МОДЕЛИРОВАНИЕ КОНТЕНТ-ПОВЕДЕНИЯ ПОЛЬЗОВАТЕЛЕЙ СОЦИАЛЬНЫХ СЕТЕЙ: ОСОБЕННОСТИ РУССКОЯЗЫЧНОГО СЕГМЕНТА

Кононова Е. Ю.

The purpose of the study is to examine the theoretical assumptions of trust models based on an analysis of users behavior. The article examines the distribution of users, their messages, friends and groups. Users clustering allowed identifying the types of their posting behavior: identified the clusters of "writers", "propagators" of information and its consumers – "readers", as well as a cluster of "indifferent" users. Analysis of clusters showed that content generators are not the main channel for diffusion of information, which contradicts the theoretical assumptions of trust model.

Ключевые слова: онлайновые социальные сети (ОСС), модели оценки доверия, типы поведения пользователей, карты Кохонена.

Введение. В условиях перехода к е-обществу количество ежедневно генерируемой информации растет стремительными темпами, при этом оценивать ее качество и достоверность становится все сложнее. Это ведет к формированию повышенного спроса на методы и модели оценки доверия как к источникам информации, так и к самой информации. К понятию доверия в последние годы обращались ученые из различных научных сфер – психологии, социологии, экономики, информатики. Однако, несмотря на растущее количество научных публикаций в данном направлении, исследования остаются фрагментарными и не дают целостного представления об особенностях формирования и распространения доверия в обществе.

Новое междисциплинарное направление – анализ данных социальных медиа, в рамках которого исследуются профили и поведение пользователей онлайновых социальных сетей, – охватывает такие области как машинное обучение, искусственный интеллект, визуализация данных, алгоритмы поиска информации, лингвистика и масштабные вычисления. Принимая во внимание тот факт, что пользователями социальных сетей сегодня является более четверти населения планеты, социальные данные представляют собой репрезентативный срез общества и могут составить основу для построения и верификации моделей оценки доверия.

Для моделирования доверия широко используются статистические [7, 10] и эвристические методы [5, 6], методы машинного обучения [1, 11] и анализа поведения агентов [3, 4, 7, 8, 9, 12, 13, 14]. Однако следует учитывать, что накопленный опыт моделирования поведения агентов ОСС базируется на западном (и частично китайском) опыте, в то время как русскоязычный сегмент имеет свои уникальные особенности.

Целью данного исследования является проверка гипотезы моделей доверия о том, что генераторы контента являются основным каналом распространения информации в русскоязычном сегменте ОСС. Исследование построено на основе анализа 248 тыс. профилей пользователей сети «ВКонтакте» (ВК) и 238 тыс. профилей сети «Одноклассники» (ОК), а также контенте – сообщениях и комментариях – этих пользователей. Собранная база данных включает следующие записи: идентификатор пользователя; список друзей пользователя; список групп, в которых состоит пользователь; открытые сообщения пользователя (посты и комментарии).

Подготовка и первичный анализ данных. На первом этапе исследования набор данных был очищен от неинформативных сообщений. Затем, используя модуль *tokenizer* библиотеки *nltk python*, в сообщениях пользователей были выделены отдельные слова и удалены посторонние символы. На основе библиотеки *pymorphy2* слова были лемматизированы, все буквы в словах переведены в нижний регистр. После предварительной подготовки для каждого пользователя были рассчитаны его характеристики и число постов на заданную тему. В результате была сформированы таблицы следующего вида (табл. 1).

Таблица 1. Фрагмент данных базы ВК

Идентификатор	Число друзей	Число групп	Число постов	Число постов друзей
195269137	49	3	992	1439
187704426	53	10	949	62
17241807	193	1	925	593
35833523	332	1	789	726
906761	300	3	773	291
23419701	273	4	754	55
51258489	536	10	735	754

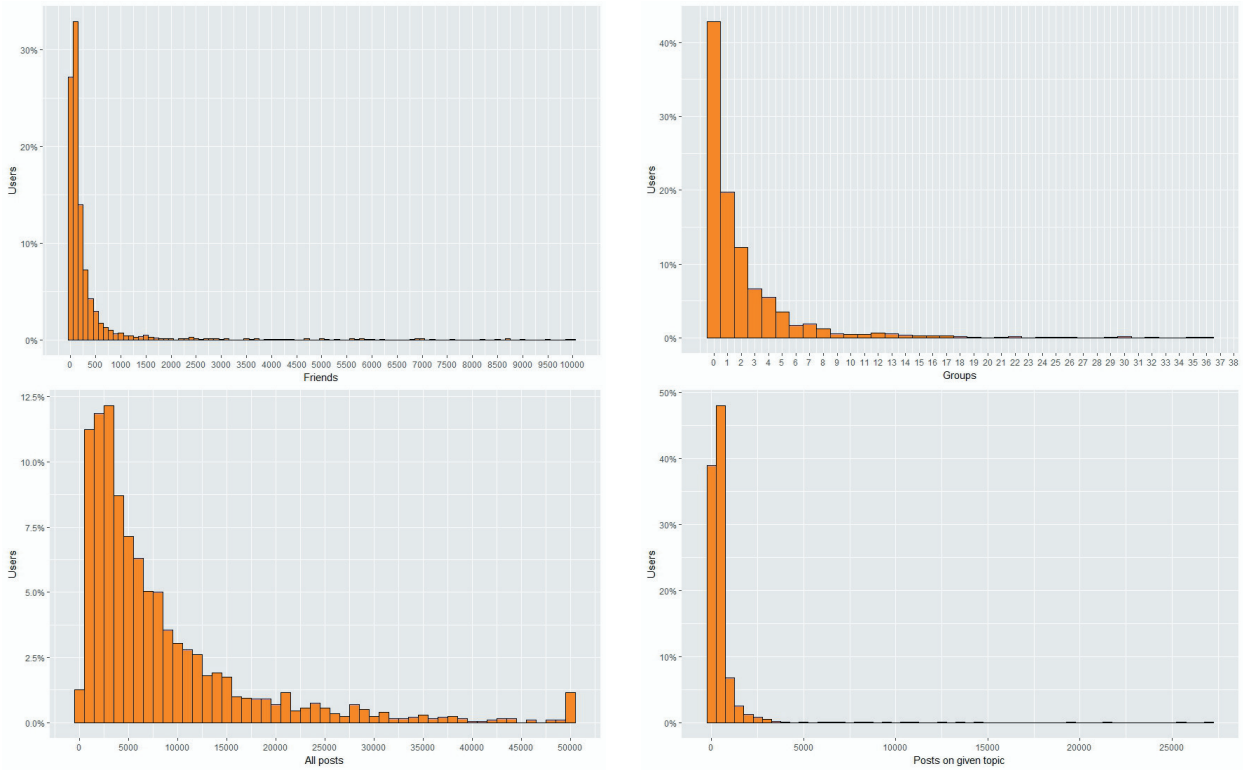


Рис. 1. Частотные характеристики пользователей ВК

Характеристики пользователей обоих сетей описываются, как и большинство распределений в Интернет [2], степенным законом (рис. 1).

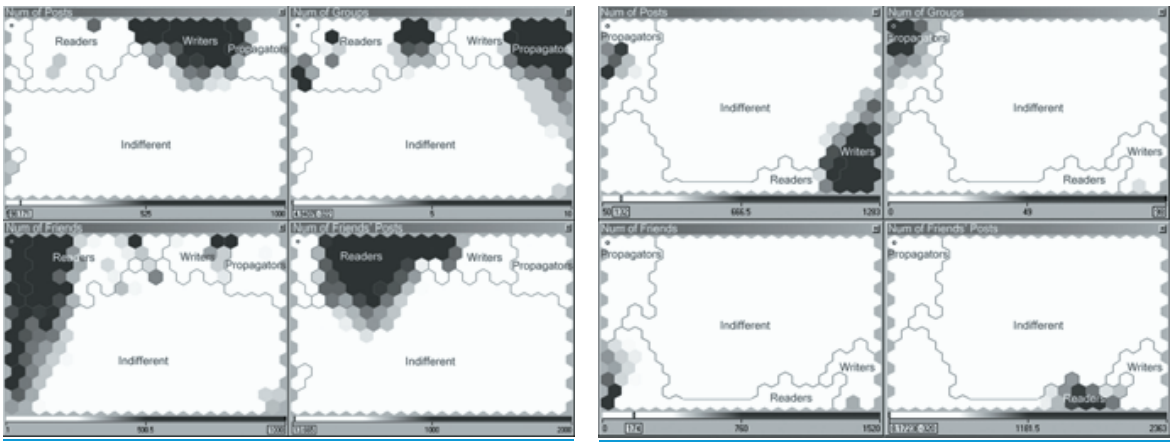
Как показало предварительное исследование, в обеих сетях часть пользователей имеет колоссальное количество друзей, некоторые состоят в большом числе групп, некоторые размещают большое число собственных постов или сообщений друзей. В связи с этим, а также опираясь на теоретические предположения моделей оценки доверия, может быть поставлен ряд вопросов: состоят ли в группах те пользователи, у кого много друзей; много ли друзей у тех, кто много пишет; все ли темы создаются идентичным образом (например, за счет написания постов или ре-постов от друзей / из групп). Для ответа на эти вопросы была проведена идентификация типов поведения пользователей ОСС.

Анализ типов поведения пользователей. Для выявления и описания типов поведения были использованы методы кластеризации (в частности карты Кохонена, позволяющие выявлять скрытые закономерности в данных). В качестве параметров кластеризации были выбраны следующие: количество друзей; количество тематических постов; количество тематических групп, в которых состоит пользователь; количество тематических постов на стенах друзей пользователя.

После серии экспериментов для обеих ОСС были выявлены аналогичные кластера, позволившие выявить паттерны поведения пользователей обеих сетей (рис. 2).

При этом структура нейронной сети идентична в обоих случаях (рис. 3).

Рис. 2. Паттерны поведения пользователей на картах Кохонена: справа ВК, слева ОК



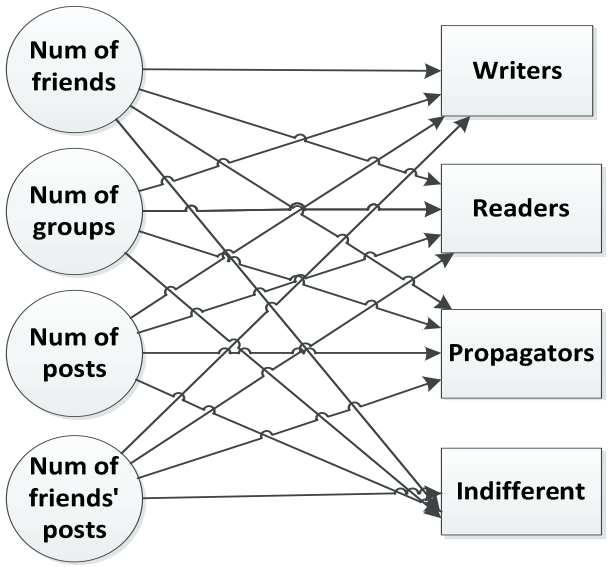


Рис. 3. Архитектура нейронной сети идентификации типов поседения пользователей

Анализ профилей кластеров, представленных на рис. 2, позволил описать следующие типы контент-поведения пользователей: «писатели», «распространители», «читатели» и «малоактивные».**ю**Кластер «писателей» в обеих ОСС самый маленький, однако его участники генерируют основной контент по заданной тематике. Кроме того, они чаще других оставляют комментарии. Следующим по численности является кластер «распространителей» – это пользователи с наибольшим количеством ре-постов из тематических групп. В отличие от других пользователей, у них мало друзей, они сосредоточены на сборе информации из тематических групп и ее дальнейшем распространении.

Среди людей, которые активно интересуются данной темой, наиболее многочисленным является кластер «читателей», новостные ленты которых состоят из сообщений друзей. В отличие от «распространителей», они ориентированы скорее на потребление информации, чем на ее распространение.

Хотя большинство пользователей обеих ОСС оказались слабо заинтересованы в исследуемой тематике (они и писали, и читали, и ре-постили относительно редко), однако некоторую активность в рамках тематики они все же продемонстрировали, отчего и попали в общий список. А значит, их можно рассматривать как потенциальную аудиторию.

Выводы. В работе проведен анализ поведения пользователей онлайновых социальных сетей. В результате кластеризации агентов с использованием сетей Кохонена идентифицированы и описаны типы их поведения: выявлены кластера «писателей», «распространителей» информации и ее потребителей – «читателей», а также кластер «малоактивных» пользователей. Исследование показало, что хотя кластер «писателей» является самым малочисленным, именно эти пользователи генерируют основную часть контента. Однако, вопреки теоретическим предположениям модели оценки доверия, друзей у пользователей этого кластера не так много, как можно было бы ожидать. Этот результат противоречит предположениям западных моделей оценки доверия и свидетельствует о целесообразности продолжения исследований русскоязычного сегмента ОСС.

ЛИТЕРАТУРА

1. Abdul-Rahman, A., & Hailes, S. (2000, January 7). Supporting trust in virtual communities. *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences (Maui, Hawaii: IEEE Computer Society)*, 1769-1777. DOI: 10.1109/HICSS.2000.926814

2. Adamic, L.A., & Huberman, B.A. (2000). Power-law distribution of the World Wide Web. *Science, 287*, 2115–2115.

3. Gruhl, D., Guha, R., Liben-Nowell, D., & Tomkins, A. (2004). *Information diffusion through blogspace.* Retrieved from http://people.csail.mit.edu/dln/papers/ blogs/idib.pdf.

4. Guo, L., Tan, E., Chen, S., Zhang, X., & Zhao, Y. (2009). *Analyzing patterns of user content generation in online social networks.* Retrieved from https://cs.gmu.edu/~sqchen/ publications/kdd09.pdf.

5. Huynh, T.D., Jennings, N.R., & Shadbolt, N.R. (2006). Certified reputation: How an agent can trust a stranger. *Proceedings of the 5th International Joint Conference on Autonomous Agents and MultiagentSystems (New York),* 1217-1224.

6. Josang, A., & Ismail, R. (2002, June 17 - 19). The beta reputation system. *Proceedings of the 15th Bled Electronic Commerce Conference (Bled, Slovenia)*, 891-900.

7. Kuan, H., & Bock, G. (2005, May 26-28). The collective reality of trust: An investigation of social relations and networks on trust in multi-channel retailers. *Proceedings of the 13th European Conference on Information Systems (Regensburg; Germany),* 1-8.

8. Kumar, R., Novak, J., Raghavan, P., & Tomkins, A. (2003). *On the bursty evolution of blogspace.* Retrieved from http:// www.disco.ethz.ch/lectures/fs12/ seminar/paper/Barbara/32.pdf.

9. Liu, J., Dolan, P., & Pedersen, E. R. (2010). *Personalized news recommendation based on click behavior.* Retrieved from http://cs.northwestern.edu/~jli156/IUI224-liu.pdf.

10. Malik, Z., Akbar, I., & Bouguettaya, A. (2009). Web services reputation assessment using a hidden markov model. *Proceedings of the 7th International Joint Conference on Service-Oriented Computing*, 576-591.

11. Mui, L. (2003). *Computational models of trust and reputation: Agents, evolutionary games, and social networks.* Retrieved from http://groups.csail.mit.edu/medg/people/lmui/ docs/phddissertation.pdf.

12. Papagelis, M., Murdock, V., & van Zwol, R. (2011). *Individual behavior and social influence in online social systems.* Retrieved from http://www.cs.toronto.edu/ ~papaggel/docs/ papers/all/HT11-Individual-Behavior-and-Social-Influence-in-Online-Social-Systems.pdf.

13. Roman, P. E., Gutierrez, M.E., & Rios, S.A. (2012). *A model for content generation in On-line social network.* Retrieved from https://www.researchgate.net/ publication/233897861.

14. Xu, Z., Zhang, Y., Wu, Y., & Yang, Q. (2012). *Modeling User Posting Behavior on Social Media.* Retrieved from http:// yaowu.co/docs/sigir12.pdf.