Галактика Zoom на РОМИП 2005: оценка модификации метода формирования инфопортрета

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

Корпорация “Галактика”
{alexa, baglei, meshkov}@galaktika.ru

Аннотация

В статье приводится описание новой модификации, реализованной в алгоритме классификации документов системы “Галактика-Зум”. Представлены результаты выполнения заданий по дорожкам тематической классификации: “Классификация Веб-страниц”, “Классификация Веб-сайтов”, “Классификация нормативно-правовых документов”, а также проведено сравнение полученных результатов.

1. Введение

Поисково-аналитическая система обработки больших объемов неструктурированных данных Галактика-Зум уже имеет опыт участия в семинаре РОМИП. Участвуя в предыдущих семинарах, мы представляли достаточно подробную информацию о системе Галактика-Зум, с которой можно ознакомиться в работах [1, 2], где описывается архитектура, принципы работы, характеристики системы. В работе [3] содержится подробная информация о примененных нами подходах к обработке массива данных РОМИП и достигнутых результатах на семинаре РОМИП’2004.

Галактика-Зум расширила свое участие на РОМИП’2005. Кроме тех дорожек, в которых ранее мы уже приобрели определенный опыт, а именно, в “Классификации нормативно-правовых документов” и “Классификации веб-сайтов”, в этом году мы приняли участие в дорожке “Классификация веб-страниц”.

Целью данной статьи является представить разработанную нами новую модификацию метода формирования инфопортрета для проведения классификации документов. Кроме того, представлены и проанализированы результаты, полученные системой “Галактика-Зум” на РОМИП 2005.

2. Анализ предыдущего опыта и развитие алгоритма классификации

2.1 Основные принципы классификации в системе Галактика-Зум и опыт предыдущего участия в семинаре РОМИП.

Как известно, основным понятием в системе Галактика-Зум является понятие Информационного портрета выборки документов (Инфопортрет, ИП). Инфопортрет представляет собой список языковых инвариантов (слов и словосочетаний), отличающих данную выборку от прочих.

При классификации в системе “Галактика-Зум”, в том числе и при обработке заданий РОМИП, принадлежность классифицируемой единицы к некоторому классу определяется с помощью вычисления меры близости инфопортрета этой единицы к инфопортрету рубрики (класса). При обработке данных по дорожке “Классификация Веб-сайтов” такой единицей выступал веб-сайт, по дорожке “Классификация нормативных документов” – нормативный документ. Данный подход в целом оправдал себя: при его использовании задача классификации была решена. Однако, при классификации веб-сайтов, метод, основанный на определении непосредственной близости инфопрортрета сайта к инфопортрету рубрики показал определенную уязвимость. Причина состояла в том, что многим веб-сайтам, свойственна перегруженность вспомогательной или иной информацией, не имеющей прямого отношения к содержанию сайта. При формировании инфопортрета такого сайта его инфопортрет также включал посторонние по отношению к теме документа элементы. Очевидно, что из-за подобных включений качество классификации ухудшалось

При классификации нормативно-правовых документов данная проблема не возникала. Как правило, такие документы монотематичны и приведены к определенному формату, то есть, не включают в себя навигационную и рекламную информацию. Поэтому описанный недостаток не повлиял на правильность формирования инфопортретов документов данного вида.

2.2 Новая модификация метода формирования инфопортрета

Очевидно, что для решения проблемы, описанной в п. 2.1 требовалась определенная “фильтрация” инфопортрета сайта, позволяющая исключить нехарактерные для рубрики элементы инфопортрета. Чтобы провести фильтрацию, мы вводим понятие “мультидокумента”, то есть, документа, в который объединяются все документы, составляющие отдельный сайт.

Таким образом, при классификации, работа системы состоит из двух этапов. На первом этапе формируются инфопортреты классифицируемых сайтов, то есть инфопрортреты выборок документов базы, отобранных по принципу принадлежности к общему домену.

Второй этап – это “фильтрация” инфопортрета. Сначала формируются мультидокументы, каждый из которых включает в себя все страницы соответствующего ему сайта. На основе выборки полученных мультидокументов, представляющей собой все сайты базы, строится инфопортрет.

Далее полученный на первом этапе инфопортрет сайта “фильтруется” с помощью инфопортрета всех мультидокументов. Элементы инфопортрета сайта, не имеющие достаточного веса в инфопортрете всех мультидокументов, исключаются из верхушки инфопортрета сайта. То есть, этап фильтрации позволяет выявить и удалить из верхушки инфопортрета сайта те элементы, которые нехарактерны для выборки, содержащей все сайты базы.

Вышеописанная технология оказала влияние не только на качество классификации по дорожке “Классификация Веб-сайтов”. Благодаря разработанной системе фильтрации инфопортрета сайта для нас стало возможным проводить классификацию отдельных веб-страниц. Ранее их классификация априори была бы неуспешной ввиду перегруженности веб-страниц второстепенной информацией и влияния этой информации на инфопортрет страниц.

2.3 Критерии близости инфопрортретов

Далее представлены результаты работы системы, состоящие из двух прогонов, условно обозначенных ГЗ 1 и ГЗ 2. Инфопортреты документов в каждом из прогонов формировались в соответствие с описанным в п. 2.2 методом. Различие между прогонами ГЗ 1 и ГЗ 2 состоит в использовании различных критериев близости инфопортретов. Критериям близости инфопортретов соответствуют собственный набор коэффициентов, с помощью которых рассчитывается мера близости.

2.4 Массив данных.

Тренировочными данными для классификации на РОМИП'2005 служила русскоязычная часть каталога dmoz.org [4], применявшаяся для РОМИП'2004. Поэтому, у нас появилась возможность сравнить результаты работы новой модификации метода классификации с подходом, применявшимся ранее при обработке заданий по дорожкам, к которым такая модификация применима, то есть, "Классификация Веб-сайтов" и "Классификация веб-страниц".

3. Участие в РОМИП 2005

В разделах, посвященных дорожкам РОМИП'2005, в которых мы приняли участие, описываются промежуточные результаты, то есть сформированные инфопортреты рубрик, с помощью которых проводилась классификация. Кроме того, приводятся оценки качества классификации, полученные после оценки результатов асессорами.

3.1 Классификация веб-сайтов

На предыдущем семинаре РОМИП мы демонстрировали сформированный инфопортрет для рубрики "Компьютеры / Программирование". Сравним его с инфопортретом той же самой рубрики, полученным с помощью модифицированного алгоритма, использовавшимся при обработке заданий РОПИП'2005.

Таблица 1. Инфопортрет рубрики "Компьютеры / Программирование", сформированный с помощью ранее использовавшегося метода и модифицированного алгоритма.

РОМИП 2004	**РОМИП 2005**
ФУНКЦИЯ	ФАЙЛ
ФАЙЛ	ФУНКЦИЯ
БАРХАТНЫЙ ПУТЬ	ЗНАЧЕНИЕ
ЗНАЧЕНИЕ	УКАЗАТЕЛЬ
СТРОКА	СТРОКА
ОКНО	ОКНО
ПРИЛОЖЕНИЕ	ПАРАМЕТР
ХОСТИНГ	ВОЗВРАЩАТЬ
УКАЗАТЕЛЬ	ПЕРЕМЕННЫЙ
ПАРАМЕТР	ИДЕНТИФИКАТОР
ВОЗВРАЩАТЬ	ПРИЛОЖЕНИЕ
ОПЕРАЦИОННАЯ	ЭТА ФУНКЦИЯ

СИСТЕМА	
ИДЕНТИФИКАТОР	ОБЪЕКТ
ФАЙЛОВАЯ СИСТЕМА	ОПЕРАЦИОННАЯ СИСТЕМА
ПЕРЕМЕННЫЙ	ПОЛЬЗОВАТЕЛЬ
КЛАСС	СИМВОЛ
СИМВОЛ	БУФЕР
СТАНДАРТНАЯ ФУНКЦИЯ	ГЛАВНОЕ ОКНО
БЕСПЛАТНАЯ УСТАНОВКА	КЛАСС
ОБЪЕКТ	СООБЩЕНИЕ

В таблице 1 элемент инфопортрета “БАРХАТНЫЙ ПУТЬ” не относится непосредственно к содержанию рубрики, и представляет собой навигационную информацию на веб-сайте. Модифицированный алгоритм формирования инфопортрета, описанный в п. 2.2 позволил “отфильтровать” данный элемент. Рассматривая более расширенную верхушку инфопортрета, мы также наблюдали улучшение в смысле соответствия его элементов рубрике.

Далее, в таблице 2 приводятся оценки по дорожке классификации веб-сайтов. Оценки относятся к двум выполненным прогонам ГЗ 1 и ГЗ 2. Отличия между данными прогонами были описаны в п. 2.3.

Таблица 2. Оценки качества классификации, присвоенные прогонам системы “Галактика-Зум” по дорожке “Классификация веб-сайтов”.

	“Слабая”оценка		“Сильная” оценка	
	ГЗ 1	ГЗ 2	ГЗ 1	ГЗ 2
Точность	0,3873	0,1999	0,2210	0,0748
Точность (макроусреднение)	0,3842	0,1976	0,1896	0,0671
F1	0,2116	0,2584	0,2027	0,1192
F1 (макроусреднение)	0,2383	0,2782	0,2203	0,1181
Полнота	0,1750	0,4805	0,2437	0,4638
Полнота (макроусреднение)	0,1727	0,4695	0,2627	0,4914

В целом, первый из прогонов показал более высокую точность классификации, второй – более высокую полноту. При этом для “слабой” оценки наблюдались лучшие показатели у первого прогона, для “сильной” оценки – у второго.

3.2 Классификация веб-страниц

В таблице 3 в качестве примера сравниваются инфопортреты рубрики “Общество / Религия и духовность”, сформированные старым методом, применявшимся на РОМИП’2004, и новой модификацией метода, использовавшейся в этом году.

Таблица 3. Инфопортрет рубрики “Общество / Религия и духовность”, сформированный ранее использовавшимся методом и его новой модификацией.

РОМИП 2004	**РОМИП 2005**
КН	ЦЕРКОВЬ
СМИРНОВ	БОГ
ПОПОВ	ХРИСТОС
ИВАН	БОЖИЙ
ПОП	ГОСПОДЬ
ИВАНОВ	ИИСУС
БОГ	СЛУЖЕНИЕ
ИВАНОВИЧ	СВЯТОЙ ДУХ
СОЛОВЬЕВ	БИБЛИЯ
ВАСИЛЬЕВИЧ	СВЯТОЙ
НИКОЛАЕВИЧ	ГРЕХ
АЛЕКСАНДРОВИЧ	МОЛИТВА
ЦЕРКОВЬ	ОТЕЦ
АЛЕКСЕЕВИЧ	ПРАВОСЛАВНЫЙ
МИХАЙЛОВИЧ	СЛУЖИТЕЛЬ
БОЖИЙ	ПИСАНИЕ
ХРИСТОС	ПАСТОР
НИКОЛАЙ	ПАВЕЛ
ПЕТРОВИЧ	МОЛИТЬСЯ
ИИСУС	ДУХОВНЫЙ

Очевидно, что использование новой модификации метода формирования инфопортрета способствует формированию более соответствующих для данной рубрики элементов в его “верхушке”. Таким образом, инфопортрет более точно отражает информацию, содержащуюся на страницах, релевантных данной рубрике. Это положительно сказывается на качестве классификации веб-страниц.

Далее приведены оценки, полученные по результатам прогонов системы “Галактика-Зум” по дорожке “Классификация Веб-страниц” для “слабого” случая. Оценивание производилось двумя способами. Первый способ заключался в отнесении неоцененных

асессорами страниц к нерелевантным страницам (ALL). При оценивании вторым способом неоцениваемые страницы не учитывались (JUDGED).

Таблица 4. “Слабые” оценки качества классификации, присвоенные прогонам системы “Галактика-Зум” по дорожке “Классификация Веб-страниц”.

	ALL		JUDGED	
	ГЗ 1	ГЗ 2	ГЗ 1	ГЗ 2
Точность	0,1283	0,3032	0,4976	0,5988
Точность (макроусреднение)	0,0527	0,0967	0,4125	0,5695
F1	0,1275	0,1500	0,3006	0,2853
F1 (макроусреднение)	0,0876	0,1336	0,3178	0,3132
Полнота	0,2923	0,2388	0,2923	0,2388
Полнота (макроусреднение)	0,2584	0,2160	0,2584	0,2160

Таблица 5. “Сильные” оценки качества классификации, присвоенные прогонам системы “Галактика-Зум” по дорожке “Классификация Веб-страниц”.

	ALL		JUDGED	
	ГЗ 1	ГЗ 2	ГЗ 1	ГЗ 2
Точность	0,0753	0,2143	0,2877	0,3805
Точность (макроусреднение)	0,0283	0,0516	0,2217	0,3044
F1	0,0900	0,1134	0,2333	0,2345
F1 (макроусреднение)	0,0515	0,0847	0,2488	0,2656
Полнота	0,3010	0,2666	0,3010	0,2666
Полнота (макроусреднение)	0,2836	0,2357	0,2836	0,2357

При классификации веб-страниц второй прогон показал лучшую точность классификации и несколько худшую полноту. При этом, при использовании второго критерия близости инфопортретов усредненные показатели классификации были выше.

3.3 Классификация нормативно-правовых документов

Новая модификация метода формирования инфопортрета не использовалась при классификации по данной дорожке, так как структура нормативно-правовых документов не позволяет его применять. В таблице 6 приведены полученные оценки для проведенных прогонов.

Таблица 6. Оценки качества классификации, присвоенные прогонам системы “Галактика-Зум” по дорожке “Классификация нормативно-правовых документов”.

	ГЗ 1	ГЗ 2
Точность	0,3529	0,3762
Точность (макроусреднение)	0,2393	0,3229
F1	0,1952	0,1618
F1 (макроусреднение)	0,1510	0,1126
Полнота	0,2508	0,1777
Полнота (макроусреднение)	0,1103	0,0682

4. Заключение

Мы разработали новую модификацию метода формирования инфопортрета, использующую фильтрацию его элементов. Благодаря новой модификации мы смогли не только принять участие, но и показать высокие результаты в новой для нас дорожке – “Классификация Веб-страниц”. Использование новой модификации метода при обработке массива данных на РОМИП’2005 полностью оправдало себя. Благодаря разработанному подходу улучшилось качество инфопортрета – ключевого объекта при проведении классификации системой “Галактика-Зум”.

Нам удалось оценить эффективность отдельных свойств поисково-аналитической системы “Галактика-Зум”, таких, как особенности использования различных критериев близости инфопортретов при классификации.

Опыт участия в семинаре РОМИП очень полезен для нас. Мы смогли получить независимую оценку качества классификации, сравнить ее с результатами прошедших семинаров и с результатами, показанными другими участниками.

Литература

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

[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.В., Козачук М.В., Мешков В.С. «Галактика-Зум: Отчет об участии в семинаре РОМИП 2004», Труды второго российского семинара по оценке методов информационного поиска. Под ред. И.С. Некрестьянова - Санкт-Петербург: НИИ Химии СПбГУ, 2004, 214 с.

[4] Каталог DMOZ. http://www.dmoz.org

[5] Сайт Галактика-Зум. http://zoom.galaktika.ru

Galaktika-Zoom at ROMIP 2005: Evaluation of a New Method of Concept Filtering

Alexander V. Antonov, Stanislav G. Baglei, Valentin S. Meshkov

This paper introduces a new method to improve text classification task developed in Galaktika-Zoom search and analytical system. We obtained classification results using the described method on the three ROMIP tracks: “Websites Classification”, “Webpages Classification”, and “Legal Documents Classification”. We present and analyze results of assessment conducted in course of ROMIP’2005 seminar.