В.Л. Арлазаров, Е.Л. Плискин, А.В. Соловьев

Определение и использование тематической дивергенции в сетях документов[1]

Аннотация. В статье предлагается сделать новый шаг для раскрытия междисциплинарного потенциала такого популярного в области прикладной статистики и машинного обучения подхода, как тематическое моделирование (ТМ). Мы вводим производное от ТМ понятие «тематической дивергенции» как меру своеобразия документа в сетевом окружении, и предлагаем его возможные междисциплинарные применения.

Ключевые слова: тематическое моделирование, сети документов, вероятностные методы, социальные сети, социология культуры, управление наукой.

Введение

Сети документов широко изучаются в литературе по прикладной статистике и машинному обучению и могут конструироваться из библиографических или полнотекстовых баз данных, Интернет-ресурсов, почтовых сообщений или социальных сетей. Узел сети может отображать как отдельную публикацию или сообщение, так и совокупность текстов одного автора. Связи могут порождаться цитатами, гиперссылками, либо вычисляться на основе некоторой меры близости документов. Направление связей в сетях документов не обязательно совпадает с направлением потоков информации. В сети коммуникаций связи направлены от автора к получателю, и в том же направлении передаётся информация. В сети цитирования связь обычно направлена от документа d_1, содержащего цитату, к цитируемому документу d_2, хотя документ d_2 исторически предшествует и может информационно влиять на автора и содержание документа d_1. Так же и в сети Интернет-ресурсов: ресурс, на который указывает ссылка, исторически предшествует и может влиять на содержание страницы со ссылкой. Обнаружение сообществ в различных моделях осуществляется с учётом как тематики текстов, так и данных о коммуникациях [6].

Сети документов – это общий подход, они могут конструироваться по-разному, в зависимости от решаемой задачи. В социальных сетях процессы распространения информационных влияний сложны. Такие процессы изучаются на более высоком уровне, чем уровень передачи элементарных информационных сообщений. Для получения общей картины, для поиска «центральных» или «влиятельных» узлов используют различные модели влияния [15].

Вероятностное тематическое моделирование (ТМ) представляет собой набор методов прикладной статистики и машинного обучения для обнаружения скрытой тематической структуры в больших коллекциях документов. Модель LDA (латентное размещение Дирихле) предполагает, что в коллекции документов есть скрытый набор тем, при этом количество тем задаётся заранее. Тема формально определяется как распределение по словарю термов. В литературе по ТМ обычно говорят о «словах» документа, подразумевая термы, которые могут быть как словами, так и словосочетаниями. Лидирующие в теме наиболее вероятные слова *чаще* встречаются совместно в одном и том же доку-

[1]Исследование выполнено при финансовой поддержке РФФИ (проекты № 14-29-05040-офи-м, № 14-29-05048-офи-м).

менте, чем можно было бы ожидать при случайном выборе слов из словаря, *без использования тем*. Модель LDA предполагает, что в каждом документе могут быть представлены различные темы, в определённых пропорциях. Алгоритм обучения модели LDA анализирует корпус текстов и пытается одновременно подобрать и темы, и распределение тем в документах.

Выделенные из корпуса документов темы можно использовать для построения как сети документов, так и тематической карты, узлами которой являются не документы, а темы. Тематические карты применяются для визуализации больших корпусов текстов «с высоты птичьего полёта», например, направлений исследований и связей между ними. При этом метки тем могут либо генерироваться автоматически из нескольких наиболее вероятных для каждой темы слов, либо, в более дружелюбном по отношению к пользователям подходе [1], извлекаться из результатов поиска после фильтрации документов по атрибутам и с учётом прав доступа пользователей.

При отсутствии априорных сведений о связях между документами, можно предсказывать связи документов, исходя из близости тем [2]. В некоторых моделях связи документов предполагаются известными и используются для тематического моделирования [4, 5], причём распределение тем в узлах сети сглаживается под влиянием смежных узлов, которое моделируется при помощи гармонической функции [14], отражающей усреднение окрестной тематики. Отсюда один шаг до нашей «тематической дивергенции», но в указанных работах решается традиционная для ТМ задача поиска латентных тем, наилучшим образом объясняющих наблюдаемое содержание документов.

В силу своей универсальности и расширяемости, современные способы тематического моделирования находят применение в широком спектре приложений [11-13], в том числе в социальных сетях: коллаборативная фильтрация в сервисах рекомендаций; построение тематических профилей пользователей форумов, блогов и социальных сетей для поиска тематических сообществ и определения наиболее активных их участников; анализ новостных потоков и сообщений из социальных сетей для определения актуальных событий реального мира и реакции пользователей на них. Привлекательной особенностью ТМ для самых различных приложений является то, что «данные говорят за себя»: классификация тем автоматически выводится из текстов, без опоры на априорные категории и без ручной обработки результатов человеком. Так, в эмпирическом исследовании [7] выявляемые при помощи ТМ в текстах патентов новые темы интерпретируются как «когнитивные прорывы», «точки роста». Хотя масштабные применения ТМ сдерживаются общей для всех методов машинного обучения вычислительной сложностью, но это препятствие отчасти преодолевается при помощи распараллеливания. Поскольку не всегда можно сделать обоснованную априорную оценку количества тем K, то были разработаны методы для автоматического подбора этого параметра [8], что ещё больше увеличивает сложность вычислений. В обзоре [9] можно найти как примеры приложений ТМ, так и ссылки на открытые программные решения. Вместе с тем, препятствием для освоения ТМ гуманитариями может быть сложный теоретико-вероятностный формализм, при отсутствии дружественного человеческого интерфейса с программными средствами. Другими словами, междисциплинарный барьер.

Примечательно, что через десятилетие после создания в 2003 году наиболее популярного в области ТМ метода LDA, развитию и применению которого посвящена обширная литература, появилась междисциплинарная работа [3], написанная автором этого метода Дэвидом Блеем вместе с социологами – коллегами по университету. В ней подход ТМ популяризируется для социологов культуры, и подчёркивается связь ТМ с такими центральными для социологии культуры понятиями, как фреймы (семантические контексты, вызывающие у читателя те или иные ассоциации), полисемия (многозначность слов), разноречие (наличие в тексте нескольких перспектив или стилей изложения), реляционность смыслов (смысл кроется не в словах, а в отношениях). Авторы [3] высказывают надежду на то, что ТМ может стать полезным инструментом для эмпирических исследований в области социологии культуры. В этом же русле лежит и наша работа: мы предлагаем сделать ещё один шаг для раскрытия культурного потенциала ТМ. Для этого мы вводим производное от ТМ понятие «тематической дивергенции» документа или узла сети и предлагаем его возможные междисциплинарные применения.

Определение тематической дивергенции

Документы и темы. Пусть заданы коллекция текстовых документов $D = \{d_1, d_2, \dots, d_N\}$ и предполагаемое количество K тем в этих документах. Каждый документ d_i состоит из последовательности слов $d_i = \langle w_{i1}, w_{i2}, \dots, w_{iN_i} \rangle$, где N_i есть количество слов в документе d_i, для $i \in \{1 \dots N\}$. Пусть $W = \bigcup_{i=1}^{N} f(d_i)$ есть словарь корпуса D, где функция $f(\cdot)$ принимает на вход текст документа и возвращает множество слов. Вероятностные тематические модели, такие как LDA, принимают на входе D и K и выдают на выходе две матрицы. Матрица $\theta \in \mathbb{R}^{N \times K}$ показывает распределение тем в документах, а матрица $\beta \in \mathbb{R}^{K \times |W|}$ показывает распределение слов в темах. Каждая строка этих матриц содержит распределение вероятностей. Предполагается следующий порождающий процесс. Слово w_{is} в s-й позиции документа d_i получается в два шага: сначала случайно выбирается тема k из вероятностного распределения тем θ_i для данного документа, а затем случайно выбирается слово из распределения слов β_k для данной темы. Каждое слово выбирается из одной темы, и все слова генерируются независимо одно от другого.

Усреднённое распределение тем документов. Пусть $DD \subset D$ некоторое подмножество документов. Усредним распределения тем: $\theta_k'(DD) = \left(\sum_{i \in DD} \theta_{i,k}\right) / \left|DD\right|$, для $k \in \{1 \dots K\}$. Вектор $\theta'(DD)$ характеризует распределение тем в DD и сам обладает свойствами распределения вероятностей, т.к. его компоненты неотрицательные и в сумме дают 1.

Расстояние между распределениями тем будем определять по методу Хеллинджера. Для двух распределений тем θ_1 и θ_2

$$H(\theta_1, \theta_2) = \frac{1}{\sqrt{2}} \sqrt{\sum_{k=1}^{K} \left(\sqrt{\theta_{1,k}} - \sqrt{\theta_{2,k}}\right)^2}$$

Так определённое расстояние находится в интервале от 0 до 1. Нулевое значение достигается для совпадающих распределений тем. Единичное значение достигается, если ни одна тема k не входит с положительной вероятностью в оба распределения, т.е. распределения состоят из совершенно разных тем. Расстояние Хеллинджера обладает благоприятными для визуализации тематических карт свойствами метрики [10] и нередко применяется в задачах ТМ [1, 4, 5].

Тематическая дивергенция $TD(i, DD)$ документа $d_i \in D$ от подмножества документов DD определяется как расстояние Хеллинджера между вектором вероятностей тем документа θ_i и усреднённым по подмножеству DD вектором вероятностей тем: $TD(i, DD) = H\left(\theta_i, \theta'(DD)\right)$. Эта величина показывает, насколько тематика документа отличается от тематики подмножества.

Тематическая дивергенция $TD(D1, D2)$ подмножества $D1 \subset D$ от подмножества $D2 \subset D$ определяется как расстояние Хеллинджера между усреднёнными по каждому подмножеству векторами вероятностей тем: $TD(D1, D2) = H\left(\theta'(D1), \theta'(D2)\right)$. Эта величина показывает, насколько тематика одного подмножества документов отличается от тематики другого подмножества.

Сеть документов есть граф $G = \langle D, E, W \rangle$, где D множество узлов, отождествляемых с документами; E множество рёбер $e = \langle d_i, d_j \rangle \in E$; W матрица смежности, содержащая веса бер, $w_{i,j} > 0$, если есть ребро из узла d_i в узел d_j. Вес ребра $w_{i,j}$ можно интерпретировать как меру влияния d_j на d_i. Например, если документ d_i цитирует документ d_j, то $w_{i,j} > 0$, при этом цитированный документ d_j предшествовал во времени и мог повлиять на текст документа d_i.

Окружение $N(i)$ (N от англ. neighborhood) **узла** d_i определяется как $N(i) = N_{out}(i) \cup N_{in}(i)$, где *исходящее окружение* $N_{out}(i) = \left\{d_j \middle| \langle d_i, d_j \rangle \in E\right\}$, а *входящее окружение* $N_{in}(i) = \left\{d_j \middle| \langle d_j, d_i \rangle \in E\right\}$.

Окружение $N(DD)$ **сообщества** DD определяется как $N(DD) = N_{out}(DD) \cup N_{in}(DD)$, где *исходящее окружение* $N_{out}(DD) = \left\{d_j \middle| \langle d_i, d_j \rangle \in E, d_i \in DD, d_j \notin DD\right\}$, а *входящее окружение* $N_{in}(DD) = \left\{d_j \middle| \langle d_j, d_i \rangle \in E, d_i \in DD, d_j \notin DD\right\}$.

Тематическая дивергенция $TDN(i)$ узла $d_i \in D$ в информационной сети определяется как $TDN(i) = TD\left(i, N(i)\right)$, где $N(i)$ окружение узла d_i в сети. Аналогично можно определить *исходящую* $TDN_{out}(i) = TD\left(i, N_{out}(i)\right)$ и *входящую* $TDN_{in}(i) = TD\left(i, N_{in}(i)\right)$ диверген-

цию узла. Эта величина показывает, насколько тематика узла отличается от непосредственного окружения. Дивергенция изолированного узла не имеет смысла, но можно считать её равной 0. Узел с нулевой [исходящей, входящей] дивергенцией будем называть **тематически сбалансированным [по исходящему, входящему окружению]**.

Тематическая дивергенция $TDN(DD)$ **сообщества** $DD \subset D$ определяется как $TDN(DD) = TD\big(DD, N(DD)\big)$, где $N(DD)$ окружение сообщества DD. Аналогично можно определить *исходящую* $TDN_{out}(DD)$ и в*ходящую* $TDN_{in}(DD)$ дивергенцию сообщества. Эта величина показывает, насколько тематика сообщества отличается от непосредственного окружения. Дивергенция изолированного сообщества не имеет смысла, но можно считать её равной 0. Что касается тематической балансировки сообщества, то можно рассматривать два определения: 1) все узлы сообщества имеют нулевую (локальную) дивергенцию или 2) сообщество в целом имеет нулевую дивергенцию относительно своего окружения.

Возможные применения тематической дивергенции

Тематическая дивергенция, как мера своеобразия документа в сетевом окружении, является скорее локальным, нежели глобальным параметром. Вряд ли имеет смысл интегрировать её по всей коллекции документов, это было бы подобно измерению «средней температуры по больнице». Её можно использовать двояко: для поиска локальных отличий и для поиска локально равновесных узлов и микро-групп.

- **Выбор типичного представителя**. В качестве кандидатов на роль типичного представителя сообщества имеет смысл рассматривать узлы социальной сети с малой дивергенцией. В сбалансированном состоянии тематика текстов автора определяется его связями. Такого субъекта можно выбирать для изучения общих мнений, для рекомендательных систем [16], для мониторинга распространения информации, или для предсказания реакции сообщества на внешние события [17].
- **Характеристика личности автора** [18]. Большая дивергенция узла социальной сети может наблюдаться в том случае, если автор предпочитает свои темы и не поддерживает чужие. Однако если в сообществе высоко ценится оригинальность, то некоторые авторы могут ради неё подражать непопулярным, забытым, или ложным образцам. В этих условиях более надёжной характеристикой личности может служить «поглотительная» дивергенция узла $TDN^{-}(i)$, вычисляемая только по таким темам k, доля которых в текстах данного автора *меньше* средней доли той же темы в его окружении: $\theta_{1,k} < \theta_k^{'}(N(i))$. Поглотительная дивергенция характеризует ослабление чужой тематики данным автором. Напротив, склонность автора активно продвигать оригинальную тематику может характеризоваться «производительной» дивергенцией узла $TDN^{+}(i)$, вычисляемой по таким темам k, доля которых в текстах данного автора *больше* средней доли той же темы в его окружении: $\theta_{1,k} > \theta_k^{'}(N(i))$.
- **Поиск точек роста** [19]. Микро-группа исследователей (из нескольких авторов), с высокой дивергенцией относительно окружения группы, но не изолированная, а включённая в коммуникации с окружением, может разрабатывать новую тематику, ещё не получившую распространение. Наличие активных коммуникаций с окружением группы может говорить о потенциальной возможности распространения её взглядов в будущем.
- **Мера нормального состояния** научной дисциплины по Т. Куну [20]. Нормальное состояние науки по Т. Куну характеризуется стабильными парадигмами, хорошо усвоенными и применяемыми в повседневной работе большинством авторов. Можно провести параллель между парадигмами и тематическим распределением. Выделение научных сообществ на основе тематического моделирования стало популярным в последние годы методом визуализации «карты» научных направлений [1]. Можно предположить, что в «нормальном» состоянии научной дисциплины мала средняя дивергенция узлов в её информационном сообществе. Здесь по-прежнему имеется ввиду сеть документов, а не тематическая карта. Другими словами, тематическая сбалансированность научного сообщества может коррелировать со стабильностью парадигм.
- **Сопоставление с центральностью** [21]**.** Интересной темой для эмпирического исследования могло бы стать сопоставление тематиче-

ской дивергенции с центральностью узлов в сети документов, в связи с задачей распространения информации. Можно предположить, что центральные, тематически сбалансированные узлы, склонные к ретрансляции тематики, могут быть полезны для поддержания желаемого стабильного информационного фона минимальными внешними усилиями.

Литература

1. Arun S. Maiya, Robert M. Rolfe. Topic similarity networks: visual analytics for large document sets. In Big Data 2014 IEEE International Conference, pp. 364-372.
2. Jonathan Chang, David M. Blei. Relational Topic Models for Document Networks. Annals of Applied Statistics, 2010, Vol. 4, No. 1, 124–150.
3. Paul DiMaggio, Manish Nag, David Blei. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding. Poetics 41, no. 6 (2013): 570-606.
4. Q. Mei, D. Cai, D. Zhang, C. Zhai. Topic modeling with network regularization. In WWW'08. New York, NY, USA: ACM, 2008, pp. 101–110.
5. Sun, Yizhou, et al. iTopicModel: Information network-integrated topic modeling. Ninth IEEE International Conference on Data Mining. IEEE, 2009, pp. 493-502.
6. Mrinmaya Sachan, Danish Contractor, Tanveer A. Faruquie, L. Venkata Subramaniam. Using content and interactions for discovering communities in social networks. In Proceedings of the 21st international conference on World Wide Web 2012 Apr 16 (pp. 331-340). ACM.
7. Kaplan, S. and Vakili, K., 2015. The double-edged sword of recombination in breakthrough innovation. Strategic Management Journal, 36(10), pp.1435-1457.
8. C. Wang, J. Paisley, and D. Blei. Online variational inference for the hierarchical Dirichlet process. Artificial Intelligence and Statistics, 2011.
9. А. Коршунов, А. Гомзин. Тематическое моделирование текстов на естественном языке. Труды Института системного программирования РАН 23 (2012).
10. Википедия. https://en.wikipedia.org/wiki/Hellinger_distance.
11. Lee, S., Song, J., and Kim, Y. An Empirical Comparison of Four Text Mining Methods. Journal of Computer Information Systems, (51:1), 2010, pp. 1-10.
12. D. Blei and J. Lafferty. Topic Models. In A. Srivastava and M. Sahami, editors, Text Mining: Classification, Clustering, and Applications. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, 2009.
13. Ali Daud, Juanzi Li, Lizhu Zhou, Faqir Muhammad. Knowledge discovery through directed probabilistic topic models: a survey. In Proceedings of Frontiers of Computer Science in China. 2010, 280-301. — перевод на русский К. В. Воронцов, А. В. Темлянцев и др.
14. X. Zhu, Z. Ghahramani, and J. D. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In ICML, pages 912–919, 2003.
15. Д. А. Губанов, Д. А. Новиков, А. Г. Чхартишвили, «Модели информационного влияния и информационного управления в социальных сетях», Проблемы управления, 2009, № 5, 28–35.
16. Perugini, Saverio; Gonçalves, Marcos André; and Fox, Edward A., «Recommender Systems Research: A Connection-centric Survey» (2004). Computer Science Faculty Publications. Paper 34. http://ecommons.udayton.edu/cps_fac_pub/34.
17. Lerman, K. and Hogg, T., 2010, April. Using a model of social dynamics to predict popularity of news. In Proceedings of the 19th international conference on World Wide Web (pp. 621-630). ACM.
18. Friedkin, N.E. and Johnsen, E.C., 1997. Social positions in influence networks. Social Networks, 19(3), pp.209-222.
19. Meadows, A.J. and O'Connor, J.G., 1971. Bibliographical statistics as a guide to growth points in science. Social Studies of Science, 1(1), pp.95-99.
20. Т.Кун. Структура научных революций. М.: Прогресс, 1977.
21. Fortunato, S., 2010. Community detection in graphs. Physics reports, 486(3), pp.75-174.

Арлазаров Владимир Львович. Руководитель отделения №2 ИСА РАН Центра. Окончил МГУ в 1961 году. Член-корреспондент РАН, профессор. Количество печатных работ: более 100. Область научных интересов: теория графов, распознавания образов, программирование. E-mail: arl@isa.ru

Плискин Евгений Львович. Ведущий научный сотрудник. ИСА ФИЦ ИУ РАН. Окончил МФТИ в 1982 году. Кандидат технических наук. Количество печатных работ: 17. Область научных интересов: автоматизированные информационные системы. E-mail: pliskin@isa.ru

Соловьев Александр Владимирович. Заместитель директора ИСА ФИЦ ИУ РАН**.** Окончил МГТУ им. Н.Э. Баумана в 1994 году. Доктор технических наук. Количество печатных работ: 53. Область научных интересов: системный анализ, системы управления базами данных, теория надежности, влияние человеческого фактора, математическое моделирование, электронный документооборот. E-mail: soloviev@isa.ru.

Measuring topic divergence in document networks and its possible applications

V.L. Arlazarov, E.L. Pliskin, A.V. Soloviev

Abstract. The paper proposes a new step to promote interdisciplinary propagation of such a popular in the fields of applied statistics and machine learning approach, as the topic modeling (TM). We introduce TM-based concept of «topical divergence» to account for the document (or its author) individuality with regard to its local network neighborhood. We propose several possible interdisciplinary applications of topical divergence related to sociology and management of science.

Keywords: topic modeling, document networks, probabilistic methods, social networks, sociology of culture, management of science.

Vladimir L. Arlazarov. Head of Department of the Institute for Systems Analysis (ISA), Federal Research Center «Computer Science and Control» of Russian Academy of Sciences (FRC CSC RAS). Corresponding Member of the Russian Academy of Sciences, Professor. Graduated from the Moscow State University in 1961. Number of publications: more than 100 articles and monographs. Research interests: graph theory, pattern recognition, and programming. E-mail: arl@isa.ru

Eugene L. Pliskin. Leading Researcher, FRC CSC RAS. Ph.D. Graduated from the Moscow Institute of Physics and Technology (MIPT) in 1982. Number of publications: 17. Research interests: automated information systems. E-mail: pliskin@isa.ru

Alexander V. Soloviev. Deputy Director of the ISA, FRC CSC RAS. Dts. Graduated from Bauman Moscow State Technology University (MSTU) in 1994. Number of publications: 53 Research interests: system analysis, database management systems, reliability theory, the human factor, mathematical modeling, and electronic documents flow. E-mail: soloviev@isa.ru