Формирование «повестки дня» в сфере электронного правительства: результаты контент-анализа новостных сообщений

Л.А. Видясова, С.Н. Кольцов, А.В. Чугунов

Университет ИТМО, НИУ ВШЭ - Санкт-Петербург
bershadskaya.lyudmila@gmail.com, kol-sergei@yandex.ru, chugunov@egov-center.ru

Аннотация

В статье представлена методика и результаты локального исследовательского проекта, выполненного в 2013 году коллективом Центра технологий электронного правительства Университета ИТМО совместно с Лабораторией интернет-исследований Санкт-Петербургского кампуса НИУ ВШЭ

С целью анализа формируемой «повестки дня» по тематике услуг электронного правительства был проведен автоматизированный контент-анализ новостных сообщений по тематике электронного правительства. В результате была исследована специфика формирования «повестки дня» по тематике электронного правительства и особенности циркулирования информации по отдельным темам. Исследованием была выявлена корреляция между тематикой публикуемых сообщений и официальными сроками реализации государственных программ и проектов, этапностью при переходе на электронные услуги.

1. Объект исследования и постановка задачи

Представляемая работа является частью серии исследований, ориентированных на изучение «повестки дня», формируемой средствами массовой информации (СМИ) и интернет-ресурсами по тематике, связанной с развитием электронного правительства, а также выявление специфики обсуждаемости процесса внедрения электронных государственных услуг в сети Интернет, в том числе в социальных медиа и блогосфере. [1] [2]

С целью анализа формируемой «повестки дня» по тематике услуг электронного правительства был проведен автоматизированный контент-анализ новостных сообщений

Объектом исследования стали около 3,5 тыс. новостных сообщений за три года (01.03.2010 - 01.05.2013). Массив информации был сформирован на основе ручного отбора новостных сообщений СМИ за указанный период. Новостные сообщения формировались с 2010 г. еженедельно в рамках текущей информационно-аналитической деятельности Центра технологий электронного правительства Университета ИТМО и служили основой для выпуска еженедельного информационного бюллетеня, выпускаемого в электронном виде. Новостные сообщения также размещаются на сайте Центра технологий электронного правительства (ЦТЭП – http://egov.ifmo.ru/) со ссылками на первоисточники.

Важным фактором, который позволяет считать эти тексты репрезентативными, т.е. освещающими весь основной поток публикаций СМИ по указанной тематике, является (а) наличие перечня источников, освещающих тематику, и подлежащих обязательному просмотру; (б) регулярность и систематичность сбора информации экспертами ЦТЭП; (в) отсутствие в массиве републикаций, т.е. в подборку включаются только самые информативные сообщения СМИ, «перепечатки» игнорируются; (г) положительные отзывы подписчиков бюллетеня - представителей органов власти и экспертов, использующих данную информацию в своей деятельности.

Массив сообщений представляет собой набор файлов, каждый из которых включает информацию за определенный месяц. Всего для анализа было подготовлено 39 файлов (в каждом от 70 до 100 новостных сообщений).

Контент-анализ проводился на основе изучения содержания новостных сообщений из следующих групп информационных источников (СМИ и информационные ресурсы, регулярно публикующие новости по тематике электронного правительства и электронных услуг):

1. Сайты органов власти и официальные тематические порталы (Сайты Президента России, Правительства РФ, Минкомсвязи и Минэкономразвития, сайт «Административная реформа в Российской Федерации», Комиссии по модернизации и технологическому развитию экономики России и др.).

2. Региональные ресурсы (Интернет-представительство глав регионов РФ «Клуб Регионов»; сайты «Электронных правительств» Астраханской и

Технологии информационного общества в науке, образовании и культуре: сборник научных статей. Материалы XVII Всероссийской объединенной конференции «Интернет и современное общество» IMS-2014, Санкт-Петербург, 19 - 20 ноября 2014 г.

Самарской областей, региональные новостные СМИ, имеющие соответствующие разделы или тематические рубрики и др. ресурсы).

3. Экспертные структуры (ВЦИОМ, ГосМенеджмент: электронный журнал, Всероссийский научно-исследовательский институт проблем вычислительной техники и информатизации, Экспертный центр электронного государства, Фонд информационной демократии и др.).

4.СМИ и новостные порталы (CNews, ComNews, Российская газета, Комсомольская правда, Известия, Независимая газета, Коммерсантъ, Ведомости, ПРАЙМ-ТАСС, РБК daily, РИА «ФедералПресс», ТАСС-Телеком, PC-Week, ИА REGNUM Новости, Портал Право.ру и др.)

5. Ресурсы в социальных сетях и блоги (блог «Госуслуги» в Livejournal, группа «Электронное правительство» в Facebook и др.).

Для контент-анализа были определены следующие единицы счета (тематики), которые представляют интерес с точки зрения выявления специфики формирования «повестки дня»:

- нормативно-правовые и организационные вопросы формирования электронного правительства;
- внедрение электронных государственных и муниципальных услуг;
- отзывы о получении электронных государственных и муниципальных услуг,
- специфика оказания электронных услуг в различных сферах (образования, здравоохранение и т.д.);
- Единый портал государственных и муниципальных услуг;
- имена и фамилии руководителей, занимающихся проектами внедрения электронного правительства и электронных услуг;
- события, мероприятия, связанные с данной темой.

2. Инструментарий и методы больших текстовых данных

Исследование больших интернет данных ставит перед социологами ряд новых задач. Во-первых, это проблема больших данных (big data), которая требует применение достаточно сложных математических алгоритмов, во-вторых, социологи должны заниматься переосмысливанием границ объектов и выбором единиц анализа, с учетом математической модели. Одной из таких моделей, которая начиная с недавнего времени, стала активно использоваться в повседневном социологическом анализе, является тематическое моделирование [3]. Тематическое моделирование (topic modeling) позволяет производить нечеткую кластеризацию документов по темам, соответственно именно тематическое многообразие становится в фокусе социологического анализа данной работы.

В компьютерной лингвистике под темой понимается совокупность общих слов в текстах, которые имеют тенденцию встречаться совместно в одних и тех же текстах. Такая интерпретация темы, позволяет сформулировать лингвистическую модель генерации контента документов коллекции, и на основании модели разработать алгоритм вычисления распределения документов и слов по темам. На данный момент разработано множество различных вариантов тематических моделей (Latent Dirichlet Allocation, LDA), однако они базируются на двух основных вариантах:

- Вариационная модель [4].
- Сэмплирование Гиббса [5].

В LDA предполагается, что существует конечное множество тем T и коллекция документов порождается дискретным распределением $p(d,w,t)$, где d - документ, w - слово, t - тема. Переменные d и w являются наблюдаемыми переменными, а переменная t — скрытой, т. е. появление каждой пары (d,w) связано с некоторой неизвестной темой t. Построить тематическую модель коллекции — означает найти множество скрытых тем T, условные распределения $p(w| t) \equiv \varphi(w,t)$ для каждой темы t и $p(t | d) \equiv \theta(t,d)$ для каждого документа d.

В рамках данной работы использовалась процедура сэмплирования Гиббса [5] для нахождения распределений документов и слов по темам по заданной коллекции документов. Выбор данной методики обусловлен вычислительной простотой.

На данный момент существует множество различных программных средств для проведения тематического моделирования [6]. Однако, как показывает анализ этих средств [13] существующим программным средствам присущи следующие недостатки:

1. Программное обеспечение дает лишь вычислительное ядро с тем или иным форматом ввода и вывода данных.

2. Отсутствуют методы препроцессинга данных для русского языка, который включал бы в себя очистку текста, лематизацию, формирование списка стоп слов и удаление стоп слов. Следовательно эти работы пользователь должен делать сам.

3. В целом отсутствуют готовые решения для анализа данных.

4. Развертывание больших систем требует создание вычислительного кластера, соответственно возникает проблема технической поддержки этого кластера.

5. Работа с такими пакетами требует специфичных навыков программирования. Исходя из этого в качестве инструмента для проведения тематического моделирования новостей по тематике электронного правительства и электронных услуг, был выбран программный продукт Topic Miner, разработанный в Санкт-Петербургском кампусе НИУ ВШЭ (разработчики С.Н. Кольцов, В.Г. Филипов) [7, 13]. Topic Miner – это информационная система, предназначенная для проведения тематического моделирования на больших массивах текстовых данных.

В состав системы входят следующие модули:

- препроцессинга данных для русского языка;
- тематического моделирования;
- анализа результатов расчета.

Модуль препроцессинга данных. Препроцессинг данных заключается в специализированной обработке текстовых данных и конвертации данных в векторный формат. Обработка включает в себя удаление html тэгов, лематизация, удаление стоп слов (слова, не несущие в себе смысла и не влияющие на его тематику, такими словами чаще всего являются предлоги, союзы и другие части речи) и подсчет частот слов (лем). Конвертация текста в векторный формат основана на векторной модели текста [8].

Модуль тематического моделирования. Входными параметрами для сэмплирования Гиббса являются: число тем, параметры описывающие распределения Дирихле (α,β), число итераций сэмплирования. Результатами моделирования являются две матрицы: (а) матрица φ(w,t) распределение слов по темам; (б) матрица θ(t,d) - распределение документов по темам. В каждой ячейке матрицы находятся вероятности принадлежности слов/документов к теме.

Модуль анализа результатов расчета. В данном модуле реализованы следующие функции анализа. (а) сортировка данных по каждой теме по величине вероятности (сортировка реализована по технологии CUDA); (б) визуализация больших массивов (технология file mapping); (в) удаление хвостов распределений (удаление низких величин вероятностей слов и документов) (г) выгрузка распределений слов и документов в различных форматах для внешних программ (выгрузки осуществляются в формате csv); (д) оценка расстояния между словами в текстах (элементы семантического анализа).

Тематическое моделирование является обобщением кластерного анализа, поэтому оно унаследовало и основные недостатки кластерного анализа. Определение количества кластеров – одна из серьезных проблем в кластерном анализе и смежных методах деления объектов на группы. Существует множество методов оценки качества кластеризации, которые можно использовать для сравнительной оценки кластерных решений с разным количеством кластеров в решениях. Однако проблема в том, что такого рода функции качества монотонно или почти монотонно изменяются с изменением количества кластеров, и если имеют перегиб или скачок, то едва заметный и не определяемый «на глаз».

Для решения этой проблемы, в рамках кластерного анализа, существует непараметрический метод из Distortion Theory [9], который позволяет трансформировать функцию качества кластерного решения так, чтобы перегиб или скачок стал явно видимым. Метод основан на использовании понятия «искажений» (distortion), которые являются оценками дисперсии внутри класса (кластера). В основе такого метода лежит оценка скорости изменения внутрикластерного (межкластерного) расстояния как функции от количества кластеров. Данный метод хорошо себя зарекомендовал в различных задачах кластерного анализа [10, 11]. Однако применение этого метода для тематического моделирования требует замены внутрикластерного расстояния на величину, которая бы хорошо характеризовала процесс тематического моделирования.

Одной из наиболее распространенных мер качества тематического моделирования является perplexity. Под этим понятием подразумевается следующая величина. Перплексити это обратная величина среднегеометрической вероятности слова принадлежности слова к темам. Данная величина показывает то насколько модель, заложенная в LDA, хорошо описывает распределения, скрытые в коллекции документов [12]. Чем меньше величина перплексити тем лучше (неоднороднее) модель LDA и тем сильнее отличие модели от начального однородного распределения. Перплексити меняется от единицы до нуля, и по сути, является аналогом внутрикластерного расстояния. В нашем исследования для вычислении оптимального количества тем мы использовали теорию скачков [9], в которой в виде функции характеризующую качество тематического моделирования в зависимости от количества тем использовали перплексити.

Исходные данные. В рамках данной работы для анализа было использовано 3444 документов, в которых число уникальных слов составляет 562. Каждая тема характеризуется набором наиболее вероятностных слов. Каждый документ включает следующие метаданные:

- название текста и его содержимое (текст статьи или сообщения СМИ);
- ссылка на первоисточник (URL документа);
- дата публикации.

В качестве входных данных использовались следующие параметры: α=0.5, β=0.1. Число тем варьировалось от 16 - 80 тем с шагом 8.

В качестве результатов тематического моделирования были получены две матрицы (для каждого тематического решения): (а) матрица распределения слов по темам; (б) матрица распределения документов по темам, и кривые перплексити.

Определение оптимального количества тем производилось следующим образом. Для заданной коллекции документов, проводилась серия тематических моделирований, в которых варьируемым параметром служила величина перплексити. В качестве окончательной величины бралось значение после 200 итерации, так как 200 итераций было достаточно для сходимости модели. Применение теории скачков для набора тематических моделей на основе перплексити показало, что оптимальным тематическим решением является 56 тем. Именно это

тематическое решение использовалось для анализа результатов тематического моделирования.

3. Результаты исследования и их интерпретация

С помощью применения данного автоматизированного инструмента имеющийся массив новостных сообщений был кластеризован на 56 тем (табл. 1).

Таблица 1. Результат кластеризации сообщений СМИ по тематике электронного правительства и электронных услуг, 2010-2013 гг.

№	Тематические группы	Кол-во тем
1	Общая информация о развитии электронного правительства, планы, программы , эффективность реализации	16
2	Электронные услуги, отраслевая информация	9
3	Технологии, порталы, информационные ресурсы, инфраструктура	17
4	Региональный и международный опыт, пилотные проекты	7
5	Персоналии и структуры	7
	ИТОГО	56

Результаты контент-анализа новостных сообщений позволяют сформулировать следующие обобщения и выводы.

1. СМИ и официальные сайты государственных структур формируют «повестку дня» по тематике электронного правительства и обеспечивают циркулирование информации по отдельным тематикам. Наиболее часто за последние 3 года в СМИ появлялись сообщения о планах по развитию электронного правительства, перспективах внедрения новых ИТ-систем в органах власти.

2. На протяжении всего исследуемого периода в СМИ постоянно появляются репортажи и сообщения о конкретных примерах получения государственных услуг, описания специфики работы портала государственных и муниципальных услуг, высказываются различные оценки (и отрицательные, и позитивные) деятельности органов власти в этой области.

3. За исследуемый период стала популярной публикация сообщений о развитии ИТ-рынка, инновациях, проектах Сколково, мобильных приложениях, которые являются новыми инструментами для реализации услуг электронного правительства.

4. Большая часть публикуемых сообщений связана с оповещением общественности о принятии новой нормативной базы, регламентирующей сферу развития электронного правительства, разработкой программ и стратегий (государственная программа «Информационное общество», отраслевые программы), формированием планов и реализации этапов перехода на электронные государственные услуги.

5. Развитие электронного правительства на региональном уровне также активно освещается в СМИ, однако отчетливо прослеживается тенденция к обсуждению опыта и проектов регионов-лидеров, активных в направлении развития электронного правительства и электронных услуг. В этом плане выделяются, по результатам контент-анализа, следующие регионы: Москва, Республика Татарстан, Санкт-Петербург и Самарская область.

Исследованием также была выявлена корреляция между тематикой публикуемых сообщений и официальными сроками реализации государственных программ и проектов, этапностью при переходе на государственные электронные услуги. В частности, в мае-июне 2011 г., когда был объявлен первый этап по подключению органов власти к системе межведомственного электронного взаимодействия (СМЭВ), в СМИ активно обсуждалась работа губернаторов разных регионов и ее результаты (имеется в виду степень готовности к выполнению тех планов, которые были утверждены на федеральном уровне).

Следует отметить, что тематики, связанные с нормативно-правовыми и организационными вопросами формирования электронного правительства, развитием портала государственных услуг, выдачей универсальных электронных карт, работой руководителей федеральных и региональных органов власти являются наиболее популярными на протяжении всего периода исследовании. В то же время, в поле медиа-контента присутствуют темы, характеризующиеся временными вспышками (1-2 месяца) новостных публикаций. К таким темам относятся: открытые данные, открытое правительство, электронное правительство в СНГ, информатизация сферы здравоохранения, образования и т.д.

В завершение важно подчеркнуть, что данное исследование носит пилотный характер и было включено в программу более широкого исследования, в рамках которого проводился и анализ запросов по тематике электронных госуслуг в поисковых системах Яндекс и Google, осуществлялся ретроспективный анализ обсуждаемости проблематики электронных государственных услуг в социальных сетях с использованием автоматизированных методов, применялись и качественные методы (экспертный опрос и другие способы получения информации).

Авторы планируют продолжение данной работы, в том числе – провести контент-анализ данного массива информации с использованием других программных средств и методик.

Отдельный интерес представляют сравнительные исследования, когда по сопоставимой методике анализируются публикации в средствах массовой информации и мнения пользователей социальных сетей. Естественно, что это связанные между собой процессы, когда довольно часто

поводом для всплеска обсуждений в социальных медиа служат публикации «традиционных» СМИ, но имеются и иные примеры, когда начавшись в социальных сетях, дискуссия перетекает на страницы газет и журналов, зачастую через их онлайновые версии.

Литература

[1] Бершадская Л.А., Чугунов А.В. Услуги электронного правительства: исследование дискуссий в социальных сетях // Межотраслевая информационная служба. 2014. № 1 (166). С. 10-17. URL: http://elibrary.ru/item.asp?id=21278501

[2] Бершадская Л.А., Биккулов А.С., Болгова Е.В., Чугунов А.В., Якушев А.В. Социальные сети и социометрические исследования: теоретические основания и практика использования автоматизированного инструментария изучения виртуальных сообществ // Информационные ресурсы России. 2012. № 4. С. 19—24. URL: http://elibrary.ru/item.asp?id=17910592

[3] Koltsova O., Koltcov S. Mapping the Public Agenda with Topic Modeling: The Case of the Russian LiveJournal // Policy & Internet. 2013. Vol. 5. No. 2. P. 207-227.

[4] Blei D.M., Ng A.Y., Jordan M.I. Latent Dirichlet allocation // Journal of Machine Learning Research. 2003. Vol. 3. P. 993–1022.

[5] Steyvers M., Griffiths T. Finding scientific topics // Proceedings of the National Academy of Sciences. 2004. Vol. 101, no. Suppl. 1. P. 5228–5235.

[6] Коршунов А., Гомзин А. Тематическое моделирование текстов на естественном языке // Труды ИСП РАН. М., 2012.

[7] Лаборатория Интернет - исследований, Национальный Исследовательский университет, Высшая Школа Экономики, Отчет о научно-исследовательской работе: Социально политические процессы в интернете. 2013г. URL: http://www.hse.ru/org/projects/79645357.

[8] Andrews N.O, Fox E.A. Recent Developments in Document Clustering, 2007. October 16. URL: http://eprints.cs.vt.edu/archive/00001000/01/docclust.pdf.

[9] Sugar C., James G. Finding the number of clusters in a data set: An information theoretic approach // J. of the American Statistical Association. 2003. Vol. 98. P. 750–763.

[10] Кольцова О.Ю., Ясавеев И.Г. Конструирование проблемы полицейского насилия в Российской блогосфере: риторика, лейтмотивы и стили // Журнал социологии и социальной антропологии. 2013. № 3. С. 81 - 100.

[11] S. Still and W. Bialek: How many clusters? An information theoretic perspective // Neural Computation. 2004. Vol. 16(12). P. 2483-2506.

[12] Newman D., Smyth P., Steyvers M. Scalable Parallel Topic Models // Journal of Intelligence Community Research and Development. 2006. Vol. 1.

[13] Koltsova O., Koltsov S., Nikolenko S.I. Comment-Based Discussion Communities In The Russian LiveJournal And Their Topical Coherence // Working papers by NRU Higher School of Economics. Series SOC "Sociology". 2013. No. WP BRP 33/SOC/2013. URL: http://www.hse.ru/data/2014/01/13/1340844763/33SOC2013.pdf

E-Governance "Agenda" Development: the Results of News Reports Content Analysis

A. Chugunov, S. Koltsov, L. Vidyasova

The article represents the results of Russian mass media content-analysis in the topics related to e-governance. The study has found a correlation between the themes and messages published on official implementation state programs and projects, stages in the transition to electronic services.