НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМ. Н. Г. ЧЕРНЫШЕВСКОГО
РОССИЙСКИЙ УНИВЕРСИТЕТ ДРУЖБЫ НАРОДОВ
ИНСТИТУТ ПРОБЛЕМ УПРАВЛЕНИЯ
им. В.А. ТРАПЕЗНИКОВА РАН

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ (ИТММ-2019)

МАТЕРИАЛЫ
XVIII Международной конференции
имени А. Ф. Терпугова
26–30 июня 2019 г.

Часть 1

ТОМСК
«Издательство НТЛ»
2019

К построению времязависимого потребительского профиля пользователя рекомендательной системы линейного телевидения

М.А. Бакланов, В.В. Поддубный

Национальный исследовательский
Томский государственный университет, г. Томск, Россия

Предлагается подход к построению потребительского профиля пользователя рекомендательной системы линейного телевидения по данным временных рядов мониторинга просмотров тегов представления контента телепередач в стандарте xml-tv. Предполагается, что система мониторинга фиксирует с постоянным малым шагом во времени наличие или отсутствие просмотра каждого тега в течение всего времени мониторинга. Эти данные преобразуются во временные ряды длительностей просмотров каждого тега на последовательных интервалах времени наблюдений (таймслотах) с большим количеством нулей, соответствующих отсутствию просмотра тега в таймслоте. Данные агрегируются по таймслотам каждого интервала времени суток каждого дня недели. По этим данным на фиксированном или скользящем интервале времени наблюдения рядов находятся эмпирические оценки статистических характеристик просмотров тегов в каждом таймслоте недели. Потребительский профиль индивидуального пользователя строится по статистически значимым превышениям средних значений длительностей просмотров тегов в недельном наборе таймслотов над соответствующими значениями, полученными в результате усреднения данных мониторинга многих пользователей.

Введение и постановка задачи

Несмотря на стремительное развитие web-телевидения (потребления ТВ-контента через интернет), линейное телевидение остаётся востребованным как и рекомендательные услуги провайдеров для пользователей ТВ. Целью этих услуг является предоставление пользователю такого контента телепередач, который обеспечивает максимальный комфорт при просмотре передач, исключающий, по возможности, ручной поиск интересной для пользователя передачи и тем самым максимально сокращающий количество ручных переключений каналов. В этой связи одной

из ключевых проблем рекомендательных систем линейного телевидения является оценка предпочтений пользователя в потреблении ТВ-контента, то есть построение потребительского профиля пользователя.

Особенностью просмотра телепередач линейного ТВ является явно выраженная зависимость от времени суток и дней недели, определяемая программами передач. Поэтому выражение потребительских предпочтений пользователя линейного ТВ и, следовательно, его потребительский профиль неминуемо зависят от времени суток и дней недели. Это обстоятельство привело разработчиков рекомендательных систем линейного ТВ к необходимости построения времязависимого потребительского профиля пользователя и времязависимой структуры рекомендательной системы. Появилось понятие таймслота, определённого интервала времени суток каждого дня недели, с которым стали связывать потребительские предпочтения пользователя и его потребительский профиль в этом интервале времени [1–7].

Одной из существенных проблем построения рекомендательных систем является учёт контекста потребления контента. В данной работе это осуществляется через профиль «усреднённого» пользователя, позволяющий частично учитывать контекст просмотра контента.

Мы предлагаем алгоритм построения времязависимого потребительского профиля пользователя на основе анализа наблюдаемых тегов представления контента телепередач в стандарте xml-tv [8] путём сравнения теговых времязависимых предпочтений пользователя с времязависимыми предпочтениями «усреднённого» пользователя, полученных по данным синхронных временных рядов мониторинга теговой телевизионной активности индивидуального пользователя и группы многих пользователей.

Структура данных мониторинга теговой ТВ-активности пользователя

Предполагается, что телевизор пользователя оснащён аппаратными средствами слежения за просмотром телепередач, способными постоянно через достаточно малые (например, порядка 1 с) промежутки времени Δt информировать провайдера о том, какая передача просматривается в текущий момент дискретного времени и, следовательно, каким тегам формата xml-tv она соответствует. Сопоставляя фактам просмотра каждого тега в этот момент времени единицу, а отсутствию просмотра – ноль, получим двоичные временные ряды исходного мониторинга

наблюдения каждым пользователем каждого тега. Это ряды вида: 0000011111111000000000111000000… Сумма n идущих подряд единиц, умноженная на Δt, даст длительность $\tau = n \cdot \Delta t$ времени просмотра данного тега данным пользователем на рассматриваемом интервале времени. Границы раздела последовательностей единиц и нулей соответствуют моментам переключения пользователя на просмотр другой телепередачи (другого тега), а границы раздела последовательностей нулей и единиц – моментам возвращения к просмотру данного тега.

Для последующего хранения и обработки данных мониторинга целесообразно агрегировать данные по таймслотам [1–7] (интервалам времени разбиения суток на равные части, значительно превышающие шаг Δt квантования времени при мониторинге). Пусть K – число таймслотов в сутках. Например, при делении суток на 4 части (ночь, утро, день, вечер) $K = 4$, а длительность таймслота равна 6 часам. Вместо 6-часовых таймслотов можно использовать часовые таймслоты ($K = 24$).

Агрегирование данных по таймслотам приводит к структуре данных в виде временных рядов с шагом в таймслот. Эти ряды выражают длительности τ просмотров каждого тега в каждом таймслоте и также содержат большое число нулей, соответствующих отсутствию просмотров тега в таймслотах. Эти ряды значительно короче исходных бинарных рядов мониторинга. Пусть τ_{il}, $i = \overline{1, I}$, $l = \overline{1, L}$, – длительность просмотра i-го тега в l-м таймслоте ряда, I – число тегов, L – число таймслотов. Если просмотр i-го тега в l-м таймслоте имеет место, $\tau_{il} > 0$, иначе $\tau_{il} = 0$. Ряды длительностей просмотров тегов образуют числовую матрицу $\tau = (\tau_{il})$ с I строками и L столбцами. С ростом времени T мониторинга растёт и число таймслотов L. Будем считать, что на интервале времени T укладывается целое число L таймслотов.

Характер просмотров телепередач пользователем можно считать различным не только в разное время суток, но и в разные дни недели. Поэтому целесообразно ввести в структуру данных недельный блок $7K$ таймслотов. Этот блок для каждого тега передач можно представить матрицей с K строками и семью столбцами, в каждой ячейке которой (в каждом таймслоте, соответствующем определённому времени суток определённого дня недели) хранится информация о длительности τ_{ikj} просмотра i-го тега в k-м интервале времени суток j-го дня недели ($k = \overline{1, K}$, $j = \overline{1, 7}$). Если просмотр i-го тега в kj-м таймслоте недельного блока таймслотов имеет место, $\tau_{ikj} > 0$, иначе $\tau_{ikj} = 0$.

Удобно считать, что в число таймслотов L укладывается целое число $N = L/(7K)$ недельных блоков таймслотов. Тогда все данные агрегированных по таймслотам наблюдений каждого тега можно представить $7K$ рядами длиной N таймслотов наблюдений длительностей просмотров тегов в таймслотах каждого времени суток каждого дня недели, выбранных из ряда наблюдений этого тега с шагом $7K$ таймслотов. Это, например, ряд длительностей всех утренних воскресных просмотров тега за N недель, всех вечерних субботних просмотров этого же тега и т.д.

Обозначим длительность просмотра i-го тега в k-м интервале времени суток j-го дня m-й недели ($m = \overline{1, N}$) через τ_{ikjm} . Если есть просмотр i-го тега в kj-м таймслоте блока m-й недели, $\tau_{ikjm} > 0$, иначе $\tau_{ikjm} = 0$.

Статистический анализ ненулевых длительностей просмотров тегов

Выделим теперь из каждого ряда $\tau_{ikj1}, \ldots, \tau_{ikjm}, \ldots, \tau_{ikjN}$ последовательность всех ненулевых (положительных) значений его членов $\tau_{ikjm} > 0$, обозначив их $x_{ikj\nu}$, где $\nu = \overline{1, n_{ikj}}$, $n_{ikj} \le N$ – число ненулевых длительностей просмотров i-го тега на k-м интервале времени суток j-го дня недели среди всех N недель, то есть n_{ikj} – число недель, в которых имеется просмотр i-го тега в kj-м таймслоте недельного блока таймслотов.

Эмпирическая оценка вероятности просмотра i-го тега в kj-м таймслоте (доля недель просмотра i-го тега в kj-м таймслоте) выражается отношением

$$P_{ikj} = n_{ikj} / N , \quad i = \overline{1, I} , \quad k = \overline{1, K} , \quad j = \overline{1, 7} . \qquad (1)$$

Проверка случайности последовательностей x_{ikj1} , …, $x_{ikj\nu}$, …, $x_{ikjn_{ikj}}$ реальных данных мониторинга просмотров ТВ-передач по критерию Уоллиса – Мура серий («фаз») знаков разностей [9, с. 354] показала, что на уровне значимости 5 % нет оснований отвергнуть нулевую гипотезу $H = 0$ об их случайности. Следовательно, последовательность $\{ x_{ikj\nu}$, $\nu = \overline{1, n_{ikj}} \}$ можно рассматривать как случайную выборку ненулевых длительностей просмотров i-го тега в kj-м таймслоте из всего ряда наблюдений. По этой выборке можно построить эмпирическую инте-

гральную функцию распределения ненулевых длительностей просмотров *i*-го тега в *kj*-м таймслоте:

$$F_{ikj} = \begin{cases} 0, & x < x_{ijk1}, \\ \nu / n_{ikj}, & x_{ikj\nu} \le x < x_{ikj,\nu+1}, \quad 1 \le \nu < n_{ikj}, \\ 1, & x \ge x_{ikj,\nu+1}. \end{cases} \qquad (2)$$

Эмпирические математическое ожидание M_{ikj} и дисперсия D_{ikj} длительности просмотра *i*-го тега в *kj*-м таймслоте выражаются формулами

$$M_{ikj} = \frac{1}{n_{ikj}} \sum_{\nu=1}^{n_{ikj}} x_{ikj\nu} \ , \quad D_{ikj} = \frac{1}{n_{ikj} - 1} \sum_{\nu=1}^{n_{ikj}} (x_{ikj\nu} - M_{ikj})^2 \ . \qquad (3)$$

Проверка гипотезы о нормальности закона распределения последовательности $\left\{ x_{ikj\nu}, \nu = \overline{1, n_{ikj}} \right\}$ по критерию Лиллиефорса [9, с. 302] показала, что в общем случае закон распределения ненулевых длительностей просмотров тегов нельзя считать нормальным. Однако с помощью копула-преобразования $y = \Phi^{-1}\left(F\left(x\right)\right)$ с эмпирическим распределением $F\left(x\right)$ вида (2) и нормальным распределением $\Phi\left(y\right)$ с параметрами (3) их можно нормализовать [10, 11].

Построение потребительского профиля пользователя линейного телевидения

Потребительский профиль индивидуального пользователя линейного ТВ можно построить, сравнивая его активность потребления ТВ контента с активностью «усреднённого» пользователя, описываемого усреднением данных мониторинга многих пользователей.

Аналогично (3), выразим эмпирические математическое ожидание $\overline{M}_{ikj}$ и дисперсию $\overline{D}_{ikj}$ ненулевых значений длительностей $\left\{ \overline{x}_{ikj\nu}, \nu = \overline{1, \overline{n}_{ikj}} \right\}$ просмотров «усреднённым» пользователем *i*-го тега в *kj*-м таймслоте недельного блока таймслотов формулами

$$\overline{M}_{ikj} = \frac{1}{\overline{n}_{ikj}} \sum_{\nu=1}^{\overline{n}_{ikj}} \overline{x}_{ikj\nu} \ , \ \overline{D}_{ikj} = \frac{1}{\overline{n}_{ikj} - 1} \sum_{\nu=1}^{\overline{n}_{ikj}} (\overline{x}_{ikj\nu} - \overline{M}_{ikj})^2 \ , \qquad (4)$$

пометив все переменные чертой сверху.

Относительное увеличение (уменьшение) предпочтения индивидуального пользователя в отношении просмотра *i*-го тега в *kj*-м таймслоте недельного блока по сравнению с «усреднённым» пользователем мы предлагаем оценивать величиной относительного увеличения (уменьшения) средней продолжительности просмотра тега индивидуальным пользователем по отношению к средней продолжительности, усреднённой по многим пользователям:

$$\delta_{ikj} = \left(M_{ikj} - \bar{M}_{ikj}\right) \Big/ \bar{M}_{ikj} \ , \ i = \overline{1,I} \ , \ k = \overline{1,K} \ , \ j = \overline{1,7} \ . \qquad (5)$$

Пространственную матрицу (5) можно представить графически матрицей столбиковых 3D-диаграмм прямоугольных матриц значений δ в таймслотах недельных блоков просмотров каждого тега. На рис. 1 приведены диаграммы относительного превышения средних длительностей просмотров тегов телепередач индивидуальным пользователем в таймслотах недельного блока по сравнению с «усреднённым» пользователем (положительный потребительский профиль пользователя).

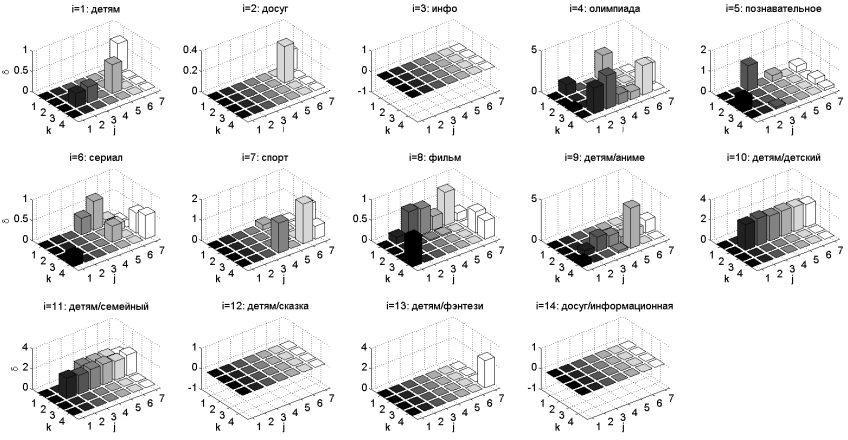

Рис. 1. Положительный потребительский профиль пользователя

Аналогично, но с отрицательным знаком ведут себя диаграммы относительного понижения средних длительностей просмотров тегов телепередач индивидуальным пользователем в таймслотах недельного блока по сравнению с «усреднённым» пользователем (отрицательный потребительский профиль пользователя).

Для проверки статистической значимости профиля пользователя можно использовать двухвыборочный асимптотически непараметрический критерий Уэлча [9, с. 248] (двухвыборочный *t*-критерий).

Построение рекомендуемого контента телепередач по потребительскому профилю пользователя

Контент телепередач, рекомендуемых пользователю рекомендательной системой линейного ТВ, должен, по-видимому, соответствовать списочному представлению потребительского профиля пользователя. Такой контент можно автоматически сформировать следующим образом.

Каждая телепередача каждого телеканала, включая имеющиеся у провайдера записи телепередач, характеризуется в стандарте xml-tv набором тегов. Каждой передаче, доступной для просмотра в *kj*-ом таймслоте недельного блока таймслотов, присваиваем рейтинг, соответствующий сумме весов тегов, входящих в неё для данного пользователя. Веса определяются потребительским профилем (5) пользователя. Полученный рейтинг можно использовать для рекомендации телевизионного контента.

Данный рейтинг, в отличие от предложенных в работах [2, 12, 13], позволяет учитывать текущий контекст просмотра путём использования «усредненного профиля», что позволяет эффективно оценивать относительную интересность передач для данного пользователя в момент трансляции.

Заключение

Предложенный алгоритм построения потребительского профиля пользователя линейного ТВ основан на сравнительном анализе средних времён просмотров тегов стандарта xml-tv ТВ-передач индивидуальным и «усреднённым» пользователями.

Алгоритм может быть эффективно использован при расчёте рейтингов ТВ-передач для индивидуального пользователя и предоставления пользователю оптимального для него контента ТВ-передач в рекомендательной системе линейного телевидения в контексте просмотров контента передач другими пользователями.

ЛИТЕРАТУРА

1. *Kim N.R.*, *Oh S.*, *Lee J.H.* A television recommender system learning a user's time-aware watching patterns using quadratic programming // Applied Sciences. 2018. V. 8(8). P. 1323.
2. *Park Y.*, *Oh J.*, *Yu H.* RecTime: Real-time recommender system for online broadcasting // Information Sciences. 2017. V. 409–410. P. 1–16.
3. *Oh S.*, *Lee J.H.* Personalized TV channel recommendation considering viewer's time dependent propensity using constrained optimization technique // Proc. 16th International Symposium Advanced Intelligent Systems, Mokpo, Korea, 4–7 November 2015. P. 817–824.
4. *Oh S.*, *Kim N.R.*, *Lee J.*, *Lee J.H.* Comparison of techniques for time aware TV channel recommendation // Proc. 2014 Joint 7th International Conference on Soft Computing and Intelligent Systems (SCIS 2014) and 15th International Symposium on Advanced Intelligent Systems (ISIS 2014), Kitakyushu, Japan, 3–6 December 2014. P. 989–992.
1. *Campos P.G.*, *Díez F.*, *Cantador I.* Time-aware recommender systems: A comprehensive survey and analysis of existing evaluation protocols // User Modeling and User-Adapted Interaction. 2014. V. 24(1–2). P. 67–119.
6. *Campos P.G.*, *Bellogin A.*, *Cantador I.*, *Diez F.* Time-aware evaluation of methods for identifying active household members in recommender systems // Advances in Artificial Intelligence: 15th Conference of the Spanish Association for Artificial Intelligence, CAEPIA 2013, Madrid, Spain, September 17–20, 2013. Proceedings. Lecture Notes in Computer Science. V. 8109. Springer, 2013. P. 22–31.
7. *Oh J.*, *Sung Y.*, *Kim J.*, *Humayoun M.*, *Park Y.H.*, *Yu H.* Time-dependent user profiling for TV recommendation // Proceedings of the 2nd International Conference on Cloud and Green Computing and 2nd International Conference on Social Computing and Its Applications, CGC/SCA 2012, Xiangtan, China, 1–3 November 2012. P. 783–787.
8. XMLTV File format. [Электронный ресурс]. URL: http://wiki.xmltv.org/index.php /XMLTVFormat.
9. *Закс Л.* Статистическое оценивание. М.: Статистика, 1976. 600 с.
10. *Айвазян С.А.*, *Фантаццини Д.* Эконометрика-2: Продвинутый курс с приложениями в финансах: Учебник. М.: Магистр; Инфра-М, 2014. 944 с.
11. *Поддубный В.В.*, *Пехтерев А.С.* Копулы сглаженных эмпирических распределений при наличии связок (совпадений) и их применение в имитационном моделировании // Труды XII Международной ФАМЭБ'2013 конференции / под ред. Олега Воробьёва. Красноярск: НИИППБ, СФУ, 2013. С. 312–321.
12. *Baklanov M.A.*, *Baklanova O.E.* Linear TV Recommender through Big Data // Data Mining and Big Data: First Int. Conf., DMBD 2016, Bali, Indonesia, June 25–30, 2016. Lecture Notes in Computer Science. V. 9714. P. 466–474.
13. *Baklanov M.A.*, *Baklanova O.E.* Methods of machine learning for linear TV recommendations // Intelligent Computing Methodologies: 12th Int. Conf., ICIC 2016, Lanzhou, China, August 2–5, 2016. Lecture Notes in Computer Science. V. 9773. P. 607–615.

Классификация текстов с использованием сверточной нейронной сети на основе векторного представления слов

И.А. Батраева, А.Д. Нарцев, А.С. Лезгян

Саратовский национальный исследовательский государственный университет им. Н.Г. Чернышевского, Саратов, Россия

Автоматизация процесса извлечения различной информации из текстов стала одной из основных проблем, связанных с информационным поиском. В частности, одной из задач анализа текстов является тематическая и жанровая классификация, которая позволяет определить принадлежность текста к определенной группе (производство, автомобили, животный мир и т.д.). Особенно актуальна такая классификация для решения задач корпусной лингвистики, так как большинство существующих на сегодняшний день корпусов такое деление по темам или жанрам текстов делают или вручную, или исходя из тематики источников текста [1]. В отдельную группу можно выделить классификацию текстов языковых корпусов, так как для них важна скорее литературная классификация текстов по темам и жанрам (песня, стихи, повествование и т.п.). Эта работа в настоящее время проводится только самими лингвистами вручную, что значительно замедляет создание электронных версий корпусов с возможностью поиска по темам и жанрам.

Дан текст на естественном языке и множество возможных жанров, к которым может принадлежать текст. Требуется определить основной жанр текста. Если текст относится к нескольким жанрам одновременно, то определить какой жанр является основным. Предполагается, что для текстов обучающего множества основная тема известна.

В качестве инструмента для решения задачи были выбраны сверточные нейронные сети (СНС), которые по результатам некоторых исследований [2] подходят для обработки текстов лучше рекуррентных нейронных сетей (РНС).

Для решения поставленной задачи требуется получить способ представления данных в виде, пригодном для обработки сверточной нейронной сетью, а именно – в виде многомерной матрицы вещественных чисел. В рамках данной работы входные данные строились на основе векторных представлений слов, полученных на основе модели word2vec.