Галактика-Zoom на РОМИП'2008

Антонов А.В. Баглей С.Г.

Мешков В.С. Стоян В.А.

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 следующих дорожках РОМИП: "Тематическая классификация нормативно-правовых документов", "Тематическая классификация Веб-страниц", "Тематическая классификация Веб-сайтов". Приведено сравнение полученных результатов с предыдущими, показанными системой.

1. Введение

Участие в семинаре РОМИП'2008 стало для нас очередным этапом развития. При обработке заданий РОМИП в этом году мы смогли оценить эффективность новой для нас меры близости, используемой при классификации документов, а также обнаружить и исправить ошибку, связанную с полнотой классификации. Все это позволило нам существенно 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 Представленние 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. Используя характеристики элементов сформированного ИнфоПортрета и собственной статистики документа, производится формирование информационного портрета отдельных документов. То есть, для каждого документа система формирует упорядоченный список слов и словосочетаний, статистически отличающих данный документ от прочих в выборке. ИнфоПортрет, связанный с документом, рассматривается как образ 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 (dot). В расчетах использовались все сформированные элементы ИнфоПортрета в отличие от опыта нашего предыдущего опыта обработки заданий, в котором, в основном, рассматривалась некоторая заданная часть ИнфоПортрета, выбранная в соответствие с полученным

ранжированием его элементов по отношению к классифицируемой рубрике. Тем самым, мы стремились максимально расширить пространство признаков для классификации.

3. Результаты классификации по отдельным дорожкам

Нами была использована экспериментальная модель системы. После выполнения заданий РОМИП в алгоритме метода мы обнаружили ошибку, связанную с отбором документов в классифицируемом массиве, которая сказалась на параметре полноты классификации в результатах по всем трем дорожкам, в которых мы приняли участие.
В результатах обработки заданий можно заметить, насколько уступает размер неоцененного пула документов для нашей системы и для других систем. Такой “несформированный” пул документов для оценки является одним из проявлений ошибки.
Мы исправили ошибку полноты классификации, но, к сожалению, на новую обработку заданий и оценку их результатов времени после исправления уже не оставалось.

Далее приведены оценки результатов классификации, полученные нашей системой.

3.1 Классификация Веб-сайтов

Для классификации Веб-сайтов были выбраны следующие модификации метода:

- метод SVM с линейным ядром. Для классификации используются все элементы полученного ИнфоПортрета. В качестве коэффициента близости используется мера Jenson-Shannon (на графике: svm_js);
- метод SVM с линейным ядром. Для классификации используются все элементы полученного ИнфоПортрета. В качестве коэффициента близости используется мера Kullback-Leibler (на графике: svm_kl);
- метод, основанный на построении матрицы близости Инфопортретов. Данный метод уже был исследован нами при обработке заданий РОМИП на предыдущих семинарах и использован в данном случае для сравнения с новым коэффициентом близости, ранее нами не используемым (на графике: ip_simil).

Рис. 1. Оценки качества классификации по дорожке
“Классификация Веб-сайтов”

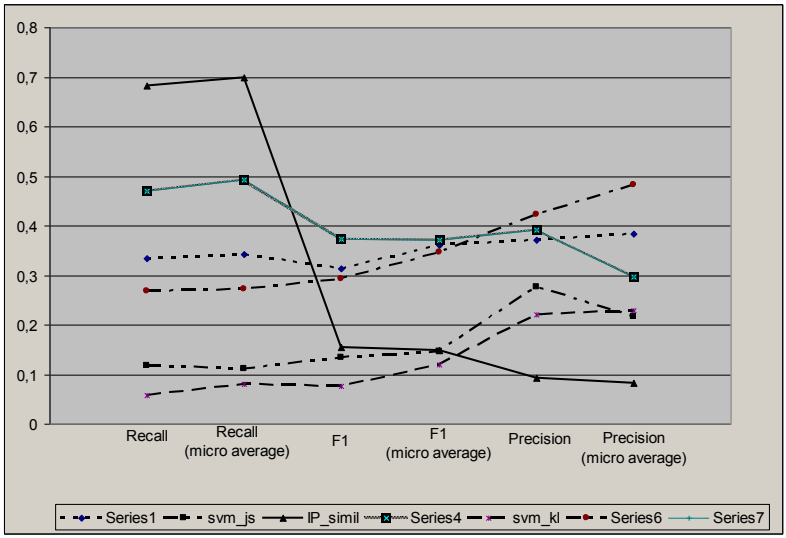

Таблица 1. Оценки качества классификации, присвоенные прогонам системы “Галактика-Zoom” по дорожке “Классификация Веб-сайтов” с использованием метода макроусреднения

	Полнота	F1	Точность
SVM dot JS “сильная” оценка	0,1172	0,1333	0,2766
SVM dot KL “сильная” оценка	0,0578	0,0765	0,2212
IP Similarity “сильная” оценка	0,6829	0,1543	0,093
SVM dot JS “слабая” оценка	0.0755	0.1093	0.3377
SVM dot KL “слабая” оценка	0.0481	0.079	0.3318
IP Similarity “слабая” оценка	0.7223	0.3085	0.2081

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

	Полнота	F1	Точность
SVM dot JS “сильная” оценка	0,1172	0,1461	0,2156
SVM dot KL “сильная” оценка	0,0804	0,1189	0,2285
IP Similarity “сильная” оценка	0,6984	0,149	0,0834
SVM dot JS “слабая” оценка	0.0803	0.1321	0.3725
SVM dot KL “слабая” оценка	0.0591	0.1031	0.4
IP Similarity “слабая” оценка	0.7293	0.3225	0.207

На наш взгляд, использованный коэффициент близости Jensen-Shannon показал лучшие результаты по точности классификации и F-мере, чем ранее использованный нами коэффициент Kullback-Leibler. При этом для обоих методов с использованием SVM характерно проявление ошибки, связанной с полнотой классификации, о которой упомянуто в п.3. Можно заметить, что метод, использующий для классификации близость ИнфоПортретов, показал неплохую оценку полноты классификации и, вместе с тем, оказался хуже обоих методов, использующих SVM, по точности классификации.

3.2 Классификация Веб-страниц

Для проведения классификации Веб-страниц была выбрана модификация метода SVM с линейным ядром, в которой рассматривается весь ИнфоПортрет и мерой близости принимается мера Jenson-Shannon (svm_js).

Для данной дорожки ошибка, связанная с полнотой классификации, проявила себя в меньшей степени, чем в других дорожках. За счет этого F-мера оказалась несколько выше, чем для других дорожек.

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 мера близости Jensen-Shannon (svm_js).

Рис. 2. Оценки качества классификации по дорожке
“Классификация Веб-страниц”

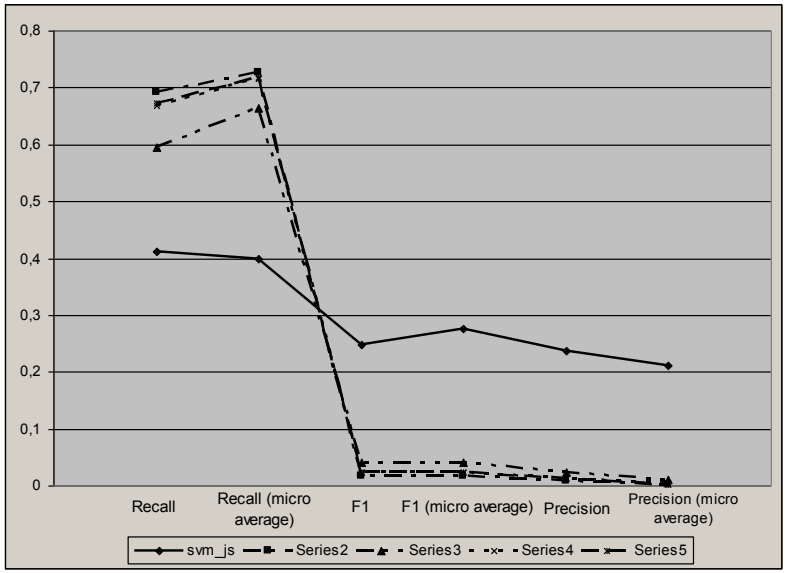

Таблица 3. Оценки качества классификации, присвоенные прогонам системы “Галактика-Zoom” по дорожке “Классификация Веб-страниц” через метод макроусреднения

	Полнота	F1	Точность
SVM dot JS “сильная” оценка	0,4122	0,247	0,2375
SVM dot JS “слабая” оценка	0.3748	0.3424	0.4443

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	0,3986	0,2753	0,2102
SVM dot JS “слабая” оценка	0.3684	0.3617	0.3552

Рис. 3. Оценки качества классификации по дорожке
“Классификация нормативно-правовых документов”

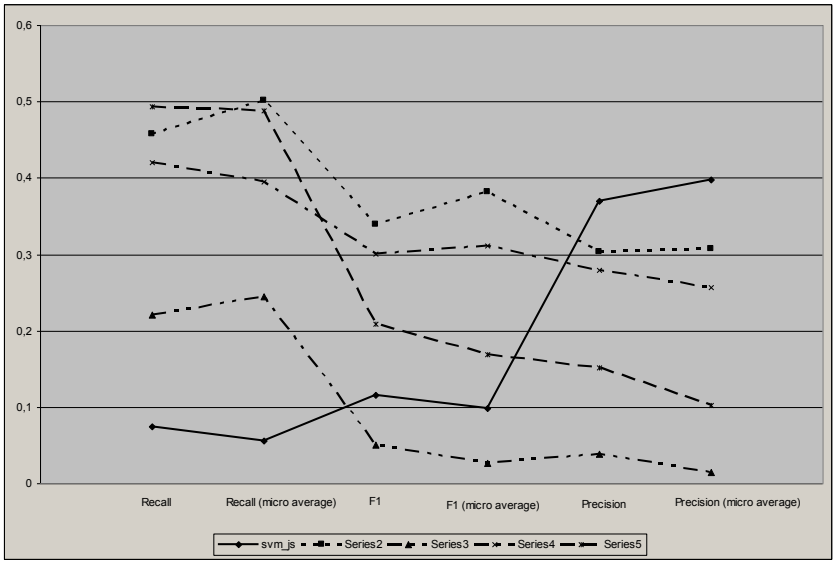

Таблица 5. Оценки качества классификации метода,
использованного системой “Галактика-Zoom” по дорожке
“Классификация нормативно-правовых документов”

	Полнота	F1	Точность
SVM dot JS макроусреднение	0.1012	0.1439	0.3751
SVM dot JS микроусреднение	0.0684	0.1167	0.3983

При обработке заданий по данной дорожке метод показал лучший результат по точности классификации. Вместе с тем, и в данном случае, проявилась ошибка, связанная с полнотой классификации.

4. Заключение

Нам удалось провести исследование эффективности новой для нас меры близости ИнфоПортретов, использованной для классификации документов с помощью метода опорных векторов в применении к системе “Галактика-Zoom”. Было проведено сравнение двух различных мер. Сравнение полученных результатов позволяет считать, что

вновь примененная мера позволяет получить более качественный результат при решении задачи классификации документов.

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 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, 1998.

[8] Joachims T. Learning to Classify Text using Support Vector Machines // Kluwer Academic Publishers, 2002.

Galaktika-Zoom at ROMIP'2008

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.