Галактика-Zoom на РОМИП'2009

Антонов А.В. Баглей С.Г.

Мешков В.С. Стоян В.Н.

Корпорация "Галактика"
{alexa, baglei, meshkov, stoyan}@galaktika.ru

Аннотация

В статье представлены результаты участия поисково-аналитической системы обработки больших объемов неструктурированных данных "Галактика-Zoom" в дорожках РОМИП: "Тематическая классификация нормативно-правовых документов", "Тематическая классификация Веб-страниц", "Тематическая классификация Веб-сайтов". Приведено сравнение полученных результатов с результатами, показанными системой в предыдущий цикл семинара.

1. Введение

Участие в семинаре РОМИП'2009 явилось новым этапом исследования возможностей и функциональности нашей системы. При обработке заданий РОМИП в этом году мы смогли оценить качество работы выбранной меры близости, используемой при классификации документов, в условиях отсутствия практических ограничений на количество признаков, использующихся в классификации, при том условии, что данные признаки выбраны системой в качестве значимых. Это позволило нам улучшить качество работы системы "Галактика-Zoom", получить независимую оценку обработки заданий, основанных на реальных текстовых массивах.

2. Методы классификации документов в системе “Галактика-Zoom”

2.1 Представление документа для задачи классификации

Основным понятием в системе “Галактика-Zoom” является понятие Информационного портрета выборки документов (ИнфоПортрета). ИнфоПортрет представляет собой список языковых инвариантов (слов и словосочетаний), отличающих данную выборку от прочих. Технология построения информационного портрета, детально описанная в работах [2, 3, 4], основана на статистических методах обработки текстовой информации. Формирование информационного портрета отдельных документов выполняется на базе характеристик элементов сформированного ИнфоПортрета, а также собственной статистики текста документа. Для каждого документа система формирует упорядоченный список слов и словосочетаний, статистически отличающих данный документ от прочих в выборке. ИнфоПортрет, связанный с документом, рассматривается в качестве образа документа для проведения классификации.

2.2 Представление множества документов

Представление отдельных рубрик для проведения классификации также формировалось через построение ИнфоПортретов этих рубрик.

После формирования ИнфоПортрета рубрики применялась объектная модель представления множества документов с помощью элементов ИнфоПортрета. Метод, использованный для построения модели, подробно описан в работе [5].

2.3 Метод опорных векторов (Support Vector Mashines)

В качестве основы для проведения классификации с помощью метода опорных векторов [8] была взята его реализация SVMLight [7].

На этапе обучения алгоритма на основе тренировочного множества документов строились ИнфоПортреты для каждой рубрики, вошедшей в задание. Далее, мы использовали пространство элементов, составляющих полученный ИнфоПортрет, для формирования представления всех документов, составляющих тренировочный массив. Элементы Инфопортрета, входящие в документы искомой рубрики с соответствующими им весами, и

элементы ИнфоПортрета всех остальных документов принимались в качестве двух тренировочных множеств для обучения алгоритма.

Для поведения классификации была выбрана модификация метода, использующая линейное ядро. В расчетах использовались все полученные элементы ИнфоПортрета. Тем самым, мы стремились максимально расширить пространство признаков для классификации.

3. Результаты классификации по отдельным дорожкам

Одной из целей исследования в рамках семинара явилось сравнение полученных данных с результатами, достигнутыми в прошлом году. Оценивая результаты РОМИП'2008, можно заметить, что в прошлом году после выполнения заданий в алгоритме метода была обнаружена ошибка, связанная с введением искусственного ограничения пространства признаков при классификации документов. Выполняя задания семинара РОМИП'2009, мы исправили данную ошибку. Сравнением прошлогодних результатов с результатами этого года, учитывая при этом вероятность отклонения полученных оценок, связанную с различием оцениваемых рубрик, мы смогли получить некоторое представление о влиянии ограничения пространства признаков на итоговые результаты классификации.

Далее приведены оценки результатов классификации, полученные нашей системой.

3.1 Классификация Веб-сайтов

Для классификации Веб-сайтов использовался метод, основанный на построении матрицы близости ИнфоПортретов. Модификация данного метода с некоторыми пороговыми отличиями уже была исследована нами при обработке заданий РОМИП на предыдущих семинарах.

Рис. 1. Оценки качества классификации по дорожке “Классификация Веб-сайтов” с использованием метода макроусреднения

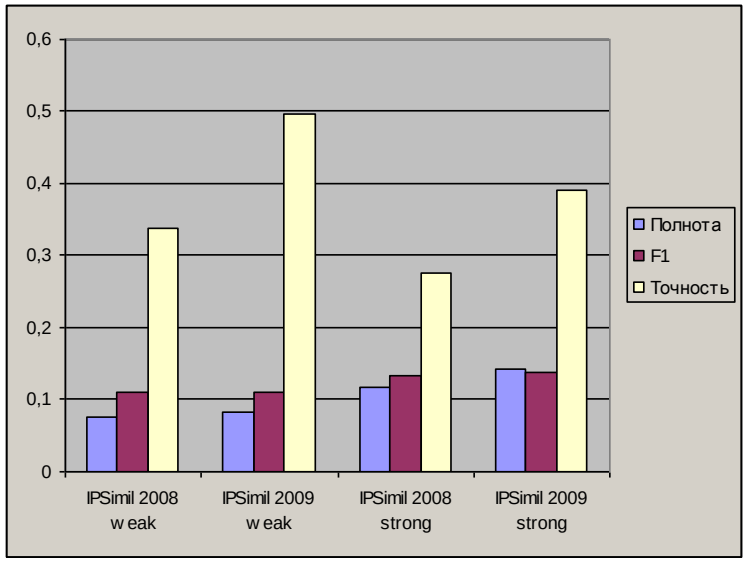

Таблица 1. Оценки качества классификации системы “Галактика-Zoom” по дорожке “Классификация Веб-сайтов” с использованием метода макроусреднения

	Полнота	F1	Точность
IPSimil “сильная” оценка, 2008	0,1172	0,1333	0,2766
IPSimil “сильная” оценка, 2009	0.1427	0.1378	0.3918
IPSimil “слабая” оценка, 2008	0.0755	0.1093	0.3377
IPSimil “слабая” оценка, 2009	0.082	0.1108	0.4959

Рис. 2. Оценки качества классификации по дорожке “Классификация Веб-сайтов” с использованием метода микроусреднения

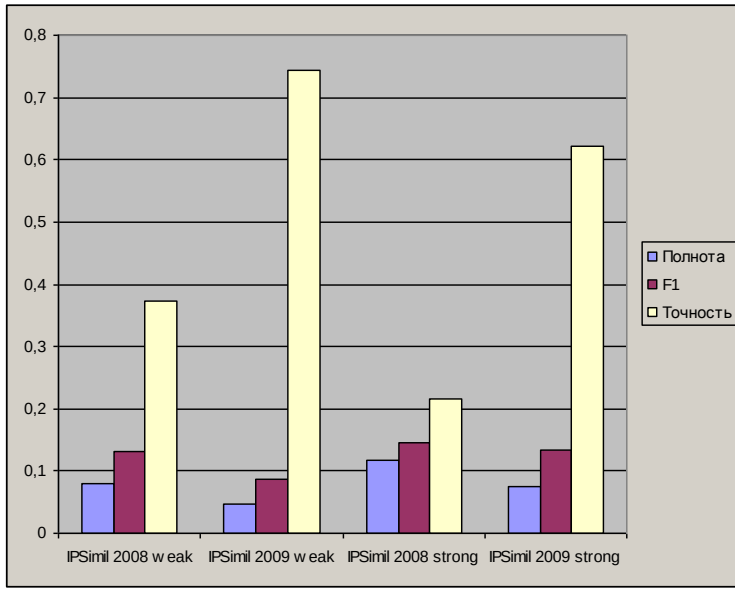

Таблица 2. Оценки качества классификации, присвоенные прогонам системы “Галактика-Zoom” по дорожке “Классификация Веб-сайтов ” с использованием метода микроусреднения

	Полнота	F1	Точность
IPSimil “сильная” оценка, 2008	0,1172	0,1461	0,2156
IPSimil “сильная” оценка, 2009	0.0746	0.1333	0.6216
IPSimil “слабая” оценка, 2008	0.0803	0.1321	0.3725
IPSimil “слабая” оценка, 2009	0.0464	0.0874	0.7432

Из результатов заметно, что среди оценок качества классификации в лучшую сторону изменились значения точности при незначительном изменении полноты. Как следствие, F-мера также изменилась незначительно.

3.2 Классификация Веб-страниц

Для классификации документов по дорожкам нормативно-правовых документов и Веб-страниц была выбрана модификация метода SVM с линейным ядром, использующая для классификации все полученные элементы ИнфоПортрета. В качестве метрики близости ИнфоПортретов использована метрика Jenson-Shannon (JS). Данная метрика близости была выбрана по результатам нашего исследования эффективности различных метрик при классификации в рамках семинара РОМИП'2008.

Рис. 3. Оценки качества классификации по дорожке "Классификация Веб-страниц" с использованием метода макроусреднения

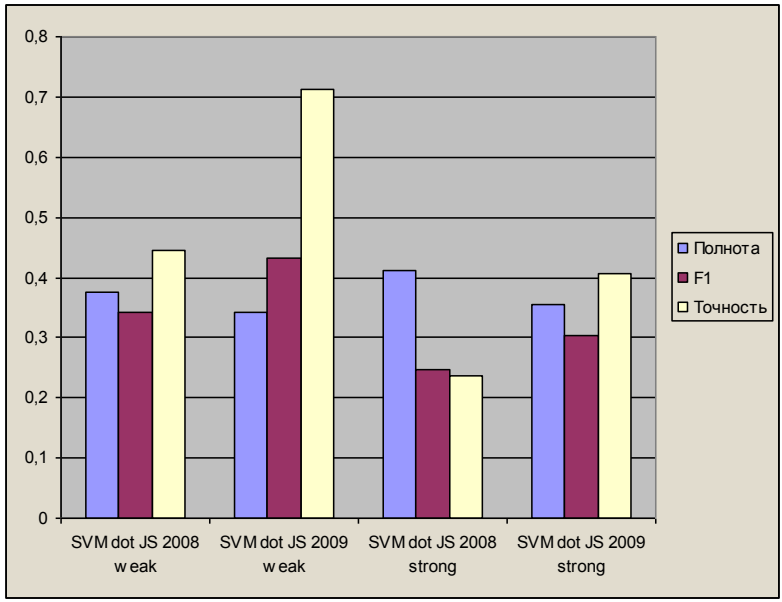

Таблица 3. Оценки качества классификации, присвоенные прогонам системы "Галактика-Zoom" по дорожке "Классификация Веб-страниц" с использованием метода макроусреднения

	Полнота	F1	Точность
SVM dot JS "сильная" оценка, 2008	0,4122	0,247	0,2375
SVM dot JS "сильная" оценка, 2009	0.3543	0.3034	0.4057

SVM dot JS “слабая” оценка, 2008	0.3748	0.3424	0.4443
SVM dot JS “слабая” оценка, 2009	0.3425	0.4327	0.7113

Рис. 4. Оценки качества классификации по дорожке “Классификация Веб-страниц” с использованием метода микроусреднения

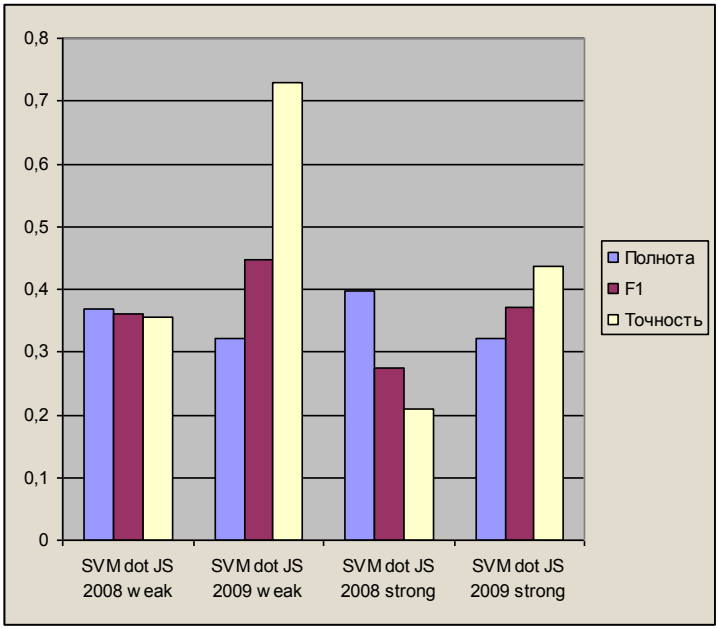

Таблица 4. Оценки качества классификации, присвоенные прогонам системы “Галактика-Zoom” по дорожке “Классификация Веб- страниц” с использованием метода микроусреднения

	Полнота	F1	Точность
SVM dot JS “сильная” оценка, 2008	0,3986	0,2753	0,2102
SVM dot JS “сильная” оценка, 2009	0.3211	0.3704	0.4377
SVM dot JS “слабая” оценка, 2008	0.3684	0.3617	0.3552
SVM dot JS “слабая” оценка, 2009	0.3225	0.4475	0.7304

Оценивая результаты классификации по данной дорожке, можно заметить очевидный прирост значений точности при не слишком сильном снижении полноты. Данное сочетание изменений позволило существенно увеличить значения F-меры по сравнению с результатами прошлого года.

3.3 Классификация нормативно-правовых документов

Для классификации так же, как и в п. 3.2, была использована модификация метода SVM с линейным ядром, при работе которой рассматривался весь ИнфоПортрет и была использована метрика близости Jensen-Shannon.

Рис. 5. Оценки качества классификации по дорожке "Классификация нормативно-правовых документов"

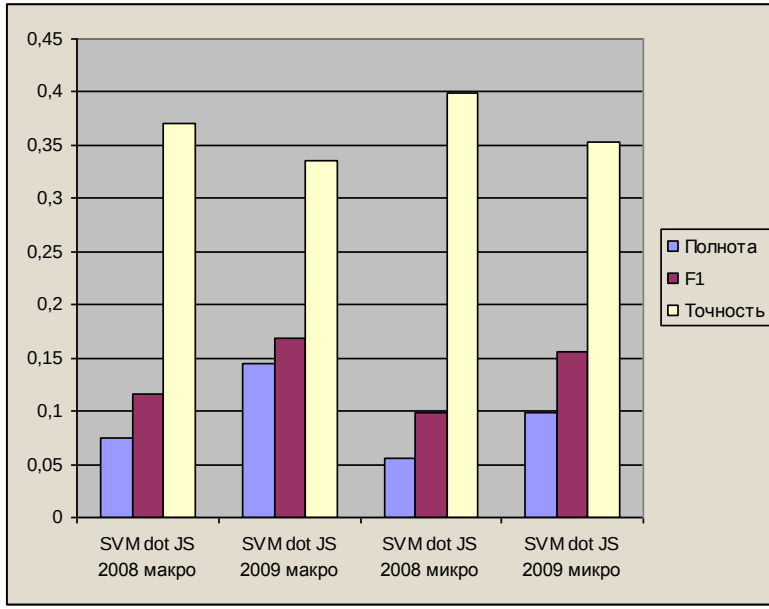

Таблица 5. Оценки качества классификации, присвоенные прогонам системы "Галактика-Zoom" по дорожке "Классификация нормативно-правовых документов"

	Полнота	F1	Точность
SVM dot JS 2008 макроусреднение	0.0748	0.1155	0.3701

SVM dot JS 2008 микроусреднение	0.0564	0.0988	0.3984
SVM dot JS 2009 макроусреднение	0.1451	0.1686	0.3353
SVM dot JS 2009 микроусреднение	0.099	0.1551	0.3523

В наших результатах по дорожке классификации нормативно-правовых документов наблюдалось увеличение полноты при несущественном снижении значений точности. Как следствие, соотношение полноты и точности стало более сбалансированным, что привело к увеличению значений F-меры.

4. Заключение

Нам удалось провести оценку качества работы метода SVM, в работе которого используется метрика близости Jenson-Shannon. В работе метода не использовались практические ограничения на количество признаков для классификации при условии, что данные признаки выбраны системой в качестве значимых для классифицируемых документов. Мы сравнили полученные результаты с результатами предыдущего цикла РОМИП, где подобное ограничение использовалось. Сравнение результатов подтвердило эффективность выбранного способа классификации.

Литература

[1] Антонов А.В. Методы классификации и технология Галактика-Zoom // сб. Международный форум по информации, Москва, ВИНИТИ, 2003. т.28.

[2] Антонов А, Курзинер Е. Автоматическое выделение предметной области большого необработанного текстового массива // Компьютерная лингвистика и интеллектуальные технологии, Труды Международного семинара Диалог-2002.

[3] Антонов А. Информационно-поисковая система Galaktika-ZOOM с элементами анализа на гипермассивах информации // Сб. ВИНИТИ №8, 2001.

[4] Антонов А., Мешков В. Современные проблемы поисковых систем и некоторые пути их преодоления // Сер. «Аналитика-Капитал», Москва, 2000.

[5] Антонов А. В., Баглей С.Г., Мешков В. С., Суханов А.В. Кластеризация документов с использованием метаинформации // Труды международной конференции Диалог'2006.
[6] Кириченко К.М, Герасимов М.Б. Обзор методов кластеризации текстовых документов // Материалы международной конференции Диалог'2001.
[7] 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.
[8] Joachims T. Learning to Classify Text using Support Vector Machines // Kluwer Academic Publishers, 2002.

Galaktika-Zoom at ROMIP'2009

Alexander Antonov, Stanislav Baglei,
Valentin Meshkov, Vitalyi Stoyan

This paper introduces test results of a new divergence modification applied to the document classification algorithm developed in Galaktika-Zoom search and analytical system. We obtained classification results using the described method based on three ROMIP tracks processing: "Websites Classification", "Webpages Classification", and "Legal Documents Classification". The results are presented and evaluated in the paper.