# РОМИП'2004: отчет организаторов

© Игорь Кураленок, Игорь Некрестьянов

Санкт-Петербургский Государственный Университет
ik@oasis.apmath.spbu.ru, nis@acm.org
http://romip.narod.ru

**Аннотация**

Эта статья описывает основные принципы и правила семинаров РОМИП, а также детали организации РОМИП'2004, включая описание дорожек и коллекций, а также вопросы связанные со сбором оценок от асессоров и некоторые результаты экспериментов с используемой методологией оценки.

## 1. Введение

Необходимость проведения мероприятий по независимой оценке методов информационного поиска в России мотивирована отсутствием публично доступных русскоязычных тестовых коллекций и редкими случаями участия российских групп в работе существующих зарубежных форумов по оценке методов информационного поиска. Как следствие, есть трудности с проведением исследований по вопросам, интересующим российские команды.

Однако, организация подобного мероприятия требует нахождения ответов для целого комплекса вопросов, включая:

- Как преодолеть взаимное недоверие тех, кто конкурирует вне РОМИП?
- Откуда брать коллекции?
- Как предотвратить нецелевое использование коллекций?
- Где найти ресурсы необходимые для проведения оценки?
- Какие методологии оценки использовать и как повысить достоверность результатов?
- Как избежать превращения в форум для рекламы коммерческих систем?

В 2003 году благодаря объединению доброй воли участников нам удалось в сжатые сроки решить все эти и многие другие проблемы,

что позволило состоятся первому российскому семинару по оценке методов информационного поиска (РОМИП). По общему мнению участников семинар в целом прошел довольно успешно [1].

Успешное завершение первого семинара создало благоприятную атмосферу для его развития в 2004 году. В этом году увеличилось число рассматриваемых задач и участвующих систем, появилась новая коллекция нормативных документов и новая обучающая выборка для задачи классификации Веб сайтов.

Значительные усилия были направлены на укрепление организационной структуры РОМИП – разработана модель накопления коллекций, позволяющая накапливать коллекции для будущего использования в целях РОМИП; подготовлено формальное соглашение для передачи коллекций участникам РОМИП; получена поддержка РФФИ для проведения семинара в 2004 году.

## 2. Основы методологии РОМИП

Принципы организации РОМИП уже неоднократно описывались нами ранее [1,4,5,6], поэтому мы лишь вкратце остановимся на них в рамках этой статьи.

Семинар РОМИП имеет циклическую природу. Для каждого годового цикла из множества реализуемых проектов по созданию тестовых наборов выбираются один или несколько проектов, которые наиболее интересны участникам. 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.

Для участия в семинаре необходимо подать заявку в оргкомитет. Пока не решен вопрос с полным внешним финансированием, участники платят оргвзнос (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.

### 2.1. Структура годового цикла

В структуре годового цикла РОМИП можно выделить следующие этапы:

1. 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.
2. 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.
3. Участник самостоятельно и на своем оборудовании выполняет поисковые задания.
4. Оргкомитет организует проведение оценки полученных ответов (с использованием независимых асессоров). Конкретная методология оценки зависит от рассматриваемой задачи и определяется на подготовительном этапе. Информация о всех оценках доступна всем участникам, но при этом для ссылок на других участников используются псевдонимы.
5. 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 - 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.
6. Подготовленные статьи представляются на очном семинаре и публикуются в его трудах.

### 2.2. Принципы проведения оценки

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.

## 3. Коллекции

Ключевыми вопросами при создании коллекции являются не только тип, объем, происхождение и формат представления данных, но также и способ соблюдения авторских прав владельца данных. Для этой цели в РОМИП было решено использовать модель с «хранителем» коллекций, подобную принятой в похожих мировых инициативах. Так, например, для TREC таким хранителем выступает LDC, а для CLEF – ELDA.

В этой модели «хранитель» заключает договора с «правообладателями» коллекций на ее использование в рамках РОМИП. В дальнейшем участники заключают отдельные соглашения с «хранителем». На данный момент «хранителем» РОМИП выступает НИВЦ МГУ.

Отделение процесса накопления коллекций позволяет заготавливать коллекции «впрок», под будущие задачи РОМИП, и облегчает общение с правообладателями.

### 3.1. Веб коллекция Narod.ru

Эта та же коллекция, что использовалась в РОМИП’2003. Она содержит порядка 22000 сайтов (около 3%) из домена narod.ru по состоянию на март 2003 года. Всего в коллекции порядка 728000 отдельных страниц. Отметим, что гипертекстовая структура этой коллекции слишком разрежена и не отражает реальной структуры Веб.

Юридическая чистота коллекции обусловлена пользовательским соглашением компании Яндекс на использование сайтов в домене narod.ru.

### 3.2. Веб коллекция DMOZ

Коллекция, созданная на основе русскоязычной части каталога dmoz.org с целью получения разумного обучающего множества для задачи классификации Веб сайтов.

В коллекцию были включены русскоязычные сайты, упоминающиеся в категориях второго уровня (начиная отсчет с World→Russian), на страницах которых не было явного запрещения копирования содержимого этих сайтов. Для снижения размеров коллекции до разумных пределов для каждого сайта в коллекцию включалось не более 500 страниц полученных обходом в ширину, начиная со стартовой страницы.

Построенная коллекция состоит из 2100 сайтов и в общей сложности содержит порядка 300000 страниц.

### 3.3. Коллекция нормативных документов

Эта коллекция была предоставлена компанией Кодекс и состоит из примерно 61000 HTML документов. Общий объем коллекции – 1.6 Гб.

В отличии от Веб коллекций эта коллекция в значительной степени более однородна и содержит нормативно-правовые документы законодательства России.

Для предотвращения несанкционированного использования данных участники подписывали соглашение об использовании данных [7].

## 4. Задачи

Программа РОМИП’2004 состояла из 5 дорожек, каждая из которых была посвящена отдельной задаче.

### 4.1. Поиск по Веб коллекции (web adhoc)

Эта дорожка являлась повторением аналогичной дорожки РОМИП’2003. Поиск производился по той же коллекции Narod.ru, но число заданий было увеличено до 24250 и для каждого из них система могла вернуть до 100 результатов.

Задания отбирались из журналов поисковых систем Рамблер и Яндекс. Большое число заданий предотвращало возможность ручной настройки системы под конкретные запросы.

### 4.2. Поиск по колекции нормативных документов (legal adhoc)

Правила этой дорожки очень похожи на правила дорожки поиска по Веб коллекции. Однако, в этом случае поиск производился по коллекции нормативных документов и число заданий было порядка 13000.

Благодаря содействию компаний Кодекс и Парк.Ру задания отбирались из журналов поисковых систем, специализирующихся на поиске по нормативным документам.

### 4.3. Тематическая классификация Веб-сайтов (web classification)

Эта задача уже рассматривалась в РОМИП'2003, но многие участники жаловались на низкое качество использовавшегося обучающего множества. Поэтому в 2004 году было решено построить новое обучающее множество на основе dmoz.org.

Итоговая таксономия содержала 247 категорий и для каждой категории было не менее 5 обучающих примеров.

Задание состояло в классификации всех сайтов коллекции narod.ru. Для каждого сайта система могла вернуть от 0 до 5 категорий к которым он по ее мнению относится.

### 4.4. Тематическая классификация нормативных документов (legal classification)

В рамках этой дорожки классификации подвергались все документы из коллекции нормативных документов. Для каждого документа система могла выбрать до 5 категорий из 163 рассматриваемых.

Для обучения классификатора было предоставлено обучающее множество из 13772 (6294 уникальных) документов, но не менее 5 документов для каждой из категорий.

### 4.5. Поиск фактов по Веб-коллекции (QA)

Задания были построены на основе информации опубликованной на сайте «Кроссворд-Кафе» (http://dilet.narod.ru/). Мы извлекли информацию о днях рождения 5052 персоналий. При подготовке описаний использовалась помощь профессионального лингвиста, предоставленная компанией «Гарант-Парк-Интернет».

Описание задания содержало ФИО персоны, известные псевдонимы и краткое описание рода деятельности персоны, позволяющее точнее ее идентифицировать. Для примера приведем несколько таких описаний:

```
<task id="qa1109">
 <variant>Владимир Ильич Ленин</variant>
 <variant>Владимир Ильич Ульянов</variant>
 <description>
   вождь мирового пролетариата
 </description>
<task id="qa1202">
 <variant>Вячеслав Иванович Иванов</variant>
 <description>
   российский поэт-символист
 </description>
</task>
```

Задание для системы формулировалось следующим образом - найти все события связанные с конкретной персоналией. Для каждого такого события система должна была вернуть фрагмент текста длинной не более 300 символов, в котором описывается это событие, и указать местоположение этого фрагмента в исходном документе. Предполагалось также, что система может также классифицировать событие и отнести его к одному или нескольким заранее определенным типам событий.

## 5. Участники

Всего мы получили 11 заявок на участие в РОМИП'2004, но только 9 заявившихся систем дошло до финиша. Следующая таблица содержит информацию о поступивших заявках и количестве сданных вариантов ответов (прочерк означает, что заявка была подана, но участник либо решил отказаться от участия в этой дорожке, либо не предоставил результаты в срок).

Отметим, что 8 из подавших заявки коллективов ранее подавали заявки для участия в РОМИП'2003. По сравнению с прошлым годом увеличилось не только количество участников и заданий, но также значительно возросло число случаев представления более одного варианта ответов системой. Всего же было получено 34 варианта ответов (в 2003 году – 14).

| Система | Поиск по Веб коллекции | Класс. Веб сайтов | Поиск нормативных документов | Класс. нормативных документов | Поиск фактов |
|---|---|---|---|---|---|
| Галактика-Zoom | | 1 | | 1 | |
| Золушка | | 3 | | | |
| ИС “Кодекс” | 2 | | 3 | | - |
| ML Классификатор 2.0 | | - | | | |
| Ментал | | | | | - |
| mnogo-Search | 1 | | | | |
| RCO | 2 | 2 | 4 | 1 | 1 |
| Синдбад | | | | 1 | |
| Sophia | | | | 3 | |
| УИС РОССИЯ | 2 | - | 2 | 3 | |
| Яндекс.Server 3.2 | 1 | | 1 | | |

## 6. Оценка результатов

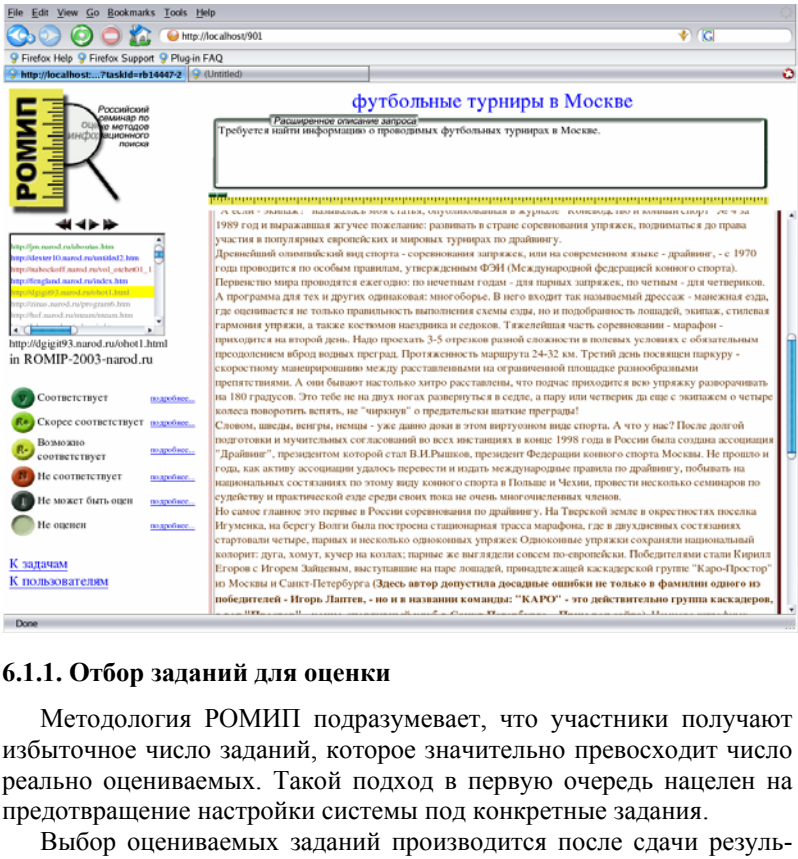
Целью процесса оценки результатов являлось построение таблиц релевантности, которые содержат информацию о правильных и неправильных ответов для каждого задания, а также вычисление итоговых официальных метрик, характеризующих качество работы системы.

### 6.1. Сбор оценок

Основным методом оценки результатов в 2004 году, как и в 2003, служили оценки асессоров, которые собирались методом «общего котла» (pooling), который используется в TREC [3, 8]. «Общий котел» — это объединенное множество первых $N_q$ ответов (дорожки $N_q$ – *«глубина»* котла) из выдачи каждой из систем для данного задания $q$. Каждый из документов попавших в такой котел далее оценивается экспертами на соответствие запросу.

Для обеих дорожек поиска глубина котла была равна 50, для дорожек классификации – 5, а для задачи поиска фактов учитывались все ответы системы.

Для сбора оценок асессоров использовался специализированный инструмент. Поскольку версия инструмента, использованная в 2003 году, вызвал массу нареканий по поводу производительности оценки (что было обусловлено неудачным выбором технологии визуализации HTML документов), то в этот раз использовалась новая версия, использующая стандартный Веб броузер для визуализации HTML (пример интерфейса для оценки дорожки поиска приведен на рисунке).




### 6.1.1. Отбор заданий для оценки

Методология РОМИП подразумевает, что участники получают избыточное число заданий, которое значительно превосходит число реально оцениваемых. Такой подход в первую очередь нацелен на предотвращение настройки системы под конкретные задания.

Выбор оцениваемых заданий производится после сдачи результатов всеми участниками и контролируется оргкомитетом. В 2004 году мы старались согласовывать выбор с участниками конкретных дорожек, чтобы повысить общую полезность оценки.

Обобщенная процедура выбора устроена следующим образом: выбирается подмножество всех заданий с «разумными» размерами котлов. На практике это означает отсечение заданий со слишком маленькими или очень большими котлами. Далее задания-кандидаты отбираются случайным образом и просматриваются человеком на предмет:

1. **Корректности**
   Отсекаются задания с грамматическими ошибками, использованием сокращений, ненормативной лексикой.

2. **Осмысленности формулировки**
   Задания рассматриваются с точки зрения их понятности для типичного асессора или возможности составить разумное расширенное описание. Это плохо формализуемый шаг и контролируется исключительно здравым смыслом того, кто их отбирает.

Процесс отбора продолжается пока общий объем котлов отобранных заданий не достигнет желаемого размера или не будет набрано желаемое число заданий.

| **Дорожка** | **Всего заданий** | **Число оцениваемых заданий** | **Источник заданий** |
|---|---|---|---|
| Поиск по Веб-коллекции | 24250 | 48 новых<br>19 повторно | Журналы Яндекс и Рамблер |
| Классификация Веб-сайтов | 247 | 38 | Таксономия dmoz.org |
| Поиск по коллекции нормативных документов | 13000 | 50 (“понятия”)<br>41 (“документы”) | Журналы Кодекс и Парк.ру |
| Классификация нормативных документов | 163 | 12 (асессоры)<br>40 (сравнение с эталоном) | Таксономия Кодекс |
| Поиск фактов | 5052 | 109 | «Кросворд-кафе» http://dilet.narod.ru/ |

Сводная информация об отобранных для оценки заданиях представлена в приведенной таблице. Полные списки отобранных для оценки заданий для каждой из 5 дорожек приведены в приложениях B-F.

Отметим, что при отборе заданий для дорожки поиска по коллекции нормативных документов на основе просмотра и анализа логов было выделено два основных типа запросов – запросы, нацеленные на поиск информации “на тему” (тип “Понятия”), и запросы, нацеленные на поиск “конкретного документа” (тип “Документ”). Поэтому отбор производился отдельно для каждой из категорий.

Отбор заданий для дорожки поиска фактов в основном производился среди заданий с размерами пулов от 10 до 20 ответов. Мы также постарались включить в число заданий некоторое количество заданий с относительно длинными описаниями (больше 3-4 слов) или задания с псевдонимами персон (всего примерно по 20 заданий каждого вида).

### 6.1.2. Формулировка заданий для асессоров

Основным подходом при формулировании заданий для асессоров было использование расширенных описаний. Расширенное описание задания используется для упразднения неоднозначности, детально описывая искомую информацию, как это понимает эксперт (тем самым уточняется одна из возможных информационных потребностей, выраженная этим запросом).

Расширенные описания заданий применялись для оценки в обеих дорожках классификации и дорожке поиска по Веб коллекции. В последнем случае была также проведена альтернативная оценка без использования расширенных описаний.

При составлении расширенных описаний для дорожки классификации использовались описания соответствующих категорий в исходных таксономиях (Кодекс и dmoz.org), если таковые были доступны. В противном случае описание составлялось самостоятельно с учетом подкатегорий рассматриваемой категории.

Для оценки дорожки поиска по коллекции нормативных документов было решено отказаться от использования расширенных описаний поскольку их составление требует знаний в прикладной области и хорошего знакомства с коллекцией. Поэтому асессорам было выдано общее описание задания для каждого из типов рассматривавшихся запросов:

- тип “Понятия”

> Представьте, что вам необходимо разобраться, что означает упомянутое понятие, какие есть нормативные документы, которые регламентируют вопросы, связанные с этим понятием.

- Тип “Документ”

> Целью поиска является упомянутый документ (или группа документов), а также документы, с ним связанные (дополнения, разъяснения, комментарии и т.п.). Отметим, что название искомого документа возможно не совсем точно сформулировано.

Для дорожки поиска фактов описание также было единым для всех заданий:

> Цель поиска - составить досье/биографическую справку на заданного человека, то есть найти _все_ события связанные с ним. Полностью релевантный ответ - это ответ содержащий описание и время возникновения события, а также ссылку на заданного человека (отсутствие части этой информации делает ответ частично релевантным). Если выделенный фрагмент текста не является идеальным (то есть можно выбрать его лучше), то необходимо также уточнить идеальные границы.

Оценка задания для этой дорожки проводилась в два шага – на первом шаге асессору показывалось описание задания и выделенный системой фрагмент, а на втором асессору показывался исходный документ и его просили уточнить идеальные границы фрагмента, описывающего факт.

### 6.1.3. Шкала оценки

В от РОМИП’2003 в этом году было решено использовать расширенную шкалу оценки со следующими градациями:

- **Соответствующий (релевантный/витальный)**
  Документ, который позволяет составить относительно полное представление о предмете или содержит ответ на поставленную поисковую задачу. Релевантный документ обязан содержать информацию, отвечающую на запрос в доступной форме, и быть Вам понятен. Еще одним критерием релевантного документа является его авторитетность, релевантными документами признаются лишь те, чья авторитетность не вызывает у Вас сомнения. Документы, ссылающиеся на искомую информацию, релевантными не являются.
- **Скорее соответствующий (релевантный+)**
  Документ, отвечающий теме запроса, но не содержащий полного ответа на поставленную поисковую задачу. Так же к этому классу относятся документы,содержащие полный ответ, но по

каким либо причинам непонятные (не путать с нечитаемыми) или не внушающие доверия. Если документ не содержит необходимой информации, но ссылается на ресурс, соответствующий запросу, то он тоже считается относящимся к описываемому классу при условии легкости нахождения этой ссылки.

- **Возможно соответствующий (релевантный-)**
  Документ, возможно не отвечающий теме запроса, но содержащий отдельные частицы информации, которые могли бы помочь в решении задачи. Не отвечающий запросу документ, содержащий полезные ссылки также относится к этому классу.
- **Не соответствующий (нерелевантный)**
  Документ, не содержащий полезной информации для выполнения поисковой задачи и не содержащий полезных ссылок.
- **Документ не может быть оценен**
  Документ не читается (представлен в некорректной кодировке или написан на непонятном языке), вызывает технические проблемы в браузере или не может быть оценен по каким-либо другим объективным причинам.

### 6.1.4. Дублирование оценок

Для того чтобы снизить влияние субъективности оценки мы собирали как минимум по две независимые оценки на пару задание-ответ (на большее не хватило ресурсов). Однако, в этом году содержимое котлов не делилось между разными асессорами, т.е. один и тот же асессор оценивал все ответы, включенные в котел.

Отметим также, что в силу организационной накладки для дорожки поиска по Веб коллекции было собрано более двух оценок для каждой пары.

### 6.1.4. Создание таблиц 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 связи с использованием в 2004 году расширенной шкалы оценок способ построения итоговых таблиц претерпел некоторые изменения по сравнению с РОМИП’2003:

- **Слабые требования к релевантности (or)**
  В этом случае результат:
  - “релевантен”, если хотя бы одна оценка превышает минимальный порог релевантности;

- “невозможно оценить”, если все оценки “невозможно оценить”;
- в остальных случаях “не релевантен”.

- **Сильные требования к релевантности (and)**
  В этом случае результат:
  - “невозможно оценить”, если все оценки “невозможно оценить”;
  - “не релевантен”, если хотя бы одна оценка не превышает минимальный порог релевантности;
  - в остальных случаях “релевантен”.

При вычислении официальных результатов РОМИП’2004 в качестве минимального порога релевантности использовалось значение «релевантный-».

### 6.2. Использование зталоного результата

Компания Кодекс любезно предоставила нам эталонную классификацию коллекции нормативных документов, которую мы могли использовать не только для подготовки обучающего множества, но также и для оценки результатов классификации.

На основе этой классификации была построена эталонная таблица релевантности (ideal) для тех 12 категорий, что оценивались вручную асессорами.

Альтернативно, случайным образом было отобрано 40 категорий, для которых в обучающем множестве было не менее 10 документов, и было произведено сравнение ответов систем с эталонной классификацией для этих 40 категорий (ideal40).

### 6.3. Официальные метрики

На основе построенных таблиц релевантности вычислялись значения стандартных метрик, формальное описание которых можно найти в приложении А:

| Задача | Поиск | Классификация | Поиск фактов |
|---|---|---|---|
| **Метрики** | Precision(5)<br>Precision(10)<br>Precision<br>AveragePrecision<br>Recall<br>11-точечный график TREC | F1<br>Recall<br>Precision<br>Error<br>Accuracy | Precision |

## 7. Результаты

Из-за временных ограничений на подготовку этой работы и большого объема доступных материалов мы решили не заниматься попытками анализа эффективности выполнения заданий всеми системами, а остановиться лишь на некоторых организационных вопросах и предварительных результатах методологических экспериментов.

### 7.1. Планируемое и реальное расписание

В следующей таблице приведена информация о планировавшемся графике РОМИП и о реальных сроках завершения этапов:

| Этап | План | Реальное завершение |
|---|---|---|
| Прием заявок | 29 марта | 27 мая |
| Начало распространения заданий | 15 апреля | 16 апреля (кроме Веб классификации) 6 июня (Веб классификация) |
| Результаты прогонов от участников | 10 июня. После переноса сроков - 1 июля. | 12 июля |
| Результаты оценки | 2 августа | 27 августа (кроме QA), 8 сентября (QA) |
| Тексты докладов | 1 сентября | 15 сентября |
| Очная часть | 1 октября | 1 октября |

Несмотря на практически своевременное распространение заданий, сбор результатов и их оценка затянулись. Поскольку дата проведения очной встречи была заранее фиксирована, опоздание пришлось компенсировать за счет времени на анализ данных и подготовку статей, а также за счет времени на печать трудов.

### 7.2. Полезны ли расширенные описания?

Применявшаяся в 2003 году модель оценки с использованием расширенных описаний вызвала неоднозначную реакцию. Частично, это было обусловлено недостаточным качеством подготовки использовавшихся расширенных описаний, но были и принципиальные возражения против использования расширенных описаний.

Для прояснения ситуации в 2004 году был проведен эксперимент с проведением независимой оценки одной и той же дорожки как с использованием, так и без использования расширенных описаний. В качестве такой дорожки была выбрана дорожка поиска по Веб коллекции.

К сожалению мы не успели детально проанализировать полученные результаты. Отметим лишь некоторые первоначальные наблюдения. При оценке без использования расширенных описаний 2025 ответов были признаны слабо релевантными, но лишь 376 признано строго релевантными. При использовании расширенных описаний соответствующие числа – 1884 и 469. Это согласуется с ожидаемым результатом – расширенные описания четче описывают цель поиска и поэтому мнения асессоров чаще совпадают, но с другой стороны какие-то документы из-за этого пропускаются.

Однако, необходимо отметить, что хотя вычисленные на основе соответствующих таблиц оценки систем и отличаются, но выводы о превосходстве методов в значительной мере совпадают. Так, при использовании слабых требований к релевантности порядок систем отсортированных по значению полноты или точности не различается[1]. При использовании сильных требований к релевантности порядок совпадает для полноты, но несколько отличается для точности.

### 7.3. Эффективность использования «не экспертов» для оценки

При подготовке дорожек использующих коллекцию нормативных документов у оргкомитета и ряда участников были сомнения в осмысленности использования для оценки результатов асессоров-«не юристов», которые слабо ориентируются в специализированном материале этой коллекции. Однако, поскольку привлечение специалистов в прикладной области для оценки результатов в этом году было явно нереализуемой задачей у нас не оставалось другого выбора.

Благодаря компании Кодекс для оценки дорожки классификации нормативных документов мы могли использовать эталонную классификацию, создание которой контролировалось экспертами в прикладной области. Это позволило нам сравнить оценки полученные на основе эталонной классификации и на основе оценок наших асессоров.

К сожалению, при проведении оценки значительная часть документов упоминающихся в эталонной классификации не попала в

[1] Рассматривались оценки полученные на суженных до глубины пула ответах систем (pd50).

оцениваемые котлы (оценено лишь 369 из 826). Поэтому для этих документов мнение асессоров осталось неизвестным и мы можем провести лишь неполное сравнение.

Интересно, что 280 (75%) из 369 оцененных эталонных документов признаны релевантными с использованием слабых требований к релевантности, а 153 (41%) – с использованием сильных требований к релевантности. Показательно также, что степень взаимного согласия асессоров на совпадающем подмножестве документов составила 54%, что заметно выше, чем в среднем для этой дорожки (29%).

Тем не менее, сравнивая итоговые результаты систем следует отметить, что прямой эквивалентности между этими способами оценки по-видимому нет. При вычислении полноты первые пять лучших прогонов при оценке на основе эталона и при использовании строгих требований к релевантности совпадают (при использовании слабых требований совпадают только первые два). Однако, при измерении точности различия при использовании сильных требований к релевантности есть уже в том, какой прогон наилучший. И хотя при использовании слабых требований согласованность лучше – совпадают два лучших прогона, да и в первой пятерке отличия минимальны, но все же сомнения остаются.

Более подробная информация о согласованности мнений между асессорами и эталонной классификацией по каждой из категорий приведена в разделе 5 приложения I.

### 7.4. Нужна ли расширенная шкала оценки?

Еще один часто возникающий в обсуждениях вопрос – это вопрос о достаточности бинарной шкалы для сбора оценок релевантности от асессоров.

Классические формулы для наиболее популярных в области информационного поиска мер для оценки, таких как точность, полнота и др., опираются на использование бинарных оценок – «релевантен»/«не релевантен». Да и известные мировые инициативы по оценке систем поиска (TREC, CLEF и т.п.) используют бинарную шкалу.

С другой стороны деление на «черное»/«белое» ставит перед асессором психологическую проблему и в таких случаях многие люди склонны давать чересчур критичные суждения, опасаясь, что множество названых «белым» будет выглядеть неоднородно.

В рамках РОМИП'2004 использовалась расширенная шкала с 4 вариантами частично релевантных документов. В этом цикле оценки основной мотивацией для введения шкалы было преодоление пси-

хологического барьера у асессоров и получение максимального объема информации о хотя бы частично релевантных ответах.

Как видно из нижеприведенной таблицы, значительная доля ответов для каждой из дорожек была оценена как частично релевантная, что косвенно подтверждает полезность расширенной шкалы оценки для максимизации «отдачи» от асессоров.

| **Дорожка** | **Витальные** | **Релевантный+** | **Релевантный-** |
|---|---|---|---|
| Поиск по Веб коллекции | 1405 | 1262 | 2095 |
| Поиск по Веб коллекции (без учета расширенных описаний) | 1164 | 1211 | 2047 |
| Классификация Веб сайтов | 2913 | 2586 | 2353 |
| Поиск по коллекции нормативных документов | 778 | 667 | 769 |
| Классификация нормативных документов | 1576 | 574 | 530 |
| Поиск фактов | 438 | 694 | 559 |

### 7.5. Прогнозируемая и реальная трудоемкость сбора оценок

При планировании объема работ по оценке мы исходили из оценок средней производительности, которые по сути были взяты практически с потолка. Конечно мы учитывали опыт оценки прошлого года, но использование нового инструмента оценки сильно изменило картину. Мы также попробовали делать предварительные тестовые замеры эффективности оценки, но поскольку они производились на «сырой» версии инструмента для оценки, то они были весьма приблизительны.

Для того чтобы осознать реальную трудоемкость работы по оценке, все результаты работы асессоров были помечены временными метками. Анализ интервалов между соседними временными метками позволяет судить о времени, которое потребовалось асессору на вынесение очередной оценки.

В следующей таблице приведены данные (в секундах) о прогнозировавшейся нами трудоемкости оценки, а также о средних фактических интервалах. Отметим, что при вычислении этих интервалов не учитывались промежутки более 5 минут, а так же все косвенные затраты связанные с получением и подготовкой данных, запуском инструмента и т.п.

| **Дорожка** | **Прогнозируемая** | **Фактическая** |
|---|---|---|
| Поиск по Веб коллекции | 30 | 13.97 |
| Поиск по Веб коллекции (без учета расширенных описаний) | 30 | 12.95 |
| Классификация Веб сайтов | 40 | 15.5 |
| Поиск по коллекции нормативных документов | 60 | 9.9 |
| Классификация нормативных документов | 60 | 21.11 |
| Поиск фактов | 120 | 9.95 |

Самая большая разница наблюдается в дорожке по поиску фактов, но на это есть ряд причин. Прогнозируемая трудоемкость включала в себя затраты на классификацию факта по заданной таксономии (по факту этого не производилось), а также предполагала, что процесс оценки потребует более детального знакомства асессора с документом.

Для нас гораздо более удивителен тот факт, что фактические затраты на проверку одного ответа для дорожки классификации нормативных документов оказались заметно меньше, чем для дорожки поиска по коллекции нормативных документов. У нас пока нет объяснения этому наблюдению.

Настораживающим фактом является также то, что заметная доля оценок асессоров (например, 1642 из 35000 для дорожки поиска по Веб коллекции) была сделана менее, чем за 2.5с. Этот вопрос также требует дополнительного исследования, хотя основные причины по-видимому ясны – сжатые сроки и сдельный принцип оплаты труда асессоров. Для будущих циклов РОМИП необходимо модифицировать методологию сбора оценок, так чтобы минимизировать потери в качестве получаемых оценок.

## 8. Заключение

Очередной цикл РОМИП завершен и хотя не все всегда было гладко, но нам кажется этот цикл был вполне успешным. Тем более, что в этом году семинар заметно вырос и как следствие заметно увеличился объем работ.

Получено большое количество экспериментальных данных, которые требуют дополнительной обработки и анализа. Мы надеемся, что это позволит улучшить или хотя бы оценить используемые в РОМИП методику.

Для дальнейшего успешного развития семинара необходимо продолжать движение в направлении укрепления и повышения эффективности его организационной структуры, а также его популяризации. В частности, необходимо завершить внедрение модели с использованием «хранителя коллекций» и проработать вопрос с возможностью участия иностранных коллективов.

## Благодарности

Огромное спасибо компании «Кодекс», предоставившей коллекцию нормативных документов для использования в РОМИП. Особо отметим личный вклад Максима Губина, проделавшего основную работу по подготовке текста соглашения об использовании данных, а также обеспечившего преобразование данных в формат РОМИП.

Большое спасибо Владу Шабанову за подготовку нового обучающего множества для дорожки Веб классификации. Мы также благодарны компании «Гарант-Парк Интернет» за их участие в подготовке дорожки по поиску фактов в Веб коллекции и компании «Парк.Ру» за предоставленную выборку из журнала запросов.

Грант РФФИ (№ 04-07-90280-в) в значительной степени упростил проведение работ по оценке результатов. Спасибо Борису Доброву, что он настоял на подаче заявки и выполнил основную работу по ее подготовке.

Отдельное спасибо Наталье Лукашевич, представлявшей РОМИП на конференции LREC’2004.

Мы также хотим поблагодарить Михаила Агеева, Александра Антонова, Илью Бояндина, Павла Браславского, Марину Некрестьянову, Екатерину Павлову, Владимира Плешко, Илью Сегаловича за их вклад в организацию РОМИП. Ну и конечно спасибо асессорам, без кропотливого труда которых безусловно ничего бы не получилось.

## Литература


[1] Труды РОМИП'2003 / Под редакцией И.С. Некрестьянова // СПб, НИИ Химии СПбГУ, 132 с.

[2] ROMIP Web site, 2004. http://romip.narod.ru

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

[4] 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.

[5] Boris Dobrov, Igor Kuralenok, Igor Nekrestyanov, Ilya Segalovich Russian Information Retrieval Evaluation Seminar //LREC'04, Apr 2004

[6] 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.

[7] Соглашение об участии в семинаре РОМИП. http://romip.narod.ru/docs/romip2004_agreement.pdf

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


## ROMIP'2004: Report from organizers


Igor Kuralenok, Igor Nekrestyanov



This paper describes basics of ROMIP'as well as details of organization of ROMIP'2004 – collections, tracks, rules and used evaluation methodology. In addition to organizational information we present preliminary results of some of our experiments with evaluation methodology.