Автоматический метод анализа множества ассоциативных правил в исследовании деятельности общественных организаций

С.В. Белим
belimsv@omsu.ru

Т.Б. Смирнова
SmirnovaTB@omsu.ru

А.Н. Мироненко
MironenkoAN@omsu.ru

Омский государственный университет им. Ф.М. Достоевского, Омск, Россия

Аннотация

В статье предложен метод выявления взаимосвязи между объектами на основе анализа множества ассоциативных правил. Ассоциативные правила формируются на основе обработки анкет. Для формирования ассоциативных правил используется алгоритм Apriori. Для множества ассоциативных правил строится граф взаимосвязей. Каждое ассоциативное правило определяет взвешенное ребро графа. Для графа выполняется поиск сообществ. Сообщества показывают наиболее сильно связанные факторы исходного множества.

Введение

Традиционное использование метода формирования ассоциативных правил из набора данных состоит в исследовании особенностей покупательского спроса на товары в торговых сетях [1, 2]. Информация о взаимосвязи интереса к покупке различных товаров, используется в дальнейшем для их позиционирования в торговых залах. В последнее время наблюдается расширение сферы использования данного метода. В качестве примеров можно привести обработку изображений [3, 4], анализ распространения биологических видов [5], поиск точек интереса горной промышленности по фотографиям поверхности [6] и другие.

Для социологических исследований поиск ассоциативных правил начал использоваться относительно недавно. Наиболее масштабные исследования предприняло Европейское социологическое общество ESS для выявления влияния страны проживания людей на их устоявшиеся стереотипы [7]. Достаточно большое число работ посвящено анализу социальных сетей с помощью поиска ассоциативных правил. Так в статье [8] на основе ассоциативных правил выявляется взаимное влияние пользователей социальных сетей. В дальнейшем выделяются наиболее влиятельные пользователи, которые могут быть использованы для эффективного распространения информации. В работах [9, 10] на основе ассоциативных правил, формируемых из записей в Facebook, установливается связь между полом студента и выбранными учебными курсами. Аналогичные исследования [11], основанными на записях в Facebook, позволили изучить влияние социальной сети на образовательный процесс студентов Турции. Аналогичные исследования в социальных сетях [12] позволили выявить факторы, влияющие на выбор хобби. В статье [13] поиск ассоциативных правил использован для анализа криминалистических данных. Авторы этой работы исследовали взаимосвязь побудительных мотивов и вида преступления. В статье [14] ассоциативные правила использованы для построения социальной сети, основанной на базе данных о террористических атаках. В статье [15]

Copyright © by the paper's authors. Copying permitted for private and academic purposes.

In: Sergey V. Belim, Nadezda F. Bogachenko (eds.): Proceedings of the Workshop on Data, Modeling and Security 2017 (DMS-2017), Omsk, Russia, October 2017, published at http://ceur-ws.org

выявлены ассоциативные правила, связывающие характер человека и его склонность к наркомании. Поиск ассоциативных правил осуществляется на базе данных наркологических клиник. Применение поиска ассоциативных правил к данным переписи населения в рамках проекта SPIN представлена в статье [16].

Данная статья посвящена определению закономерностей в деятельности общественных организаций российских немцев на основе базы данных, полученной с помощью анкетирования. База данных анкет используется для построения ассоциативных правил и выявления закономерностей между различными аспектами деятельности общественных организаций.

1 Постановка задачи и методы решения

Для использования метода построения ассоциативных правил необходимо сформировать множество возможных записей и список транзакций. В данной работе анализировалась деятельность общественных организаций российских немцев в различных регионах. Исходная информация бралась из анкет, заполнявшихся руководителями организаций. В качестве транзакции выбиралась одна анкета. Каждая анкета содержала вопросы различного формата. Первый тип вопросов предполагал два варианта ответа «Да» или «Нет». Второй тип вопросов допускал выбор одного из четырех или пяти вариантов, причем допускался дополнительный ответ «прочее», который не кодировался, так как содержал неопределенность. Также в анкете присутствовали вопросы с недетерминированным ответом, но они не учитывались при формировании транзакций, так как все ответы были различными и не могли привести к выявлению ассоциативных правил со сколько-нибудь заметной поддержкой. Каждый вариант ответа кодировался своей записью. Вопрос обозначался идентификатором из одной или двух латинских букв. Например, вопрос, обозначаемый «A», имел два варианта ответа, кодируемых «A1» и «A2». В вопросах, содержащих выбор из двух альтернатив, нельзя ограничиваться кодированием только одной из них, не смотря на возможность однозначного восстановления второй. Такое ограничение может приводить к потере ассоциативных правил.

Пусть I – множество всех ответов, которые могут присутствовать в транзакции. Каждая транзакция T – это набор элементов из I ($T \subseteq I$). D – множество всех транзакций. Говорят, что транзакция T содержит набор элементов X, если $X \subseteq T$ и $X \subseteq I$. Ассоциативным правилом называется импликация $X \Rightarrow Y$, где $X \subseteq I$, $Y \subseteq I$ и $X \cap Y = \emptyset$.

Каждое ассоциативное правило характеризуется некоторым набором параметров. Первый параметр, называемый поддержкой, показывает частоту встречаемости данного правила в имеющемся наборе транзакций. Поддержка правила $X \Rightarrow Y$ вычисляется как процент транзакций, содержащий множество $X \cup Y$:

$$supp(X \Rightarrow Y) = \frac{N(X \cup Y)}{|D|} \cdot 100\%,$$

где $N(X \cup Y)$ – количество транзакций, содержащих множество $X \cup Y$.

Достоверность правила показывает, с какой вероятностью из X следует Y. Достоверность ассоциативного правила $X \Rightarrow Y$ вычисляется как процент транзакций, содержащих как X, так и Y, в множестве транзакций, содержащих X:

$$conf(X \Rightarrow Y) = \frac{supp(X \Rightarrow Y)}{supp(X)}.$$

Задача поиска ассоциативных правил состоит в нахождении наборов элементов, поддержка которых не ниже чем minsupport. Из найденных наборов выделяются правила с достоверностью не ниже minconfidence.

2 Ассоциативные правила

В общей сложности было обработано 107 анкет, каждая из которых рассматривалась как независимая транзакция. После кодирования были получены транзакции с различным числом записей от 24 до 50. На основе данных транзакций был осуществлен поиск ассоциативных правил с поддержкой не менее 60% и достоверностью не менее 80%. Для поиска ассоциативных правил был использован алгоритм APriori [17]. Ассоциативные правила, удовлетворяющие данным свойствам, приведены в Таблицах 1 и 2.

Ассоциативные правила в Таблицах 1 и 2 необходимо интерпретировать в формате:

«**Если** Предпосылка, **то** Следствие».

Анализ показывает, что поиск ассоциативных правил позволяет выявить как достаточно очевидные взаимосвязи между различными аспектами деятельности общественных организаций, так и достаточно неожиданные влияния

Таблица 1: Ассоциативные правила по деятельности общественных организаций российских немцев

N	Предпосылка	Следствие	*supp*	*conf*
1	Более 50% посетителей центра являются российскими немцами	Интернет используется в работе центра несколько раз в день	61,68	80,72
2	Более 50% посетителей центра являются российскими немцами	В организации используются языковые курсы для взрослых	60,75	80,72
3	Более 50% посетителей центра являются российскими немцами	В организации знают о том, что Германия осуществляет специальные программы	62,62	80,72
4	Более 50% посетителей центра являются российскими немцами	Знания немецкого языка за последние 10 лет улучшились	62,62	81,93
5	Интернет используется в работе центра несколько раз в день	Знания немецкого языка за последние 10 лет улучшились	62,62	80,00
6	В организации используются языковые курсы для взрослых	Более 50% посетителей центра являются российскими немцами	63,55	82,72
7	В организации используются языковые курсы для взрослых	Интернет используется в работе центра несколько раз в день	62,62	82,72
8	В организации используются языковые курсы для взрослых	Знания немецкого языка за последние 10 лет улучшились	63,55	85,19
9	Знания немецкого языка за последние 10 лет улучшились	Более 50% посетителей центра являются российскими немцами	62,62	80,95
10	Знания немецкого языка за последние 10 лет улучшились	Интернет используется в работе центра несколько раз в день	62,62	80,95
11	Знания немецкого языка за последние 10 лет улучшились	В организации используются языковые курсы для взрослых	64,49	82,14
12	Партнерами в работе являются организации в России	Интернет используется в работе центра несколько раз в день	62,62	83,54
13	Партнерами в работе являются организации в России	Знания немецкого языка за последние 10 лет улучшились	63,55	82,28

Таблица 2: Ассоциативные правила по деятельности общественных организаций российских немцев (продолжение)

N	Предпосылка	Следствие	*supp*	*conf*
14	Сотрудничество с МСНК осуществляется постоянно с высокой степенью эффективности	Интернет используется в работе центра несколько раз в день	63,55	88,16
15	В организации знают о том, что Германия осуществляет специальные программы	Более 50% посетителей центра являются российскими немцами	64,49	80,72
16	Для изучения немецкого языка используются детские и молодежные языковые клубы	Знания немецкого языка за последние 10 лет улучшились	50,47	90,00
17	Партнерами в работе являются организации в России **и** Сотрудничество с МСНК осуществляется постоянно с высокой степенью эффективности	Интернет используется в работе центра несколько раз в день	51,40	90,16
18	Сотрудничество с МСНК осуществляется постояннос высокой степенью эффективности **и** Предлагаются программы целевой направленности для всех возрастов	Интернет используется в работе центра несколько раз в день	51,40	96,49
19	Предлагаются программы целевой направленности для всех возрастов **и** Интернет используется в работе центра несколько раз в день	Сотрудничество с МСНК осуществляется постоянно с высокой степенью эффективности	51,40	90,16
20	Сотрудничество с МСНК осуществляется постоянно с высокой степенью эффективности **и** Знания немецкого языка за последние 10 лет улучшились	Интернет используется в работе центра несколько раз в день	52,34	90,32

Таблица 3: Обозначения для утверждений, встречающихся в ассоциативных правилах

v_1	Более 50% посетителей центра являются российскими немцами
v_2	Интернет используется в работе центра несколько раз в день
v_3	В организации используются языковые курсы для взрослых
v_4	Знания немецкого языка за последние 10 лет улучшились
v_5	Партнерами в работе являются организации в России
v_6	Сотрудничество с МСНК осуществляется постоянно с высокой степенью эффективности
v_7	В организации знают о том, что Германия осуществляет специальные программы
v_8	Для изучения немецкого языка используются детские и молодежные языковые клубы

факторов друг на друга. При этом правила с одним утверждением в предпосылке обладают большей поддержкой, но меньшей достоверностью. Ассоциативные правила, содержащие конъюнкцию двух утверждений в предпосылке, характеризуются меньшей поддержкой, но очень высокой достоверностью. Следует отметить, что в формировании ассоциативных правил участвует всего 9 утверждений из 173 возможных. Между остальными утверждениями ассоциативные правила с достаточно высокими поддержкой и достоверностью отсутствуют.

3 Теоретико-графовый анализ

Построим граф связей на основе выявленных ассоциативных правил. В таблице 3 приведены обозначения для утверждений, встречающихся в ассоциативных правилах.

На рисунке 1 приведен граф, построенный на основе ассоциативных правил с одной предпосылкой. В качестве веса ребер использованы значения достоверности.

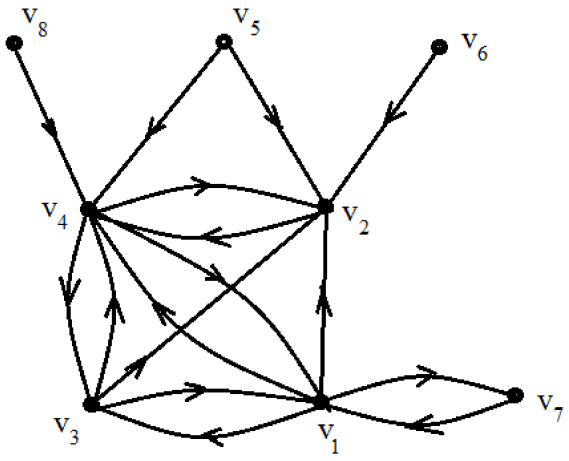

Рис. 1: Граф связи утверждений ассоциативными правилами

Матрица весов построенного графа приведена в Таблице 4. Вес отсутствующих дуг равен нулю.

Применим к данному графу алгоритм поиска сообществ (community) [18]. Для этого необходимо предварительно получить приведенный вид матрицы весов $e = E/m$, где

$$m = \sum_{i,j=1}^{N} E_{ij}.$$

N – количество вершин. В приведенной матрице весов элемент e_{ij} показывает долю веса заданного ребра в общем весе графа. В дальнейшем под матрицей весов будет пониматься именно приведенный вид. Легко увидеть, что

$$\sum_{i,j=1}^{N} e_{ij} = 1.$$

Таблица 4: Матрица весов графа ассоциативных правил

	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8
v_1	0	80,72	80,72	81,93	0	0	80,72	0
v_2	0	0	0	80,00	0	0	0	0
v_3	82,72	82,72	0	85,19	0	0	0	0
v_4	80,95	80,95	82,14	0	0	0	0	0
v_5	0	83,54	0	82,28	0	0	0	0
v_6	0	88,16	0	0	0	0	0	0
v_7	80,72	0	0	0	0	0	0	0
v_8	0	0	0	90,00	0	0	0	0

Для выявления сообществ используется функция модульности (modularity), показывающая оптимальность разбиения графа на подграфы:

$$Q(e) = \sum_{i=1}^{N} e_{ii} - \sum_{i=1}^{N} p_{ii},$$

где p_{ii} – «ожидаемая связность». В канонической модели [18] p_{ii} определяется через исходящую степень вершины a_i, и входящую степень вершины b_i:

$$p_{ii} = a_i b_i.$$

В этом случае модульность записывается в виде:

$$Q(e) = \sum_{i=1}^{N} e_{ii} - \sum_{i=1}^{N} a_i b_i,$$

где

$$a_i = \sum_{j=1, j\neq i}^{N} e_{ij}, \quad b_i = \sum_{j=1, j\neq i}^{N} e_{ji}.$$

Для поиска сообществ на графах используется алгоритм образования стяжек. Выделим в графе G подграф G' и заменим все его вершины одной вершиной, при этом вершины подграфа $G\backslash G'$ остаются неизменными. Образованная вершина связана дугами с теми вершинами графа $G\backslash G'$, с которыми были связаны вершины, вошедшие в стяжку. Вес вершины, вошедшей в стяжку равен сумме весов вершин и дуг, вошедших в стяжку.

Под сообществом будем понимать подграф исходного графа, который при образовании из него стяжки максимизирует функцию модульность графа $Q(e)$. Нашей задачей является выявление сообществ на графе ассоциативных правил. В силу того, что исходный граф имеет малое количество вершин, задача может быть решена полным перебором.

Функция модульности исходного графа равна $Q = -0,1497$. Объединение вершин v_1, v_3, v_4 в одно сообщество приводит к значению функции модульности $Q_{1,3,4} = -0,0361$, то есть такое объединение является выгодным и показывает тесную связь этих вершин. Объединение в одно сообщество вершин v_1, v_3, v_4 и v_2 приводит к значению модульности $Q_{1,2,3,4} = -0,0233$. Остальные варианты объединения не повышают функцию модульности, то есть не являются выгодными.

Из этого анализа можно сделать вывод, о тесной связи таких аспектов деятельности общественных организаций российских немцев: «Более 50% посетителей центра являются российскими немцами», «Интернет используется в работе центра несколько раз в день», «В организации используются языковые курсы для взрослых», «Знания немецкого языка за последние 10 лет улучшились». Эти четыре направления деятельности наиболее тесно взаимосвязаны между собой и их надо рассматривать в совокупности.

4 Выводы

Предложенный в данной статье подход, основанный на поиске ассоциативных правил и дальнейшем представлении связей между вопросами анкеты в виде ориентированного графа, позволяет выявить закономерности, проявляющиеся в деятельности общественных организаций. Анализ графа взаимосвязей с помощью поиска сообществ вершин дает возможность определять наиболее тесно взаимосвязанные аспекты деятельности общественных организаций.

Список литературы

[1] R. Agrawal, T. Imielinski, A. Swami. Mining Association Rules between Sets of Items in Large Databases. *In Proc. of the 1993 ACM SIGMOD International Conference on Management of Data*, Washington DC, USA, 207–216, 1993.

[2] M. Shaheen, M. Shahbaz, A. Guergachi. Context Based Positive and Negative Spatio Temporal Association Rule Mining. *Elsevier Knowledge-Based Systems*, 261–273, 2013.

[3] S.V. Belim, A.O. Mayorov-Zilbernagel. Image Restoration With Static Gaps On The Basis Of Association Rules. *Herald of computer and information technologies*, 12:18–23, 2014.

[4] S.V. Belim, A.O. Mayorov-Zilbernagel. Algorithm for Searching the Broken Pixels and Eliminating Impulse Noise in Images Using a Method of Association Rules. *Science and Education of the Bauman MSTU*, 12:716–737, 2014. URL: `http://technomag.bmstu.ru/doc/744983.html`.

[5] M.S. Atepalikhin, B.Yu. Kassal, S.V. Belim. The identification of interrelation of habitation of species by association rules. *Herald of Omsk university*, 2(72):25–29, 2014.

[6] I. Lee, G. Cai, K. Lee. Mining Points-of-Interest Association Rules from Geo-tagged Photos. *46th Hawaii International Conference on System Sciences*, 1580–1588, 2013.

[7] European Social Survey. *Sampling for the European Social Survey Round VI: Principles and Requirements Mannheim, European Social Survey*, GESIS, 2012.

[8] F. Erlandsson, P. Brodka, A. Borg, H. Johnson. Finding Influential Users in Social Media Using Association Rule Learning. *arXiv:1604.08075v2*, 2016.

[9] F. Erlandsson, A. Borg, H. Johnson, P. Brodka. Predicting User Participation in Social Media. *In Advances in Network Science*, Springer International Publishing: Cham, Switserland, 126–135, 2016.

[10] P. Nancy, G.R. Ramani, S. Jacob. Mining of Association Patterns in Social Network Data (Face Book 100 Universities) through Data Mining Techniques and Methods. *In Advances in Computing and Information Technology*, Springer: Berlin/Heidelberg, Germany, 178:107–117, 2013.

[11] A.S. Bozkir, S.G. Mazman, E.A. Sezer. Identification of User Patterns in Social Networks by Data Mining Techniques: Facebook Case. *IMCW 2010*, 145–153.

[12] X. Yu, H. Liu, J. Shi, J.N. Hwang, W. Wan, J. Lu. Association Rule Mining of Personal Hobbies in Social Networks. *In Proceedings of the 2014 IEEE International Congress on Big Data (BigData Congress)*, Anchorage, AK, USA, 310–314, 2014.

[13] B.L. Pereira, W.C. Brandao. ARCA: Mining crime patterns ising association rules. *IADIS International Conference Applied Computing 2014 (IADIS AC2014)*, 2014.

[14] J. Gorecki, K. Slaninova. Building synthetic social network using association rules and clustering methods: case study on global terrorism database. *Acta academica karviniensia*. URL: `http://www.slu.cz/opf/cz/informace/acta-academica-karviniensia/casopisy-aak/aak-rocnik-2013/docs-3-2013/Gorecki_Slaninova.pdf`.

[15] F. Zahedi, M.R. Zare-Mirakabad. Employing data mining to explore association rules in drug addicts. *Journal of AI and Data Mining*, 2(2):135–139, 2014.

[16] D. Malerba, F. Esposito, F.A. Lisi. Mining Spatial Association Rules in Census Data. *Research in Official Statistics*, 1:19–45, 2002.

[17] R. Agrawal, R. Srikant. Fast Discovery of Association Rules. *In Proc. of the 20th International Conference on VLDB*, Santiago, Chile, 487–499, 1994.

[18] M.E.J. Newman. Mixing patterns in networks. *Phys. Rev. E.*, 67:026126-1–026126-13, 2003.

Automatic Method of the Associative Rules Set Analysis in a Public Organizations Activities Research

Sergey V. Belim, Tatyana B. Smirnova, Anton N. Mironenko

In article the correlation between objects detection method on the basis of the associative rules set analysis is suggested. The associative rules are created on the basis of processing of questionnaires. For a associative rules formation the algorithm Apriori is used. For the associative rules set the correlations graph is constracted. Each associative rule defines the weighed edge in the graph. For the graph a communities search is executed. Communities show the most strongly connected factors in the initial set.