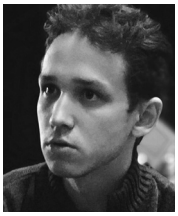
“Цифровые следы” и измерения ценности особенных благ

Статья рекомендована И.Ю. Алексеевой 17.09.2019.



БУЛЫГИН Денис Игоревич
Преподаватель департамента информатики факультета «Санкт-Петербургская школа физико-математических и компьютерных наук», Национальный исследовательский университет «Высшая школа экономики»

МУСАБИРОВ Илья Леонидович
Старший преподаватель департамента информатики факультета «Санкт-Петербургская школа физико-математических и компьютерных наук» Национального исследовательского университета «Высшая школа экономики»

Аннотация
Рынок опытных благ в рамках информационного общества активно растет. При этом, несмотря на существующую дискуссию в области экономической социологии и растущий объем данных, наблюдается недостаток эмпирических работ, изучающих механизмы оценки таких товаров. В статье обсуждаются “цифровые следы” выбора и оценки опытных благ (на примере достопримечательностей, отелей и брендов), формируемые внутри интернет-сервисов, на которых потребители выражают мнение о благах и связанным с ними опыте. Описаны наиболее популярные методы анализа суждений людей, приведены примеры их использования, выделены основные достоинства и недостатки.

Ключевые слова:
цифровые следы, особенные блага, измерение ценности особенных благ, анализ текстовых данных, частотный анализ текстов, нетнография.

Введение

Развитие цифрового общества сопровождается ростом рыночной роли товаров и услуг, направленных на опыт потребителей. Например, в 2019 году продажи на рынках музыки и киноиндустрии достигли 20 и 42 миллиарда долларов, и к 2021 году ожидается их рост на 10%. Объем рынка игр превосходит и кино- и музыкальную индустрию и составляет 152 миллиарда долларов. Близок к нему по объему рынок отельного бизнеса, оборот которого в 2018 году оценивался в 147 миллиардов долларов, и, согласно прогнозам, увеличится до 211 миллиардов к 2026 году. Несмотря на то, что представленные рынки распространяют разную продукцию, их всех можно отнести к опытным благам (experience goods) с точки зрения экономики и к особенным благам (singularities в терминах Л. Карпика) с точки зрения социологического анализа формирования ценности.

Механизмы выбора и оценивания таких благ являются предметом активного обсуждения в рамках социологии формирования ценности [1-3]. Классическим примером особенных благ является рынок вин [4], в котором цена вина формируется не только с точки зрения классических экономических механизмов, но и под влиянием престижности, географии производства вина, года его выпуска и других символических характеристик.

Особенно ярко выражена необходимость фокуса на социальном процессе конструирования ценности товаров в ситуации, когда функциональная ценность товаров не ясна или мала, например, на рынках искусства, антиквариата, вина,

моделей и т.п. В последнее время на стыке с исследованиями формирования ценностей в экономической социологии развивается область оценочных исследований (valuation studies) [1], которая фокусируется на социально-психологических аспектах потребления товаров и опыта, используя экономико-социологический подход к объяснению оценки и формирования ценности товаров.

Люсьен Карпик, анализируя процессы формирования ценности, предлагает концепцию "особенных благ" (singularities) [2,5], у которых нет единой шкалы определения ценности (и цены) товара. При этом в переплетении социальных факторов формирования ценности порождается несоизмеримость ценности товаров, которая требует от потребителей прибегать к инструментам оценочных суждений (judgement devices) [2,5]: персональным и неперсональным сетям, экспертам (гайдам, рецензиям, спискам топ-10), рейтингам и брендам [2].

При этом появляются работы, которые организованы в схожей перспективе и пытаются выделить компоненты формирования ценности "особенных благ". Такие исследования изучают формирование ценности опосредованно через различные рейтинги и опросные данные. Главное ограничение этого подхода заключается в том, что исследователь изучает уже известные измерения опыта, потенциально игнорируя другие, пока неоткрытые аспекты.

Многообещающим решением этой проблемы является использование интернет-данных, которые играют всё большую роль в социальных науках и получают признание в академической среде [6-8]. Особенно можно отметить текстовые данные, содержащие рефлексию пользователей о пережитом опыте.

С помощью сервисов с отзывами люди делятся своими ощущениями от приобретенного и пережитого опыта или продукта. Например, пользователи сайта TripAdvisor оставляют отзывы об отелях, в которых они останавливались. Такие истории о пережитом опыте могут быть разделены на компоненты, например, качестве сервиса, дизайне номеров и дополнительных услугах, которые выражают ценность товара или услуги в разных её аспектах.

Анализ текстовых данных является альтернативой опросным и дневниковым техникам, широко применяемых в социологии [9] или данным, основанным на автоматизированно собираемых журналах действий или сетях дружбы [10]. Фокус на текстах в данном случае позволяет анализировать восприятие товаров и опытов в интернет-сообществах и прослеживать их трансформацию, что может быть затруднено при использовании более традиционных техник. Преимущество таких методов перед, например, рейтингами и опросами заключается в том, что исследователь не ограничивает людей заранее заданной схемой аспектов опыта, как это делают опросы, ограниченные существующими теоретическими представлениями, принятыми исследователем. Например, анализируя отзывы на отели, авторы [11] выявили аспекты, ранее не поднимаемые в теоретических и эмпирических статьях.

Методы анализа «цифровых следов»

Проблемой текстового анализа является невозможность или затрудненность непосредственного статистического анализа содержания текста [12]. Поэтому для вычислительного текстового анализа необходимо трансформировать текст в числовые

характеристики, пригодные для статистического анализа. В рамках статьи мы описываем несколько базовых методов вычислительного анализа текстов: частотный анализ слов и последовательностей слов (n-грамм), оценка эмоциональной окраски текста, семантические сети, а также более вычислительно сложный метод — тематическое моделирование. Следует отметить, что на текущий момент существуют и более вычислительно сложные методы анализа текстов, которые могут иметь применение в этой области, в частности некоторые методы суммаризации текстов, и методы на основе дистрибутивно-семантических моделей, которые не рассмотрены в настоящей статье.

Частотный анализ текстов

Базовым способом трансформации текста в данные является частотный анализ, при котором последовательность слов в текстовом документе и их лексическое значение игнорируется, а остается только информация о том, как часто каждое слово встречается в тексте [13], то есть тексты рассматриваются как "мешки слов" ("bags-of-words").

В такой ситуации исследователь сталкивается с небольшим набором слов (50-100 слов), которые встречаются в нашей речи повсеместно, и большим набором слов, которые встречаются гораздо реже, с общим распределением, подчиняющимся закону Ципфа [14]. Но в зависимости от цели текста, выбранной темы, бэкграунда автора и других факторов частоты некоторых слов будут систематически отличаться [15].

Частотный анализ может выявить слова, характерные для всего текстового корпуса (например, какими словами описывают номера отелей), но наибольшую пользу он приносит при сравнении нескольких корпусов либо выделении разных групп слов.

В первом случае один и тот же набор слов выделяется из двух корпусов (например, корпуса отзывов отелей с двумя и четырьмя звездами). При этом разница может считаться как в абсолютных величинах, так и по специальным метрикам, основанным на частотах слов, таких как TF-IDF или отношения правдоподобия.

Во втором случае составляются словари с разными наборами слов, связанных общей характеристикой. Исследователь может создавать словари по любому принципу, в зависимости от исследовательского вопроса. Например, можно составить словари, раскрывающие разные аспекты опыта проживания в отелях (интерьер номеров, качество проживания, наличие каких-то услуг), посчитать их выраженность в текстах и сравнить их между собой.

Другим примером анализа с использованием словарей является словарный анализ тональности (эмоциональной окраски) текста [16]. В данном случае составляются словари, выражающие эмоциональный характер слов, то есть связанные с негативной или позитивной окраской текста, или с разными эмоциями (радость, страх, злость и т.д.). Вдобавок к этому, каждое слово обладает разной степенью эмоциональной окраски. Таким образом, можно определить, насколько эмоционально окрашен текст и какие именно эмоции он выражает.

Методы, основанные на соприсутствии слов

Одним из недостатков фокуса анализа на отдельных словах является отсутствие учета омонимии [17], когда одно и то же слово может иметь несколько разных значений. Кроме того, учет одиночных слов не позволяет рассматривать ситуации, когда слова упоминаются с частицей “не” или другими словами, которые меняют смысл слова. Несмотря на то, что эти проблемы решаются на уровне анализа n-грамм, часто требуются более эффективные способы анализа контекста. В таком случае используются, например, семантические сети или сети со-встречаемости [18], позволяющие анализировать связи между словами.

Другим примером методов, основанного на соприсутствии слов, является тематическое моделирование [19]. Тематическое моделирование — это современный вычислительный метод кластеризации слов на основе их со-встречаемости в текстах. Слова, которые встречаются вместе чаще, чем остальные, объединяются в темы (topics), к которым эти слова имеют высокую вероятность принадлежности, в то время как остальные слова имеют вероятность, близкую к нулю. При этом выраженность каждой темы в документе характеризуется пропорцией принадлежащих к ней слов.

Тематическое моделирование генерирует темы и вычисляет их пропорции в текстах, однако их интерпретация зависит от исследователя. Используя самые вероятные слова темы (или другие метрики важности слов для темы) и тексты с наивысшей пропорцией выраженности темы, исследователь может аннотировать каждую из тем.

Метод тематического моделирования подходит, в первую очередь, для эксплораторного анализа, при котором у исследователя не обязательно есть полная заранее заданная схема того, что обсуждается в текстах. Таким образом, метод позволяет выявлять неочевидные для исследователя паттерны, а возможность считать пропорции тем в текстах позволяет оценивать относительную важность этих тем для обсуждающих [20]. Некоторые расширения классических методов тематического моделирования, например, — структурные тематические модели — позволяют оценивать статистически связь ковариат с выраженностью тем.

Нетнография

Описанные методы используются не только для чисто статистического анализа, но и могут быть связаны с качественным анализом текстов. Методологически анализ обсуждений интернет-сообществ связан с нетнографическим подходом [21], главная особенность которого заключается в смешении количественных и качественных методов анализа данных. В данном случае количественный анализ текстов помогает исследователю ориентироваться в материале, выявлять интересные паттерны и выбирать тексты, требующие более глубокого осмысления.

У современного нетнографического подхода есть несколько преимуществ, в том числе в задачах анализа процесса формирования ценностей особенных благ: вычислительные методы позволяют быстро выделить ключевые паттерны обсуждений в корпусах текстов, а следование нетнографическому подходу позволяет

интерпретировать выявленные связи с учетом локальной культуры интернет-сообществ, которая может быть изначально непрозрачной для исследователя, что влечет затруднения в интерпретации.

Измерения ценности особенных благ и “цифровые следы”

Применение описанных в статье подходов позволяют использовать “цифровые следы” выбора и оценки опытных благ, накапливаемые в рамках интернет-сервисов. К таким следам, в первую очередь, относятся отзывы, так как они непосредственно направлены на осмысление опыта потребителя, однако и обсуждения в свободной форме являются не менее ценным источником информации о конструировании ценности.

В первую очередь, отзывы могут быть быть источником информации о позитивно или негативно оцениваемом опыте, полученным от взаимодействия с “особыми благами”. Для решения таких задач может использоваться оценка эмоциональной окраски текста. Так Джурафски с коллегами [16] продемонстрировали, что в отзывах на рестораны с низкой оценкой пользователи используют язык, характерный для описания “пережитой травмы”, в то время как в отзывах с высокой оценкой пользователи склонны использовать язык, связанный с “зависимостью“ и чувственным удовольствием.

Однако эмоциональная окраска не является единственным измерением опыта. Например, авторы [22], используя частотный анализ слов, анализируют как преподносится город в официальных описаниях туристических объектов (сторона предложения услуг) и отзывах туристов на TripAdvisor (сторона потребления услуг).

Помимо этого, результатом таких исследований может являться формирование глобальных и локальных моделей измерений ценности (dimensions) для групп благ и тематических областей. Например, в работе [11] авторы используют тематическое моделирование для анализа измерений опыта гостей отелей в отзывах на сайте TripAdvisor. С помощью базового алгоритма тематического моделирования Latent Dirichlet Allocation [13] авторы выделили 30 измерений опыта гостей, среди которых 9 измерений были описаны впервые, тем самым показав важность учета возможности выхода реальных практик оценки за рамки классических моделей.

Совмещая тематическое моделирование с анализом преобладания тех или иных слов в подкорпусах отзывов, авторы [23] сконцентрировались на восприятии отелей Санкт-Петербурга и связанного с городом туристического опыта на основе отзывов русскоязычных пользователей сайта TripAdvisor. Помимо 29 измерений туристического опыта, авторы сфокусировались на сравнении их преобладания в отелях разного уровня обслуживания, показывая локальные аспекты формирования ценностей. Так что локация отеля, качество уборки и завтрак оказались более важны для посетителей 2-3 звездочных отелей, в то время как для посетителей 4-5 звездочных отелей оказались более важны event-менеджмент и интерьер отеля и номеров. Кроме того, авторами показано, что часть опыта связана не только с расположением отеля, но и с городскими локациями, интересующими посетителей. Так, Аврора, Марсово поле и Спас-на-Крови преобладают в отзывах к 2-3

звездочным отелям, а такие локации как Александро-Невская лавра и Исаакиевский собор преобладают в отзывах к 4-5 звездочным отелям.

Еще более интересной областью применения новых методов является анализ формирования ценности для виртуальных товаров, существующих только в информационных системах (например, онлайн-играх), но продаваемых и покупаемых за реальные деньги. Несмотря на существование классических моделей ценности таких товаров [24,25], исследования на основе анализа "цифровых следов" позволяют более полно описать процессы взаимодействия оценочных механизмов, реконструируя их на основе анализа обсуждений пользователей. Так, авторы [26] анализируют обсуждения на сайте Reddit.com, посвященные виртуальным товарам популярной онлайн-игры Dota 2, фокусируясь на аспектах формирования ценности нефункциональных товаров и показывая преобладание социально-конструируемых аспектов ценности и их непрерывное конструирование и реконструирование в неразрывной связке с игровым опытом. Работа [27] описывает предварительные результаты решения более сложной задачи — анализа взаимодействия личного и командного бренда киберспортсменов, используя момент перехода игроков между командами в качестве "точки разборки" этих двух типов брендов.

Выводы

В рамках данного обзора мы фокусируем внимание на "цифровых следах" и вычислительном текстовом анализе как новом подходе к исследованию механизмов формирования ценности особенных благ. Главным преимуществом описанных методов является возможность выявить новые механизмы и измерения опыта использования особенных благ, не обязательно предусмотренные теорией, но являющиеся важной его частью; а также наблюдать механизмы в их взаимодействии и столкновениях [28], позволяя ученым в области социальных наук более глубоко анализировать отдельные аспекты информационного общества.

При этом каждый подход и метод имеет свою область применения и ограничения. Частотный анализ, являющийся основой остальных методов, позволяет выявлять частое (или редкое) употребление слов (или групп слов) и сравнивать между собой группы текстов, но отдельные слова или n-граммы могут плохо передавать контекст употребления слов. Частотный анализ со словарями позволяет выявлять выраженность тем, но требует схемы, заданной заранее. Анализ эмоциональной окраски текста, зачастую осуществляемый на основе словарей, помогает выявить (не без ограничений, особенно характерных для анализа текстов на русском языке) характер (радость, грусть, гнев) или направление эмоции (положительный или негативный), но плохо схватывает возможное разнообразие измерений потребительского опыта. Тематические модели помогают выделять группы слов, но интерпретация этих групп может быть затруднена из-за субъективности исследователя. При этом, несмотря на появление и более технически "продвинутых" вычислительных моделей, фокус на анализе и интерпретации результатов требует особого внимания к учету контекста исследования (например, на основе нетнографического подхода) и тщательного выбора сочетания методов, соответствующих задаче.

Статья подготовлена в ходе проведения исследования 18-01-0002 в рамках Программы «Научный фонд Национального исследовательского университета „Высшая школа экономики“ (НИУ ВШЭ)» в 2018-2019 гг. и в рамках государственной поддержки ведущих университетов Российской Федерации «5-100».

ЛИТЕРАТУРА

1. HELGESSON C.–F., **Muniesa F. For what it's worth: An introduction to valuation studies** // Valuat. Stud. 2013. Vol. 1, № 1. P. 1-10.

2. KARPIK L. **The economics of singularities.** Princeton University Press, Princeton, 2010.

3. BECKERT J., MUSSELIN C. **Constructing Quality: The Classification of Goods in Markets.** OUP Oxford, 2013. 356 p.

4. BECKERT J., RÖSSEL J., SCHENK P. **Wine as a Cultural Product Symbolic Capital and Price Formation in the Wine Field** // Sociol. Perspect. 2016.

5. РОЩИНА Я. М. **Как на рынках «Особенных благ» формируются суждения о качестве?** Рецензия на книгу: Karpik L. 2010. Valuing the Unique: the Economics of singularities. Princeton; Oxford. Princeton University Press // Экономическая Социология. 2015. Vol. 16, № 4.

6. СТРЕБКОВ Д.О. ET AL. **Социальные факторы выбора контрагентов на бирже удалённой работы: исследование конкурсов с помощью «больших данных»** // Экономическая Социология. 2019. Vol. 20, № 3. P. 25-65.

7. ТОЛСТОВА Ю. **Социология и компьютерные технологии** // Социологические Исследования. 2015. № 8. P. 3-13.

8. ТОЛСТОВА Ю. Н. **Новые информационные технологии как фактор повышения эффективности социологического исследования** // Математическое Моделирование Социальных Процессов Сб Трудов. 2015. № 17. P. 210-228.

9. ВОЛЧЕНКО О. В. **Измерение практик использования интернета в социальных науках: обзор основных методов** // Информационное Общество. 2017. № 1. P. 47-54.

10. МАРАРИЦА Л.В., ТИТОВ С. М. **Социальный мир человека в эпоху виртуальных социальных сетей** // Информационное Общество. 2017. № 2. P. 30-36.

11. GUO Y., BARNES S. J., JIA Q. **Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation** // Tour. Manag. 2017. Vol. 59. P. 467-483.

12. KOPPEL M., ARGAMON S., SHIMONI A. R. **Automatically categorizing written texts by author gender** // Lit. Linguist. Comput. 2002. Vol. 17, № 4. P. 401-412.

13. BLEI D.M., NG A. Y., JORDAN M. I. **Latent Dirichlet Allocation** // J Mach Learn Res. 2003. Vol. 3. P. 993-1022.

14. PIANTADOSI S. T. **Zipf's word frequency law in natural language: A critical review and future directions** // Psychon. Bull. Rev. 2014. Vol. 21, № 5. P. 1112-1130.

15. STUBBS M. **Three concepts of keywords** // Keyness Texts. 2010. P. 21-42.

16. JURAFSKY D. ET AL. **Narrative framing of consumer sentiment in online restaurant reviews** // First Monday. 2014. Vol. 19, № 4.

17. ROLL U., CORREIA R. A., BERGER-TAL O. USING MACHINE LEARNING TO DISENTANGLE HOMONYMS IN LARGE TEXT CORPORA // CONSERV. BIOL. 2018. VOL. 32, № 3. P. 716-724.

18. RULE A., COINTET J.–P., BEARMAN P. S. **Lexical shifts, substantive changes, and continuity in State of the Union discourse,** 1790-2014 // Proc. Natl. Acad. Sci. 2015. Vol. 112, № 35. P. 10837-10844.

19. BLEI D. M. **Probabilistic topic models** // Commun. ACM. 2012. Vol. 55, № 4. P. 77-84.

20. DIMAGGIO P., NAG M., BLEI D. **Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding** // Poetics. 2013. Vol. 41, № 6. P. 570-606.

21. KOZINETS R. V. **Netnography // The International Encyclopedia of Digital Communication and Society** / ed. Ang P. H., Mansell R. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2015. P. 1-8.

22. GORGADZE A., GORDIN V., BELYKOVA N. **Semantic Analysis of the Imperial Topic: Case of St. Petersburg** // E-Rev. Tour. Res. 2019. Vol. 16, № 2/3.

23. KASPRUK N., SILYUTINA O., KAREPIN V. **Hotel Value Dimensions and Tourists' Perception of the City. The Case of St. Petersburg** // Digital Transformation and Global Society / ed. Alexandrov D. A. et al. Springer International Publishing, 2017. P. 341-346.

24. LEHDONVIRTA V. **Virtual Item Sales as a Revenue Model: Identifying Attributes that Drive Purchase Decisions:** SSRN Scholarly Paper ID1351769. Rochester, NY: Social Science Research Network, 2009.

25. HAMARI J., LEHDONVIRTA V. **Game Design as Marketing: How Game Mechanics Create Demand for Virtual Goods:** SSRN Scholarly Paper ID1443907. Rochester, NY: Social Science Research Network, 2010.

26. MUSABIROV I. ET AL. **Deconstructing Cosmetic Virtual Goods Experiences in Dota 2** // Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: ACM, 2017. P. 2054-2058.

27. MARCHENKO E., MUSABIROV I. **Mediametrics in Esports: The Case of Dota 2** // Proceedings of the 2019 Annual Symposium on Computer-Human Interaction in Play. 2019.

28. KORNBERGER M. **Brand society: How brands transform management and lifestyle.** Cambridge University Press, 2010.