РОМИП'2007: отчет организаторов

© И. Некрестьянов, М. Некрестьянова

romip@romip.ru

Аннотация

В статье описаны детали организации РОМИП'2007 с упором на особенности организации семинара в 2007 году. 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 2007 году оргкомитет получил 11 заявок на участие, из которых до финиша добралось 8 участников. Подробная информация о заявках, и полученных результатах приведена в таблице 1.

Как видно из таблицы, несколько участников не дошло до финиша и из-за этого ряд дорожек был отменен. Большое число опозданий привело к тому, что труды не были готовы к началу семинара.

Но были и позитивные тенденции: возросло число заявок на получение доступа к наборам данных РОМИП со стороны сторонних исследователей; в арсенале РОМИП появились три новые большие текстовые коллекции и новая большая таксономия для классификации нормативных документов.

Далее мы расскажем о некоторых этих и других аспектах организации семинара более подробно.

2. Новые текстовые коллекции

Благодаря помощи компаний Кодекс, КМ онлайн и Яндекс, у РОМИП в 2008 году появилось три новых коллекции.

Компания Кодекс предоставила базу нормативно правовых документов “Законодательство РФ, Москвы и Санкт-Петербурга” по состоянию на декабрь 2006. Эта коллекция является обновлением коллекции Legal-2004, но имеет ряд важных отличий:

- Значительно больший объем. В коллекции около 300.000 документов объемом около 1.7Гб в сжатом виде.
- Наличие документов с большим количеством разных редакций (“нечетких” дублей).
- Новая крупная таксономия с 1904 рубриками и информация о принадлежности документов коллекции рубрикам (более полутора миллионов пар документ-рубрика).

Расширились и Веб коллекции. Компания “КМ онлайн” предоставила коллекцию, основанную на содержимом портала KM.RU (около 90% по состоянию на май 2007 года). Коллекция основана на 57 сайтах, содержит более 3.000.000 документов общим объемом 13.7Гб сжатого текста.

Дорожка	Заявившихся участников	Предоставивших их результаты	Общее число прогонов
Поиск по по Web	6	5	8
Поиск по нормативной коллекции	4	4	5
Поиск по смешанной коллекции	3	1	1
Поиск по документу образцу	4	0	0
Классификация нормативных документов	4	1	1
Классификация Веб сайтов	4	2	5
Классификация Веб страниц	4	2	5
Кластеризация новостного потока	4	2	3
Контекстно-зависимое аннотирование	1	1	1

Таблица 1. Сводная статистика о РОМИП'2007

С помощью компании Яндекс была построена коллекция BY.Web, которая состоит из выборки сайтов из домена .by по состоянию на май 2007 года. С каждого сайта брались ссылки на глубину 3 от стартовой. Процент ссылок внутрь коллекции в этой коллекции достигает 25% и есть надежда, что ее можно использовать для экспериментов со ссылочным ранжированием. Коллекция содержит 1.500.000 документов, Общим объемом 8Гб сжатого текста.

3. Поисковые дорожки

В 2007 году состоялось три дорожки по поиску текстовой информации:

- Поиск по Веб коллекции;
- Поиск по нормативной коллекции;
- Поиск по смешанной коллекции.

Также была заявлена дорожка поиска по документу-образцу, но она была отменена в связи с отказом участников.

Поисковые дорожки использовали новые коллекции, а задания для них отбирались из журналов соответствующих поисковых систем. Всего было 19628 заданий для Веб дорожки и 14797 для дорожки поиска по нормативной коллекции. Набор заданий для поиска по смешанной коллекции состоял из объединения этих заданий.

При постановке задачи для дорожки поиска по Веб коллекции предполагалось, что участники будут выполнять поиск по объединению коллекций By.Web и KM.RU. Однако, на практике получилось, что часть участников следовала правилам, а часть сдала ответы для одной из коллекций. Это затрудняет сравнение разных систем.

Для того чтобы упростить ситуацию с анализом результатов мы рассчитали результаты, как для объединения, так и их сужения на каждую из Веб коллекций. В набор рассчитываемых метрик была также включена метрика bpref [2], предназначенная для использования в условиях, когда нет полной информации о релевантности всех документов. Интерпретацию же результатов мы оставили участникам.

Интересно, что при оценке мнения разных асессоров о релевантности документов By.Web совпадали несколько реже, чем по KM.RU. Соотношения количества слабо/сильнорелевантных документов (то есть релевантных по мнению одного из асессоров

или обоих асессоров) для этих коллекций - 1329/660 и 976/238 соответственно.

К сожалению, в 2007 году нам не удалось привлечь экспертов-юристов к оценке заданий по нормативной коллекции, и эта дорожка оценивалась асессорами без юридического образования.

Для повышения качества результата оценки, были предприняты следующие шаги:

- Расширенные описания, “понятные” обычным асессорам.
- Переиспользованы описания составленные экспертами от 35 ранее оценивавшихся запросов (они оценивались на старой коллекции).
- Эксперты с юридическим образованием составили расширенные описания для 15 новых запросов.
- Каждое задание оценивало 2 разных асессора.

Тем не менее, коэффициент согласия асессоров составил всего 0.72 (против 0.9 для дорожек поиска по Веб коллекции), что подтверждает сложность этого задания для асессоров без профильного образования.

Также отметим, что в этом году было выявлено много релевантных документов среди оценивавшихся – более 40% признано слаборелевантными, а 20% сильнорелевантными. Для сравнения, в 2006 и 2005 годах, когда оценку производили эксперты с юридическим образованием доля релевантных документов была - 16.5% и 29.5% (в 2006 и 2005 года была собрана только одна оценка для каждого задания).

4. Дорожки по классификации

Дорожки классификации также использовали новые коллекции. Однако, для классификации Веб сайтов и Веб страниц в качестве обучающего множества и таксономии использовались теже таксономия и обучающее множество DMOZ, что и в предыдущие годы.

Для классификации нормативных документов было построена новая таксономия и обучающее множество на основе таксономии предоставленной Кодекс. Были отобраны все категории-листья, для которых в эталоне присутствует >50 документов и для обучения для каждой из них было выбрано по 50 случайных примеров. Поскольку в каталоге Кодекс один и тот же документ может относиться ко многим категориям, то было принято решение использовать каждый документ в обучающем множестве только один раз (и игнорировать повторные вхождения). Всего было отобрано 727 категорий.

Такой подход вызвал справедливую критику участников – обучающее множество не отражает реальной “мощности” категорий, а наивное исключение дубликатов приводит к тому, что оценка обучающего множества как прогона не дает идеальных оценок, что выглядит неестественно.

Для дорожек Веб классификации оценивалось 19 категорий, для которых было доступно не менее 5 обучающих примеров. При проведении оценки мы столкнулись с проблемой слишком больших котлов. Некоторые участники отнесли очень много документов к каждой из категории. Все оценить невозможно в силу трудоемкости оценки, но как выбрать, что оценивать?

Для решения этих проблем мы использовали следующий подход:

- Для оценки отбирались категории с относительно небольшими котлами по дорожки классификации Веб сайтов – до 400. При этом мы старались представить по нескольку (не менее трех) категорий из каждого затронутого “крупного” раздела таксономии.
- Для дорожки классификации Веб страниц отбиралось 200 страниц на категорию (из котла для этой категории).

Однако, отбор части содержимого котла может оказывать значительное влияние на оценку – одни прогоны могут оказать недооцененными по сравнению с другими.

Выбор пересечения разных прогонов вероятно помогает получить подмножество с максимальной плотностью правильных ответов, но для оценки полноты важно также проанализировать “особенные” ответы, которые были найдены только одной из систем.

Мы использовали следующий подход. Документы выбирались таким образом, что бы из каждого прогона в пул попала равная доля документов. Примерно половина оценивавшихся документов была взята из пересечения всех прогонов, остальное – уникальные для какого-то отдельного прогона документы.

Еще одной проблемой было то, что один из участников предоставил результаты, в которых документы были упорядочены по степени близости к категории. Для этого участника в котел попали документы из верхушки списка и неудивительно, что доля релевантных документов для этого прогона выше – 33% против 15% (+100 документов) для сильных требований к релевантности и: 50% против 32.5% (+250) для слабых.

5. Кластеризация новостного потока

Задачей участника дорожки кластеризации потока новостей является разбиение потока новостных сообщений на событийные сюжеты. **Событие** (event) – это 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). Сюжеты могут быть связаны в сюжетные линии ассоциативными связями (причинно-следственно-пространственно-временными).

Попытка провести оценку дорожки кластеризации новостного потока не увенчала успехом. Почему? Вроде бы была сформулирована методология оценки и подготовлен инструмент, но получить результат не удалось.

В 2006 году идея оценки состояла в полной разметке сообщений за первые два дня для каждой недели (что отражает взгляд “редактора” новостного потока). На практике это означало, что асессор должен был проанализировать и структурировать множество из примерно 3500 документов (для каждой недели!). Однако, в таком наборе данных, очевидно есть сотни событий и сюжетов, держать их в памяти и ориентироваться в уже созданных событиях – очень сложная задача для неподготовленного человека.

Поскольку все события невозможно удержать в голове, то асессоры должны были просматривать уже созданные кластеры на предмет наличия подходящего существующего кластера.

Как следствие, распределение размеров построенных кластеров оказалось сильно непохоже на ожидаемое, наблюдалось низкое согласие между разными асессорами, а выборочная проверка результатов показала, что внутренняя структура сюжетов зачастую была сильно недоработана (очевидные дубли не склеены).

В 2007 году мы решили попробовать другой подход, который отталкивается от результатов систем. В основе лежали следующие соображения:

- проверка результатов систем поможет оценить хотя бы точность;
- объединение результатов систем даст приближение полноты;
- хочется контролировать объем работы и сложность задачи.

Для каждого сообщения были построены “событийный” и “сюжетный” котлы, в которые включались все сообщения, которые

хоть в одном прогоне были отнесены к тому же событию/сюжету, что и данное. И были вычислены “ядра” котлов – множества сообщений, для которых котлы совпадают.

Задача асессора состояла из трех этапов:

- Этап 1
  - Дан событийный пул.
  - Разбить сообщения на группы, задав каждой группе имя события.
  - В одной группе должны быть одинаковые новости про одно событие (эхо события из реального мира).
- Этап 2
  - Даны списки названий групп (можно посмотреть состав) .
  - Объединить те, что представляют одно событие.
- Этап 3
  - Даны списки названий групп.
  - Сгруппировать их в сюжеты.

Практически была завершена только оценка на этапе один для нескольких сюжетных колов разного размера (от 20 до 1000 элементов). Предварительные результаты продемонстрировали, что такой подход позволяет обнаружить ряд проблем с группировкой событий. Например, ошибочное объединение:

- повторяющихся событий за разные даты (курсы валют);
- географически несвязанных событий (терракт в Чечне/Ираке);
- относящихся к разному объекту (курс USD/EUR);
- ложное объединение вокруг “обзорных” сообщений (главное за неделю);

Тем не менее, такой подход к оценке тоже оказался не идеален и для больших котлов сложность задачи приводила к получению незавершенных разбиений. Большое число событий значительно затрудняет решение о том, является ли данное сообщение дубликатом какого-то из еще просмотренных ранее. Создание отдельных кластеров на каждое одиночное сообщение также требует времени.

В будущем мы планируем модифицировать подход к оценке и перейти от парадигмы “обработка ленты входящих сообщений” к парадигме “выявление групп сообщений про одно событие из множества сообщений”.

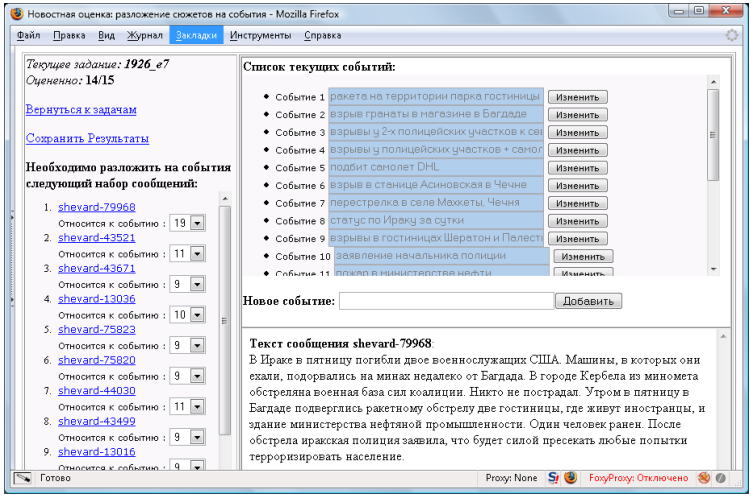

Рисунок 1. Новый инструмент оценки для дорожки кластеризации новостного потока.

Заключение

2007 был сложным годом для РОМИП. Значительно расширился арсенал РОМИП – появились новые коллекции и задания, но возросло и число проблем, связанных со срывом сроков.

Проведение РОМИП было бы невозможно без помощи многих людей и мы благодарны им за помощь. Надеемся, что сложности 2007 года – это временное явление и нам удастся разрешить эти проблемы в рамках РОМИП'2008.

Литература

[1] Труды РОМИП онлайн. http://romip.ru

[2] *C. Buckley , E. M. Voorhees*. Retrieval evaluation with incomplete information, Proc. of the SIGIR'2004, July 25-29, 2004.

ROMIP 2007: Report of organizers

Marina Nekrestyanova, Igor Nekrestyanov

This report describes details of ROMIP'2007 from organizers perspective. We focus on specifics of this year – new collections, changes in the methodology, problems and workarounds.