

Оптимизация параметров алгоритма поиска на основе анализа оценок экспертов*

© М.С. Агеев, Б.В. Добров

Научно-исследовательский вычислительный центр
МГУ им. М.В.Ломоносова,
АНО Центр информационных исследований
{ageev, dobroff}@mail.cir.ru

Аннотация

В статье описываются подходы, использованные коллективом разработчиков Университетской информационной системы РОССИЯ (УИС РОССИЯ, <http://www.cir.ru>), для выполнения заданий РОМИП 2005 по поиску в коллекции правовых документов и поиску в web коллекции. В этом году мы применили подход к оптимизации параметров поисковой системы, основанный на учете оценок релевантности документов, полученных в рамках предыдущего семинара РОМИП. Полученные результаты сравниваются с «классическими» методами.

1. Введение

Коллектив разработчиков информационно-поисковой машины и ряда лингвистических технологий, используемых в Университетской информационной системе РОССИЯ (далее – УИС РОССИЯ, <http://www.cir.ru> [4]) принимает участие в РОМИП с 2003 года.

* Авторы выражают признательность Российскому фонду фундаментальных исследований за поддержку данной работы (грант № 04-07-90280-в).

В 2003 году основной целью являлось получение «basic line» (отправной точки) для дорожки по поиску, если использовать классическую схему поиска TF*IDF [2].

В 2004 году мы провели эксперименты по добавлению к классическому алгоритму поиска учета следующих факторов [1]:

- учет расстояния между словами, встретившимися в запросе;
- повышение веса слов, встретившихся в заголовках документа.

Кроме того, были проведены эксперименты по автоматической классификации текстов.

Анализ полученных в рамках РОМИП 2003 и 2004 года оценок позволил выявить перспективные возможности для улучшения качества работы поисковой системы УИС РОССИЯ – за счет учета расстояния между словами запроса в документе и реализация поиска документов, частично удовлетворяющих запросу [5, 6].

В 2005 году мы исследовали алгоритмы поиска, основанные на учете следующих факторов:

- поиск на основе векторной модели TF*IDF;
- учет расстояния между словами, встретившимися в запросе;
- «мягкий» поиск, позволяющий искать документы, содержащие не все слова запроса.

При этом возникает задача выбора параметров алгоритмов, регулирующих соотношение указанных факторов поиска. Для выбора оптимальных параметров алгоритма мы воспользовались известными оценками релевантности документов, полученными в РОМИП'2004 (по запросам, оцененным в РОМИП'2004).

Тестирование экспериментальных алгоритмов производилось на дорожках поиска по нормативной коллекции и поиска по web-коллекции. Результаты показали преимущество оптимизированного алгоритма как на запросах 2004 года, так и на новых запросах 2005 года.

2. Описание алгоритмов, использованных для выполнения заданий

2.1. Поиск на основе векторной модели TF*IDF

Мы использовали классический алгоритм поиска по словам с ранжированием документов по формуле TF*IDF в формулировке INQUERY [2, 7] (также известной под наименованием BM25 Okapi).

Каждый запрос представлялся в виде последовательности словоформ, соединяемых условием «AND», для каждой словоформы вырабатывались леммы, соединяемые условием «OR».

2.2. Нечеткий поиск с учетом расстояния между словами

Данный алгоритм учитывал следующие факторы при ранжировании документов:

- вес слов по модели $TF*IDF$;
- учет расстояния между словами, встретившимися в запросе;
- количество слов запроса, найденных в документе.

Производился поиск документов, содержащих не менее, чем $T(Q)$ слов запроса. Порог $T(Q)$ вычислялся по следующей формуле:

$$T(Q) = \lceil |Q|^{0.75} \rceil \quad (1)$$

где $|Q|$ — количество слов в запросе Q .

Найденные документы упорядочивались по убыванию ранга, вычисляемого по следующей формуле:

$$\text{Rank}_D(Q) = \frac{FF-1}{|Q|} + \frac{1}{|Q|} \cdot \frac{V_D(Q) + \text{Near}_D(Q)}{2} \quad (2)$$

где $|Q|$ — количество слов запроса, $\text{Rank}_D(Q)$ — ранг соответствия документа запросу; FF — количество слов запроса, которые содержатся в данном документе, $V_D(Q)$ — ранг, вычисленный по формуле $TF*IDF$ [2].

Ранг по близости $\text{Near}_D(Q)$ вычисляется следующим образом: производится поиск минимального «куска» документа, в котором содержатся все слова запроса. Пусть длина этого куска документа в словах равна $\lambda_D(Q)$. Тогда

$$\text{Near}_D(Q) = \frac{1}{\ln(\lambda_D(Q) - |Q| + 4)} \quad (3)$$

Отметим, что данный алгоритм ранжирования упорядочивает документы, в первую очередь, в порядке убывания количества найденных слов запроса, затем — по убыванию ранга, учитывающего $TF*IDF$ и «кучность» слов запроса в документе.

2.3. Оптимизация параметров ранжирования

Формулы ранжирования 1-3 были выбраны нами из эвристических соображений и допускают включение параметров,

влияющих на вес различных факторов ранжирования. Мы предположили, что можно улучшить качество поиска, если подобрать оптимальные весовые коэффициенты для различных факторов ранжирования.

Для оценки качества поиска мы воспользовались матрицами релевантности, полученными для запросов РОМИП 2004 года. В качестве критерия оценки использовались:

- 1) метрика Average Precision [3];
- 2) метрика Average Precision по документам пула — то есть найденные документы, для которых нет оценки ассессора, удалялись из результатов запроса.

Оптимизация параметров производилась для следующих классов функций ранжирования:

$$1) \text{Rank}_D(Q) = \frac{\text{FF}-1}{|Q|} + \frac{1}{|Q|} \cdot \frac{V_D(Q) + \beta \cdot \text{Near}_D(Q)}{1 + \beta}$$

— однопараметрическое семейство функций с параметром $\beta > 0$, задающим относительный вес кучности слов запроса;

$$2) \text{Rank}_D(Q) = \frac{\text{FF}-1}{|Q|} + \frac{1}{|Q|} \cdot \frac{1 + \beta}{\frac{1}{V_D(Q)} + \frac{\beta}{\text{Near}_D(Q)}}$$

— однопараметрическое семейство функций с параметром $\beta > 0$, задающим относительный вес кучности слов запроса;

$$3) \text{Rank}_D(Q) = V_D(Q) \cdot \frac{\text{FF}}{|Q|} + \beta \cdot \text{Near}_D(Q) \cdot 2^{-\alpha \cdot (|Q| - \text{FF})}$$

— двухпараметрическое семейство функций с параметрами $\beta > 0$ (относительный вес кучности слов запроса) и $\alpha > 0$ (зависимость веса кучности от количества найденных слов);

$$4) \text{Rank}_D(Q) = \frac{V_D(Q) + \beta \cdot \frac{1}{\ln^\gamma(\lambda_D(Q) - |Q| + 4)} + \alpha \cdot \text{FF}}{1 + \beta + \alpha \cdot |Q|}$$

— трёхпараметрическое семейство функций с параметрами $\beta > 0$ (относительный вес кучности слов запроса), $0 < \alpha \leq 1$ (вес количества найденных слов запроса) и $\gamma > 0$ (жесткость функции ранжирования по кучности слов).

Отметим, что «не оптимизированный» алгоритм поиска (п. 2.2) соответствует частному случаю для каждого из четырех классов функций.

Перебор различных значений параметров для каждого из четырех классов функций позволил выбрать оптимальные параметры для дорожек поиска:

- для дорожки поиска по нормативной коллекции оптимальными оказались:
 - о класс функций 3 с параметрами $\alpha = 0.5$, $\beta = 0.3$;
 - о класс функций 1 с параметром $\beta = 0.3$ (с небольшим отставанием)
- для дорожки поиска по web-коллекции оптимальными оказались:
 - о класс функций 3 с параметрами $\alpha = 0.5$, $\beta = 0.1$;
 - о класс функций 1 с параметром $\beta = 0.1$ (с небольшим отставанием).

Для выполнения заданий РОМИП'2005 мы использовали алгоритмы из класса функций 1 (меньше параметров — меньше шансов «переобучения» алгоритма) с соответствующими оптимальными параметрами.

3. Описание результатов

Оценка результатов прогонов производилась на основе сравнения результатов систем с мнением экспертов РОМИП с использованием общепринятых метрик. Среди предоставленных организаторами матриц релевантности мы выделили матрицы с глубиной пула 50 (все присланные документы получили оценки экспертов), таблица релевантности «OR» (хотя бы один эксперт оценил документ как релевантный). Использовались наборы запросов:

- только 2004 года — для оценки результатов на использованных в процессе обучения запросах;
- только 2005 года — для оценки результатов на новых запросах, которые заранее не были известны.

На следующих графиках отображены результаты различных прогонов по коллекциям нормативных документов и web-коллекции. Наши прогоны обозначены:

- tfidf — поиск на основе векторной модели TF*IDF (п. 2.1);

- noopt — не оптимизированный алгоритм поиска с учетом расстояния между словами (п. 2.2). Данный прогон мы выполняли только по коллекции нормативных документов.
- opt — алгоритм поиска с оптимизированными весовыми коэффициентами (п. 2.3).

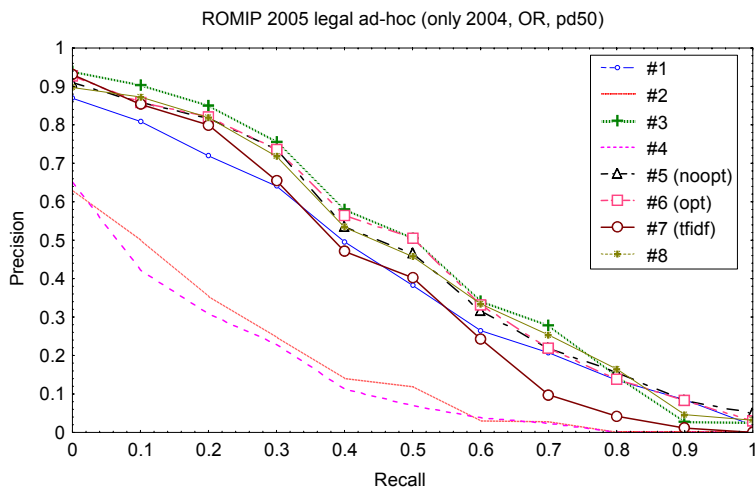


Рис. 1. 11-точечный график полноты/точности для различных прогонов дорожки legal adhoc (запросы 2004 года).

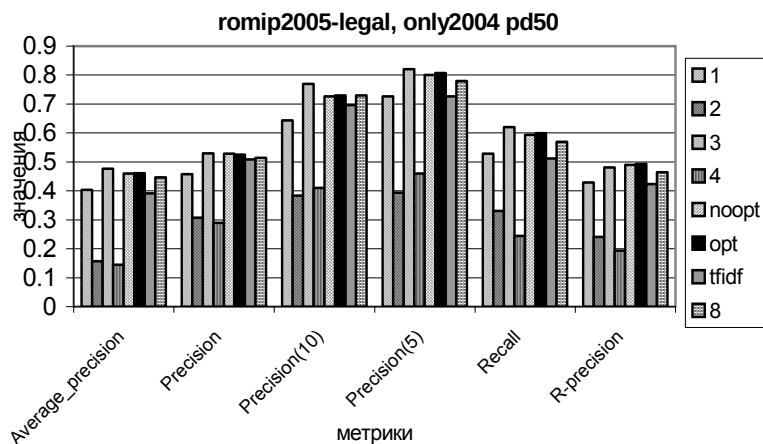


Рис. 2. Метрики качества поиска для различных прогонов дорожки legal-adhoc (запросы 2004 года).

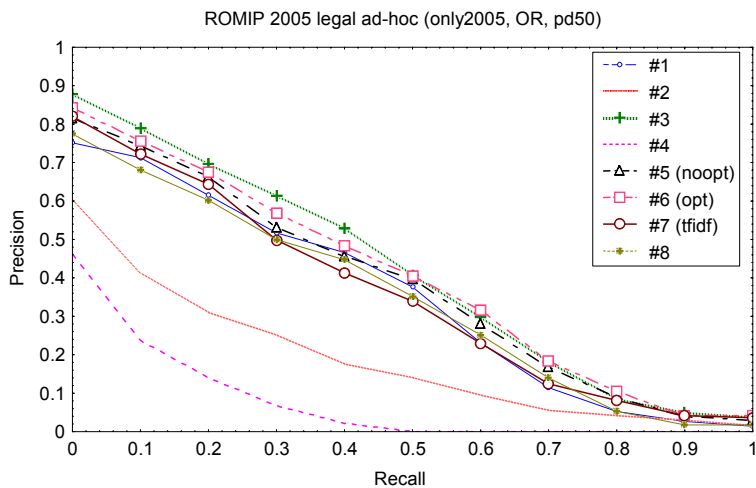


Рис. 3. 11-точечный график полноты/точности для различных прогонов дорожки legal adhoc (запросы 2005 года).

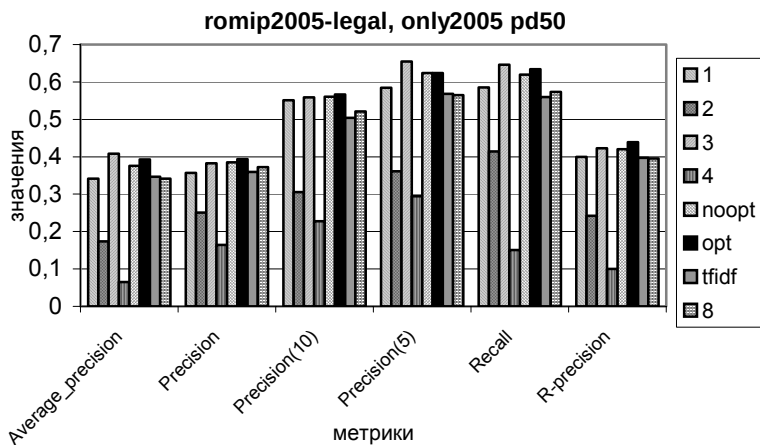


Рис. 4. Метрики качества поиска для различных прогонов дорожки legal-adhoc (запросы 2005 года).

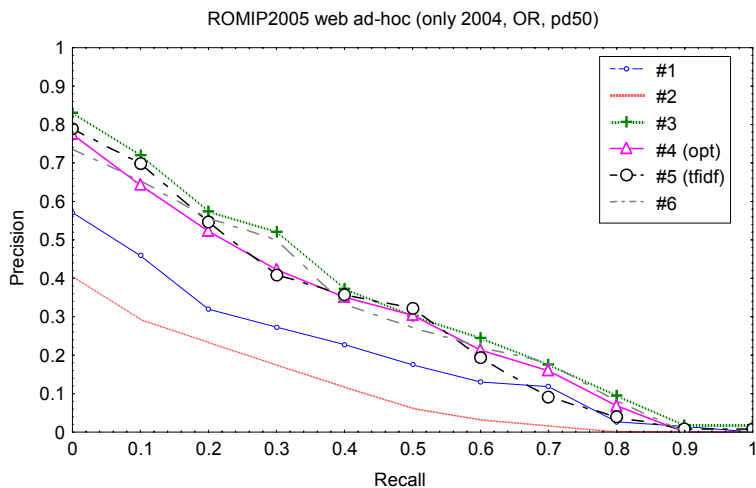


Рис. 5. 11-точечный график полноты/точности для различных прогонов дорожки web adhос (запросы 2004 года).

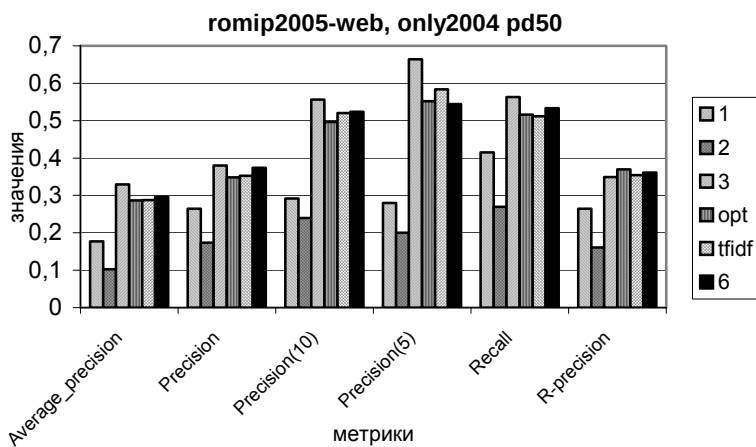


Рис. 6. Метрики качества поиска для различных прогонов дорожки web-adhoc (запросы 2004 года).

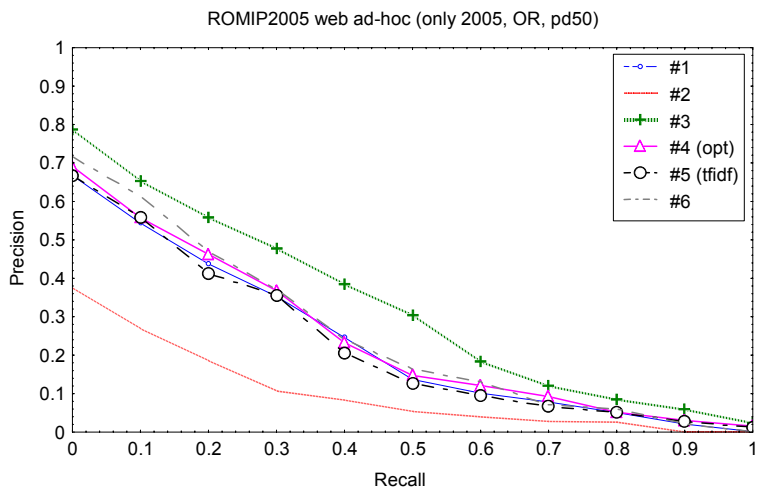


Рис. 7. 11-точечный график полноты/точности для различных прогонов дорожки web adhос (запросы 2005 года).

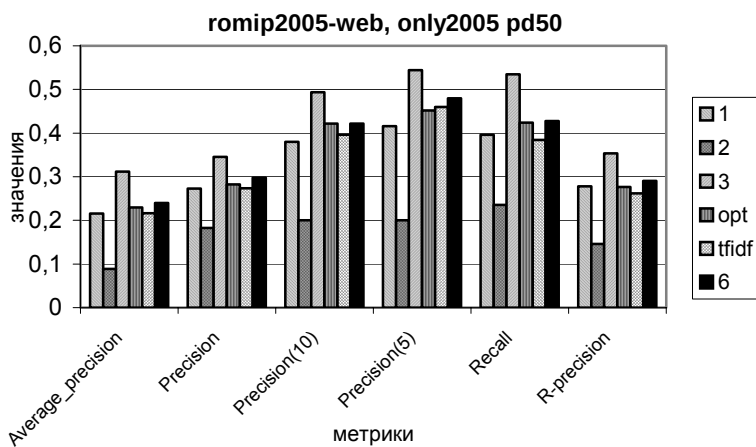


Рис. 8. Метрики качества поиска для различных прогонов дорожки web-adhос (запросы 2005 года).

Из приведённых графиков можно сделать следующие выводы:

- 1) Учет расстояния между словами и «мягкий» поиск повышают качество поиска;
- 2) Оптимизированный алгоритм поиска показывает более высокие результаты, чем не оптимизированный как на известных алгоритму запросах (2004 года), так и на новых запросах (2005 года);

Заключение

Мы описали метод улучшения характеристик поисковой системы с использованием накопленных материалов оценки — результатов прошедших семинаров РОМИП. Результаты тестирования в рамках РОМИП'2005 показали эффективность предложенного метода автоматической настройки параметров алгоритма поиска.

Предложенная методика может быть использована для повышения эффективности поисковых машин для конкретных коллекций документов.

В дальнейшем мы планируем развивать данный подход в направлении расширения набора параметров, влияющих на работу поискового алгоритма.

Литература

- [1] Агеев М.С., Добров Б.В., Лукашевич Н.В., Сидоров А.В. Экспериментальные алгоритмы поиска/классификации и сравнение с «basic line» // Российский семинар по Оценке Методов Информационного Поиска (РОМИП 2004) — Пущино, 2004. — стр. 62-89
- [2] Агеев М.С., Добров Б.В., Лукашевич Н.В., Сидоров А.В., Штернов С.В. "Отправная точка" для дорожки по поиску в РОМИП (предварительный анализ) // Труды РОМИП'2003 (Российский семинар по Оценке Методов Информационного Поиска) — НИИ Химии СПбГУ / Под ред. И.С.Некрестянова — Санкт-Петербург, 2003 — стр. 87-110.
- [3] Агеев М.С., Кураленок И.Е. Официальные метрики РОМИП'2004 // Российский семинар по Оценке Методов Информационного Поиска (РОМИП 2004) — Пущино, 2004.
- [4] Агеев М.С., Добров Б.В., Журавлев С.В., Лукашевич Н.В., Сидоров А.В., Юдина Т.Н. Технологические аспекты организации доступа к разнородным информационным ресурсам в университетской информационной системе РОССИЯ. // Электронные библиотеки, 2002 — Том.5 — Выпуск 2

- [5] Губин М.В. Участие ИПС "Кодекс" в семинаре РОМИП 2004 // Российский семинар по Оценке Методов Информационного Поиска (РОМИП 2004) — Пушино, 2004.
- [6] Сегалович И.Е. Яндекс на РОМИП-2004. Некоторые аспекты полнотекстового поиска и ранжирования в Яндекс // Российский семинар по Оценке Методов Информационного Поиска (РОМИП 2004) — Пушино, 2004.
- [7] Callan J.P., Croft W.B. and Harding S.M., The INQUERY Retrieval System // A.M. Tjoa and I. Ramos (eds.), Database and Expert System Applications. Proceedings of {DEXA}-92, 3rd International Conference on Database and Expert Systems Applications. - Springer Verlag, New York. - 1992. - pp.78-93.

Optimization of Parameters of Search Algorithm Based on ROMIP Relevance Assessments

Mikhail S. Ageev, Boris V. Dobrov

This article describes approaches used by team of UIS RUSSIA (University Information System of Russian inter-University Social Science Information and Analytical consortium, <http://www.cir.ru/eng/>) search engine for RIRES2005 (Russian Information Retrieval Evaluation Seminar) tracks. We participated in ad hoc track on collection of legal documents and web collection. We optimize parameters of our search engine by using relevance assesments obtained from previous RIRES seminar. Optimized algorithm showed advantage over non-optimized algorithm.