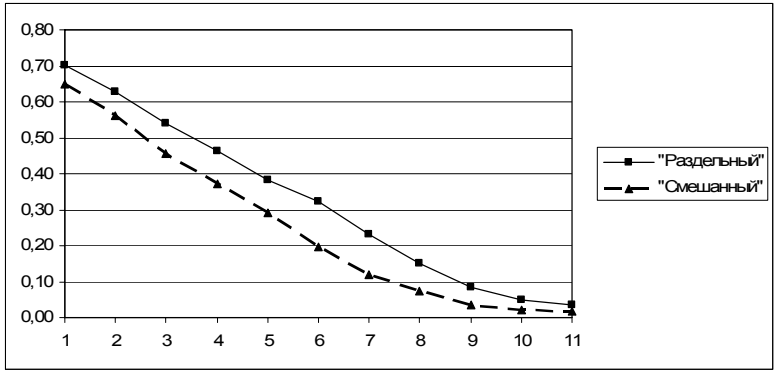

Рисунок 1. 11 точечный график для mixed-adhoc

Мы считаем, что результаты с использованием «раздельного подхода» можно было улучшить, если была проведена предварительная подгонка алгоритма с использованием подборки специального взвешивающего коэффициента, назначаемого для каждой из коллекций.

3. Контекстно-зависимое аннотирование

Формирование аннотаций для найденных документов является одной из стандартных функций современных поисковых систем. Короткая аннотация, выведенная в результате поиска, должна позволять пользователю значительно быстрее оценивать релевантность документа запросу по сравнению с чтением текста документа. Именно поэтому мы предлагали постановку эксперимента, когда сравниваются оценки релевантности, выставленные асессором для документов и для их аннотаций. Подобный подход используется в ряде экспериментов, проводимых в рамках конференции SUMMAC[3].

3.1 Использованные подходы

Более подробно описание использованных подходов и наши внутренние эксперименты по контекстно-зависимому аннотированию описаны в статье [2].

В РОМИП2005 мы экспериментировали с 3 алгоритмами формирования аннотаций:

1. Алгоритм, который анализирует в документе только слова запроса и отбирает фрагмент с их наибольшей плотностью. В дальнейшем этот алгоритм мы будем называть базовым.

2. Алгоритм, где кроме слов запроса при отборе фрагмента учитывались и другие слова, имеющие высокую частоту встречаемости в тексте в некотором окружении вокруг фрагмента. Этот алгоритм в статье будет называться Freq.

3. Алгоритм, где кроме слов запроса при отборе фрагмента использовались слова, отобранные с помощью алгоритма LRU-K, учитывающего повторяемость слов в тексте. Этот подход к формированию аннотаций мы будем обозначать LRU-K.

3.2 Экспериментальные результаты

Первоначально, получив экспериментальные результаты, мы были несколько удивлены близость полученных значений. Мы ожидали, что значения будут близкими, так как подобное наблюдается и в результатах SUMMAC [3], однако в случае РОМИП расхождения составили тысячные. Результаты для строгой релевантности (and-and) и слабой релевантности (or or) приведены в таблице 1.

	базовый	Freq	LRU-K
And/And	0,789832	0,788395	0,788395
Or/Or	0.893758	0.890610	0.890610

Таблица 1. Точность для двух вариантов оценки релевантности

Различия в значениях для трех использованных подходов можно считать пренебрежимо малым, то есть его не удалось выявить в этих экспериментах.

По нашему мнению, столь близкие полученные результаты можно объяснить следующим:

В отличие от наших внутренних экспериментов, где основное внимание было уделено аннотированию больших документов (сотни килобайт), в экспериментах РОМИП большинство аннотированных документов были достаточно короткими (единицы килобайт), и любая достаточно длинная фраза документа оказалась достаточной для определения его релевантности асессором.

Несовпадение результатов РОМИП с нашими внутренними экспериментами можно так же объяснить тем, что в постановке задачи асессору в наших внутренних экспериментах ставилась несколько иная задача – оценить, полезна ли ему аннотация, с точки зрения

запроса, что не совсем совпадает с постановкой эксперимента в РОМИП.

Значения оценок других систем отличаются от полученных нами так же на сотые, то есть экспериментами не были выявлены различия. Получив в результатах аннотации, сформированные другими системами, мы обратили внимание, что они очень похожи и часто представляют собой одни и те же фрагменты документов, иногда отличающиеся сдвигом на несколько слов, что объясняет столь малые отличия.

4. Выводы и планы

Эксперименты РОМИП позволили объективно проверить наши идеи с помощью методик, которые приняты не только нами, но и другими группами.

«Раздельный» подход к поиску по смешанной коллекции показал слегка лучшие результаты, но не столь хорошие как мы ожидали. Очевидно, что эта идея требует дополнительных доработок.

Эксперименты по контекстно-зависимому аннотированию показали, что, по крайней мере, наша реализация соответствует уровню других участников.

5. Предложения по проведению семинара в 2005 году

В этом году организация семинара была практически идеальной. Однако, из опыта прошедших и этого семинара, для нас основной проблемой является повторное использование результатов и методик РОМИП для работы вне семинара. Если форматы коллекций и результатов представлены в XML подобном формате, который достаточно удобно обрабатывать, то результаты поступают в виде текстовых файлов, которые не всегда удобны для обработки. Желательно, чтобы результаты также получались в XML подобном формате. Кроме того, идеально было бы получить утилиты для вычисления характеристик экспериментов, чтобы не «изобретать велосипед», а использовать готовые проверенные утилиты расчета.

Литература

[1] Губин М.В. Информационный поиск в коллекции разнородных документов. Материалы *RCDL2005,* 2005

[2] Губин М.В., Меркулов А.И. Эффективный алгоритм формирования контекстно-зависимых аннотаций документов при поиске. В *Материалы Диалог 2005*, стр. 116-120, 2005

[3] Inderjeet Mani and others. SUMMAC: A Text Summarization Evaluation**.** *Natural Language Engineering*, Vol. 8, No. 1, pp. 43–68, 2002

The Kodek Information System at ROMIP 2005

Maxim Gubin

We present the Kodeks Information System evaluation at ROMIP2005. This year we focused our efforts on mixed ad-hoc and summarization track. The article describes used methods and analysis of experimental results.