Классификация запросов и оптимизация факторов для поиска нормативных документов

© М.С. Агеев[1, 2], Б.В. Добров[1, 2], Н.В. Лукашевич[1, 2],
А.В. Сидоров[2], С.В. Штернов[1, 2]

[1] Научно-исследовательский вычислительный центр
МГУ имени М.В.Ломоносова
[2] АНО Центр информационных исследований
{ageev, dobroff, louk, alexeys}@mail.cir.ru,
sergey@shternov.ru

Аннотация

В статье описываются подходы, использованные коллективом разработчиков Университетской информационной системы РОССИЯ (УИС РОССИЯ http://www.cir.ru), для выполнения заданий РОМИП 2009 по поиску нормативно-правовых документов.

В цикле РОМИП 2009 мы основное внимание уделили дорожке по поиску в коллекции нормативно-правовых документов.

В этом году мы развили модель обработки длинных тематических запросов, примененную нами в 2008 году [1], с учетом выявленных недостатков. Основными нововведениями были следующие:

- классификация запросов (4 класса) и различная обработка запросов для разных классов;
- расширение набора используемых факторов ранжирования до 18 факторов;
- оптимизация весов факторов ранжирования на основе максимизации метрики AveragePrecision по запросам 2008 года.

В 2008 году мы разработали модель обработки запросов, ориентированную на обработку длинных (в том числе очень длинных) поисковых информационных запросов. Данная модель использует [1]:

- двухшаговую процедуру обработки запросов: сначала используется грубая, но быстрая модель поиска, затем первые 100 результатов переранжируются с использованием более точной модели;
- комбинированную векторную модель текста, построенную на двух индексах – индексе лемм и индексе концептов Общественно-политического тезауруса [4];
- учет расстояния между словами и концептами Тезауруса, расположенными в одном или соседних предложениях текста.

Концепты тезауруса дают возможность дополнительно учесть три дополнительных фактора:

- синонимию терминов,
- лексическую многозначность – производится предварительный выбор наиболее подходящего по контексту значения слов и выражений,
- близкое расположение в тексте компонентов многословных терминов и выражений.

Поэтому результаты работы двух видов векторных моделей могут достаточно серьезно различаться.

Анализ результатов 2008 года показал, что предложенная модель действительно показывает высокие результаты на длинных тематических запросах, но при этом проигрывает другим методам на коротких запросах и навигационных запросах, для которых требуется точный поиск по реквизитам документа.

В этом году мы использовали следующие факторы ранжирования:

1) класс запроса:
   - класс 1 «Похожие на заголовки без цифр» — запрос не включает в себя цифры, имеет длину не менее 2 слов, текст запроса похож (нечеткий поиск фразы) на заголовок хотя бы одного документа;
   - класс 2 «С цифрами» — текст запроса содержит цифры, в большинстве случаев это – реквизиты искомого документа;

- класс 3 «Короткие тематические запросы» — запросы, не относящиеся к классам 1-2 и состоящие из 1-2 слов (не считая предлогов);
- класс 4 «Длинные тематические запросы» — остальные запросы;

2) «gr» — вес документа по модели обработки длинных тематических запросов 2008 года [1];
3) BM25 — вес документа по модифицированной формуле BM25 [3];
4) BM25 с нормализацией — вес слов запроса и документа нормализуется по евклидовой норме (сумма квадратов весов слов документа равна 1);
5) MW — вес документа по формуле «минимального окна» [2];
6) P1, P3, P5 — вес документа по парам слов запроса, расположенных на расстоянии 1 (непосредственно рядом), 3 и 5 [2];
7) PH — фактор, равный 1, если запрос встретился в документе в виде точной фразы, и 0 – иначе;
8) факторы BM25, MW, P1, P3, вычисленные для заголовка документа;
9) фактор PH, вычисленный для заголовка документа;
10) факторы BM25, MW, P1, P3, PH, вычисленные для заголовка документа, объединенного со всеми подзаголовками документа.

Поиск оптимальной формулы ранжирования производился в классе линейных комбинаций факторов 2-10 отдельно для каждого класса запросов. Для подбора оптимальных весов использовался элементарный алгоритм оптимизации, основанный на методе покоординатного спуска:

1) оптимальный весовой коэффициент для первого фактора вычисляется перебором значений весов первого фактора;
2) аналогично производится подбор 2-го, 3-го, ... 18 весового фактора;
3) процедура повторяется до тех пор, пока есть возможность улучшения результата хотя бы по одной координате.

В качестве оптимизируемой метрики использовалась AveragePrecision на запросах дорожки поиска по нормативным документам 2008 года (95 запросов).

Нами были подготовлены следующие прогоны:

1) gr2008 — алгоритм ранжирования, примененный нами в 2008 году; использовалась формула

$$R = GR + 0.25 \cdot \mathrm{BM25}$$

где R – вес соответствия документа запросу;

2) newgr — использовалась смесь факторов GR и BM25, но вес фактора BM25 выбирался оптимальным для данного типа запросов:

$$R = GR + w(T) \cdot \mathrm{BM25}$$

где T – класс запроса;

3) 11f1c_opt — использовалась смесь факторов 2-6, 8 (11 факторов), без использования классификации запросов, веса факторов выбирались оптимизацией метрики AveragePrecision;
4) 11f4c_opt — использовалась смесь факторов 2-6,8 и классификация запросов, веса факторов выбирались оптимизацией метрики AveragePrecision отдельно для каждого класса;
5) 11f3c_gr4c — аналогично предыдущему прогону, но для 4-го класса запросов (длинные тематические) использовались параметры прогона gr2008 — в предположении, что для длинных тематических запросов оптимальна модель 2008 года;
6) 18f4c_opt — использовалась смесь факторов 2-10 (18 факторов) и классификация запросов, веса факторов выбирались оптимизацией метрики AveragePrecision отдельно для каждого класса.

При использовании нескольких индексов (индекс по полному тексту, индекс по заголовкам и индекс по оглавлению документов) в пул документов для ранжирования по одному запросу попадали все документы, найденные среди первых 100 результатов по одному индексу. При этом значения факторов ранжирования для остальных индексов устанавливались равными нулю. Например, если документ найден среди первых 100 документов по индексу заголовков и оглавления, но не найден в первых 100 по индексу полного текста, то значение фактора «BM25 по полному тексту» устанавливалось равным нулю.

Еще 3 прогона являются вариацией прогонов 1, 4, 5, но с другим способом учета попадания документа в разные индексы: для каждого документа, попавшего в первые 100 документов хотя бы по

одному из индексов, вычислялись значения всех факторов ранжирования.

Анализ результатов

В дорожке поиска по нормативно-правовой коллекции модель 11f4c_opt показала лучший результат из 18 представленных алгоритмов по метрике Precision(10) и один из лучших результатов – по остальным метрикам (см. рис.1).

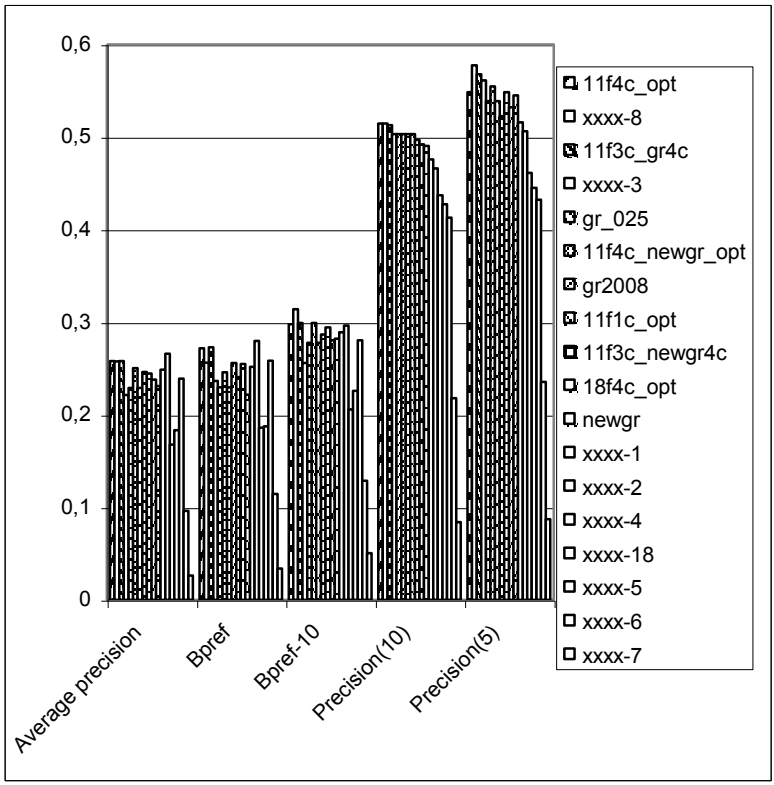

Рис.1. Результаты дорожки 2009 Legal adhoc, pd35, прогоны отсортированы по убыванию метрики Precision(10).

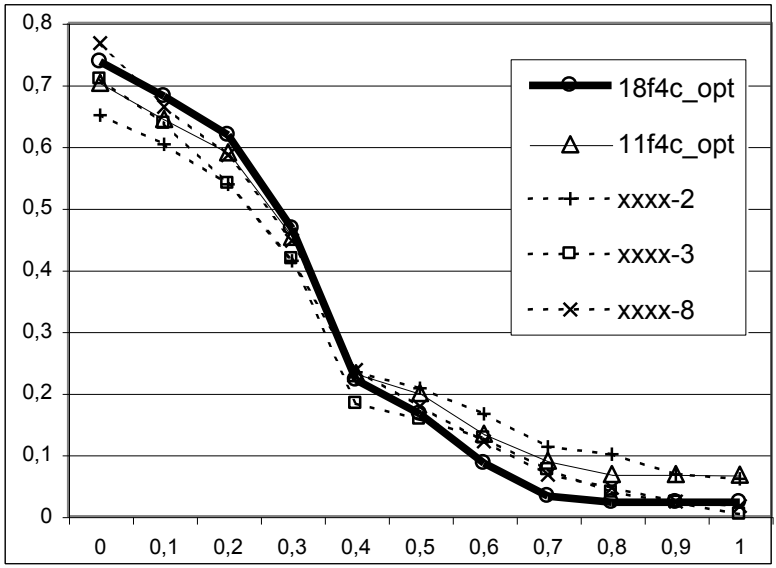

Рис.2. 11-точечный график полноты-точности для
5 лучших прогонов ROMIP legal 2009

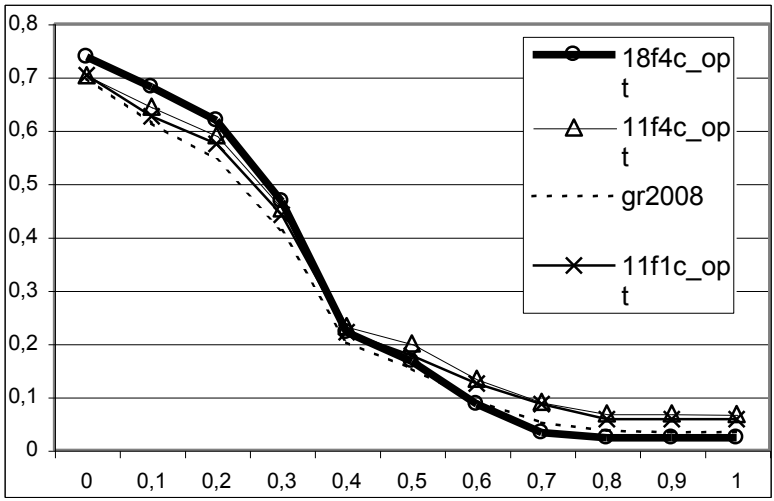

Рис.3. 11-точечный график полноты-точности для различных
прогонов (2 лучших УИС РОССИЯ,
прогон 2008, а также без классификации запросов)

На рис. 2 и 3 представлены 11-точечные графики полноты-точности для двух лучших наших прогонов

а) по сравнению с другими лучшими прогонами, представленными в данной дорожке;
б) по сравнению с нашим алгоритмом 2008 года «без классификации, 2 фактора ранжирования» и алгоритмом «11 факторов, без классификации запросов».

Мы проанализировали сравнительную эффективность различных прогонов на коллекции обучения (95 запросов 2008 года) и коллекции тестирования (68 запросов этого года). Результаты показаны в таблице 1.

	Данные 2008 г.		Данные 2009 г.	
	Average Precision	Улучшение по сравнению с gr2008	Average Precision	Улучшение по сравнению с gr2008
gr2008	0.2923	1.0000	0.2290	1.0000
newgr	0.3144	1.0756	0.2309	1.0080
11f1c_opt	0.3253	1.1130	0.2462	1.0750
11f3c_gr4c	0.3368	1.1523	0.2578	1.1257
11f4c_opt	0.3408	1.1659	**0.2580**	**1.1265**
11f4c_newgr_opt	**0.3433**	**1.1745**	0.2504	1.0931
18f4c_opt	0.2886	0.9873	0.2378	1.0384

Таблица 1. Сравнение эффективности различных алгоритмов на коллекции обучения (2008) и коллекции тестирования (2009).

Видно, что результаты на коллекции обучения не всегда коррелируют с результатами на коллекции тестирования, но в целом можно сказать, что добавление новых факторов ранжирования улучшает качество поиска.

Для того чтобы объяснить различие в поведении алгоритмов на коллекциях обучения и тестирования, мы провели анализ эффективности алгоритмов на отдельных запросах.

Анализ показал, что результаты можно улучшить, если подключить алгоритмы исправления опечаток. Кроме того, мы выявили определенные проблемы в интерпретации понятия «релевантность» при проведении оценки результатов дорожки поиска.

Критический анализ результатов оценки

Одним из критериев согласованности оценки асессоров, также известным как гипотеза кластерности (см. C. J. Van Rijsbergen [5]), является предположение, что похожие документы должны получать одинаковые оценки релевантности, за исключением случаев, когда различие документов действительно влияет на соответствие документов запросу.

Мы подготовили инструмент проверки согласованности мнений асессоров для похожих документов по одному и тому же запросу.

Анализ результатов оценки показал, что имеется множество ярких примеров несогласованности. В частности, при пороге 95% близости (это практически идентичные документы) мы выделили 70 кластеров с разно оцененными документами, содержащих 304 документа, при пороге 90% - 93 кластера и 383 документа, при пороге 80% - 108 кластеров и 535 документов.

По нашему мнению, наличие информации о кластеризации близких документов позволяет очень быстро – в течение одного-двух часов единообразно оценить практически одинаковые документы.

Нас огорчает, что хотя мы предоставили данные о близости документов в коллекции организаторам, но они не были учтены.

В результате:

1) итоговые метрики могут измениться на величины около 10%, что может существенно влиять на оценку перспективности тех или иных методов;
2) но более важно другое – возникает существенный вопрос о возможности поиска оптимальных сочетаний параметров выбранных методов. Даже применение различных техник очистки данных будет бессмысленно при такой «случайной» оценке результатов.

Мы приведем два достаточно простых типичных примера.

Пример 1. Запрос «статья 353».

Для данного запроса есть расширенное описание: «Кодексы содержащие статью 353».

Рассмотрим один из больших кластеров — это различные версии уголовно-процессуального кодекса РСФСР (естественно, очень похожие друг на друга), каждый такой документ содержит статью 353 и ссылки на нее.

Более детальное рассмотрение показывает, что данная статья изменялась 7 августа 2000 года. Однако асессоры, очевидно, не

сошлись во мнении о релевантности этих документов запросу (см. таблицу 2) ни до указанной даты, ни после.

#	Релевантность	Заголовок
1	релевантен	Уголовно-процессуальный кодекс РСФСР (с изменениями на 18 февраля 1993 года)
2	не релевантен	Уголовно-процессуальный кодекс РСФСР (с изменениями на 29 апреля 1993 года)
3	не релевантен	Уголовно-процессуальный кодекс РСФСР (с изменениями на 1 июля 1993 года)
4	релевантен	Уголовно-процессуальный кодекс РСФСР (с изменениями и дополнениями на 27 августа 1993 года)
5	не релевантен	Уголовно-процессуальный кодекс РСФСР (с изменениями и дополнениями на 1 июля 1994 года)
6	не релевантен	Уголовно-процессуальный кодекс РСФСР (с изменениями и дополнениями на 28 апреля 1995 года)
7	не релевантен	Уголовно-процессуальный кодекс РСФСР (с изменениями и дополнениями на 18 мая 1995 года)
8	не релевантен	Уголовно-процессуальный кодекс РСФСР (с изменениями и дополнениями на 19 июля 1995 года)
9	релевантен	Уголовно-процессуальный кодекс РСФСР (с изменениями и дополнениями на 2 февраля 1996 года)
10	релевантен	Уголовно-процессуальный кодекс РСФСР (с изменениями и дополнениями на 28 ноября 1996 года)
11	релевантен	Уголовно-процессуальный кодекс РСФСР (с изменениями и дополнениями на 21 декабря 1996 года)
12	не релевантен	Уголовно-процессуальный кодекс РСФСР (с изменениями и дополнениями на 17 марта 1997 года)
13	релевантен	Уголовно-процессуальный кодекс РСФСР (с изменениями на 10 апреля 2000 года)
14	не релевантен	Уголовно-процессуальный кодекс РСФСР (с изменениями на 27 июня 2000 года)
15	релевантен	Уголовно-процессуальный кодекс РСФСР (с изменениями на 7 августа 2000 года)
16	не релевантен	Уголовно-процессуальный кодекс РСФСР (с изменениями на 9 марта 2001 года)
17	релевантен	Уголовно-процессуальный кодекс РСФСР (с изменениями на 20 марта 2001 года)
18	релевантен	Уголовно-процессуальный кодекс РСФСР (с изменениями на 9 апреля 2002 года)

Таблица 2. Различные оценки для документов по запросу «статья 353».

Пример 2. Запрос «Закон о прокуратуре».

Для данного запроса есть расширенное описание: «федеральный закон о прокуратуре».

В таблице 3 приведен пример 4 заголовков документов, имеющих различные оценки асессоров. Каждый из этих документов содержит ссылку на закон о прокуратуре и примерно одинаковое содержание Постановления.

#	Релевантность	Заголовок
1	не релевантен	О порядке исчисления выслуги лет, назначения и выплаты пенсий работникам органов и учреждений прокуратуры Российской Федерации и их семьям (с изменениями на 24 марта 2000 года)
2	не релевантен	О порядке исчисления выслуги лет, назначения и выплаты пенсий работникам органов и учреждений прокуратуры Российской Федерации и их семьям (с изменениями на 6 февраля 2004 года)
3	релевантен	О порядке исчисления выслуги лет, назначения и выплаты пенсий работникам органов и учреждений прокуратуры Российской Федерации и их семьям (с изменениями на 29 октября 2005 года)
4	не релевантен	О порядке исчисления выслуги лет, назначения и выплаты пенсий работникам органов и учреждений прокуратуры Российской Федерации и их семьям (с изменениями на 20 октября 2006 года)

Таблица 3. Различные оценки для документов по запросу «Закон о прокуратуре».

Предполагаем, что следует увеличить мощность программных средств контроля действий асессоров, например, особого контроля за обоснованием различных оценок для похожих документов.

Заключение

Наш коллектив использует возможность участия в РОМИП как эффективный способ уточнения параметров применяемых нами методов.

В этом году мы исследовали следующие методы улучшения качества поиска:

- классификация запросов и различная обработка запросов разных типов;

- введение в рассмотрение широкого разнообразия факторов ранжирования (18 факторов) и оптимизация весов факторов методом машинного обучения.

Мы оптимизировали параметры алгоритмов поиска при помощи оптимизации метрики AveragePrecision на коллекции обучения и получили улучшение данной метрики по сравнению с базовым алгоритмом:

- на коллекции обучения – на 17%
- на коллекции тестирования – на 13%.

Вместе с тем для нас явилось неожиданным то, что метрики качества различных методов на коллекции обучения слабо коррелируют с метриками на коллекции тестирования. В частности, лучший результат показал не тот прогон, который мы ожидали.

Для анализа результатов оценки нами была разработана система поиска несогласованных оценок асессоров, основанная на предположении о том, что похожие документы обычно должны иметь одинаковые оценки релевантности.

Проведенный нами анализ показывает, что имеется класс расхождений в оценках очень близких документов. Данные расхождения могут быть легко (в течение нескольких часов) учтены (возможно появление оценок типа OR и AND).

К сожалению, точность текущей оценки (ориентировочно около 10%) затрудняет адекватную интерпретацию полученных результатов.

Мы настойчиво рекомендуем, чтобы в дальнейшем аналогичные системы контроля асессоров были внедрены в процесс оценки результатов РОМИП, что позволит повысить качество оценки и увеличит осмысленность получаемых результатов.

Еще одним важным моментом считаем определение «ответственного» за проведение дорожки, например, из участников. Который бы согласовывал выбор оцениваемых запросов (конечно после сдачи результатов участниками), а также контекст их оценки асессорами заранее.

Также, в традиционных для РОМИП дорожках можно дополнительно устраивать «тематические» поддорожки, например, проведение более детальной оценки по заранее определяемым классам запросам, например, коротким или, наоборот, длинным.

Литература

[1] Агеев М.С., Добров Б.В., Лукашевич Н.В., Штернов С.В. УИС РОССИЯ в РОМИП 2008: поиск и классификация нормативных документов // Российский семинар по Оценке Методов Информационного Поиска. Труды РОМИП 2007-2008: Семинар в рамках Всероссийской науч. конф. RCDL'2008. 9 окт. 2008 г., Дубна - изд-во СПбг: НУ ЦСИ, 2008, 258 с. - С.44-58.

[2] Агеев М.С., Добров Б.В., Красильников П.В., Лукашевич Н.В., Павлов А.М., Сидоров А.В., Штернов С.В. УИС РОССИЯ в РОМИП'2007: поиск и классификация // Российский семинар по Оценке Методов Информационного Поиска. Труды РОМИП 2007-2008: Семинар в рамках Всероссийской науч. конф. RCDL'2007. 18 окт. 2007 г., Переславль-Залесский - изд-во Санкт-Петербург: НУ ЦСИ, 2008, 258 с. - С.199-220.

[3] 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. Труды второго российского семинара РОМИП'2004 (Пущино, 01.10.2004) – СПб: НИИ Химии СПбГУ. – 2004. – С.62-89.

[4] Лукашевич Н.В., Добров Б.В., Тезаурус русского языка для автоматической обработки больших текстовых коллекций // Компьютерная лингвистика и интеллектуальные технологии: Труды Международного семинара Диалог'2002 / Под ред. А.С.Нариньяни – М.: Наука – 2002. – Т.2 - С.338-346.

[5] C. J. Van Rijsbergen, Information Retrieval, Butterworth-Heinemann, Newton, MA, 1979

UIS RUSSIA at ROMIP 2009:
Ad Hoc Search of Legal Documents

Mikhail S. Ageev, Boris V. Dobrov, Natalia V. Loukachevitch,
Alexey V. Sidorov, Sergey V. Shternov

In the paper we describe methods used by the team of UIS RUSSIA (http://www.cir.ru/eng/) search engine for ROMIP 2009 tracks. We participated in the ad hoc track on a legal documents collection. We used complex retrieval model consisting of 1) sophisticated Thesaurus-based model for processing of long information queries 2) query classification 3) 18 query-text features 4) learning to rank optimization methods.