Анализ «лабораторной» парадигмы оценки систем поиска

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 отчете представлены предварительные результаты экспериментального анализа некоторых методологических аспектов «лабораторной» парадигмы оценки систем информационного поиска. Исследование оценки методом «общего котла» проводилось на основе материалов семинара РОМИП за 2003 и 2004 годы. Рассматривались следующие вопросы:

- Насколько эффективен метод «общего котла»?
- Как параметры эксперимента влияют на выводы?
- До какой степени можно использовать полученные таблицы релевантности для оценки других систем?
- Насколько на результат влияет «человеческий фактор»?

Большинство опубликованных исследований схожих методологических вопросов проводились на основе данных TREC. Целью этой работы являлось не только получение новых результатов, но также и проверка уже опубликованных выводов на альтернативных данных.

1. Введение

Эксперименты по оценке систем информационного поиска, проводимые в рамках таких инициатив как TREC, CLEF, NTCIR и с недавних пор РОМИП [1, 3, 9], 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” [8, 17].

Оригинальная версия этой парадигмы оценки основана на следующих предположениях [17]:

- 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 [8], 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 [23].

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, 11, 13].

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? Какой 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 [15].

В течение последних пяти лет интерес к теме изучения границ применимости «лабораторной парадигмы» вообще и метода «общего котла» в частности резко возрос [7, 17, 21]. Во многом это обусловлено возможностью использовать материалы прошедших семинаров TREC для систематизации, обобщения и анализа применяемой в TREC методологии оценки [5, 16, 18, 23].

Материалы семинара РОМИП дают возможность изучения этих вопросов на альтернативных данных. При этом интересно не только исследовать новые методологические вопросы, но и проверить справедливость уже опубликованных результатов. Возможность прямого переноса выводов методологических исследований на другие эксперименты по оценке не очевидна и это обуславливает значимость проверки на альтернативном TREC материале.

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[1] 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.

---

[1] Частично результаты из разделов 3-5 уже были опубликованы в работе [4].

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. 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.

2.3. 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. Подробную информацию о методологии и результатах РОМИП можно найти в трудах семинара, опубликованных на Веб сайте семинара.

В контексте этой работы нам важно отметить следующее отличие РОМИП 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.

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.

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 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.

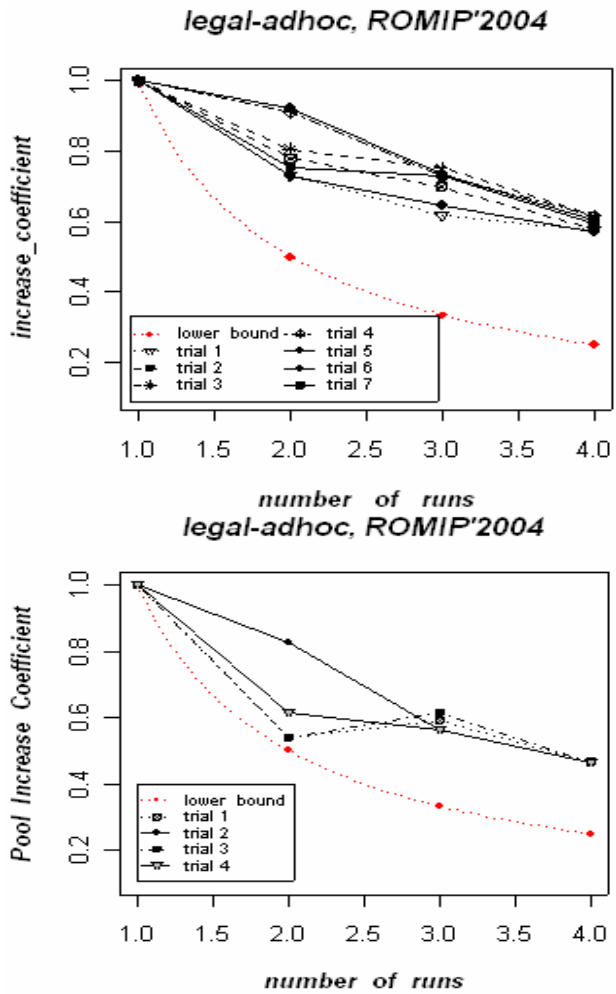

Рисунок 1. Зависимость коэффициента роста котла от числа учтенных прогонов. Сверху – рассматриваются прогоны разных систем, снизу – несколько прогонов одной системы. Коэффициент роста котла вычислялся по формуле:

$$IncreaseCoefficient == \frac{\sum_q \text{размер котла для запроса } q}{\sum_q \sum_k \text{число результатов системы k в котле q}}$$

Для оценки покрытия множества релевантных документов мы построили зависимость числа новых релевантных документов от глубины котла (рис. 2). Такая зависимость изучалась в работе [23] 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 [23]).

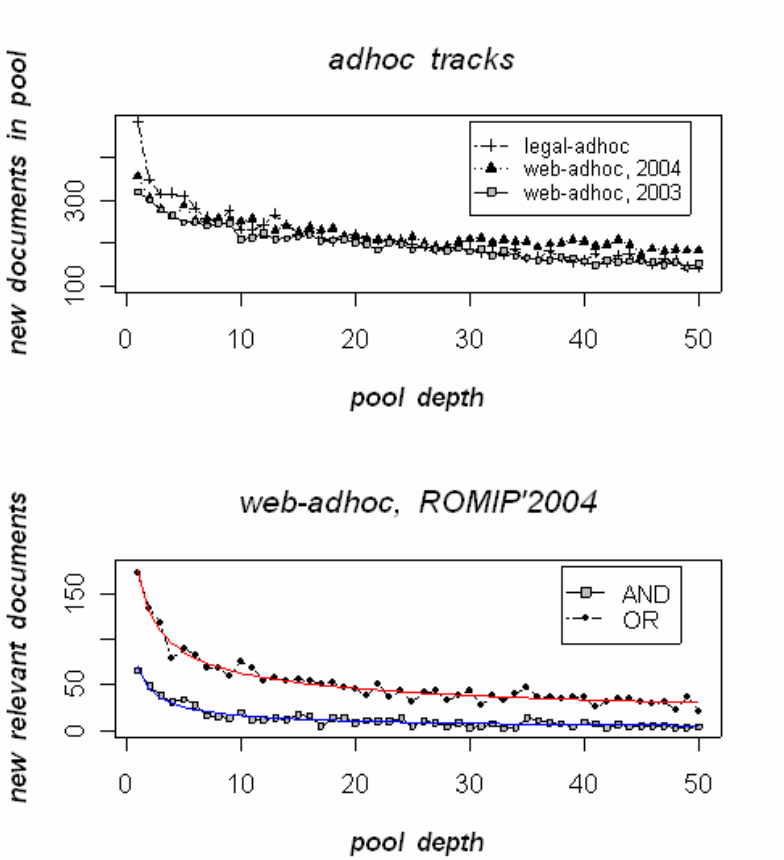

Рисунок 2. Сверху – число новых документов на заданной глубине котла для трех дорожек поиска. Снизу - число новых релевантных документов, обнаруживаемых на заданной глубине котла.

Наблюдаемые нами результаты в принципе также хорошо аппроксимируются такой формулой, хотя в нашем случае коэффициенты другие. Например, для дорожки поиска по Веб коллекции РО-

МИП'2004 – $C = 177.17$, $s = -0.443$ при использовании слабых требований к релевантности и $C = 71.51$, $s = -0.619$ 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 1.34 раза, а слабо релевантных в 1.52. Отметим, что для дорожки поиска по нормативным документам эти коэффициенты оказались очень похожи – 1.35 и 1.5 соответственно.

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 1434 таких документов (то есть при глубине пула 50 выявлено 41.7%). При использовании слабых требований релевантности - 1.9% (из 128800). Для нормативной коллекции предсказание заметно хуже – только 22% из 5920 сильно релевантных документов выявлено при глубине пула в 50, а оценка числа слабо 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 [5, 7, 16, 17, 21, 23].

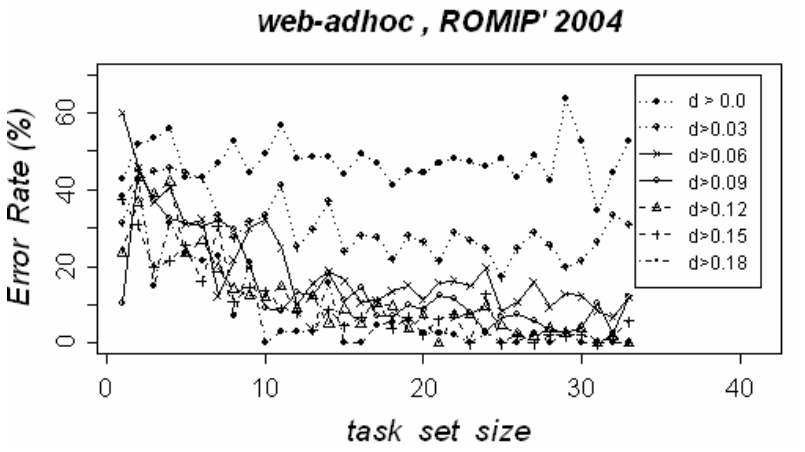

Рисунок 3. Зависимость вероятности сделать ошибку в выводе от числа заданий для дорожки «Поиск по Веб», 2004 при наблюдаемой разнице в абсолютных оценках (d) в пределах заданного «узкого» диапазона. (AveragePrecision, сильные требования к релевантности)

Для оценки стабильности наблюдаемых результатов мы опирались на подход предложенный в работе [18]. 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 «*Пусть по заданной метрике на данном наборе в k запросов система A превосходит систему B на некое абсолютное значение d. 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 наборах запросов для каждого значения *k*. Подробное алгоритм вычисления описан в [18].

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.

	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	0.12 (0.348)	0.17 (0.326)	0.14 (0.264)	0.2 (0.32)	0.17 (0.162)	0.19 (0.714)
Веб поиск 2004, OR	0.14 (0.394)	0.17 (0.424)	0.2 (0.537)	0.2 (0.59)	0.16 (0.384)	0.17 (0.657)
Поиск по нормативной коллекции, AND	0.13 (0.444)	0.18 (0.428)	0.13 (0.446)	0.15 (0.53)	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).

В работе [18] 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 Веб поиска, РОМИП'2004 и сильных требований к релевантности.

Полученные нами графики не настолько гладкие, как графики, построенные на основе данных 8 циклов TREC (1994-2001) представленные в [18]. По-видимому, это можно объяснить меньшим объемом доступных нам материалов – эксперименты в работе [18] 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 [5], наиболее стабильной метрикой оказалась средняя точность (AveragePrecision). В нашем случае эта метрика также показала относительно неплохую стабильность, хоть и не является явным лидером.

Отметим, что поскольку вычисление минимально требуемого превосходства производилось на половине запросов, то наблюдаемые абсолютные значения могли значительно отличаться от тех, что получаются на полном наборе запросов. Так, например, для дорожки Веб поиска (РОМИП-2004, сильные требования к релевантности) необходимое превосходство для P_{50} превышает наилучший из зафиксированных результатов на полном 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%, 11.8% 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 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 С, 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 [23] 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.

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).

При этом наблюдаемые значения считались «не сравнимыми», если разница не превышала 5% от большего значения.

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.

N	Сильные требования к релевантности			Слабые требования к релевантности		
	Пропущено/ Всего релевантных	**Доля (%)**	**Изменения выводов (A/B)**	**Пропущено/ Всего релевантных**	**Доля (%)**	**Изменения выводов (A/B)**
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-Prec - 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-Prec - 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.

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 наблюдае-

мая 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.

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

6. Поведение асессоров

Согласно лабораторной парадигме, асессоры моделируют реальных пользователей. Однако, в отличие от реального пользователя ИПС асессор решает другую задачу. В частности:

- Пользователь самостоятельно *формулирует* запрос, руководствуясь своей неосознанной информационной потребностью. Асессор начинает с готовой формулировки и *восстанавливает* информационную потребность.
- Пользователь заинтересован в поиске ответа на его информационную потребность, то есть:
  - Найдя требуемую информацию, он не мотивирован продолжать просмотр документов и прекращает процесс поиска.
  - Ему не обязательно просматривать каждый документ, он может смотреть лишь на наиболее «перспективные» и не анализировать документы внимательно.

  Асессор же должен идентифицировать *все* релевантные документы и одинаково внимательно изучать каждый из них.

Разные постановки решаемых задач обуславливают потенциальные расхождения в поведении реальных пользователей и асессоров. К сожалению, прямо оценить масштаб этого расхождения и его влияние на результаты оценки затруднительно. На данном этапе наших исследований мы изучали только два аспекта этой темы:

- Наблюдаются ли закономерности в поведении асессоров?
- Как субъективность оценки влияет на результаты?

6.1. Поведение асессоров

Наблюдаемые закономерности в поведении асессоров формируют основу для проведения сравнения с поведением реальных пользователей. С прикладной точки зрения аномалии в поведении асессоров теоретически могут помочь выявлять «халтуру» (в силу потери концентрации или по другим причинам)

Асессор	Процент заданий	Время на выставление оценки (сек.)			
		Vital	R+	R-	NotRel
1	**14.5**	24,1 (8,2)	41,5 (17,7)	36,7 (25,9)	21,3 (6,6)
2	**23.5**	20,1 (15,2)	25,3 (14,3)	23,8 (12,8)	10,9 (3,8)
3	**15.9**	8,7 (2,0)	11,8 (2,5)	12,5 (1,5)	7,3 (4,1)
4	**15.8**	15,5 (8,8)	20 (11,4)	17,4 (8,2)	13,7 (4,9)
5	**15.9**	19,3 (7,7)	17,6 (8,7)	16,9 (5,5)	13,8 (3,8)
6	**98.2**	23,8 (13,7)	21,3 (12,6)	24,2 (11,9)	10,7 (2,9)
7	**98.2**	27,4 (15,2)	32,6 (27,8)	35,4 (27,1)	12,3 (3,8)
8	**19.0**	29,9 (10,1)	22,6 (13,0)	23,4 (7,8)	12,5 (8,5)
9	**13.2**	49,9 (13,2)	55,6 (24,3)	61,8 (35,6)	27,9 (7,8)

Таблица 3. Среднее и стандартное отклонение времени принятия решения асессорами. (Дорожка Веб поиска, 2004 г)

Гипотеза 1.

Выставление нерелевантной оценки в среднем требует меньше всего времени. Принятие решения о частичной релевантности документа наиболее медленное.

Первая часть этой гипотезы получила строгое подтверждение, то есть она оказалась верна для *каждого* из асессоров, принимавших участие в оценке поисковых дорожек РОМИП'2004. Вторая оказалась верна лишь в 7 из 9 случаев. Более подробную информацию можно почерпнуть из таблицы 3.

Гипотеза 2.

Со временем асессор начинает быстрее принимать решения о соответствии документа запросу.

Интуитивно, можно предположить, что, просмотрев несколько документов на предмет их соответствия заданному запросу, асессор начинает лучше понимать суть задания и, следовательно, может быстрее принимать решения.

Мы проанализировали поведение для 60 случайных запросов, исключив из рассмотрения все аномально длинные (>60 сек) интервалы между оценками (таковых было в пределах нескольких процентов). Типичный график приведен на рисунке 4.

В 59 случаях линейная аппроксимация убывает и лишь в одном случае она возрастает. Углы наклона (в градусах) для всех 60 случаев приведены на рисунке 5.

Мы также исследовали, но нам не удалось подтвердить (как впрочем, и опровергнуть) следующие гипотезы:

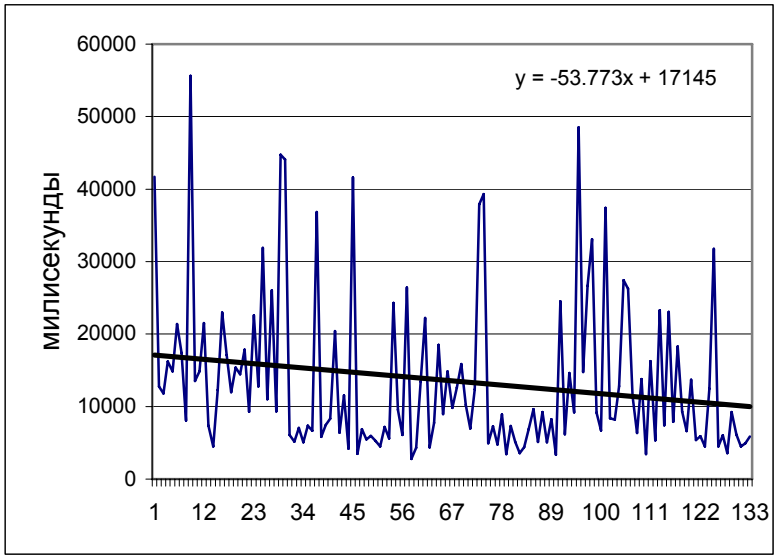

Рисунок 4. Время, затраченное на оценку N-го документа (пример для одного из асессоров).

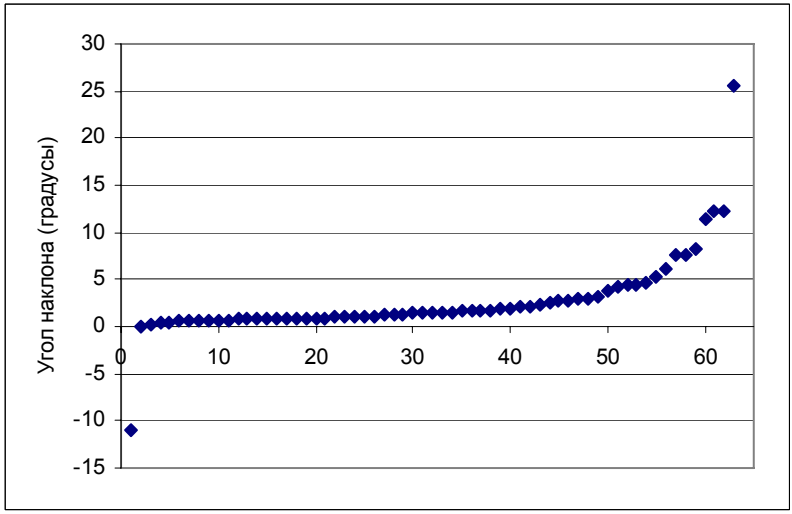

Рисунок 5. Степень уменьшения времени принятия решения асессором по мере выполнения задания.

- Чем больше асессор ставит релевантных оценок, тем дольше он читает документы. Или наоборот, кто дольше в среднем читает, ставит больше релевантных документов.
- Чем меньше времени тратится на выставление оценок (менее тщательная оценка), тем выше расхождение с мнением других асессоров.
- Асессор, у которого среднее время принятия решения для релевантных и нерелевантных оценок отличается меньше, чем у других, имеет минимальный коэффициент согласия.

Частично это объясняется тем, что абсолютная разница во времени выставления оценок зависит как минимум от технических условий проведения оценки в случае конкретного асессора, а также от особенностей асессора – например, скорости чтения, быстроты восприятия и даже навыков работы с ПК.

Переход к относительному измерению времени позволил получить косвенное подтверждение второй и третьей из этих гипотез. Однако, делать из этого выводы без проведения дополнительных исследований на расширенных наборах данных нам кажется преждевременно.

6.2. Роль «человеческого фактора»

Основной вопрос, на который мы хотим здесь найти ответ – «насколько изменились бы выводы, если бы изменились бы условия получения оценок от асессоров». В частности:

- Улучшается ли стабильность выводов с увеличением числа дублирующих оценок асессоров?
- Насколько выводы могли бы измениться, если бы оценку выполняли другие люди?
- Насколько «брак» в работе асессоров влияет на выводы?

Поскольку в оценке РОМИП участвовало небольшое число асессоров, то этих материалов конечно недостаточно для получения статистически обоснованного подтверждения гипотез. Поэтому, результаты, описываемые в этом разделе, следует рассматривать как предварительные наблюдения.

Во всех описываемых в этом разделе экспериментах считалось, что системы «не сравнимы» по данной метрике, если абсолютная разница не превышала 5% от большего значения.

6.2.1. Влияние «брака»

Для оценки влияния «брака» в работе асессоров на выводы мы провели серию экспериментов по внесению случайного шума в «идеальную» матрицу результатов и анализу влияния этого шума на

выводы. Эксперименты проводились на основе материалов дорожки поиска по Веб-коллекции, РОМИП-2004.

Рассматривались следующие подходы к внесению шума:

1. Случайное подмножество (5/10/20%) релевантных оценок заменялось нерелевантными.
2. Случайное подмножество нерелевантных оценок заменялось релевантными.

В первом случае внесение до 20% шума не привело к значительным изменениям выводов. Единственная метрика, для которой мы наблюдали изменение выводов на противоположные, – это R-precision при использовании строгих требований к релевантности. При 5% уровне ошибок это случилось в одном эксперименте из 5 для одной пары систем. При уровнях ошибки в 10 и 20% - такой эффект наблюдался дважды.

Заметим, что абсолютные значения метрик изменялись весьма значительно. Так сильных требованиях к релевантности максимальные отклонения R-precision составили – 13, 20 и 33% для уровней ошибки 5, 10 и 20% соответственно а P_5 – 10, 19 и 29% соответственно. Меньше всего изменялась полнота, что ожидаемо при такой схеме внесения шума.

Отметим также, что при использовании в экспериментах слабых требований к релевантности изменений выводов не наблюдалось, хотя абсолютные значения метрик также менялись значительно. Например, максимальные отклонения P_5 составили 13, 20 и 29%.

При внесении шума путем замены нерелевантных оценок релевантными ситуация изменилась. При 5% уровне шума значительных изменений также не наблюдалось (лишь в одном случае в одном эксперименте с использованием слабых требований к релевантности вывод изменился на противоположный по метрикам R-precision и AvrPrecision). Однако, в большинстве экспериментов с уровнем ошибки в 10 и 20%, как минимум один из выводов по метрикам R-precision, AvrPrecision и Recall менялся на противоположный, если использовались слабые требования к релевантности. При использовании сильных требований изменение наблюдалось только однажды. Этот результат объясняется тем, что при использовании такой схемы генерации шума таблица релевантности, построенная на основе слабых требований к релевантности, меняется в значительной степени. А если используются сильные требования, то изменения в таблице релевантности заметно менее значительны.

Тем не менее, поскольку «случайные» ошибки асессоров вряд ли могут превысить даже 5% барьер, то, по-видимому, значительно влияния на выводы это оказать не может.

6.2.2. Влияние числа оценок

В 2004 для дорожки Веб-поиска было получено по 3 оценки для каждой пары документ-запрос (с использованием расширенных описаний). При этом двое асессоров оценили почти все задания[2].

Для изучения вопроса о влиянии числа оценок мы проверили насколько изменяются выводы, если использовать только одну или две оценки. В каждой серии экспериментов использовалась по 4 подмножества оценок.

	1	**2**	**3**	**4**	**Итого**
Сильные требования к релевантности					
AvgPrec	3/0	3/0	3/0	2/0	11/0
P_5	4/0	4/1	3/1	4/0	15/2
P_{10}	3/0	2/1	1/0	2/1	8/2
P_{50}	1/0	1/0	1/0	1/0	4/0
R-Precision	3/1	3/1	5/1	3/1	14/4
Recall	2/1	4/1	4/0	5/0	15/2
Итого	16/2	17/4	17/2	17/2	
Новых релевантных	139%	138%	147%	134%	
Слабые требования к релевантности					
AvgPrec	1/0	2/0	1/0	2/0	6/0
P_5	2/0	4/0	3/0	2/0	11/0
P_{10}	1/0	2/0	1/0	2/0	6/0
P_{50}	1/0	1/0	1/0	1/0	4/0
R-Precision	0/1	2/1	2/1	2/1	6/4
Recall	3/0	1/0	1/0	1/0	6/0
Итого	8/1	12/1	9/1	10/1	
Пропущено релевантных	42%	42%	40%	43%	

Таблица 4. Изменения в выводах при использовании одной оценки асессора для каждой пары документ-запрос по сравнению с полными таблицами (в ячейках – изменения типов A/B)

[2] Более 98% пар документ-запрос.

В случае одной оценки выбор подмножеств основывался на случайном упорядочивании всех асессоров. Для каждой пары документ-запрос использовалась оценка того асессора, кто раньше упоминался в этом списке.

Для случая двух оценок подмножества строились следующим образом. В первых двух вариантах учитывалась оценка только одного из оценивавшего все задания асессоров. В третьем - учитывались только их оценки (в редких случаях, когда их не было, использовались оценки еще 2-х асессоров). Четвертый вариант - пара оценок выбиралась случайно независимо для каждой пары документ-запрос.

	1	**2**	**3**	**4**	**Итого**
Сильные требования к релевантности					
AvgPrec	1/0	1/0	2/0	1/0	5/0
P_5	0/0	3/1	2/0	3/0	8/1
P_{10}	1/0	2/1	3/0	3/0	9/1
P_{50}	0/0	0/0	2/0	0/0	2/0
R-Precision	2/0	1/1	2/1	2/0	7/2
Recall	2/0	4/0	1/0	2/0	9/0
Итого	6/0	11/3	12/1	11/0	
Новых релевантных	19%	54%	39%	29%	
Слабые требования к релевантности					
AvgPrec	0/0	1/0	1/0	0/0	2/0
P_5	3/0	1/0	1/0	0/0	5/0
P_{10}	1/0	0/0	3/0	2/0	6/0
P_{50}	0/0	2/0	0/0	0/0	2/0
R-Precision	0/0	0/0	1/0	0/0	1/0
Recall	0/0	1/0	1/0	0/0	2/0
Итого	4/0	5/0	7/0	2/0	
Пропущено релевантных	21%	6%	19%	14%	

Таблица 5. Изменения в выводах при использовании двух оценок асессора для каждой пары документ-запрос по сравнению с полными таблицами (в ячейках – изменения типов A/B)

Общее число сравниваемых пар систем составило 28. Сводные результаты этого эксперимента приведены в таблицах 4 и 5.

Во-первых, отметим, что выбор асессоров может в значительной мере повлиять на результаты. Так, например, даже при использовании двух оценок разница между первым и вторым экспериментами весьма значительна, а в этом случае рассматриваемые наборы оценок фактически получались заменой одного асессора другим.

Увеличение числа дублирующих оценок повышает надежность выводов. Это можно объяснить насыщением матрицы релевантности, значительные изменения которой коррелируют с заметными изменениями в выводах.

Эти эксперименты также демонстрируют хорошую стабильность средней точности – ни в одном из экспериментов вывод по этой метрике не изменился на противоположный. Это согласуется с наблюдениями других исследователей [5, 16]. В случае слабых требований к релевантности неплохая стабильность также у оценки полноты и, что удивительно, у P_{50}.

Близкие вопросы обсуждались в ряде других исследований [7, 10, 16]. Все они сходятся в том, что субъективность асессоров компенсируются усреднением по числу запросов и выводы получаются стабильными вне зависимости от числа оценок. При этом 50 запросов считается вполне достаточным [10,16]. Это не совсем согласуется с нашими наблюдениями – у нас получилось, что справедливость этого утверждения зависит от используемой метрики.

В отличие от нас, авторы последних исследований и, в частности, наиболее масштабной работы [16], использовали несколько другой подход к оценке стабильности. Так, все результаты считались сравнимыми, вне зависимости от наблюдаемой разницы абсолютных значений, а стабильность оценивалась на основе коэффициента корреляции Кэндела τ[3] (оценивает разницу двух упорядоченных списков). Они также фокусировались только на средней точности.

При использовании такого подхода мы также получаем, что $\tau>0.9$ для средней точности вне зависимости от числа оценок. Однако, для ряда других метрик в случае одной оценки наблюдались заметно меньшие значения. Например, для R-precision – 0.71 или P_5 – 0.77. Более того – не использование «допуска» при сравнении значений делает такой подход чувствительным к небольшим изменениям близких результатов. Так, в одном из случаев $\tau=0.79$ для P50, хотя ни одного значимого изменения не произошло.

---

[3] В [20] считалось, что при $\tau > 0.9$ ранжирования можно считать эквивалентными, а $\tau < 0.8$ – значительно различающимися.

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.
- Показано, что «случайные» ошибки асессоров не оказывают заметного влияния на выводы экспериментов.
- Продемонстрировано, что результат оценки может в значительной степени зависеть от того, кто был асессором. Дублирование оценок позволяет улучшить стабильность выводов.
- Сделана попытка выявить закономерности в поведении асессоров. Однако удалось подтвердить лишь часть рассматривавшихся гипотез.

Многие из полученных результатов требуют дополнительного анализа и новых экспериментов. Ряд вопросов так и остался открытым. Так, например, осталось неясным насколько высокое качество поиска с точки зрения асессоров согласуется с мнением реальных пользователей. Мы планируем продолжить исследование этих вопросов.

Тем не менее, полученные результаты нам представляются весьма полезными и могут быть, в частности, использованы для улучшения методологии оценки РОМИП.

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] И. Некрестьянов, М. Некрестьянова, А. Нозик. К вопросу об эффективности метода «общего котла». Принято к публикации в трудах RCDL-2005.

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

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

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

[8] 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)

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

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

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

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

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

[14] Singhal and M. Kaszkiel. A case study in web search using TREC

algorithms. In *Proc. of the WWW2001,* pp. 708-716, 2001.

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

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

[17] 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.

[18] 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.

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

[20] Ellen M. Voorhees Evaluation by highly relevant documents. *In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74-82, 2001.

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

[22] 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.

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

Analysis of Cranfield approach to IR system evaluaton on ROMIP data

Igor Nekrestyanov, Marina Nekrestyanova, Anna Nozik
Saint Petersburg State University
http://ir.apmath.spbu.ru
{igor, marina}@meta.math.spbu.ru, blake_@mail.ru

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 [5, 16, 18, 23]. 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).

Four 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?
- To which extent "human factors" (such as subjectivity) can have an effect on the evaluation of retrieval results? What may change if other assessors will be judging system results? Does increasing number of assessors help to improve stability?

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.

Most of our results are in line with previous research. Therefore we mention only few specific ones here:

- Calculated quantitative estimations of minimal difference in scores for main metrics on ROMIP data.
- Shown that evaluation of system omitted from the pool is reasonably reliable for ROMIP data.
- Proved that random errors in assessor judgments do not have significant impact on conclusions.
- Demonstrated that individual assessor can make the difference in the conclusions and redundant assessment helps to improve stability.