РОМИП’2006: отчет организаторов

© И. Некрестьянов, М. Некрестьянова

Санкт-Петербургский Государственный Университет
romip@meta.math.spbu.ru
http://romip.narod.ru

Аннотация

В этой статье кратко описаны основные детали организации РОМИП’2006, включая описание дорожек и коллекций, а также процедуры сбора оценок от асессоров. Основное внимание уделено особенностям организации РОМИП в 2006 году. Принципы РОМИП и дополнительные подробности о деталях организации «постоянных» дорожек можно найти в отчетах организаторов РОМИП за прошлые годы [1, 2, 3, 4].

1. Введение

Как и ранее, в 2006 году РОМИП структурно состоял из набора «дорожек» - секций, посвященных конкретным проектам (с фиксированной задачей и правилами оценки). Для участия в семинаре требовалось подать заявку в оргкомитет и подписать необходимые соглашения об использовании данных. Участники свободны в выборе набора дорожек, в которых они хотят участвовать.

Традиционно, правила проведения дорожек обсуждаются во время их формирования. Так, например, в этом году участники во многом определили правила проведения вопросно-ответной дорожки. Оргкомитет лишь координирует проведение семинара и пытается воплощать выбранные методы оценки в жизнь.

Выполнение заданий РОМИП производится участниками самостоятельно и на своем оборудовании. Оценка результатов происходит независимо и контролируется оргкомитетом. Участники могут помочь в проведении оценки, но в этом случае они привлекаются к оценке дорожек, в которых они не принимали участия. 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 [5, 8, 9, 10], поскольку это повышает уверенность в получении надежных результатов [6, 7, 11, 13].

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 2006 году состоялся четвертый семинар РОМИП. Семинар состоял из 10 дорожек и в нем приняло участие 12 систем. В этой отчете 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. Созданна 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 и расширена (примерно на треть) в 2006 году. Коллекция содержит все новостные сообщения из 25 источников (список источников опубликован на Веб-сайте РОМИП) для трех недельных временных интервалов. Текущий объем коллекции – около 75Mb, примерно 31500 документов.

Для предотвращения несанкционированного использования данных участники подписывают специальные соглашения.

Отметим, что коллекции (как, впрочем, задания и таблицы релевантности) 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. В 2006 году этой возможностью воспользовалось порядка 10 человек, некоторые из которых в итоге подали заявки на участие в РОМИП'2006.

3. Задачи

В программе было запланировано 11 дорожек (5 в 2004 году, 10 в 2005), каждая из которых была посвящена отдельной задаче. Однако по факту одна из дорожек (дорожка фактографического поиска) была отменена в связи с тем, что ни один из заявленных участников не дошел до финиша. Итого, реально РОМИП'2006 состоял из 10 дорожек.

В этом разделе мы вкратце 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.

   Как и в 2005 году, 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».

   Важно отметить, что в любом случае предполагалось, что информационная потребность пользователя описывается запросом и текст документа не используется при проведении оценки результатов асессорами. Отметим также, что в качестве документов образцов использовались только строго релевантные документы.

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.

3.2. 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 и 2005 годов с изменением списка оценивающихся 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 асессоров.

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

Повторение дорожки появившейся в программе РОМИП в прошлом году. Дорожка 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.

Правила и процедура оценки не имели значительных изменений по сравнению с прошлым годом.

Формально, рассматривалась 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.

Асессоры оценивали релевантность документа запросу по аннотации этого документа (сам документ асессорам не предоставлялся). При оценке асессор видел заголовок документа (не более 100 символов) и текст аннотации (до 300 символов).

Набор заданий состоял из 42587 заданий, построенном на основе 261 (117 и 144 соответственно) задании дорожек поиска по Веб и по нормативной коллекции за 2004 и 2005 годы. В набор включались все оценивавшиеся пары документ-запрос.

3.4.Структуризация новостного потока

Повторение дорожки по правилам 2005 года. Однако, в 2005 году всего один участник смог сдать результаты и оргкомитет не сумел подготовить инструмент для оценки в срок. Так что, вторая часть этой дорожки (оценка и анализ) выполнялась в этом году впервые.

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. Вопросно-ответная дорожка

Эта дорожка – «новичок» РОМИП, посвящена задаче поиска ответов на вопросы на естественном языке. Близкий аналог в TREC – QA дорожка [12]. Однако, отличия ограничиваются использованием вопросов на русском языке и другой коллекции. В частности, в РОМИП применялся другой принцип формирования заданий и видоизмененная процедура оценки.

Проведение этой дорожки было разбито на несколько этапов:

- Участниками был предложен и после совместного обсуждения зафиксирован список типов вопросов (см. табл. 1)
- Участники и оргкомитет предложили свои наборы запросов для оценки (в итоге всего заданий было 615 – по 200 от участников и 215 от оргкомитета)

- Задания выдавались участникам партиями по 100-175 на короткие промежутки времени (1-2 дня)
- Результаты оценивались асессорами

Вопросы к определению, к подлежащему:
- Что такое? (Что такое анафора?)
- Кто такой? (Кто такой Набоков?)
- Кто сделал что-то? (Кто изобрел велосипед?)
- Какой (-ая,-ое...)/какова? (Какая страна приняла участие в Олимпиаде?)

Вопросы к прямому дополнению:
- Что сделал кто-то? (Что изобрел Томсон?)

Вопросы к обстоятельству:
- Сколько? (Сколько человек живет в Москве?)
- Какую длину/площадь/высоту?
- Какова длина/площадь/высота...? (Какова площадь помещений, построенных в прошлом году?)
- Когда? В какой день? В каком месяце? В каком году? Как долго? (В каком году (месяце,...) случился пожар?, Как долго проходили проверки?)
- Куда? В какую страну/город? На какой континент? (Куда был отправлен груз 18 мая?)
- Откуда? Из какой страны/города? (Из какой страны прибыл груз 18 мая?)
- Где? В какой стране/городе? На каком континенте? С какого континента? (В каком городе находится Эйфелева башня?)
- Почему? (Почему случился пожар?)
- Как? (Как убрать пятно с ковра?)

Вопросы к косвенному дополнению:
- Предлог + <что,в чем, на чем, из чего> (Из чего состоит вода?)
- Какую (-ого,-ое...) + слово с известной семантикой?
- Какую (-ого,-ое...)/какова + слово с неизвестной семантикой?

Вопрос к прямому дополнению:
- Какой (-ие,-ую...)? (Какую страну посетил Путин?)

Таблица 1. Типы вопросов дорожки вопросно ответного поиска

	Поиск				Классификация			News	QA	Аннотирование	Вопр. Ответ.
	Веб	**Legal**	**Mixed**	**feedback**	**Sites**	**Pages**	**Legal**				
ACK				3						1	
ClusterRetrieve2006		-				-		-			
Exactus	-	-	-			-			-	-	1
eXtragon										-	
Kallimachos						-	3			-	
RCO							12		-		
SearchInform	-	-	-	-							
Specs				-	3	2	6				
Stocona	-	-							-		1
ThematicSearch2006		-				-					
Золушка								3		-	
Кодекс	1	1	1							-	
Поиск@Mail.ru	1	1	1							1	
Галактика-Зум				1	2	2	5	-			
Атлас-Альфа					-	3					
УИС Россия	-	-		-	-	-					
Яндекс	6	1	-			-		1	-		

Таблица 2. Участники РОМИП'2006.

4. Участники

Всего мы получили 17 заявок на участие в РОМИП'2006, но только 12 из этих систем дошло до финиша (17 и 14 в 2005 году, 11 и 9 в 2004). Более подробная информация о полученных заявках и вариантах ответа представлена в таблице 2 (прочерк означает, что заявка была подана, но участник либо решил отказаться от участия в этой дорожке, либо не предоставил результаты в срок).

5. Оценка результатов

Основным методом оценки результатов в 2006 году, как и в прошлые годы, служили оценки асессоров, которые собирались методом «общего котла» (pooling) [5, 6, 10].

«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 большинства дорожек мы собирали как минимум по две независимые оценки на пару задание-ответ (исключение - дорожка поиска по нормативной коллекции, где ресурсов хватило лишь на сбор одной оценки). Как и в прошлые годы, при слиянии оценок асессоров 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 основном, при проведении оценки мы стремились использовать те же подходы и инструменты, что применялись в РОМИП'2005. Однако, новые дорожки – вопросно-ответный поиск и

кластеризация новостных документов, потребовали создания специализированных подходов к сбору экспертных оценок.

Далее в этом разделе мы кратко опишем основные особенности оценки каждой из дорожек. Подробная информация об оценивавшихся заданиях и метриках доступна в приложениях и на сайте семинара.

5.1. Поисковые дорожки

Как и в РОМИП'2005 мы постарались совместить оценку поисковых дорожек, где это было возможно. В частности, это означает, что задания для оценки разных дорожек выбирались не совсем независимо, а с учетом их «осмысленности» в контексте других близких дорожек. Однако, накладываемые ограничения были не очень большими и на наш взгляд не должны оказывать заметного влияния на результат оценки.

Технически совмещение производилось следующим образом. В тех случаях, когда 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.

Для сбора оценок использовался тот же самый инструмент, что и в 2005 году. Глубина котлов (50) и шкала оценки (витальный, релевантный+, релевантный-, нерелевантный) также сохранились. Как и в прошлом году, отбор заданий для оценки и составление расширенных описаний производилось самими асессорами.

В 2006 году удалось привлечь эксперта с неполным высшим юридическим образованием (чьи оценки выборочно контролировались другими экспертами (профессиональными юристами)) к оценке результатов для дорожки поиска по нормативной коллекции. Как и в прошлом году, на первом этапе асессоры юристы выбирали потенциально интересные запросы и подготавливали для них расширенные расписания, затем оргкомитет отбирал подмножество из них, которое реально оценивалось (отбор основывался на размерах котлов).

К сожалению, в этом году оценка от эксперта-юриста опять не дублировалась, но по результатам выборочного контроля больших претензий к ее качеству не возникло.

5.2. Дорожки классификации

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. В отличие от прошлых лет, отбор категорий для оценки не производился случайным образом. Число не оценивавшихся ранее категорий, для которых в обучающем множестве было не менее 10 примеров, составило всего 34. Этот набор был расширен 6 категориями с числом обучающих примеров не менее 8.

Основная оценка для дорожек классификации Веб-сайтов и Веб-страниц производилась на основе 24 категорий (см. пример расширенного описания в приложении С). Поскольку, при классификации Веб страниц число документов, отнесенных к заданной категории хотя бы одной из систем, зачастую было весьма значительным, то, как и в прошлом году, оценивалось порядка 10% случайно отобранных из них (но не более 75 от одной системы).

Так как при таком подходе не для всех Веб-страниц отнесенных системой к заданной категории были собраны оценки, то (как и прошлом году) 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.

Печальной особенностью этого года стало большое количество опозданий и накладок, сказавшихся на процессе оценки. Наиболее заметная проблема была в дорожке классификации Веб страниц. После завершения оценки выяснилось, что прогоны одного и участников были построены по неправильному обучающему множеству и оценку пришлось доделывать (что плохо, поскольку мнение асессоров о том, что такое правильный ответ, неизбежно изменяется во времени)

5.3. Аннотирование

Отбор заданий для оценки производился среди тех заданий, для которых все системы предоставили варианты аннотаций. Для каждого из запросов отбиралось от 10 до 80 документов для Веб запросов и от 10 до 40 для запросов к коллекции нормативных документов. Как и в 2005 году мы старались, чтобы 50% составляли строго

релевантные документы (то есть документы, признанные релевантными при использовании сильных требований к релевантности в 2005 году), 25% - слабо релевантные и 25% - нерелевантные. Итого было отобрано **1470** документов для 48 Веб-запросов и **1630** документов для 48 запросов по коллекции нормативных документов.

На предварительном этапе обсуждалась возможность видоизменения прошлогодней процедуры оценки, поскольку итоговые результаты разных систем незначительно различались. Однако, поскольку было получено всего два прогона и достойной альтернативы процедуре оценки так и не было найдено, процедура оценки повторяла прошлогоднюю:

- Сформировано 2 набора для оценки, каждый из которых содержал не более одного варианта аннотации для конкретной пары запрос-документ.
- Асессор мог участвовать в оценке одного или двух этих наборов. В случае двух наборов, задания асессору выдавались частями, так чтобы к моменты оценки разных вариантов аннотаций для одной и той же пары документ-запрос были максимально разнесены во времени.
- Асессор оценивал каждую аннотацию индивидуально (не видя текста других аннотаций).

Как и в прошлом году асессор должен был оценить вероятность того, что полезная информация есть в аннотируемом документе. Шкала оценки имела следующую семантику:

- **Соответствующий (релевантный/витальный)**
  Совершенно очевидно, что полезная информация в документе есть. Например, таким случаем является упоминание ответа или частей ответа в самой аннотации.
- **Скорее соответствующий (релевантный +)**
  Собственно детального ответа в аннотации нет, но есть частичный ответ или что-то, что позволяет предположить наличие ответа в тексте документа. Эта оценка означает, что, увидев такую аннотацию в результатах поиска, вы бы непременно открыли этот документ для ознакомления.
- **Возможно соответствующий (релевантный)-**
  По аннотации кажется, что документ на близкую заданной потребности тему и возможно в нем есть что-то полезное. Вы бы скорее всего в него заглянули, если бы не нашлось лучших ответов.

- **Не соответствующий (нерелевантный)**
  По аннотации кажется, что в документе нет ничего полезного. Например, документ явно посвящен другим вопросам.
- **Документ не может быть оценен**
  Аннотация не читается (представлена в некорректной кодировке или на непонятном языке), вызывает технические проблемы в браузере или не может быть оценен по каким-либо другим объективным причинам.

Интересно, но в этом году результаты систем не были столь разительно похожи, как это было в прошлом году. Вот, например, результаты участников при использовании сильных требований релевантности (и для аннотаций и для документов, см. описание метрик в приложении Н):

Метрика	Прогон 1	Прогон 2
AnnotationAccuracy	0.878	0.829
AnnotationError	0.466	0.463
PrecisionAnnotations (macro)	0.180	0.205
PrecisionAnnotations	0.189	0.218
PrecisionDocuments (macro)	0.565	0.565
PrecisionDocuments	0.557	0.557

Мы пока не знаем, чем объясняется прошлогодний феномен – менее качественной оценкой, ошибками в инструменте вычисления оценок или какими-то еще причинами. Но мы планируем обязательно изучить этот вопрос.

5.4. Вопросно-ответная дорожка

Для проведения оценки по этой дорожке был разработан отдельный инструмент, не интегрированный в среду работы асессора РОМИП. Это временное решение, вызванное техническими сложностями с интеграцией в короткое время. Хотя примененный способ оценки оказался удобен в этом году, когда оценивалось всего два прогона, вероятно при увеличении числа прогонов и, как следствие, вариантов ответов потребуется видоизменить подход к оценке.

В отличие от других дорожек, где асессор оценивает соответствие пары – задание/результат (например, запрос/документ, запрос/аннотация, тема/сайт), в этом случае асессору предлагалось выставлять оценки для всех доступных ответов на заданный вопрос. Поскольку заданиями являлись вопросы на естественном языке, то расширенные описания не использовались. Список доступных ответов показывался асессору в случайном порядке. Асессору демонстрировалось только текстовое содержание ответа, возможность посмотреть исходный документ не предоставлялась. Отметим, что хотя формально ответы систем должны ограничиваться 300 символов, практически это ограничение не соблюдалось (было обнаружено уже в процессе оценки).

Для каждого ответа асессора просили выставить оценку по следующей шкале:

1. Точный ответ
2. Частичный ответ
3. Ответа нет, но из текста похоже, что он рядом
4. Нет ответа

Для ускорения процесса оценки поля ввода ответов, по умолчанию инициализировались значением 4 (“нет ответа”).

Для вычисления метрик правильными ответами считались оценки 1 и 2. Введение двух типов оценок для «правильных» ответов объяснялось наличием разных типов вопросов – некоторые вопросы подразумевают ответ в виде конкретного факта (“Кто открыл Америку?”), в других случаях вероятно существование дополняющих друг друга ответов (“Почему украинцы воруют российский газ?”).

Оценки типа 3 были изначально введены для того, чтобы как-то отмечать явные промахи выделения фрагментов, когда практически весь ответ присутствует в приведенном фрагменте, но главная часть его была обрезана. В некоторых случаях это кажется очевидным из текста (“В 1990 году чемпионом мира по футболу стала сборная …”), а иногда среди ответов встречались очень схожие, но более корректно выделенные фрагменты, которые содержали правильный ответ.

Однако, в процессе выполнения оценки стало ясно, что оценки типа 3 используются не только с этой целью. Например, известно, что асессоры также использовали их, чтобы отметить близкие к теме вопроса фрагменты текста в тех случаях, когда среди ответов не было ни одного точного или частичного ответа.

Вообще, критерии, по которым асессоры принимали решения, видимо несколько варьировались от вопроса к вопросу. Так, напри-

мер, для вопросов, для которых было найдено много правильных ответов, оценки типа 3 выставлялись не так либерально, как в тех случаях, когда не нашлось ни одного правильно ответа.

Отметим, что в текущем виде полученные оценки пока фактически не пригодны для повторного использования, Изначально, мы пытались собрать от асессоров «каноническое представление» правильных ответов, но для многих вопросов оно не очевидно и было принято решение отказаться от этого усложнения процесса. Интересно, что для трети вопросов асессоры не нашли ни одного правильно ответа.

На этапе формирования списка вопросов, для ряда вопросов были сразу предоставлены варианты правильных ответов и даже указаны документы их содержащие. Эта информация не использовалась при оценке, но возможно на основе ее и дополнительного анализа собранных результатов удастся подготовить что-то пригодное для повторного использования (например, в виде наборов регулярных выражений или шаблонов для каждого запроса).

5.5. Структуризация новостного потока

Основная сложность при проведении этой дорожки – это отсутствие четкого понимания, как правильно организовывать оценку для такой задачи (и как сделать результаты повторно используемыми).

Обсуждение разных подходов продолжалось (с большими паузами) более двух лет (см. архив списка рассылки РОМИП). По сути, все обсуждавшиеся варианты находятся между двумя крайностями:

- **Полная разметка потока вручную**
  Имитация работы редактора асессорами, что формально наиболее близко к процессу, выполняемому системой.
- **Проверка результатов работы системы**
  Выборочная проверка разбиения на события и сюжеты предложенного системой с целью узнать мнение асессора по вопросам типа:
  - Должны ли эти события быть вместе?
  - Нет ли чего лишнего в этом кластере?

Как это было точно подмечено, эти подходы отражают концептуально разные позиции - «редакторов» и «читателей».

Технически, полная разметка подразумевает работу асессора с большим количеством индивидуальных сообщений и необходимость принятия сложных решений. Очевидно, что в силу субъективности асессорских мнений (и отсутствия редакторской практики) вероятны значительные расхождения в разбиениях построенных разными асессорами.

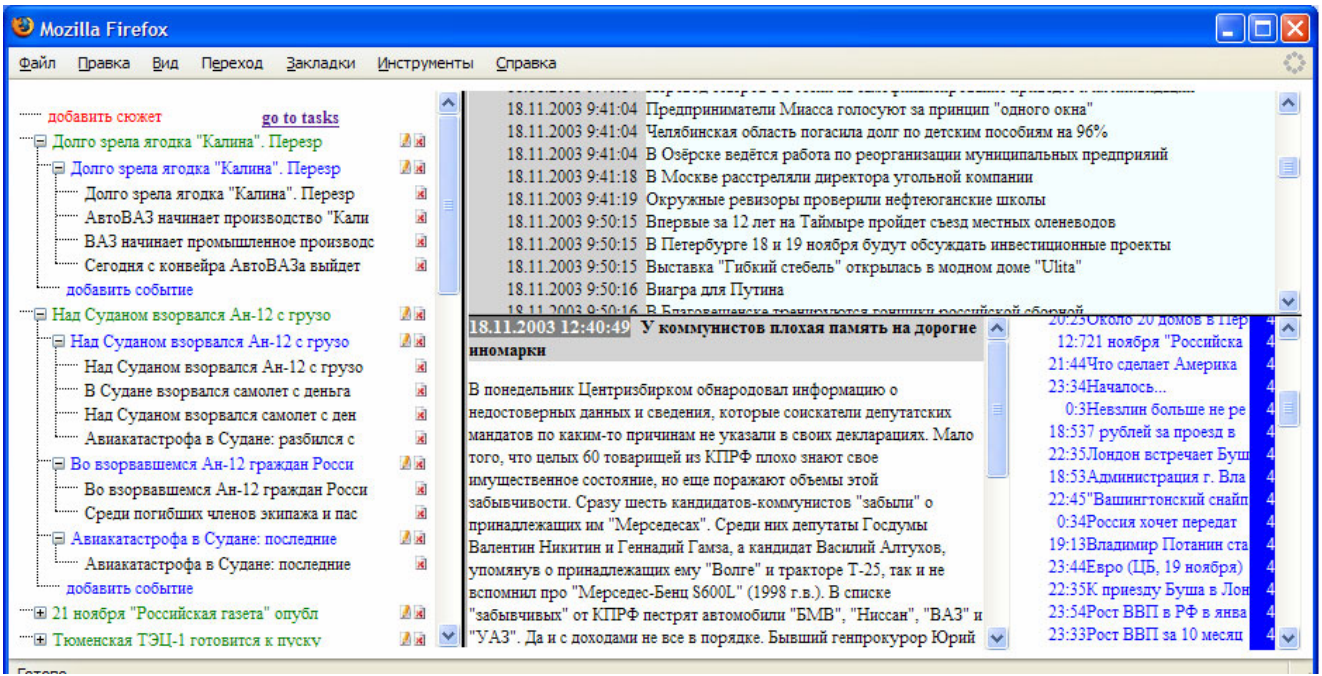

Рисунок 1. Инструмент сбора экспертных оценок для дорожки структуризации новостного потока.

Дорожка	Коллекция	Число заданий	Оценивалось заданий	Из них повторно	Объем оценки (часов)	Согласие асессоров
Поиск по Веб	Narod.Ru	24595	70 запросов, 17813 док.	20	340 (включ. mixed)	0.94 (витал. – 0.32)
Поиск по норм. коллекции	Legal	12925	50 запроса, 7886 док.	-	Н.д.	-
Поиск по смеш. коллекции	Mixed	37091	158 запросов, 30053 док.	-	Включен в другие дорожки	0.94 (витал. – 0.32)
Поиск «по образцу»	Mixed	11358	233 задания	-	Включен в поиск по Веб	-
Классификация Веб сайтов	Narod.Ru + DMOZ	247	24 категории, 2906 сайтов	0	110	0.90 (витал. – 0.24)
Классификация Веб страниц	Narod.Ru + DMOZ	247	24 категории, 7695страниц	-	200 (без доп. оценки)	0.88 (витал. -0.38)
Классификация норм. док.	Legal	183	40 категорий	0	0	-
Аннотирование	Mixed	42587	1470 + 1630 док, 2939+3260 аннот.	-	110	0.74 (vital-0.53)
Вопросно-ответная дорожка	Narod.Ru	615	615, 11654 ответа	-	140	н.д.
Структуризация новостного потока	News	-	~10500 сообщений	-	240	н.д.

Таблица 3. Дорожки РОМИП'2005.

Выборочная проверка результатов может быть реализована с использованием подхода общего котла. Это позволяет лучше контролировать объем оценки и варьировать сложность заданий, которые решает асессор (чем больше альтернатив, тем ниже коэффициент согласия асессоров). Однако, такой подход помогает проверить лишь отсутствие мусора в построенных кластерах, но не позволяет обнаружить пропущенные «не склеенные» события. Это вероятно особенно актуально, если число оцениваемых прогонов невелико.

В итоге, мы решили попробовать реализовать некоторый промежуточный подход. Формально, асессор выполнял работу «редактора» по структуризации входящих сообщений в двухуровневую структуру. Однако, предполагалось, что при работе с конкретным сообщением асессор будет видеть подсказки о том, к каким другим событиям/сюжетам может относиться это сообщение, и эти подсказки будут построены на основе прогонов участников.

Для этой цели был разработан новый модуль для среды работы асессора РОМИП, проиллюстрированный на рисунке 1. Входящая лента сообщений представлена в верхнем правом фрейме, при выборе конкретного сообщения его содержимое показывалось в нижнем центральном фрейме, а справа от него показывались подсказки (если были). Приняв решение, асессор мог перетащить сообщение в одно из существующих событий в левом фрейме или добавить новое событие или сюжет. Узлы первого уровня в дереве соответствуют сюжетам, второго – событиям и листья – это конкретные сообщения.

Отметим, что названия событий и сюжетов можно корректировать, и, в случае изменения решения о каком-либо сообщении, его можно вернуть в список необработанных сообщений или перенести в другое событие/сюжет.

Планировалось, что асессоры обработают все документы за первые два дня от каждой недели. Каждому из 6 участвующих асессоров необходимо было оценить два дня от одной недели, то есть оценки дублировались.

К сожалению, разработка инструмента затянулась и внедрение его происходило с большим количеством технических накладок. В частности, непосредственно перед началом оценки выяснилось, что работа с полной лентой сообщений за два дня (порядка 4000 тысяч) практически невозможна из-за проблем с производительностью. Было принято решение разбить ленту на кусочки по 1000 (а позднее и по 500) сообщений, но выдавать их разным асессорам с перекрытием (одному 0-1000, 1000-2000, … другому – 500-1500, 1500-2500), чтобы можно было попробовать склеить события и сюжеты, которые захватывают несколько разных кусков.

Был выявлен ряд других досадных проблем, так что за время выполнения оценки асессоры трижды обновляли используемое ПО, а часть работы по оценке пришлось переделывать. В последней версии были решены основные проблемы с производительностью, но значительная часть оценки к этому моменту уже была проведена.

Предварительный анализ полученных результатов показывает не очень высокое качество полученной разметки. В причинах еще предстоит разобраться. Во многом, это, безусловно, связано с техническими проблемами инструмента (более поздние результаты на первый взгляд несколько более качественные). Однако, как нам кажется дело не только в этом. В частности, асессоры довольно часто формировали сюжеты в которых присутствуют дублирующиеся события. Вероятно, использование в будущем небольшого демонстрационного задания, для которого будет предоставлен «эталонный» ответ поможет лучше донести суть задания.

Нам также кажется, что текущий процесс слишком сложен для асессора. Последовательная обработка сообщений означает, что асессорам зачастую приходится выбирать среди сотен уже созданных событий и сюжетов.

Одной из перспективных модификаций процесса представляется следующий вариант. Задача асессора не разбить все сообщения на сюжеты, а выявить максимальное число нетривиальных (>1 сообщения) сюжетов среди данных сообщений. При этом рекомендованная процедура - это не просто последовательный поиск сообщений, но также и использование поиска по заголовкам для быстрого перехода к другим потенциальным кандидатам. Для упрощения работы со списком сообщений, где будут оставаться все тривиальные сюжеты, полезно поддерживать несколько флагов для пометки сообщений (как минимум прочитанное/непрочитанное сообщение).

6. Сложности организации

За несколько лет организации семинара у нас накопился изрядный опыт решения разнообразных проблем, возникающих в процессе его проведения. О многих из них мы уже говорили в прошлые годы, но поскольку они все еще актуальны, то мы хотим опять остановиться на некоторых наших наблюдениях.

Ключевая проблема – уже ставшее традиционным отставание от графика проведения, прямыми следствиями которого являются:

- зачастую весьма ограниченное время на осмысление результатов, которое остается участникам;
- высокая пиковая нагрузка на оргкомитет и асессоров;

- «поспешность» при внедрении методологий оценки отсутствие запаса времени не позволяет ставить небольшие эксперименты по сбору оценок модифицированными способами, проанализировать удобство и качество получаемых результатов (например, необходимость править «по живому» в этом году возникала и при оценке вопросно-ответной и новостной дорожек).

Конкретных причин возникновения отставания много, грубо их можно классифицировать в следующие группы:

- **«Авральный» тип соблюдения графика**
  Национальная особенность, присущая не только многим участникам РОМИП, но, к сожалению, и оргкомитету. Например, нередко участники начинают реально работать над заданиями РОМИП после напоминания, что срок сдачи результатов уже прошел.

  Мы стараемся предоставлять участникам максимально допустимое время, для того чтобы повысить ценность дорожек. К сожалению, это стало нормой, а не исключением, поэтому многие рассчитывают на то, что результаты можно будет сдать позже (а без получения всех результатов по дорожке невозможно начало оценки).

  Информация об отказе от участия в дорожках часто не сообщается оргкомитету оперативно, что вкупе с политикой «растяжимых сроков» обуславливает дополнительные задержки.

- **Технические проблемы**
  - *Несоблюдение правил дорожки.*
    Например, использование неправильных заданий или наборов данных.
  - *Неточное соблюдение формата результатов.*
    Зачастую используются прошлогодние версии форматов или даже прошлогодние идентификаторы.
  - *Частичное «искажение» данных*
    Например, в 2006 году в одном из полученных прогонов все идентификаторы были переведены в нижний регистр, а в другом – все знаки ‘-‘ в идентификаторах почему-то оказались замененными на ‘/’.
  - *Недостаточная обкатка инструментов оценки.*
    Довольно большое число досадных ошибок было обнаружено в процессе реальной оценки (новых дорожек).

- **Накладки коммуникации**
  В большинстве случаев нам удается получить оперативный ответ от участников и мы сами стараемся реагировать оперативно. Однако, некоторые сложности все же остаются:
  - *Этап планирования дорожек*
    Например, запуск дорожки QA затянулся поскольку ряд важных деталей согласовывался до середины июня.
  - *Отпускной период*
    Нередко, когда участники сдают результаты прогонов, получают от оргкомитета отмашку о получении и со спокойной совестью уходят в отпуск. Однако, при попытке запустить оценку оргкомитету приходится самостоятельно придумывать как разрешить обнаруженные технические проблемы.
  - *Распределенная организация семинара*
    При решении многих как публичных, так и внутренних организационных вопросов зачастую требуется участие множества лиц. Относительно небольшие задержки на этапах обсуждения могут в сумме быть весьма заметны.
- **Отсутствие информации о прогрессе делегированной оценки**
  При отсутствии прямого контакта с асессорами (как, например, происходит при привлечении экспертов-юристов) или выдаче крупных наборов заданий оргкомитет сталкивается с тем, что все что обычно можно достоверно сказать о текущем состоянии – это «оценка завершена» или «оценка в процессе».

Во многом, эти и другие проблемы по-видимому неизбежны, поэтому необходимо делать больший запас задержек в расписании семинара (но к сожалению оно тоже не «резиновое»). Представляется полезным также:

- Ввести обязательное выполнение предварительных (микро) заданий с целью (автоматической) проверки корректности форматов данных в результатах участников.
- Начинать подготовку и тестирование требуемых средств оценки как можно раньше.
- Собирать предпочтения участников касательно порядка сдачи результатов (например, «классификацию Веб-сайтов» мы скорее всего будем готовы сдавать в последнюю очередь).
- Разнести сроки сдачи заданий еще сильнее, так чтобы участие в РОМИП меньше конфликтовало с другими проектами.

Заключение

В 2006 году не наблюдалось такого активного роста семинара как в прошлые годы. Число участников семинара не увеличилось, а до финиша дошло даже меньше, чем в прошлом году.

Фактически, семинар продолжил развиваться – завершилось оформление юридического статуса РОМИП, к участию подключились новые интересные коллективы, расширилась новостная коллекция, была реализована и внедрена методика оценки для дорожки по структуризации новостного потока и появилась новая интересная дорожка по вопросно-ответному поиску.

Организация оценки по дорожкам вопросно-ответного поиска и, в особенности, дорожки структуризации новостной коллекции поставила перед нами ряд интересных методологических проблем. Результаты еще предстоит осмыслить, но полученный опыт безусловно пригодится для дальнейшего развития этих дорожек.

Мы надеемся, что и в этом году семинар был полезен и интересен всем участникам, и, что в следующем году семинар будет продолжать успешно двигаться дальше.

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] Труды РОМИП'2005. *Под ред. И.С. Некрестьянова,* Санкт-Петербург: НИИ Химии СПбГУ, 224 с, октябрь 2005.

[4] *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.

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

[6] *И. Некрестьянов, М. Некрестьянова, А. Нозик.* К вопросу об эффективности метода «общего котла». Труды RCDL'2005.

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

[8] *Kando N., Kuriyama K., Yoshioka M.*, Overview of Japanese and English Imformation 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.

[9] *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.
[10] *Voorhees E.*, Overview of TREC 2001 NIST Special Publication 500-250: The Tenth Text REtrieval Conference (TREC 2001) - pp. 1-15.
[11] *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.
[12] *E. Voorhes.* The TREC-8 Question Answering Track Report. In Proc. of the TREC-8, 1999.
[13] *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'2006

Igor Nekrestyanov, Marina Nekrestyanova
romip@meta.math.spbu.ru

This paper describes major details of organization of ROMIP'2006 – collections, tracks, rules and evaluation methodology. Primary focus is on things specific to 2006 year. Detailed description of generic rules is available in the proceedings for previous years.