Галактика-Зум: Отчет об участии в семинаре РОМИП 2004

© Антонов А.В, Козачук М.В., Мешков В.С.

Корпорация «Галактика»
{alexa, kozachuk, meshkov}@galaktika.ru kozachuk

Аннотация

В статье описывается опыт участия системы Галактика-Зум в семинаре РОМИП 2004 (задания по классификации веб-коллекции и тематической классификации коллекции законодательных документов). Приводятся краткие сведения о системе, описания экспериментов, полученные результаты.

1. Информация о системе

Поисково-аналитическая система Галактика-Зум (ГЗ) предназначена для аналитической обработки больших объемов неструктурированной текстовой информации. Программный комплекс сочетает в себе возможности классической поисковой системы, системы сбора текстовых данных (text mining) и системы аналитической обработки информации. Разработка системы ведется в Отделе поисковых систем Корпорации «Галактика» начиная с 1999 года. Более подробная информация о системе представлена на сайте [7].

Базовым 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. Инфопортрет строится на основе статистических методов обработки текстовой информации, при этом применяются модифицированные байесовские методы. Инфопортрет используется для решения задач быстрого представления результатов запроса без чтения текстов документов, ускоренной

навигации внутри полученной выборки документов, автоклассификации и автореферирования.

2. Участие в РОМИП

Обычный режим эксплуатации системы Галактика-Зум подразумевает непрерывный диалог с пользователем, поэтому возможности ее использования для решения большинства задач, предлагаемых в рамках РОМИП, особенно в дорожках, посвященных поиску, были существенно ограничены. Наименьших затрат на адаптацию системы требовали дорожки, посвященные автоматической классификации. Функция автоматической классификации появилась в системе ГЗ сравнительно недавно, около года назад [1][2], и нам было важно получить независимые оценки качества ее работы. Исходя из этих соображений, мы принимали участие в двух дорожках РОМИП: классификации коллекции веб-документов и тематической классификации документов (на основе коллекции нормативных документов).

2.1 Классификация коллекции веб-документов

Исходными данными для дорожки классификации веб-коллекции служило множество сайтов домена narod.ru. В качестве классификатора первоначально предполагалось использовать рубрики из каталога narod.ru [4], т.е. повторялось задание РОМИП 2003. Именно на этих исходных данных проводились первые эксперименты по классификации.

Качество тренировочного массива было не слишком высоким. Вероятно, это вызвано тем, что каталог narod.ru не модерируется и заполняется самими создателями сайтов. Вследствие этого содержимое многих сайтов тренировочного массива с большой натяжкой можно было отнести к заявленной их создателями рубрике. Поэтому организаторы РОМИП 2004 заменили тренировочный массив, взяв за основу русскоязычную часть каталога dmoz.org [5].

Таблица 1. Характеристики исходного массива данных

Общее количество сайтов	22359
Общее количество веб-страниц	588077
Общее количество рубрик	247

На основе данных исходного и тренировочного массивов была сформирована база данных ГЗ (ROMIP). Для загрузки веб-страниц в

базу использовался стандартный загрузчик HTML документов, входящий в состав системы ГЗ. При этом каждая страница сайта рассматривалась как отдельный документ. Связи между страницами и структура HTML документов не учитывались. В таблице 2 приводится перечень полей, используемых для описания веб-страницы в базе.

Таблица 2. Перечень полей базы ROMIP

Поле	Тип	Описание
Message	Текст	Текст web-страницы
CATEGORY_ID	Текст	Для документов из тренировочного массива dmoz.org поле содержит строку DMOZ. Для остальных документов не заполняется.
Label	Текст	Заголовок страницы
$Fileinvno$	Текст	Для страниц, не относящихся к массиву dmoz.org, поле содержит номер страницы в массиве РОМИП.
$HREF$	Текст	HTTP адрес web-страницы

Для хранения дополнительной информации, связанной с задачей классификации (перечень и наименования рубрик, имена сайтов относящихся к определенной рубрике и т.д.), была создана отдельная реляционная база данных (использовалась СУБД MS SQL Server 2000).

Таблица 3. Характеристики базы данных ГЗ

Размер базы ГЗ (Мб)	8288
Объем словаря слов	8859735
Число словомест	556880329
Объем словаря словосочетаний	3120022
Число мест словосочетаний	40163957

Классификация документов web-коллекции РОМИП осуществлялась по следующему алгоритму. На основе

тренировочного набора документов были построены Информационные портреты каждой классификационной рубрики. В Инфопортреты включали как слова, так и словосочетания. Полученные Информационные портреты были сохранены для последующего использования.

В таблице 4 приводится фрагмент Информационного портрета рубрики Компьютеры/Программирование. Здесь, W – уровень значимости элемента Инфопортрета.

Таблица 4. Фрагмент Информационного портрета рубрики

Значение	W
ФУНКЦИЯ	55
ФАЙЛ	27.1
ЗНАЧЕНИЕ	16.8
СТРОКА	15.7
ОКНО	14.6
ПРИЛОЖЕНИЕ	14
ХОСТИНГ	13
УКАЗАТЕЛЬ	12.4
ПАРАМЕТР	12.1
ВОЗВРАЩАТЬ	11.1
ОПЕРАЦИОННАЯ СИСТЕМА	10
ИДЕНТИФИКАТОР	9.95
ФАЙЛОВАЯ СИСТЕМА	9.45
ПЕРЕМЕННЫЙ	7.98
КЛАСС	7.97
СИМВОЛ	7.09

Для каждого сайта основного массива строился Информационный портрет, который поочередно сравнивался с Информационными портретами всех рубрик. В результате относящимися к определенной рубрике считались сайты, у которых коэффициент близости с рубрикой составлял не менее 0.45. Полученная информация сохранялась в реляционной базе данных для последующей обработки.

Общее время работы процедуры классификации составило около 30 часов. Задача запускалась на компьютере со следующими характеристиками: процессор - Intel Pentium IV 2.6 ГГц, ОЗУ - 3 Гб.

Оценки качества результатов классификации веб-коллекции приводятся в Таблице 5. Методика расчета оценок описывается в документе [6].

Таблица 5. Оценки качества классификации

Оценка	weak	strong
F1 (macro average)	0.511	0.369
Recall	0.579	0.544
Precision (macro average)	0.517	0.285
Error	0.009	0.008
F1	0.49	0.341
Recall (macro average)	0.505	0.526
Accuracy	0.991	0.991
Precision	0.559	0.316

2.2 Тематическая классификация (коллекция нормативных документов)

В качестве исходных данных использовались нормативные документы из базы данных КОДЕКС [1]. Обучающим массивом служила часть классификатора КОДЕКС.

Таблица 7. Характеристики исходного массива данных

Общее количество документов	60294
Общее количество рубрик	183
Количество классифицированных рубрик (содержат хотя бы один документ)	162

На основе данных исходного и тренировочного массивов была сформирована база данных ГЗ (ROMIPN). Для загрузки документов в базу использовался стандартный загрузчик HTML документов, входящий в состав системы ГЗ. Связи между документами исходного массива и структура HTML файлов не учитывались. В таблице 8 приводится перечень полей, используемых для описания документов нормативной коллекции в базе.

Таблица 8. Перечень полей базы ROMIPN

Поле	Тип	Описание
Message	Текст	Текст документа
CATEGORY_ID	Текст	Список номеров категорий, к которым отнесен документ тренировочного массива РОМИП. Для документов классификационного массива поле не заполнялось.
Label	Текст	Заголовок документа
$HREF$	Текст	Номер документа в массиве РОМИП

Для хранения дополнительной информации, связанной с задачей классификации (перечень и наименования рубрик, результаты классификации и т.д.), использовалась реляционная база данных.

Таблица 9. Характеристики базы данных ГЗ

Размер базы ГЗ (Мб)	816
Объем словаря слов	1017360
Число словомест	87073269
Объем словаря словосочетаний	340879
Число мест словосочетаний	10171889

Классификация документов нормативной коллекции РОМИП осуществлялась по алгоритму, сходному с алгоритмом классификации веб-коллекции. На первом этапе на основе тренировочного набора документов были построены Информационные портреты классификационных рубрик. В Инфопортреты включали слова и словосочетания.

В отличие от веб-коллекции, в которой каждый сайт состоял из множества документов (веб-страниц) и представлял собой выборку документов, в случае коллекции нормативных документов каждому документу исходного массива соответствовал один документ в базе данных ГЗ. Поэтому для определения степени близости документа к

Инфопортрету рубрики использовалась процедура ранжирования документов по заданному Инфопортрету. Данная процедура является менее затратной по времени, чем сравнение Инфопортретов. При ранжировании документов выборки по заданному Инфопортрету каждому документу присваивается ранг (число в диапазоне от 0 до 1), характеризующий степень близости документа к Инфопортрету рубрики.

Таким образом, для каждой рубрики производилось ранжирование всех документов исходного массива по Информационному портрету рубрики. При этом документы с рангом менее 0.45 исключались из результатов классификации.

Общее время работы процедуры классификации составило 4 часа. Задача запускалась на компьютере со следующими характеристиками: процессор - Intel Pentium IV 2.6 ГГц, ОЗУ - 3 Гб.

Оценки качества результатов тематической классификации коллекции нормативных документов приводятся в таблице 10. В колонке Ideal содержатся результаты автоматической оценки, полученной путем сравнения результатов классификации с содержимым рубрикатора КОДЕКС. Нужно отметить, что эти оценки существенно отличаются от оценок, полученных «ручным» способом. Точность классификации оказалась заметно выше, зато полнота резко упала.

Таблица 10. Оценки качества классификации

Оценка	weak	strong	Ideal
F1 (macro average)	0.438	0.216	0.174
Recall	0.535	0.46	0.214
Precision (macro average)	0.37	0.13	0.379
Error	0.03	0.028	0.033
F1	0.328	0.141	0.214
Recall (macro average)	0.537	0.64	0.113
Accuracy	0.97	0.97	0.967
Precision	0.283	0.092	0.442

3. Заключение

Это был первый год нашего полноценного участия в семинаре. К сожалению, в прошлом году, по техническим причинам, до представления итоговых результатов дело не дошло. На этот раз начатое удалось довести до логического завершения.

Для нас было несколько неожиданно то, что результаты, показанные системой при классификации веб-коллекции (точность и полнота), оказались заметно лучше результатов выполнения заданий по тематической классификации, хотя работа с данными из Интернета не является основной специализацией системы. Возможно, такое расхождение объясняется тем, что объем исходных данных дорожки классификации веб-коллекции значительно превосходил объем коллекции нормативных документов, а статистические методы, положенные в основу системы ГЗ, лучше работают на бо́льших коллекциях документов.

Участие в семинаре РОМИП 2004 дало нам очень полезный опыт. Мы получили возможность провести независимую экспертизу некоторых наших разработок в области автоматической классификации, а также познакомиться с результатами других групп, ведущих исследования в данной области. Надеемся, что взаимовыгодное сотрудничество будет расширяться и в дальнейшем.

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] Информационно-правовой портал КОДЕКС. http://www.kodeks.ru

[4] Каталог narod.ru. http://narod.yandex.ru/rubrics/

[5] Каталог DMOZ. http://www.dmoz.org

[6] Описание официальных оценок РОМИП. http://romip.narod.ru/docs/romip_metrics.pdf

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

[8] Сайт РОМИП. http://romip.narod.ru

Galaktika-Zoom: report on participation in RIRES 2004

Antonov A.V., Kozachuk M.V., Meshkov V.S.

This article presents a report on experiments in web site and legal documents classification tasks that were made as part of RIRES initiative. The system's brief information, experiments description and obtained results are described.