Социальные алгоритмы онлайн-сообществ: аналитический обзор

Статья рекомендована И.Ю. Алексеевой 11.06.2019.

СУВОРОВА Алена Владимировна
Кандидат физико-математических наук, доцент департамента информатики, Национальный исследовательский университет «Высшая школа экономики», Санкт-Петербург

БАХИТОВА Алина Асылановна
Преподаватель департамента информатики, НИУ ВШЭ, Санкт-Петербург; стажер-исследователь Научно-учебной лаборатории «Социология образования и науки», НИУ ВШЭ, Санкт-Петербург

КУЗНЕЦОВА Анастасия Дмитриевна
Преподаватель департамента информатики, Национальный исследовательский университет «Высшая школа экономики», Санкт-Петербург

ГУЛЯЕВ Павел Романович
Преподаватель департамента информатики, Национальный исследовательский университет «Высшая школа экономики», Санкт-Петербург

Аннотация
Онлайн-сообщества и социальные медиа являются важной системообразующей частью информационного общества. В последнее время все большее внимание приковано к служащим основой формирования таких сообществ алгоритмам и к их свойству усиления эффектов социальных процессов, в первую очередь, негативных, приводящих к усилению неравенства и дискриминации. В настоящей статье приведен обзор классов социальных алгоритмов и их исследований. Делается вывод о необходимости алгоритмической культуры как составляющей части цифровой грамотности. Даются рекомендации по применению инструментов интерпретации алгоритмов машинного обучения в формировании алгоритмической культуры.

Ключевые слова:
социальные алгоритмы, рейтинги, рекомендательные системы, машинное обучение.

Концепция социальных алгоритмов

Социальные алгоритмы [1], в том числе рейтинги, являются инструментами, облегчающими процесс принятия решения во множестве контекстов [2,3]. В последнее время в ряде исследований поднимаются вопросы о влиянии социальных алгоритмов (ранжирования, рекомендаций, взаимного оценивания, отзывов, формирования репутации) на социальные системы.

Рост интереса к изучению алгоритмов ранжирования и механизмов репутации отражается в нарастающей дискуссии между учеными в области социальных наук [1,4]. Обсуждение данных вопросов стимулируется социальными эффектами алгоритмов, которые наблюдаются в сообществах и имеют видимое влияние на участников. Подобные алгоритмы, с одной стороны, ориентированы на отражение реальных оценок репутации, вклада (например, для репутационных и коллаборативных систем), значимости или предпочтений (для формирования рекомендаций). С другой стороны, широкое внедрение таких алгоритмов

приводит к изменению реальности, в частности, формированию поведения, нацеленного именно на получение высоких оценок, или неявного ограничения доступности источников информации из-за особенностей их ранжирования. Как следствие, конструктивное исследование подобных алгоритмов, включая проведение эмпирических исследований, имеет значимые научные перспективы, как с точки зрения оценивания влияния алгоритмов на реальность, так и с точки зрения развития этих алгоритмов при учете выстраивающихся вокруг них социальных механизмов и практик. Часто такие исследования посвящены этическим проблемам, связанным с подобным влиянием («алгоритмическая дискриминация») [5].

Рассмотрим классы систем, связанных с концепцией алгоритмического управления (algorithmic governance).

Коллаборативные системы

Основой современного социального компьютинга являются системы, организованные для обеспечения совместной работы людей, они все чаще становятся объектом исследований, производимых в рамках изучения человеко-компьютерного взаимодействия. Исследование результатов совместного человеческого производства внутри виртуального пространства является важным инструментом, с помощью которого ученые могут выявлять специфические характеристики и паттерны реального мира. Так, в совместной работе британских и испанских исследователей [6] выявление районов города с повышенным уровнем депривации осуществляется при помощи анализа данных о контенте, производимом людьми, живущими в том или ином районе города.

Также изучение поведения людей внутри коллаборативных систем позволяет выявить их структуру и механизмы, формирующие процессы, связанные с взаимодействием людей. Большое количество современных работ по данному направлению исследуют подобные интеракции людей на основе одной из основных виртуальных систем, основанной на совместном труде, — электронной энциклопедии Википедии. Киттур, Су, Пендлтон, Чи (Kittur, Suh, Pendleton, Chi) [7] обнаружили, что взаимодействие людей, занимающихся производством и редактурой материала для электронной энциклопедии, может порождать как элементы координации, так и развитие конфликтов. Исследователи разрабатывают модель, которая дает им возможность на основе материала, создаваемого посредством коллективного труда, предсказывать образование конфликтов между производителями материала. Ученые также утверждают, что данная модель может быть применима для изучения паттернов взаимодействия внутри коллективов в более структурно сложных системах.

Коллаборативные системы также могут служить фундаментом для организации коллективного действия в реальном мире. Асад (Asad) и Ле Данде (Le Dandec) [8] из Технологического института Джорджии (Georgia Tech) обнаружили, что интеракции, образованные внутри виртуальных систем, напрямую влияют на повышение среди участников систем гражданского самосознания и более сильную их включенность в деятельность локальных сообществ.

Управление на основе социальных алгоритмов [9] является важной чертой коллаборативных систем, в частности, авторы [10] описывают возникновение

такого управления в Википедии как постепенный процесс превращения социальных механизмов в алгоритмические. Этот процесс, однако, не всегда успешен. Например, в [11] проанализирован неудачный опыт алгоритмического управления в децентрализованной системе. Одним из ключевых видов алгоритмического управления в коллаборативных системах являются репутационные системы.

Репутационные системы

Благодаря развитию интернета взаимоотношения между незнакомцами стали очень распространены: это и покупка-продажа товаров, аренда жилья, сервисы такси, поиск ответа на специализированных форумах и т.д. Необходимость взаимодействия, причем чаще всего дистанционного, с теми, кого мы не знаем, порождает ряд проблем, связанных с доверием. Именно поэтому крупные интернет-сервисы заинтересованы в создании репутационных систем [3]. Их цель заключается в ответе на вопрос: можно ли верить данному человеку или нет? Например, пользуясь E-Bay или Amazon, можно оценивать продавцов, на популярных сервисах вопросов и ответов обычно можно ставить оценки как ответам, так и вопросам, а ориентируясь на оценки и комментарии других пользователей, можно делать предварительные выводы об адекватности ответов, добросовестности продавцов или качестве их клиентского сервиса.

Чтобы репутационная система функционировала, в ней должны присутствовать «долгоживущие» объекты и субъекты с хорошей репутацией, вдохновляющие новых пользователей на «правильные» поступки. Также репутационная система должна фиксировать все действия достаточно детально и полно, чтобы помогать людям выбирать, кому они могут довериться [12].

Репутационные системы привлекают интерес исследователей из компьютерных и социальных наук, но их прикладная важность ничуть не меньше.

Одной из важных практических работ в этой области является платформа «Building Web Reputations Systems» [13], где описаны практические рекомендации по построению репутационных систем онлайн-сообществ. Авторы часто ссылаются на реальные примеры, подчеркивая отсутствие универсального рецепта эффективной репутации и очерчивая диапазон от систем, где эффективным будет только учет зафиксированных поступков пользователя, до тех, что основываются исключительно на оценках других пользователей. Хотя скорее всего для эффективной репутационной системы придётся сделать комбинированный алгоритм. Ванг (Wang) и Васильева (Vassileva) проанализировали репутационные системы, составили их классификацию и выдвинули идеи для улучшения различных алгоритмов [14]. Одним из важных предложений было делать репутационные системы более децентрализованными: по их классификации децентрализация означает, что репутацию человека формирует всё сообщество. Репутационные системы могут быть глобальными, когда пользователь получает общую оценку других, или персонализированными, где пользователь может узнать, как оценен человек какой-то определённой группой. Исследование того, какой из вариантов будет более правильным или полезным, актуально и для «реального», офлайн-мира.

Изучаются в этом контексте и *рекомендательные системы*, в которых поставленная пользователем оценка (или отношение, выраженное другим способом,

например просмотром товара) используется не только для характеристики оцененного объекта (товара, ответа, другого пользователя), но и для предложения этому пользователю других объектов (схожих товаров, мест посещения, тем обсуждений). Подобные исследования зачастую основываются на применении двух важных психологических и социологических конструктов — индивидуальных предпочтений и социального влияния. Для создания рекомендательных систем используется некоторое количество техник, которые можно формально разделить на методы, опирающиеся на контент, методы коллаборативной фильтрации и гибридные [15]. «Контентные» методы рекомендации исторически связаны с задачами информационного поиска [16] и предлагают пользователям контент, похожий на тот, что они уже предпочли в прошлом. Тогда как методы коллаборативной фильтрации предсказывают пользовательские интересы через раскрытие сложных и нетривиальных паттернов по его предыдущему поведению и предлагают пользователю то, что выбирали другие пользователи с похожими на его предпочтениями и интересами [17,18].

Возрастающий интерес приобретают и *рейтинговые системы с географической привязкой*. Появление в наборе данных отсылки к размещению того или иного объекта в системе географических координат приводит к появлению новых возможностей для аналитики, которые зачастую были бы недоступны ранее. С помощью таких данных можно строить сложные модели, учитывающие пространственное распределение исследуемых объектов и свойств окружающего их пространства [19]. При этом географические данные могут быть использованы в работе с большим спектром прикладных исследовательских задач, включая, например, изучение влияния пространственных факторов на выбор школы [20] и определение мест концентрации криминальной активности [21].

Работа с пространственными данными, не будучи методологическим направлением сама по себе, подразумевает возможность применения широкого набора инструментов — от анализа социальных сетей до продвинутых методов машинного обучения — оставляя при этом простор для разработки собственного методологического инструмента. На настоящий момент существует целое направление, объединяющее различные математические методы анализа данных, разработанные специально для работы с учетом географической привязки, например, различные метрики центральности узлов, подходящие для решения задач, в которых приемлема проекция географической составляющей в виде сети [22-24].

Рейтинговые механизмы на мезо- и макроуровне

В социальных науках классической работой, изучающей влияние систем ранжирования на оцениваемый объект, стала работа Эспеланд (Espeland) и Сауде (Sauder) «Rankings and Reactivity», рассматривающая рейтинги юридических вузов [25]. Исследователи описывают, с одной стороны то, как рейтинги влияют на восприятие публикой «топовых» школ, с другой стороны, как меняется административная политика юридических вузов после публикации рейтинга. В другой работе авторы описывают более глобальные последствия от публикаций все большего числа рейтингов университетов, в частности, изменение и появление новых образовательных стандартов, повышенный запрос на прозрачность деятельности [26].

Другими словами, сначала рейтинги составляются для оценки деятельности, а затем возникают предпосылки для выстраивания механизмов деятельности для достижения более высоких позиций в рейтинге.

Помимо исследований действия социальных эффектов, перспективным направлением является сравнение алгоритмов ранжирования per se [1]. Это позволяет обратить внимание на те части объекта, где результаты ранжирования значительно расходятся, что, зачастую, говорит о сложной природе явления. Таким образом, методологическая постановка проблемы может быть плодотворной для получения содержательных инсайтов об исследуемом объекте. Так было показано, что большинство рейтингов присваивают одинаковые позиции европейским университетам, а наибольший «уровень несогласия» достигается при оценке университетов в Азии [27], что говорит, в частности, о фокусировании рейтингов на различных аспектах деятельности, которые по-разному выражены в университетах различных типов. Близкой к вышеупомянутой проблеме является разработка устойчивых алгоритмов ранжирования [28]. Так публикация в новом журнале, как правило, привлекает высококвалифицированных специалистов, таким образом приводя к сильному смещению результатов ранжирования. Авторами рассматриваются методы для снижения воздействия данного эффекта на результаты ранжирования университетов, основанного на библиографических данных. Более того, важной частью является обобщение результатов ранее опубликованных исследований, использующих рейтинговые механизмы. Так для сравнения влияния лучших университетов на научную продуктивность сотрудников, перед исследователями встала задача, соответственно, определения лучших университетов на основе результатов ранжирований из предыдущих работ [29].

Алгоритмическое управление и дискриминация

Последней, привлекающей все большее внимание, темой, связанной с социальными алгоритмами онлайн-сообществ, является тема этических аспектов и негативных эффектов применения социальных алгоритмов, в первую очередь, в их влиянии на социальные процессы («алгоритмическая дискриминация» [5]). Растущая важность этой темы в научной и популярной дискуссии свидетельствует о росте места, занимаемого социальными алгоритмами в повседневности информационного общества.

Одной из важных точек в популяризации проблемы стала предложенная Паризером концепция «пузыря фильтров» (filter bubble) [30], описывающая побочный эффект алгоритмов рекомендательных систем, применяемых в большинстве социальных медиа. Трактовка материалов как «неинтересных» пользователю с точки зрения рекомендательных алгоритмов с соответствующим понижением вероятности их попадания в фокус внимания пользователя приводит к сокращению его доступности для альтернативных точек зрения, и как следствие, к поляризации мнений. Паризер уделяет особое внимание Фейсбуку, играющему важную роль в американской политике, так как следствием пузыря фильтров и поляризации мнений может быть и поляризация политическая. Как показали впоследствии исследователи из Фейсбук [4], индивидуальный выбор пользователя играет

[1] «как есть», «самих по себе»

большую роль по сравнению с алгоритмами алгоритмического ранжирования, однако и их роль нельзя недооценивать.

«Пузырь фильтров», однако, не является единственной проблемой, связанной с социальными алгоритмами. Автоматизированное принятие решений на основе алгоритмов больших данных имеет в качестве побочного эффекта алгоритмическую дискриминацию [31, 32], в которой косвенные сигналы о характеристиках пользователей, вкусах, вехах жизненного пути и т.д., содержащиеся в больших данных, в том числе социальных, могут усиливаться алгоритмически и приводить к отрицательным последствиям в виде дискриминирующих решений для не вписывающихся в «типовой» паттерн, или даже попросту тех пользователей, о которых отсутствуют некоторые данные [33]. Более того, не только атрибуты пользователя, но и его позиция в сетях социальных связей может стать причиной алгоритмической дискриминации [34]. Причиной этой проблемы часто выступают существующие в социальных системах искажения, попадающие «на вход» алгоритмов в качестве обучающих выборок [35], целевой функционал алгоритмов машинного обучения, призванных отделять «сигнал» в данных от «шума» и усиливать его, отсеивая таким образом «нестандартные» случаи, а также возможное злоупотребление алгоритмами [36]. Поиск подходов к решению этой проблемы является важной междисциплинарной задачей. Ученые предлагают как технические [37], так и организационно-политические решения, выдвигающие требования к создателям и операторам алгоритмических систем в части прозрачности алгоритмов [38]. К последнему направлению относится и разработка инструментов интерпретации алгоритмов машинного обучения, позволяющих на основе модели типа «черный ящик» (т.е. с неизвестными правилами вывода) создавать интерпретируемые правила [39]. Применение подобных инструментов приводит к раннему выявлению дискриминации [40] (например, если вывод был сделан из-за отсутствия той или иной информации) и повышению доверия к системам вследствие их большей понятности [41].

Заключение

Все более широкое проникновение различных информационных систем в повседневность приводит к тому, что алгоритмы, использующиеся в подобных системах, начинают сильнее влиять на решения и действия, в том числе, за пределами этих информационных систем. В дополнение к исследованию подобных алгоритмов и формируемых ими сообществ как модельного примера с относительно доступными данными для обоснования выводов о схожих сообществах в «реальном» мире, сейчас становятся актуальными исследования, направленные на изучение влияния, которое социальные алгоритмы оказывают на общество. Необходимость выявления различных побочных эффектов обусловлена в том числе и тем, что знание о них позволит выработать меры по снижению подобных эффектов: от алгоритмических (антидискриминационные алгоритмы) до образовательных (знание об ограничениях как стимул для рассмотрения альтернативных вариантов).

Статья подготовлена в ходе работы (проект № 17-05-0024, Научно-учебная группа «Машинное обучение и социальный компьютинг») в рамках Программы «Научный

фонд Национального исследовательского университета „Высшая школа экономики" (НИУ ВШЭ)» в 2017 – 2018 гг. и в рамках государственной поддержки ведущих университетов Российской Федерации «5-100».

ЛИТЕРАТУРА

1. LAZER D. **The rise of the social algorithm** // Science. 2015. Vol. 348, № 6239

2. BAROCAS S., HOOD S., ZIEWITZ M. **Governing Algorithms: A Provocation Piece** // SSRN Electronic Journal. 2013. P. 1-12.

3. LAMPE C. **The role of reputation systems in managing online communities** // The reputation society. How online opinions are reshaping the offline world. 2012. P. 77-88.

4. BAKSHY E., MESSING S., ADAMIC L. A. **Exposure to ideologically diverse news and opinion on Facebook** // Science. 2015. Vol. 348, № 6239. P. 1130-1132.

5. WINTER J. **Algorithmic Discrimination: Big Data Analytics and the Future of the Internet** // The Future Internet: Alternative Visions / ed. Winter J., Ono R. Cham: Springer International Publishing, 2015. P. 125-140.

6. VENERANDI A. ET AL. **Measuring urban deprivation from user generated content** // Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing. ACM, 2015. P. 254-264.

7. KITTUR A. ET AL. **He says, she says: conflict and coordination in Wikipedia** // Proceedings of the SIGCHI conference on Human factors in computing systems. ACM, 2007. P. 453-462.

8. ASAD M., LE DANTEC C. A. **Illegitimate civic participation: supporting community activists on the ground** // Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing. ACM, 2015. P. 1694-1703.

9. PITT J. ET AL. **Collective intelligence and algorithmic governance of socio-technical systems** // Social Collective Intelligence. Springer, 2014. P. 31-50.

10. MÜLLER-BIRN C., DOBUSCH L., HERBSLEB J. D. **Work-to-rule: The Emergence of Algorithmic Governance in Wikipedia** // Proceedings of the 6th International Conference on Communities and Technologies. New York, NY, USA: ACM, 2013. P. 80-89.

11. DUPONT Q. **Experiments in algorithmic governance: A history and ethnography of «The DAO,» a failed decentralized autonomous organization** // Bitcoin and Beyond. Routledge, 2017. P. 157-177.

12. RESNICK P. ET AL. **Reputation systems** // Communications of the ACM. 2000. Vol. 43, № 12. P. 45-48.

13. FARMER R., GLASS B. **Building Web Reputation Systems.** O'Reilly Media, Inc, 2010.

14. WANG Y., VASSILEVA J. **Trust and reputation model in peer-to-peer networks** // Peer-to-Peer Computing, 2003.(P2P 2003). Proceedings. Third International Conference on. IEEE, 2003. P. 150-157.

15. ADOMAVICIUS G., TUZHILIN A **Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions** // IEEE Transactions on Knowledge & Data Engineering. 2005. № 6. P. 734-749.

16. BAEZA-YATES R., RIBEIRO-NETO B. **Modern information retrieval.** ACM press New York, 1999. Vol. 463.

17. KOREN Y. **Factorization meets the neighborhood: a multifaceted collaborative filtering model** // Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2008. P. 426-434.

18. SU X., KHOSHGOFTAAR T. M. **A survey of collaborative filtering techniques** // Advances in artificial intelligence. 2009. Vol. 2009. P. 1-19.

19. LARSON R.R., **Frontiera P. Spatial ranking methods for geographic information retrieval (GIR) in digital libraries** // International Conference on Theory and Practice of Digital Libraries. Springer, 2004. P. 45-56.

20. HENIG J. R. **Geo-spatial analyses and school choice research** // American Journal of Education. 2009. Vol. 115, № 4. P. 649-657.

21. RATCLIFFE J.H., MCCULLAGH M. J. **Hotbeds of crime and the search for spatial accuracy** // Journal of geographical systems. 1999. Vol. 1, № 4. P. 385-398.

22. MIAOU S. – P., SONG J. J. **Bayesian ranking of sites for engineering safety improvements: decision parameter, treatability concept, statistical criterion, and spatial dependence** // Accident Analysis & Prevention. 2005. Vol. 37, № 4. P. 699-720.

23. PÁEZ A., SCOTT D. M. **Spatial statistics for urban analysis: a review of techniques with examples** // GeoJournal. 2005. Vol. 61, № 1. P. 53-67.

24. WEN T. – H. **Geographically modified PageRank algorithms: Identifying the spatial concentration of human movement in a geospatial network** // PloS one. 2015. Vol. 10, № 10.

25. ESPELAND W.N., SAUDER M. **Rankings and reactivity: How public measures recreate social worlds1** // American journal of sociology. 2007. Vol. 113, № 1. P. 1-40.

26. MARGINSON S., VAN DER WENDE M. **To rank or to be ranked: The impact of global rankings in higher education** // Journal of studies in international education. 2007. Vol. 11, № 3-4. P. 306-329.

27. AGUILLO I. ET AL. **Comparing university rankings** // Scientometrics. 2010. Vol. 85, № 1. P. 243-256.

28. CROOK M.D., WALKUP B. R. **Rankings and Trends in Finance Publishing: An Iterative Approach** // Journal of Financial Research. 2016. Vol. 39, № 3. P. 291-322.

29. KIM E.H., MORSE A., ZINGALES L. **Are elite universities losing their competitive edge?** // Journal of Financial Economics. 2009. Vol. 93, № 3. P. 353-381.

30. PARISER E. **The filter bubble: What the Internet is hiding from you.** Penguin UK, 2011.

31. BAROCAS S., SELBST A. D. **Big data's disparate impact** // Cal. L. Rev. 2016. Vol. 104. P. 671.

32. ROMEI A., RUGGIERI S. **A multidisciplinary survey on discrimination analysis** // The Knowledge Engineering Review. 2014. Vol. 29, № 5. P. 582-638.

33. WILLIAMS B.A., BROOKS C. F., SHMARGAD Y. **How Algorithms Discriminate Based on Data they Lack: Challenges, Solutions, and Policy Implications** // Journal of Information Policy. 2018. Vol. 8. P. 78-115.

34. BOYD D., LEVY K., MARWICK A. **The networked nature of algorithmic discrimination** // Data and Discrimination: Collected Essays. Open Technology Institute. 2014. P. 53-57.

35. CALISKAN A., BRYSON J. J., NARAYANAN A. **Semantics derived automatically from language corpora contain human-like biases** // Science. 2017. Vol. 356, № 6334. P. 183-186.

36. BRUNDAGE M. ET AL. **The malicious use of artificial intelligence: Forecasting, prevention, and mitigation** // arXiv preprint arXiv:1802.07228. 2018.

37. HAJIAN S., BONCHI F., CASTILLO C. **Algorithmic bias: From discrimination discovery to fairness-aware data mining** // Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2016. P. 2125-2126.

38. SANDVIG C. ET AL. **Auditing algorithms: Research methods for detecting discrimination on internet platforms** // Data and discrimination: converting critical concerns into productive inquiry. 2014. P. 1-23.

39. SAMEK W., WIEGAND T., MÜLLER K. R. **Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models** // arXiv preprint arXiv:1708.08296. 2017.

40. KILBERTUS N., CARULLA M. R., PARASCANDOLO G., HARDT M., JANZING D., SCHÖLKOPF B. **Avoiding discrimination through causal reasoning** // Advances in Neural Information Processing Systems. 2017. P. 656-666.

41. YU K., BERKOVSKY S., CONWAY D., TAIB R., ZHOU J., CHEN, F. **Do I Trust a Machine? Differences in User Trust Based on System Performance** // Human and Machine Learning. 2018. P. 245-264