

Обзор РОМИП'2005

© И. Кураленок, И. Некрестьянов, М. Некрестьянова

Санкт-Петербургский Государственный Университет

romip@meta.math.spbu.ru

<http://romip.narod.ru>

Аннотация

Эта статья описывает основные детали организации РОМИП'2005, включая описание дорожек и коллекций, а также процедуры сбора оценок от ассессоров.

1. Введение

РОМИП - это инициатива по созданию плацдарма для проведения независимой оценки методов информационного поиска, ориентированных на работу с русскоязычной информацией. Как часть этого проекта проводится ежегодный практический семинар, участники которого имеют возможность протестировать разработанные ими методы решения различных задач информационного поиска. Принципы организации РОМИП уже неоднократно описывались нами ранее [1,2,3], поэтому мы лишь вкратце остановимся на них в рамках этой статьи.

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

Для участия в семинаре необходимо подать заявку в оргкомитет и подписать необходимые соглашения об использовании данных. Участник свободен в определении набора дорожек, в которых он хочет участвовать, и может напрямую влиять на правила проведения этих дорожек во время их формирования. Оргкомитет лишь координирует проведение семинара.

Выполнение заданий РОМИП производится участниками самостоятельно и на своем оборудовании. Оргкомитет организует проведение оценки результатов выполнения заданий. Предполагается, что участники самостоятельно анализируют результаты оценки и подготавливают доклад, описывающий принципы и результаты их экспериментов.

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 [4, 7, 8, 9], поскольку это повышает уверенность в получении надежных результатов [5, 6, 10, 11].

Проект имеет некоммерческий характер и осуществляется силами сообщества российских исследователей и разработчиков, занимающихся информационным поиском. Затраты на подготовку и проведение тестирования частично компенсируются грантом РФФИ (№ 04-07-90280), а частично возмещаются за счёт взносов участников. Результаты тестирования предназначены для использования только в исследовательских целях и не могут быть использованы в маркетинговых или коммерческих целях.

В 2005 году состоялся третий семинар РОМИП. Семинар состоял из 10 дорожек и в нем приняло участие 14 систем. В этой статье дан краткий обзор семинара, более подробную информацию о котором можно найти на сайте romip.narod.ru.

2. Коллекции

В настоящее время участники РОМИП могут выполнять исследования со следующими коллекциями полнотекстовых документов:

- **Веб-коллекция Narod.ru**
Предоставлена ООО «Яндекс» в 2003 году. Состоит из 728 тысяч документов с 22 тысяч Веб сайтов из домена narod.ru.
- **Веб-коллекция DMOZ**
Предоставлена ООО «Рамблер Интернет Холдинг» в 2004 году. Создана на основе русскоязычной части каталога dmoz.org с целью получения обучающего множества для задачи классификации Веб-сайтов, содержит более 300 тысяч документов с 2087 сайтов.

- **Коллекция нормативных документов**
Предоставлена ИК «Кодекс» в 2004 году. Содержит 60 тысяч основных правовых документов законодательства России, изданных федеральными органами власти по состоянию на начало 2004 года.
- **Новостная коллекция**
Предоставлена ООО «Яндекс» в 2005 году. Состоит из 24000 сообщений из 17 новостных источников. Более подробно коллекция описана в разделе 2.1.

Для предотвращения несанкционированного использования данных участники подписывают специальные соглашения.

Отметим, что коллекции доступны не только участникам РОМИП - доступ к ним с целью исследования методов информационного поиска может получить любой желающий после обращения в оргкомитет и подписания необходимых соглашений с правообладателями.

2.1. Новостная коллекция

Коллекция содержит все новостные сообщения из 17 источников (список источников опубликован на Веб-сайте РОМИП) для трех временных интервалов:

- с 18.11.2003 по 24.11.2003 включительно (8 дней)
Ключевой сюжет: "отставка Шеварднадзе"
- с 01.12.2003 по 09.12.2003 включительно (8 дней)
Ключевые сюжеты: "Взрыв в Ессентуках - Выборы в Госдуму"
- с 31.03.2004 по 7.04.2004 (8 дней)
"Обычная неделя", нет ярко выраженного ключевого сюжета.

Общий объем коллекции – около 50Mb, 24 тысячи документов.

Права на использование в целях РОМИП предоставлены компанией Яндекс, являющейся владельцем коллекции. Для получения доступа к коллекции необходимо подписать соглашение об использовании данных.

3. Задачи

Программа РОМИП'2005 состояла из 10 дорожек (5 в 2004 году), каждая из которых была посвящена отдельной задаче. В этом разделе мы кратко опишем дорожки этого года. Детальное описание

правил дорожек, включая файлы с наборами заданий, можно найти на сайте семинара (romip.narod.ru).

3.1. Поисковые дорожки

В этом году 4 дорожки было посвящено задаче поиска документов по запросу:

1. Поиск по Веб коллекции (web adhoc)

Повторение прошлогодней дорожки, с незначительно сокращенным набором запросов. Отличие состояло в наборе заданий, которые отбирались для оценки.

2. Поиск по коллекции нормативных документов (legal adhoc)

Также повторение дорожки РОМИП'2004. Новый набор оцениваемых заданий и методология сбора оценок.

3. Поиск по смешанной коллекции (mixed adhoc)

В качестве набора заданий использовалось объединение заданий дорожек поиска по Веб и поиска по коллекции нормативных документов. Оценка производилась для тех же заданий, что оценивались для этих двух дорожек.

4. Поиск по «документу-образцу» (mixed feedback)

Дорожка, посвященная оценке методов поиска по документу образцу и методов, учитывающих обратную связь от пользователя. В качестве коллекции использовалась объединение коллекции Narod.Ru и коллекции нормативных документов. Более подробно дорожка описана в разделе 3.1.1.

Для всех поисковых дорожек системы выполняли большое число заданий, из которых впоследствии отбирались задания для оценки. Такой подход предотвращал возможность ручной настройки системы под конкретные запросы.

Ответ системы для каждого из заданий представлял собой упорядоченный список из не более 100 документов. Первые 50 из них использовались для проведения оценки.

3.1.1. Поиск по документу образцу (mixed feedback)

Для того чтобы совместить оценку этой дорожки с оценкой других поисковых дорожек задача формулировалась следующим образом: «Для задания, состоящего из запроса и релевантного ему документа, вернуть упорядоченный список релевантных документов».

Участник сам вправе решать какую информацию использовать. Используя только запрос, задача сводилась к классической задаче

поиска, только документ – поиска по документу образцу, а в случае если использовались обе части задания – задаче поиска с использованием обратной связи от пользователя.

Важно отметить, что в любом случае предполагалось, что информационная потребность пользователя описывается запросом и текст документа не используется при проведении оценки результатов ассессорами.

Задания были построены на основе запросов, которые оценивались в предыдущих циклах РОМИП. Всего было 8205 заданий построенных по 156 (67 + 89) запросам. В качестве документов образцов использовались только строго релевантные документы.

3.2. Дорожки по тематической классификации

В программе семинара было три дорожки посвященных задаче тематической классификации:

1. Классификация Веб-сайтов
2. Классификация Веб-страниц
3. Классификация нормативных документов

Дорожки связанные с Веб-сайтами и нормативными документами – это повторение дорожек РОМИП'2004 с изменением списка оцениваемых категорий. Для каждого классифицируемого объекта система могла вернуть от 0 до 5 категорий.

Отметим, что в случае нормативных документов оценка производилась путем сравнения с эталонным каталогом, который составлялся вручную экспертами компании «Кодекс». То есть оценка производилась без участия ассессоров.

Дорожка по классификации Веб страниц по своим правилам повторяла дорожку классификации Веб-сайтов, за исключением того, что объектом классификации являлись отдельные страницы, а не сайты в целом. В частности, использовалось то же самое обучающее множество из 247 категорий на основе каталога DMOZ.

3.3. Контекстно-зависимое аннотирование

Эта дорожка была посвящена оценке качества аннотирования в приложении к задаче представления результатов для поисковой системы. В контексте этой задачи аннотации используются для краткого описания найденных документов, по которому пользователь принимает решение о загрузке полных текстов документов.

Системам, участвовавшим в этой дорожке, предлагалась следующая задача: по паре запрос и документ, составить аннотацию

этого документа по этому запросу. Размер аннотации не должен превышать 300 символов. Использование HTML-разметки в аннотации не допускается.

Ассессоры оценивали релевантность документа запросу по аннотации этого документа (сам документ ассессорам не предоставлялся). При оценке ассессор видел название документа (не более 100 символов) и текст аннотации (до 300 символов).

Набор заданий состоял из 21872 заданий, построенном на основе 158 (67 и 91 соответственно) заданий дорожек поиска по Веб и по нормативной коллекции РОМИП'2004. В набор включались все оценивавшиеся пары документ-запрос.

3.4. Организации новостного потока

Рассматривалась задача разбиения потока новостных сообщений на событийные сюжеты. Сюжеты могут быть связаны в сюжетные линии ассоциативными связями (причинно-следственно-пространственно-временными).

Формально:

- **Событие** (event) - нечто, происходящее в определенное время в определённом месте наряду со всеми необходимыми причинами и всеми неотвратимыми последствиями
- **Событийный сюжет** (event-based topic) - отражение события в потоке новостных сообщений (то есть набор новостных сообщений, посвященных соответствующему событию).

Задачей системы является структуризация потока сообщений новостной коллекции в набор сюжетов, связанных ассоциативными связями в «надсюжеты».

Для того, чтобы проиллюстрировать задачу, приведем пару примеров:

- *«Траур по кончине папы»* и *«Перенос в связи с кончиной папы очередного тура футбольного чемпионата Италии»* - это разные сюжеты, но между ними есть ассоциативная связь.
- *«Олег Табаков остается на посту художественного руководителя МХАТа»* и *«Александр Домогаров выступит в главной роли в музыкальном спектакле»* - это не только разные сюжеты, но между ними нет ассоциативной связи, поскольку нет причинно-следственной связи (хотя есть тематическая схожесть).

3.5. Поиск фактов (QA)

В отличие от 2004 года в этом году фактографическая дорожка проводилась на основе новостной коллекции и без использования фиксированного списка фигурантов.

По результатам обсуждения с участниками было выделено две подзадачи:

- **Выделение именованных сущностей.**

Для каждого документа коллекции система должна построить список именованных сущностей, которые в нем встречаются. Желательно указать класс сущности из следующего списка:

- персона,
- организация,
- географический объект,
- другой класс.

- **Выделить факты заданных типов.**

От системы требовалось выявить все вхождения фактов одного из следующих двух типов во всех документах коллекции:

- Где работал/работает данный человек?
- Кто владеет или владел данной организацией?

Факт должен был иметь следующую структуру:

- 2 фигуранта
- отношение, их связывающее
- фрагмент текста (до 500 байт), где факт описывается.

В правилах дорожки также было указано, что:

- Референтные упоминания не учитываются.
- Нормализация имен собственных не обязательна, но и не запрещается.
- Число именованных сущностей и число фактов, которые можно вернуть для одного документа не ограничивается.

4. Участники

Всего мы получили 17 заявок на участие в РОМИП'2005, но только 14 из этих систем дошло до финиша (11 и 9 в РОМИП'2004). Более подробная информация о полученных заявках и вариантах ответа представлена в таблице 1 (прочерк означает, что заявка была подана, но участник либо решил отказаться от участия в этой дорожке, либо не предоставил результаты в срок).

	Поиск				Классификация			Ново-сти	QA	Анно-тиро-вание
	Веб	Legal	Mixed	По образцу	Веб сайты	Веб стр.	Норм. док.			
АСК				19						
AStars						-	-			
Exactus	-	1	-							-
eXtragon										1
NNCS					1					
RCO					4	3	24	-	1+1	1
SearchInform	1	1	1	1			-	-	-	-
Specs						1	4			
WHO.Integrum									1+1	
ДЕКОМ		-					-			
Золушка	-								-	3
Кодекс	1	1	2							3
Поиск@Mail.ru	1	1	1		1	1	2	-		
Галактика-Зум					2	2	2			
Синдбад					-	-				
УИС Россия	2	3				-	-			
Яндекс	1	1			-	1		1		

Таблица 1. Участники РОМИП'2005.

5. Оценка результатов

Основным методом оценки результатов в 2005 году, как и в прошлые годы, служили оценки ассессоров, которые собирались методом «общего котла» (pooling) [4, 5, 9].

«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 того чтобы снизить влияние субъективности оценки мы собирали как минимум по две независимые оценки на пару задание-ответ (на большее не хватило ресурсов). При слиянии оценок ассессоров использовались два метода:

- **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”.

В отличие от прошлого года, отбор заданий для оценки был более интерактивным. После предварительного отбора заданий оргкомитетом, каждый из участников дорожки мог исключить 5% отобранных заданий. Окончательный выбор производился из оставшихся заданий.

К сожалению, на момент подготовки трудов РОМИП оценка результатов не была закончена в полном объеме. В частности, была завершена лишь частичная оценка второго этапа для дорожки поиска фактов и не выполнена проверка результатов для дорожки организации потока новостей. По мере завершения этой работы, процедура оценки этих дорожек будет описана более детально на сайте семинара.

Далее в этом разделе мы кратко опишем основные особенности оценки каждой из дорожек. Более подробную информацию об оценивавшихся заданиях, использовавшихся метриках и полученных результатах можно найти в приложениях к трудам РОМИП'2005, а также на сайте семинара.

5.1. Поисквые дорожки

Для сокращения затрат оценка поисковых дорожек была совмещенной. То есть в том случае где задания дорожек пересекались, полученные от систем ответы использовались для построения единого котла. Так, например, для Веб-запроса в котел включались документы, встречающиеся в ответах не только на соответствующие задание дорожки поиска по Веб, но и из соответствующих заданий дорожек поиска по смешанной коллекции и поиска по документу-образцу.

Методология сбора оценок ассессоров для поисковых дорожек во многом повторяла методологию 2004 года. В частности, для оценки использовались расширенные описания запросов, глубина котла 50 и использовалась та же шкала оценки: витальный, релевантный+, релевантный-, нерелевантный.

Значительно изменился этап отбора и подготовки заданий для оценки. К отбору заданий и составлению для них расширенных описаний были привлечены ассессоры. Четверым ассессорам было поставлено задание отобрать и составить описания для максимум 35 запросов из некоторой случайной выборки выполнявшихся заданий. Каждый ассессор работал со своей выборкой, что позволило избежать коллизий в списках отобранных заданий. В дальнейшем, каждый из этих ассессоров производил оценку для всех «своих» заданий.

Список из 121 отобранного запроса с описаниями был предоставлен участникам соответствующих дорожек, которые забраковали 10 из них. Из оставшихся запросов оргкомитетом было отобрано 50 для проведения оценки и еще 25 оценивалось повторно.

Значительным изменением в процедуре оценки стало привлечение экспертов с юридическим образованием для оценки заданий поиска по коллекции нормативных документов.

Как и в случае с Веб-запросами, отбор новых заданий для оценки производился ими самостоятельно. В полученный список из 80 запросов были добавлены 30 задания для повторной оценки, 27 заданий было исключено в силу разных причин.

Как было отмечено координатором оценки: «Запросы, представленные для оценки, касаются самых разных тем и экспертам не всегда сразу было понятно, о чем идет речь. Но знание основ права,

юридической силы нормативных актов и т.д. позволяло достаточно быстро вникнуть в проблему и, как мне кажется, достаточно качественно проставить оценки. Качественно - значит так, чтобы потенциальный потребитель правовой информации (в первую очередь это юрист) согласился с нашими оценками».

К сожалению, трудозатраты при использовании экспертов значительно превысили ожидаемые. Поэтому, в этом году удалось собрать лишь по одной оценке для каждого из заданий. Тем не менее, есть надежда, что качество полученных оценок значительно выше, чем при использовании ассессоров без юридического образования.

5.2. Дорожки классификации

Как и в прошлом году, для оценки качества классификации нормативных документов использовался эталонный каталог, предоставленный компанией «Кодекс». Оценивались все документы, отнесенные к одной из 40 выбранных категорий какой-либо из систем. Отбор категорий для оценки производился случайным образом среди категорий, которые не оценивались в прошлом году, и для которых в обучающем множестве было не менее 10 примеров.

Для дорожек классификации Веб-сайтов и Веб-страниц оценка производилась для одних и тех же 26 категорий. Однако, при классификации страниц число документов, отнесенных к заданной категории хотя бы одной из систем, было весьма значительным. Поэтому оценка производилась не для всех этих документов, а лишь для 10% случайно отобранных из них.

Таким образом, получилось, что не для всех Веб-страниц отнесенных системой к заданной категории были собраны оценки. Поэтому, при расчете численных результатов было вычислено два набора значений:

- Учитывались все документы и те, для которых нет оценок релевантности, считались нерелевантными.
- Учитывались только те документы, для которых есть оценки релевантности.

В отличие от прошлого года расширенные описания категорий были построены единым способом и явно перечисляли все соседние категории второго уровня. Пример описания приведен в приложении С.

5.3. Аннотирование

Отбор заданий для оценки производился среди тех заданий, для которых все системы предоставили варианты аннотаций. Для каж-

дого из запросов отбиралось от 10 до 80 документов, так чтобы 50% составляли строго релевантные документы (то есть документы, признанные релевантными при использовании сильных требований к релевантности в 2004 году), 25% - слабо релевантные и 25% - нерелевантные. Итого было отобрано 960 документов для 39 Веб-запросов и 1970 документов для 56 запросов по коллекции нормативных документов.

По правилам дорожки, процедура оценки аннотаций должна была быть похожей на процедуру оценки результатов для задачи поиска, а в идеале имитировать процесс работы пользователя с результатами поисковой системы.

Необходимо также учитывать следующие ограничения:

- Демонстрируя ассессору несколько вариантов аннотаций для одной и той же пары документ запрос, есть риск, что ассессор слишком строго будет судить не лучшие, но, тем не менее, хорошие аннотации (поскольку он видит результат лучше).
- Результаты разных систем могут значительно различаться по качеству. Демонстрация их в едином списке может негативно сказаться на оценках менее читабельных аннотаций.
- Желательно, чтобы каждый ассессор оценивал примерно одинаковое число заданий от каждой системы.

Исходя из этого, использовалась следующая процедура оценки:

- Удалены все точные дубликаты аннотаций (из 8 вариантов ответа для каждого задания уникальных было 6-7).
- Сформировано 7 наборов для оценки, каждый из которых содержал не более одного варианта аннотации для конкретной пары запрос-документ.
- Ассессор мог участвовать в оценке одного или двух этих наборов. В случае двух наборов, задания ассессору выдавались частями, так чтобы к моменты оценки разных вариантов аннотаций для одной и той же пары документ-запрос были максимально разнесены во времени.
- Ассессор оценивал каждую аннотацию индивидуально (не видя текста других аннотаций).

Еще одним важным отличием от оценки результатов дорожки поиска является семантика шкалы оценки. Вместо, оценки наличия полезной информации в документе, ассессор должен был оценить вероятность того, что полезная информация там есть.

5.4. QA

На этапе обсуждения методологии оценки качества выделения фактов предполагалось, что будет использован следующий подход:

- Оргкомитет отберет несколько фигурантов, для которых во всех документах коллекции было найдено заметное число фактов.
- Участники предложат альтернативные именованные фигуранты, которые будут использованы для построения соответствующих котлов фактов.
- Каждый из фактов будет проверен ассессорами.

Несомненным достоинством такого подхода является возможность оценить полноту сбора «досье». Однако, этап согласования псевдонимов требует дополнительно координации.

В силу жестких временных ограничений при проведении оценки был использован альтернативный вариант оценки:

- Оргкомитет отобрал некоторое множество документов.
- Оценивались все факты, отобранные из этих документов.

Такой подход не позволяет оценить полноту досье, но позволяет оценить полноту выделения фактов из документов. Безусловно, эти характеристики не эквивалентны и мы планируем провести дополнительную оценку согласно изначально запланированному подходу.

Для оценки были отобраны документы, где было выделено не менее 10 фактов, а также примерно 300 случайно отобранных документов, в которых был найден только один факт только одной из систем. Общее число оценивавшихся фактов – 3000.

Одной из методологических сложностей является получение результатов оценки, которые можно повторно использовать. Проблема состоит в том, что использование разных схем выделения фигурантов и нормализации их наименований не позволяет сравнивать факты, выделенные разными системами автоматически. Поэтому, процедура оценки была направлена на получение «эталонного» множества фактов для данного набора документа. Факты, возвращенные системами использовались в качестве опорных точки, для последующего уточнения.

К сожалению, на момент написания этой статьи оценка еще не полностью закончена. В частности не проведена оценка качества выделения именованных сущностей и унификация фактов.

Дорожка	Коллекция	Число заданий	Оценивалось заданий	Из них повторно	Объем оценки (часов)	Согласие ассессоров
Поиск по Веб	Narod.Ru	24166	75 запросов, 16418 док.	25	500 (включ. mixed)	0.89
Поиск по норм. коллекции	Legal	12925	83 запроса, 12213 док.	30	400 (включ. mixed)	-
Поиск по смеш. коллекции	Mixed	37091	158 запросов, 17175 док.	-	Включен в другие дорожки	0.89
Поиск «по образцу»	Mixed	8205	58 заданий	-	Включен в другие дорожки	-
Классификация Веб сайтов	Narod.Ru + DMOZ	247	26 категорий, 6941 сайт	0	240	0.91
Классификация Веб страниц	Narod.Ru + DMOZ	247	26 категорий, 9734 страницы	-	240	0.88
Классификация норм. док.	Legal	183	40 категорий	0	0	-
Аннотирование	Mixed	21872	960+1970 док, 5760+12060 аннотаций	-	240	0.72
Поиск фактов	News	-	3000 фактов	-	50+150	0.94
Новостная дорожка	News	-	-	-	0+200	н.д.

Таблица 2. Дорожки РОМИП'2005.

Одной из причин задержки с оценкой являются значительные технические изменения в инструменте оценки, которые были произведены недостаточно оперативно.

Работа ассессоров затруднялась расплывчатыми ответами систем – в случае указания широкого фрагмента текста, где встречается много фактов, ассессору сложно провести коррекцию ошибки или уточнение факта. Отметим, что формат представления результатов оказался не очень удачным – нам кажется, что оценка была бы качественней при более четком указании мест, где упоминаются фигуры.

Заключение

В 2005 году семинар снова значительно вырос – увеличилось как число участников, так и число рассматриваемых задач, появилась новая коллекция. Значительные усилия были направлены на улучшение методологии оценки.

Значительное увеличение семинара имело и негативные последствия. В частности, «проблемы роста» обусловили значительные сдвиги сроков завершения отдельных этапов.

Отчасти это обусловлено несоблюдением сроков сдачи результатов, но во многом причиной является наложение сроков оценки разных дорожек, что значительно затрудняло как техническую подготовку, так и координацию процесса. Нам представляется, что для улучшения ситуации в будущем стоит:

- Еще сильнее разнести во времени сроки сдачи результатов.
- Активнее делегировать выполнение оценки.
- Больше внимания уделять подготовке требуемых технических инструментов до начала оценки.

Тем не менее, мы надеемся, что семинар в 2005 году оказался полезным и интересным для всех его участников и что в будущем развитие семинара не прекратится.

Литература

- [1] Труды РОМИП'2003. Под ред. И.С. Некрестьянова, Санкт-Петербург: НИИ Химии СПбГУ, 132 с, октябрь 2003.
- [2] Труды РОМИП'2004. Под ред. И.С. Некрестьянова, Санкт-Петербург: НИИ Химии СПбГУ, 214 с, сентябрь 2004.
- [3] М.С. Агеев, М.В. Губин, Б.В.Добров, И.Е. Кураленок, И.С. Некрестьянов, В.В. Плешко, И.В.Сегалович, В.И.Шабанов. Российский семинар по оценке методов информационного поиска (РОМИП) в 2004 году, Труды Диалог'05, июнь 2005.
- [4] И. Кураленок, И. Некрестьянов. Оценка систем текстового поиска. Программирование, 28(4): 226-242, 2002.
- [5] И. Некрестьянов, М. Некрестьянова, А. Нозик. К вопросу об эффективности метода «общего котла». Труды RCDL'2005.
- [6] C. Buckley and E.M. Voorhees. Retrieval evaluation with incomplete information. In *Proc. of the SIGIR'04*, pp. 25-32, 2004.
- [7] Kando N., Kuriyama K., Yoshioka M., Overview of Japanese and English Information Retrieval Tasks (JEIR) at the Second NTCIR Workshop In *Proc. of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization May 2000 - March 2001*. - National Institute of Informatics, 2001.
- [8] C.Peters, M. Braschler, J.Gonzalo, M.Kluck (Eds.) Evaluation of Cross-Language Information Retrieval Systems - Second Workshop of the Cross-Language Evaluation Forum (CLEF-2001). Revised papers. - Lecture Notes in Computer Science 2406, Springer 2002.
- [9] Voorhees E., Overview of TREC 2001 NIST Special Publication 500-250: The Tenth Text REtrieval Conference (TREC 2001) - pp. 1-15.
- [10] E.M. Voorhees and C. Buckley. The Effect of Topic Set Size on Retrieval Experiment Error. In *Proc. of the SIGIR'02*, p. 316-323, 2002.
- [11] J. Zobel. How reliable are the Results of Large-Scale Information Retrieval Experiment? In *Proc. of the SIGIR'98*, p.307-314, 1998.

Overview of the ROMIP'2005

Igor Kuralenok, Igor Nekrestyanov, Marina Nekrestyanova

This paper describes details of organization of ROMIP'2005 – collections, tracks, rules and evaluation methodology.