Mail.Ru на РОМИП-2006

© Костин Михаил, Проскурин Андрей,

Федоровский Андрей

Mail.Ru
kostin@corp.mail.ru, proskurin@corp.mail.ru,
fedorovsky@corp.mail.ru

Аннотация

Статья посвящена участию компании Mail.Ru в семинаре РОМИП-2006. Особое внимание уделено алгоритмам, использованным для решения задачи контекстно-зависимого аннотирования.

1. Введение

При участии в семинаре во второй раз, нам, как и в прошлом году, было очень интересно и полезно получить объективную независимую оценку качества наших разработок в области поисковых технологий.

В данной статье рассказывается о результатах нашего участия в РОМИП-2006 и описываются алгоритмы, использованные нами для решения задачи контекстно-зависимого аннотирования.

2. Дорожка контекстно-зависимого аннотирования

По условиям этой дорожки, в которой мы участвовали впервые, участники должны были предоставить аннотацию размером не более 300 символов по каждой из предложенных пар запрос-документ.

Данная задача очень близка к задаче построения аннотаций документов (сниппетов) в выдаче поисковой системы. В качестве единственного существенного отличия можно назвать строго заданную длину аннотации (предоставлять аннотацию меньшего размера не воспрещалось, однако это не имело особого смысла по условиям задачи). При построении аннотации в поисковой системе приходится решать еще и задачу поиска оптимального баланса между ее ин-

формативностью и размером (количеством строк), занимаемым ею на экране, а также делать выбор между стабильной длиной аннотации, предпочтительной с точки зрения дизайна, и стремлением к оптимальному соотношению информативности и размера аннотации для каждого конкретного документы. Поэтому, при выполнении задания по этой дорожке нами использовалась функциональность формирования сниппетов нашей поисковой системы, лишь незначительно модифицированная для строгого соблюдения ограничения по длине аннотации.

2.1 Принципы составления аннотации

При составлении аннотации мы придерживаемся следующих основных принципов:

- Включение в аннотацию фрагмента текста, релевантного запросу в целом, если таковой содержится в документе
- Включение в аннотацию контекста всех слов поискового запроса
- Максимальная информативность контекста для наиболее значимых слов запроса
- Включение в аннотацию, по возможности, законченных по смыслу фрагментов текста, за счет учета знаков препинания и html-форматирования документа

Как особый случай рассматривалась ситуация, когда слова запроса отсутствуют в теле документа (присутствуют только в заголовке). В этом случае в качестве аннотации выдавалось начало текста документа.

Рассмотрим перечисленные принципы подробнее.

2.2 Релевантные запросу в целом фрагменты

Наиболее полное представление о документе в контексте запроса дает фрагмент документа, релевантный запросу в целом. Если таковой имеется, то он обязательно включается нами в аннотацию.

Для поиска релевантных всему запросу фрагментов документа нами используется алгоритм поиска релевантных пассажей, реализованный в модуле ранжирования нашей поисковой системы [1]. Релевантным пассажем там считается фрагмент текста, не превышающий заданного ограничения по длине, и содержащий представительное множество слов запроса. Представительное множество зависит от количества слов в запросе и их частотности в коллекции, для достаточно коротких (два-три слова) запросов это, как правило,

все слова запроса за исключением стоп-слов. Каждому пассажу присваивается вес, величина которого зависит от целого ряда параметров, главными из которых являются длина пассажа, соответствие порядка слов в запросе и пассаже и совпадение грамматических форм слов запроса и пассажа. По этому весу производится отбор одного или нескольких лучших пассажей для аннотации.

Найденный пассаж включается в аннотацию вместе с его окружением, включающим заданное количество слов с обеих сторон. В зависимости от длины пассажей в аннотацию может быть включен один или несколько таких фрагментов. Для дорожки контекстно-зависимого аннотирования настройки были выбраны таким образом, что в аннотацию включались один или два фрагмента с пассажами.

2.3 Контекст всех слов запроса

С нашей точки зрения, для информативности аннотации очень важно показать в ней контекст всех слов запроса, за исключением стоп-слов. Если количество слов в запросе велико, то наименее значимые слова могут быть опущены, но и в этом случае мы стремимся максимизировать количество слов запроса встречающихся в аннотации.

2.4 Зависимость контекста от значимости слов

Под значимостью слов мы понимаем в данном случае обратную частотность терма в коллекции (inverted document frequency – IDF), вычисленную по стандартной логарифмической формуле. Наиболее значимые слова с большей вероятностью являются ключевыми в запросе пользователя и, соответственно, именно их ему наиболее важно увидеть в максимально информативном контексте. В качестве параметров, определяющих информативность контекстного фрагмента для определенного слова, мы рассматриваем его длину (в словах) и наличие в нем других слов запроса.

В то же время, мы избегаем и слишком короткого контекста для менее значимых слов, с тем чтобы информативность цитаты для таких слов оставалась на разумном уровне. Для этого мы используем ограничение на минимальную длину контекстного фрагмента.

Для выбора величины контекстного фрагмента в качестве базовой нами используется следующая формула:

$$L_{term} = L_{\min} + \frac{L_a - N_q \cdot L_{\min}}{L_{\min}} \cdot IDF_{term} \quad (1), \text{ где}$$

L_{term}– длина контекстного фрагмента (слов)

L_{min} – минимальная длина контекстного фрагмента (слов)

L_a – оптимальная длина аннотации (слов)

N_q – количество термов в запросе, за вычетом стоп-слов

IDF_{term} – обратная частотность терма в коллекции

При выборе контекстного фрагмента для каждого из слов, предпочтение отдается фрагментам, включающим в себя другие слова запроса. Размеры фрагментов в этом случае вычисляются несколько более сложным, чем только что описанный, образом, но принцип зависимости величины контекста от значимости слов соблюдается и в этом случае.

2.5 Выравнивание фрагментов аннотации

Для повышения читаемости и информативности аннотации мы корректируем ее фрагменты с учетом знаков препинания и html форматирования документа, стараясь включать в нее, по возможности, целые предложения или их законченные части.

В то же время, мы рассматриваем это требование как менее важное, чем перечисленные выше, и применяем его не в ущерб их соблюдению.

2.6 Результаты

Методика оценки результатов для дорожки контекстного аннотирования была основана на проверке соответствия оценок релевантности документа и аннотации. По результатам этих оценок вычислялись метрики *Accuracy* и *Error*. Первая из них соответствует доле аннотаций, оцененных как релевантные, которым соответствуют релевантные документы; второй – доле аннотаций, оцененных как нерелевантные, которым соответствуют релевантные документы.

В таблице приведены результаты нашей системы для различных уровней релевантности и способов оценки.

Уровень релевантности	Relevant -		Vital	
Способ оценки	OR-OR	AND-AND	OR-OR	AND-AND
Accuracy	0.829	0.795	0.878	0.922
Error	0.378	0.409	0.545	0.503

Таблица 1. Оценки по дорожке контекстно-зависимого аннотирования.

К сожалению, из–за недостатка времени мы не смогли пока достаточно тщательно проанализировать результаты данной дорожки. Однако, сразу обратил на себя внимание относительно высокий процент ошибок вида «релевантная аннотация - нерелевантный документ» у нашей системы. Возможно, это явилось следствием нашего принципа включения в аннотацию контекста всех слов запроса, встретившихся в документе. Можно предположить, что соседство в аннотации фрагментов, включающих в себя разные слова запроса, создавало в некоторых случаях иллюзию взаимной близости этих фрагментов в документе. Однако, это требует дополнительного исследования.

3. Поисковые дорожки

В этот раз, как и год назад, мы принимали участие в трех поисковых дорожках – по веб-коллекции, по коллекции нормативно-правовых документов и по смешанной коллекции.

Используемые нами алгоритмы были достаточно подробно описаны в прошлогоднем докладе [1]. На этот раз принципиальных изменений в них внесено не было.

Заслуживают того, чтобы быть отмеченными два дополнительных фактора, которые мы учитывали в поиске по веб-коллекции: тексты ссылок на страницу (то есть ссылочное ранжирование) и релевантность запросу сайта в целом.

Польза от ссылочного ранжирования для данной коллекции оказалась, как мы и предполагали, самой минимальной. Это объясняется, в первую очередь, относительно небольшим объемом коллекции, а также ее незамкнутым характером, следствием которого является почти полное отсутствие ссылок между составляющими коллекцию сайтами. Также играет здесь свою роль и почти исключительно информационный характер запросов, входящих в тестовое множество.

Учет релевантности всего сайта также не сказался существенным образом на полученных результатах. Основной причиной тут является то, что этот метод эффективен, прежде всего, при показе результатов поиска в сгруппированном по сайтам виде. В данном случае, условиями задачи подобная группировка не предусматривалось. Кроме того, как и в предыдущем случае, не способствовал применимости этого метода достаточно узкий информационный характер большинства запросов.

3.1 Поиск по веб-коллекции

Задача дорожки поиска по веб-коллекции представляла из себя стандартную задачу ad hoc поиска. Подробное описание используемых наборов данных и методики оценки результатов можно найти в материалах прошлогодних семинаров [2,3].

Хорошо заметной особенностью результатов дорожки веб-поиска оказалось в этом году лидерство разных систем по разным метрикам. Аналогичная картина наблюдается и в результатах поиска по нормативно-правовым документам. Множества характеристик, по которым наблюдается преимущество одной из систем, хорошо иллюстрируют исследованную нами в прошлом году закономерность – метрики, используемые в поисковых дорожках РОМИП, можно разделить на две основные группы, внутри которых наблюдается значительная взаимная корреляция (это исследование не вошло в печатную версию нашего доклада, однако доступно на слайдах на сайте РОМИП [4]). В одну группу входят метрики, характеризующие релевантность на всем объеме поисковой выдачи, наиболее показательной из которых, на наш взгляд, является *Average Precision*, в другую метрики, характеризующие релевантность первой страницы результатов поиска, наиболее важной из которых логично считать метрику *P(10)*. Таким образом, можно сделать вывод, что эта особенность связана с различиями в использованных участниками алгоритмах ранжирования, а не со случайными флуктуациями, вызванными особенностями набора запросов или решениями асессоров.

Ниже приведены результаты участников дорожки веб-поиска для способов оценки AND (сильные требования к релевантности) и OR (слабые требования к релевантности).

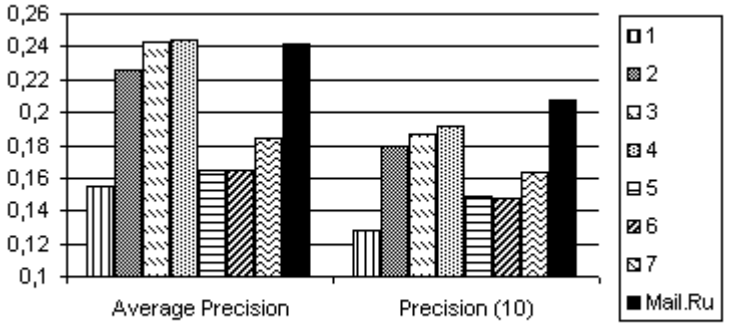

Рисунок 1. Результаты участников дорожки веб-поиска. AND, pd50.

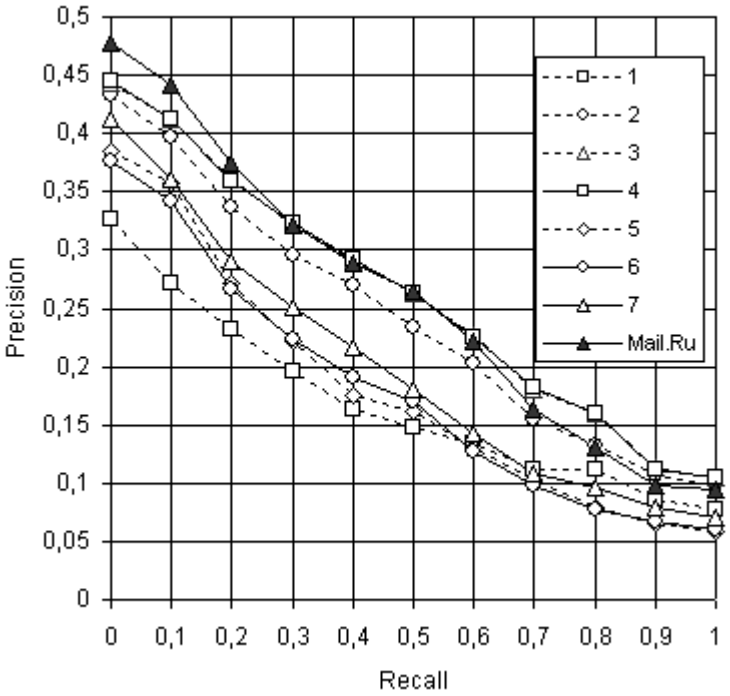

Рисунок 2. 11- точечные графики TREC. Веб-поиск, AND, pd50

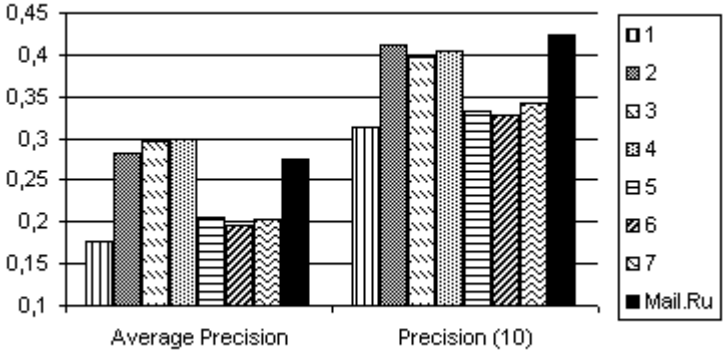

Рисунок 3. Результаты участников дорожки веб-поиска. OR, pd50

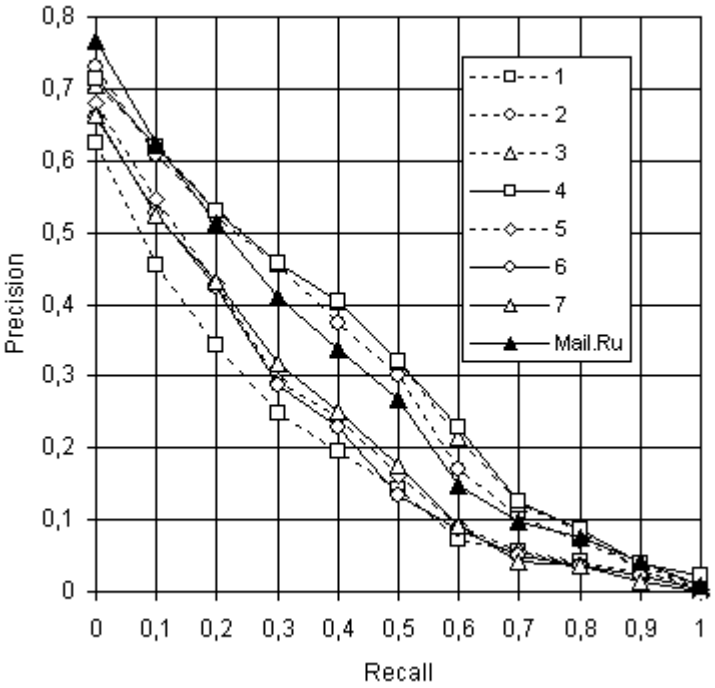

Рисунок 4. 11-точечные графики TREC. Веб-поиск, OR, pd50.

3.2 Поиск по коллекции нормативно-правовых документов

Хотя количество участников данной дорожки было в этом году очень небольшим, в ее результатах можно увидеть те же характерные особенности, что и в результатах дорожки веб-поиска.

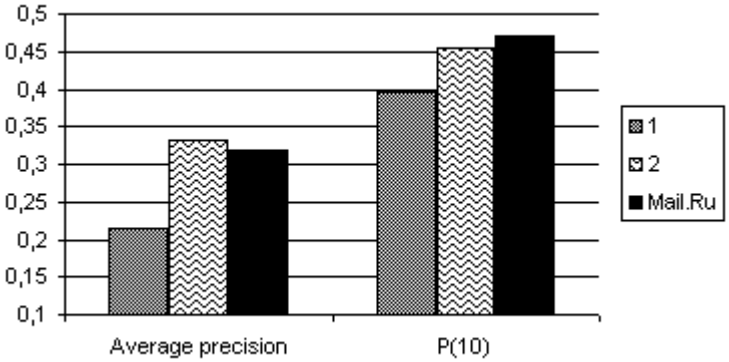

Таблица 2. Результаты участников дорожки поиска по нормативно-правовым документам.

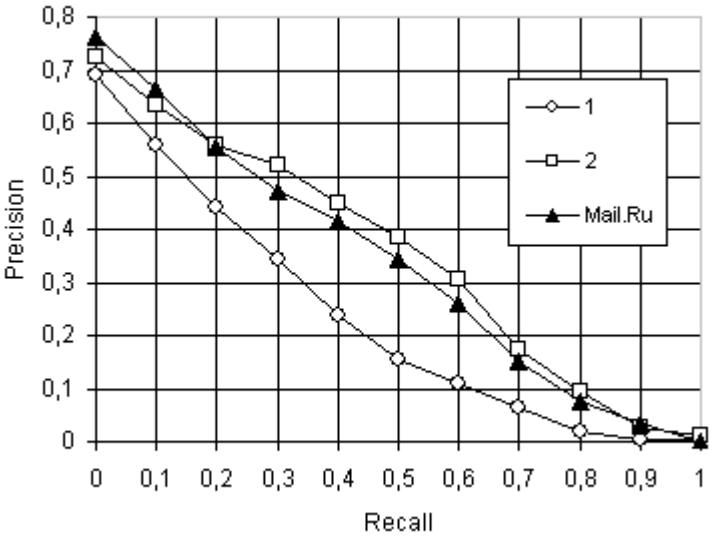

Рисунок 5. 11-точечные график TREC участников дорожки поиска по коллекции нормативно-правовых документов.

3.3 Поиск по смешанной коллекции

На гистограмме приведены результаты дорожки поиска по смешанной коллекции.

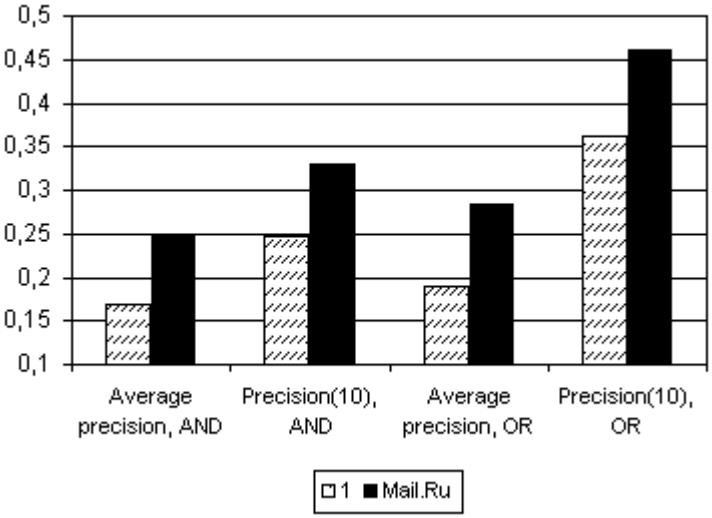

Рисунок 6. Результаты участников дорожки поиска по смешанной коллекции

Более полные результаты этой и других поисковых дорожек могут быть предоставлены по запросу.

4. Заключение

По предварительным итогам анализа, хочется порадоваться прогрессу результатов по поисковым дорожкам: наш сильный прошлогодний отрыв сменился в этом году паритетом. Хочется верить, что это только подхлестнет участников семинара и приведет к еще большему улучшению качества результатов. Надеемся, что алгоритмы, опубликованные нами в прошлом году, также помогут в этом исследователям.

Дорожка контекстно-зависимого аннотирования, надеемся, будет развиваться и привлекать дополнительных участников. Результаты, полученные на ней, пока спорны и требуют дальнейшего изучения и анализа.

Литература

[1] А.Федоровский, М. Костин, А. Проскурин. Mail.Ru на РОМИП-2005. Труды третьего российского семинара по оценке методов информационного поиска. Под ред. И.С. Некрестьянова - Санкт-Петербург: НИИ Химии СПбГУ, 2005.

[2] Труды второго российского семинара по оценке методов информационного поиска. Под ред. И.С. Некрестьянова - Санкт-Петербург: НИИ Химии СПбГУ, 2004

[3] Труды третьего российского семинара по оценке методов информационного поиска. Под ред. И.С. Некрестьянова - Санкт-Петербург: НИИ Химии СПбГУ, 2005.

[4] Mail.Ru на РОМИП-2005. Слайды. http://romip.narod.ru/romip2005/slides/mailru.pdf

Mail.Ru at RIRES-2006

A. Fedorovsky, M. Kostin, A. Proskurin

Mail.Ru

fedorovsky@corp.mail.ru, kostin@corp.mail.ru, proskurin@corp.mail.ru

The paper presents information retrieval system Search@Mail.Ru developed by Mail.Ru company at RIRES-2006. We participated in ad hoc tracks on various collections and in summarization track. The article describes used methods and experimental results.