Галактика Zoom на РОМИП’2006

© Антонов А.В., Баглей С.Г., Мешков В.С.

Корпорация “Галактика”
{alexa, baglei, meshkov}@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

Система “Галактика-Zoom” принимает участие в семинаре РОМИП уже четвертый раз. От семинара к семинару нами накапливался необходимый опыт, позволяющий не только регулярно сравнивать полученные результаты с достижениями других участников семинара, но и оценить качество развития нашей системы от этапа к этапу.

На РОМИП’2006 “Галактика-Zoom” расширила свое участие – кроме проведения традиционной тематической классификации мы приняли участие в дорожке поиска похожих документов по документу-образцу. Это стало возможным благодаря использованию новых для нашей системы методов и подходов, примененных к задачам классификации и поиска документов.

Далее описываются особенности использования подходов в системе “Галактика-Zoom” с точки зрения сочетания преимуществ самой системы с возможностями, которые предоставляются при использовании методов.

Кроме того, приводится сравнение полученных результатов с нашими результатами, показанными на РОМИП’2005.

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].

Мы провели эксперименты с несколькими методами и их сочетаниями, основанными на работе с объектной моделью представления документов и рубрик. Далее следует описание этих методов.

2.3 Метод опорных векторов (Support Vector Mashines)

В качестве основы для проведения классификации с помощью метода опорных векторов [9] была взята его реализация SVMLight [8]. Как известно, работа метода опорных векторов включает в себя два основных этапа – обучение на тренировочных данных и предсказание, то есть, разбиение всего массива на классы.

На этапе обучения алгоритма на основе множества тренировочных документов строился ИнфоПортрет некоторой рубрики. Далее, используя пространство полученного ИнфоПортрета, формировалось представление всех остальных документов, составляющих тре-

нировочный 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 были выбраны следующие разновидности метода:

1. Метод SVM с линейным ядром (dot).
   В расчетах использовались 100 элементов ИнфоПортрета с наибольшим весом для рубрики.
2. Метод SVM с ядром, основанным на радиальных базисных функциях (RBF, radial based functions).
   В расчетах использовались 100 элементов ИнфоПортрета с наибольшим весом для рубрики.
3. Метод SVM с линейным ядром (dot) в режиме регрессии.
4. Метод SVM с линейным ядром (dot) в сочетании с алгоритмом латентного семантического анализа (LSA, Latent Semantic Analysis).
   В расчетах использовались 100 элементов ИнфоПортрета с наибольшим весом для рубрики.

В качестве дополнительного фильтра для отбора элементов ИнфоПортрета, соответствующих отдельной рубрике, выбран известный алгоритм кластеризации LSA/LSI, использующий принципы факторного анализа для выявления латентной структуры объектов. Задачей факторного анализа является выделение главных факторов из пространства элементарных. Выбор данного алгоритма обусловлен рядом причин. Во-первых, LSA не нуждается в обучении. То есть, при кластеризации формируется такая структура кластеров, которая зависит исключительно от обрабатываемых данных. Кроме того, не требуется проведения этапа предварительной настройки алгоритма. Во-вторых, из опыта предыдущих работ [7], метод LSA признается лучшим для выявления латентных зависимостей в структуре объектов.

С помощью метода LSA на основе ИнфоПортрета рубрики формируется множество кластеров документов на основе всех документов массива. Ограниченное подмножество наиболее “близких” к ИнфоПортрету рубрики кластеров объединяются, после чего документы, их составляющие, рассматриваются в качестве множества классифицируемых документов для метода SVM. Таким образом, идея данного подхода заключается в попытке ограничения всего множе-

ства классифицируемых документов для улучшения качества работы алгоритма SVM.

5. Метод SVM с линейным ядром (dot) в сочетании с алгоритмом латентного семантического анализа. В расчетах использовался весь сформированный ИнфоПортрет.

3. Результаты классификации по отдельным дорожкам

Нами была использована экспериментальная модель системы. После выполнения заданий РОМИП в алгоритме метода мы обнаружили ошибку, которая проявила себя при классификации по некоторым рубрикам. В дальнейшем ошибка была найдена и исправлена, однако, в данной работе приведены результаты обработки до внесения исправлений, что не позволило наглядно продемонстрировать улучшение качества рубрикации для отдельных прогонов.

Несмотря на этот факт, большинство из обработанных системой заданий было успешно выполнено.

	Полнота	F1	Точность
SVM dot 100 IP “сильная” оценка	0,8865	0,3880	0,3143
SVM dot 100 IP “слабая” оценка	0,5108	0,4861	0,6306
SVM RBF “сильная” оценка	0,3552	0,3155	0,4467
SVM RBF “слабая” оценка	0,1250	0,1299	0,4137
ГЗ’2005 “сильная” оценка	0,4914	0,1181	0,0671
ГЗ’2005 “слабая” оценка	0,4695	0,2782	0,1976

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

3.1 Классификация Веб-сайтов

Для классификации Веб-сайтов были выбраны следующие модификации метода:

- SVM с линейным ядром, рассматриваются верхние 100 элементов ИнфоПортрета (SVM dot 100 IP);
- SVM с ядром RBF, рассматривается весь полученный ИнфоПортрет (SVM RBF).

Несколько лучший результат для данной дорожки показала модификация SVM с линейным ядром. Можно отметить, что большая разница между модификациями метода наблюдалась по полноте классификации.

При этом достигнуто существенное улучшение параметров классификации по сравнению с результатами, показанными системой "Галактика-Zoom" на РОМИП-2005.

	Полнота	F1	Точность
SVM dot 100 IP "сильная" оценка	0,7528	0,4213	0,2925
SVM dot 100 IP "слабая" оценка	0,3931	0,4901	0,6506
SVM RBF "сильная" оценка	0,3146	0,2901	0,2692
SVM RBF "слабая" оценка	0,1319	0,2070	0,4807
ГЗ'2005 "сильная" оценка	0,4638	0,1192	0,0748
ГЗ'2005 "слабая" оценка	0,4805	0,2584	0,1999

Таблица 2. Оценки качества классификации, присвоенные прогонам системы "Галактика-Zoom" по дорожке "Классификация Веб-сайтов " методом микроусреднения.

3.2 Классификация Веб-страниц

Для проведения классификации Веб-страниц были выбраны следующие модификации метода SVM:

- SVM с линейным ядром, рассматривается весь ИнфоПортрет (SVM dot);
- SVM с ядром RBF, рассматривается весь ИнфоПортрет (SVM RBF).

	Полнота	F1	Точность
SVM dot “сильная” оценка	0,5968	0,2422	0,2223
SVM dot “слабая” оценка	0,4544	0,3663	0,4130
SVM RBF “сильная” оценка	0,8804	0,2513	0,1889
SVM RBF “слабая” оценка	0,6603	0,3627	0,2976
ГЗ’2005 “сильная” оценка	0,2357	0,0847	0,0516
ГЗ’2005 “слабая” оценка	0,2160	0,1336	0,0967

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

	Полнота	F1	Точность
SVM dot “сильная” оценка	0,6153	0,2282	0,1400
SVM dot “слабая” оценка	0,5317	0,3829	0,2991
SVM RBF “сильная” оценка	0,9018	0,1751	0,0970
SVM RBF “слабая” оценка	0,7556	0,3173	0,2008
ГЗ’2005 “сильная” оценка	0,2666	0,1134	0,2143
ГЗ’2005 “слабая” оценка	0,2388	0,1500	0,3032

Таблица 4. Оценки качества классификации, присвоенные прогонам системы “Галактика-Zoom” по дорожке “Классификация Веб-сайтов ” методом микроусреднения.

Для данной дорожки методы показали сравнимые усредненные результаты. При этом,можно отметить, что SVM с линейным ядром был лучше по точности классификации, тогда как SVM с ядром RBF показал лучшие показатели полноты.

Примененные методы намного улучшили качество классификации по данной дорожке по сравнению с результатами системы “Галактика-Zoom”, показанными на РОМИП’2005.

3.3 Классификация нормативно-правовых документов

Для классификации были использованы следующие модификации метода:

- SVM с линейным ядром, рассматривается весь ИнфоПортрет (SVM dot);
- SVM с ядром RBF, рассматривается весь ИнфоПортрет (SVM RBF).
- SVM с линейным ядром, с учетом всего ИнфоПортрета в режиме регрессии (SVM regr);
- комбинация алгоритмов LSA и SVM с линейным ядром, с учетом 100 верхних элементов ИнфоПортрета (LSA/SVM 100 IP);
- комбинация алгоритмов LSA и SVM с линейным ядром, с учетом всего ИнфоПортрета (LSA/SVM full IP).

	Полнота	F1	Точность
SVM dot	0,0478	0,0722	0,3453
SVM RBF	0,0660	0,0968	0,3132
SVM regr	0,0794	0,0087	0,3081
LSA/SVM 100 IP	0,1712	0,1431	0,2043
LSA/SVM full IP	0,0151	0,0266	0,2477
ГЗ'2005	0,0682	0,1126	0,3229

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

	Полнота	F1	Точность
SVM dot	0,0406	0,0731	0,3670
SVM RBF	0,0682	0,1101	0,2855
SVM regr	0,0207	0,0097	0,0063
LSA/SVM 100 IP	0,1468	0,1793	0,2303
LSA/SVM full IP	0,0125	0,0239	0,2412
ГЗ'2005	0,1777	0,1618	0,3762

Таблица 6. Оценки качества классификации нормативно-правовых документов в системе "Галактика-Zoom". Метод микроусреднения.

При обработке заданий по данной дорожке ошибка в алгоритме, упомянутая в п.3, сказалась наибольшим образом. Предположительно, вследствие более однородного распределения характерных слов и словосочетани в массиве нормативных документов, чем в Вебколлекции. В данных условиях лучшим оказался алгоритм, использующий комбинацию алгоритмов LSA и SVM с линейным ядром, учитывающий всвоей работе 100 верхних элементов сформированного ИнфоПортрета. Данный метод показал некоторое улучшение по сравнению с результатами, достигнутыми системой на РОМИП'2005.

4. Заключение

Нам удалось провести исследование эффективности нового подхода к классификации документов в применении к системе "Галактика-Zoom". Была проведена оценка разновидностей метода SVM и некоторых его комбинаций с методом LSA. Сравнение полученных результатов позволяет считать, что примененные методы являются эффективными для решения задачи классификации документов.

Благодаря использованию подхода мы смогли расширить свое участие на РОМИП'2006, приняв участие, помимо дорожек тематической классификации, в дорожке поиска документов по документу-образцу.

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] Deerwester, S., Dumais, S., Landauer, T., Furnas, G. and Harshman, R. Indexing by latent semantic analysis // Journal of the Society for Information Science, 1990, vol. 41(6), 391-407.
[8] 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.
[9] Joachims T. Learning to Classify Text using Support Vector Machines // Kluwer Academic Publishers, 2002.

Galaktika-Zoom at ROMIP'2006

© Alexander V. Antonov, Stanislav G. Baglei, Valentin S. Meshkov
{alexa, baglei, meshkov}@galaktika.ru

This paper introduces a new modification to 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.