РОМИП

Инициативный проект Российского семинара по оценке методов информационного поиска (РОМИП)

http://romip.narod.ru

romip@yahoogroups.com

РОМИП

Что такое РОМИП?

РОМИП=

(КОРПУС + ЗАДАЧИ + ОЦЕНКА) +

ОРГАНИЗАЦИОННЫЕ ПРОЦЕДУРЫ +

СЕМИНАР

РОМИП

Международные аналоги

- **CLEF (Cross-Language Evaluation Forum) – европейский форум по многоязычному поиску на европейских языках**
- **NTCIR – японский семинар с интернациональными участниками по многоязычному поиску**
- **SUMMAC – конференция по оценке качества автоматического аннотирования**
  MUC (Message Understanding Conference) – серия конференций, направленных в основном на определении в текстах объектов
- **TDT (Topic Detection and Tracking) – проект по обнаружению новых тем в потоке новостей и отслеживанию их развития**
- **DUC (Document Understanding Conference) – конференция по вопросам автоматического аннотирования**

РОМИП

Задачи РОМИП

- создание общедоступных корпусов (тексты + задания + оценки) с возможностью повторного использования;
- независимая оценка методов ИП;
- объединение профессионалов;
- формирование «правил игры».

РОМИП

Принципы семинара

- Равноправие систем
- Анонимность источника результата
- Использование апробированных подходов

РОМИП

Корпус narod_romip

- Источник – narod.ru
- Общий объем – 7 Гб +
- Документов – 600 000 +
- Число сайтов – 20 000+
- Лицензия основана на пользовательском соглашении Яндекса

РОМИП

Задачи (tracks)

Поиск по произвольному запросу (*ad hoc*)

- 10000 запросов из лога Яндекса
- Выдача – 100 документов

Тематическая классификация

- Классификация документов по 70 категориям категориям второго уровня каталога narod.ru
- Обучающая выборка – сайты каталога narod.ru (модерируемый самоввод), не менее 5 для каждой категории

РОМИП

Оценка

Метод «общего котла» (pooling)

- ~N_T первых документов из выдачи
- Оценка общего числа документов для проверки ~ $T^{0.7} \cdot N_T$
- T – количество участников
- Полнота рассчитывается по числу релевантных документов в пуле

РОМИП

Объективность оценки

- ~50 неизвестных участникам запросов из 10 000
- 5 неизвестных участникам категорий из 70
- расширенное описание запроса составляется экспертом
- оценщик не знает «происхождение» и ранг документа в выдаче
- троекратная оценка каждого документа

РОМИП

Участники 2003 года

- Russian Context
- Алхимик
- Кодекс
- Золушка
- Ключи к Тексту
- Галактика-Zoom
- Яндекс.Software 3.0