РОМИП’2010: отчет организаторов

© И. Некрестьянов, М. Некрестьянова

romip@romip.ru

Аннотация

В статье описаны особенности организации РОМИП’2010 – дорожки, коллекции, процедуры оценки результатов и другие аспекты проведения семинара. Подробности о принципах РОМИП и базовых подходах к оценке и более детального описания дорожек, которые традиционно являются частью программы РОМИП, можно найти в трудах РОМИП прошлых лет [1], где они неоднократно подробно описывались.

1. РОМИП 2010 в фактах

Программа

- объявлено 15 дорожек, состоялось 9
  - меньше двух заявок
    - поиск по нормативной коллекции
    - поиск по смешанной коллекции
    - поиск по новостной коллекции
    - извлечение фактов из новостной коллекции
  - исчерпалась таксономия для коллекции нормативных документов
    - классификация нормативных документов
  - нет сданных результатов
    - поиск подобных документов

- новые дорожки
  - построение текстовых меток изображений (2 заявки, 1 участник, 1 прогон)
  - вопросно-ответный поиск (4 заявки, 2 участника, 4 прогона)
  - свободная дорожка: исследования на основе материалов РОМИП без общей оценки (2 заявки, 1 участник)

Участие

- 22 заявки, 12 финишировало и сдало 63 прогона
- заявки от 7 новых участников (не считая разных коллективов из одной компании) , 4 финишировало
- 52 заявки на участие в дорожках, только по 23 были сданы прогоны
  - только 6 заявок не реализовалось из-за отмены дорожек
- 2-3 участника финишировало для большинства дорожек
- самая популярная дорожка
  - по заявкам: классификация Веб страниц (7 заявок, 3 финишировало)
  - по участию: поиск картинок по подобию (5 заявок, 4 финишировало)
  - по числу прогонов: классификации Веб сайтов (17)
- максимум прогонов от одного участника по одной дорожке – 11
- самый активный участник
  - по заявкам: УИС РОССИЯ (8 дорожек)
  - по прогонам: УИС РОССИЯ (19)
  - по участию: Exactus (4 дорожки), Яндекс (3 команды, 5 дорожек)
  - по прогонам для одной дорожки: УИС РОССИЯ (11)
- Подробная информация о заявках и полученных результатах приведена в таблице 1.

Коллекции

- Набор коллекций в 2010 году не изменился
- Подробная информация о наборе коллекций РОМИП приведена в таблице 2.

Оценка

- вся оценка завершена до печати трудов ☺
- 2+ оценки для всех заданий
  - 3 оценки для дорожки поиска изображений по образцу
- 1800 человекочасов, 24 асессора
  - 636 человекочасов отработали участники
  - 93 человекочаса - безвозмездная помощь асессоров Яндекса
- Дополнительные координаторы оценки (огромное спасибо!)
  - Поиск по тексту – Михаил Агеев
    - Координация асессоров для дорожки поиск по By.Web – Роман Поборчий
  - Картиночные дорожки – Наталья Васильева
    - Включая создание инструмента оценки для дорожки построения текстовых меток для изображений

Инструментарий и организация

- Значительные обновления инструментов оценки для
  - построения текстовых аннотаций
  - вопросно-ответного поиска
  - построения текстовых меток изображений
  - поиска нечетких дубликатов изображений
- Большинство инструментов перенесено в публичное хранилище
- Активное использование Google Docs для совместной работы над проектом

Дорожка	Заявившихся участников	Предоставивших результаты	Общее число прогонов
Поиск по By.Web	6 (10)	3 (8)	10 (17)
Поиск по KM.RU	6 (9)	3 (7)	7 (12)
Классификация Веб сайтов	6 (3)	3 (2)	17 (4)
Классификация Веб страниц	7 (4)	3 (3)	14 (6)
Контекстно-зависимое аннотирование	2 (3)	2 (3)	2 (7)
Вопросно-ответный поиск	4	2	4
Поиск по визуальному подобию	5 (2)	4 (2)	6 (4)
Поиск нечетких дубликатов изображений	3 (4)	2 (4)	2 (4)
Построение меток изображений	2	1	1

Таблица 1. Сводная статистика о РОМИП'2010. (в скобках данные за 2009 год)

2. Текстовый поиск

В 2010 году в программе РОМИП было запланировано 5 дорожек для задач поиска по текстовой коллекции. Эти дорожки присутствовали в программе РОМИП и в прошлые годы:

Коллекция	Состав	Размер	Предоставлена
Narod.Ru 2003	Веб-сайты из домена narod.ru	728 000 док. 22 000 сайтов	Яндекс
Legal 2004	Законодательство РФ	60 000 док.	Кодекс
DMOZ 2003	Веб-сайты из русской части DMOZ	300 000 док. 2087 сайтов	Рамблер
News 2006	Все новости за три периода из 17 источников	31 500 док. 75 Мб	Яндекс
By.Web 2007	страницы домена .by из индекса Яндекс (май 2007)	1 524 676 док. 8 Гб	Яндекс
KM.RU 2007	~90% от объема www.km.ru на май 2007 (57 сайтов)	3 010 455 док. 13.7 Гб	КМ Онлайн
Legal 2007	Законодательство РФ, Москвы и Санкт-Петербурга (декабрь 2006)	300 000 док. 1.7 Гб	Кодекс
Flickr 2008	подмножество www.flickr.com	20 000 фот.	Flickr
ImageDupl 2008	стоп-кадры из 15 часов видеоматериала	37 800 изобр.	Оргкомитет РОМИП

Таблица 2. Набор коллекций РОМИП

- Классическая задача поиска по запросу по:
  - коллекции нормативно-правовых документов;
  - Веб коллекции BY.Web;
  - Веб коллекции KM;
  - смешанной коллекции.

- Поиск похожих документов по документу-образцу или фрагменту текста.

Фактически состоялись только дорожки поиска по Веб коллекциям By.Web и KM.RU. Остальные дорожки были отменены в связи с отсутствием минимального числа заявок или отсутствием сданных результатов.

Две дорожки поиска по Веб коллекции имеют много общих черт. В частности единый набор заданий из 29231 запросов, полученных из журналов поисковых систем. При оценке результатов для каждой дорожки использовались только запросы из журнала запросов, соответствующего данной коллекции.

Асессоры использовали один и тот же инструмент оценки и руководствовались единой инструкцией, регламентирующей описания информационных потребностей и примеры оценок (см. Приложение В).

Для каждого оцениваемого задания оценки собирались от двух асессоров. При этом асессор мог отказаться от оценки запроса, если задание было ему совершенно непонятным или предлагаемый запрос был порнографическим. Замена асессора в таком случае не производилась (этой возможностью воспользовались 12 раз).

Перед началом работы над заданием и после его окончания асессоры заполняли небольшую анкету, характеризующую их понимание задания. Список вопросов в анкете не изменился с прошлого года [1].

2.1 Особенности дорожки поиска по BY.Web

В этом цикле для дорожки поиска по By.web было оценено 550 запросов – 500 новых запросов и 50 оценивавшихся в прошлом году. Как и в прошлые годы, глубина пула составила 20 документов на запрос. Итого, на данный момент в коллекции By.web размечено 1560 запросов.

Новые запросы выбирались следующим образом:

- сделана выборка из 800 случайных запросов из журнала запросов для BY.web
- ручная фильтрация плохих запросов (порно, непонятные, «противные», а также навигационные на сайты, которых явно нет в коллекции) и специфичных для России, которые вряд ли пользователь искал бы по белорусскому интернету (например, "тур.фирмы г. Краснодара", "турбазы Воронеж")
- из оставшихся выбраны первые 500 запросов с ненулевым пулом

Инструкция асессора, как и в 2009 году, содержала явное указание на необходимость оценки с позиции пользователя, находящегося на территории Белоруссии.

Особенностью оценки в этом году была внешняя и более тщательная координация работы асессоров. В частности, первые варианты ответов асессоров проверялись и при необходимости уточнялось их понимание задачи. Также был рассчитан ряд новых метрик (см. Приложение А).

В этом году, как и в предыдущие годы, каждый из асессоров составлял расширенное описание самостоятельно, и тем не менее уровень согласия асессоров высок – 0,864. Это на 4.5% выше, чем в прошлом году и видимо является следствием более тщательного контроля работы асессоров

2.2 Особенности дорожки поиска по KM.RU

Особенностью оценки дорожки поиска по KM.RU состоит в относительно большой глубине пула (50), что позволяет лучше аппроксимировать полноту. Однако, число оцениваемых запросов значительно меньше, чем для коллекции By.Web.

В этом году оценивалось 100 запросов, из которых 70 новых и 30 ранее оценивавшихся. Процедура отбора запросов схожа с процедурой, которая была описана в предыдущем разделе. Изначально было отобрано порядка 120 запросов, которые были отфильтрованы вручную на предмет плохих запросов, и из оставшихся были выбраны 70 первых запросов не с пустым пулом.

Для запросов, которые оценивались ранее, оба асессора использовали существующее написанное расширенное описание. Для новых запросов описание составлял один из асессоров, а второй переиспользовал его. Полученные результаты имеют уровень согласия асессоров – 0,88, что все еще несколько выше, чем для коллекции by.web, где расширенные описания не переиспользуются.

Всего на данный момент размечено 240 запросов для коллекции KM.RU.

3. Текстовая классификация

Изначально мы не планировали менять набор дорожек для задач текстовой классификации по сравнению с прошлым годом, так как есть устойчивый интерес ко всем дорожкам. Однако, резервы готовой таксономии для нормативной коллекции практически исчерпаны, так что требовалось проведение полномасштабной ручной оценки для этой дорожки. Мы к этому не были готовы

организационно и решили отменить эту дорожку в 2010 году. Надеемся, что в будущем году ее удастся вернуть в программу семинара.

Основные факты для проводившихся дорожек представлены в последующих разделах.

3.1 Классификация Веб сайтов

Для дорожки классификации сайтов постановка задачи и правила проведения дорожки остались прежними, как и в предыдущие годы. Использовалось тоже обучающее множество на основе подкатолога DMOZ, содержащее 247 рубрик и 2116 обучающих Веб сайтов.

Традиционно для этой дорожки оценка построена на методе общего котла, но в этом году полные котлы для многих категорий были очень большими (настолько, что в некоторых случаях запланированных ресурсов не хватило бы на оценку даже одного полного котла).

Поэтому для оценки мы применили метод общего котла на «сужении коллекции»:

- Оценка производилась по подколлекции, то есть выбиралось некоторое подмножество сайтов и все результаты участников сужались на это подмножество. Далее вычислялись оценки традиционным способом.
- Подмножество для сужения строилось исходя из следующих принципов:
  - фиксировался набор оцениваемых категорий и примерная итоговая трудоемкость оценки
  - на первом этапе множество строилось как объединение случайных выборок по 5 сайтов для каждой пары категория/прогон (это гарантирует минимальную оценку каждого прогона)
  - далее множество расширялось за счет случайно выбранных сайтов до тех пор, пока общая итоговая трудоемкость не достигла заданного уровня
  - для сайтов из построенного множества оценивались все пары категория/сайт, в которых сайт был отнесен к одной из оцениваемых категорий хотя бы в одном из прогонов. Например, если сайт aaa.by попал в множество потому что система1 отнесла его к оцениваемой категории2, но в то же время он также был отнесен системой2 к другой оцениваемой категории, то эта пара документ/категория тоже оценивалась.

Всего оценивалось 4000 пар сайт-категория в 20 категориях. Из них 2058 попали в множество на первом этапе, а остальные пары были добавлены в результате случайного расширения "подколлекции".

Как и в прошлые годы при проведении оценки крупные сайты (более 200 страниц) передавались асессорам в сокращенном виде. В сокращенное множество включались только документы, которые в сжатом виде занимали не более 150000 байт. Всего отбиралось 200 документов, так чтобы было хотя бы по одному документу для каждой папки верхнего уровня или скрипта (то есть для каждой строки, получаемой из url путем обрезания по первому символу ‘/’ или ‘?’ после имени сайта).

Интересно, что, как показывает анализ журналов активности асессоров, во многих случаях решение по видимому принимается исключительно по стартовой странице сайта. Если при формировании журналов не было допущено ошибки, то среднее число просмотренных страниц с одного сайта не превышает двух. Означает ли это, что работа асессоров недостаточно качественная, и насколько это влияет на качество оценки? На эти вопросы у нас пока нет ответов.

3.2 Классификация Веб страниц

С 2008 года особенностью этой дорожки является использование не множеств документов, отнесенных к данной категории, а списков документов, упорядоченных по близости к категории. Для оценки из каждого такого списка берутся только первые N документов (в 2010 году - 50) от каждой системы. В результате повышается процент релевантных документов в котле, что дает лучшее представление о качестве работы систем. Однако, это не позволяет аккуратно оценить полноту.

В этом году оценка производилась по тем же 20 категориям, что были выбраны для оценки классификации Веб-сайтов. При этом согласие асессоров для классификации страниц (0,74) оказалось ниже, чем для классификации сайтов (0,83).

4. Контекстно-зависимое аннотирование

Постановка задачи для этой дорожки была практически такой же, как и в прошлом году. Набор заданий был построен на основе оценивавшихся в дорожках поиска документов и содержал 193676 заданий.

Ответом системы для задания является фрагмент текста не более 300 символов без HTML разметки (в том числе,
 и <p>). В дополнение к аннотации асессор видел заголовок документа (содержимое тега title, до 100 символов). Размер заголовка не учитывается в размере аннотации. Асессор НЕ видит сам документ.

Мы получили заявки от 4 участников, но только двое добрались до финиша и предоставили по одному варианту ответов.

В процедуру отбора заданий для оценки было внесено одно существенное изменение:

- оценивались только аннотации для релевантных документов (согласно слабой таблице релевантности содержащей данные для этого запроса)

Методологически мы следуем духу оценки в 2008 и 2009 годах и пытаемся оценить аннотацию по двум критериям:

- *Информативность:* критерий характеризует, насколько эта аннотация понятна для принятия решения о полезности документа в контексте этого запроса.
- *Читабельность*: дает ответ на вопрос "Аннотации зачастую состоят из обрывков приложений и отдельных словосочетаний. Мешает ли вам это понимать их смысл?"

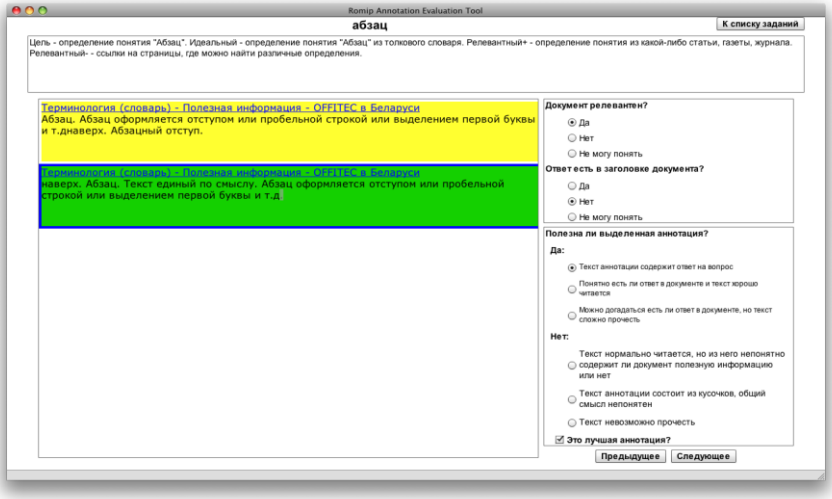

Рисунок 1. Инструмент асессора для дорожки аннотирования.

Но, как и 2009, году оцениваются критерии не прямо, а косвенно, в контексте того, насколько “правильно” эта аннотация передает представление о релевантности документа.

Это было сделано с целью дать асессору возможность «различить» аннотации, даже если формально у них будут одинаковые оценки. Предварительный анализ результатов показывает, что эта возможность полезна для сравнения прогонов – один из прогонов был признан лучшим в 391 случае, а другой в 728. Однако требуется дополнительный анализ, чтобы лучше понять, насколько выводы на основе этой оценки дублируют выводы на основе других вопросов асессору.

Всего оценивалось 1757 пар документ-запрос. Как и в прошлом году оценку производили 2 асессора, и инструкция по оценке в основном осталась прежней (см. Приложение D) с одним дополнением:

- асессор мог (опциональная возможность) пометить один из вариантов аннотаций для заданной пары документ-запрос как “лучшая аннотация”

Несмотря на то, что для оценки были отобраны только релевантные документы около 47% полученных индивидуальных оценок для документов (оценка для документа выставляется на основе всех доступных для этого документа аннотаций) были отмечены как «нерелевантно». При этом в 597 случаях оба асессора согласились, что документ нерелевантный, и в еще в 433 случаях хотя бы один асессор признал документ нерелевантным.

Наблюдаемый процент «противоположных» выводов значительно выше, чем коэффициент несогласия асессоров при оценке дорожек поиска (обычно не превышает 15%), что служит подтверждением как важности качества аннотации при поиске информации, так и осмысленности такого подхода к оценке (оценки привязаны к действиям пользователя и в то же время позволяют различить разные варианты прогонов).

5. Вопросно-ответный поиск

В этом году рассматривалась следующая постановка задачи:

- Система-участник получает коллекцию BY.web и набор запросов в виде вопросов на русском языке.
- Выдачей системы на каждый вопрос является упорядоченный список ответов длиной не более 5.

- Ответ систем состоит из 3 частей - первоисточник, краткий ответ на вопрос, фрагмент текста из первоисточника (до 300 символов), содержащий ответ.

Такая постановка не ограничивает размер “краткого” ответа и в некоторых прогонах, что мы получили для оценки между кратким ответом и фрагментом текста не было разницы.

Набор заданий состоял из 9617 запросов, которые были автоматически из набора заданий, которые используются для других задач поиска. В набор включались все запросы, где использовалось хоть одно из списка вопросительных слов или вопросительный знак.

Из четырех поданных заявок до финиша дошло два участника, которые предоставили 4 прогона.

При отборе заданий для оценки нам пришлось отклониться от случайного выбора в силу следующих причин:

- большинство прогонов содержало ответы лишь на некоторое подмножество вопросов
  (1 прогон – 108 запросов, 2 прогона отвечали на порядка 750 запросов и один на 9200)
- процедура формирования заданий обуславливает значительное число запросов

Процедура отбора заданий для оценки была такова:

- Построен упорядоченный список кандидатов на оценку:
  - множество общих запросов для всех прогонов (~100)
  - случайная выборка запросов (~400)
- Список вручную фильтровался на предмет
  - мусора
  - порно
  - запросов со значительными опечатками (например, слова в неправильной кодировке)
  - очевидно не фактографических запросов (например, цитата фрагмента текста на 300 символов)
  - запросов, не похожих на вопрос
- Всего было оставлено первых 250 запросов (процент отсева был около 50%)
  - непустые списки ответов получились для 246, которые далее и оценивались

В результате проведения оценки мы хотели получить ответы на следующие наши вопросы:

- Осмысленно ли задание
  - Ищут ли факт? Понятно ли вообще задание?
  - Есть ли в коллекции ответ?
- Качество ответа
  - Есть ли он в результате? В расширенном результате?
  - Ответ полный или частичный? Много ли “лишней” информации в ответе?

При этом предполагалось, что асессор не оценивает корректность факта, а только наличие его в тексте. Более подробно с правилами оценки можно ознакомиться в приложении, содержащем инструкцию асессора.

Оценка каждого задания (вопрос с 1-20 вариантами ответа) производилась двумя асессорами, и всего в оценке приняло участие 4 разных асессора. При оценке использовался инструмент, изображенный на рисунке 2.

По результатам оценки оказалось, что только для 14 из 246 заданий были ответы, которые хотя бы один из асессоров признал «точными». И только для 60 заданий из 246 мы предполагаем, что в коллекции есть точный ответ (ответ либо был найден, либо хотя бы один из документов был помечен, как содержащий ответ). Возможно, это обусловлено тем, что участники не смогли обнаружить необходимые документы, но возможно набор запросов не идеально соответствует коллекции. Так, в 15% случаев документы, из которых извлекался ответ, были признаны даже отдаленно не близкими теме вопроса.

В будущем мы бы хотели повысить КПД усилий по оценке и оценивать запросы, для которых ответ можно найти, но у нас нет готового идеального решения. Можно, например, использовать запросы, оценивавшиеся в других дорожках, для которых есть релевантные документы, но таких запросов мало. Другой вариант – составить запросы вручную руководствуясь наличием ответов, и подмешать их в набор заданий, но это чревато появлению искусственных шаблонов в наборе заданий.

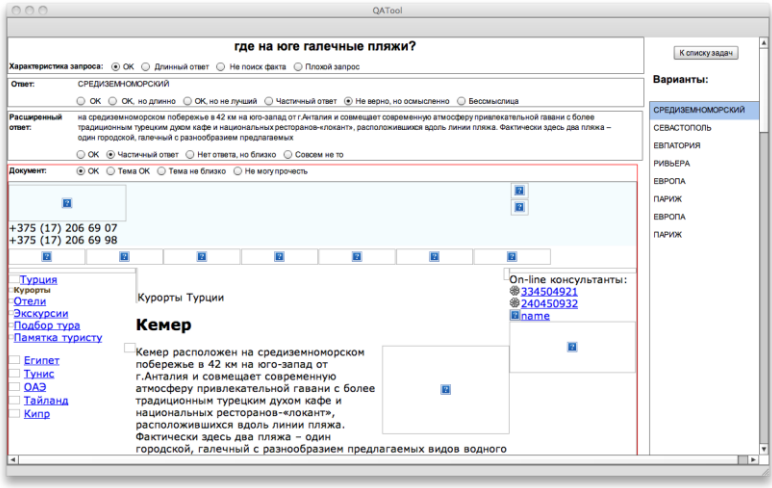

Рисунок 2. Инструмент оценки для дорожки вопросно-ответного поиска.

Список ответов признанных «точными» приведен ниже:

- *где снимали фильм сталкер?*
  РОССИЯ
- *чем питаются мальки рыб?*
  Первые дни новорожденные питаются микроводорослями, через 2 3 суток их потихоньку начинают подкармливать разведенными пекарскими дрожжами, живой «пылью», постепе
- *где отдыхали летом?*
  АЛУШТА
- *что такое акдс?*
  АКДС адсорбированная коклюшно дифтерийно столбнячная вакцина. Адсорбированная коклюшно дифтерийно столбнячная вакцина (АКДС). Прогрессирующие заболевания нервн
- *кто входит в совет федерации?*
  С 1995 г. в Совет Федерации входили губернаторы и спикеры законодательных собраний субъектов РФ. Депутат ГД Алексей Розуван и член Совета Федерации Андрей Ищук подго

- *где сейчас играет футболист марат измайлов?*
  МОСКВА
- *23 февраля, выходной день?*
  23 февраля Ѓ не красный день календаря,. Не красный, знать, не выходной,. Не красный, нет, он золотой. 23 ФЕВРАЛЯ ОТ А ДО Я! День рождения.
- *кто построил парфенон???*
  — Аристотель, Диоген, Фидий, который построил Парфенон в Древней Греции, а потом был жестоко наказан за то, что изобразил там где – то в углу свое собственное лицо: т
- *биография чингиз айтматов умер где?*
  Чингиз Торекулович Айтматов. Чингиз Торекулович Айтматов родился в 1928 г. в кишлаке Шекер в Киргизии. Наиболее известные работы Айтматова: романы И дольше века длит
- *март по гороскопу кто ?*
  ОВЕН
- *водитель-экспедитор кто это?*
  Экспедитор обязан контролировать силами водителя процесс погрузки (выгрузки), включая пересчет грузовых мест, внешнее состояние упаковки, порядок погрузки, крепл
- *2. в какой стране изобрели панамы?*
  ЭКВАДОР
- *кто автор знаменитой статуи дискобол?*
  Мирон Дискобол. 15. Статуя Шэду. 6. Статуя писца Каи.
- *15 апреля кто по гороскопу?*
  ОВЕН
- *кто автор знаменитой статуи дискобол?*
  Знаменитая статуя скульптора Мирона изображает атлета, готовящегося к метанию диска. Знаменитая статуя скульптора Поликлета (датированная второй половиной 5 в. до

Несколько из этих ответов не являются «полными» или «минимальными» и согласно инструкции не должны были быть помечены как «точными». Мы используем этот опыт для уточнения инструкции.

6. Поиск изображений по образцу

Повторение дорожки, добавленной в программу РОМИП в 2008 году, – та же коллекция (Flickr, 20000 фотографий), набор заданий и правил, не изменившийся инструмент оценки (см. Рисунок 3).

Рассматривается задача поиска по содержанию изображений (content-based image retrieval) в коллекции разнородных фотографий типичных для персональных непрофессиональных фотоархивов.

Участникам необходимо было отобрать изображения, похожие на изображение-образец визуально и семантически с точки зрения человека. Релевантными изображениями считались как глобально похожие, так и обладающие локальным сходством.

Отметим, что фотографии в коллекции не связаны с какой-либо дополнительной информацией (такой как аннотации, теги или другой контекст).

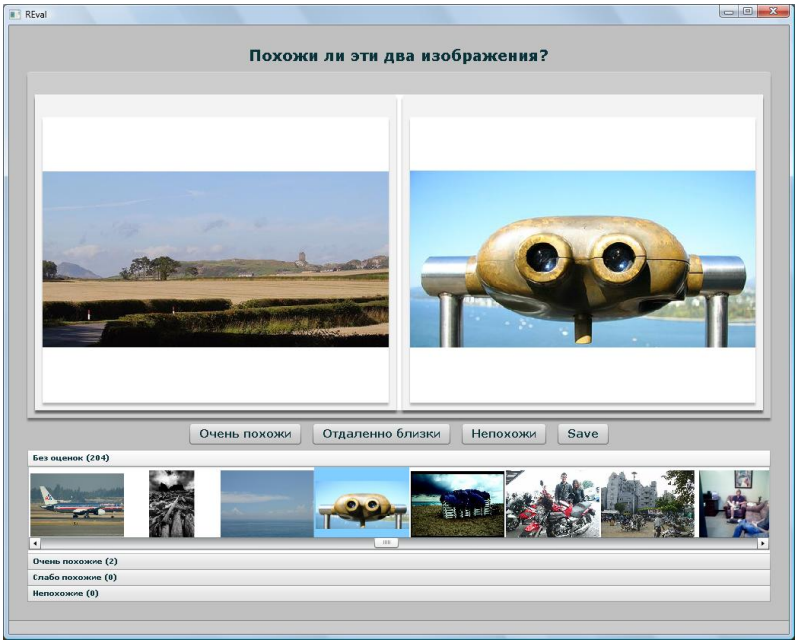

Рисунок 3. Инструмент оценки для дорожки поиска по визуальному подобию.

В связи с наличием трех оценок к двум традиционным схемам слияния оценок (AND и OR) был добавлен промежуточный вариант – VOTE, при котором ответ считается релевантным, если большинство асессоров так считают. Основным отличием правил дорожки этого года является значительно более детальная инструкция асессора (см. Приложение E), составленная при участии всех участников дорожки.

Остальная процедура оценки не изменилась по сравнению с прошлым годом. Для оценки было случайным образом отобрано 250 заданий (то есть изображений образцов). В котлы попало по 20 первых результатов в каждом прогоне. Каждое задание оценивало 3 асессора, и всего собрано 78309 оценок (то есть оценено 26103 задания).

По результатам оценки число слаборелевантных ответов составило – 377 (AND), 1086 (VOTE) и 3052 (OR). Это примерно соответствует доле слаборелевантных ответов в прошлом году – в 2009 году при двух оценках для 36000 ответов число слаборелевантных составило 1473 (AND) и 4403 (OR).

7. Выявление нечетких дубликатов в коллекции изображений

Эта дорожка посвящена оценке методов поиска нечетких естественных дубликатов в коллекции фотографий и в 2009 году являлась повторением дорожки 2008 года. В отличие от прошлого года оценка производилась двумя асессорами, которые не были знакомы с содержимым коллекции (см. Приложение F).

Коллекция состоит из 37800 изображений, полученных путем выборки случайных кадров из 15 часов видеоматериала в нескольких различных разрешениях. Качество изображений варьируется в широких пределах.

По определению дубликатами считаются фотографии одной и той же сцены или объекта, сделанные в разных условиях или разного качества. В частности, дубликатами являются фотографии, снятые в разном масштабе или с разных точек, с различиями в фокусном расстоянии, освещении, с незначительными изменениями фона (движение волны в море или листьев на дереве).

Примеры "естественных" дубликатов из постановки задачи:




Примеры визуально и/или семантически похожих изображений, не являющихся при этом дублями:




Система-участник должна определить имеющиеся группы дублей в коллекции. Допускается, что одно изображение входит в

несколько различных групп дублей одновременно. Ограничений на размеры групп нет.

Оценка производилась путем выделения кластеров дублей из окрестностей 100 случайных изображений (точек в видеопотоке). Из каждой окрестности можно было выделить до 25 кластеров.

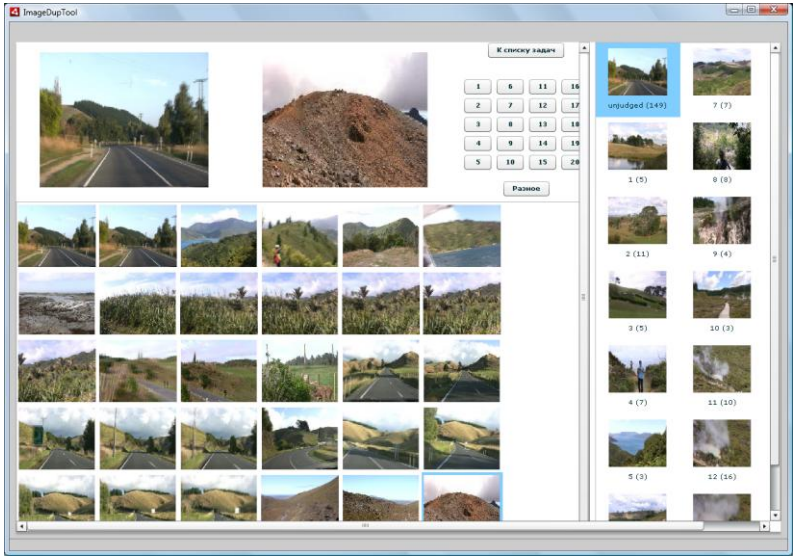

Рисунок 4. Инструмент оценки для дорожки поиска нечетких дубликатов изображений.

Формально окрестность строилась как результат объединения:

- Всех изображений из прогонов участников, которые попали в кластеры с этим изображением.
- Изображений из временной окрестности этого изображения (используя информацию о связи этого изображения с кадрами в исходном видеопотоке).

В оценке принимало участие два асессора. По результатам оценки выяснилось, что позиции асессоров значительно отличаются. Один асессор выделил 1793 группы, а второй только 934.

Поэтому при расчете результатов было решено не пытаться сливать столь непохожие разметки, а рассчитать два набора оценок (с позиции каждого из асессоров). Мы также добавили в расчет оценок прогоны прошлых лет (поскольку они не участвовали в формировании пулов, то их оценки могут быть несколько занижены) и «искусственные» прогоны, имитирующие мнение асессоров.

При этом в силу особенностей алгоритма расчета (сравнение всегда происходит с кластером максимального размера из тех, что содержит данное изображение) оценка прогона асессора по своей же таблице релевантности может не быть 100%.

Результаты подтверждают, что один из асессоров был склонен выделять кластеры меньшего размера, чем другой. Так, в одном случае полнота одного асессора относительно другого – 96.5%, а наоборот, только 65%. Точность же – 40% и 92% соответственно. Но даже полнота в 65% это почти на 20% лучше ближайшего результата.

Интересно, что выводы на основе оценок разных асессоров дают практически такие же тройки лидеров для точности и полноты (дальнейшие списки и другие метрики не сравнивались), хотя абсолютные значения метрик значительно отличаются (для полноты лидер получил оценки 0.68 и 0.46 соответственно). Единственное отличие в порядке у двух прогонов, чьи оценки отличаются на 0.5-1%, что, вероятно, ниже границы статистической значимости результата для этой дорожки.

8. Построение текстовых меток для изображений

Эта дорожка проводилась в 2010 году впервые, и мы специально решили не регламентировать правила очень строго, чтобы дать возможность попробовать разные методы и подходы к решению задачи.

Задача участников дорожки состояла в том, чтобы предоставить набор из не более чем 15 текстовых меток для каждого из изображений тестового набора (всего 2000 тестовых изображений). При этом предполагалось, что метки будут построены на основе анализа содержания изображений (а не извлечены из текста вокруг изображений).

Мы не накладывали жестких ограничений на семантику аннотаций (в будущем, возможно, все же стоит несколько четче регламентировать этот аспект). Предполагалось, что аннотации, к примеру, могут быть построены путем распознавания некоторого фиксированного набора объектов или при помощи классификаторов общего плана (indoor/outdoor, city/landscape, ...) и аннотирования изображений названиями соответствующих им категорий.

Для обучения были разрешены любые подходы. Принципиальным являлось отсутствие единого для всех участников размеченного обучающего множества. В качестве тестовой коллекции использовалась коллекция изображений Flickr.

Рис. 6. Инструмент оценки для дорожки построения текстовых аннотаций для изображений

На участие в дорожке было подано две заявки, однако результаты предоставил только один из заявленных участников.

При планировании дорожки предполагалось, что для оценки будет случайным образом отобрано некоторое подмножество изображений из тестового набора, для каждого из которых будет построен котел из меток, предоставленных участниками.

Однако, фактически результаты были предоставлены только одним из участников и число меток на одно изображение не превышало двух, всего метки были предоставлены для 775

изображений из 2000. Поэтому ввиду небольшого объема оценки было решено оценить все предоставленные метки. Каждая из меток оценивалась двумя асессорами при помощи инструмента, изображенного на рис. 4.

По результатам оценки были вычислены значения для следующих метрик для каждого из тестовых изображений: точность; полнота; число корректных меток, предоставленных участником; общее число меток, предоставленных участником; общее число корректных меток среди всех меток участников; общее число меток в котле. Далее были вычислены макро и микро усреднения оценок, среднее число корректных меток на одно изображение, предоставленных участником.

Значения метрик были рассчитаны по слабой и сильной схеме (OR и AND) объединения оценок асессоров. Отметим, высокую согласованность оценок асессоров, а также довольно высокую точность ответов участника. Так, значение точности при микроусреднении по схеме AND составило 0.7046, по схеме OR – 0.7397.

Заключение

Несмотря на высокий процент схода заявившихся участников в 2010 году, семинар продемонстрировал рост по ряду параметров – появились новые дорожки и участники, вырос объем и качество оценки, обновились инструменты и методология. Организация семинара все больше децентрализуется, что добавляет ей стабильности.

Однако, общее количество финишировавших участников для каждой из дорожек в этом году оказалось невелико. Надеемся, это временная проблема и в следующем году ее не будет.

Семинар не мог бы состояться без активного участия и помощи многих людей – членов оргкомитета, участников, асессоров и многих других. Мы вам очень за это признательны и надеемся на дальнейшее сотрудничество!

Литература

[1] Труды РОМИП онлайн. http://romip.ru
[2] И.Некрестьянов, М. Некрестьянова. РОМИП'2006: отчет организаторов. РОМИП'2006.
[3] Li Chen, F. W.M. Stentiford. Comparison of Near-Duplicate Image Matching. Visual Media Production, 2006. CVMP 2006, p. 38-42, 2006.

ROMIP 2010:
Report from Organizers

Igor Nekrestyanov, Marina Nekrestyanova

This report describes details of ROMIP'2010 evaluation activities from organizers perspective. We briefly describe track rules and focus on specifics of evaluation procedures used for last yearly cycle.