Об одной модели цифровых привычек

А.П. Рыжов, П.А. Новиков

В работе предложена формализация способов использования человеком информационно-коммуникационных технологий на основе анализа логов смартфона. Полученная модель взаимодействия человека с другими людьми и информационными ресурсами (цифровым миром) может быть использована для персонализации такого цифрового окружения и, соответственно, оптимизации такого взаимодействия. Предложен сценарий использования модели для персонализации новостной ленты.

Ключевые слова: персонализация, цифровой след, нечеткая кластеризация.

1. Введение.

Широкое использование информационно-коммуникационных технологий в повседневной жизни открывает новые возможности для государств (e-government, smart city), экономики (цифровая экономика, цифровая трансформация), оптимизации различных функций бизнеса - маркетинга (цифровой маркетинг), продаж (cross-selling, up-selling) и многих других аспектов функционирования бизнеса и жизни человека. Интересующемуся читателю можно порекомендовать ознакомиться с [1], где эти возможности представлены достаточно полно и описаны достаточно убедительно и глубоко.

Основой значительной доли таких возможностей является обычный мобильный телефон. Еще в 2014 году количество активных SIM-карт, используемых в телефонах, смартфонах и планшетах, впервые в истории превысило число живущих на Земле людей [2]. То же самое произошло в 2016 году с мобильными телефонами [3]. Интересная статистика распределения телефонов по странам представлена в [4]: от 11,6 на 100 человек на Кубе до 240,2 на 100 человек в Гонконге.

Приведенные цифры говорят о том, что мобильный телефон является действительно персональным устройством для большинства людей, и стиль его использования во многом характеризует самого человека, которому он принадлежит. Это давно подметили крупные компании -

ритейл, банки, страховые, энергетические и многие другие компании, которые научились использовать это в сегментации клиентов, маркетинге, продажах и во многих других бизнес-процессах. Читатель может без труда найти много таких примеров в интернете, мы не будем их анализировать. Заметим лишь, что в этой парадигме человек - источник данных для всех подобных алгоритмов - является объектом, его классифицируют, сегментируют, а потом еще и навязывают различную рекламу на основе проведённых классификаций и сегментаций. Авторы видят в этом несправедливость. Настоящая работа - первый шаг сделать человека субъектом в этом процессе, когда он на основе своей активности формирует «персональный API» и указывает различным бизнесам что, когда, и в каком формате ему предлагать. Эта концепция глубокой персонализации (deep personalization) нами только осмысливается, она частично описана в [5 - 7]. Мы видим ее большой потенциал как для людей, так и для компаний.

Из вышесказанного вытекает две особенности решаемой задачи. Во-первых, большинство сервисов, решающих задачи персонализации и построения рекомендаций принципиально ограничены обработкой индивидуальных событий (“пользователь купил товар”, “пользователь высоко оценил просмотренный фильм” и т.п.), которые отражают лишь статические характеристики интересов отдельных пользователей (исключение составляет информация об эволюции этих интересов, которая рассматривается, наепример, в [20]), либо обобщённые характеристики больших групп пользователей. Данные же об использовании телефона открывают возможность рассматривать привычки пользователя в привязке ко времени.

Во-вторых, поскольку мы предполагаем, что запрос на персонализацию поступает от пользователя, необходим инструмент, позволяющий описать данные, передаваемые сервисам, в понятных человеку терминах. В таком контексте естественным выбором, позволяющим совместить информационную нагруженность с понятностью, является язык нечёткой логики.

Итак, изучая логи использования мобильного телефона, мы можем что-то сказать о человеке и его привычках работы с информацией. Описываемая ниже модель, несмотря на массу допущений, дает представление о потенциале подобного анализа.

2. Модель.

2.1. Предварительные замечания.

Ниже мы будем пытаться извлечь полезную для решения определённой задачи информацию из логов смартфона определённого человека. Это имеет отношение к анализу цифровых следов человека. В настоящее время нет четкого определения термина «цифровой след», это больше журналистский прием (см., например, [16]). Предлагаемые определения (например, [17] - «Цифровой след (или цифровой отпечаток; англ. digital footprint) — совокупность информации о посещениях и вкладе пользователя во время пребывания в цифровом пространстве. Может включать в себя информацию, полученную из Интернета, мобильного Интернета, веб-пространства и телевидения») вызывают много вопросов, поэтому мы не будем пытаться их анализировать и критиковать, а будем использовать более нейтральный термин - цифровые привычки пользователя.

Под цифровыми привычками будем понимать набор лингвистических переменных [18], определяемых на логах смартфона, достаточный для решения определённой задачи. Например, *Активность* в терминах «активный», «средней активности», «не активный»; *Предпочтения типа контента* в терминах «голос», «картинки», «видео»; *Предпочтения размера контента* в терминах «большой», «средний», «маленький»; *Предпочтения времени* в терминах «утро», «день», «вечер» и т.п. Набор лингвистических переменных определяется задачей (то есть нас не будет волновать абстрактная полнота набора лингвистических переменных, но будет важно понимать, как их строить).

Формально, будем считать, что наш объект описывается конечным набором признаков $A = \{A_1, ..., A_n\}$. Каждому признаку A_q ставится в соответствие U_q множество его «физических» значений и множество $\{a_{1q}, ..., a_{n_q q}\}$ лингвистических значений $(1 \leq q \leq n)$ (то есть признак - это лингвистическая переменная). Каждому такому лингвистическому значению a_{wq} ставится в соответствие функция принадлежности $\mu_{a_{wq}}(u_q)$ в универсальном множестве U_q $(1 \leq w \leq n_q)$. Множества U_q определяются набором доступных данных. Такой набор данных (dataset) был предоставлен компанией ООО “Технологии обработки данных” и включает в себя запись истории запуска приложений смартфона (логов) 800 пользователей за 4 недели, в формате <идентификатор, транзакция>, где транзакция имеет формат <время начала транзакции, приложение, время окончания транзакции>, где приложение - имя одного из приложений, установленных на смартфоне пользователя, с которым происходила работа в указанное время. Из такого dataset мы можем извлекать различные данные (множества U_q), например, среднее количество звонков, среднее количество использования телефона в сутки / за неделю и над

ними (множествами U_q) строить лингвистические значения, описывающие интересующие нас признаки A_q $(1 \leq q \leq n)$.

2.2. Обозначения.

Обозначим через X_i $(1 \leq i \leq N)$ каналы коммуникации, которые мы умеем измерять. Это могут быть звонки, смс, чаты, мессенджеры, социальные сети и пр. Единицы измерения - время (для звонков), количество символов (для смс и мессенджеров), время нахождения в приложении и т.п. должны быть определены.

Обозначим через Δt некоторый промежуток времени, «естественный» для задачи. Ниже будем считать этот промежуток равным суткам. Разобьём сутки на элементарные единицы - минуты и обозначим через Δt_j $(1 \leq j \leq 1440)$ промежуток времени, соответствующий j минуте.

Обозначим через $v_{i,j}^k$ количество единиц, потраченных k-ым пользователем по каналу i на j-ой минуте $(1 \leq k \leq K)$. Для времени $v_{i,j}^k$ - число из отрезка $[0, 1]$, соответствующее активности в рассматриваемый промежуток, для других единиц измерения (например, символов), это может быть усредненная величина за время набора текста.

Замечание 1. Нами не учитывается направленность коммуникации, только интенсивность. Это упрощение модели позволяет тем не менее решать достаточно важные задачи.

2.3. Разбиение универсального множества.

Допустим, нам будет интересно понимание готовности конкретного пользователя потреблять информацию в определенное время суток. Сколько таких осмысленных интервалов суток существует? Общего ответа нет, поэтому правильно выбрать такое количество «естественных интервалов», которое обеспечит лучшее качество покрытия ими универсального множества $[0, 1440]$. Качество определяется различными показателями - ниже мы будем использовать, аналогично [8], дисбаланс классов и степень нечеткости.

Для этого соберем все активности всех пользователей за определенное время, то есть вычислим для каждого Δt_j величину $\bar{v}_j^k = \sum_{i=1}^{N} \bar{v}_{i,j}^k$, где $\bar{v}_{i,j}^k = 1$, если k-ый пользователь использовал i-ый канал коммуникации во временя Δt_j и $\bar{v}_{i,j}^k = 0$, если у k-го пользователя не было никакой активности в промежуток времени Δt_j. Для разбиения времени суток выберем «типичного» пользователя (например, центр кластера после кластеризации пользователей)и кластеризацию объединения мно-

жеств $\{i|\bar{v}_i > 0\}$ стандартным алгоритмом c-means для разного количества кластеров и посчитаем качество кластеризации.
Замечание 2. Мы неявно считаем, что все каналы коммуникаций имеют одинаковую ценность. Это еще одно упрощение модели.
Замечание 3. 1 секунда оказалась не самым удачным разбиением суток - человек если и делает что-то регулярное, то не с точностью до секунды. Проведенные нами эксперименты показали наилучшие результаты для интервала времени 15 мин., что вполне разумно. Поэтому ниже мы будем использовать именно такую агрегацию времени, то есть $(1 \leq j \leq 95)$ Результаты собраны в таблице 1.

Таблица 1. Характеристики качества кластеризации времени использования

Количество кластеров	Степень нечеткости	Дисбаланс классов
2	0.17368421	0.230382367079
3	0.29473684	0.199661103355
4	0.32631579	0.389432788809
5	0.33684211	0.488659684614

Как видно из таблицы 1, оптимальным с точки зрения дисбаланса классов является разбиение на 3 класса, которые можно интерпретировать как «утро», «рабочее время» и «вечер» (центры кластеров 22.84787803, 54.46499606, 78.82015172). Будем далее использовать именно такое разбиение суток (рис. 1).

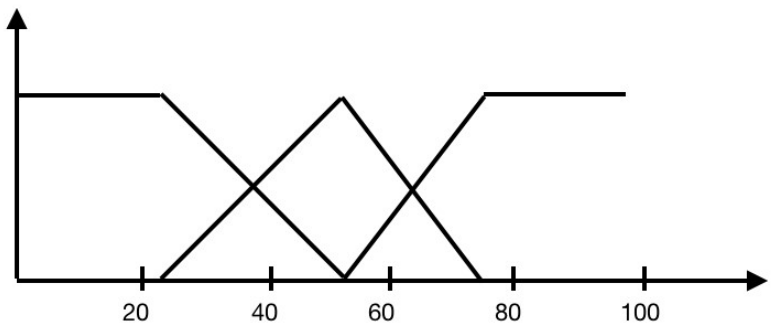

Рис. 1. Кластеризация времени использования для трех кластеров.

Аналогично мы можем определять, например, активность пользователя:

1. Вычисляем среднее значение потраченных единиц за сутки для каждого пользователя.

2. Проводим кластеризацию пользователей.

3. Интерпретируем полученные кластеры.

Если нас интересует описание пользователя в терминах «Активный/ не активный», «Любит читать новости утром/ днем/ вечером», «Любит видео/ текст» и пр., то, используя аналогичные рассуждения и при наличии данных, мы можем построить такие описания. Эти описания (лингвистические переменные, построенные на универсальных множествах, полученных из dataset) будем называть первичными. На основе полученных первичных описаний, мы можем проводить классификацию пользователей с помощью нечеткого классификатора [9] и получать вторичные описания. Заметим, что мы можем делать это с минимальной неопределенностью для большого количества случаев [9].

Такие описания и являются описанием цифровых привычек пользователя. Архитектура системы построения цифровых привычек представлена на рис. 2

Рис. 2. Архитектура системы построения цифровых привычек

Описания цифровых привычек, или профили, могут быть использованы для персонализации различных сервисов (например, рекомендательных).

3. Сценарий использования.

В качестве возможного примера использования такой формализации информационных привычек пользователя, рассмотрим персонализацию новостной ленты. Эта задача активно обсуждается [10 – 15], и, по мне-

нию многих аналитиков, является трендом развития интернет СМИ и новостных агрегаторов.

Пусть у нас есть поток новостей, и мы хотим упорядочить их так, чтобы пользователь испытывал минимальных дискомфорт при просматривании новостей. Каждая новость представляет собой файл, про который известно:

- Наличие и размер текста
- Наличие и размер картинок
- Наличие и размер видеоклипов
- Тематика (по классификатору агрегатора) и другие параметры (зависит от агрегатора или СМИ)

Локальной задачей является оптимизация времени подачи новости. Давно подмечено, что время существенно влияет на эффективность коммуникации (даже существует термин prime-time). Если мы умеем вычислять какую информацию предпочитает пользователь утром, а какую - вечером, то, соединяя это знание с параметрами новости, мы можем определить насколько она будет соответствовать его привычкам. Например, если пользователь утром предпочитает короткие текстовые файлы (мы можем это видеть из используемых им утром приложений и времени их использования), а вечером - большие видеофайлы (информация так же доступна), то мы можем вычислить насколько конкретная новость соответствует этим критериям (вычисляя степень принадлежности к понятиям «короткий текст» и «большой видео»). Добавляя степень принадлежности текущего времени к понятиям «утро» и «вечер», мы можем вычислить рейтинг новости в данный момент времени для данного пользователя (как t-норму упомянутых выше степеней принадлежности [19]) и на основе этого - позицию новости в списке новостей для данного пользователя. В результате мы будем иметь «персональную газету», которая будет отличаться для пользователей с разными цифровыми привычками, и, более того, в разное время для одного и того же пользователя персональная лента будет разной. Архитектура такого персонализатора новостей представлена на рис. 3.

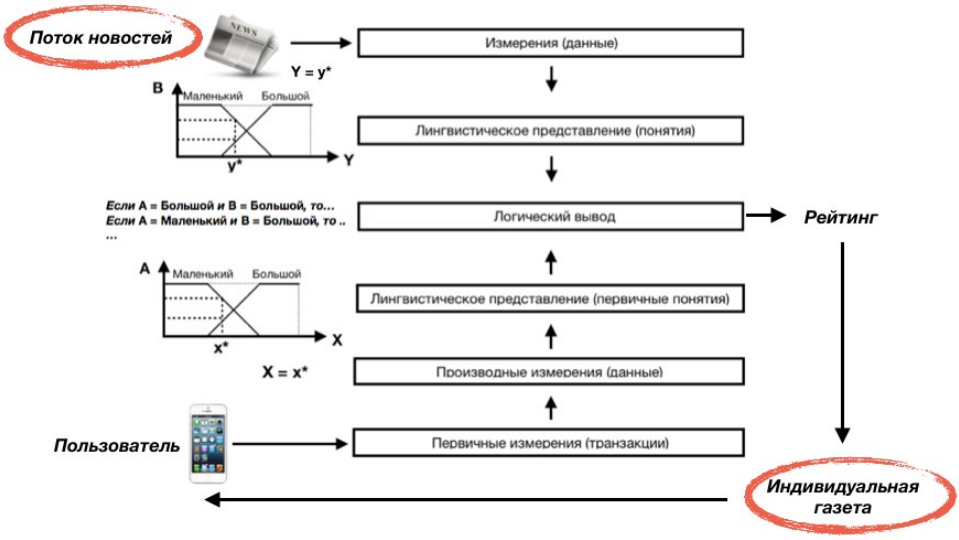

Рис. 3. Архитектура персонализатора новостной ленты.

Подобные дополнения могут сделать более персонализированными (а, следовательно, и более эффективными) любые коммуникации с клиентами, включая рекомендации, рекламные кампании и т.п.

Список Литературы

[1] James Manyika, Michael Chui, Jacques Bughin, Richard Dobbs, Peter Bisson, Alex Marrs. Disruptive technologies: Advances that will transform life, business, and the global economy. McKinsey Global Institute (MGI), May 2013, 176 p. - *http* : *//www.mckinsey.com/insights/business*$_t$*echnology/disrupti*

[2] Eric Mack. There are now more gadgets on Earth than people - https://www.cnet.com/news/there-are-now-more-gadgets-on-earth-than-people/

[3] Мобильных телефонов – больше, чем людей на планете - http://apps4all.ru/post/10-09-14-mobilnyh-telefonov-bolshe-chem-lyudej-na-planete

[4] Количество мобильных телефонов по странам мира. 30-04-2017. -

http://total-rating.ru/1970-kolichestvo-mobilnyh-telefonov-po-stranam-mira.html

[5] Рыжов А.П. Некоторые задачи оптимизации и персонификации социальных сетей. Saarbrucken, LAP, 2015, 88 с. (ISBN: 978-3-659-68661-0)

[6] Alexander Ryjov. Personalization of Social Networks: Adaptive Semantic Layer Approach. In: Social Networks: A Framework of Computational Intelligence. Witold Pedrycz and Shyi-Ming Chen (Eds.). Springer International Publishing Switzerland 2014, pp. 21-40.

[7] Alexander Ryjov. Towards an optimal task-driven information granulation. In: Information Granularity, Big Data, and Computational Intelligence. Witold Pedrycz and Shyi-Ming Chen (Eds.). Springer International Publishing Switzerland 2015, pp. 191-208.

[8] Рыжов А. П., Журавлев А. Д., Вахов А. Н., Кривцов В. В. Об одном подходе к персонализации обучения в рамках компьютерных обучающих систем. Интеллектуальные Системы. Теория и приложения. Т. 20, Вып. 3, 2016, с. 180-185.

[9] Рыжов А.П. О качестве классификации объектов на основе нечетких правил. Интеллектуальные системы, Т.9, вып. 1-4, Москва, МНЦ КИТ, 2005, с. 253 – 264.

[10] Персонализация в новостях - агрегатор Top Story - http://www.mforum.ru/phones/tests/113536.htm

[11] Новостной агрегатор Gong использует новый алгоритм персонализации -https://hightech.fm/2016/11/25/gong-news

[12] Новостной агрегатор SmartNews собрал $38 млн. - https://hightech.fm/2016/07/08/smartnews

[13] Мифы и факты об алгоритмах формирования новостной ленты Facebook - https://vc.ru/6615-facebook-news-feed

[14] «ВКонтакте» изменит алгоритм формирования ленты новостей - http://www.rbc.ru/rbcfreenews/56f502f39a79478da23507bd

[15] Instagram изменит порядок показа публикаций - http://www.rbc.ru/technology_and_media/16/03/2016/56e914e09a794700ce80e798

[16] Стечкин И. Кому принадлежит наш «цифровой след»? - http://jrnlst.ru/komu-prinadlezhit-nash-cifrovoy-sled

[17] Цифровой след. Материал из Википедии — свободной энциклопедии. https://ru.wikipedia.org/wiki/Цифровой_след

[18] Заде Л.А. Понятие лингвистической переменной и его применение к принятию приблизительных решений. М., Мир, 1976. - 165 с.

[19] Рыжов А.П. Элементы теории нечетких множеств и измерения нечеткости. Москва, Диалог-МГУ, 1998, 116 с.

[20] Li L, Zheng L, Yang F, Li T. Modeling and broadening temporal user interest in personalized news recommendation. Expert Systems with Applications. 2014;41:3168-77.

One model of digital habits

A.P. Ryjov P.A. Novikov

The paper proposes the formalization of ways of using the information technologies on the basis of the analysis of smartphone's logs. The resulting model of human interaction with other people and information resources (digital world) can be used to personalize such a digital environment and, accordingly, to optimize such interaction. A scenario for using the model to personalize the news feed is described.

Keywords: personalization, digital footprint, fuzzy clustering.