ROMIP'2004: отчет организаторов

И. Кураленок, И. Некрестьянов
http://romip.narod.ru

План доклада

- Что такое РОМИП?
- Обзор РОМИП'2004
  - Коллекции
  - Задачи
  - Методология оценки
- РОМИП'2004: первые наблюдения
- Открытые вопросы

Зачем это надо?

- Отсутствие публично доступных русскоязычных тестовых коллекций
- Российские группы редко участвуют в работе существующих зарубежных форумов по оценке методов информационного поиска
- Как следствие, трудности с проведением исследований по вопросам, интересующим российские команды

Коллекции РОМИП'2003 и РОМИП'2004 **доступны** для тех, кто не участвовал в цикле.

Проблемы организации

- Участники
  - Как преодолеть взаимное недоверие тех, кто конкурирует вне РОМИП?
- Методология
  - Как оценивать?
  - Как повысить достоверность результатов?
- Ресурсы
  - Откуда брать коллекции?
  - Где найти ресурсы необходимые для проведения оценки?
- Проблемы с законодательством
  - Как предотвратить нецелевое использование коллекций?
  - Как избежать превращения в форум для рекламы коммерческих систем?

Терминология

- **Дорожка**
  секция РОМИП, посвященная решению конкретной задачи поиска на конкретной коллекции
- **Асессор**
  человек, принимающий решение о соответствии конкретного ответа поставленной задаче
- **Таблица релевантности**
  таблица, содержащая (неполную) информацию об «идеальных» ответах
- **Тестовая коллекция**
  коллекция + задания + таблицы релевантности + инструмент вычисления оценок

Принципы

- **Равноправие систем**
  - Предпочтения асессоров не должны сказываться на результате
- **Анонимность источника результата**
  - Как при проведении оценки, так и при распространении результатов
- **Возможность повторного использования**
  - Результаты работы экспертов можно использовать для оценки и после завершения цикла РОМИП
- **Использование апробированных подходов**
  - Опыт TREC, CLEF и других мировых форумов
  - Повышает уверенность в получении надежных результатов

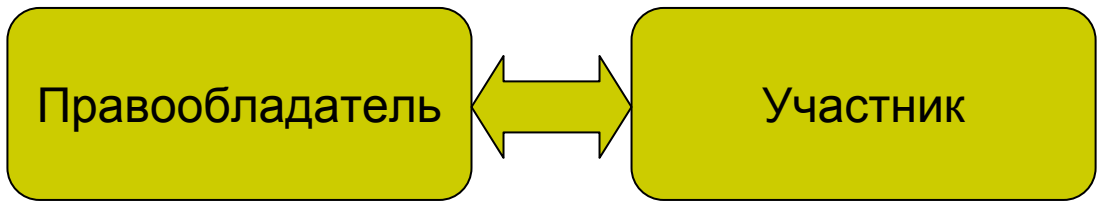
Годовой цикл

- **Подготовительный этап**
  - Фиксируется: список дорожек, методология, создания тестовых коллекций и оценки
  - Прием заявок на участие
- **Подготовка и распространение заданий**
  - Соглашение об участии ограничивает возможности использования набора данных и результатов РОМИП.
- **Выполнение заданий участниками**
  - Самостоятельно и на своем оборудовании
- **Независимая оценка полученных ответов на задания**
  - Методология зависит от задачи
- **Предоставление результатов оценки участникам**
  - Анонимные псевдонимы для ссылок на других участников
- **Заключительный очный семинар**

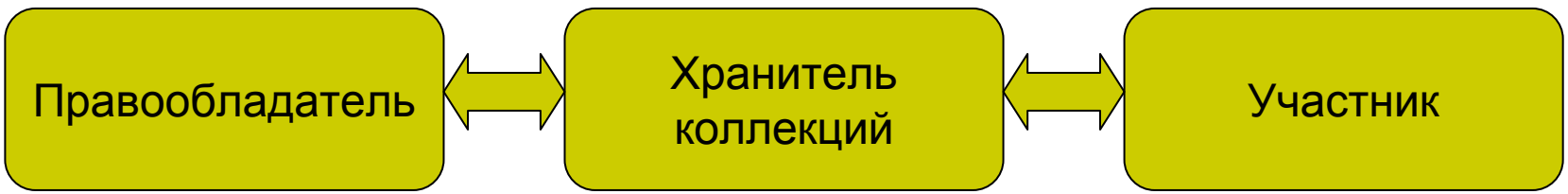
Оценка: метод «общего котла»

- «Общий котел» — это объединенное множество первых N_q ответов из выдачи каждой из систем для данного задания q.
- N_q – «*глубина* котла»
- Ответы из «котла» оцениваются асессорами
  - Один асессор оценивает все ответы из «котла»
  - Асессор НЕ знает чей это ответ
  - Асессор НЕ знает на какой позиции был этот ответ
  - Порядок выдачи документов случаен

РОМИП: правовая основа



Идея: отделение процесса накопления



- Накопление коллекции под будущие задачи
- Проще договариваться с правообладателями
- Аналоги: ELDA для CLEF, LDC для TREC

РОМИП: эволюция

2003

- 1 коллекция (Веб)
- 2 дорожки
  - Поиск
  - Классификация
- Участники
  - 9 заявок
  - 7 дошло до финиша
  - 14 прогонов
- 550 часов оценки

2004

- 2 коллекции
- 5 дорожек
- Участники
  - 11 заявок
  - 9 дошло до финиша
  - 34 прогона
- 1300 часов оценки
- Грант РФФИ
- Соглашение об использовании коллекций РОМИП

Коллекции

Коллекция DMOZ

300 000 страниц
(не более 500 страниц с одного сайта)
Область применения: обучающее множество

Задачи: поиск

Задание:

Для каждого запроса вернуть упорядоченный список (до 100) документов

Глубина котла - 50

Narod.ru

24250 запросов из журналов Яндекс и Рамблер.
Оценка: 48 + 19 повторно.
Альтернативная оценка с и без учета расширенных описаний.

Legal

13000 запросов из журналов Кодекс и Парк.Ру .

Классы запросов:
50 “понятия” + 41 “документы”

Задачи: классификация

Задание:

Задан список категорий и обучающая выборка.
Для каждого сайта/документа вернуть список до 5 категорий к которым он относится.

DMOZ/Narod.ru

Классификация ***Веб-сайтов***
247 категорий из каталога DMOZ

Оценка: 38 категорий

Legal

163 категории
13772 обучающих примера

Оценка: 12 вручную,
40 «сравнение с эталоном»

Задачи: поиск фактов

Задача:

найти все события связанные с персоной.
Ответ - фрагмент текста до 300 символов, описывающий это событие
(текст + ссылка на его положение в оригинале).

- Источник:

  «Кроссворд-кафе»
- 5052 задания
- Оценивалось: 109
- Оценка включала в себя проверку границ выделения

```
<task id="qa1109">
 <variant>Владимир Ильич Ленин</variant>
 <variant>Владимир Ильич Ульянов</variant>
 <description>
  вождь мирового пролетариата
 </description>
```

Оценка: детали

- **Многозначная шкала:**
  - Соответствующий (релевантный/витальный)
  - Скорее соответствующий (релевантный+)
  - Возможно соответствующий (релевантный-)
  - Не соответствующий (нерелевантный)
  - Документ не может быть оценен
- **Все оценки дублировались**
- **Использование расширенных описаний** (Веб-поиск, обе дорожки классификации)
  - Цель – упразднить неоднозначность трактовки

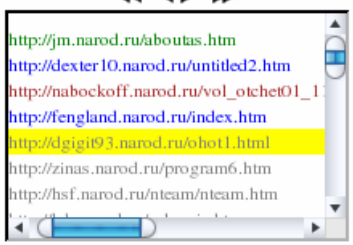
РОМИП

Российский семинар по оценке методов информационного поиска

футбольные турниры в Москве

Расширенное описание запроса

Требуется найти информацию о проводимых футбольных турнирах в Москве.



http://dgigit93.narod.ru/ohot1.html
in ROMIP-2003-narod.ru

- V Соответствует подробнее...
- R+ Скорее соответствует подробнее...
- R- Возможно соответствует подробнее...
- N Не соответствует подробнее...
- I Не может быть оцен подробнее...
- Не оценен подробнее...

К задачам
К пользователям

«А если - экипаж?» называлась моя статья, опубликованная в журнале "Коневодство и конный спорт" № 4 за 1989 год и выражавшая жгучее пожелание: развивать в стране соревнования упряжек, подниматься до права участия в популярных европейских и мировых турнирах по драйвингу.

Древнейший олимпийский вид спорта - соревнования запряжек, или на современном языке - драйвинг, - с 1970 года проводится по особым правилам, утвержденным ФЭИ (Международной федерацией конного спорта). Первенство мира проводятся ежегодно: по нечетным годам - для парных запряжек, по четным - для четвериков. А программа для тех и других одинаковая: многоборье. В него входит так называемый дрессаж - манежная езда, где оценивается не только правильность выполнения схемы езды, но и подобранность лошадей, экипаж, стилевая гармония упряжи, а также костюмов наездника и седоков. Тяжелейшая часть соревновании - марафон - приходится на второй день. Надо проехать 3-5 отрезков разной сложности в полевых условиях с обязательным преодолением вброд водных преград. Протяженность маршрута 24-32 км. Третий день посвящен паркуру - скоростному маневрированию между расставленными на ограниченной площадке разнообразными препятствиями. А они бывают настолько хитро расставлены, что подчас приходится всю упряжку разворачивать на 180 градусов. Это тебе не на двух ногах развернуться в седле, а пару или четверик да еще с экипажем о четыре колеса поворотить вспять, не "чиркнув" о предательски шаткие преграды!

Словом, шведы, венгры, немцы - уже давно доки в этом виртуозном виде спорта. А что у нас? После долгой подготовки и мучительных согласований во всех инстанциях в конце 1998 года в России была создана ассоциация "Драйвинг", президентом которой стал В.И.Рышков, президент Федерации конного спорта Москвы. Не прошло и года, как активу ассоциации удалось перевести и издать международные правила по драйвингу, побывать на национальных состязаниях по этому виду конного спорта в Польше и Чехии, провести несколько семинаров по судейству и практической езде среди своих пока не очень многочисленных членов.

Но самое главное это первые в России соревнования по драйвингу. На Тверской земле в окрестностях поселка Игуменка, на берегу Волги была построена стационарная трасса марафона, где в двухдневных состязаниях стартовали четыре, парных и несколько одноконных упряжек Одноконные упряжки сохраняли национальный колорит: дуга, хомут, кучер на козлах; парные же выглядели совсем по-европейски. Победителями стали Кирилл Егоров с Игорем Зайцевым, выступавшие на паре лошадей, принадлежащей каскадерской группе "Каро-Простор" из Москвы и Санкт-Петербурга (**Здесь автор допустила досадные ошибки не только в фамилии одного из победителей - Игорь Лаптев, - но и в названии команды: "КАРО" - это действительно группа каскадеров,**

Таблицы релевантности

Мнения асессоров не совпали –
что считать «правильным ответом»?

Использовавшийся подход:

- Сведение к бинарным оценкам асессора:
  релевантно - все что выше некоторого порога («релевантный-»)
- Слияние оценок:
  - Слабые требования (OR):
    релевантно – если кто-то так считает
  - Сильные требования (AND):
    релевантно – если все так считают

Степень согласия асессоров

Дорожка	and	or	доля
Веб поиск	637	2618	0.24
Веб поиск (без описаний)	504	2743	0.18
Поиск по Legal	1960	5587	0.35
Классификация Веб сайтов	910	1723	0.53
Классификация по Legal	488	1653	0.30
Поиск фактов	379	1095	0.35

Асессоры «не эксперты»

- Асессоры РОМИП не эксперты в области нормативных документов
- Сравнение с «идеальной» классификацией Кодекс:
  - В котлы попало лишь лишь 369 из 826 ☹
  - OR - 280 (75%)
  - AND - 153 (41%)
  - Степень взаимного согласия – 55% (>> 30%)
- Однако, выводы на основе идеальной классификации и оценки не всегда совпадают!

Расширенная шкала

Дорожка	витальный	релевантный+	релевантный-
Веб поиск	1405	1262	2095
Веб поиск (без описаний)	1164	1211	2047
Поиск по Legal	2913	2586	2353
Классификация Веб сайтов	778	667	769
Классификация по Legal	**1576**	574	530
Поиск фактов	**438**	694	559

Трудоемкость оценки

Дорожка	Прогноз	Факт
Веб поиск	30	13.97
Веб поиск (без описаний)	30	12.95
Поиск по Legal	40	15.5
Классификация Веб сайтов	60	21.11
Классификация по Legal	60	**9.9**
Поиск фактов	120	9.95

- Не учитываются косвенные затраты
- Есть оценки сделанные быстрее 2 секунд

Открытые вопросы

- Что изменить в организации для снижения
- Как улучшить эффективность асессоров?
- Как снизить порог для участия в РОМИП?
- Как правильно вычислять оценки для прогонов, которые не учитывались при построении котла?
- «Готовы ли мы» и «как правильно» привлекать зарубежных участников?
- Возможна ли оценка с возможностью повторного использования для интерактивной дорожки? Для дорожки аннотирования?

Как поучаствовать?

romip.narod.ru

Труды семинара доступны онлайн