Особенности организации и проведения РОМИП'2008

© И. Некрестьянов, М. Некрестьянова

romip@romip.ru

Аннотация

В статье описаны детали организации РОМИП'2008 – дорожки, коллекции, процедуры оценки результатов и другие аспекты проведения семинара. Основное внимание уделено особенностям РОМИП'2008. Подробности о принципах РОМИП и базовых подходах к оценке можно найти в трудах РОМИП прошлых лет [1], где они неоднократно подробно описывались.

1. Введение

В 2008 году оргкомитет получил заявки на участие от 19 исследователей, исследовательских коллективов и компаний, из которых до финиша добралось 14 участников. Подробная информация о заявках, и полученных результатах приведена в таблице 1.

Набор коллекций РОМИП в этом году расширился за счет появления двух новых коллекций изображений (см. Разделы 5 и 6). Текстовые коллекции не изменялись. Сводная статистика о наборе коллекций РОМИП приведена в таблице 2.

Основными особенностями РОМИП'2008 являются:

- Появление в программе семинара дорожек, посвященных задачам поиска изображений, которые вызвали довольно большой интерес и потребовали значительных усилий по подготовке коллекций, правил и оценке результатов;
- Экспериментальное изменение подхода к оценке результатов поиска по текстовому запросу на примере коллекции BY.WEB – оценивалось значительно большее число запросов с меньшей глубиной пула;
- Новый подход к оценке результатов дорожки контекстно-зависимого аннотирования;

- Отмена дорожки поиска похожих документов по документу-образцу в связи со сходом заявившихся участников;
- Завершение оценки дорожки кластеризации новостного потока ДО очного семинара ☺
- Массовое привлечение “временных” асессоров к оценке;
- Активная реакция участников РОМИП на результаты оценки. В оргкомитет поступила масса критики, советов, предложений и сообщений об ошибках.

В последующих разделах мы подробно расскажем об особенностях проведения дорожек РОМИП в этом году.

Дорожка	Заявившихся участников	Предоставивших результаты	Общее число прогонов
Поиск по By.Web	10	6	15
Поиск по KM.RU	10	7	15
Поиск по нормативной коллекции	7	6	6
Поиск по смешанной коллекции	5	1	1
Поиск по документу-образцу	5	0	-
Классификация нормативных документов	4	3	5
Классификация Веб сайтов	4	2	7
Классификация Веб страниц	4	2	5
Кластеризация новостного потока	3	2	3
Контекстно-зависимое аннотирование	2	2	3
Поиск по визуальному подобию	3	3	6
Поиск нечетких дубликатов изображений	5	4	6

Таблица 1. Сводная статистика о РОМИП'2008

2. Текстовый поиск

В 2008 году в программе РОМИП было заявлено 4 дорожки, которые присутствовали в программе РОМИП и в прошлые годы:

- Классическая задача поиска по запросу по
  - коллекции нормативно-правовых документов;
  - Веб коллекции;
  - смешанной коллекции.
- поиск похожих документов по документу-образцу или фрагменту текста

По факту программа несколько отличалась от запланированного. Дорожка по поиску по Веб коллекции распалась на две – поиск по коллекции BY.Web и поиск по коллекции KM.RU, а дорожка поиска похожих документов была отменена в связи со сходом заявившихся участников.

2.1 Поиск по коллекции нормативных документов

В отличие от РОМИП'2007 в этом году оценка результатов дорожки поиска по нормативно-правовой коллекции производилась экспертами с юридическим образованием.

Для каждой пары документ-запрос была собрана одна оценка (за небольшим числом исключений, обусловленным техническими причинами).

Отбор запросов для оценки производился следующим образом. Экспертам были предоставлены два списка:

- 50 прошлогодних запросов
- 300 случайно отобранных ранее не оцениваемых запросов (без какой-либо фильтрации)

Экспертов попросили отобрать 25 запросов из первой группы и 75 из второй, руководствуясь следующими принципами:

- не выбирать запросы с опечатками;
- выбирать запросы, по которым эксперту понятно, что искали

Во второй группе асессоров попросили отобрать:

- 25 запросов на поиск с каким-то упоминанием номера документа (статья такая-то, и т.д.), предпочтительно разных, чтобы нельзя было выделить один и тот же шаблон, и менялся бы тип информационной потребности. Например:
  - приказ ФТС России от 29 ноября 2006 года N 1252
  - Комментарий ст.14.5 КоАП РФ
  - Форма КС-2

- 50 запросов, где смысл можно понять из текста (не требуется применение шаблонов и т.п., чтобы понять определить правильный ответ)

Для всех отобранных запросов экспертов составили расширенные описания, которыми руководствовались асессоры при оценке (множество асессеров и экспертов сильно пересекалось).

По техническим причинам 5 запросов было исключено, и оценка производилась по 95 запросам. Глубина пула по этой дорожке в 2008 году была 35 документов, что обусловленно сжатыми сроками и временем, которое асессоры могли посвятить участие в оценке.

2.2 Поиск по Веб коллекции

В РОМИП'2007 предполагалось, что участники будут выполнять поисковые задания на объединении коллекций KM.RU и BY.WEB, но на практике получилось, что многие участники использовали только одну из двух коллекций. Кроме того, оценивавшиеся запросы могли плохо подходить содержимому коллекции, так как происходили из лога запросов для другой коллекции и в этой коллекции тематика запроса могла быть совсем не отражена.

В 2008 году оценка дорожки поиска по Веб-коллекции дорожка была разделена на два отдельных задания – один и тот же набор запросов надо было выполнить для каждой из коллекций отдельно. Набор запросов был расширен на почти 10000 запросов (до 29231) за счет дополнения случайной выборкой запросов к Яндекс с белорусских IP адресов.

Изменились также и другие аспекты процедуры проведения оценки. В частности, при оценке задания использовались только запросы, которые происходили из журнала запросов соответствующего данной коллекции.

Новая инструкция для асессоров включала описание и примеры нескольких типов разных информационных потребностей (поиск информации, поиск ресурса и др., 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.

Перед началом работы над заданием и после его окончания, асессоры заполняли небольшую анкету, характеризующую их понимание задания.

2.2.1 Поиск по BY.Web

Особенностью этого задания было значительно большее число оценивавшихся запросов, а именно 500 (что почти на порядок больше, чем в прошлом году). Рост числа запросов напрямую влияет на объем работ по оценке. Чтобы закончить оценку в реалистичные сроки, глубина пула была сокращена до 20.

Запросы для оценки отбирались случайным образом, а потом производилась их легкая фильтрация (нечитаемая кодировка, порнографические запросы). Всего было отобрано 600 запросов, из которых было отфильтровано около 40. Оценивались первые 500 из оставшихся.

При оценке задания каждый из асессоров должен был составить свое описание информационной потребности, соответствующей запросу. Первая версия описания составлялась до начала оценки, то есть асессор видел только текст запроса, но не видел ни одного из оцениваемых документов. Асессор мог исправить расширенное описание по окончании выполнения оценки этого запроса.

2.2.2 Поиск по KM.RU

Правила проведения этого задания во многом повторяют правила проведения дорожек по поиску прошлых лет.

Оценивалось 60 запросов, глубина пула – 50. Из случайного набора отсеивались мусор, опечатки, явно навигационные запросы на что-то, чего нет в коллекции (например, одноклассники.ру). Благодаря более строгой фильтрации запросов всего для одного запроса оценка не была произведена из-за непонимания исходной информационной потребности асессорами.

При оценке этой дорожки расширенное описание для задания составлял первый оценивавший его асессор, а второй асессор должен был переиспользовать это описание.

Интересно, что согласованность оценок асессоров для этого задания выше, чем при оценке поиска по By.Web (0.9 против 0.8).

2.3 Поиск по смешанной коллекции

Оценка, как и в предыдущие годы, была совмещена с другими поисковыми дорожкам и по тем же запросам, что использовались для оценки этих дорожек. Однако, мы решили отказаться от совмещения оценки с BY.WEB, поскольку это повлекло бы слишком заметное дополнительное увеличение объема оценки - прибавилось бы 3000 новых пар документ-запрос к оцениваемым 50000. То есть

всего оценка производилась для двух групп запросов – 95 запросов по нормативной коллекции, 60 запросов по КМ.

3. Текстовая классификация

3.1 Нормативных документов

По результатам обсуждения с участниками обучающее множество было расширенно и включало информацию о принадлежности упоминающихся документов ко всем (а не к одной как раньше) категориям, к которым они относились в эталоне от Кодекс. То есть в обучающем множестве один и тот же документ мог относится к более чем одной категории.

Оценка была произведена по 74 случайно выбранным категориям, методом сравнения с рубрикацией предоставленной компанией Кодекс вместе с коллекцией. При вычислении оценок НЕ учитывались документы, которые были использованы для обучения по данной категории (и такие документы НЕ включены в итоговую таблицу релевантности).

3.2 Веб сайтов

Оценка производилась по 15 категориям, по три категории для 5 категорий первого уровня. Два асессора на каждое задание.

При построении заданий для асессоров выяснилось, что включение сайтов полностью приводит к получению очень больших архивов для некоторых из оценивавшихся категорий (более 200Мб). Такой размер архивов оказался проблематичен как для распространения, так и для оценки.

Для того чтобы разрешить эту проблему, было принято решение включать сайты не целиком, а в сокращенном виде. Сокращение касалось только сайтов, для которых в коллекции присутствует более 200 документов. 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 образцу для каждой папки верхнего уровня или скрипта (то есть для каждой строки получаемой из url путем обрезания по первому символу ‘/’ или ‘?’ после имени сайта).

3.3 Веб страниц

В отличие от предыдущих лет в правилах этой дорожки было следующее изменение: ответ должен был состоять из

упорядоченных по близости к теме рубрики списков документов для каждой рубрики.

Оценка производилась по 15 категориям. Оценивалось только 75 первых документов для каждой категорий с каждого прогона. Поэтому не для всех возвращенных документов из каждого прогона систем есть оценки релевантности, что, безусловно, сказывается на общих оценках систем.

Для получения более полной картины мы как всегда рассчитали два набора оценок:

- Учитывались все документы и те документы, для которых нет оценки, считались нерелевантными.
- Учитывались только оцениваемые документы. Документы, для которых нет оценок, не учитывались (как будто система не относила их к оцениваемым категориям).

Отметим, что в таблице релевантности РОМИП'2008 по этой дорожке значительно больше релевантных документов, чем в предыдущие годы. Вероятно, это обусловлено изменением формата ответа, в результате чего в котле для оценки попадают наиболее вероятные ответы с точки зрения систем-участников, и доля ошибок в верхушке это списка ниже, чем для почти случайной выборки из результата работы системы.

4. Контекстно-зависимое аннотирование

Набор заданий для этой дорожки был построен на основе таблиц релевантности дорожек по поиску РОМИП'2007. Всего было 20510 заданий, соответствующих 105 разным запросам к коллекции нормативных документов и Веб коллекции.

Опыт организации этой дорожки в прошлые годы показал, что традиционное сравнение предсказания о релевантности документа по его аннотации и истинной аннотации документа приводит к получению очень близких оценок для разных прогонов и участников, что затрудняет их дальнейший анализ [2]. Поэтому в 2008 году мы решили попробовать новый подход.

4.1 Отбор заданий для оценки асессорами

Отбор заданий производился следующим образом:

- были отобраны задания, соответствующие коллекциям KM и BY;

- для каждого запроса было выбрано 50 случайных заданий (документов, которые необходимо было аннотировать);
- для всех этих документов была запущена процедура извлечения названий документов из тега title. Все задания, для которых тег title был пустой (или его не удалось выделить), были удалены из рассмотрения.

Всего оценивалось 1896 заданий, соответствующие 60 запросам.

4.2 Процедура оценки

Оценка была организована следующим образом:

- Задания группировались в наборы по запросам. Набор состоит из последовательности заданий, каждое из которых соответствует одному документу.
- Асессор видел ВСЕ доступные аннотации в случайном порядке (заголовок документа + текст аннотации, обрезанный до 300 символов).
- Для каждой аннотации необходимо было выставить 2 оценки:
  - информативность
  - читабельность
- Асессор также должен был ответить на два вопроса, характеризующие его понимание о релевантности документа и о роли заголовка:
  - исходя из полученной информации, считаете ли вы, что документ содержит релевантную информацию?
  - приняли ли бы вы такое же решение, используя ТОЛЬКО заголовок документа?

Оценка каждого задания производилась двумя асессорами с использованием инструмента оценки, изображенного на рисунке 1.

Оценки по критериям информативность и читабельность выставлялись по трехбалльной системе – ПЛОХАЯ, ХОРОШАЯ, ОТЛИЧНАЯ. Однако, технически асессору было доступны промежуточные значения (шкала имела 9 градаций), чтобы он мог указать на небольшое превосходство одного варианта аннотации над другим.

Информативность характеризует, насколько эта аннотация понятна для принятия решения о полезности документа в контексте этого запроса.

Критерий читабельности определял ответ на следующий вопрос “Аннотации зачастую состоят из обрывков приложений и отдельных словосочетаний. Мешает ли вам это понимать их смысл?”.

При этом мы полагали, что высокой информативностью и читабельностью могут обладать и аннотации для нерелевантных документов, так как цель аннотации помочь пользователю принять правильное решение о полезности документа.

Полную инструкцию асессора по этой дорожке можно найти в приложение D.

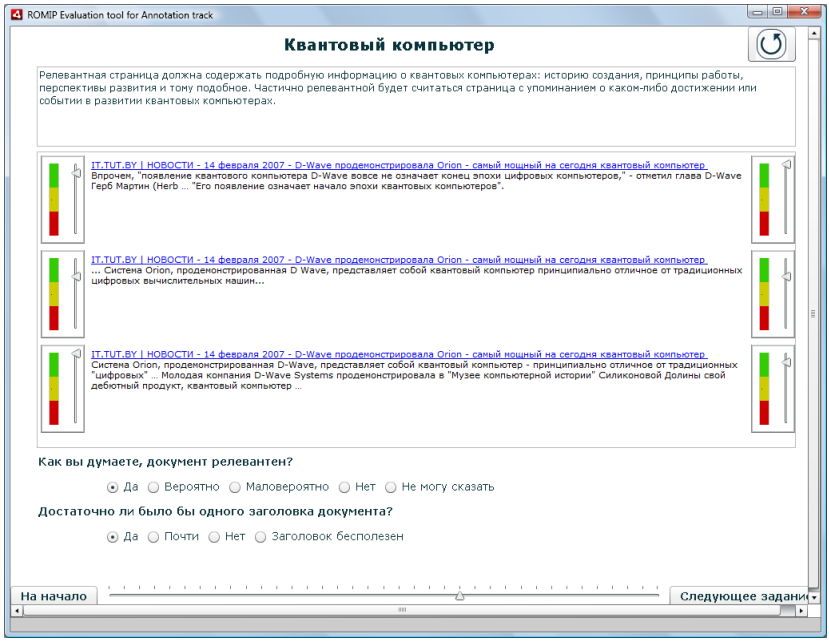


Рисунок 1. Инструмент оценки для дорожки контекстно-зависимого аннотирования.

4.3 Метрики

При вычислении метрик оценки асессоров (для информативности и читабельности) были отображены в трехзначную шкалу 1/2/3 (градации 1-3 в 1, 4-6 в 2, 7-9 в 3).

При объединении оценок использовалось 2 подхода:

- [min] выбор минимальной оценки
- [max] выбор максимальной оценки

Расчет метрик производился как на всем наборе оценок, так и на нескольких сужениях:

1. все оцененные документы;
2. только релевантные документы по каждой из шкал OR и AND для adhoc дорожки;
3. только нерелевантные документы по каждой из шкал OR и AND для adhoc дорожки;
4. документы, для которых предсказание релевантности по аннотации совпадает с "истинной" релевантностью (из таблицы релевантности прошлых лет)
5. для каждого запроса в отдельности;

В качестве метрик использовались число оценок каждого типа и средние оценки по каждому из критериев.

4.4 Первые наблюдения

Мы еще не успели проанализировать результаты оценки, но уже можно отметить несколько первых наблюдений:

- Оценки разных прогонов разнятся значительно существеннее, чем при использовании предыдущего подхода к оценке.
- У оргкомитета и некоторых участников было разное исходное понимание того, могут ли быть полезными аннотации для нерелевантных документов.
- Задача оценки для асессоров значительно сложнее и были случаи, когда асессор неправильно понимал постановку задачи, несмотря на подробную инструкцию и примеры.

Тем не менее, нам кажется, что опыт получился познавательным и успешным. Мы планируем более подробно изучить результаты оценки в будущем.

4. Кластеризация новостного потока

Эта дорожка уже не первый год присутствует в программе РОМИП, но 2008 год стал первым годом, когда оценка этой дорожки была наконец-то проведена до проведения очной части семинара!

Основной причиной таких проблем с оценкой этой дорожки является отсутствие четкого понимания, как правильно оценивать результаты по этой дорожке. В докладе об организации РОМИП

2007 года мы описывали возможные подходы к оценке с точки зрения позиций “читателя” и “редактора”.

Предыдущие попытки использовать подход “редактора” или гибридные подходы не увенчались успехом. Задача разбиения всей коллекции слишком сложна для асессора, и поэтому он не может выполнить работу качественно. Во многом, по-видимому, это проблема, связанная с используемыми инструментами оценки.

В этом году мы несколько изменили подход к оценке и использовали новый инструмент оценки (см. Рисунок 2). Задачей асессора было все также выделить события/сюжеты, определение которых осталось неизменным.

Однако, были и важные отличия:

- Асессор видел ленту всех сообщений за неделю, без привязки к ответам участников.
- Асессор **мог** использовать поиск по заголовкам.
- Приоритетом при оценке было невыявление максимального количества дублей/сюжетов или полная разметка непрерывного интервала времени, а полнота выявления всех сообщений, относящихся к данному информационному дублю/сюжету.
- Асессор сам решал, какие дубли/сюжеты он выделяет.
- Асессор мог строить иерархию произвольной высоты (не обязательно уровня два).
- Асессор присваивал событиям и кластерам символьные имена, которые служили для того, чтобы ему было проще ориентироваться среди уже созданных кластеров.
- Асессор явным образом помечал кластеры-"события" (то есть дубли)
- Следуя этому подходу, была проведена оценка недели shevard одним асессором. Построенное дерево содержало 337 листьев и охватывало 707 сообщений.
- Отметим, что процедура оценки подразумевает, что сообщения, не включенные в построенное дерево, не относятся ни к одному из кластеров в дереве (если не учитывать погрешности асессора). Это делает оценки полноты более достоверными.

Отметим также, что символьные метки, присвоенные узлам, могут быть использованы в дальнейшем для анализа результатов, так как они отражают логику решений асессора.

Поскольку результатом работы асессора является многоуровневое дерево, не повторяющее в точности условие задачи для этой дорожки, то для вычисления метрик прогоны и идеальное разбиение трансформировались в общее пространство. Было рассчитано 3 набора оценок на основе следующих таблиц релевантности (задающих одноуровневое разбиение):

- **Events**: кластером являлась группа сообщений, отнесенных асессором к одному событию (дублю).
- **Topics**: кластер - группа сообщений, отнесенных асессором к одному узлу в дереве, включая все сообщения из непосредственных потомков, соответствующих событиям.

  Условно говоря, это "узкие сюжеты, куда в явном виде включены информационные дубли".
- **BroadTopic**: кластер объединяет все сообщения, которые отнесены асессором к одной группе на верхнем уровне.

При вычислении оценок оказалось, что участники по-разному понимают постановку задачи. А именно, может ли одно и то же сообщение относится к нескольким кластерам. В результате, было рассчитано два набора оценок:

- При вычислении оценки, в качестве кластера, куда в прогоне отнесено сообщение m, использовался случайный кластер из прогона, содержащий m.
- Вместо случайного кластера использовалось множество всех кластеров из этого прогона, которые содержали m.

При расчете оценок мы использовали все поданные прогоны по этой дорожке за 2006-2008 года. Сводные результаты приведены в таблице 2 (макроусреднение точности и полноты для каждого способа оценки).

	Events		Topics		BroadTopics	
Прогон	P	R	P	R	P	R
1	0.574	0.500	0.533	0.704	0.890	0.085
2	0.333	0.724	0.809	0.610	0.784	0.192
3	0.386	0.659	0.796	0.595	0.799	0.152
4	0.293	0.472	0.705	0.606	0.577	0.098
5	0.672	0.443	0.825	0.567	0.889	0.080
6	0.672	0.443	0.727	0.612	0.889	0.080
7	0.996	0.233	0.996	0.477	0.996	0.020

Таблица 2. Сводные результаты за 2006-2008 год по дорожке кластеризации новостного потока.

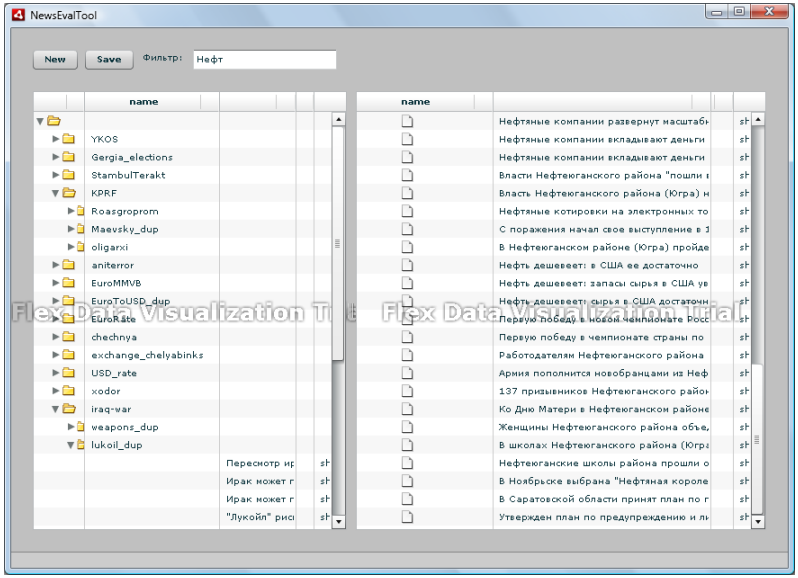


Рисунок 2. Инструмент оценки для дорожки кластеризации новостного потока.

5. Поиск изображений по образцу

Эта дорожка впервые проводилась в 2008 году и поэтому необходимо было решить полный набор вопросов по ее организации:

- Какими должны быть правила?
- Как построить коллекцию?
- Какой нужен инструмент оценки и что оценивать?

5.1 Постановка задачи

Дорожка посвящена оценке методов решения задачи поиска по содержанию изображений (content-based image retrieval) на коллекции разнородных фотографий, типичных для персональных непрофессиональных фотоархивов.

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

Изображения глобально похожи, если на них представлены одинаковые сцены (например, два снимка ночного города).

Изображения обладают локальным сходством, если на них представлены похожие объекты на разном фоне.

Глобальное сходство играет решающую роль в случае, когда на изображениях сложно выделить центральный объект (в основном, пейзажные фотографии), в то время как локальное сходство важно для фотографий с явно выраженным объектом съемки (портретное фото, съемка животных).

Для локального сходства не требуется идентичности объектов на различных снимках, объекты должны быть одного вида, одной природы. Так, два портретных изображения разных людей могут быть признаны похожими при наличии некоторого визуального сходства (одна поза, одинаковая длина и цвет волос, и т.д.).

Ниже приведены примеры похожих изображений:




Примеры изображений, обладающих частичным сходством:



Примеры изображений, не обладающих визуальным или семантическим сходством в должной мере:




5.2 Коллекция Flickr

Коллекция была подготовлена Натальей Васильевой, на основе договоренности с Flickr об использовании материалов в исследовательских целях (при условии обязательно ссылки на источник данных).

Тестовая коллекция состоит из 20000 фотографий без единой темы и разного качества. В коллекции есть фотографии, сделанные в помещении и на улице, включая портреты, пейзажы, городские сцены и другие типы фотографий. Размерность картинок не превышает 500 пикселов (типичный размер 500x375).

Фотографии не связаны с какой-либо дополнительной информацией (такой как аннотиции, теги или другой контекст). Коллекция иммитирует задачи поиска в персональных коллекциях непрофессиональных фотографов.

5.3 Оценка

Оценка производилась во многом аналогично оценке для задач текстового поиска. Для оценки было случайным образом отобрано 250 запросов (то есть изображений образцов). В котлы попало по 20 первых результатов в каждом прогоне. Каждое задание оценивало 2 асессора, используя инструмент изображенный на рисунке 3.

В процессе расчета оценок выяснилось, что часть участников включала изображения образец в ответ. Поэтому, было решено расчет окончательных оценок произвести после исключения изображений образцов из ответов участников и таблицы релевантности. Для того чтобы все прогоны были в равных условиях и имели равное число результатов, то расчет оценок проводился на глубине 19.

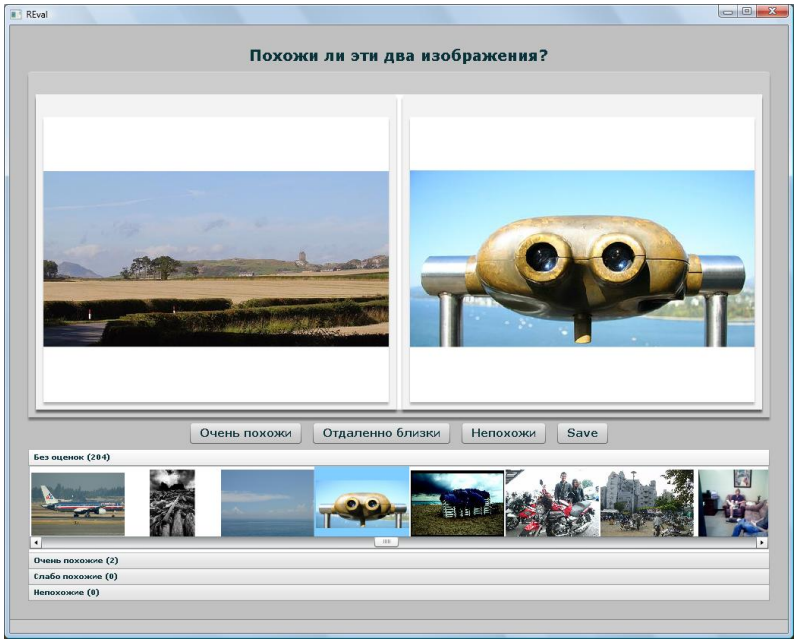


Рисунок 3. Инструмент оценки для дорожки поиска по визуальному подобию.

6. Выявление нечетких дубликатов в коллекции изображений

Эта дорожка также впервые проводилась в 2008 году и для ее проведения необходимо было разрешить те же вопросы, что и для дорожки поиска изображений по визуальному подобию:

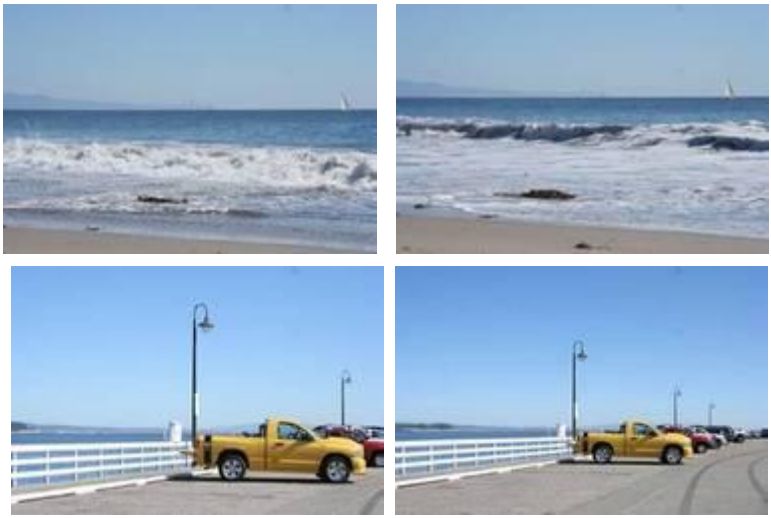
- Какими должны быть правила?
- Как построить коллекцию?
- Какой нужен инструмент оценки и что оценивать?

Мы подробно обсудим этим вопросы в последующих разделах.

6.1 Постановка задачи

Дорожка посвящена оценке методов поиска дубликатов в коллекции фотографий. 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:



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




Первоначальная постановка задачи звучала следующим образом:

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, но оцениваться будут только группы из верхушки списка, отсортированного по убыванию размера групп.

Однако, при проведении оценки последнее ограничение не соблюдалось.

Предполагалось, что коллекция будет содержать большое число "естественных" дублей, что отличает рассматриваемую задачу от задачи поиска трансформированных изображений.

6.2 Коллекция

Создание коллекции для этой дорожки неожиданно оказалось нетривиальной задачей.

Планировалось, что хорошую коллекцию можно получить, объединив результаты поиска нескольких поисковых систем, которые ищут изображения в Веб. Однако, выборочный анализ построенной таким образом коллекции показал, что ее вряд ли можно назвать удачной коллекцией для этой дорожки.

В результате, был использован следующий подход. Коллекция была построена на основе коллекции персональных видеофильмов, методом записи случайных стоп-кадров. Выбор кадра производился в два шага:

- Из каждого интервала в 30 секунд выбирался один опорный кадр.
- Из последующих 20 секунд выбиралось случайное число кадров (от 0 до 20).

Всего использовалось около 15 часов видеоматериала, которое также было продублировано в нескольких разрешениях (получен с камеры) - 720x576, 352x288 и 176x144 и построенная коллекция содержит 37800 изображений.

Построенная таким образом коллекция, конечно, содержит довольно много изображений низкого качества, но в то же время представляет собой интересную коллекцию естественных дублей.

К тому же, информация о связи изображения с исходным видео (о положении кадра во времени) может быть использована для грубой оценки качества результата. Маловероятно, что разнесенные во времени изображения действительно являются дубликатами.

6.3 Оценка

Во многом, проблемы при оценке этой дорожки схожи с проблемами при оценке результатов кластеризации новостного потока. Конечно, решение о схожести изображений принять легче, но и объем коллекции здесь больше.

Для оценки результатов был выбран следующий подход. Асессор выделял кластеры в окрестностях 45 случайно выбранных 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).

Построенный таким образом котел анализировался асессором, целью которого было выделить до 20 групп дубликатов. При этом:

- размер группы не ограничивался;
- групп могло быть меньше;
- не обязательно было оценивать все изображения, даже еще оставались неиспользованные группы.

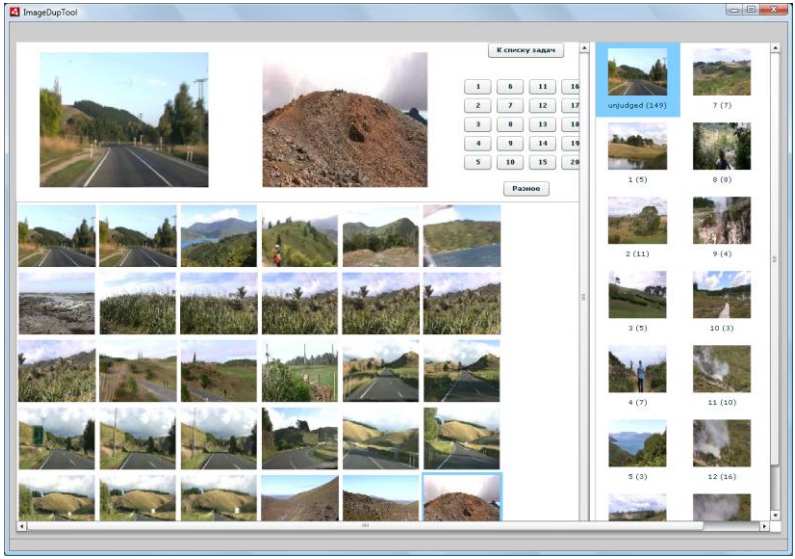


Рисунок 4. Инструмент оценки для дорожки поиска нечетких дубликатов изображений.

Основная идея такого подхода была в том, что относительно большие кластеры в таком котле должны быть выделены более-менее полно (за счет временной окрестности). Точность же выделения “маленьких” кластеров должна быть ниже – это могут быть случайно попавшие в этот котел картинки, дубли которых могут существовать и просто не попали в этот котел.

Для идентификации полученных кластеров были автоматически выбраны "изображения-маркеры" групп. В качестве маркеров использовались изображения, которые имеет "среднюю" временную метку среди элементов группы.

В оценке принимал участие один асессор и всего было получено 526 групп, включающих 3765 изображений.

Отметим, что такой подход к оценке не гарантирует:

- отсутствия пересечений между построенными кластерами. Хоть в качестве отправной точки и использовались разнесенные во времени стартовые изображения, но у нас нет контроля над происхождением изображений, которые попадают в котел из прогона участника.

- того, что асессор просмотрел все картинки из всех затронутых кластеров из ответа участника. Это, безусловно, добавляет погрешности полученным оценкам.

Отметим также, что выделение четких границ нечетких дубликатов во многих случаях нетривиальная задача для асессора. Соседние во времени изображения часто можно назвать дубликатами, но это отношение не является транзитивным и сложно сформулировать, в какой момент накопленная разница становится критической, чтобы признать их разными и как правильно разбить входное множество. Возможно, стоило разрешить асессорам относить одно и то же изображение к нескольким группам при оценке одного котла.

Заключение

РОМИП'2008 вырос - в числе дорожек, объеме оценки, количестве заявок и финишировавших участников. Появились не только новые задачи, но и новые подходы к оценке ранее рассматривавшихся задач.

Конечно, были и проблемы – традиционные отставания от расписания, усугубленные ростом объема работ по оценке, что обусловило использование предварительных результатов оценки для некоторых дорожек при подготовке трудов РОМИП'2008.

Из позитивных тенденций также хочется отметить повышение активности участников в этом году. Конечно, это доставило нам больше хлопот, но это были приятные хлопоты. Надеемся, что в будущем участники РОМИП будут еще больше влиять и непосредственно участвовать в работе семинара (с самых ранних стадий).

Несмотря на разнообразные трудности и накладки, мы считаем, что в РОМИП'2008 нам удалось сделать шаг вперед в развитии семинара, и надеемся, что участники семинара разделяют наше мнение.

Благодарности

Мы хотим выразить благодарность всем, кто активно помогал проводить семинар: участвовал в создании новых коллекций, совершенствовании правил проведения семинара, помогал находить ошибки в инструментах оценки и участвовал в оценке результатов.

Отдельное спасибо Максиму Губину, чей вклад в существование НП РОМИП невозможно переоценить.

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 2008 Evaluation: Rules, Methodology and Adhoc Decisions

Marina Nekrestyanova, Igor Nekrestyanov

This report describes details of ROMIP'2008 evaluation activities from organizers perspective. We focus on specifics of this year – new tracks, new collections, new evaluations methodologies, statistics, complications and adhoc decisions.