К вопросу об эффективности метода «общего котла»[*]

© И. Некрестьянов, М. Некрестьянова, А. Нозик

Санкт-Петербургский Государственный университет
http://ir.apmath.spbu.ru
{igor, marina}@meta.math.spbu.ru, blake@mail.ru

Аннотация

В работе рассматриваются вопросы, связанные с методологией оценки систем поиска методом «общего котла» на основе опыта использования этого подхода в рамках семинара РОМИП в 2003 и 2004 годах.

1 Введение

Эксперименты по оценке систем информационного поиска, проводимые в рамках таких инициатив как TREC, CLEF, NTCIR и с недавних пор РОМИП [1, 3, 8], заметно стимулируют развитие методов решения задач поиска. Методологическая основа этих экспериментов была заложена при работе над проектом Cranfield-2 в середине 1970-х годов и получила название “лабораторной парадигмы оценки” [7, 16].

Оригинальная версия этой парадигмы оценки основана на следующих предположениях [16]:

- Релевантность можно аппроксимировать тематической схожестью.
- Набор оценок асессора репрезентативно представляет пользователя ИПС.
- Для каждого задания известны все релевантные документы.

Каждое из этих предположений имеет ряд важных последствий. Например, из первого предположения следует, что все релевантные документы одинаково интересны пользователю, что релевантность одного документа не зависит от других и что информационная потребность пользователя не изменяется во времени.

Оригинальный вариант подразумевает, что асессоры производят оценку всей коллекции. Поскольку в то время считалось [7], что размер коллекции не играет значительной роли, а наиболее важным фактором представлялась доступность для всех документов детальной информации об оценке. Однако, рост объемов данных, с которыми приходится работать поисковым системам, обусловил необходимость проведения более масштабных экспериментов.

Для коллекций, состоящих из миллионов документов, разметка всей коллекции асессорами – задача практически невыполнимая, и третье предположение требует корректировки. Для решения этой проблемы в TREC используется метод “общего котла”, при котором оцениваются лишь первые несколько десятков документов, возвращенных каждой из систем. Считается, что такой «котел» позволяет достаточно хорошо аппроксимировать выводы, полученные по полной коллекции [21].

Конечно же, в общем случае предположения, лежащие в основе «лабораторной парадигмы», не верны, и использование такой упрощенной картины делает процесс оценки зашумленным. Отметим, что «лабораторная парадигма» - это не единственный подход к оценке систем информационного поиска [2], хотя и наиболее широко распространенный. Альтернативными подходами являются, например, различные методы аналитического сравнения систем информационного поиска или методы оценки, ориентированные на оценку удовлетворенности конкретного пользователя, рассматривающие поиск как интерактивный процесс [2, 10, 12].

Вне зависимости от используемого подхода к оценке, основной целью сравнения является получение ответа на вопрос “Какая из систем A или B лучше решает данную поисковую задачу?”. Более формально, вопрос можно переформулировать так: “Какая из систем A или B лучше справляется с удовлетворением информационных потребностей пользователей, сформулированных данным образом?” (например, в виде запросов к поисковой системе).

Ключевой методологической проблемой при проведении экспериментальной оценки является вопрос о степени достоверности сделанных выводов. Поскольку, при использовании «лабораторного подхода» сравнение различных поисковых систем производится в одинаковых условиях – при решении одинаковой задачи на одинаковой коллекции - и результаты разных систем оцениваются одним человеком, то на первый взгляд кажется, что нет повода сомневаться в достоверности полученных выводов.

Однако это не так. Будет ли вывод таким же при других параметрах проведения эксперимента? Ка-

кой должна быть разница оценок, чтобы можно было бы с уверенностью сделать вывод о превосходстве одной из них? Изменилось бы что-нибудь, если бы оценку производили другие люди? Насколько полученные выводы соответствуют реальной ситуации, т.е. при практическом использовании этих же систем в реальных условиях?

Например, известно, что проведенный в 2001 году эксперимент по сравнению передовых методов поиска согласно экспериментам TREC и коммерческих ИПС для поиска Веб показал, что при решении типичной для Веб задачи поиска конкретных страниц (домашних страниц или сайтов компаний) методы TREC оказываются значительно менее эффективными [14].

В течение последних пяти лет интерес к теме изучения границ применимости «лабораторной парадигмы» вообще и метода «общего котла» в частности резко возрос [6, 16, 19]. Во многом это обусловлено возможностью использовать материалы прошедших семинаров TREC для систематизации, обобщения и анализа применяемой в TREC методологии оценки [4, 15, 17, 21].

Благодаря проведению семинара РОМИП у нас появилась возможность исследования этих вопросов на альтернативных данных. При этом нам интересно не только исследовать новые методологические вопросы, но и проверить справедливость уже опубликованных результатов. Отметим, что проверка на альтернативном TREC материале является важной задачей, ведь возможность прямого переноса выводов методологических исследований на другие эксперименты по оценке не очевидна.

В частности, нас интересуют следующие вопросы:

- **Насколько эффективен метод «общего котла»?**
  Действительно ли сокращается объем работы по оценке? Является ли найденное множество релевантных документов достаточно хорошей аппроксимацией множества всех релевантных документов в коллекции?

- **Каковы должны быть параметры эксперимента для получения стабильных выводов?**
  Сколько должно быть запросов? Какова должна быть глубина котла?

- **Насколько влияет на результат «человеческий фактор»?**
  Что изменилось бы, если бы документы оценивали другие люди? Насколько поведение асессоров похоже на поведение реальных пользователей поисковых систем?

- **До какой степени можно использовать накопленную информацию для оценки систем, которые не участвовали в исходной оценке (при формировании «котлов»)?**
  Какие метрики и при каких условиях позволяют сделать относительно надежные выводы?

В данной работе представлены предварительные результаты некоторых наших исследований на основе материалов семинаров РОМИП в 2003 и 2004 году [3].

2 Лабораторная парадигма

Основным принципом «лабораторной парадигмы» оценки является сравнение различных поисковых систем в *одинаковых* (контролируемых) условиях. В этом разделе мы вкратце опишем основные принципы метода «общего котла», который на данный момент является наиболее популярным вариантом применения этой парадигмы на практике, а также представим семинар РОМИП, материалы которого используются нами далее для проведения оценки.

2.1 Метод «общего котла»

Формально, «общий котел» (pooling) - это объединенное множество первых N_q документов из выдачи каждой из систем для данного запроса q (параметр N_q называется глубиной пула) [2]. Такой «котел» строится для каждого из оцениваемых заданий, и все документы из этого котла в дальнейшем оцениваются асессором, т.е. человеком, который решает, релевантен или не релевантен данный документ исходной информационной потребности.

Отметим, что асессор оценивает документы, не зная, какой системой они были возвращены, т.е. в случайном порядке. Тем самым гарантируется непредвзятость оценки.

На основе оценок асессора строится таблица релевантности, содержащая информацию о том, какие документы были признаны релевантными, а какие нет. Используя эту таблицу для каждой из систем можно вычислить оценки ее эффективности. До тех пор, пока не требуется использование информации о документах за пределами глубины пула, вычисленные оценки не отличаются от тех, что были бы получены при оценке всех документов коллекции. Например, к этому классу метрик относится оценка точности на заданном уровне.

Поскольку полной оценки коллекции не производится, то точное число релевантных документов в коллекции узнать невозможно. В качестве его аппроксимации используется общее число релевантных документов в «котле». Такой подход позволяет получить аппроксимацию оценки полноты ответа.

Поскольку качество результата поиска во многом зависит от конкретного запроса, то вывод о превосходстве того или иного метода делается на основе усреднения по некоторому множеству запросов, представляющему популяцию всех возможных запросов. Отметим, что кроме усреднения абсолютных характеристик качества результата, можно также сравнивать эффективность методов на отдельных запросах и усреднять уже эту информацию.

2.2 Российский семинар по оценке методов информационного поиска (РОМИП)

Инициатива РОМИП (http://romip.narod.ru) состоит в регулярном проведении семинаров, каждый из которых посвящен сводной оценке качества русского текстового поиска и смежных технологий. Целью ее, кроме обмена опытом российских разработчиков, является создание и поддержание общедоступных «канонических» русскоязычных коллекций текстов, запросов и оценок, с помощью которых будущие исследователи смогут настраивать и развивать свои системы. Методология проведения семинаров РОМИП основывается на передовом зарубежном опыте подобных мероприятий TREC, CLEF и т.п. На данный момент успешно завершено два годовых цикла семинара РОМИП, и идет работа в рамках третьего.

В первом цикле РОМИП в 2003 году приняло участие 9 исследовательских коллективов, и рассматривались 2 поисковые задачи – классификация и поиск по Веб коллекции. Всего было получено 14 вариантов ответов. В РОМИП'2004 приняло участие уже 11 коллективов, и для 5 рассматриваемых задач в общей сложности было получено 34 варианта ответов. Отметим, что эти цифры, конечно, значительно меньше соответствующих характеристик TREC (где число ответов для одной задачи зачастую превышает 50), но в TREC нет русскоязычных коллекций, и на данный момент РОМИП – это наиболее крупномасштабный проект по оценке методов поиска на основе русскоязычных коллекций.

С точки зрения проведения исследований методологии оценки, важно, сколько альтернативных ответов было получено для каждой из оценивавшихся задач. Сводная информация о задачах, которые мы будем рассматривать далее, представлена в следующей таблице:

Название задачи	Число систем	Число вариантов ответов
РОМИП'2003		
Поиск по Веб-коллекции	5	9
Классификация Веб-сайтов	4	5
РОМИП'2004		
Поиск по Веб-коллекции	5	8
Классификация Веб-сайтов	3	6
Поиск по нормативной коллекции	4	10
Классификация нормативных документов	5	9

В отличие от TREC, в РОМИП каждый документ оценивается не менее, чем двумя независимыми асессорами. Поскольку их оценки субъективны, то они не всегда совпадают. Поэтому в РОМИП рассматривается две схемы объединения их в единую таблицу релевантности:

- *Сильные требования к релевантности (AND)* Документ считается релевантным, если все асессоры признали его релевантным.
- *Слабые требования к релевантности (OR)* Документ считается релевантным, если хотя бы один асессор признал его релевантным.

2.3 Что такое стабильный вывод?

Качество результата поиска зависит не только от используемого метода поиска, но также и от коллекции документов и заданий, на основе которых производится оценка. Полученные абсолютные характеристики качества результата имеют ограниченную ценность вне контекста конкретного эксперимента. Поэтому обычно основной целью проведения экспериментальной оценки является получение относительных результатов, т.е. результатов сравнения нескольких разных подходов к решению одной и той же задачи.

Целью экспериментальной оценки является получение вывода о том, какая из систем A или B лучше решает данную поисковую задачу (на заданной фиксированной коллекции). Формально, получить этот вывод можно путем вычисления абсолютных характеристик качества результата и их сравнения. Но насколько должны различаться абсолютные значения, чтобы можно было сделать вывод, что «лучше», а что «хуже»?

Даже при выполнении одних и тех же заданий можно выделить ряд параметров эксперимента, которые ограничивают выборку, на которой производится оценка, и, следовательно, могут влиять на получаемые результаты. Например, изменился ли бы вывод, если бы для оценки использовалось в 10 раз больше запросов?

Вывод «стабилен», если при изменении параметров эксперимента, которое *не уменьшает* выборку (по которой производится оценка) сам вывод остается неизменным. Отметим, что при этом абсолютные значения характеристик могут изменяться. Например, если возрастет общее число известных релевантных документов, то полнота ответов систем A и B снизится.

Из-за «зашумленности», обусловленной упрощенной моделью оценки, абсолютно надежный вывод получить нельзя. Можно лишь говорить о вероятности сделать неправильный вывод и выбирать параметры эксперимента, гарантирующие заданный уровень правдоподобности выводов.

3 Характеристики котлов

Теоретически, использование «общих котлов» выгодно, поскольку:

- Сокращается объем оценки по сравнению с независимой оценкой систем за счет удаления дубликатов. Причем чем больше систем участвует, тем больше удельная выгода.

- Строится хорошая аппроксимация множества релевантных документов.

Отметим, что, как объем оценки, так и качество аппроксимации зависят от числа систем N и глубины «котла» N_q. Безусловно, выбор конкретных запросов, как и алгоритмы, используемые в системах-участниках, также влияют на оценку «выгодности». Однако, имеющиеся в нашем распоряжении материалы не позволяют исследовать влияние этих параметров.

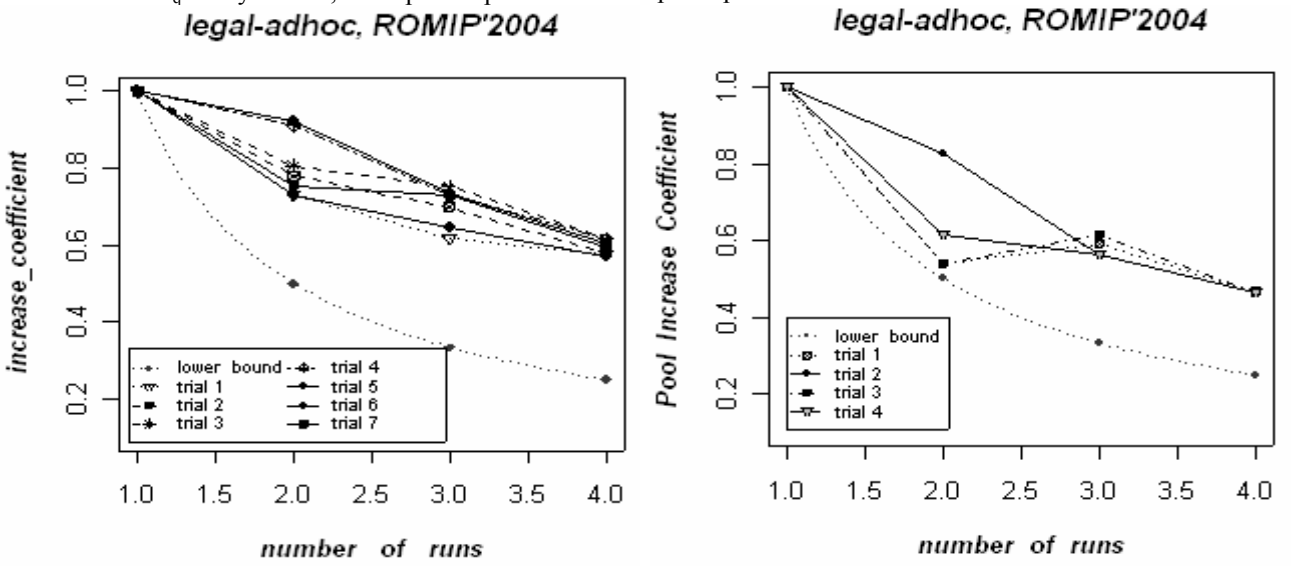

Рисунок 1. График зависимости коэффициента роста котла от числа учтенных прогонов. Слева – рассматриваются прогоны разных систем, справа – несколько прогонов одной системы. Для вычисления коэффициента роста общего котла использовалась следующая формула:

$$IncreaseCoefficient = \frac{\sum_q \text{размер котла для запроса q}}{\sum_q \sum_k \text{результатов системы k, включенных в котел для запроса q}}$$

Для того, чтобы проверить, насколько сокращается объем оценки, мы вычислили, насколько объем «котла» меньше, чем суммарное число документов, возвращенных системами. В частности, на рис. 1 представлены результаты для дорожки поиска по нормативным документам.

Если все системы возвращают уникальные документы, то коэффициент роста котла равен 1. С другой стороны, минимально возможное значение достигается, если ответы все время повторяются – *1/k*. Фактически, этот коэффициент показывает насколько меньше в среднем ресурсов потребуется на оценку одного варианта ответа по сравнению с его оценкой отдельно от других вариантов ответов.

Интуитивно ясно, что разные варианты ответа от одной и той же системы вероятно более схожи, чем ответы от разных систем. Эта гипотеза подтверждается результатами наших экспериментов на основе поисковых дорожек РОМИП'2004. На приведенных графиках в обоих случаях котел строился для 91 оценивавшегося в РОМИП'2004 запроса и коэффициент вычислялся после добавления одного ответа. Всего рассматривалось по 4 ответа - в первом случае ответы выбирались случайным образом среди вариантов, поданных разными системами, а во втором использовались 4 варианта одной и той же системы.

Очевидно, что порядок добавления систем сказывается на наблюдаемых значениях, поэтому мы рассматривали несколько случайных порядков. Разброс значений на графике слева демонстрирует влияние конкретных алгоритмов на абсолютные значения коэффициента. Тем не менее, эти графики наглядно иллюстрируют значительное сокращение затрат на оценку методом общего котла для *всех* участников с добавлением еще одного ответа.

Для оценки покрытия множества релевантных документов мы построили зависимость числа новых релевантных документов от глубины котла (рис. 2). Такая зависимость изучалась в работе [21] и ее авторы предположили, что она может быть описана формулой вида

$$N = CN_{depth}^{s} - 1,$$

где C и s - это некоторые константы, зависящие от числа прогонов и алгоритмов систем ($C = 382.5$, $s = -0.6182$ в [21]). Наблюдаемые нами результаты в принципе также хорошо аппроксимируются такой формулой, хотя в нашем случае коэффициенты другие. Например, для дорожки поиска по нормативной коллекции РОМИП'2004 – $C = 358.3798$, $s = -0.4145$ при использовании слабых требований к релевантности и $C = 205.451$, $s = -0.613$ при использовании сильных требований.

Наличие аналитической зависимости позволяет предсказывать полезность изменения глубины пула. Получается, что увеличение глубины пула в два раза с 50 до 100 привело бы к увеличению числа обнаруженных сильно релевантных документов в 1.35 раза, а слабо релевантных в 1.5.

К сожалению, для предсказания общего числа еще не обнаруженных документов построенные зависимости не очень пригодны. При использовании сильных требований к релевантности получается, что всего в коллекции 5920 таких документов (то есть при глубине пула 50 выявлено 22%).

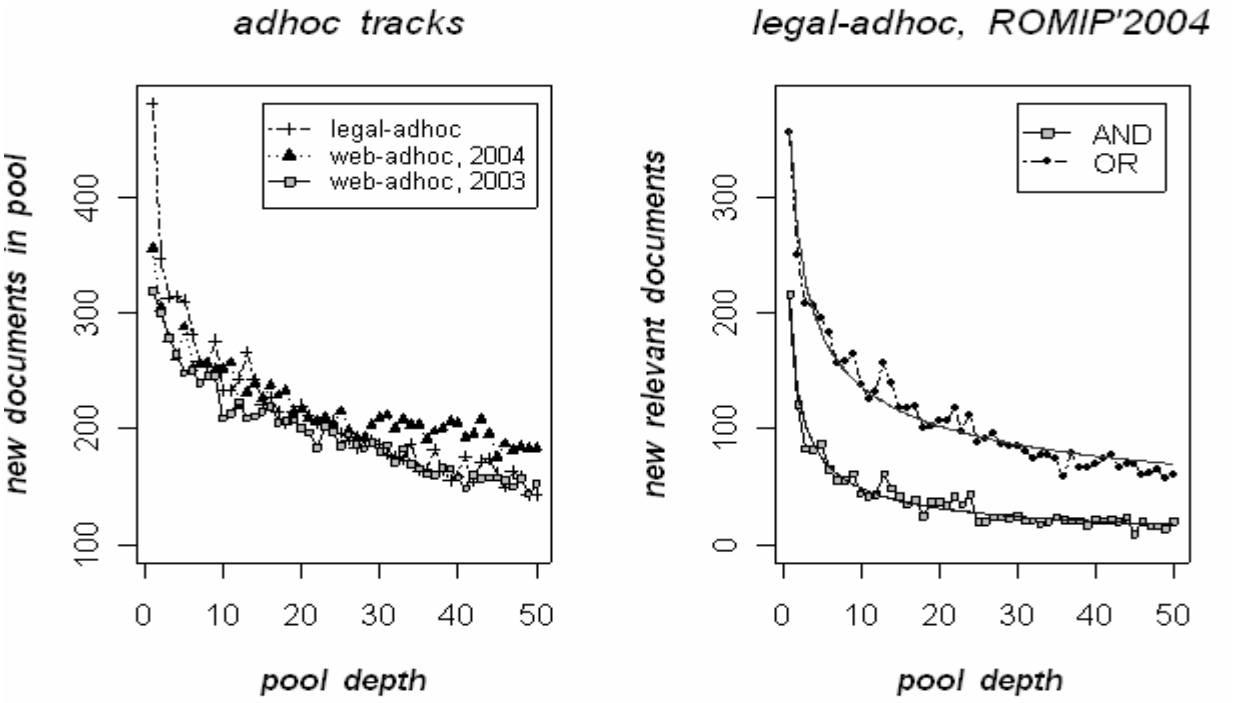

Рисунок 2. Слева – число новых документов на заданной глубине котла для трех дорожек поиска. Справа - число новых релевантных документов, обнаруживаемых на заданной глубине котла. («Поиск по нормативно-правовой коллекции», РОМИП'2004)

При использовании же слабых требований число релевантных документов превышает размер коллекции.

4 Стабильность выводов

На результат эксперимента по оценке методом «общего котла», кроме используемой для оценки метрики, влияет еще ряд параметров:

- Размер и состав набора заданий
- Глубина «котла»
- Субъективность оценки асессора
- Величина «допуска», которая используется при принятии решения «лучше»/«хуже»

Формально, на абсолютные характеристики влияет также и то, сколько и каких результатов учитывалось при построении «котлов». Однако, несложно показать, что хотя добавление еще одного варианта ответа может вызвать изменение вычисляемых оценок для других ответов (учтенных при построении «котла»), но на порядке результатов это не сказывается. Поэтому далее в этом разделе мы этот параметр не рассматриваем.

Влияние некоторых этих параметров исследовалось ранее на материалах TREC [4, 6, 15, 16, 19, 21].

Для оценки стабильности наблюдаемых результатов мы опирались на подход предложенный в работе [17]. Оценивалась зависимость вероятности сделать ошибку в выводе относительно числа запросов, по которым производится оценка, и «допуска», который используется для принятия решения.

Доля ошибок вычислялась для гипотез вида «*Если на данном наборе в k запросов, система A лучше, чем система B, на d по заданной метрике. Означает ли это, что на другом наборе из k запросов система A будет лучше системы B по этой же метрике?*". Проверка осуществляется многократным повторением эксперимента на разных множествах запроса для каждого значения *k*. Более подробное описание алгоритма можно найти в [17].

Отметим, что для того чтобы симулировать использование множества всех запросов рассматриваемые пары наборов из *k* запросов не должны пересекаться, и как следствие зависимость может быть экспериментально построена лишь для половины от общего числа оценивавшихся запросов.

Для наших экспериментов мы использовали данные поисковых дорожек за 2003 и 2004 годы. Вычисление вероятности сделать ошибку производилось по результатам 50 экспериментов для каждого значения *k*. Максимальные значения *k* составили 27 и 33 для дорожек поиска по Веб-

коллекции в 2003 и 2004 году соответственно и 45 для дорожки поиска по нормативной коллекции.

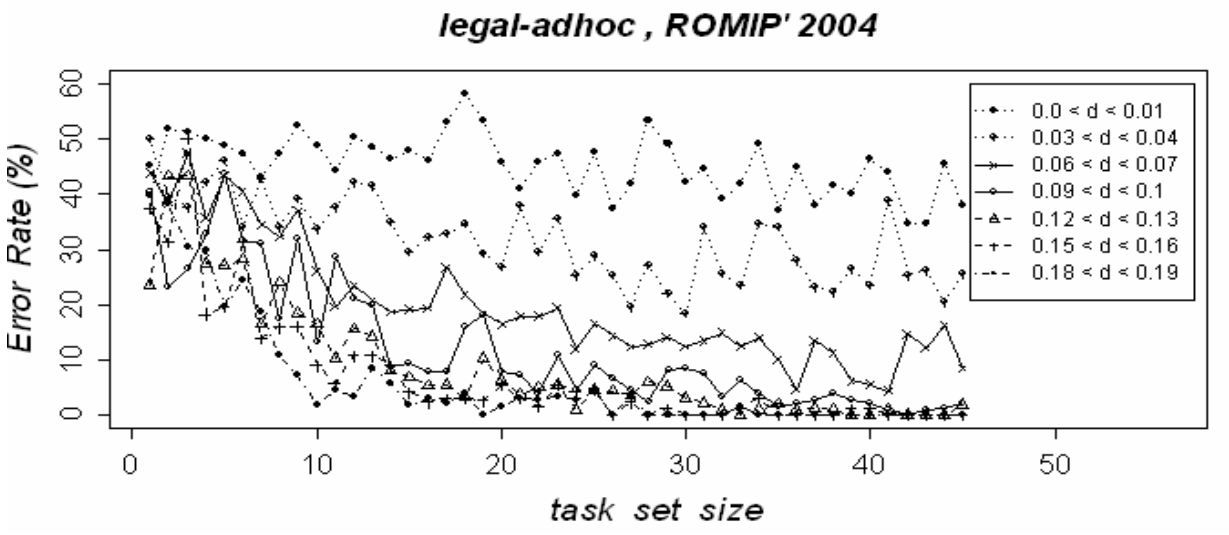

Рисунок 3. Зависимость вероятности сделать ошибку в выводе от размера набора заданий для дорожки «Поиск по нормативной коллекции» при наблюдаемой разнице в абсолютных оценках (d) в пределах заданного «узкого» диапазона. (метрика AveragePrecision, сильные требования к релеватности)

В работе [17] рассматривались зависимости для «узких» диапазонов «абсолютного превосходства» (например, от разницы в пределах 0.01 до 0.02). Пример таких зависимостей продемонстрирован на рисунке 3 для метрики AveragePrecision при использовании материалов дорожки поиска по нормативным документам и сильных требований к релевантности.

Полученные нами графики не настолько гладкие, как графики, построенные на основе данных 8 циклов TREC (1994-2001) представленные в [17]. По-видимому, это можно объяснить меньшим объемом доступных нам материалов – эксперименты в работе [17] проводились на основе 484 вариантов ответов систем. Более того, авторы этого исследования исключили из рассмотрения ряд ответов, которые они сочли плохими.

С точки зрения практического применения наибольший интерес представляет вопрос: «*Насколько велико должно быть наблюдаемое превосходство, чтобы можно было сделать вывод о превосходстве одной из систем с небольшой вероятностью ошибки?*». В таблице 1 приведены ответы на этот вопрос для нескольких метрик и дорожек РОМИП. Для иллюстрации в скобках указано максимальное абсолютное значение, показанное системами, принимавшими участие в этой дорожке. Необходимо отметить, что оценки системам в РОМИП вычислялись на вдвое большем числе запросов и очевидно, что при большем числе запросов минимальные требования к наблюдаемому превосходству скорее всего несколько снизятся.

По данным полученным в работе [4] наиболее стабильной метрикой оказалась средняя точность (averagePrecision). В нашем случае эта метрика также показала относительно неплохую стабильность.

	Average Precision	**R-precision**	**P_{10}**	**P_5**	**P_{50}**	**Recall**
Веб поиск 2003, AND	0.1 (0.176)	0.13 (0.151)	0.06 (0.138)	0.09 (0.16)	0.04 (0.133)	0.2 (0.618)
Веб поиск 2003, OR	0.07 (0.256)	0.1 (0.29)	0.08 (0.319)	0.14 (0.37)	0.07 (0.209)	0.13 (0.578)
Веб поиск 2004, AND	0.18 (0.348)	0.2 (0.326)	– (0.264)	0.2 (0.324)	0.16 (0.162)	0.2 (0.714)
Веб поиск 2004, OR	0.16 (0.394)	– (0.424)	0.2 (0.537)	– (0.585)	0.19 (0.384)	0.2 (0.657)
Поиск по нормативной коллекции, AND	0.13 (0.444)	0.18 (0.428)	0.13 (0.446)	0.15 (0.527)	0.19 (0.26)	0.17 (0.765)
Поиск по нормативной коллекции, OR	0.1 (0.519)	0.1 (0.529)	0.15 (0.747)	0.16 (0.79)	– (0.546)	0.15 (0.7)

Таблица 1. Минимальные требования к наблюдаемому абсолютному превосходству при использовании данной метрики для получения вывода с 5% вероятностью ошибки. В скобках указан наилучший результат показанный в этой дорожке (это значение вычислялось на вдвое большем числе запросов).

Для ряда других метрик мы не смогли вычислить необходимые значения гарантирующие 5% уровень ошибки на 25 запросах – такие ситуации обозначены прочерком в таблице (в частности, это означает что абсолютное превышение должно быть не ниже 0.2 при таком числе запросов)).

Полученные оценки показывают, например, что если бы оценка систем в РОМИП'2004 производилась лишь по вдвое меньшему набору запросов, то при использовании сильных требований к релевантности вероятность ошибиться в выводе, сделанном на основе метрики *averagePrecision*, при сравнении первого и третьего результата в поисковых дорожках составили бы 11.9%, 23.2% и 3.4% для поиска по Веб-коллекции в 2003 и 2004 годах и поиска по нормативной коллекции соответственно.

5 Переиспользование результатов

Одним из ключевых вопросов, связанных с проведением масштабных экспериментов по оценке, является возможность повторного использования их результатов (например, таблиц релевантности) в будущем. В частности, наиболее важными являются два следующих сценария:

- Сравнение методов A и B. Оба метода не участвовали в эксперименте.
  (Например, в случае, когда задачей является выбрать оптимальные параметры для нового метода поиска)
- Сравнение метода C, который не участвовал в эксперименте, с теми, что участвовали. (Например, хочется узнать, как новый метод поиска выглядит на фоне уже применяющихся).

Возможно ли проведение таких сравнений, так чтобы у нас была какая-то уверенность в выводах? Ключевая проблема состоит в том, что из-за неучастия метода в эксперименте, часть возвращенных им документов могла не попасть в «котел» и поэтому осталась неоцененной.

Очевидно, что наличие неоцененных документов в ответе системы сказывается на абсолютных оценках. Например, пропуск одного релевантного документа означает погрешность в 10% при оценке точности на уровне 10.

Этот эффект получил название «system omission». Ранние циклы TREC показали, что наткнуться на его проявление – вполне реальная ситуация. Интуитивно ясно, что с ростом числа участвующих систем вероятность возникновения такой ситуации уменьшается. Так, в работе [21] показано, что для TREC-5 среднее улучшение оценки эффективности системы, после добавления ее в «общий котел», составило лишь 0.5%, а для более раннего TREC-3 – 2.2%. Это показывает, что очень важно использовать адекватную глубину пула.

Эксперимент	Сильные требования к релевантности				Слабые требования к релевантности			
	Попущено релевантных	Всего релевантных	Доля (%)	Изменения выводов	Попущено релевантных	Всего релевантных	Доля (%)	Изменения выводов
1	10	138	7.2		54	453	11.9	AvgPrec – 0/1
2	7	223	3.1		80	805	9.9	P_{50} - 0/1
3	24	334	7.2	R-Prec -0/1 P_{50} - 0/2 Recall - 0/1	96	1030	9.3	Recall – 0/1
4	11	342	3.2	P_{50} - 0/1 Recall – 0/1	80	1091	7.3	R-Precision - 0/1 P_5 - 0/1 P_{50} – 0/1
5	18	213	8.5	AvgPrec – 0/1	95	700	13.6	
6	20	256	7.8		83	834	10	P_5 - 0/2 P_{10} - 0/1
7	53	132	4.0		300	552	54.3	Recall – 1/0
8	27	356	7.6		193	1149	16.8	R-Precision - 0/1 P_{50} - 1/1

Таблица 2. Качество оценки ответа, не учтенного при построении котла на примере дорожки поиска по Веб-коллекции, 2004 год. Каждая строка соответствует отдельному эксперименту.

Для изучения этого эффекта мы провели несколько экспериментов следующего вида. Фиксировался один прогон run_{new} , который играл роль нового прогона. Множество всех остальных доступных прогонов использовалось для построения матрицы релевантности и вычисления оценок всех прогонов, включая run_{new}. Альтернативный набор оценок вычислялся на основе таблицы релевантности построенной по всем прогонам.

Нас интересовало, насколько отличаются результаты попарного сравнения run_{new} с другими ответами при использовании этих альтернативных оценок. Мы различали 2 типа изменений:

A. вывод меняется на противоположный (т.е. в одном случае X лучше Y, а в другом Y лучше X)
B. вывод становится более/менее четким (в одном из случаев одна из систем превзошла другую, но в другом они показали примерно одинаковый результат).

Для этого эксперимента наблюдаемые значения считались примерно одинаковыми, если разница не превышала 5% от большего из них.

В таблице 2 приведены результаты нескольких экспериментов для дорожки поиска по Веб коллекции 2004 для нескольких разных прогонов. В колонке «Изменения выводов» перечислены число наблюдаемых изменений типов A и B соответственно для тех метрик, где они наблюдались. Кроме этого мы также привели общее число релевантных документов в run_{new} и число тех, которые не попали в котел, если run_{new} не рассматривался при его построении.

На первый взгляд можно предположить, что использование сильных требований к релевантности, по-видимому, позволяет получить более правдоподобные результаты, так как в 5 случаях из 8 изменений выводов не произошло.

Отметим также, что изменения выводов коснулись всех метрик. Наиболее часто менялись выводы для точности на заданном уровне (P_N), где максимальная наблюдаемая погрешность при использовании сильных требований к релевантности составила 40.2% (N=50), 19.2% (N=10), 12% (N=5). Для сравнения, максимальная наблюдаемая погрешность для средней точности (*AveragePrecision)* – 19.3%, R-точности (*R-precision)* – 23.5%, полноты (*Recall*) – 30.8%. При использовании слабых требований к релевантности наблюдаемые максимальные погрешности были значительно (вплоть до 20%) выше.

Тем не менее статистически эти предположения на имеющемся наборе данных не подтвердить и этот вопрос требует дополнительных исследований.

Заключение

Уверенность в осмысленности выводов является важнейшей составляющей процесса оценки. Эксперименты на основе данных TREC позволили прояснить методологические вопросы, связанные с применением метода «общего котла». В этой работе представлены предварительные результаты экспериментов с данной методологией на основе опыта семинаров РОМИП 2003 и 2004 годов.

В частности, получены следующие результаты:

- На реальных данных продемонстрирована экономия ресурсов при проведении оценки методом общего котла.
- Наблюдаемые на данных РОМИП зависимости роста числа релевантных документов и роста размера пула согласуются с данными полученными для коллекций TREC.
- Для ряда метрик получена количественная оценка минимального превосходства необходимого для получения выводов с вероятностью ошибки не более 5%
- Проведен ряд экспериментов по анализу погрешности при оценке ответа, не учтенного при построении котла на основе которого строились таблицы релевантности. Несмотря на наблюдаемую разницу в абсолютных результатах это не сказалось на выводах.

Литература

[1] Б.В. Добров, И.С. Некрестьянов, И.В. Сегалович, В.И. Шабанов Результаты первого Российского семинара по оценке методов информационного поиска (РОМИП-2003), *Труды Диалог'04*, июнь 2004.

[2] И. Кураленок, И. Некрестьянов. Оценка систем текстового поиска. Программирование, 28(4): 226-242, 2002.

[3] Труды РОМИП'2004. *Под ред. И.С. Некрестьянова* - Санкт-Петербург: НИИ Химии СПбГУ, сентябрь 2004, 214 с.

[4] C. Buckley and E.M. Voorhees. Evaluating evaluation measure stability. In Proc. of the SIGIR'2002, p. 33-40, 2002.

[5] C. Buckley and E.M. Voorhees. Retrieval evaluation with incomplete information. In *Proc. of the SIGIR'04*, pp. 25-32, 2004.

[6] R. Burgin. Variations in relevance judgments and the evaluation of retrieval performance. *Information Processing and Management*, 5(28):619-627,

[7] C.W. Cleverdon. The Cranfield tests on index language devices. In *Aslib Proceedings,* volume 19, pp. 173-192, 1967. (Reprinted in *Readings in Information Retrieval,* K. Sparck-Jones and P.Willett, editors, 1997)

[8] D. Harman. What we have learned, and not learned, from TREC. In *Proc. of the BCS IRSG'2000*, pp. 2-20, 2000.

[9] M. Lesk, G. Salton. Relevance assessments and retrieval system evaluation. *Information Processing and Management*, 3(4):343-358, 1968.

[10] Robins D. Interactive Information Retrieval: Context and Basic Notions. *Informing Science*, 3(2):57-62, 2000.

[11] M. Sanderson and H. Joho. Forming test collections with no system pooling. In *Proc. of the SIGIR'04*, pp. 33-40, 2004.

[12] T. Saracevic. Evaluation of evaluation in Information retrieval. In *Proc. of the SIGIR'95*, pp. 135-146, 1995.

[13] Singhal and M. Kaszkiel. A case study in web search using TREC algorithms. In *Proc. of the WWW2001,* pp. 708-716, 2001.

[14] I. Soboroff. On evaluating Web Search With Very Few Relevant Documents. In *Proc. of the SIGIR'04*, pp. 530-531, 2004.

[15] E.M. Voorhees. Variation in relevance judgments and the measurement of retrieval effectiveness. Information Processing and Management, 36:697-716, 2000.

[16] E. M. Voorhees. The philosophy of Information Retrieval Evaluation. *Revised Papers from the Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems*, pp. 355 – 370, 2001.

[17] E.M. Voorhees and C. Buckley. The Effect of Topic Set Size on Retrieval Experiment Error. In *Proc. of the SIGIR'02*, p. 316-323, 2002.

[18] E. M. Voorhees. Measuring Ineffectiveness. In *Proc. of the SIGIR'04*, pp. 562-563, 2004..

[19] J. W. Wilbur. The knowledge in multiple human relevance judgments. *TOIS*, 16(2):101-126, Apr. 1998.

[20] M. Wu, M. Fuller and R. Wilkinson. The Role of a Judge in a User based Retrieval Experiment. In *Proc. of the SIGIR'00*, pp. 331-333, 2000.

[21] J. Zobel. How reliable are the Results of Large-Scale Information Retrieval Experiment? In Proc. of the SIGIR'98, p.307-314, 1998.

Pooling revisited: ROMIP-based experiments

Igor Nekrestyanov, Marina Nekrestyanova,
Anna Nozik

This work focuses on evaluation of pooling-based methodology widely used to evaluate information retrieval systems [2]. Number of previous works studied pooling characteristics and impact based on TREC data [4, 15, 17, 21]. In our research we are using results of first two years of Russian Information Retrieval Seminar (ROMIP) [1, 3] (see also http://romip.narod.ru).

Three main groups of questions are considered:

- Is pooling effective way to reduce evaluation costs for all participants? Does it provide good approximation of set of relevant documents?
- How reliable are results of such experiments? Will conclusions change if some experiment parameters will be changed? E.g. if other queries will be judged.
- Are resulted collections and relevance tables are reusable? Can they be used to reasonable evaluate system run omitted from pool?

Some of these questions were considered earlier using TREC data. We are interested to verify some of published results as well as to see if ROMIP-based dependencies are similar to TREC ones.

* Эта работа поддержана грантом Яндекс N 103126.