Вестник Омского университета. Серия «Социология». 2014. № 1/2. С. 12–16.

УДК 311.216

ПОСТРОЕНИЕ ТЕМАТИЧЕСКОГО ПРОФИЛЯ ОМСКИХ ИНТЕРНЕТ-СМИ

BUILDING A THEMATIC PROFILE OF OMSK INTERNET MEDIA

О.С. НАГОРНЫЙ

O.S. NAGORNY

Омский государственный университет им. Ф.М. Достоевского

Приводятся результаты автоматического выделения тем методом LDA из текстов новостных статей омских интернет-СМИ. Показывается последовательность шагов, необходимых для предварительной обработки данных. Обосновывается выбор количества тем для построения модели тематического моделирования. Полученные темы ранжируются по распространённости и комментируемости.

In this paper we present the results of the automatic selection of texts by the LDA news articles Omsk online media. Shows the sequence of steps required for pre-processing data. The choice of the number of topics to construct a model of thematic modeling. The resulting threads are ranked by prevalence and commented.

Ключевые слова: латентное размещение Дирихле, тематическое моделирование, интеллектуальный анализ текста, интернет-СМИ, новости, веб-скрапинг, обработка текста, Омск.

Key words: LDA, latent Dirichlet allocation, topic modelling, text-mining, online mass media, news, web scraping, text processing, Omsk.

В условиях, когда благодаря распространению сети Интернет большинство людей имеют возможность в невиданных ранее объёмах потреблять и создавать информацию, социологам следует использовать соответствующие этим условиям подходы к сбору и анализу данных. Для описания нового характера таких данных, которые отличаются большим объёмом, высокой скоростью роста и значительным многообразием свои форм, был введён термин «большие данные» (big data). Феномен «больших данных» создал потребность в новых методах обработки и анализа, способных извлекать из этих, как правило, неструктурированных данных полезное знание. Совокупность таких методов обозначается термином «интеллектуальный анализ данных» (data mining).

Из этой совокупности выделяется подмножество, специализирующееся на анализе текстовых данных – «интеллектуальный анализ текстовых данных» (text mining). Кроме того, выделяется группа методов, которые выполняют задачу тематического моделирования – построения статистических моделей, определяющих тематическую принадлежность каждого документа из корпуса.

Тематическое моделирование может быть успешно использовано для решения социологических задач. С их помощью, например, можно определить разницу в освещении информации правительственными и оппозиционными СМИ [1] или увидеть, как менялось отношение СМИ к культурной политике государства [2]. В данном исследовании мы определим, какие темы составляют тематический профиль омских интернет-СМИ, выделим самые популярные темы у редакторов и самые комментируемые у читателей.

Сбор данных

Генеральную совокупность в данном исследовании составили новостные статьи интернет-СМИ города Омска, опубликованные с 1 сентября 2013 г. по 1 сентября 2014 г. омским интернет-СМИ считался веб-сайт, ставящий своей задачей выполнение функции средства массовой информации в сети Интернет и ориентированный на людей, живущих в Омской области. Из всей совокупности данных СМИ были выбраны четыре самых

популярных, на долю которых приходится 65 % просмотров [3]: Город 55, БК55, НГС Омск и Омск-Информ. Новостными статьями считались те, которые публиковались на данном ресурсе в разделе «Новости» или в соответствующем ему. Статьи из категорий «Работа», «Объявления», «Блоги» и другие в исследовании не участвовали.

Новостная статья – не просто текст. Это документ, который имеет свою структуру. В его структуре нас интересовали такие элементы, как собственно текст, название, дата публикации, комментарии и принадлежность к тому или иному СМИ. Данные элементы представлены в статьях каждого из рассматриваемых нами СМИ и являются достаточными для решения исследовательских задач.

Для сбора статей использовался язык программирования Python.

Количество собранных таким образом статей составило 33887 единиц.

Предварительная обработка данных

Обработка данных – чрезвычайно важный этап интеллектуального анализа текста. Цель исследователя на этом этапе – удаление несущественных и вносящих помехи данных и преобразование данных к удобному для анализа виду.

Удаление специфических признаков

Данный этап предварительной обработки данных заключается в удалении из каждой статьи признаков, свидетельствующих о её принадлежности к какому-либо источнику. Если внимательно посмотреть на полученные тексты, то можно увидеть, что редакция каждого СМИ устанавливает собственные правила оформления статей. Эти правила касаются способа оформления ссылок на источники данных, упоминания имён репортёров, специфических для данного СМИ условных обозначений. В случае если отличительные черты не будут устранены, алгоритмы тематического моделирования, которые мы в дальнейшем собираемся применить к собранному корпусу текстов, будут стремиться образовать темы вокруг источников. Процедура унификации статей из различных источников достаточно трудоёмка и требует ручного анализа множества статей, позволяющего выявить специфические черты для каждого сайта и написать программу, которая избавится от них.

После устранения специфической информации данные из различных источников объединялись в единый корпус и подвергались дальнейшей обработке.

Токенизация

Следующим этапом предварительной обработки текста является токенизация. Именно с неё обычно начинается обработка естественного языка. Под токенизацией понимают процесс сегментации текста на отдельные части, называемые токенами. Именно токены являются теми первичными элементами, которые непосредственно участвуют в процессе анализа. Два основных признака токена – это лингвистическая значимость и методологическая полезность.

Нами было протестировано несколько алгоритмов токенизации. Корректнее всех выделял токены изначально не предназначенный для работы с русским языком токенайзер из программы Pattern. Например, он единственный интерпретировал url'ы как цельные токены, не выделяя в них отдельные сегменты на основе знаков препинания.

Стемминг и лемматизация

После токенизации и удаления токенов, являющихся знаками препинания, мы перешли от представления документов как набора символов к документам как списку слов. Дальнейшие наши шаги будут направлены на уменьшение длинны этого списка, т. е. на снижение общего количества токенов. Необходимость таких шагов обусловлена желанием снизить вычислительную сложность анализа данных.

Для компьютера различные формы одного и того же слова являются совершенно разными словами. Существует два способа приведения различных словоформ к одной лексеме. Первый, самый простой, – стемминг. Он состоит в отсечении слово- и формообразующих частей (префиксов, суффиксов, окончаний), в результате чего остаётся основа слова – неизменная часть, выражающая его лексическое значение.

Более сложным подходом к решению проблемы унификации словоформ является лемматизация. Лемматизация – это процесс приведения словоформы к лемме – её нормальной (словарной) форме. В русском языке нормальная форма имени существительного имеет именительный падеж и единственное

число, для прилагательных добавляется требование мужского рода, а глаголы, деепричастия и причастия в нормальной форме должны стоять в инфинитиве.

Для постановки слова в нормальную форму необходимо иметь словарь, где для каждого слова определены его характеристики, т. е. часть речи, падеж, число, род, форма глагола (если это глагол). Создание такого словаря требует колоссальных трудов. Стемминг же предполагает наличие лишь списка приставок, суффиксов и окончаний, количество которых исчисляется несколькими десятками. К счастью, для русского языка существует так необходимый для лемматизации словарь, созданный в рамках проекта OpenCorpora. Использующая этот словарь программа pymorphy2 позволяет приводить слова к нормальной форме.

Между вышеозначенными способами мы выбрали лемматизацию, поскольку получаемые в результате этого процесса леммы интерпретировать легче, чем усечённые основы слов, значение которых не всегда легко восстановить.

Удаление стоп-слов

Дальнейшие усилия по уменьшению количества токенов связаны с удалением так называемых стоп-слов. Эти слова, сами по себе почти не неся полезного смысла, тем не менее необходимы для нормального восприятия текста. Чаще всего к разряду стоп-слов относятся служебные части речи – предлоги, союзы, частицы. Будучи широко распространёнными в любых текстах, они мало могут сказать о его специфике, а следовательно, и о теме.

В качестве базы для списка стоп-слов были использована русские стоп-слова из программы NLTK. Однако такой список нельзя считать достаточно полным. Включая в себя 151 слово, он покрывает лишь самые основные случаи. Для его пополнения мы обратились к собранным ранее данным. На их основе был составлен список часто встречающихся в корпусе токенов, среди них выбраны несколько десятков, подходящих под описание стоп-слов (*это*, *который*, *такой*, *некоторый*, *другой*, *тот* и др.), которые затем были добавлены в соответствующий список и удалены из статей.

Тематическое моделирование

Построение тематической модели может рассматриваться как задача одновременной кластеризации документов и слов по одному и тому же множеству кластеров, называемых темами. В терминах кластерного анализа тема – это результат бикластеризации, то есть одновременно кластеризации и слов, и документов по их семантической близости. Обычно выполняется нечёткая кластеризация, то есть документ может принадлежать нескольким темам в различной степени. Таким образом, сжатое семантическое описание слова или документа представляет собой вероятностное распределение на множестве тем. Процесс нахождения этих распределений и называется тематическим моделированием.

Что касается конкретных методов тематического моделирования, то в данном исследовании будет использован метод латентного размещения Дирихле (latent Dirichlet allocation, LDA). Созданный в 2003 г. [4], сейчас LDA, безусловно, лидирует среди других вероятностных тематических моделей.

На выходе после обучения модели LDA получаются векторы, показывающие, как распределены темы в каждом документе, и распределения, показывающие, какие слова более вероятны в тех или иных темах. Таким образом, из результатов LDA легко получить для каждого документа список встречающихся в нём тем, а для каждой темы – список характерных для неё слов, т.е. фактически описание темы.

Для тематического моделирования мы использовали программы Gensim и Mallet.

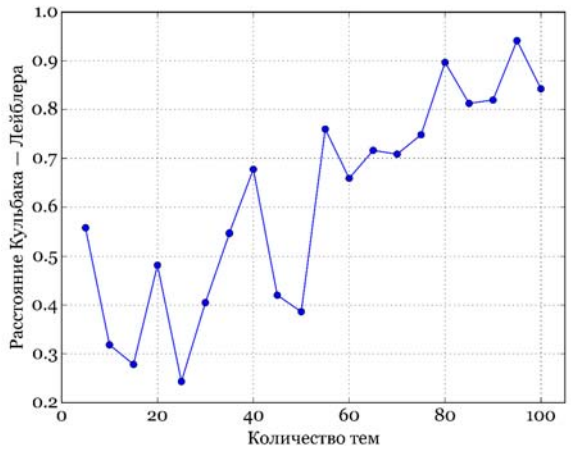
Однако прежде чем приступать к тематическому моделированию, необходимо произвести предварительную обработку данных, специфичную для данного этапа, а именно удаление редко встречающихся токенов. До обработки мы имеем 118718 уникальных токенов, что может быть причиной медленной работы алгоритма. Однако токены, встречающиеся в корпусе всего лишь один раз, не влияют на построение тематической модели, так что мы легко можем от них избавиться, сократив количество уникальных токенов до 69447. Удалённые токены представляли собой слова с ошибками, цифры, гиперссылки, английские слова (в том числе написанные транслитом), имена собственные и просто редкие слова.

Определение оптимального количества тем

Определение оптимального числа тем – важная подзадача в тематическом моделировании, поскольку её решение существенно влияет на осмысленность получаемого набора тем. Занижение числа тем приводит к чрезмерно общим результатам. Завышение приводит к невозможности разумной интерпретации. Оптимальное число тем зависит от числа документов в анализируемом корпусе: в малых корпусах оптимальным является, как правило, меньшее число тем. Согласно оригинальному исследованию, оптимальное число тем для корпуса из 16333 новостных статей составило 100, тогда как для корпуса из 5225 аннотаций научных статей – 50. Однако не существует однозначного метода определения оптимального количества тем, и часто это количество определяется «на глазок», исходя из личного мнения исследователя.

Тем не менее можно говорить о том, что самым распространённым способом является расчёт перплексии. Данная мера основана на значении правдоподобия, именно она использовалась в оригинальном исследовании [4], где впервые был представлен метод LDA.

Однако в нашем случае расчёт перплексии не дал однозначных результатов, поэтому мы использовали другую меру качества модели, основанную на расстоянии Кульбака – Лейблера [5]. Её результаты показаны на рисунке. Чем меньше указанное расстояние, тем лучше модель.



Значение меры качества на основе расстояния Кульбака – Лейблера для разного количества тем

В конечном итоге после сравнения моделей с разным количеством тем мы выбрали модель с 50 темами. Данное количество является достаточно большим, чтобы максимально полно описать все важные темы, и достаточно маленьким, чтобы этим темы могли быть отличимы друг от друга и интерпретированы.

Популярность тем в СМИ

Название темы определялись после анализа слов, которые эта тема генерирует с наибольшей вероятностью. Ниже показано, как программа описывает одну из тем. Рядом с каждым словом указана вероятность, с которой оно генерируется данной темой. Как мы можем понять, эта тема имеет отношение к прогнозу погоды в области:

*0.030*омск + 0.018*температура +*
*+ 0.017*день + 0.015*снег +*
*+ 0.014*погода + + 0.014*воздух +*
*+ 0.012*градус + 0.011*ветер +*
*+ 0.010*область + 0.010*днём +*
*+ 0.009*ожидаться + 0.009*дождь +*
*+ 0.008*ночью + 0.007*выходной +*
*+ 0.006*составить + 0.006*неделя +*
*+ 0.006*управление + 0.006*м/с +*
*+ 0.005*атмосферный + 0.005*тёплый.*

Также мы можем рассчитать вероятностное тематическое распределение для каждого отдельного документа, выявив наиболее связанные с ним темы. Так как в LDA используется нечёткая кластеризация, каждый документ с определённой вероятностью можно отнести к любой теме.

Итак, о чём пишут в омских интернет-СМИ? Для ответа на этот вопрос разумно сложить вероятности всех тем из всех документов и разделить получившееся для каждой темы на количество документов. Таким образом мы рассчитаем среднее вероятностное распределение тем на всех документах.

Получается, что пять самых популярных у редакторов интернет-СМИ тем новостей – это ДТП, преступления, пожары и другие ЧП, местная власть и экономика области. Самые непопулярные темы касаются военных учений, Крыма (хотя вроде бы связанная с ней тема Украины находится на шестом месте) и телевидения (табл. 1).

Таблица 1

Популярность тем новостей для редакторов интернет-СМИ

Место темы в рейтинге популярности	*Название темы*	*Вероятность, с которой документы можно отнести к данной теме*
1	ДТП	0.0478
2	Преступления	0.0448
3	Пожары и другие ЧП	0.0389
4	Местная власть	0.0383
5	Экономика области	0.0378
6	Украина	0.0369
…		
49	Крым	0.0032
50	Военные учения	0.0062

Таблица 2

Популярность тем новостей для читателей интернет-СМИ

Место темы в рейтинге комментируемости	*Название темы*	*Значение комментируемости, %*
1	Украина	100
2	ДТП	72
3	Контроль и регулирование на предприятиях	67
…		
48	Правоохранительные органы	9
49	Военные учения	6
50	Продажа автомобилей	5

Популярность тем у читателей

Зная количество комментариев к новостным статьям, мы можем определить самые резонансные темы, подсчитав, статьи какой тематики комментируют чаще всего, а какой – реже.

Мы можем получить показатель комментируемости темы путём сложения рассчитанных для каждого документа произведений вероятности присутствия темы в документе на количество комментариев в данном документе. Полученное таким образом значение само по себе ничего не значит, важно лишь то, как оно различается от темы к теме. Поэтому для наглядности мы можем без последствий принять наибольшее значение комментируемости за 100 %, а для остальных тем рассчитать долю, которую они составляют от этих 100 %.

Таким образом, было выявлено, что самые комментируемые темы связаны с Украиной, ДТП, контролем и регулированием на предприятиях и экономикой области. Причём интересно, что тема, связанная с событиями на Украине, почти вполовину обгоняет ближайшую конкурирующую тему. Безразличнее всего читатели отнеслись к темам о продаже автомобилей, военных учениях и правоохранительных органах (табл. 2).

Дальнейшая работа будет связана с анализом тональности комментариев и определением тем, к которым отношение наиболее положительное или отрицательное.

1. *Causey Charles*. The Battle for Bystanders: Information, Meaning Contests, and Collective Action in the Egyptian Uprising of 2011. – 2012.

2. *DiMaggio Paul, Nag Manish, Blei David.* Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding // Poetics. – 2013. – 12. – Vol. 41. – № 6. – P. 570–606.

3. Рейтинг АРИ омских интернет-СМИ. Сводка. – URL: http://omsk-journal.ru/publ/9-1-0-116 (дата обращения: 25.09.14).

4. *Blei David M., Ng Andrew Y., Jordan Michael I.* Latent Dirichlet Allocation // J. Mach. Learn. Res. – 2003. – March. – Vol. 3. – P. 993–1022.

5. On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations / R. Arun, V. Suresh, C.E. Veni Madhavan, M.N. Narasimha Murthy // Proceedings of the 14th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. – Volume Part I. – PAKDD'10. – Berlin, Heidelberg: Springer-Verlag, 2010. – P. 391–402.