Статистическая обработка общественно-политических текстов о регионах России

© Н. Н. Абрамова

ФГУП «НИЦИ при МИД России»
Москва
NAbramova@mid.ru

Аннотация

В статье описываются методы и результаты обработки большого массива общественно-политических текстов о регионах России, собранного за 15 лет, начиная с 2001 года. Статистическая обработка осуществлялась с помощью программы анализа данных AtteStat в среде электронных таблиц Microsoft Excel. Результаты работы могут быть полезны системным администраторам, планирующим размещение информации на серверах, а также экспертам, оценивающим общественно-политическую жизнь регионов России.

1 Введение

В последние годы заметно вырос интерес к мониторингу и оценке региональной информации, представленной в СМИ. Известно более 200 систем мониторинга социальных медиа. Одна из таких систем мониторинга, анализирующая интернет-тексты по теме «Социально-политическая жизнь регионов Российской Федерации» с помощью лингвистического процессора, описывается в статье [1]. Успеха в этой области добилось также российское агентство «Смыслография», которое разрабатывает коммуникационные рейтинги регионов на основе контент-анализа с использованием материалов информационно-аналитической службы Factiva.com в Топ-100 ведущих англоязычных СМИ [2]. Каждому региону присваивается общий балл, учитывающий количество его упоминаний и долю благоприятных публикаций.

В представленной работе предлагается использовать методы математической статистики для составления рейтингов регионов по количеству публикаций в СМИ. Был обработан большой массив публикаций российских СМИ по региональной проблематике за 15 лет, каждый документ которого соотнесен с одним или несколькими регионами, что позволило выявить статистические закономерности при публикации материалов о различных регионах.

2 Формирование баз данных по регионам России

2.1 Источники для формирования баз данных

Базы данных по регионам России формируются путем отбора информации из множества информационных ресурсов:

- сообщений ведущих информационных агентств России: ИТАР-ТАСС, ИНТЕРФАКС, REGNUM, РБК, ПРАЙМ, АК&М, Россия сегодня, Росбалт, lenta.ru, vz.ru, Newsru, Полит.Ру, Утро.Ру, inopressa.ru, expert.ru;
- статей из основных центральных и московских газет и журналов: АИФ, Известия, Новые известия, Независимая газета, РБК-daily, КоммерсантЪ-Daily, Комсомольская правда, ИМ Ведомости, Российская газета, Парламентская газета, Новая газета, Советская Россия, Московская правда, Московский комсомолец, Вечерняя Москва, Огонек и т.д.;
- официальной российской информации, поступающей с сайтов российских органов государственной власти (пресс-служб президента, правительства, генеральной прокуратуры, министерств, агентств, служб, надзоров, управлений).

Объем информационных ресурсов, использованных для формирования баз данных региональной информации, составил около 8 млн. документов.

2.2 Фильтрация информации

Во избежание дублирования информации исключались источники, содержащие обзоры и дайджесты. Также не допускалось попадание в базы данных расписаний спортивных мероприятий, турнирных таблиц, результатов спортивных соревнований. Кроме того, сопровождающий базу региональной информации сотрудник ежедневно просматривает и при необходимости корректирует автоматически распределенную по регионам

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

информацию. Это позволяет, в частности, более точно формировать информацию, относящуюся к региону «Москва».

2.3 Методика отбора информации

В качестве среды для функционирования баз данных используется платформа IBM Lotus Domino/Notes, располагающая встроенным языком программирования Lotus Script [4]. Специальная программа, разработанная на языке Lotus Script, отбирает информацию из информационных ресурсов, содержащих региональную информацию, с помощью запросов, составленных для каждого из 85 регионов России.

Пример запроса для региона «Калужская область» приводится ниже:
((Калужск*|Калуга|Калуге|Калуги|Калугой|Калугу|Обнинск*|Балабаново|Медын*|Малоярославец*) **AND NOT** ((Калужск* **sentence** площад*)|(Калужск* **sentence** шоссе)|(Калужск* застав*)|(метро Калужск*)|метро **sentence** "Калужская"|"м. Калужская"|Калужско-Рижск*|Синема парк на Калужской|"Калужская пл."|"на Калужской")))

В данном запросе используются следующие шаблоны и логические операторы:

* (звездочка) - замена нескольких символов в указанной позиции слова;

«» (кавычки) – точное написание фразы;

AND NOT – запрещенные слова и словосочетания;

| (или) – содержит хотя бы одно из слов;

sentence (предложение) – разделенные оператором слова находятся в одном предложении.

В поисковых запросах также могут быть использованы другие логические операторы (ранг соответствия, абзац, операторы полей, верхнего регистра и веса).

Региональная информация распределялась также по тематическим рубрикам с помощью запросов, составленных для каждой рубрики. Пример запроса для тематики «Религиозно-политические проблемы» приводится ниже:
(религиозн*|православ*|старообряд*|церков*|ислам*|фундаменталист*|клерикал*|миссионер*|католи*|РПЦ|РПЦЗ|христиан*|епископ*|архиер*|Ватикан*|иуд*|пап* римск*|пресвитер*|священник*|диакон*|патриарх*|экумени*|конфесси*|администратур*|межконфессион*|внутриконфессион*|далай-лам*|конфуцианств*|сект*&!секто*|саентолог*|мормон*|мечет*|мусульман*|шариат*|тоталитарн* братств*|митрополит*|священнослужител*|служител* культ*|суфийск* орден*|католикос*|протестант*|раввин*|дацан*|лютеран*|будди*|баптист*|масон*|пятидесятник*|евангели*|монастыр*|синод*|собор*|паломни*|хамбо-лам*|курия|курией|курии|курию)

3 Методы статистической обработки

При статистической обработке использовалась программа анализа данных AtteStat, версия 13, предназначенная для работы в среде электронных таблиц Microsoft Excel 4.1 [3]. Использовались модули программного обеспечения: проверки нормальности распределения Normal Distribution Check for Excel (NDC), корреляционного анализа Correlation Analysis for Excel (CORA), кластерного анализа Cluster Analysis for Excel (CLA).

3.1 Проверка нормальности распределения

По всем годам, начиная с 2001 до 2015, было подсчитано количество поступивших в базы данных документов и определен закон распределения документов. Всего в базы данных поступило 2653060 документов, что составляет ~ 33% от общего объема использованных информационных ресурсов.

На рис.1 приведен график зависимости количества документов от года поступления (по оси x цифра 1 соответствует 2001 году, 2 – 2002 году, 15 – 2015 году).

Проверка нормальности распределения проводилась по модифицированным критериям Колмогорова, Смирнова и хи-квадрат Фишера [3]. Оказалось, что количество поступивших за год документов в базу данных по регионам (случайная величина) подчиняется нормальному закону распределения (по всем трем критериям гипотеза о нормальности не отклонялась). Это позволяет строить прогнозы о размере региональных баз данных, что имеет значение при планировании размещения информации на серверах и закупке серверов.

3.2 Корреляционный анализ

Анализировалась теснота связи (корреляционная зависимость) между количеством документов по регионам, поступившим за два года, с помощью коэффициента корреляции Фехнера. Вычисления производились по формуле

$$r_F = \frac{C - H}{C + H} \quad ,$$

где C - число пар совпадающих знаков отклонений количества документов в анализируемые годы от соответствующих средних значений, H - число пар несовпадающих знаков.

В таблице 1 приведены значения коэффициента Фехнера для некоторых сравниваемых периодов. Значения коэффициента Фехнера указывают на наличие корреляционной связи, причем теснота связи выше для пар, у которых годы идут подряд.

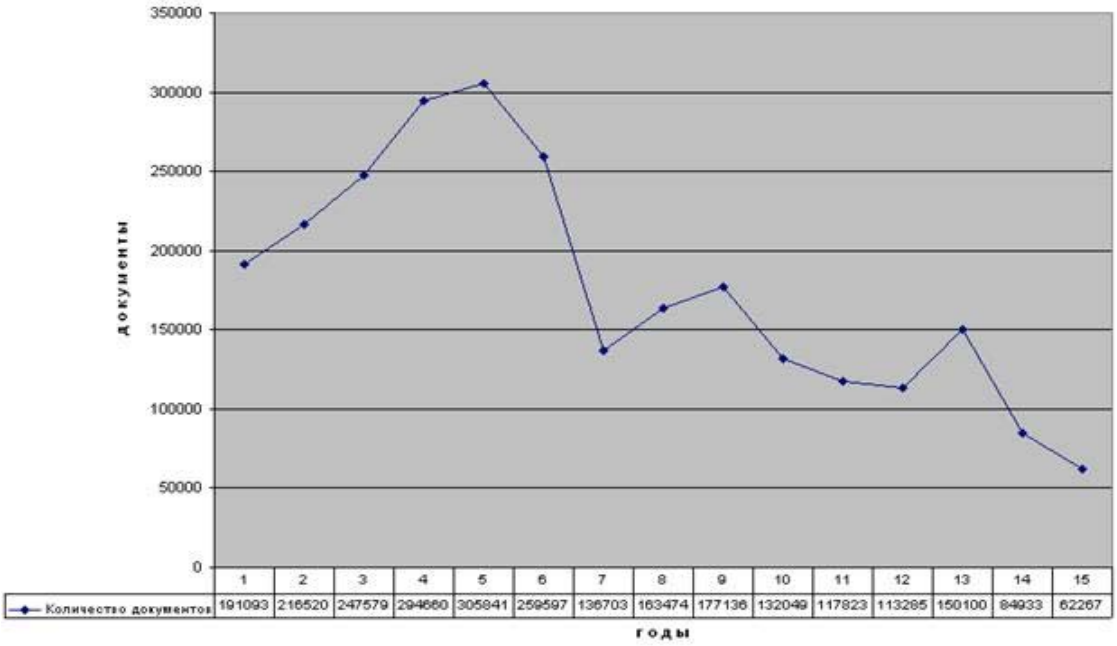

Рисунок 1 Зависимость количества документов от года поступления

Таблица 1 Корреляционная связь между количеством поступающих документов

Период (пары лет)	Коэффициент Фехнера
2001, 2015	0,6867
2002, 2003	0,9036
2002, 2013	0,7108
2004, 2005	0,8072
2008, 2009	0,8313
2009, 2010	0,9277
2012, 2013	0,8313
2013, 2014	0,8313
2014, 2015	0,7349

3.3 Кластерный анализ

Методами кластерного анализа решается задача разбиения множества регионов таким образом, чтобы все регионы, принадлежащие одному кластеру, были более похожи друг на друга по степени освещения их в прессе, чем на объекты (регионы) других кластеров. Так как оцениваются только количественные признаки, был выбран метод Уорда. Каждый объект (регион), описываемый k признаками (количество документов за один конкретный год), может быть представлен как точка в k-мерном пространстве на одном из n объектов. Сходство с другими объектами определялось как Евклидово расстояние между кластерами:

$$\rho(X_i, X_j) = \sqrt{\sum_l (x_{il} - x_{jl})^2},$$

где: X_i , X_j - координаты i-го и j-го объектов в k-мерном пространстве;
x_{il} - x_{jl} - величина l-той компоненты у i-го (j-го) объекта (l=1,2,...,k; i,j=1,2,...,n).

Сущность этого метода заключается в том, что на первом шаге каждый объект рассматривается как отдельный кластер. Процесс объединения кластеров происходит последовательно: на основании матрицы расстояний объединяются наиболее близкие объекты. Сначала объединяются два ближайших кластера. Для них определяются средние значения каждого признака и рассчитывается сумма квадратов отклонений:

$$\delta_l = \sum_i \sum_j (x_{ij} - x_{jl})^2 ,$$

где: l - номер кластера,
i - номер объекта (i = 1,2, ... , n_l),
n_l - количество объектов в l - том кластере,
j - номер признака (j = 1,2, ..., k),
k - количество признаков, характеризующих каждый объект.

В дальнейшем объединяются те объекты или кластеры, которые дают наименьшее приращение величины δ_l.

Кластеризация проводилась по 83 регионам за период с 2001 по 2013 гг. (табл. 2) и по 85 регионам (после вхождения в состав России двух новых субъектов) за 2015 гг. (табл.3).

В таблицах регионы обозначены кодами, которые используются во всех государственных органах России [5]. Например, код «77» относится к Москве, «78» - Санкт-Петербургу, «91» - Республике Крым, «92» - Севастополю и т.д.

Сравнение результатов кластеризации за 2001-2013 и 2015 годы показывает, что регионы, входящие в первую десятку, в целом, сохранили свои позиции, за исключением Чеченской республики. В группу лидеров вошли также два новых российских региона.

Можно составить рейтинг регионов по каждому году, исходя из количества документов, в которых упоминается регион. Так, в 2015 году абсолютными лидерами были Москва (1место) и С.-Петербург (2

место), затем следовали Краснодарский край, Московская область, Приморский край, Республика Крым, Севастополь, Новосибирская область, Красноярский край, и замыкал десятку Татарстан. Регионы-аутсайдеры представляли Ненецкий АО (81 место) и области: Ульяновская (82 место), Тульская (83 место), Магаданская (84 место), Липецкая (85 место). Любопытно, что Башкортостан, занявший седьмое место в рейтинге регионов в зарубежных средствах массовой информации по итогам 2015 года [2], оказался только на 16-ом месте. Такие различия можно объяснить разными подходами к отбору информации.

Таблица 2 Результаты кластеризации регионов (2001-2013 гг.)

Номер кластера	Численность кластера	Коды субъектов РФ
1	1	77
2	1	78
3	3	20,25,50
4	8	16,23,24,27,38,39,54,66
5	19	02,05,06,15,26,34,36,41, 52,53,55,59,61,63,64,65, 72,73,74
6	24	07,10,14,22,28,30,31,32, 40,42,46,47,48,49,51,56, 60,62,69,70,71,75,76,83
7	27	01,03,04,08,09,11,12, 13,17,18,19,21,29,33,35, 37,43,44,45,57,58,67,68, 79, 86,87,89

Таблица 3 Результаты кластеризации регионов за 2015 год

Номер Кластера	Численность кластера	Коды субъектов РФ
1	1	77
2	1	78
3	3	23,25,50
4	3	54,91,92
5	3	16,24,66
6	18	02,05,26,27,34,36,38,39, 41,52,55,59,61,63,65,70, 72,74
7	7	14,20,22,28,47,75,89
8	11	03,15,30,31,32,42,51,56, 60,64,69
9	22	04,06,07,10,11,17,18,19, 29,33,35,37,43,45,46,57, 58,67,68,86,87,89
10	16	01,08,09,12,13,21,40,44, 48,49,53,71,73,76,79,83

Зарубежные СМИ традиционно проявляют интерес к спортивным и статусным мероприятиям в российских регионах, таким как подготовка к чемпионату мира по футболу в 2018 году, российским спортивным первенствам, проведению саммитов ШОС и БРИКС, а в рассматриваемых нами базах данных наибольший интерес представляет общественно-политическая информация о жизни регионов.

При анализе информации о регионах важны не только количественные характеристики о числе документов, в которых упоминается регион, но и качественные, т.е. тематический характер. В качестве примера были выбраны три наиболее востребованные нашими пользователями тематики («Религиозно-политические проблемы», «Права человека», «Охрана окружающей среды» (ООС)), для которых построены графики зависимостей количества документов от года поступления (см. рис.2).

Нами была проведена кластеризация документов за 2014-2015 годы по трем перечисленным выше тематикам. В табл.4 представлены регионы, которые попали в первые четыре кластера, хотя бы по одной тематике (указывается номер кластера и в скобках рейтинг региона), а по другим тематикам они могут занимать даже последние позиции. Документы с упоминанием Москвы составили отдельные кластеры по всем тематикам, а с упоминанием Санкт-Петербурга - по двум тематикам.

4 Заключение

Проведенный анализ может быть использован для прогнозирования размера региональных баз, составления рейтинга регионов по освещению в прессе в целом и с учетом тематики, а также просмотра изменений рейтингов в динамике.

В работе показано, что наряду с методами контент-анализа, семантико-ориентированного лингвистического анализа для мониторинга публикаций о регионах можно использовать методы математической статистики – корреляционный и кластерный анализ. Эти методы не требуют затрат, связанных с работой экспертов. Однако для статистического анализа нужно располагать большим объемом информации.

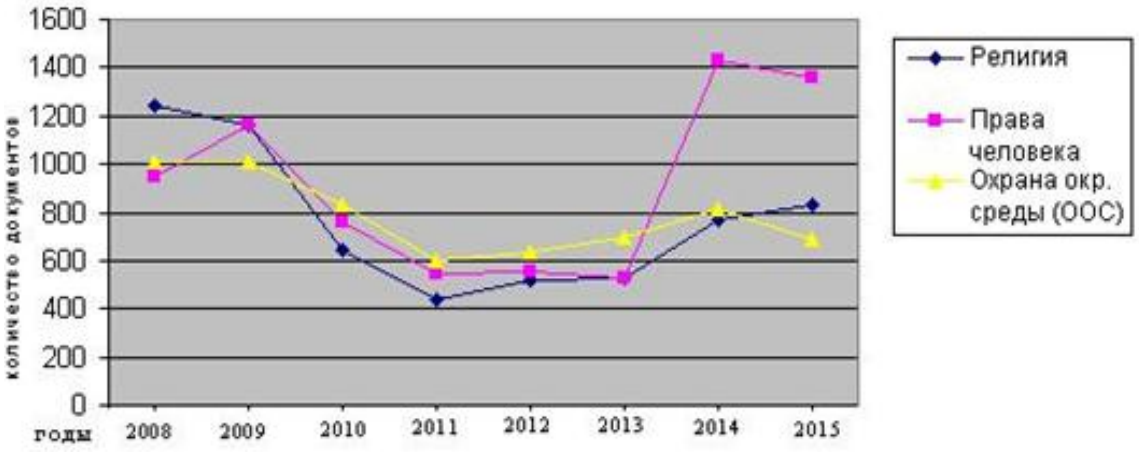

Рисунок 2 Распределение числа документов по тематикам за 2008-2013 гг.

Таблица 4 Регионы-лидеры по тематическим областям (2014 - 2015гг.)

Субъект РФ	**Религия**	**Права человека**	**ООС**
Москва	1 (1)	1 (1)	1 (1)
С.-Петербург	2 (2)	2 (2)	2 (2)
Московская обл.	3 (3)	3 (4)	2 (4)
Приморский край	5 (8)	4 (11)	2 (3)
Республика Крым	5 (9)	3 (3)	3 (7)
Севастополь	5 (10)	3 (5)	3 (9)
Краснодарский край	5 (6)	3 (6)	4 (10)
Забайкальский край	6 (25)	7 (39)	3 (5)
Красноярский край	6 (18)	5 (15)	3 (8)
Республика Бурятия	7 (45)	7 (75)	3 (6)
Новосибирская обл.	4 (5)	4 (9)	5 (17)
Чечня	4 (4)	4 (7)	7 (69)
Татарстан	4 (7)	4 (8)	6 (21)
Свердловская обл.	6 (15)	4 (12)	6 (20)
Нижегородская обл.	6 (14)	4 (10)	6 (19)
Иркутская область	6 (21)	6 (23)	4 (11)
Камчатский край	6 (23)	7 (73)	4 (12)
Хабаровский край	6 (24)	7 (40)	4 (13)

Литература

[1] Е. Б.Козеренко и др. Создание системы мониторинга интернет-текстов по теме «Социально-политическая жизнь регионов Российской Федерации»//Материалы III Международной научно-практической конференции (Москва, 18–19 сентября 2014): Сборник статей и тезисов. – М.: МГГУ им. М. А. Шолохова, 2014. С. 51–55. [Электронное издание] http://mggu-sh.ru/sites/default/files/sb_2014.pdf.

[2] О.Васильева. Коммуникационный рейтинг как инструмент оценки федерального образа региона// СОВЕТНИК №5 (185), 2011. С. 12-13 [Электронное издание] http://s-graph.ru/upload/iblock/faf/faf7de0c213d06b561daa433868163d1.jpg

[3] Программное обеспечение анализа данных AtteStat. Руководство пользователя. Версия 13//Авторское право © И.П. Гайдышев, 2002–2012. http://биостатистика.рф/files/13.pdf

[4] Н.Н.Ионцев, Е.В.Поляков, О.Г.Таранцев. Программирование в Lotus Domino R5: формулы и функции, язык Lotus Script, встоенные классы Lotus Script и Java. – М.:изд. «Светотон», 2000, 935 с.

[5] Классификатор субъектов Российской Федерации. Материалы из Википедии. https://ru.wikipedia.org/wiki/Коды_субъектов_Российской_Федерации.

Statistical processing of the social and political texts about Russian regions

N.N. Abramova

The article describes the methods and results of the processing of the social and political texts about the Russian regions that was collected since 2001 during 15 years. Statistical processing was carried out using the data analysis program AtteStat in the environment of Microsoft Excel spreadsheets. The results can be used by system administrators for planning the placement of information on the servers, as well as by experts evaluating the social and political life of the Russian Federation regions.