УИС РОССИЯ в РОМИП 2008: поиск и классификация нормативных документов

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

[1] Научно-исследовательский вычислительный центр МГУ им. М.В.Ломоносова
[2] АНО Центр информационных исследований
ageev@mail.cir.ru, dobroff@mail.cir.ru, louk@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 2008 по поиску и классификации нормативно-правовых документов.

1. Введение

В цикле РОМИП 2008 мы принимали участие в дорожках по поиску в коллекции нормативно-правовых документов и классификации нормативно-правовых документов.

Для дорожки ad hoc поиска документов по запросу описание проведенных экспериментов и выводы описаны в разделе 2, для дорожки классификации – в разделе 3.

2. Дорожка ad hoc поиска по коллекции нормативно-правовых документов

В этом году мы тестировали новую модель обработки запросов, ориентированную на обработку длинных (в том числе очень длинных) поисковых информационных запросов. Основной

направленностью разработки модели была обработка длинных информационных запросов, то есть запросов, которые имеют длину более 3 слов, и выражают некоторую информационную потребность. Информационные запросы условно противопоставляются навигационным запросам, суть последних в нормативно-правовой коллекции заключается в получении документа путем задания его формальных реквизитов: типа документа, номера документа, даты выхода, заголовка.

Для поиска документов по запросам в нормативно-правовой коллекции использовалась двухшаговая процедура.

На первом этапе исполнялась комбинированная векторная модель, построенная на двух индексах – индексе лемм и индексе концептов Общественно-политического тезауруса [2].

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.

Результаты работы векторных моделей замешиваются с помощью параметра α_1, то есть каждый документ получает вес по следующей формуле:

$$W_d = \alpha_1 W_{word} + (1- \alpha_1) W_{conc},$$

где W_{word} – вес документа по пословной векторной модели (модификация формулы BM25 [1]),

W_{conc} – вес документа по векторной модели, выполненной на основе концептов тезауруса [3].

Из документов, найденных по смешанной векторной модели, отбирается 100 документов – эти документы будут дополнительно анализироваться и дополнительно переупорядочиваться на **втором этапе** обработки запроса.

На втором этапе обработки запроса найденные документы переупорядочиваются по следующему принципу:

Максимальное число элементов запроса (слов и терминов) должно быть найдено не разбросанными по всему тексту, а сосредоточены в двух парах соседних предложений.

Коэффициент α_2 оценивает относительную весовую значимость лемм и концептов тезауруса в предложениях.

Получение нового веса документа можно представить как двухпроходный процесс. Сначала подсчитываются веса отдельных предложений, которые получаются суммированием весов лемм и концептов из запроса, найденных в предложении:

$$W_s = \alpha_2 \sum w_{wordi} + (1- \alpha_2) \sum w_{concj}$$

где w_{wordi}, w_{concj} – веса слов и концептов предложения.

На втором проходе вычисляется «усиленный» вес каждого предложения: если не все элементы запроса найдены в текущем предложении, то проверяется, нет ли недостающих элементов в соседнем предложении или в еще одной паре предложений документа. Веса дополнительных элементов найденных в других предложениях домножаются на параметрические коэффициенты α_4 (для присоединения элементов из соседнего предложения) и α_5 (для присоединения элементов из другой пары рядом лежащих предложения).

Таким образом, формула «усиленного» веса предложения имеет следующий вид:

$$W_{s1+} = W_1 + \alpha_4 W_{2-} + \alpha_5 [W_{3-} + \alpha_4 W_{4-}],$$

где W_1 - вес «главного» предложения, W_{2-} – вес следующего предложения, W_{3-} , W_{4-} - веса еще одной пары смежных предложений. Причем для каждого следующего предложения учитываются только те слова и концепты, ассоциируемые с запросом, которые еще не были учтены для предыдущих предложений.

Параметры модели оптимизировались на материалах дорожки нормативно-правового поиска romip-legal-2005. Оптимизировалось максимальное число релевантных документов в первых пяти документах выдачи, то есть показатель Precision(5).

2.1. Анализ результатов

В дорожке поиска по нормативно-правовой коллекции представленная модель показала лучший результат из 6 представленных алгоритмов, получив на первых 35 документах, которые были полностью оценены асессорами, показатель средней точности - 0.296 (см. рис.1), который превышает показатель следующего участника – 0.276 на 7%.

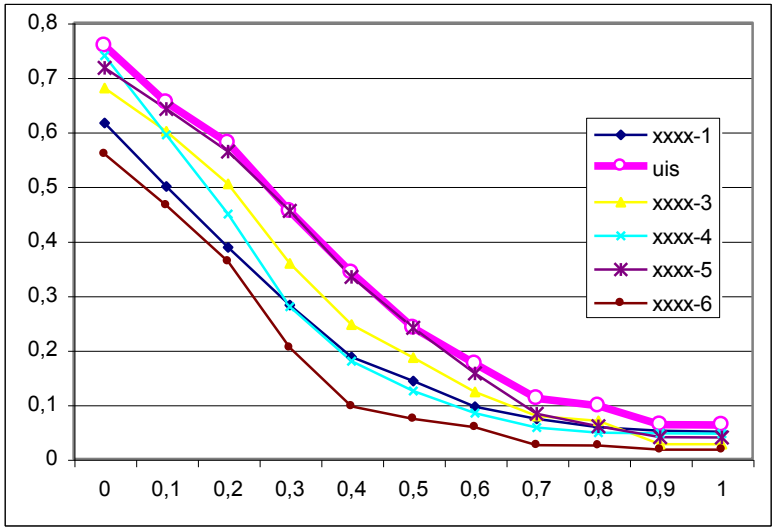

Рис.1. Результаты дорожки 2008 Legal adhoc, pd35.

Чтобы проанализировать, насколько хорошо модель отработала на целевом множестве длинных информационных запросов, мы разбили запросы на 4 группы:

1) запросы, явно указывающие выходные реквизиты документа: номер, дата, фрагмент названия или полное название, например, «*N 765 от 25.07.2006*»;
2) запросы с формулировками названий документов, например, «*Закон О валютном регулировании и валютном контроле*». По сути, это специальный случай поиска документов по выходным реквизитам;
3) короткие запросы, например, «*ипотека*». Особенностью обработки коротких запросов в нормативно-правовой области, на наш взгляд, является то, что при наличии большого количества документов, в которых встречаются слова запроса, важными критериями упорядочения документов являются не только частотные характеристики употребления слова в документе и коллекции, но и статус самого документа, а также, возможно, и вхождение слов запроса в название документа;

4) длинные информационные запросы, например, *уплата налога на прибыль организацией при отсутствии затрат*.

Пользуясь этой классификацией мы разделили все оцененные запросы этой дорожки на соответствующие группы и оценили среднюю точность участников по этим группам (оценка производилась на первых 35 документах, которые были полностью оценены асессорами) – см. Таблицу 1.

Таблица 1.

группа запр.	число запр.	1	uis	3	4	5	6
1	21	**0.33**	0.30	0.18	0.25	0.12	0.17
2	15	0.11	0.25	0.27	0.16	**0.35**	0.15
3	32	0.14	0.23	0.22	0.17	**0.28**	0.12
4	27	0.19	**0.36**	0.28	0.24	0.32	0.15
Всего	95	0.20	**0.296**	0.243	0.212	0.276	0.15

Действительно, можно видеть, что система показала наилучший результат на 4 группе запросов – длинных информационных запросах. Этот результат значительно превышает средний результат системы.

Проведенный анализ качества работы системы на разных группах запросов показывает, что важно уметь автоматически классифицировать поступающие запросы, и, в зависимости от класса запроса, применять несколько разные алгоритмы поиска.

2.2. Критический анализ результатов оценки

К сожалению, детальный по-документный анализ результатов оценки запросов вызывает у нас в настоящий момент много вопросов.

Мы проанализировали результаты выполнения нами запроса «*Компенсация за использование для служебных поездок личного легкового автомобиля*» (см. Таблицу 2).

Таблица 2. Результаты оценки релевантности некоторых документов по запросу «*Компенсация за использование для служебных поездок личного легкового автомобиля*»

№1. Нерелевантен

МИНИСТЕРСТВО ФИНАНСОВ РОССИЙСКОЙ ФЕДЕРАЦИИ
ПИСЬМО
от 21 июля 1992 года N 57
Об условиях выплаты компенсации работникам за использование ими личных легковых автомобилей для служебных поездок

Релевантный фрагмент:

Марка автомобиля	Норма компенсации в месяц (рублей)
ЗАЗ	427
ВАЗ (кроме ВАЗ-2121)	595
АЗЛК, ИЖ	638
ГАЗ, УАЗ, ВАЗ -2121	730

№2. Нерелевантен

МИНИСТЕРСТВО ФИНАНСОВ РОССИЙСКОЙ ФЕДЕРАЦИИ
ИНСТРУКТИВНОЕ ПИСЬМО
от 16 июня 1993 года N 74
"О предельных нормах компенсации за использование личных легковых автомобилей и мотоциклов для служебных поездок" (с изменениями от 26 августа 1994 года)

Релевантный фрагмент:

Марка автомобиля	Норма компенсации в месяц (рублей)
ЗАЗ	9900
ВАЗ (кроме ВАЗ-2121)	12700
АЗЛК, ИЖ	14000
ГАЗ, УАЗ, ВАЗ-2121	16300

№3. Нерелевантен

СОВЕТ МИНИСТРОВ – ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ПОСТАНОВЛЕНИЕ
от 24 мая 1993 года N 487
О предельных нормах компенсации за использование личных легковых автомобилей и мотоциклов для служебных поездок

Релевантный фрагмент:

марка автомобиля	норма компенсации в месяц (рублей)
ЗАЗ	3392
ВАЗ (кроме ВАЗ-2121)	4725
АЗЛК, ИЖ	5067
ГАЗ, УАЗ, ВАЗ-2121	5799

№4. Релевантен (но предыдущая версия этого документа нерелевантна)

МИНИСТЕРСТВО ФИНАНСОВ РОССИЙСКОЙ ФЕДЕРАЦИИ
ИНСТРУКТИВНОЕ ПИСЬМО
от 16 июня 1993 года N 74
"О предельных нормах компенсации за использование личных легковых автомобилей и мотоциклов для служебных поездок"
(с изменениями на 4 февраля 2000 года)

Релевантный фрагмент:

Марка автомобиля	Предельные нормы компенсации в месяц (рублей)
ЗАЗ	116
ВАЗ (кроме ВАЗ-2121)	148
АЗЛК, ИЖ	173
ГАЗ, УАЗ, ВАЗ-2121	221
Мотоциклы (для работников органов местного самоуправления сельской местности)	91

Как нетрудно видеть, очень похожие документы, которые по нашему мнению должны были быть отмечены как релевантные, получили у асессоров различные оценки.

Предполагаем, что для оценки было привлечено недостаточное количество асессоров. Кроме того, следует увеличить мощность программных средств контроля действий асессоров, например, особого контроля за обоснованием различных оценок для похожих документов.

3. Дорожка классификации нормативно-правовых документов

Задание состояло в построении процедуры автоматической классификации текстов для коллекции нормативных документов законодательства Российской Федерации из БД СПС «Кодекс».

В этом году оргкомитет РОМИП предоставил новую коллекцию документов, состоящую из 348410 документов Законодательства Российской Федерации, Москвы и Санкт-Петербурга по состоянию на декабрь 2006 года.

Множество рубрик, по которым требовалось выполнить классификацию, состояло из 726 рубрик 2-4 уровня, являющихся подмножеством большого иерархического рубрикатора нормативных документов.

Примеры названий рубрик:

- Арбитражный процесс. Отдельные виды споров
    - *Арбитражное судопроизводство*
- Трудовое право. Социальное обеспечение и социальное страхование
    - Занятость населения и социальная защита при безработице
        - *Трудоустройство*
- Трудовое право. Социальное обеспечение и социальное страхование
    - Социальное страхование и социальное обеспечение
        - *Социальное страхование и социальное обеспечение (частные вопросы)*
        - *Социальное страхование и социальное обеспечение (общие вопросы)*
- Государственное право (государственное устройство)
    - Конституционный строй
        - Правотворческая деятельность органов государственной власти
            - *Порядок опубликования и вступления в силу нормативных правовых актов*

Для каждой рубрики было предоставлено множество примеров документов, относящихся к рубрике. Минимальное количество примеров для рубрики – 50, максимальное – 615.

Из 726 рубрик для оценки было выбрано 75 случайных рубрик.

Мы применили два классических алгоритма автоматической классификации текстов на основе машинного обучения:

1) алгоритм k ближайших соседей;
2) алгоритм SVM.

3.1. Алгоритм k ближайших соседей

Алгоритм k ближайших соседей известен как один из эффективных методов машинного обучения, применяемых для автоматической классификации текстов [5, 6].

Первым этапом работы алгоритма является преобразование текстов документов в векторы в пространстве признаков. Все слова, встретившиеся в документе, приводятся к нормальной форме. Веса слов вычисляются по формуле TF*IDF в формулировке BM25 INQUERY [1].

Для каждого документа d, подлежащего классификации, производится поиск k «ближайших соседей» — документов обучающей выборки с наибольшим рангом по метрике

$$\cos(d,x) = \frac{\sum_{i=1}^{n} d_i \cdot x_i}{\sqrt{\sum_{i=1}^{n} d_i^2} \cdot \sqrt{\sum_{i=1}^{n} x_i^2}} \qquad (1)$$

где

- d — документ-вектор, подлежащий классификации;
- x — документ-вектор из обучающей выборки;
- n — размерность пространства признаков (количество различных слов в коллекции);
- d_i, x_i — TF*IDF-вес i-го признака документа d и x соответственно.

В поставленных экспериментах использовался параметр k=10.

Для каждой рубрики c_j вычисляется релевантность документа по формуле

$$S(d,c_j) = \sum_{x \in \{\text{k nearest neighbours}\}} \cos(d,x) \cdot \theta(d,c_j) \qquad (2)$$

Документ считается принадлежащим рубрике при условии $S(d, c_j) \geq s_j$, где s_j — порог, выбранный на основе обучающей выборки по условию максимизации F-меры:

$$s_j = \arg\max_{s_j} \mathrm{F}(c_j, s_j) \qquad (3)$$

где $\mathrm{F}(c_j, s_j)$ — метрика F-мера, подсчитанная на обучающей выборке для категории c_j с порогом s_j.

Известной проблемой метода k ближайших соседей является высокая вычислительная сложность на этапе поиска ближайших соседей. Нам удалось оптимизировать алгоритм вычисления так, что обработка документов на этом этапе производилась со скоростью 6 документов в секунду на обычном офисном компьютере 3GHz CPU, 1G RAM. При увеличении коллекции до ~1 000 000 документов скорость обработки падает до 2 документов в секунду, что также вполне приемлемо.

3.2. Алгоритм SVM

Алгоритм SVM известен как один из наиболее эффективных по качеству классификации алгоритмов, применяемых для задач классификации текстов [4-6]

Векторная модель документа использовалась та же, что и для прогона 1. Для классификации использовалась известная реализация SVM — svm_light v. 6.01 [4].

Первый прогон (uis_svm_0) представляет собой результат работы svm_light с параметрами по умолчанию, без оптимизации параметров.

Для второго прогона (uis_svm_opt) производилась оптимизация порога релевантности документов. Порог выбирался по критерию максимизации F-меры на обучающей выборке. Этот метод хорошо зарекомендовал себя в наших исследованиях на коллекции нормативных документов РОМИП 2004 года.

К сожалению, при обработке входных данных была допущена ошибка, и в результате метрики по официальному прогону (uis2) не отражают поведение алгоритма.

После получения результатов мы обнаружили и исправили ошибку. Результаты по исправленному прогону приведены в следующем разделе.

3.2. Описание результатов

На рис. 2 приведены метрики результатов участников дорожки классификации нормативных документов РОМИП2008. Результат метода ближайших соседей обозначен uis_knn, результат прогона 2 (с ошибкой) обозначен uis2.

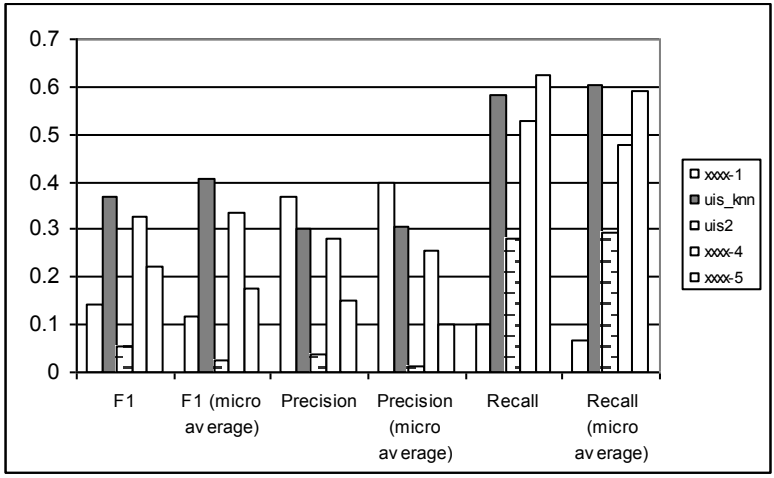

Рис.2 РОМИП2008: классификация нормативных документов

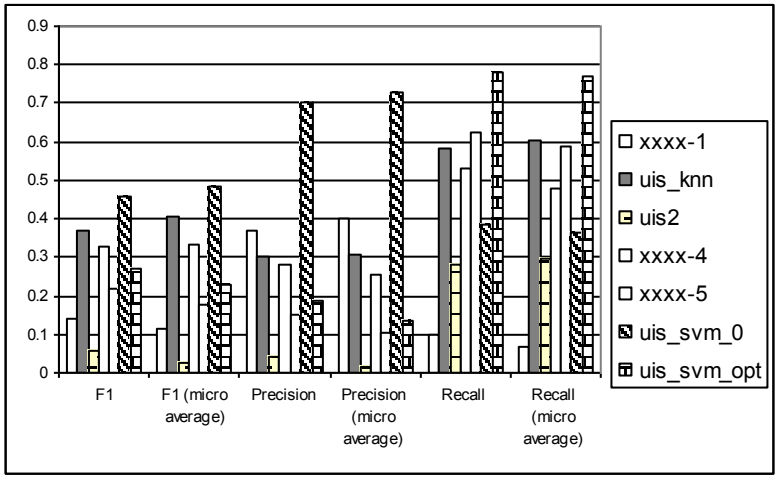

Рис.3 РОМИП2008: классификация нормативных документов + 2 (неофициальных) прогона

На рис. 3 — те же результаты, плюс два прогона — результат работы исправленных алгоритмов SVM с параметрами по умолчанию (uis_svm_0) и SVM с оптимизацией порога (uis_svm_opt).

Видно, что метод SVM с параметрами по умолчанию показал лучший результат по сравнению с другими представленными методами, что согласуется со многими исследованиями эффективности алгоритмов машинного обучения для задачи классификации текстов.

Метод ближайших соседей показал результат хуже SVM, но также вполне применим.

Метод SVM с оптимизацией порога показал результат значительно хуже, чем SVM с параметрами по умолчанию, что не согласуется с нашими ожиданиями. На наш взгляд, это связано с особенностями формирования обучающей выборки для данной дорожки РОМИП.

Стоит отметить что, не смотря на сравнительные преимущества методов SVM и kNN, общее качество классификации — на уровне 40%-45% F-меры, не очень высоко. По нашему мнению, этот факт мог бы объясняться противоречивостью обучающей коллекции. В этом случае метод kNN имеет определенные методологические преимущества, так как может явно указать на существующие противоречия – разная классификация похожих документов, что позволяет организовать эффективную процедуру очистки обучающей коллекции.

3.3. Критический анализ состава обучающей выборки

В этом году метод формирования обучающей выборки для дорожки классификации нормативных документов отличался от используемых ранее.

Во-первых, для тестирования были выбраны только рубрики, для которых есть не менее 50 документов в коллекции.

Во-вторых, набор документов в коллекцию для обучения производился не равномерно, а следующим образом: для каждой рубрики, имеющей не менее 50 документов, были выбраны случайные 50 релевантных документов.

В результате, распределение документов по рубрикам получилось существенно различным в обучающем и тестовом множестве документов.

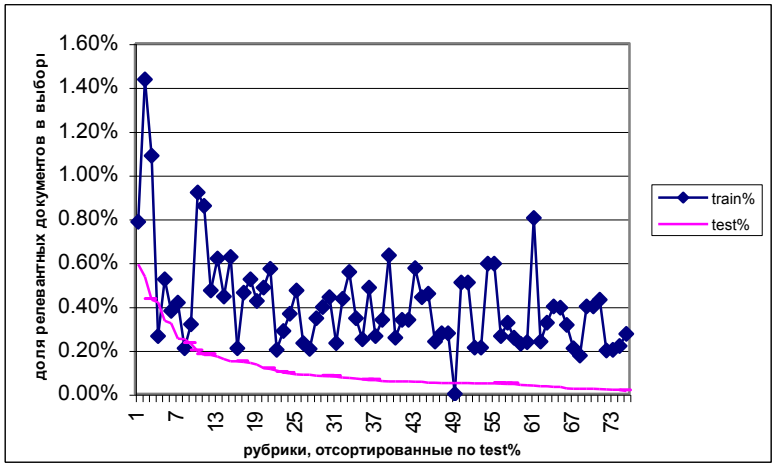

Рис.4 Распределение доли документов, релевантных рубрике, в обучающем и тестовом множестве

На рис. 4 показана гистограмма распределения процента документов, принадлежащих рубрике в тестовой и обучающей выборках (только для рубрик, попавших в оба множества).

Видно, что, с одной стороны, количество документов распределено по рубрикам неравномерно даже в обучающей выборке (от 51 до 420). С другой стороны, по количеству документов в обучающей выборке невозможно оценить популярность рубрики. Это делает невозможным корректный подбор параметров, влияющих на соотношение полнота/точность.

На наш взгляд, такой способ формирования обучающей выборки не соответствует реальным задачам классификации текстов, и предпочтительнее использовать в качестве обучающей выборки случайное подмножество документов коллекции.

Заключение

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

Опыт показывает, что для получения высоких результатов в дорожках РОМИП является полезным более полный учет специфики решаемых задач, настройка на особенности коллекций.

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

[2] 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.

[3] Лукашевич Н.В., Автоматическое рубрицирование потоков текстов по общественно-политической тематике // НТИ. Сер.2. - 1996. - N 10. - С.22-30.

[4] Joachims T., Making large-scale support vector machine learning practical // Advances in Kernel Methods: Support Vector Machines / B.Schölkopf, C. Burges, A. Smola (eds.) - MIT Press: Cambridge, MA" – 1998. (http://svmlight.joachims.org/)

[5] Yang Y., Liu X., A re-examination of text categorization methods // Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval / M.A.Hearst, F.Gey, R.Tong (eds.) - ACM Press: New York, Berkeley, 1999 – pp. 42—49

[6] Joachims T., Text Categorization with Support Vector Machines: Learning with Many Relevant Features // Proceedings of ECML-98, 10th European Conference on Machine Learning, 1998.

UIS RUSSIA at ROMIP 2008:
Ad Hoc Search and Text Categorization of Legal Documents

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

In the paper we describe methods used by the 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 ROMIP 2008 (Russian Information Retrieval Evaluation Seminar) tracks. We participated in the ad hoc track and text categorization on a legal documents collection. In the ad hoc track we used a retrieval model intended for processing of long information queries. The model consists of two processing stages: 1) combined vector retrieval model based on words and thesaurus concepts; 2) reranking of found documents according to the distribution of query elements in document sentences. In text categorization track we used well-known algorithms: kNN and SVM.