RCO на РОМИП 2009

© Поляков П.Ю., Плешко В.В., Ермаков А.Е.
info@rco.ru

Аннотация

Настоящая работа является отчетом об экспериментах, проведенных в рамках семинара РОМИП 2009 года. Проведены исследования влияния метода рубрикации в задаче классификации web-страниц и сайтов. Также апробирован новый метод преобразования поисковых запросов на коллекции нормативно-правовых документов.

1. Введение

Система тематической классификации RCO, основанная на методе опорных векторов, стабильно показывает хорошие результаты как на дорожках классификации нормативно-правовых документов, так и на дорожках классификации веб-сайтов и веб-страниц [1-3]. Выбор метода опорных векторов в качестве основы нашей системы был сделан по двум причинам. Во-первых, этот метод считается наиболее подходящим для классификации текстовых данных. Во-вторых, результаты по дорожкам классификации РОМИП, полученные другими методами, оказывались более слабыми по сравнению с SVM [4,5]. Однако, в работе [3] показано, что качество классификации, прежде всего, зависит от выбора классификационных терминов, а не от настроек SVM. Кроме того, наши предыдущие эксперименты, связанные с улучшением экстраполяционных свойств профиля, построенного методом SVM путем отбрасывания из него терминов с отрицательными весами, навели на мысль, а настолько ли SVM лучше своих более простых собратьев, как все говорят. Чтобы проверить это преположение, мы сопоставили в данной работе результаты SVM и некоторых модификаций Байеса, полученных на одном и том же наборе классификационных признаков.

Также в рамках цикла РОМИП-2009 нами был опробован новый способ преобразования запросов на естественном языке в языки запросов поисковых машин, основанный на машинном анализе синтаксических связей между словами и их отображении на соответствующие операторы языка поисковой машины с максимальным сохранением смысла исходного запроса [11].

2. Тематическая классификация веб-страниц и веб-сайтов

2.1 Постановка задачи

Участникам было предложено подмножество интернет-каталога dmoz.org (300000 страниц), используя которое в качестве обучающей выборки, требовалось соотнести с категориями каталога dmoz.org (247 категорий) страницы из домена .by (600714 страниц).

Среди особенностей задачи следует отметить «зашумленность» обучающей выборки. Если сайт из обучающей выборки принадлежал категории, то все его страницы относились к положительным примерам этой категории.

2.2 Методы классификации

Исследования проводились в рамках векторной модели представления документов. В сравнении участвовали метод опорных векторов с линейным ядром и байесовский классификатор с пуассоновской функцией плотности распределения вероятностей.

2.2.1 Метод опорных векторов с линейным ядром

Мы остановились на линейном ядре SVM, так как в этом случае облегчается интерпретация и настройка профилей рубрик, и как показано в работе [2], использование различных вариантов ядра в методе опорных векторов не оказывает заметного влияния на качество классификации.

Линейному ядру соответствует обычный линейный классификатор вида $\mathbf{d}*\mathbf{c} > h$, где **d** – вектор документа, **c** – вектор профиля рубрики, h – пороговое значение, которое должно превысить скалярное произведение упомянутых векторов для отнесения документа к рубрике. Размерность векторов равна числу терминов, задействованных в классификации. При классификации веб-сайтов вектор **d** представлял собой суперпозицию веб-страниц сайта.

Использовалась реализация SVM-Light [6] с параметрами:
b = 1, j = (число положительных примеров в обучающей выборке) / (число отрицательных примеров в обучающей выборке).

2.2.2 Байесовский классификатор с распределением Пуассона.

Из всех возможных вариантов байесовских классификаторов мы остановились на модификации, использующей распределение Пуассона в качестве функции плотности распределения вероятностей слов в тексте. Формулы для вычисления весов слов в этом случае оказываются достаточно простые, а пуассоновское распределение соответствует реальному распределению частот слов в тексте [7].

Согласно теореме Байеса, вероятность того, что документ d_j принадлежит классу с можно представить в виде:

$$p(c \mid d_j) = \frac{\exp(z_{jc})\,p(c)}{\exp(z_{jc})\,p(c) + p(\bar{c})}, \tag{1}$$

где

$$z_{jc} = \log \frac{p(d_j, c)}{p(d_j, \bar{c})}. \tag{2}$$

Будем считать, что документ d_j описывается вектором переменных f_{ij} – частотой i-го термина в j-м документе. Если преположить, что термины появляются в документе независимо друг от друга и с вероятностями $P(f_{ij})$, тогда

$$p(d_j) = \prod_{i=1}^{|V|} P(f_{ij}), \tag{3}$$

где $|V|$ – размер словаря терминов. Исходя из распределения Пуассона частот слов в документе

$$P(f_{ij}) = \frac{e^{-\lambda}\lambda^{f_{ij}}}{f_{ij}!}, \tag{4}$$

где λ – мат. ожидание частоты i-го термина. Подставляя (3) и (4) в (2), представим функцию z_{jc} в виде

$$z_{jc} = \sum_{i=1}^{|V|} \log \frac{P(f_{ij}, c)}{P(f_{ij}, \bar{c})} = \sum_{i=1}^{|V|} \log \frac{e^{-\lambda_i}\lambda_i^{f_{ij}}}{e^{-\mu_i}\mu_i^{f_{ij}}} = \sum_{i=1}^{|V|} \left(f_{ij} \log \frac{\lambda_i}{\mu_i} - \lambda_i - \mu_i \right), \tag{5}$$

где λ_i и μ_i – мат. ожидания частоты i-го термина в документах, принадлежащих и не приналежащих классу c соответственно.

Согласно определению распределения Пуассона, f_{ij} – частота термина в документе фиксированной длины. Поэтому перед использованием этой величины в формуле (5), необходимо

произвести нормировку к длине документа. Кроме того, во многих ранних работах по информационному поиску пишут о том, что можно построить более качественную модель, если сгладить частоту термина в документе. Приняв во внимание оба этих аспекта, получим следующую оценку величины f_{ij} в формуле (5):

$$\tilde{f}_{ij} = \frac{x_{ij} + \theta}{dl_j + \theta \cdot |V|}, \qquad (6)$$

где x_{ij} – действительная частота i-го термина в j-м документе, dl_j – число слов в j-м документе, θ – параметр сглаживания, который в наших экспериментах был равен 0.01.

При вычислении λ_i и μ_i также следует пользоваться вместо величины f_{ij} её нормированным значением $\tilde{f}_{ij}$:

$$\tilde{\lambda} = \frac{1}{|D_c|} \sum_{d_j \in D_c} \tilde{f}_{ij}, \qquad \tilde{\mu} = \frac{1}{|D_{\bar{c}}|} \sum_{d_j \in D_{\bar{c}}} \tilde{f}_{ij}, \qquad (7)$$

где D_c и $D_{\bar{c}}$ – множества документов обучающей выборки, принадлежащих и не принадлежащих классу c соответственно.

Для классификации документов не требуется знать в точности вероятность отнесения документа к классу, необходимо лишь определить, превышает или нет эта величина некий порог h, то есть проверить условие $z_{jc} > h$. Более того, если порог h вычисляется отдельным алгоритмом (в нашем случае максимизацией F-меры на обучающей выборке), то в (5) можно опустить члены, не зависящие от частоты терминов. В итоге мы приходим к обычному линейному классификатору вида $\mathbf{d}*\mathbf{c} > h$, где вектор $\mathbf{d}$ для j-го документа имеет компоненты $\tilde{f}_{ij}$, а вектор $\mathbf{c}$ имеет компоненты $\log(\tilde{\lambda}_i / \tilde{\mu}_i)$.

2.3 Отбор терминов

Отбор терминов проводился из набора положительных примеров для каждой из категорий. Отбирались как однословные, так и многословные термины. В качестве однословных терминов выделялись все слова документа за исключением служебных частей речи, числительных и дат. Многословные термины выделялись при помощи алгоритма синтактико-семантического анализа [8] и представляли собой простые именные группы (напр. «подоходный налог, база налогообложения»). Именные группы были усложнены включением в их структуру конструкций с предлогами в соответствии с моделями управления [9] (напр. «налог на добавленную стоимость»).

Выделенные термины подвергались фильтрации для каждой категории отдельно. Фильтрация производилась по информационной значимости термина. За основу информационной значимости термина был выбран коэффициент IG (information gain, см. например [10]):

$$IG(t_i, c_k) = \sum_{c \in \{c_k, \bar{c}_k\}} \sum_{t \in \{t_i, \bar{t}_i\}} P(t,c) \cdot \log(\frac{P(t,c)}{P(t)P(c)}) \cdot \quad (8)$$

Нами были взяты только первое и четвертое слагаемые, характеризующие описательную способность термином рубрики. Фильтрация по признаку информационной значимости проводилась следующим образом:

1. Термины упорядочивались по убыванию информационной значимости;

2. Далее были отобраны первые N терминов, сумма информационной значимости которых составила 80% от общей суммы по всем терминам. Влияние числа терминов на качество результата следующее: при увеличении N качество слегка падает, а при уменьшении появляется неустойчивость (существенно усиливается разброс индивидуальных показателей рубрик).

2.4 Взвешивание терминов

В методе опорных векторов использовался частотный способ взвешивания термина, так как именно он давал наилучшие результаты в предыдущих работах. При расчете весов терминов в документе создавался вектор, ненулевыми элементами которого служили частоты терминов в документе, отобранных для данной категории. Затем вектор документа приводился к единичной длине.

В байесовском классификаторе вес термина в документе вычислялся непосредственно по формуле (6).

2.5 Описание прогонов

На оценку было представлено 3 прогона:

1. SVM – метод опорных векторов из п. 2.2.1;

2. Bayes – байесовский классификатор из п. 2.2.2, в котором вес термина в профиле вычисляется по формуле $\log(\tilde{\lambda}_i / \tilde{\mu}_i)$;

3. Bayes-IG – байесовский классификатор из п. 2.2.2, в котором производится дополнительная нормировка профилей рубрик. Эта нормировка позволяет придать больший вес терминам, имеющим большую информационную значимость для данной рубрики. Вес *i*-го термина в профиле *k*-й рубрики вычисляется по формуле

$$\frac{IG(t_i, c_k)}{\sum_i IG(t_i, c_k)} \log(\tilde{\lambda}_i / \tilde{\mu}_i) \cdot \quad (8)$$

2.6 Результаты оценки классификации веб-страниц

micro	and			or		
	r	p	F1	r	p	F1
SVM	0.26	0.35	0.30	0.26	0.57	0.35
Bayes	0.36	0.41	0.39	0.34	0.65	0.45
Bayes-IG	0.44	0.45	0.45	0.41	0.68	0.51

macro	and			or		
	r	p	F1	r	p	F1
SVM	0.26	0.35	0.25	0.27	0.57	0.34
Bayes	0.40	0.35	0.30	0.32	0.59	0.39
Bayes-IG	0.43	0.40	0.35	0.39	0.63	0.45

Таблица 1. Полнота, точность и F1 рубрик с сильными (and) и слабыми (or) требованиями к релевантности для микро- и макро-усреднения.

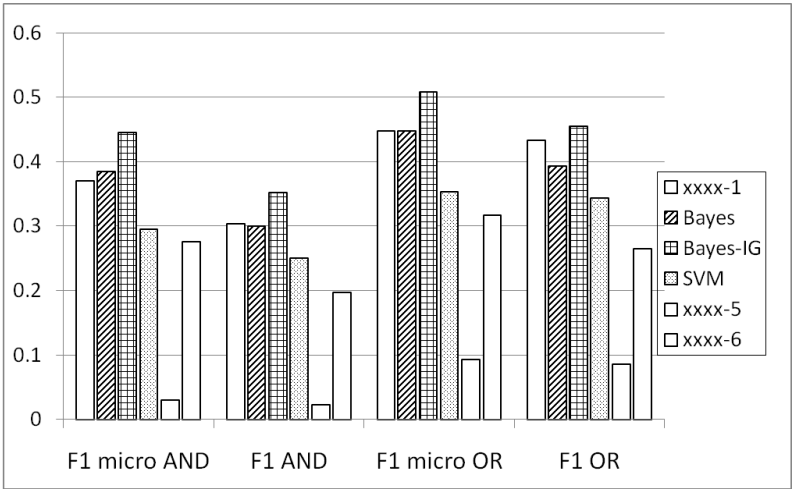

Рисунок 1. Значения F1(micro) и F1(macro) с сильными и слабыми требованиями к релевантности для участников дорожки классификации web-страниц.

Как видно из таблицы 1 и рисунка 1, при классификации веб-страниц результаты упорядочиваются следующим образом: Bayes-

IG > Bayes > SVM. Причем преимущество лучшего прогона над худшим достигает от 20% до 30%.

2.7 Результаты оценки классификации веб-сайтов

micro	and			or		
	R	p	F1	r	p	F1
xxx-1	0.08	0.62	0.13	0.05	0.74	0.09
Bayes-IG	0.51	0.43	0.47	0.41	0.68	0.51
SVM	0.58	0.46	0.51	0.44	0.68	0.53

macro	and			or		
	r	p	F1	R	p	F1
xxx-1	0.14	0.39	0.14	0.08	0.49	0.11
Bayes-IG	0.36	0.39	0.32	0.27	0.61	0.35
SVM	0.44	0.44	0.39	0.34	0.68	0.41

Таблица 2. Полнота, точность и F1 рубрик с сильными (and) и слабыми (or) требованиями к релевантности для микро- и макро-усреднения.

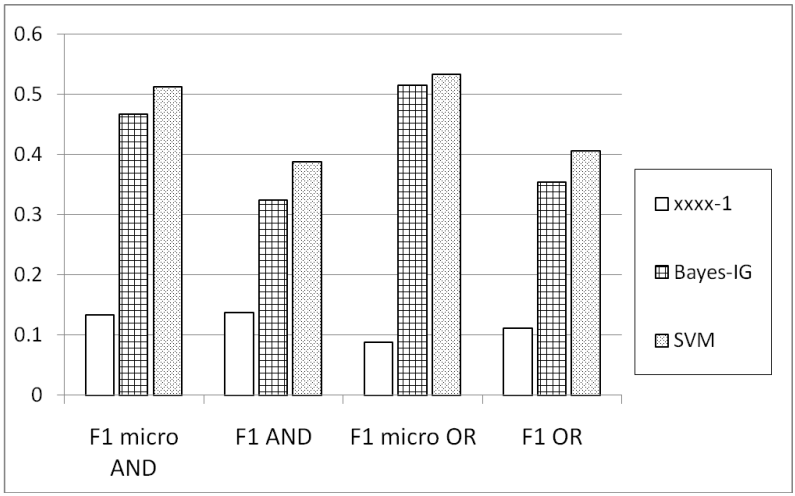

Рисунок 2. Значения F1(micro) и F1(macro) с сильными и слабыми требованиями к релевантности для участников дорожки классификации web-сайтов.

При классификации веб-сайтов ситуация меняется на противоположную – SVM > Bayes-IG. Как видно из таблицы 2 и рисунка 2, результаты упорядочиваются следующим образом: SVM > Bayes-IG > xxx-1. Причем преимущество лучшего прогона над

худшим в данном случае не такое большое – не более 10% (см. таблицу 2).

3. Поиск по коллекции нормативно-правовых документов

3.1 Постановка задачи

Система-участник получает коллекцию нормативных документов (всего - 300 000) и набор запросов. Ответом системы на каждое задание является упорядоченный список документов, длиной не более 100 ссылок.

3.2 Описание метода

Исследуемый нами в данной работе подход является метапоисковым и призван повысить качество работы уже готовой поисковой машины (базового метода). Поэтому выбор поисковой машины в рамках данного эксперимента не является существенным.

3.2.1 Базовый метод

В качестве базового метода была выбрана одна из вариаций взвешивания tf*idf, реализованная в одной из распространенных коммерческих СУБД.

3.2.2 Метод преобразования запроса

Основная идея метода [11] заключается в использовании синтаксических связей между словами запроса для применения различных поисковых ограничений, отражающих силу этих связей, а также в последующем пошаговом ослаблении ограничений (вплоть до tf*idf) для достижения желаемой полноты.

Мы использовали отображение связей между словами на операторы из следующего множества: И, РЯДОМ, РЯДОМ_УПОРЯДОЧЕННО, ФРАЗА, ВО_ВСЕХ_ФОРМАХ. Рассмотрим пример конструирования запроса к поисковой машине для запроса *авансовые платежи налог на прибыль предприятий*. Соответствующее выражение на некотором языке запросов может выглядеть так:

(m:АВАНСОВЫЙ m:ПЛАТЕЖ) and (m:НАЛОГ near_ord (НА near_ord (ПРИБЫЛЬ ПРЕДПРИЯТИЙ=ПРЕДПРИЯТИЯ))).

Здесь *m:* означает учет словоформ, = означает возможность присутствия любого из перечисляемых слов в указанной позиции,

near означает, что слова расположены рядом в тексте, near_ord – рядом в указанном порядке.

Затем мы строили последовательность запросов, ослабляя поисковые ограничения путем следующих преобразований:

- удаление зависимых слов в словосочетаниях (с наибольшим расстоянием в дереве синтаксического разбора от главного слова, в случае равенства использовалось расстояние от главного слова в строке запроса);
- удаление общеупотребимых слов (возможна настройка на коллекцию путем придания классам слов и терминов весовых коэффициентов и удаления на каждом шаге слов с наименьшим весом);
- замена операторов на «более слабые».

При прочих равных условиях из запроса отбрасывалось слово с наибольшим порядковым номером (стоящее справа).

Пример применения преобразований:

1. (m:ПЛАТЕЖ) and (m:НАЛОГ near_ord (НА near_ord (ПРИБЫЛЬ)));
2. (m:АВАНСОВЫЙ m:ПЛАТЕЖ) and (m:НАЛОГ near (m:ПРИБЫЛЬ ПРЕДПРИЯТИЙ=ПРЕДПРИЯТИЯ));
3. (m:АВАНСОВЫЙ m:ПЛАТЕЖ) or (m:НАЛОГ and (m:ПРИБЫЛЬ ПРЕДПРИЯТИЙ=ПРЕДПРИЯТИЯ));
4. (m:ПЛАТЕЖ) or (m:НАЛОГ and m:ПРИБЫЛЬ).

Последним в последовательности всегда стоял базовый запрос (tf*idf).

Каждый запрос в порядке ослабления ограничений последовательно выполнялся, и новые документы, найденные при помощи очередного запроса, добавлялись в конец выборки. Таким образом, если преобразования запроса не давали ни одного документа, результат выполнения запроса сводился к базовому методу.

Мы рассчитывали, что данный метод будет иметь преимущество над базовым на длинных запросах.

3.3 Результаты оценки

В таблице 3 приведены результаты оценки Базового прогона, прогона с использованием метода преобразований, а также лучшего из прогонов других участников. Применение преобразований запроса дает улучшение всех показателей в 3-4 раза по сравнению с базовым методом.

Однако выбранный (видимо, ошибочно) способ ранжирования в базовом методе показал результаты значительно хуже результатов прогонов других участников. Это не позволяет сделать окончательных выводов об эффективности метода преобразований.

Показатель	Базовый	Преобразования
Precision(10)	0.083	0.22
Bpref-10	0.064	0.17
Bpref	0.046	0.15
Recall	0.13	0.24
Average precision	0.03	0.12
Precision	0.049	0.11
R-precision	0.055	0.16
Precision(5)	0.087	0.23

Таблица 3. Результат оценки and для дорожки legal-2007 adhoc

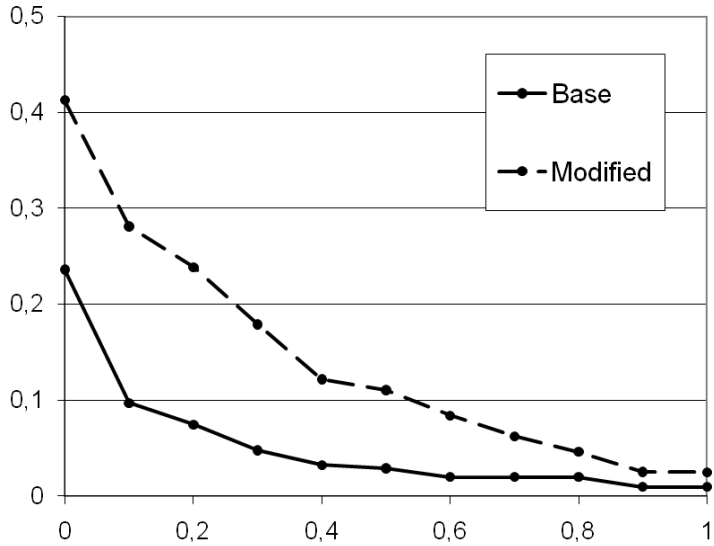

Рисунок 3. 11-точечные графики для Базового метола и Метода преобразований запроса для дорожки legal-2007 adhoc (оценка and)

Как показано на рисунке 3, точность, полученная для модифицированных запросов, значительно превосходит точность базового метода на всей области определения 11-точечного графика.

В принципе, эффект от преобразований должен проявляться на длинных запросах. Поэтому при сравнении подходов следовало бы рассматривать результаты только на длинных запросах. Если пересчитать оценки на множестве длинных запросов, эффект от преобразований запроса будет проявляться ещё сильнее.

4. Заключение

В работе показано, что Байесовский метод классификации с распределением Пуассона не уступает по качеству методу SVM при классификации web-страниц и сайтов. Показано, что дополнительная нормировка профилей рубрик с использованием величины информационного выигрыша улучшает качество рубрикации Байесовского классификатора.

Также опробован метод преобразования запросов на основе синтаксических связей между словами запроса [11]. Показано значительное повышение точности на длинных запросах в сравнении с методом, использующим векторную модель. Однако полученные результаты не позволяют сделать окончательные выводы об эффективности метода.

Литература

[1] *Плешко В.В., Ермаков А.Е., Голенков В.П., Поляков П.Ю.* RCO на РОМИП 2005 // Труды третьего российского семинара РОМИП'2005. (Ярославль, 6 октября 2005г.). - Санкт-Петербург: НИИ Химии СПбГУ - 2005 - с. 106-124.

[2] *Поляков П.Ю., Плешко В.В.,* RCO на РОМИП 2006 // Труды четвертого российского семинара РОМИП'2006. (Суздаль, 19 октября 2006г.). - Санкт-Петербург: НУ ЦСИ – 2006 - с. 72-79.

[3] *Плешко В.В., Поляков П.Ю.,* RCO на РОМИП 2008 // Труды РОМИП 2007-2008. (Дубна, 9 октября 2008г.). - Санкт-Петербург: НУ ЦСИ, 2008 - с. 96-107.

[4] *Максаков А.*, Сравнительный анализ алгоритмов классификации и способов представления Web-документов. // Труды третьего российского семинара РОМИП'2005. (Ярославль, 6 октября 2005г.). - Санкт-Петербург: НИИ Химии СПбГУ - 2005- с. 63-73.

[5] *Агеев М.С., Добров Б.В., Лукашевич Н.В., Штернов С.В.*, УИС РОССИЯ в РОМИП 2008: поиск и классификация нормативных документов. // Труды РОМИП 2007-2008. (Дубна, 9 октября 2008г.). - Санкт-Петербург: НУ ЦСИ, 2008 - с. 44-58.

[6] *Joachims T.* Making large-scale support vector machine learning practical // Advances in Kernel Methods: Support Vector Machines /

B.Scholkopf, C. Burges, A. Smola (eds.) - MIT Press: Cambridge, MA" – 1998.
[7] *Antic G., Stadlober E., Grzybek P., Kelih E*. Word Length and Frequency Distributions in Different Text Genres // Studies in Classification, Data Analysis, and Knowledge Organization - Springer Berlin Heidelberg - 2006.
[8] *Ермаков А.Е.* Значимость элементов текста в свете теории синтаксической парадигмы // Русский язык: исторические судьбы и современность. II Международный конгресс исследователей русского языка. Труды и материалы. - Москва: МГУ - 2004.
[9] *Ермаков А.Е*. Эксплицирование элементов смысла текста средствами синтаксического анализа-синтеза // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог'2003 (Протвино, 11-16 июня, 2003 г.). – Москва, Наука, 2003
[10] *H. Avancini, A. Lavelli, F. Sebastiani, R. Zanoli*. Automatic expansion of domain-specific lexicons by term categorization // ACM Transactions on Speech and Language Processing (TSLP) Discovery – 2006, V.3, No.1 – pp.1-30.
[11] *Ермаков А.Е., Плешко В.В.* Обработка естественно-языковых запросов к поисковой машине на основе их лингвистического анализа // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009» (Бекасово, 27-31 мая 2009 г.). Вып. 8 (15).– М.: РГГУ, 2009. 620 с.

RCO at RIRES 2009

Polyakov P.Yu., Pleshko V.V., Ermakov A.E.

This article presents report on experiments in IR that were driven as a part of RIRES seminar. The research was taken on different comparison of Bayes and SVM methods that affect quality of web-site and web-page classification task. Also an approach to query transformation based on mapping of syntactic links between terms to proximity search operators was studied on legal adhoc task.