

## Результаты первого Российского семинара по оценке методов информационного поиска (РОМИП-2003)

Б.В. Добров, НИВЦ МГУ, [dobroff@mail.cir.ru](mailto:dobroff@mail.cir.ru)  
И.С. Некрестьянов, СПбГУ, [nis@acm.org](mailto:nis@acm.org),  
И.В. Сегалович, ООО Яндекс; [iseg@yandex-team.ru](mailto:iseg@yandex-team.ru),  
В.И. Шабанов, Рамблер Интернет Холдинг, [vs@rambler-co.ru](mailto:vs@rambler-co.ru)  
<http://romip.narod.ru>

Инициатива РОМИП 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 общедоступной «канонической» русскоязычной коллекции текстов, запросов и оценок, с помощью которой будущие исследователи смогут настраивать и развивать свои системы. Методология проведения семинаров РОМИП основывается на передовом зарубежном опыте подобных мероприятий TREC, CLEF и т.п. Первый семинар проводился в 2003 году и состоял из двух дорожек – поиска и классификации на основе Веб коллекции.

Проведение семинара стало возможным благодаря объединению усилий разных организаций и компаний. В результате работы РОМИП'2003 отработана процедура организации, разработано ПО для сбора экспертных оценок, создан первый крупный русскоязычный тестовый корпус для оценки методов информационного поиска и классификации, подготовлены таблицы релевантности.

В этой работе описан опыт проведения и основные результаты работы семинара за 2003 год, а также план работы РОМИП на 2004 год.

### Введение

Решение проблем текстового поиска опирается на эффективную реализацию используемых лингвистических моделей. Поскольку понятия «эффективности поиска», «качество поиска» не определены, количественная оценка результатов поиска не тривиальна. Традиционный подход в подобных случаях – сравнение с «эталонным» результатом, в контексте задач поиска плохо применим из-за невозможности построения эталонного ответа, зависимости от конкретного набора данных и поискового задания. Поэтому в задачах поиска оценка обычно строится на основе аппроксимированного эталона [7].

Проведение оценки самостоятельно обычно сильно затрудняется ограниченностью доступных ресурсов. Более того, полученные результаты оценки имеют ограниченную полезность, поскольку, как правило, результаты невозможно повторно использовать или даже воспроизвести. Тем самым наблюдаемая эффективность не может быть сравнима с данными от других исследователей.

В качестве решения этой проблемы мировое научное сообщество рекомендует использовать для оценки широко доступные (и де-факто стандартные) наборы данных или организовывать форумы для совместной оценки. В мире существует ряд таких международных форумов по оценке методов информационного поиска и близкой тематике – TREC, CLEF, NTCIR, DUC, SUMMAC, TDT, MUC.

Инициатива РОМИП – это российская инициатива подобного рода, наиболее близкими являются TREC (Text Retrieval Conference, <http://trec.nist.gov>) и CLEF (Cross-Language Evaluation Forum, <http://clef.iei.pi.cnr.it:2002>). В рамках РОМИП предполагается регулярное проведение семинаров, каждый из которых посвящен оценке решения одной или нескольких задач текстового поиска. Создание нового семинара было обусловлено несколькими причинами:

- отсутствие публично доступных русскоязычных тестовых коллекций;
- малое представительство российских групп в работе существующих зарубежных форумов по оценке методов информационного поиска;
- как следствие, трудности с проведением исследований по вопросам, интересующим российские команды.

Проведение семинара по оценке требует наличия четкой методологии, а также материальных ресурсов необходимых для проведения оценки.

Организация подобного мероприятия в РФ усложняется пробелами в законодательстве и культуре обмена информацией. В частности, одним из важнейших вопросов являлось преодоление взаимного недоверия участников, которые могут являться прямыми конкурентами вне РОМИП. Поэтому, задачей минимум в 2003 году было *просто* провести первый в РФ семинар, и эта задача была решена объединением *доброй воли* участников.

Концепция проведения семинара РОМИП была представлена в трудах Диалог'2003 [2], поэтому здесь мы лишь вкратце остановимся на общей методологии проведения и больше внимания уделим опыту и результатам полученным за 2003 год.

## 1. Общая методология организации семинара РОМИП

Семинар РОМИП имеет циклическую природу. В рамках годового цикла из множества реализуемых проектов по созданию тестовых наборов выбираются один или несколько наборов, которые наиболее интересны участникам. Эти отобранные проекты реализуются, а по завершении этапа с учетом накопленного опыта и текущих приоритетов участников выбираются новые проекты.

Структурно семинар состоит из набора «дорожек» - секций, посвященных конкретным проектам (с фиксированной задачей и правилами оценки).

Важнейшим принципом РОМИП является совместное с участниками определение задач для оценки и формирование правил проведения оценки. Оргкомитет лишь координирует проведение секций.

### Структура годового цикла:

На подготовительном этапе определяется список участников, уточняется список рассматриваемых задач и методология создания тестовых коллекций и оценки. Оговариваются форматы и способы обмена данными, официальные метрики для оценки. Фиксируется график проведения. Все участники получают псевдонимы, которые будут использоваться для анонимной оценки и публикации результатов.

Оргкомитет формирует тестовые наборы данных, заданий и распространяет их участникам. В зависимости от происхождения данных может требоваться оформление соглашения о нераспространении и ограничении возможностей использования набора участником.

Участник самостоятельно и на своем оборудовании выполняет поисковые задания.

Оргкомитет организует проведение оценки (с использованием независимых экспертов) полученных ответов. Конкретная методология оценки зависит от рассматриваемой задачи и определяется на подготовительном этапе. Информация о всех оценках будет доступна всем участникам, но эта информация будет использовать псевдонимы для ссылок на участников.

Участники самостоятельно анализируют полученные результаты и готовят статью, описывающую (общие) принципы их подхода и наблюдаемые результаты. При этом не обязательно раскрывать свое инкогнито и все детали реализации (это зависит от доброй воли участника) - достаточно в общих чертах описать какие известные методы использовались и что отличает их подход от других. Предоставление более подробной информации о системах, результатах и проблемах приветствуется.

Подготовленные статьи представляются на очном семинаре и публикуются в его трудах.

Процедура оценки различается для различных задач информационного поиска и формируется для конкретных дорожек, 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.

Для участия в семинаре необходимо подать заявку в оргкомитет. Пока не решен вопрос с внешним финансированием, участники платят вступительный взнос (компенсирующий начальные затраты на создание, распространение наборов данных, проведение оценки), а также подписывает необходимые соглашения (лицензии).

Участник свободен в определении набора дорожек, в которых он хочет участвовать, и может напрямую влиять на правила проведения этих дорожек во время их формирования. Приветствуется также предложение новых вариантов дорожек на общее обсуждение.

Результаты работы семинара в целом и каждой из дорожек публично доступны, как в виде трудов семинара, так и используемых корпусов и наборов заданий, а также построенных таблиц релевантности и созданных инструментов. Большинство этих материалов публикуется на сайте семинара (<http://romip.narod.ru>), а доступ к корпусам РОМИП можно получить после обращения в оргкомитет и подписания необходимых соглашений.

## 2. РОМИП в 2003 году

В 2003 году семинар РОМИП состоял из двух дорожек - поиска и классификации, для участия в которых были поданы заявки на 9 систем, а дошло до финиша 7. Общая координация и организация процесса оценки производилась исследовательской группой из СПбГУ ([ir.apmath.spbu.ru](http://ir.apmath.spbu.ru)).

Название системы	Дорожка по поиску	Дорожка по классификации
Алхимик [4]	1	1
ИС "Кодекс" [3]	1	
Ключи к Тексту [5]		1
ML Классификатор 2.0 [10]		2
RCO Classifier & Russian Context Server [9]	1	1
УИС РОССИЯ [1]	3	
Яндекс.Software 3.0 [8]	3	

Таблица 1. Распределение количества прогонов по участникам для каждой из дорожек.

### 2.1. Корпус narod.ru

Набор данных является одной из важнейших составляющих процесса оценки. От него напрямую зависят наблюдаемые результаты. Существует два принципиальных подхода к созданию набора данных – использование «естественной» коллекции целиком (например, архива газеты) или формирование новой коллекции. [7, 14]

В качестве основы для тестового набора данных использовалась выборка из примерно 3% сайтов домена *narod.ru*, предоставленная компанией Яндекс. Сайты этого домена распределены по нескольким физическим серверам случайным образом, поэтому был выбран один из серверов и отобраны все сайты, которые хранились на нем. При этом учитывались только HTML файлы и не рассматривались сайты, которые были созданы исключительно на основе стандартных шаблонов оформления страниц, которые предлагаются на [www.narod.ru](http://www.narod.ru). Сайты, созданные только и исключительно при помощи шаблонных страниц, составляют от 30 до 40 процентов на [www.narod.ru](http://www.narod.ru), их специфика (единообразная структура, повторяющиеся слова и выражения) сильно отличает их от средних характеристик глобальной Веб коллекции, которые пытались эмулировать организаторы.

По своей природе домен *narod.ru* содержит сайты различных тематик и тем самым является неплохим отражением Рунет в целом. Отметим, что сайты в домене *narod.ru* относительно редко ссылаются друг на друга. Используемая нами процедура построения коллекции еще более усугубляет ситуацию и поэтому гипертекстовая структура построенного набора конечно же имеет отдаленное сходство со структурой ссылок в Рунет.

Документы в наборе кроме HTML-содержимого содержат только информацию о URL откуда была получена эта страница и внутренний идентификатор документа в рамках коллекции РОМИП.

Построенная коллекция содержит более 728000 HTML страниц с 22000 сайтов общим объемом более 7 Гб (4,5 миллиона словоформ, более 130 миллионов уникальных пар словоформа/документ). Размер ее по отношению ко всему российскому сегменту Интернет невелик, по видимому, всего лишь 0.12-0.30%, так как число уникальных, то есть «сильно» отличающихся, страниц в Рунет составляет по самым широким оценкам от 250 до 600 миллионов.

### 3.2. Дорожка по поиску

Дорожка по поиску посвящена классической проблеме поиска по запросу и методология ее проведения во многом опиралась на многолетний опыт проведения дорожки «ad-hoc retrieval» в рамках TREC [16, 18]. В частности сбор экспертных оценок был организован на основе подхода "общего котла".

Источником запросов послужил журнал запросов поисковой системы Яндекс. Из него было отобрано 15000 запросов, которые и выполнялись каждой из систем участниц. Отбиралась русскоязычные запросы без опечаток, не использующие ненормативную лексику и операторы языка запросов Яндекс. Запросы, в основном, представляли короткие двух-трехсловные выражения, при этом примерное распределение запросов в выборке по длине и прагматике соответствовало общему потоку.

Требование получения ответов для столь большого числа запросов является отличительной особенностью РОМИП. Такой подход позволяет понизить вероятность подстройки систем под конкретные задания, за что иногда критикуют аналогичные конференции.

Для каждого из 15000 запросов система могла вернуть до 100 упорядоченных по убыванию релевантности документов, из которых первые 50 участвовали в дальнейшей оценке. Оценка результатов проводилась ассессорами из числа участников и организаторов для случайно отобранных, заранее неизвестных участникам, 54 запросов, при этом оценки дублировались.

Каждый запрос был дополнен экспертной интерпретацией, снимающей неоднозначность в трактовке отобранного запроса. Особенностью трактовок в цикле РОМИП-2003 была общая ориентация на понимание релевантного ответа как некоммерческого и общепознавательного. В сущности планки оценки результатов были завышены организаторами сознательно, с целью сохранять логику в перспективе будущего развития РОМИП

Полученные итоговые результаты проиллюстрированы на рисунке 1 в виде 11-точечных графиков TREC. Более подробная информация по отдельным запросам доступна в трудах РОМИП'2003 [11].

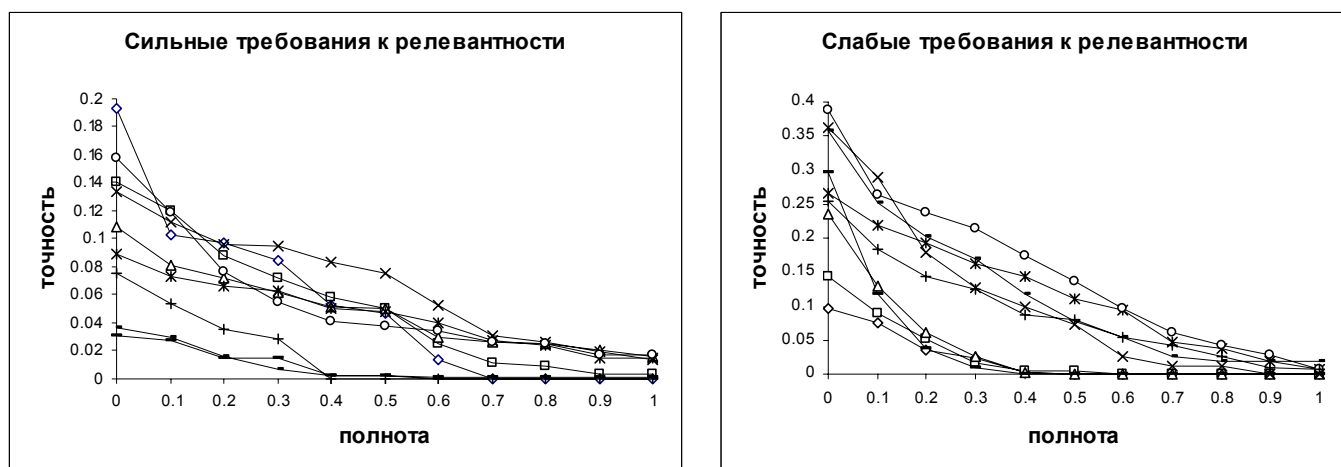


Рисунок 1. Итоговый 11-точечный график TREC для каждого из прогонов дорожки по поиску.

### 3.3. Дорожка по классификации

Эта дорожка посвящена задаче тематической классификации Веб сайтов. Насколько нам известно, это первая попытка оценить эффективность решения этой задачи в рамках форума по оценке систем информационного поиска. Формально сайтом в данном эксперимент считалось полное и неделимое содержание домена третьего уровня на narod.ru.

Специфика традиционной задачи тематической классификации для веба состоит в том, что львиная доля веб-страниц носит служебный, управляющий или навигационный характер, и не содержит тематических признаков *per se*. В то же время задача выявления тематики страниц остается актуальной для отнесения документов к тем или иным тематическим рубрикам при поиске, для повышения качества анализа веб и хост-графа, для логической сегментации доменов и поддоменов и многих других задач. В сущности веб-сайт в нашей постановке рассматривается как один сложный гетерогенный документ, созданный единым автором или группой авторов, и следовательно, содержащий все необходимые тематические признаки.

В основе дорожки по классификации РОМИП'2003 кроме корпуса *narod.ru* лежит саморегулируемый каталог *narod.ru*. Процедура отнесение сайта к рубрике происходит по заявке автора сайта, с последующим просмотром странички модератором и верификацией заявки. При этом разрешается отнесение сайта не более чем к двум рубрикам. Рубрики второго уровня из этого каталога составили основу таксономии, а сайты, которые были отнесены к этим рубрикам в каталоге *narod.ru*, сформировали обучающее множество. Так как, в корпус *narod.ru* представляет собой подмножество домена *narod.ru*, то не все упоминавшиеся в каталоге сайты присутствовали в корпусе. Поэтому из таксономии были удалены рубрики, для которых было менее 5 обучающих примеров. Итоговая таксономия состояла из 163 рубрик.

По условиям рассматриваемой задачи, один и тот же сайт мог принадлежать нескольким разным категориям. Поэтому для каждого из 22000 сайтов система могла вернуть список рубрик к которым она отнесла этот сайт длиной до 5 элементов.

Оценка производилась для 17 категорий, которые были заранее неизвестны участникам. Согласно инструкции для ассессоров сайт считался релевантным, если хотя бы одна из его страниц относилась к заданной категории. В таблице 2 представлены сводные оценки для каждого из 5 прогонов, более подробную информацию о результатах можно найти в трудах РОМИП [11].

Относительно невысокий уровень полученных результатов объясняется сложность рассматриваемой задачи. В дополнение к нетрадиционной для многих систем постановке задачи и высокой степени разнородности классифицируемой информации в процессе работы РОМИП всплыл ряд сложностей с обучающей выборкой:

- выборка содержала сайты, относящиеся к нескольким рубрикам;
- выборка была в значительной степени некорректной – из 320 сайтов обучающей выборки, которые были проверены в процессе оценки, 119 (37%) были признаны не релевантными;
- из 205545 страниц в выборке лишь 182022 являлись уникальными;
- многие сайты содержали разделы, не относящиеся к основной тематике: сборники анекдотов, рефераты и т. д.

Низкое качество обучающей выборки обусловлено природой каталога *narod.ru*, модераторы которого зачастую не анализируют содержание всего сайта, ограничиваясь просмотром одной двух страниц, да и смена содержания и, как следствие, темы сайта или его страниц не отслеживается систематическим образом.

Таким образом, в данной дорожке тестировались не только способность систем-участников к классификации и обучению на заданных образцах, но и способность фильтровать дубликаты, исключать из рассмотрения малозначимые части сайтов, выявлять и удалять элементы дизайна и навигации. В планы РОМИП входит введение дополнительной дорожки по классификации нормативных документов, в которой указанные выше сложности не так заметны.

Номер прогона	Микроусреднение			Макроусреднение		
	Точность	Полнота	$F_1$	Точность	Полнота	$F_1$
<b>Сильные требования к релевантности</b>						
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	<b>0.21</b>	0.10	0.13

4	0.07	0.14	0.09	<b>0.21</b>	0.14	0.10
5	<b>0.15</b>	<b>0.51</b>	<b>0.22</b>	0.14	<b>0.58</b>	<b>0.23</b>
<b>Слабые требования к релевантности</b>						
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	<b>0.38</b>	0.06	0.11
4	0.15	0.13	0.10	0.20	0.12	0.15
5	<b>0.38</b>	<b>0.55</b>	<b>0.42</b>	<b>0.38</b>	<b>0.56</b>	<b>0.45</b>

Таблица 2. Сводные результаты дорожки по классификации.

### 3.4. Оценка

Для проведения оценки было решено использовать метод «общего котла» (pooling), который активно используется в TREC [16]. «Общий котел» - это объединенное множество первых  $N_q$  документов из выдачи каждой из систем для данного запроса  $q$ . Для дорожки классификации вместо документов были сайты, а вместо запросов категории.

Каждый из элементов «котла» оценивался на соответствие запросу вручную как минимум двумя независимыми экспертами. Согласованность оценок экспертов в значительной мере зависит от того, насколько совпадают их понимания исходной информационной потребности [17, 19]. Как короткие запросы в дорожке поиска, так и названия категорий в дорожке классификации допускают множество возможных трактовок.

Для того, чтобы обойти эту проблему в РОМИП'2003 использовался следующий подход. Небольшое множество экспертов, отбирающих запросы для оценки, для каждого отобранного задания создают расширенную версию задания, которая содержит более детальное описание искомой информации, как его понимает эксперт<sup>1</sup> (тем самым уточняется одна из возможных информационных потребностей, выраженная этим запросом). Именно эта, «расширенная», версия и использовалась в дальнейшем для сбора оценок релевантности.

Для проведения оценки был разработан инструмент для сбора оценок ассессора. Мы стремились избежать шума, который может внести разные способы визуализации HTML документов программами просмотра на различных платформах, и поэтому использовали технологии java для обеспечения независимости от конкретной платформы. Отметим, что у ассессора есть возможность вернуться к уже просмотренному документу и изменить свое решение. Внешний вид разработанного инструмента представлен на рисунке 2.

Оценка проводилась по шкале «Релевантен»/«Не релевантен»/«Невозможно оценить». К последней группе относятся документы, которые не смогли быть визуализированы инструментом для оценки по техническим причинам.

Несмотря на принятые меры, собранные экспертные оценки в значительной мере расходились (что соответствует наблюдениям других исследователей [19]). Как наилучшим образом объединить оценки разных ассессоров в единый итоговый результат – открытая исследовательская проблема [19, 20]. В РОМИП'2003 использовались два способа построения таблиц релевантности. В случае таблицы «слабой» релевантности документ считался релевантным, если хотя бы один из оценщиков признал его таковым. Другая крайность – считать документ релевантным, только если все оценщики признали его таковым. Подобным образом строится таблица «строгой» релевантности.

<sup>1</sup> Для того чтобы различать разные роли в РОМИП используется термин «эксперт» для лиц, фиксирующих информационную потребность, и «оценщик» для лиц, реально производящих оценку руководствуясь расширенным заданием.

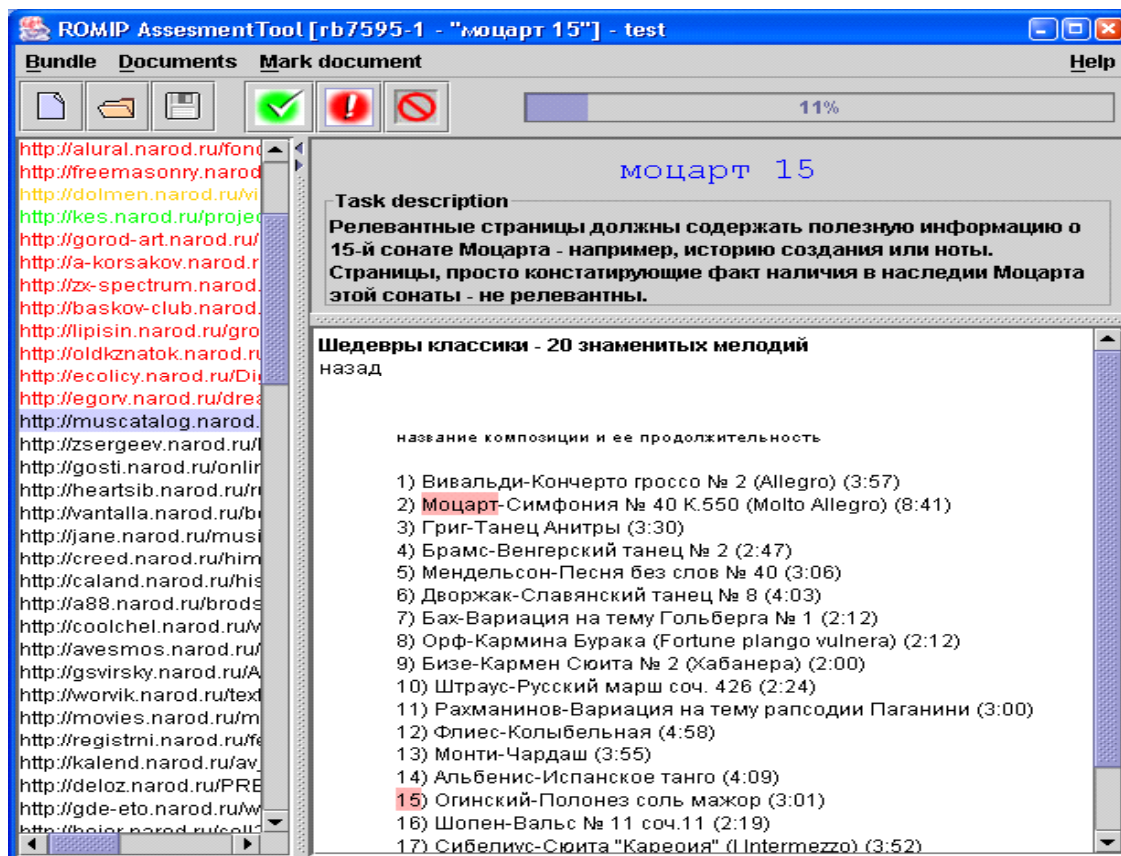


Рисунок 2. Инструмент сбора оценок (на примере дорожки по поиску)

Для вычисления итоговых оценок использовались стандартные меры:

- **Дорожка по поиску:** точность, полнота, средняя точность и 11-точечный график зависимости точности от полноты, используемый в TREC;
- **Дорожка по классификации:** точность, полнота и  $F_1$  с микро- и макро- усреднением.

Поскольку во многом мы следовали стандартным методам оценки систем информационного поиска, то несмотря на недостаточность накопленных нами результатов для самостоятельного анализа стабильности наблюдений, мы можем предполагать значимость наблюдаемых результатов исходя из анализа опыта оценки в TREC [13, 21]. Детальное описание процедуры оценки можно найти в трудах РОМИП [6].

	Хотя бы один эксперт признал релевантным	Все эксперты признали релевантным	Всего оценено
Дорожка по поиску	1187	391	10084
Дорожка по классификации	906	338	3060

Таблица 3. Распределение релевантных документов для каждой из дорожек.

### 3. Итоги и планы

Главной целью РОМИП в 2003 году было *просто* провести первый в России семинар, посвященный оценке методов информационного поиска для русскоязычных коллекций. Пробелы в законодательстве РФ и культуре обмена информацией осложняют достижение этой цели. Достижение этой цели стало возможным благодаря доброй воле участников.

В 2003 году был проведен комплекс работ, в результате которого:

- удалось объединить усилия разных организаций и продемонстрировать работоспособность методологии проведения семинара;
- создан первый крупный корпус русскоязычных документов для оценки методов информационного поиска;
- подготовлен набор заданий для двух задач информационного поиска – поиска по запросу и классификации;
- создано специализированное ПО для сбора экспертных оценок релевантности результатов;
- собрано более 27000 экспертных оценок, по которым построены таблицы слабой и сильной релевантности;
- определена политика распространения результатов.

Отметим, что построенный корпус *narod.ru* может быть интересен и для исследователей-лингвистов, поскольку он представляет собой корпус «живого» русского языка.

### 3.1. Планы РОМИП на 2004 год

В 2004 году программа семинара будет включать как уже апробированные в 2003 году дорожки поиска и классификации по Веб коллекции, так и новые дорожки с использованием новых коллекций. Повторение дорожек 2003 обусловлено рядом причин. В частности:

- в 2003 году участники по ряду причин применяли простейшие методы - чтобы гарантировано выполнить задание, не раскрывать своих секретов и т.п. Это позволило получить хороший стартовый уровень. В 2004 году некоторые участники планируют использовать новые методы, которые будет интересно сравнить с простыми;
- при анализе результатов РОМИП 2003 возникли интересные для всех участников РОМИП вопросы о природе «релевантности». В 2004 году хотелось бы продолжить исследование этого феномена на большей базе экспериментальных оценок.

В качестве новых дорожек предполагается включение дорожек поиска по коллекции нормативных документов и дорожки фактографического поиска. Использование коллекции нормативных документов даст возможность оценить эффективность методов поиска в «чистом» окружении. Окончательный список дорожек будет сформирован на основе поданных заявок на участие.

Еще одной важной задачей на 2004 год является расширение списка корпусов, которые могут быть использованы в работе последующих семинаров. Решение этого вопроса подразумевает обеспечение легальной возможности получения прав на использование в исследовательских целях РОМИП коллекций текстовых документов.

Семинар открыт для присоединения новых участников. К участию приглашаются все заинтересованные лица - как создатели поисковых систем, так и исследователи, занимающиеся проблемами информационного поиска, автоматической обработкой текстов.

### Благодарности

Проведение РОМИП'2003 стало возможным благодаря совместным усилиям ряда разных организаций и персоналий. Мы благодарны всем им за помощь. В частности, мы хотели бы отметить М.С. Агеева, А.В. Антонова, П.И. Браславского, М.В. Губина, В.Ю. Добрынина, И.Е. Кураленка, Е.Ю. Павлову, Н.В. Пантелееву, В.В. Плешко за их вклад в организацию РОМИП.

Отдельное спасибо оргкомитету RCDL'2003, оказавшему неоценимую помощь в организации очной части РОМИП, и ассессорам, без кропотливого труда которых безусловно бы ничего бы не получилось.

### Литература

1. Агеев М.С., Добров Б.В., Лукашевич Н.В., Сидоров А.В., Штернов С.В., «Отправная точка» для дорожки по поиску в РОМИП (предварительный анализ) // Труды РОМИП'2003, октябрь 2003, – СПб: НИИ Химии СПбГУ - стр. 87-109.



2. Браславский П.И., Губин М.В., Добров Б.В., Добрынин В.Ю., Кураленок И.Е., Некрестьянов И.С., Павлова Е.Ю., Сегалович И.В., Инициативный проект Российского семинара по оценке методов информационного поиска (РОМИП) // Труды Диалог'2003, июнь 2003.
3. Губин М.В., Опыт участия ИС «Кодекс» в РОМИП 2003 // Труды РОМИП'2003, октябрь 2003, – СПб: НИИ Химии СПбГУ - стр. 31-41.
4. Дерновой Г., Семантический анализ и РОМИП // Труды РОМИП'2003, октябрь 2003, – СПб: НИИ Химии СПбГУ – стр. 80-86.
5. Крейнес М.Г., Афонин А.А., Тихонов А.В., Технология смыслового анализа и поиска информации КЛЮЧИ К ТЕКСТУ // Труды РОМИП'2003, октябрь 2003, – СПб: НИИ Химии СПбГУ - стр. 72-79.
6. Кураленок И.Е., Некрестьянов И.С., Павлова Е.Ю., РОМИП 2003: Опыт организации. // Труды РОМИП'2003, октябрь 2003, – СПб: НИИ Химии СПбГУ - стр. 9-30.
7. Кураленок И., Некрестьянов И., Оценка систем текстового поиска. // Программирование. 2002, 28(4):226-242.
8. Маслов М., Алгоритм поиска Яндекс // Труды РОМИП'2003, октябрь 2003, – СПб: НИИ Химии СПбГУ - стр. 86-87.
9. Плешко В.В., Ермаков А.Е., Митюнин В.А., RCO на РОМИП 2003: отчет об участии в семинаре по оценке методов информационного поиска // Труды РОМИП'2003, октябрь 2003, – СПб: НИИ Химии СПбГУ - стр. 42-51.
10. Шабанов В.И., Андреев А.М., Метод классификации текстовых документов, основанный на полнотекстовом поиске // Труды РОМИП'2003, октябрь 2003, - СПб: НИИ Химии СПбГУ - стр. 52-71.
11. Труды РОМИП'2003 / Под редакцией И.С. Некрестьянова // СПб, НИИ Химии СПбГУ, 132 с.
12. Broder A., A taxonomy of web search. // ACM SIGIR Forum, 2002, 36(2):3-10.
13. Buckley C., Voorhees E., Evaluating evaluation measure stability // Proc. of the SIGIR'00 - pp. 33-40.
14. Cormack G.V., Palmer C.R., Clarke C.L.A., Efficient construction of large test collections // Proc. of the SIGIR'98 - pp. 282-289.
15. Dobrov B., Kuralenok I., Nekrestyanov I. and Segalovich I., Russian Information Retrieval Evaluation Seminar // In Proc. of LREC'04, April 2004 (accepted for publication).
16. Harman D., What we have learned, and not learned, from TREC // Proc. of the BCS IRSG'2000 - pp 2-20.
17. Mizzaro S., How many relevancies in information retrieval? // Interacting with Computers, 1998, 10(3), pp. 303-320.
18. Sparck Jones K., Reflections on TREC // Information Processing and Management, 1995, 31(3):291-314.
19. Voorhees E., Variations in relevance judgments and the measurement of retrieval effectiveness. // Proc. of the SIGIR'98 - pp. 315-323.
20. Wilbur J. W., The knowledge in multiple human relevance judgments. // TOIS, 1998, 16(2):101-126.
21. Zobel J., How reliable are large-scale information retrieval experiments? // Proc. of the SIGIR'98 - pp. 308-315.