Е.Д. Корнилина, А.П. Михайлов, А.П. Петров

О применении анализа данных к изучению программ политических партий

Аннотация. Работа посвящена изложению основанной на латентно-семантическом анализе методики определения близости политических позиций, содержащихся в предвыборных программах политических партий, а также других документах, публикуемых партиями для привлечения избирателей. Развивается предложенный ранее подход, в соответствии с которым близость политических позиций проявляется как синтагматическая близость текстов программ. Представлено подробное описание алгоритма, включающего в себя предварительную обработку текста, разбиение его на фрагменты, построение матрицы «фрагмент-слово», ее нормализацию, применение сингулярного разложения, и составление диаграммы сходства документов. Кратко очерчены содержательные политологические результаты, полученные в результате данного анализа.

Ключевые слова: латентно-семантический анализ, предвыборная программа партии, политическая позиция, математическая модель.

Введение

Выявление политический позиций, содержащихся в программных документах политических партий, является актуальной задачей политической науки. Результаты такого рода востребованы, в частности, при проведении политологических исследований с помощью математических методов и моделей. Приведем лишь два примера.

1. При определении индексов влияния парламентских фракций рассматриваются так называемые выигрывающие коалиции, т.е. объединения фракций, контролирующие необходимое для принятия решения количество голосов. Влияние фракции тем выше, чем больше количество выигрывающих коалиций, в которых эта фракция является ключевой (т.е коалиция перестает быть выигрывающей, если данная фракция из нее выходит). На этой идее построена теория индексов влияния (например [1]). При этом, базовые варианты индексов предполагают, что вероятности образования коалиций не зависят от политических позиций фракций. Например, коалиция левоцентристской и крайне левой партий предполагается столь же вероятной, как коалиция правоцентристской и крайне левой партий. Более современные варианты учитывают, что у каждой фракции имеются свои предпочтения по выбору партнеров для образования коалиций. Вопрос о том, какие коалиции являются более предпочтительными, чем другие, может быть решен на основании различных подходов. Одним из них опирается на анализ результатов состоявшихся ранее голосований [2], другой – на анализ политических позиций, занимаемых партиями. Чем идеологически ближе находятся две партии, тем чаще их парламентские фракции будут вступать в коалицию [3]. Измерение идеологической близости иногда проводится без какой-либо четкой методики, на основании качественных рассуждений о том, какая партия является самой леворадикальной, самой праворадикальной и т.д.

2. При построении некоторых динамических моделей политических процессов [4,5] рассматривается переход власти от одной партии к

[1]Работа выполнена при поддержке РФФИ (проект 13-01-00392)

другой, либо изменение некоторой переменной (связанной, например, с отношением бюджета к ВВП) на лево-правой шкале (в англоязычной литературе часто именуемой Rile-шкалой, от Right-Left). Для насыщения этих моделей эмпирическими данными требуется, в частности, определить взаимное расположение партий в некотором «пространстве политик», под которым может пониматься, например, евклидово пространство с измерениями, соответствующими параметрам, относительно которых правящая партия принимает решения (налоговая ставка, доля бюджета, направляемая на инвестиции в инфраструктуру и т.д.).

Выявление политических позиций партий проводится, как правило, на основании их манифестов (под которыми понимаются не только официальные предвыборные программы, но и прочие тексты, публикуемые партиями с целью привлечения избирателей) с помощью процедур, так или иначе основанных на подходе контент-анализа, и требующих привлечения экспертной оценки. Наиболее известная методика такого рода представлена в проекте «Manifesto» [6], ее основным элементом является специального вида рубрикатор. Анализируя тексты, выражающие предвыборные партийные манифесты, эксперты определяют позиции партий по различным пунктам рубрикатора, и в результате агрегирования получают для каждой из партий определенное числовое значение на Rile-шкале, имеющей пределы -100 (крайне левая партия) и 100 (крайне правая партия). Таким образом, тексты обрабатываются вручную, что неизбежно влечет наличие определенного «экспертного субъективизма».

С лингвистической точки зрения, методика проекта «Manifesto» лежит в русле так называемого парадигматического подхода, в рамках которого «семантическую» близость следует определять, опираясь на данные о значении, хранящиеся «внутри» языкового знака, а не за его пределами…» [7].

Альтернативным по отношению к парадигматическому является синтагматический подход, использующий «данные о значении слова, хранящиеся «вне» языкового знака» [7]. Можно сказать, что синтагматический подход к измерению расстояний в тексте – это понимание значения слова через контекст, в котором оно употреблено. При этом измерение расстояния происходит через сопоставление синтагматических свойств. Этот подход был реализован в методе латентно-семантического анализа (ЛСА) – запатентованный в 1988 году метод, предложенный американскими учёными [8]. ЛСА основан на «гипотезе о том, что между отдельными словами и обобщённым контекстом (предложениями, абзацами и целыми текстами), в которых они встречаются, существуют неявные (латентные) взаимосвязи, обуславливающие совокупность взаимных ограничений» [7].

Данная работа посвящена описанию методики определения близости политических позиций, выраженных в программных документах политических партий, основанной на том, что близость этих позиций понимается как синтагматическая близость соответствующих текстов. Некоторые базовые положения ее раннего варианта были кратко изложены нами в [9], а политологическое обсуждение полученных с ее помощью результатов – в работах [22, 23]. Настоящая работа и статьи [22, 23] являются взаимодополняющими в том смысле, что здесь излагается разработанная методика, с подробным описанием алгоритма, предназначенным для специалистов в области анализа данных, а статьи [22, 23] предназначены для политологической аудитории.

1. Описание методики

Исходное предположение заключается в том, что взаимосвязи между словами и контекстом являются различными в текстах, выражающих различные политические позиции. Так, нетрудно представить себе, что фамилия политического деятеля употребляется, как правило, в положительном контексте его сторонниками, и в негативном – противниками. Это относится не только к фамилиям и названиям партий, но также к отдельным политическим событиям, проектам и т.д. Тем самым, тексты, выражающие политические позиции, могут быть классифицированы путем выделения контекста, в который эти тексты погружают отдельные слова.

Контекст, в понимании ЛСА – это слова, близкие к данному слову по фактическому расположению в тексте. Более конкретно: исследуемый текст в целях проведения анализа нарезается на фрагменты (синтагмы), и контекст образуется всеми словами, входящими в один фрагмент с данным. Далее, ЛСА устанавливает

для каждых двух фрагментов меру их близости, которая в работе называется синтагматической близостью. Более подробно о синтагматическом подходе к измерению семантических расстояний в тексте и между тестами [7].

Основой для применения метода латентно-семантического анализа к данному кругу задач является выдвинутая в [9] гипотеза: близость политических позиций проявляется в синтагматической близости текстов (фрагментов, образующих тексты), выражающих эти позиции. Эта гипотеза была заложена в основу предлагаемой методики и реализована нами в виде программного комплекса.

Программные реализации по обработке текстов на иностранных языках (например, реализации ЛСА [10,11]; кластеризация контекстов [12]) в данном случае не могут быть применены в чистом виде, т. к. не учитывают особенностей русского языка (например, морфологию).

Для русскоязычных текстов примеров решения подобных задач существенно меньше, отметим некоторые из них. Semanticanalyzer [13] позволяет делать морфологический анализ и выделяет семантические облака документов (наборы смысловых слов). Автоматическая классификация лексики [14] представляет собой синтез ЛСА с алгоритмами кластеризации и предназначен для автоматической кластеризации слов. Варианты ЛСА – вероятностный, инкрементальный и иерархический, применяются для автоматического прогнозирования пользовательских интересов исходя из накопленной информации о вкусах и интересах пользователей [15, 16]. Также различные разработки по кластеризации слов, выделению смысловой части текста и т. д. совместно с российскими коллегами ведёт словацкий Social Network Research Center [17].

Основная суть реализованного в настоящей работе алгоритма состоит в следующем. Анализируемый текст, представляющий собой, в общем случае, неразмеченную совокупность текстов различных авторов, подвергается предварительной обработке.

Предварительная обработка текста заключается в:

- удалении семантически не нагруженных слов (стоп-слов);
- удалении слов, встречающихся по одному разу;
- лемматизации [18].

Список стоп-слов был построен с помощью выделенных по тематике текстов из национального корпуса русского языка [19]. Использование стандартных списков (как, например, 350 слов [7]) приводило к зашумлённым результатам.

После предобработки текст разбивается на небольшие (150–200 слов) фрагменты одинаковой длины, размер фрагмента (число слов) остаётся постоянным внутри одного исследования. Затем строится матрица, строки которой соответствуют словам исходного текста, а столбцы – фрагментам, таким образом, каждый элемент матрицы характеризует число вхождения встречаемых слов в каждый фрагмент. Производится энтропийная нормализация элементов матрицы:

$$a_{ij} = g_i \log\left(tf_{ij} + 1\right)$$

где a_{ij} - элементы нормализованной матрицы, tf_{ij} – число вхождений слова i во фрагмент j (term frequency),

$$g_i = 1 - \sum_n \frac{p_{ij} \log p_{ij}}{\log n}, \quad p_{ij} = \frac{tf_{ij}}{gf_i}$$

gf_i – общее количество появлений слова i во всём тексте (general frequency).

Полученная матрица разлагается по сингулярным векторам (векторам, соответствующим сингулярным числам) [20]. Такое разложение существует всегда, по теореме о сингулярном разложении: любая вещественная матрица может быть представлена в виде произведения трёх матриц: $A = U\Sigma V^T$, где U и V – ортогональные матрицы, а Σ – диагональная матрица. Столбцы матрицы U называются левыми сингулярными векторами, а столбцы матрицы V - правыми сингулярными векторами матрицы A.

По теореме о наилучшей аппроксимации с понижением ранга [21], если исходная матрица A задана сингулярным разложением вида $A = \sum_{i=1}^{r} u_i \sigma_i v_i^T$, то для любого целого $k : 1 \le k \le r$ подматрица $A = \sum_{i=1}^{k} u_i \sigma_i v_i^T$, соответствующая первым k наибольшим сингулярным числам σ_i, является наилучшим приближением исходной матрицы A с точки зрения евклидовой нормы в k-мерном подпространстве. Это свойство сингулярного разложения

является ключевым для метода латентно-семантического анализа: оно позволяет выбором глубины разложения по сингулярным векторам осуществлять переход от исходной матрицы большой размерности к матрице гораздо меньшей размерности, отражающей основную структуру исходной.

Соотношение близости между фрагментами исходного текста понимается как соотношение близости между векторами – строками этой новой матрицы; при этом будем говорить о синтагматической близости фрагментов. В свою очередь, в качестве синтагматического расстояния (меры близости r_{ij} между векторами) был выбран косинус угла между ними.

Таким образом, на вход анализирующего алгоритма подаётся единый текст, являющийся, вообще говоря, набором последовательно соединённых документов (например, предвыборных программ политических партий), которые в ходе выполнения программы разбиваются на небольшие фрагменты. На выходе получается квадратная матрица корреляции каждого фрагмента с каждым, либо каждого слова с каждым словом.

Для измерения близости не только отдельных фрагментов текста, но и целых документов (в данном случае – партийных программ) между собой, предлагается эмпирическая метрика, основанная на следующих исходных положениях. При сравнении двух документов содержательный смысл имеет корреляция между сопоставляемыми попарно фрагментами документа 1 и документа 2, в сравнении с внутренней корреляцией фрагментов документа 1 и документа 2.

Пусть к документу 1 относится фрагменты с номерами $1,..,N_1$, а к документу 2 - фрагменты $N_1+1,..,N_1+N_2$. Введём величины

$$A=\frac{\sum_{i=1}^{N_1}\sum_{j=1}^{N_1}\left(\frac{1+r_{ij}}{2}\right)+\sum_{i=N_1+1}^{N_1+N_2}\sum_{j=N_1+1}^{N_1+N_2}\left(\frac{1+r_{ij}}{2}\right)}{N_1^2+N_2^2},$$

$$B=\frac{1}{N_1N_2}\sum_{i=1}^{N_1}\sum_{j=N_1+1}^{N_1+N_2}\left(\frac{1+r_{ij}}{2}\right).$$

Мерой синтагматической близости документов будем называть отношение $R=200B/A$.

При этом значения R от 0 до 50 являются скорее гипотетическими, так как соответствуют преобладанию отрицательных значений r_{ij}. Значения R>100 также являются гипотетически возможными, это означало бы, что фрагменты программы первой партии ближе к программе второй партии, чем к своей. Для реальных текстов следует ожидать значения R от 50 для наиболее далеких программ до 100 для наиболее близких. Отметим, что попадание эмпирических значений R в данный интервал может рассматриваться как аргумент в пользу валидности методики.

В качестве примера полученных таким образом результатов приведем значения R для предвыборных партийных программ четырех российских партий на выборах в Государственную Думу 2007 г. (Табл.1.)

Табл. 1. Синтагматическая близость партийных программ 2007 года

2007	ЕР	КПРФ	ЛДПР	Яблоко
ЕР	0	99	92	87
КПРФ	99	0	95	91
ЛДПР	92	95	0	91
Яблоко	87	91	91	0

В работах [22, 23] приведены более подробные результаты анализа предвыборных программ данных партий на выборах в Государственную Думу 2007 и 2011 годов - в частности, диаграммы сходства программ и результаты иерархической кластеризации. Показано, в частности, что произошла определенная перестройка партийной системы. Именно: если в 2007 году наиболее близкими (из данных четырех партий) были программы КПРФ и ЛДПР, а наиболее далекой от трех других – программа ЕР, то в 2011 наиболее близкими стали программы КПРФ и ЕР, а наиболее далекой от остальных – программа Яблока. Упрощенно, говоря, в 2007 г. имела место консолидация «против партии власти», а в 2011 г – консолидация «против либеральной партии».

Заключение

Латентно-семантический анализ имеет многочисленные приложения, обзор которых представлял бы самостоятельную и довольно мас-

штабную задачу. Однако, работы посвященные применению ЛСА для выявления политических позиций нам неизвестны. Вероятно, отсутствие таких работ можно объяснить отсутствием некоторой исследовательской гипотезы, позволяющей связать специфику этого подхода со спецификой политических текстов. Предложенная нами гипотеза о том, что близость политических позиций проявляется в синтагматической близости текстов, выражающих эти позиции, призвана занять это место. Она позволила разработать описанную в данной статье методику, с ее помощью провести анализ предвыборных программ российских политических партий, и получить ряд содержательный выводов.

Литература

1. Aleskerov F. Power Indices Taking into Account Agents' Preferences // Mathematics and Democracy. Recent Advances in Voting Systems and Collective Choice. Berlin; Heidelberg: Springer, 2006. P. 1–18.
2. Алескеров Ф.Т., Благовещенский Н.Ю., Сатаров Г.А., Соколова А.В., Якуба В.А. Оценка влияния групп и фракций в российском парламенте (1994—2003 гг.). Препринт ГУ Высшая Школа Экономики, WP7/2003/01, Москва, 2003
3. F. Aleskerov, M. J. Holler, R. Kamalova. Power Distribution in the Weimar Reichstag in 1919–1933 : Working paper WP7/2010/08 /; The University – Higher School of Economics. – Moscow: Publishing House of the University – Higher School of Economics, 2010. – 54 p.
4. Ахременко А.С., Петров А.П. Политические институты, эффективность и депривация: математическая модель перераспределения политического влияния // Полис (Политические исследования). 2012. №6. С.81-100
5. Ахременко А.С., Петров А.П. Математическая модель перераспределения политического влияния: результаты и перспективы // Математическое моделирование социальных процессов. Вып. 15. Под ред. А.П.Михайлова – М.: МАКС Пресс, 2013. С.4-21
6. A.Volkens, O.Lacewell, P.Lehmann, S.Regel, H.Schultze, A.Werner (2011): The Manifesto Data Collection. Manifesto Project (MRG/CMP/MARPOR), Berlin: Wissenschaftszentrum Berlin fur Sozialforschung (WZB).
7. Митрофанова О.А. Семантические расстояния: проблемы и перспективы // XXXIV Международная филологическая конференция: Вып. 21. Прикладная и математическая лингвистика. СПб., 2005.
8. T. Landauer, P.W. Foltz, D. Laham Introduction to Latent Semantic Analysis. Discourse Processes 25: 259–284 (1998).
9. Корнилина Е.Д., Петров А.П. О приложении латентно-семантического анализа к исследованию текстов социально-политической тематики // Теория активных систем - 2011 (ТАС-2011). Труды международной научно-практической конференции. ИПУ РАН, Москва, 14-16 ноября 2011 г. Том 2. с.266-269.
10. http://lsa.colorado.edu. – LSA @ CU Boulder.
11. http://cran.at.r-project.org/web/packages/lsa/index.html. – Fridolin Wild LSA.
12. http://www.d.umn.edu/~tpederse/senseclusters.html. – Sense Clusters.
13. http://semanticanalyzer.info/blog/. – Semantic analyzer.
14. Mitrofanova O., Mukhin A., Panicheva P., Savitsky V. Automatic word clustering in Russian texts. – Heidelberg: Springer Berlin. – C. 85–91. – 2007.
15. Vinokourov A., Girolami M. Probabilistic Framework for the Hierarchic Organisation and Classification of Document Collections // Information Processing and Management. – 2002.
16. Лексин В.А., Воронцов К.В. Анализ клиентских сред: выявление скрытых профилей и оценивание сходства клиентов и ресурсов // ММРО-13. – М.: МАКС Пресс. – С. 488-491. – 2007.
17. http://sonet.webnode.cz/. – Social Network Research Center.
18. Сегалович И. В. Как работают поисковые системы // Мир Интернет. – 2002. – №10.
19. Martin L.W., Vanberg G., A robust transformation procedure for interpreting political text, Political Analysis 16 (1), 93-100.
20. Berry M. Large scale singular value computations // International Journal of Supercomputer Applications. — 1992. — V. 6, No. 1. — P. 13 — 49.
21. Golub G.H., van Loan C.F. Matrix computations. 3ed., JHU, 1996.
22. Петров А.П., Корнилина Е.Д. Исследование близости политических позиций методом латентно-семантического анализа // XII Международная научная конференция по проблемам развития экономики и общества. Книга 2. М.: Издательский дом высшей школы экономики. 2012. С.334-342
23. Корнилина Е.Д., Петров А.П. Латентно-семантический анализ предвыборных партийных программ на выборах в Государственную Думу 2007 и 2011 годов // М.: Вестник МГУ. Сер. 12: Политические науки, Издательство МГУ. – 2013. №2. С. 90-98.

Корнилина Елена Дмитриевна. Научный сотрудник Института прикладной математики им. М.В. Келдыша РАН: Окончила в 2009 г. ВМК МГУ им. М.В. Ломоносова. Кандидат физико-математических наук. Автор 12 научных работ. Область научных интересов: математическое моделирование в социальных науках, анализ данных. E-mail: ekornilin@gmail.com.

Михайлов Александр Петрович. Заведующий сектором ИПМ им. М.В. Келдыша РАН. Зав. лабораторией математического моделирования социальных процессов социологического ф-та МГУ им. М.В. Ломоносова. В 1974 г. окончил МФТИ. Доктор физико-математических наук, профессор. Автор более 200 научных работ. Область научных интересов: математическое моделирование в социальных науках. E-mail: apmikhailov@yandex.ru. .

Петров Александр Пхоун Чжо. Ведущий научный сотрудник ИПМ им. М.В. Келдыша РАН. В 1993 г. окончил физический факультет МГУ им. М.В. Ломоносова. Доктор физико-математических наук. Автор более 100 научных работ. Область научных интересов: математическое моделирование в социальных науках. E-mail: petrov.alexander.p@yandex.ru