РОМИП 2003: Опыт организации

Игорь Кураленок,
Игорь Некрестьянов,
Екатерина Павлова
Санкт-Петербургский Государственный Университет
romip@oasis.apmath.spbu.ru
http://ir.apmath.spbu.ru

1. *Введение*

Инициатива РОМИП нацелена на регулярное проведение семинаров РОМИП, каждый из которых посвящен оценке эффективности решения одной или нескольких задач текстового поиска [1]. Список рассматриваемых задач определяется на основе обсуждения с участниками семинара и возможностей реализации этих проектов (доступности наборов данных и других ресурсов).

Структурно семинар представляет собой набор «дорожек» — секций, посвященных конкретным проблемам (с фиксированной задачей и правилами оценки). В 2003 году семинар состоял из двух дорожек — *поиска* и *классификации*, для участия в которых были поданы заявки на 9 систем, а дошло до финиша 7.

В качестве основы для тестового набора данных в 2003 году использовалась коллекция Веб страниц из домена `narod.ru`. Поскольку общий объем страниц в домене `narod.ru` значительно превышал предполагаемый размер набора, то коллекция была сформирована на основе случайной выборки сайтов из домена. Коллекция содержит порядка 600000 HTML страниц с 22000 сайтов общим объемом более 7 Гб.

Такой выбор обусловлен не только высокой актуальностью задач поиска в контексте Веб, но в значительной мере также и легальностью использования этого набора данных участниками семинара (легальность обеспечивается пользовательским соглашением Яндекс, которое регулирует правила использования информации на сайтах `narod.ru`). Рассматривавшиеся альтернативные варианты не Веб-коллекций были отклонены, в основном, по причине невозможности обеспечить легальный доступ за столь ограниченное время.

Отметим, что такой объем коллекции не означает, что в семинаре могут принимать участие только большие ИПС. На самом деле за-

дания 2003 года вполне могли быть выполнены коллективом из одного-двух человек на персональном компьютере.

Для того чтобы набор данных был максимально приближен к реальной задаче поиска в Веб, было решено не подвергать содержимое страниц никакой модификации (подавляющее большинство страниц на русском языке в кодировке cp1251).

В целях упрощения обмена данными с участниками все процедуры обмена использовали XML в качестве основы форматов представления данных. Оргкомитет также предоставил участникам часть инструментов необходимых для работы с этими данными, включая инструмент для извлечения данных из распространявшейся тестовой коллекции, и инструмент для вычисления оценки.

Более подробно организация семинара и каждой из дорожек 2003 года описываются в следующих разделах.

2. Организация семинара

Общий цикл проведения семинара состоит из следующих этапов:

0. Подготовительный:
Формируется набор дорожек, включая методологию создания тестовых наборов данных, фиксируется график проведения. Производится прием заявок от участников, назначаются псевдонимы для обеспечения анонимности оценки.

1. Подготовка и распространение тестовых наборов данных:
Подготовка и распространение осуществляется оргкомитетом, от участников может потребоваться подписание соглашения об использовании данных.

2. Проведение экспериментальных прогонов своей системы:
Участником самостоятельно и на своём оборудовании выполняет поисковые задания. Результаты предоставляются оргкомитету, используя псевдоним.

3. Оценка полученных ответов:
Организуется оргкомитетом с соблюдением анонимности источника результата. Вся накопленная информация о правильных/неправильных ответах предоставляется каждому из участников. Участники также получают итоговые оценки ответов своей системы по некоторым метрикам.

4. *Очная встреча*
Целью очной встречи является не только обсуждение наблюдаемых результатов и обмен опытом, но также обсуждение методологии и процесса проведения семинара, включая формирование предварительной программы на следующий цикл.

В 2003 году финансирование затрат на проведение семинара производилось участниками совместно, включая затраты на подготовку и распространение тестовых наборов данных, организацию очной встречи и работу асессоров. Расходы на работу асессоров (а это наиболее значительная часть затрат) могли компенсироваться путем предоставления асессоров для выполнения своей доли работ по оценке.

2.1. Методология оценки

Процедура оценки, безусловно, различается для различных задач информационного поиска и формируется отдельно для каждой конкретной дорожки, но можно выделить ряд общих основополагающих соображений:

- ***Равноправие систем.*** Процедура оценки должна по возможности гарантировать равноправие систем при оценке результатов;
- ***Анонимность источника результата.*** При проведении оценки должна соблюдаться анонимность источника результата - то есть, те, кто оценивают результат, не должны знать, какая система выдала этот результат;
- ***Использование апробированных подходов.*** Предпочтительным является использование апробированных методологий оценки, поскольку это повышает уверенность в получении надежных результатов;
- ***Избыточность для понижения влияния субъективности.*** Экспертные оценки правильности ответа конечно же отражают субъективную точку зрения конкретного эксперта, которая зачастую может не совпадать с мнением другого человека. Для понижения влияния субъективности оценки на качество результатов предпочтительно использовать несколько независимых оценок разных экспертов и по ним выводить итоговое решение.

Важным принципом РОМИП является предоставление участникам не только итоговых сводных оценок результатов работы их системы, но также и полных таблиц правильных ответов, что позволяет самостоятельно проводить дополнительные эксперименты и оценивать их.

2.2. Оргкомитет

Согласно идеологии РОМИП лишь координирует процесс проведения семинара. Задачи, решаемые оргкомитетом, включают:

- Координация формирования программы семинара на следующий цикл (обсуждения и голосования).
- Распространение приглашений к участию и популяризация семинара.
- Подготовка и распространение наборов данных.
- Организация обмена данными с участниками (тестовые наборы, результаты прогонов и оценки).
- Подготовка ПО для асессоров и вычисления итоговых оценок.
- Отбор и подготовка заданий для асессоров.
- Распределение заданий для асессоров и дальнейшая координация их работы.
- Организация очной части семинара.

В 2003 году оргкомитет РОМИП состоял из 8 человек представляющих пять различных организаций. При этом 5 человек (из двух организаций) не связаны с системами, которые принимали участие в дорожках РОМИП, и именно они контролировали процесс подготовки наборов данных, подготовку заданий для экспертной оценки и сам процесс экспертной оценки.

3. Дорожка по поиску

Эта дорожка посвящена классической проблеме поиска по запросу и методология ее проведения во многом следует опыту проведения дорожки по ad-hoc в рамках TREC (1992 – 1998 годы), в частности используя подход “общего котла” к организации сбора экспертных оценок [4].

В 2003 году для участия в дорожке по поиску было подано 7 заявок, но добрались до финиша лишь 5 участников, которые предоставили организаторам результаты 9 прогонов.

3.1. Правила проведения

Формально рассматриваемая задача формулировалась следующим образом:

По данному запросу пользователя вернуть упорядоченный список документов из коллекции `narod.ru`, *которые наилучшим образом удовлетворяют информационные потребности пользователя. Максимальная длина списка — 100 документов.*

Для того чтобы приблизить задачу к реальной задаче поиска в Веб решаемой поисковыми системами общего назначения, запросы пользователей отбирались из журнала поисковой системы Яндекс полностью автоматическим образом — начиная с некоторой точки, берутся все запросы, удовлетворяющие критериям отбора:

1. русскоязычные;
2. без явных грамматических ошибок;
3. без употребления ненормативной лексики.

Всего было отобрано 10000 запросов, которые и выполнялись каждой из систем участниц. Такой подход позволяет понизить вероятность подстройки систем под конкретные задания и обеспечивает гибкость при выборе заданий для экспертной оценки.

Оценка основана на методе «общего котла» (pooling), который используется в TREC [3, 4]. «Общий котел» — это объединенное множество первых N_q документов (для этой дорожки N_q=50) из выдачи каждой из систем для данного запроса q. Каждый из документов попавших в такой котел далее оценивается экспертами на соответствие запросу.

3.2. Выбор заданий для оценки

Отбор заданий для оценки производился независимо 4 экспертами, после того как были собраны результаты от всех систем участниц.

Процедура отбора была организована следующим образом: из общего списка в 10000 запросов случайным образом выбиралось 20, просматривая которые эксперт мог отобрать не более одного запроса. Основным критерием отбора была способность эксперта предоставить разумную трактовку исходной информационной потребности пользователя. Выходом процедуры отбора являются расширенные версии исходного запроса, представляющие собой пары “запрос+трактовка”. Всего было отобрано 62 запроса.

Используемый нами для оценки метод “общего котла” подразумевает, чтобы все оценки релевантности для одного и того же запроса производятся исходя из одного и того же понимания информационной потребности (а иначе собранные оценки будут несогласованны). Однако, короткие запросы, используемые в этой дорожке, зачастую могут трактоваться несколькими разными способами. Рас-

ширенная версия запроса используется для упразднения неоднозначности, детально описывая искомую информацию, как это понимает эксперт[1] (тем самым уточняется одна из возможных информационных потребностей, выраженная этим запросом). Для примера приведем расширенное описание одного из запросов (полный список опубликован в приложении к этому сборнику):

<definition type="***Relevance Judgement***" id="***13894***">
<query>***ученый древности***</query>
<description>
Документ должен содержать информацию о хотя бы одном ученом древности (годы жизни, области научных знаний, основные достижения). Простая ссылка на достижения ученых древности не делает документ релевантным.
</description>
</definition>

На втором шаге процедуры отбора выбиралось подмножество расширенных запросов так, чтобы затраты на сбор экспертных оценок не превышали объем доступных ресурсов. Было отобрано 54 запроса, суммарный объем котлов которых составил 10084 пар документ-запрос. Полный список отобранных запросов приведен в "Приложении А" этого сборника.

3.3. Сбор оценок асессоров

Экспертные оценки о релевантности документа данному запросу выставляются асессорами на основе расширенного представления запроса.

Для того чтобы снизить влияние субъективности оценки мы собирали по две независимые оценки на пару документ-запрос (на большее не хватило ресурсов). Кроме этого каждый "котел" делился между тремя асессорами, так чтобы на долю каждого приходилось 70% содержимого котла (таким образом пересечение оценок между любыми двумя составляло 40%).

Всего в оценке принимало участие 13 асессоров — 9 асессоров от оргкомитета и 4 от участников, выполнявших часть работы по

[1] Для того, чтобы различать разные роли в РОМИП используется термин «эксперт» для лиц, фиксирующих информационную потребность, и «асессор» для лиц, реально производящих оценку руководствуясь расширенным заданием.

оценке самостоятельно (таковых было 2-е). К сожалению, нам не удалось обеспечить равномерное распределение нагрузки между асессорами — она варьировалась в диапазоне от 809 до 2813 оценок от одного асессора. Отметим, что множества асессоров и экспертов не пересекались.

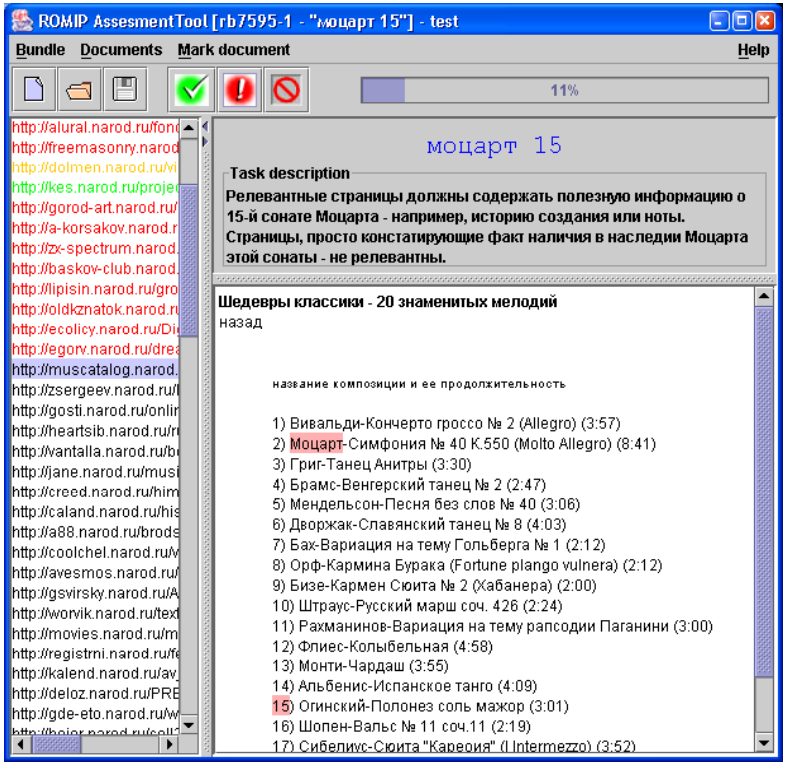


Рисунок 1. Инструмент асессора - дорожка поиска.

Для оценки документов был разработан инструмент, который использовался всеми асессорами. Инструмент обеспечивал одновременный просмотр оцениваемой пары документ/запрос и выставление оценки по шкале — “релевантно”/“не релевантно”/“невозможно оценить”. Отметим, что у асессора есть возможность вернуться к уже просмотренному документу и изменить свое решение. Для облегчения работы асессора инструмент подсвечивал в документе все словоформы слов из исходного запроса. Внешний вид разработанного инструмента представлен на рисунке 1.

Реализация инструмента основана на технологиях Java для обеспечения платформной независимости способов отображения документов. К сожалению, отображение HTML-документов средствами Java оказалось очень медленным для большого количества документов. Это значительно тормозило процесс оценки и поэтому была создана модификация инструмента, которая использовала Microsoft Windows Internet Explorer для отображения документов. Около 50% оценок было получено с помощью этой версии.

Инструкция для асессора содержала следующую информацию:

Инструкции для асессоров
(дорожка по поиску)

Задачей асессора является оценить соответствие документов заданной информационной потребности. В этом документе описаны некоторые общие принципы, направленные на получение более адекватных и согласованных результатов оценки.

Шкала оценки

Вообще, решение о степени соответствия (релевантности) документа запросу принимается асессором в соответствии с его собственным мнением.

Короткая форма запроса зачастую допускает неоднозначную трактовку того, что является искомой информацией, и поэтому все запросы снабжены комментариями, поясняющими, какую информацию хочется найти.

Для оценки документов используется следующая шкала:

Релевантный (RELEVANT)

Исходя из общих соображений релевантный документ должен содержать искомые сведения (хотя бы частично) или информацию, почему искомые сведения невозможно найти по такому запросу (например, для запроса "зимняя олимпиада 1999" релевантной будет страница объясняющая, что Олимпиады проводятся раз в 4 года и упоминающая, что олимпиада в Нагано была в 1998 году).

Не релевантный (NOT RELEVANT)

Очевидно, что документ не является релевантным, если он содержит только информацию не относящуюся прямо к исходному запросу. Также не релевантными являются документы, содержащие через чур общую или, наоборот, детальную информацию, чем под-

разумевается текстом запроса. Не релевантными также являются и документы, которые содержат

Невозможно оценить (CAN NOT JUDGED)

К этому классу относятся документы которые не удалось прочитать с помощью инструмента для оценки. Возможные причины - документ не отображается из-за особенностей html разметки (симптом - панель документа пустая) или документ в странной кодировке (на другом языке).

Документы, которые полагаются на картинки для передачи информации и почти не содержат текста (картинки не будут загружены) к этой группе не относятся, а являются нерелевантными!

Всего было собрано 20589 оценок. Исходя из оценки производительности в 60 оценок/час, общие трудозатраты на оценку составили 343,15 человеко-часа. На практике эффективность в значительной мере определялась скоростью отображения документа, а не скоростью его восприятия асессором. Поэтому реальная производительность асессоров сильно зависела от используемой версии инструмента (для некоторых документов время отображения версией с IE уменьшалось в десятки раз), а также от производительности оборудования, и варьировалась от 5-10 до 100 оценок в час.

3.4. Таблицы релевантности

Таблица релевантности — это таблица, содержащая информацию о том, какие документы считаются релевантными данным запросам, а какие нет (то есть это “эталонный ответ”). Содержимое таблицы определяется оценками асессоров.

Однако, в нашем лишь в 32% случаев второй асессор поддерживал мнение первого о том, что документ релевантен. Высокий процент расхождений в мнениях асессоров не удивителен и согласуется с результатами других исследований (в работах [5, 6, 7] цифры варьируются в диапазоне от 30 до 50%). Для разрешения неоднозначных ситуаций мы использовали два альтернативных подхода[2] (и таким образом далее используются две альтернативных таблицы релевантности):

- ***“Слабая согласованность” (weak)***
  В этом случае результат:

[2] Интересны промежуточные варианты, которые учитывают попарные коэффициенты согласия экспертов и коэффициенты доверия к оценкам экспертов. Это тема для дальнейших исследований.

- “релевантенный”, если хотя бы одна оценка “релевантный”;
- “невозможно оценить”, если все оценки “невозожно оценить”;
- в остальных случаях “не релевантен”.

- ***“Сильная согласованность” (strong)***

  В этом случае результат:
  - “невозможно оценить”, если все оценки “невозможно оценить”;
  - “не релевантенный”, если хотя бы одна оценка “не релевантенный”;
  - в остальных случаях “релевантеный”.

В случае слабой согласованности 1187 пар документ-запрос были признаны релевантными, а в случае строгой — лишь 391. В обоих случаях 80 документов были признаны “не оцениваемыми”.

3.5. Метрики для вычисления оценок результатов прогонов

Оргкомитет организовал централизованную оценку результатов прогонов участников по нескольким метрикам. Большинство использовавшихся метрик — это классические метрики для оценки результатов поиска, которые часто используются как на конференциях по оценки систем, так и в самостоятельных исследованиях по этой тематике.

Для оценки результатов работы по отдельному запросу использовались следующие метрики:

Точность: доля релевантных документов в ответе.

$$p = \frac{relevant}{total}$$

Полнота:

отношение найденых релевантных документов к общему их количеству.

$$r = \frac{relevant}{total\_relevant}$$

Здесь *relevant* — это число релевантных документов в ответе системы из *total* документов, а *total_relevant* — это общее число известных релевантных документов для данного запроса.

Точность на уровне α (P_α)***:***

Популярная вариация критерия точности, которая вычисляется как точность для ответа содержащего первые α документов. Такая метрика, например, довольно хорошо характеризует насколько качество поиска, если пользователь просматривает лишь несколько первых результатов (как это часто бывает при поиске в Интернет).

R-Точность *(R_p)**:***

Эта метрика также представляет из себя оценку точности, но не на абсолютном уровне, а относительно числа известных релевантных документов для данного запроса.

$$R_p = P_{total\_relevant}$$

Средняя точность *(averageP)**:***

это одна из наиболее распространенных характеристик эффективности поиска. Для вычисления средней точности используется набор множеств, соответствующих появлению новых релевантных документов в выдаче системы. Для каждого релевантного документа *d* вычисляется точность на уровне *level* равному порядковому номеру документа в выдаче. Эти значения усредняются по общему количеству релевантных документов для данного запроса. Более формально:

$$averageP = \sum_{level:d_{level} \in relevant} P_{level}$$

Отметим, что согласно недавним исследованиям Бакли и Вурхес [8] средняя точность является оценкой с наилучшим соотношением стабильность/дескриптивность

Отметим, что при вычислении оценок на уровнях больше глубины пула не для всех документов были известны оценки релевантности. Все такие документы, а также документы которые были помечены “невозможно оценить” считались нерелевантными.

Итоговые оценки для прогона вычислялись усреднением вышеуказанных метрик по отдельным запросам (микроусреднение) и с помощью двух вариантов графиков зависимости точности от полноты.

Первый вариант графика — классический 11-точечный график используемый в конференции TREC [5]. Для его вычисления использовалась следующая форма:

$$r_{\lambda}^{TREC} = \begin{cases} \frac{1}{n}\sum_{i=1}^{n} \max_{r(level)\geq\lambda} p(level),\ \exists\, level : r(level) \geq \lambda \\ 0,\ \ \forall\, level : r(level) < \lambda \\ 1,\ \ total\_relevants = 0 \end{cases}$$

где *p(level)* и *r(level)* точность и полнота на уровне *level*. Эти значения вычисляются на 11 уровнях полноты от 0.1 до 1, причем если заданный уровень полноты не достижим, то значение в этой точке интерполируется .

В качестве альтернативы классическому 11-точечному графику от TREC мы также строили его модифицированный вариант, который возможно меньше дискриминирует системы с короткими ответами и позволяет лучше оценивать поведение систем на близких к нулю уровнях полноты. Отметим, что эта гипотеза требует обоснования и поэтому не стоит спешить с выводами на основе модифицированных 11-точечных графиков. Формально:

$$r_{\lambda}^{RIRES} = \begin{cases} \frac{1}{n}\sum_{i=1}^{n} p(level), \\ \qquad level : relevants_{level} = [\lambda \cdot total\_relevants] \\ 1, \quad total\_relevants = 0\ \&\ total = 0 \\ 0, \quad total\_relevants\ \&\ total \neq 0 \end{cases}$$

3.6. Сводные результаты систем

Информация об итоговых оценках для каждого из прогонов, полученных при использовании слабой и строгой таблиц релевантности, представлена в таблицах 1 и 2.

Отметим, что лидерство зависит от типа используемой таблицы релевантности и конкретной метрики.

Мы также приводим 11-точечные графики для наглядного сравнения результатов работы участвовавших систем. Относительно низкие абсолютные результаты, показанные системами обуславливаются сложностью задания и довольно жесткими критериями к релевантности ответов предъявляемыми экспертами. Это косвенно подтверждается значительным расхождением во мнениях асессоров, которое наблюдается в полученных результатах оценки.

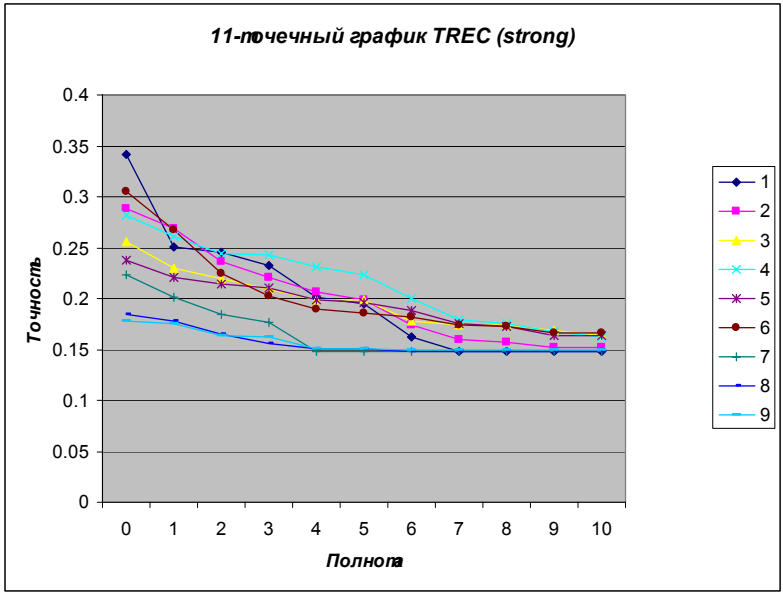
	P	*R*	*averageP*	R_p	P_{10}	P_5
1	***0.19***	0.26	***0.10***	***0.20***	0.13	***0.15***
2	0.05	0.50	0.05	0.19	0.06	0.05
3	0.04	0.67	0.04	0.18	0.03	0.04
4	0.06	***0.68***	0.05	0.19	0.05	0.05
5	0.05	0.63	0.04	0.17	0.03	0.02
6	0.16	0.43	0.07	0.18	***0.14***	***0.15***
7	0.09	0.19	0.03	0.17	0.07	0.08
8	0.04	0.21	0.01	0.16	0.03	0.03
9	0.02	0.23	0.01	0.15	0.01	0.01

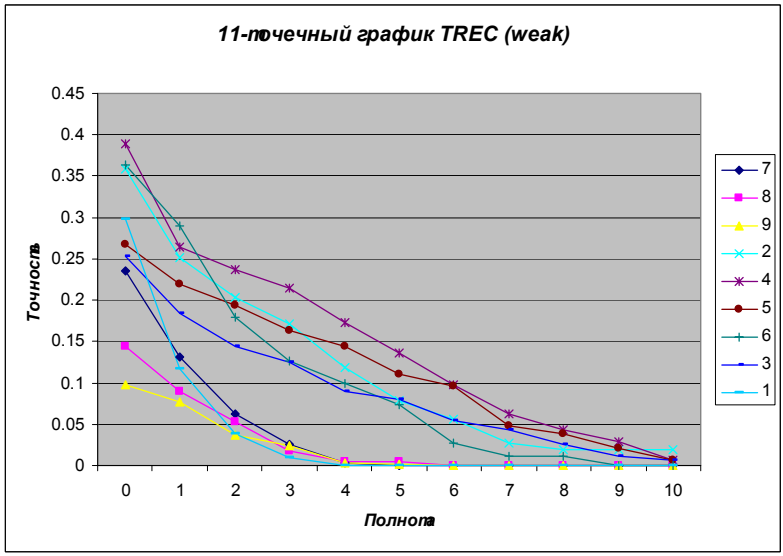
Таблица 1. Сводные результаты по дорожке поиска (сильная таблица релевантности)

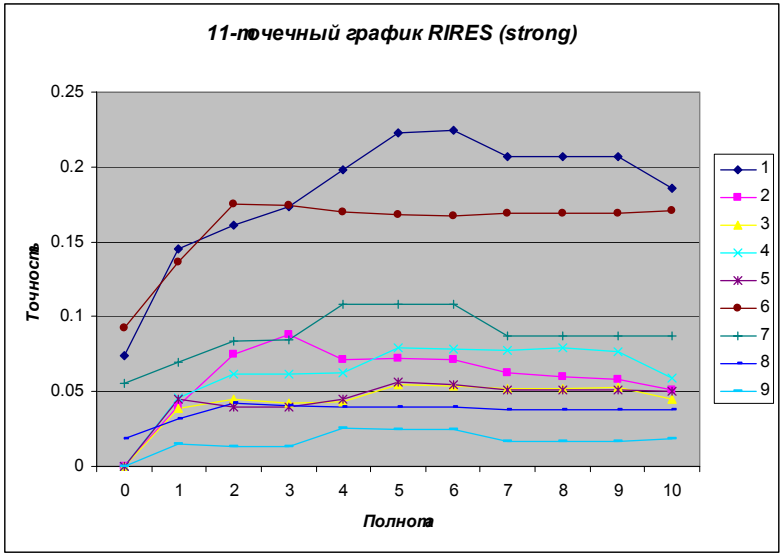
	P	*R*	*averageP*	R_p	P_{10}	P_5
1	0.17	0.06	0.17	0.05	0.09	0.14
2	0.15	0.42	0.16	***0.15***	0.15	0.16
3	0.11	***0.59***	0.10	0.10	0.09	0.10
4	0.16	***0.59***	0.16	***0.15***	***0.16***	0.16
5	0.14	0.55	0.13	0.13	0.10	0.10
6	***0.21***	0.25	***0.22***	0.13	0.14	***0.18***
7	0.12	0.07	0.13	0.05	0.07	0.08
8	0.07	0.10	0.07	0.05	0.05	0.05
9	0.06	0.11	0.05	0.04	0.04	0.03

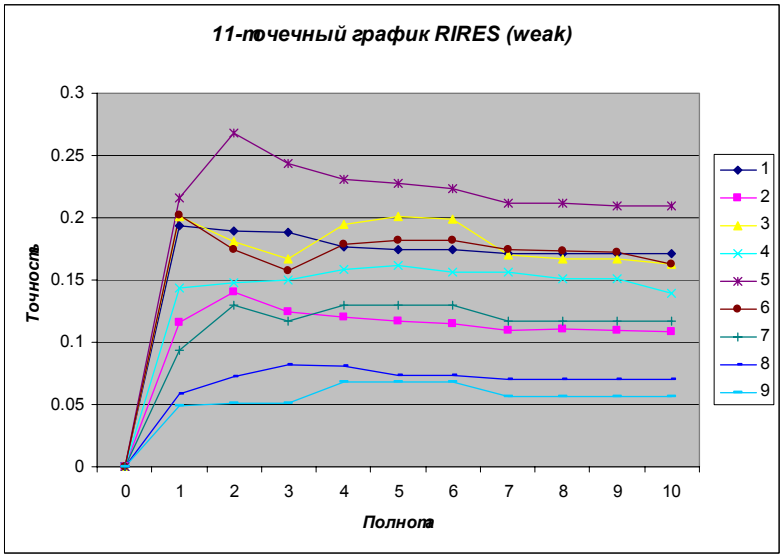
Таблица 2. Сводные результаты по дорожке поиска (слабая таблица релевантности)

Необходимо отметить, что оценки в 11-точечном графике TREC завышены в силу дефекта вычисления оценки. Дело в том, что для семи запросов в строгой таблице релевантности не было ни одного релевантного документа, и в этом случае значение точности на любом уровне считалось единичным, что и обусловло завышение результата на 0.129.


11-точечный график TREC (strong)
Точность
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0 1 2 3 4 5 6 7 8 9 10
Полнота
1
2
3
4
5
6
7
8
9


11-точечный график TREC (weak)
Точность
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0 1 2 3 4 5 6 7 8 9 10
Полнота
7
8
9
2
4
5
6
3
1


11-точечный график RIRES (strong)
Точность
Полнота
0
0.05
0.1
0.15
0.2
0.25
0 1 2 3 4 5 6 7 8 9 10
1
2
3
4
5
6
7
8
9


11-точечный график RIRES (weak)
Точность
Полнота
0
0.05
0.1
0.15
0.2
0.25
0.3
0 1 2 3 4 5 6 7 8 9 10
1
2
3
4
5
6
7
8
9

4. Дорожка по классификации

Эта дорожка посвящена задаче классификации Веб сайтов.

В 2003 году для участия в дорожке было подано 7 заявок, но добрались до финиша лишь 4 участника, которые предоставили организаторам результаты 5 прогонов.

4.1. Правила проведения

Официальная постановка задачи выглядела следующим образом:

> Задан список категорий, обучающая выборка и множество сайтов (не документов!). Надо присвоить каждому из сайтов коллекции категорию из этого списка с учётом обучающей выборки.
>
> Один и тот же сайт может относиться сразу к нескольким категориям. Поэтому ответом является упорядоченный список (до 3-5 категорий) для каждого из классифицируемых сайтов.
>
> Отметим, что сайт может не относится ни к одной из категорий и в этом случае идеальным ответом является пустой список назначенных категорий

Проведение этой дорожки также основывалось на наборе данных narod.ru. Множество классов сформировано на основе категорий второго уровня каталога narod.ru (catalog.narod.ru), который основан на принципе “модерируемого самоввода” — создатели сайтов сами выбирают категории для включения сайтов, но все заявки модерируются редакторами каталога. Тем не менее контроль за качеством данных в каталоге конечно же не очень жесткий.

Обучающая выборка была построена путем пересечения списка сайтов уже присутствующих в каталоге narod.ru и списка сайтов попавших в коллекцию используемую РОМИП’2003.

После этого из списка рассматриваемых классов были удалены те, для которых было доступно менее пяти обучающих примеров. Итоговый список классов состоял из 164 элементов.

Оценка результатов прогонов участников также была основана на методе “общего котла”, как и для дорожки по поиску (см. 3.1). В этом случае котел формировался для каждой категории и содержал все сайты, которые были отнесены к этой категории хотя бы одной из систем.

4.2. Выбор заданий для оценки

При отборе оцениваемых категорий учитывались следующие соображения:

- *Общий объем — порядка 3000 пар документ/категория*: из-за ограниченности доступных ресурсов суммарный объем оцениваемых котлов ограничивался цифрой в 3000 элементов (при сборе двух независимых оценок — это 6000 оценок).
- *Широта охвата*: хотелось по возможности расширить множество категорий для которых оценивалось качество полученных результатов.
- *Разумная плотность выбранного подмножества*: возможность проанализировать насколько хорошо различаются близкие категории весьма полезна при анализе результатов. разумный анализ результатов репрезентативность категорий первого уровня:
- *Возможность внятно описать категорию для асессора*: в отличии от эксперта асессор не знает о существовании других категорий и поэтому из описания должно быть понятно для каких сайтов эта категория лучше всего подходит, а для каких вероятно есть более подходящие варианты.

Процедура выбора была устроена следующим образом. Для всех 164 категорий были вычислены размеры общих котлов и из списка были удалены все категории, к которым было отнесено более 500 сайтов. Трое независимых экспертов случайным образом выбирали из оставшихся в списке категорию и пытались подготовить описания для асессора. Если это удавалось, то эта категория вносилась в множество кандидатов. Таким образом было выбрано 22 категории.

На втором этапе из множества исключили категории, которые были единственными представителями соответствующих категорий первого уровня. Из оставшхся отбирали категории для оценки, стараясь чтобы было по 3-4 представителя для каждой выбранной категории первого уровня и общий объем работы не превысил указанных ограничений.

Всего было отобрано 17 категорий, полный список которых приведен в “Приложении В” этого сборника.

4.3. Сбор оценок асессоров

Для оценки результатов асессоры использовали тот же инструмент, что и для дорожки поиска (см. раздел 4.3), однако в этом режиме работы был ряд важных отличий:

- Оценка выставлялась одна на весь сайт, а не для каждого документа;
- В навигационной панели страницы с одного сайта групировались вместе;
- Была разрешена навигация по локальным ссылкам, если соответствующие документы были включены в коллекцию;
- Не было подсветки слов.

Отметим, что при оценке результатов по классификации сайтов технические проблемы с отображением страниц возникали значительно реже (вероятно потому, что средний размер документов был значительно меньше) и поэтому для дорожки классификации все оценки были получены при использовании версии инструмента, которая не использует IE.

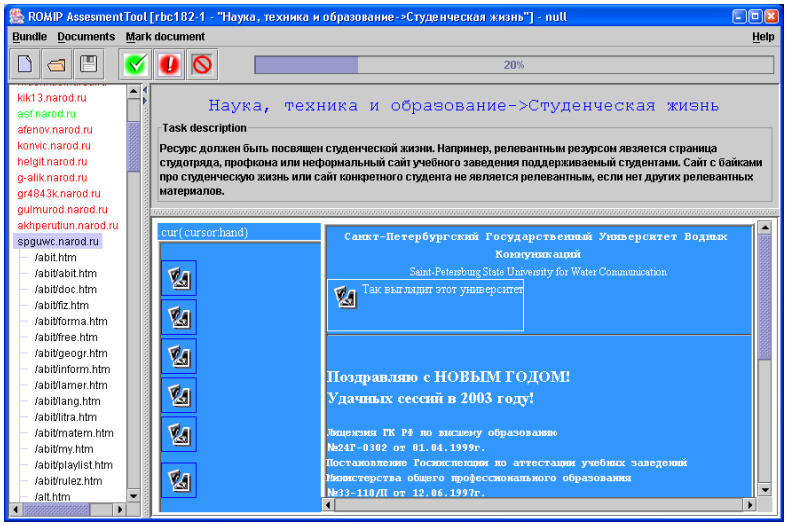


Рисунок 2. Инструмент асессора — дорожка классификации.

Всего оценивалось 3060 пар сайт/категория и для каждой пары запрашивалось не менее двух оценок. В результате мы собрали 6402

оценки — 1268 “релевантно”, 5066 "не релевантно” и 68 “невозможно оценить”.

Всего в оценке приняло участие 6 асессоров — 4 от оргкомитета и 2 от участников, самостоятельно производивших оценку результатов. Нагрузка на одного асессора варьировалась в пределах от 401 до 2130 пар сайт/категория. Предварительная оценка производительности в 30 пар сайт/категория в час (она оказалась довольно точной и отражающей реальные затраты) и общий объем работы составил 213,4 человеко-часа.

Инструкция для асессоров выглядела следующим образом:

Инструкции для асессоров
(дорожка по классификации)

Задачей асессора является выяснить “Относится ли содержимое данного ресурса к заданной тематической области?”.

При оценке ресурса вам будет показана его стартовая страница, а также список всех страниц (на левой панели), относящихся к этому ресурсу. Для принятия решения вы должны просмотреть достаточное для уверенной оценки количество страниц ресурса, выбирая их из списка или переходя по ссылкам на текущей странице.

Названия страниц в списке слева содержат полные имена страниц, включая имена директорий, — пожалуйста, попытайтесь посмотреть хотя бы по 1-2 странице из каждой директории (или группы страниц со схожими названиями), чтобы лучше представить его содержание.

Сколько страниц смотреть — решать вам, но цель — оценка ресурса в целом, а не только 1-2 его страниц. Помните, что ресурс может быть разнородным по содержанию. Некоторые представленные в наборе сайты не имеют стартовой страницы — в этом случае вы увидите сообщение об ошибке при показе страницы. Однако, этого недостаточно чтобы объявить сайт не оцениваемым — попробуйте посмотреть еще несколько страниц с этого сайта, выбрав из списка в левой панели.

Для того, чтобы ресурс считался относящимся к заданной теме необходимо, чтобы "заметная" часть его содержимого была ей посвящена (например, логический раздел). Для маленьких ресурсов большая часть материалов должна относится к заданной теме, для более крупных сайтов требуемая доля конечно же меньше.

Вообразите, что вашей задачей является сбор материалов по данной теме для того, чтобы составить каталог полезных ссылок.

Для того чтобы четче описать границы тематической области все задания содержат поясняющий комментарий.

Шкала оценки:

Релевантный (RELEVANT)

Ресурс относится к тематике.

Не релевантный (NOT RELEVANT)

На ваш взгляд ресурс не может быть отнесен к данной тематике.

Невозможно оценить(CAN NOT JUDGE)

Эта оценка должна использоваться в том случае, если вы не можете оценить содержимое сайта, поскольку вы не можете прочитать его содержимое (ни одну из страниц).

4.4. Таблицы релевантности

Как и в случае дорожки поиска мнения экспертов значительно расходились (второй асессор подтверждал положительную оценку первого лишь в 37% случаев) и мы строили две таблицы релевантности — сильную и слабую (см. раздел 3.4).

В случае слабой согласованности 906 сайтов были признаны правильно классифицированными, а в случае строгой — лишь 338. В обоих случаях 31 сайт был признан “не оцениваемым” и для 2120 случаев эксперты согласились, что решение было неправильным.

4.5. Метрики для вычисления очков

Для вычисления оценок прогонов по дорожке классификации использовались не только общие метрики — полнота и точность (см. раздел 3.5), но также и некоторые другие популярные в задаче классификации метрики [3].

В частности, вычислялись значения функции F_1. В общем виде функция F_α представляет собой комбинированную оценку полноты и точности ответа:

$$F_\alpha = \frac{(\alpha + 1) \cdot rp}{\alpha p + r}$$

F_1 — это наиболее часто используемая функция из этого семейства, равноправно трактующая важность критериев полноты и точности.

Кроме этого для каждой метрики вычислялись усредненные значения с использованием как микро, так и макроусреднения [3]. При микроусреднении, вычисленные оценки получаются в результате усреднения оценок на конкретных запросах. При макроусреднении же, напротив, сначала вычисляются суммарные базовые параметры

(количество релевантных документов в ответе, общее количество документов и т.д.) а затем на этих данных строятся окончательные оценки. (более формального описание можно найти, например, в [3])

4.6. Сводные результаты

Сводные оценки результатов прогонов участвовавших систем приведены в таблицах 3 и 4. В отличие от дорожки поиска в один из прогонов выглядит заметно лучше остальных, уступая другим лишь по макроусредненному значению точности в случае сильной таблицы резевантности.

	Точность	*Полнота*	F_1	*Точность (макро)*	*Полнота (макро)*	F_1 *(макро)*
1	0.13	0.18	0.11	0.10	0.27	0.14
2	0.08	0.09	0.09	0.14	0.13	0.14
3	0.13	0.08	0.10	***0.21***	0.10	0.13
4	0.07	0.14	0.09	***0.21***	0.14	0.10
5	***0.15***	***0.51***	***0.22***	0.14	***0.58***	***0.23***

Таблица 3. Результаты классификации (строгая таблица релевантности)

Отметим, что значительной сложностью в задачи дорожки по классификации явилось невысокое качество обучающей выборки. Для проверки качества выборки мы включили ее в оцениваемые котлы и лишь 33% элементов выборки были признаны релевантными всеми асессорами, и еще 33% были признаны релевантными частью асессоров.

	Точность	*Полнота*	F_1	*Точность (макро)*	*Полнота (макро)*	F_1 *(макро)*
1	0.28	0.21	0.20	0.25	0.26	0.25
2	0.20	0.08	0.11	0.27	0.09	0.13
3	0.28	0.06	0.10	***0.38***	0.06	0.11
4	0.15	0.13	0.10	0.20	0.12	0.15
5	***0.38***	***0.55***	***0.42***	***0.38***	***0.56***	***0.45***

Таблица 4. Результаты классификации (слабая таблица релевантности)

5. Наблюдения и планы

Методология и процедура проведения семинара РОМИП в 2003 году была не лишена недостатков. На то есть ряд причин — "быстрый" старт семинара, "пилотность" проекта и, как следствие, отсутствие опыта и как следствие недооценка объема работ как организаторами, так и у большинством участников. В частности, это обусловило перенос некоторых сроков и то, что ряд участников не смог выполнить все планировавшиеся эксперименты.

Однако, несмотря на все сложности в организации семинара, мы, безусловно, довольны, что первый семинар РОМИП состоялся, и очень надеемся, что он окажется не последним.

Мы полагаем, что накопленные результаты работы семинара — тестовый корпус и результаты оценки, исходные тексты инструментов использовавшихся для оценки, а также тексты трудов необходимо сделать свободно доступными. Этот шаг в первую очередь нацелен на популяризацию идей РОМИП и расширение круга его будущих участников — не только за счет опытных специалистов в области информационного поиска, но также и начинающих/академических коллективов.

Литература

[1] П.И. Браславский, М.В. Губин, Б.В. Добров, В.Ю. Добрынин, И.Е. Кураленок, И.С. Некрестьянов, Е.Ю. Павлова, И.В. Сегалович. Инициативный проект Российского семинара по оценке методов информационного поиска (РОМИП). Труды "Диалог-2003", Протвино, 2003.

[2] ROMIP Web site, 2003. http://romip.narod.ru

[3] И. Кураленок, И. Некрестьянов. Оценка систем текстового поиска. Программирование.28(4):226-242, 2002. http://ir.apmath.spbu.ru/ru/papers.html

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

[5] Voorhees E. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proc. of the SIGIR'98*, pp. 315-323, Aug. 1998.

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

[7] Zobel J. How reliable are large-scale information retrieval experiments? In *Proc. of the SIGIR'98*, pp. 308-315, Aug. 1998.

[8] Buckley C., Voorhees E. Evaluating evaluation measure stability. In *Proc. of the SIGIR'00*, pp. 33-40, 2000.