DOI: 10.25728/tas.2019.50.4.2

ОБ ОДНОМ ПОДХОДЕ К ВЫЯВЛЕНИЮ ИНФОРМАЦИОННЫХ ПРЕДПОЧТЕНИЙ ПОЛИТИЧЕСКИ ВОВЛЕЧЕННЫХ ПОЛЬЗОВАТЕЛЕЙ ОНЛАЙНОВОЙ СОЦИАЛЬНОЙ СЕТИ

Бойко Л. М., Губанов Д. А.
(*Институт проблем управления РАН, Москва*)
boiko.lilia@gmail.com, dmitry.a.g@gmail.com

Рассматривается и решается задача идентификации информационных предпочтений информационных источников пользователей онлайновой социальной сети на примере ВКонтакте.

Ключевые слова: анализ социальных сетей, определение категорий информационных предпочтений пользователей, ДЛС-модель.

1. Введение

В данной работе проведен анализ информационных предпочтений пользователей, проявляющих свои идейно-политические интересы. Эта задача представляется довольно важной, поскольку информационные предпочтения пользователя во многом отражают его убеждения, а следовательно, и его поведение, позволяя тем самым в дальнейшем планировать и осуществлять управляющие информационные воздействия [2].

Задача идентификации информационных предпочтений может быть решена различными способами, в частности, при помощи анализа публикуемого пользователями контента в сети. Как показывают наши исследования [3], довольно небольшой процент пользователей, активно публикует политически релевантный контент, поэтому в данной работе предлагается иной способ

– идентификация предпочтений на основе подписок на информационные источники в социальной сети.

2. Подход к идентификации информационных предпочтений политически вовлеченных пользователей онлайновой социальной сети

Введем несколько необходимых для работы определений.

Во-первых, *политически вовлеченными* считаются пользователи онлайновой сети, которые подписаны хотя бы на один из выделенных экспертом учетных записей (*маркерных аккаунтов*), принадлежащих тем или иным политикам. Политические взгляды этих политиков таковы, что могут быть отнесены однозначно ровно к одному из трех классов модели Державник-Либерал-Социалист [1]. В рамках такой модели считается, что граждане РФ, проявляющие свои идейно-политические предпочтения в онлайновых социальных сетях, придерживаются этих позиций в той или иной степени.

Во-вторых, *информационными источниками* пользователей считаются публичные страницы или страницы блогеров, на которых они подписаны.

Наконец, *информационными предпочтениями* пользователей считаются интерпретируемые экспертом кластеры (классы) информационных источников.

Поставим теперь задачу идентификации информационных предпочтений политически вовлеченных пользователей онлайновой социальной сети.

Пусть задан набор источников информации $x_1, \ldots, x_n$ и некоторая неотрицательная и симметричная мера сходства s_{ij} между источниками информации. Тогда содержательно задача идентификации информационных предпочтений состоит в следующем: разбить источники информации на несколько групп таким образом, чтобы источники из одной группе были похожими, а из разных групп – непохожими друг на друга. Если нет дополнительной информации помимо сходства источников, то хорошим способом представления данных является неориентированный граф

сходства $G = (V, E)$ со множеством вершин $V = \{v_1, \ldots, v_n\}$. Вершинами этого графа являются информационные источники: каждой вершине v_i соответствует информационный источник x_i. Ребра графа соединяют такие вершины, для которых сходство s_{ij} между источниками x_i и x_j положительно или превышает определенный порог (при этом вес ребра w_{ij} равен s_{ij}).

Тогда можно переформулировать задачу идентификации предпочтений: найти такое разбиение графа сходства, для которого вес ребер между вершинами разных групп был небольшим, а вес ребер между вершинами одной группы – большим. Кроме того, желательно, чтобы размеры групп были разумно большими. Тогда формально эту задачу можно записать в виде минимизации следующей функции [7]:

$$NCut(A_1,...,A_k) = \frac{1}{2}\sum_{i=1}^{k}\frac{W(A_i, A_i \setminus A)}{vol(A_i)}$$

где $A_1, \ldots, A_k$ искомое разбиение вершин для заданного числа кластеров k, $vol(A_i) = \sum_{j \in A_i} d_j$ (d_j – степень j-ой вершины),

$$W(A,B) = \sum_{i \in A,\, j \in B} w_{ij}\ ,\ A, B \subset V\ .$$

К сожалению, такая задача минимизации является NP-трудной, поэтому для ее решения часто используется аппроксимирующий алгоритм спектральной кластеризации [5, 6].

Для решения задачи кластеризации необходимо предпринять дополнительный ряд шагов. Во-первых, необходимо выбрать меру сходства, которых насчитывается не один десяток [4]. Во-вторых, необходимо доопределить способ конструирования графа сходства [5], выбрав между графом ε-окрестности (две вершины соединяются, если сходство между ними больше заданного ε), графом k ближайших соседей (вершина соединяется с k ближайшими соседями) или полным графом (в таком графе функция близости сама должна моделировать отношение локального соседства).

3. Выборка данных

Экспертно были выделены 26 маркерных аккаунтов в онлайновой социальной сети ВКонтакте (9 державников, 9 либералов и 8 социалистов). Далее были выделены политически вовлеченные пользователи (233 тысячи), подписанные хотя бы на один из этих маркерных аккаунтов. Информационные источники таких пользователей (общим числом 2.6 млн.) и определяют в конечном итоге их информационные предпочтения.

4. Идентификация информационных предпочтений

Для выбора меры сходства и типа графа сходства далее используется следующее соображение. Мера сходства и тип графа сходства должны приводить к таким кластерам маркерных аккаунтов, которые максимально совпадали бы с эталонным разбиением – классами, заданными экспертом (державники, либералы и социалисты).

Проведем оценку качества кластеризации маркерных аккаунтов при различных комбинациях параметров (см. табл. 1).

Таблица 1. Комбинации параметров

Тип графа	Мера сходства	Число соседей k	ε-окрестность
полный граф	функция гауссова несходства $s_{ij} = \exp(-\|\|x_i - x_j\|\|^2/(2\sigma^2))$	–	–
граф ε-окрестности	Dice, Jaccard, Kulsinski, Rogers-Tanimoto, Russell-Rao, Sokal-Michener, Sokal-Sneath, Yule [5]	–	0.1, 0.2, …, 1.0
граф k ближайших соседей		1, 2, …, 20	–
граф k ближайших взаимных соседей		1, 2, …, 20	–

Наилучшие результаты получаются для графа k ближайших взаимных соседей (k = 14) и меры сходства Yule. Значение качества кластеризации (используется нормированная взаимная информация, NMI) при этом равно 0.73 (чем больше, тем лучше; максимум, равный единице, достигается, если эталонное разбиение полностью совпадает с полученной кластеризацией). Оказалось, что либеральные информационные источники выявляются четко, а между социалистами и державниками возникает частичное смешение.

Важно отметить, что популярный метод кластеризации k-means, примененный к маркерным аккаунтам, дает значение NMI близкое к 0.12.

Примечание. Мера сходства Yule [5] равна отношению $(ad-bc)/(ad+bc)$, где a – число совпадающих компонент двух векторов с значением 1, и d – число совпадающих компонент с значением 0, числа b и c определяют несовпадающие компоненты (комбинации (1, 0) и (0, 1) соответственно).

Имея «оптимальные» значения параметров, можно перейти к задаче кластеризации наиболее значимых информационных источников (расширенного перечня, включающего маркерные аккаунты) политически вовлеченных пользователей. Наиболее значимыми информационными источниками считаются те, которые имеют не менее 3000 подписчиков из числа политически вовлеченных пользователей.

Для оценки оптимального количества кластеров среди ряда вариантов ($k \in \{20, 40, 60, 80, 100\}$) использована мера силуэт (silhouette), оптимальным оказалось число кластеров k = 40 и 60. Содержательный анализ кластеров показал, что для k = 60 наблюдается более высокая однородность внутри кластеров.

Далее все кластеры были просмотрены и содержательно проинтерпретированы. Выявлены однородные кластеры, связанные с бизнесом, здоровьем, искусством, развлечениями, политикой, СМИ, религией, домоводством, строительством и ремонтом, психологией и эзотерикой, спортом, автомобилями, интересными фактами и т. д. Процент ошибочно причисленных кластерам источников составил 12%.

Таблица 2. Примеры информационных предпочтений.

	Предпочтение	Примеры источников	Число источников
1	Славянский и русский мир	"Славянское Движение", "СЛАВЯНСКИЙ ПУТЬ | РУСЬ • ТРАДИЦИИ • ИСТОРИЯ", "Вечная Слава"	38 (1.6%)
2	Ситуация на Украине	"РУССКАЯ ВЕСНА (rusvesna.su)", "Партизаны Новороссии/ДНР/ЛНР", "Сопротивление Новороссии"	41 (1.8%)
3	Религия (православие)	"Мудрые советы святых православных старцев", "Цитаты Отцов Церкви", "Верю, надеюсь, люблю †"	71 (3.0%)
4	Психология и эзотерика	"Путь Души - Саморазвитие / Психология", "Психология отношений", "Астрология Эзотерика"	175 (7.5%)
5	Домоводство, кулинария, красота	"Рецепты - Просто Повар", "Райская кухня", "Малогабаритка | Идеи для маленьких квартир"	220 (9.4%)

Работа выполнена при частичной поддержке РФФИ (проект 18-29-22042мк).

Литература

1. БЫЗОВ Л. Г. Динамика идейно-политических предпочтений за 25 лет. Три этапа трансформации общественного сознания // Россия XXI. 2019. №1. С. 6–29.

2. ГУБАНОВ Д.А., НОВИКОВ Д.А., ЧХАРТИШВИЛИ А.Г. *Социальные сети: модели информационного влияния, управления и противоборства*. 3-е изд., перераб. и дополн. М.: МЦНМО, 2018. – 224 с.
3. ГУБАНОВ Д.А., ЧХАРТИШВИЛИ А.Г. *Влиятельность пользователей и метапользователей социальной сети* // Проблемы управления. 2016. № 6. С. 12-17.
4. CHOI S. S., CHA S. H., TAPPERT C. C. *A survey of binary similarity and distance measures* //Journal of Systemics, Cybernetics and Informatics. 2010. Т. 8. №. 1. С. 43-48.
5. FOUSS F., SAERENS M., SHIMBO M. *Algorithms and models for network data and link analysis*. Cambridge University Press, 2016.
6. NG A. Y., JORDAN M. I., WEISS Y. *On spectral clustering: Analysis and an algorithm* //Advances in neural information processing systems. – 2002. – С. 849-856.
7. SHI J., MALIK J. *Normalized cuts and image segmentation* //Departmental Papers (CIS). – 2000. – С. 107.