DOI 10.37882/2223–2966.2022.02.22

ЦИФРОВАЯ ТРАНСФОРМАЦИЯ И ГРАФИЧЕСКАЯ ИНТЕРПРЕТАЦИЯ СОЦИАЛЬНЫХ ИНТЕРЕСОВ СТУДЕНТОВ ВУЗА

DIGITAL TRANSFORMATION AND GRAPHIC INTERPRETATION OF SOCIAL INTERESTS OF UNIVERSITY STUDENTS

S. Makhnev
N. Dekanova

Summary. The proposed work is the initial stage of a study devoted to the analysis of textual data obtained from social networks in order to identify relationships between graduates of various specialties of the university and incoming applicants. The semantic-graphical method of digital transformation and graphical interpretation of unstructured text data is considered. The basis of the semantic method is a combination of empirical and theoretical approaches to determining the amount of information K Shannon and Poisson's uniform distribution. The resulting characteristics make it possible to single out the most valuable social interests of the study participants. The use of trigonometric and exponential functions as guides in the graphical interpretation of weight coefficients is proposed. The totality of the obtained matrices of weight coefficients and raster images is the input data for the subsequent formation of a trained neural network designed to recognize the proximity of the applicant's social interests to the collection of interests of graduates of certain specialties.

Keywords: text element, semantic analysis, amount of information, weight factor, term set, set collection, graphical representation, bitmap.

Махнев Сергей Александрович
Аспирант, ФГБОУ ВО «Иркутский государственный университет путей сообщения»
still-1994@mail.ru
Деканова Нина Петровна
Д.т.н, профессор, ФГБОУ ВО «Иркутский государственный университет путей сообщения»
dekhan@yandex.ru

Аннотация. Предлагаемая работа является начальной стадией исследования, посвященного анализу текстовых данных, получаемых из социальных сетей с целью выявления взаимосоответствий между выпускниками различных специальностей ВУЗа и поступающими абитуриентами. Рассматривается семантико-графический метод цифровой трансформации и графической интерпретации неструктурированных текстовых данных. Основа семантического метода — комбинация эмпирического и теоретического подходов к определению количества информации К Шеннона и равномерного распределения Пуассона. Получаемые характеристики позволяет выделить наиболее ценные социальные интересы участников исследования. Предложено использование тригонометрических и показательных функций, как направляющих в графической интерпретации весовых коэффициентов. Совокупность полученных матриц весовых коэффициентов и растровых изображений составляет входные данные для последующего формирования обучаемой нейронной сети, предназначенной для распознавания близости социальных интересов абитуриента к коллекции интересов выпускников отдельных специальностей.

Ключевые слова: текстовый элемент, семантический анализ, количество информации, весовой коэффициент, набор терминов, коллекция наборов, графическое представление, растровое изображение.

В современном мире широкое распространение получили практические задачи, методы решения которых опираются на обработку больших объемов данных с целью выявления определенных закономерностей. К таким задачам относятся проблемы, связанные с обработкой естественного языка, например, поиск информации, распознавание образов или речи, обнаружение нечетких дубликатов для текстовых документов и ряд других приложений. Одним из актуальных направлений является анализ структуры социальных графов и текстовых данных, получаемых из социальных сетей, с целью исследования взаимодействий между участниками сети. В ряде работ приводятся методы сбора и статистического анализа данных социальных сетей [2, 1]. В сфере образования несомненный интерес представляет распознавание отличительных социальных образов для студентов различных специальностей высших учебных заведений. Близость социальных интересов абитуриента к социальному «портрету» группы студентов определенной специальности может служить дополнительным аргументом для абитуриента в выборе направления обучения.

Основополагающей задачей в решении данной проблемы является цифровая трансформация и графическая интерпретация представленных в сети интернет социальных интересов выпускников различных специальностей, успешно прошедших обучение в университете.

Задачи обработки и анализа текстовой информации основываются на семантических методах. Важным фактором при работе с текстовой информацией является точность и полнота выделяемой содержательной ча-

сти, надёжное распознавание элементов и поиск смысловой пары. Существует множество методов семантического анализа текстов, основанных на смысловых значения единиц языка, как среди отечественных, так и среди зарубежных разработок [5–8]. Задача поиска зависимостей среди неструктурированных социальных интересов студентов, обучающихся по различным специальностям, является сложной задачей. Социальные интересы участника сети интернет включают слова и словосочетания из разных сфер жизнедеятельности (спорт, музыка, различные группы по интересам). Далее под словом «термин» понимается слово, или словосочетание, представляющее собой единицу информации в исследовании; набор терминов — уникальный набор социальных интересов отдельного участника сети; коллекция — совокупность наборов терминов, относящиеся к студентам некоторой специальности. Термины в наборе не структурированы и расположены в порядке добавления по времени.

Статистическая интерпретация специфичности терминов основана на методах математической статистики и теории информации. Алгоритм включает расчет частоты повторений терминов в рассматриваемом наборе и в коллекции в целом. Пусть имеется коллекция, состоящая из N наборов терминов, представляющих социальные интересы студентов-выпускников, относящихся к некоторой специальности. Совокупность различных терминов коллекции составляет множество терминов данной коллекции $R = \{r_1, r_2, \dots, r_K\}$. Обозначим через $cr_k^i, \forall i = 1,2, \dots, N; k = 1,2, \dots, K$ — число повторений термина r_k в i-м наборе терминов. Величина cr_k^i представляет собой оценку важности термина в пределах отдельного набора терминов [9]. Строится частотный словарь терминов коллекции

$$C = \{[r_1, CR_1], [r_2, CR_2], \dots, [r_K, CR_K]\},$$

где CR_k — число вхождений некоторого термина $r_k, k = 1, 2, \dots, K$ в коллекцию равно:

$$CR_k = \sum_{i=1}^{N} cr_k^i. \quad (1)$$

Эмпирическая оценка количества информации, содержащейся в сведении о том, что данный термин $r_k, k = 1, 2, \dots, K$ входит в некоторый набор терминов коллекции по меньшей мере один раз — величина IDF_k^E, согласно определению К. Шеннона, формируется следующим образом:

$$IDF_k^E = -log_2 Pr_k(cr_k^i \geq 1) = -log_2\left(\frac{D_k}{N}\right), \quad (2)$$

где $Pr_k(cr_k^i \geq 1)$ — доля наборов коллекции, в которые входит термин r_k хотя бы один раз; D_k — число наборов коллекции, включающих термин r_k хотя бы один раз. В информационном поиске принято определять вес термина, исходя из величины IDF_k^E, однако, как показано в [10], лучшие результаты обеспечивает величина, представляющая собой разницу между эмпирическим значением IDF_k^E и значением, предсказанным согласно модели распределения Пуассона. Если предположить, что термины в коллекции распределяются случайным и независимым образом, равномерно рассеиваясь с некоторой средней плотностью, то соответствующее количество информации равно:

$$IDF_k^P = -log_2 Pr_k^P(cr_k^i \geq 1) = \\ = -log_2\left(1 - e^{-\frac{CR_k}{N}}\right), \quad (3)$$

где $\frac{CR_k}{N}$ —

оценка среднего значения и дисперсии распределения Пуассона, CR_k вычисляется согласно (1).

Разность между эмпирической оценкой количества информации (2) и оценкой, полученной в результате предположения равномерного рассеяния терминов в коллекции (3), представляет собой прирост информации R_IDF_k, содержащейся в реальном распределении термина в коллекции по сравнению с равномерно-случайным распределением Пуассона:

$$R\_IDF_k = IDF_k^E - IDF_k^P = \\ = -log_2\left(\frac{D_k}{N}\right) + log_2\left(1 - e^{-\frac{CR_k}{N}}\right). \quad (4)$$

Такая оценка повышает ценность значимых терминов, так как значимые, осмысленные термины должны быть распределены неравномерно среди относительно небольшого числа наборов интересов, а бессодержательные термины будут равномерно рассеяны по всей коллекции [11]. Далее для каждого термина в наборе вычисляется его «вес» wt_k^i по формуле:

$$wt_k^i = TF_k^i \cdot R_{IDF_k},$$

$$\forall i = 1,2, \dots, N; k = 1,2, \dots, K,. \quad (5)$$

Величина R_IDF_k рассчитывается по формуле (4), а TF_k^i согласно формуле:

$$TF_k^i = 0{,}5 + 0{,}5 \cdot \frac{cr_k^i}{tf\_max_k},$$

где tf_max_k — наибольшее число повторений k-го термина в наборах коллекции. Так как значения коэффициентов $wt_k^i \in [-1,1]$, то произведём дополнительную математическую операцию:

Таблица 1. Фрагмент входных и выходных данных

Набор	Термин	Вес	Термин	Вес	Термин	Вес
1	38RUS	0,47	Иркутск	0,47	Irkdtp	0,49
2	Zloyshkolnik	0,49	Ненормально	0,48	Nenorm	0,69
3	Новосибирск	0,33	Incidentnsk	0,32	Новосибирская	0,43

$$wtn_k^i = 0{,}5 + 0{,}5 \cdot wt_k^i, \quad (6)$$

в результате которой весовые коэффициенты $wtn_k^i \in [0,1]$.

Представленный семантический метод позволяет для всех наборов исследуемой коллекции оценить весовой коэффициент каждого термина, содержащегося в наборе, то есть выполнить цифровую трансформацию социальных интересов студентов.

Следующий шаг — графическая интерпретация i-го (i = 1, 2, …, N) набора терминов на основе соответствующих терминам числовых значений весовых коэффициентов. Термины в наборах коллекции являются значимыми, если соответствующие им значения весовых коэффициентов wtn_k^i не менее порогового значения. Предлагается элементы, чей вес wtn_k^i выше порогового предела, графически отразить в соответствии с некоторым графиком функций. В качестве функций для коллекций, относящихся к студентам различных специальностей, могут быть выбраны, например, тригонометрические или показательные функции. Для графического отражения имеющихся данных потребуется растровое изображение размерностью $n \times n$ ячеек, где $n = integer\left(\sqrt{dl\_max}\right) + 1$. Величина dl_max определяется, исходя из формулы:

$$dl\_max = \max_{i=1,2,\ldots,N} dl^i,$$

где dl^i — количество терминов в i-м наборе коллекции. Очевидно, что часть ячеек матрицы окажется невостребованной в связи с округлением в большую сторону, а также с учетом того, что $dl^i \leq dl_{max}, \forall\, i \in [1, N]$. Ячейки матрицы заполняются оттенками серого цвета следующим образом. Значениям весовых коэффициентов ставится в соответствие градация оттенков серого цвета: наименьшие значения отражаются светлыми оттенками, наибольшие — темными. Весовые коэффициенты, значения которых выше порогового, равного 0,5, распределяются согласно выбранной для коллекции графической модели, остальные значения распространяются в порядке очереди от верхнего левого угла матрицы вправо и вниз. Невостребованные ячейки матрицы заполняются ячейками белого цвета.

Экспериментальные исследования

Данными эксперимента являются интересы выпускников нескольких специальностей ВУЗа. Сведения получены из социальной сети Vkontakte и представляют массив данных о подписках, книгах, играх и прочих интересах выпускников. Данные обезличены согласно 155-ФЗ о персональных данных. В экспериментальных исследованиях использовано 879 наборов социальных интересов выпускников четырех специальностей. Для графической интерпретации наборов терминов в коллекциях использованы в качестве направляющих тригонометрические функции $y = \pm sin(x)$ и $y = \pm cos(x)$. Наиболее представительной является коллекция, состоящая из N = {35} наборов. Перед статистической обработкой выполнен необходимый препроцессинг — очистка от специальных символов, удаление коротких слов, длиной три и менее букв, лишних пробелов. Рассмотрим набор, включающий наибольшее число терминов $dl\_max = 1179$. Веса терминов, составляющих набор, получены согласно алгоритму (1)-(6). Результат работы алгоритма представлен в виде электронной таблицы и доступен онлайн [12]. В табл. 1 представлены образцы терминов из трех наборов и соответствующие им весовые коэффициенты.

Графическая интерпретация полученных числовых значений коэффициентов основана на функции $y = sin(x)$. Для графического отражения имеющихся данных требуется растровое изображение размерностью 35x35 ячеек, так как $n = integer\left(\sqrt{dl\_max}\right) + 1 = 35$.

В соответствии со значениями полученных весовых коэффициентов матрица заполняется следующим образом: значения выше порогового, равного 0,5, распределяются ячейками темного цвета согласно графической модели $y = sin(x)$, остальные значения распространяются в порядке очереди. Невостребованные ячейки матрицы имеют белый цвет (рис. 1).

На графике видно, что термины, имеющие максимальный вес, распределены согласно выбранной функции и имеют цвет близкий к чёрному. Светлые оттенки серого цвета соответствуют терминам с более низкими значениями весов. Ячейки белого цвета соответствуют пустым клеткам.

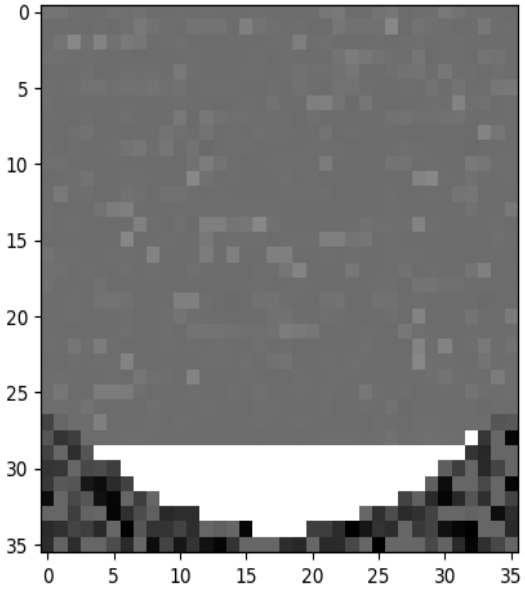

Рис. 1. Графическое представление матрицы

Заключение

В работе рассмотрен семантический алгоритм обработки неструктурированных текстовых данных, основу которого составляет комбинация эмпирического и теоретического подходов к определению количества информации К Шеннона и равномерного распределения Пуассона. Сравнение получаемых характеристик позволяет выделить наиболее ценные социальные интересы студентов-выпускников определенной специальности. Согласно цифровой трансформации текстовых данных для каждого термина в наборе интересов студента формируется весовой коэффициент его значимости. Предложено использование тригонометрических или показательных функций, как направляющих в графической интерпретации весовых коэффициентов, значения которых превышают некоторый порог.

В дальнейшем исследовании предполагается совокупность весовых коэффициентов и растровых изображений использовать в качестве входных данных нейронной сети, с целью формирования цифровых и графических образов социальных «портретов» выпускников рассматриваемых специальностей. Обученная нейронная сеть предназначается для распознавания близости набора социальных интересов некоторого, поступающего в ВУЗ абитуриента к коллекции интересов студентов-выпускников отдельных специальностей. В целом, разработанный подход можно успешно использовать в процессе обработки несвязанной текстовой информации, не только в рамках образовательной сферы, но и для решения задач других прикладных областей, имеющих подобную постановку задачи и неструктурированные текстовые данные.

ЛИТЕРАТУРА

1. Черногорова, Ю.В. Методы распознавания образов / Ю.В. Черногорова // Молодой ученый. — 2016. — № 28(132). — С. 40–43.
2. Деканова Н.П., Махнев С.А. Анализ социальных сетей — поддержка абитуриентов в профессиональной ориентации // Информационные и математические технологии в науке и управлении. 2019. No 4 (16). С. 131 141.
3. Ефимова, И.Н. Анализ мотивации абитуриентов при выборе вуза (на примере Нижегородской области) / И.Н. Ефимова // Университетское управление: практика и анализ. — 2011. — № 6(76). — С. 60–68.
4. Окунев, С.В. Рассмотрение способов формирования наборов данных для обучения нейронных сетей / С.В. Окунев // Вестник науки и образования. — 2020. — № 2–3(80). — С. 16–19.
5. Орлова, Ю.А. Аннотирование и визуальное представление текстовой информации в графическом виде / Ю.А. Орлова, В.Л. Розалиев, А.В. Заболеева-Зотова // Известия Волгоградского государственного технического университета. — 2015. — № 13(177). — С. 74–85.
6. C. Manning, P. Raghavan, H. Schütze. Introduction to Information Retrieval. Cambridge University Press. 2008. 117–120 с. 155–156 с.
7. Furu Wei. A document-sensitive graph model for multi-document summarization / Furu Wei, Wenjie Li, Qin Lu, Yanxiang He // Knowledge and Information Systems, February 2010. Volume 22, Issue 2. — Pp. 245–259

8. Sheng-Tun Li. Constructing tree-based knowledge structures from text corpus / Sheng-Tun Li, Fu-Ching Tsai // Applied Intelligence, August 2010.— Volume 33, Issue 1.— Pp. 67–78
9. Короленко, В.А. Применение анализа тональности текстов для распознавания фейковых новостей / В.А. Короленко.— Текст: непосредственный // Молодой ученый.— 2020.— № 22 (312).— С. 36–44.— URL: https://moluch.ru/archive/312/70957/ (дата обращения: 01.12.2021).
10. K. Church, W. Gale. Poisson mixtures. Natural Language Engineering, 1995, 1(2):163–190.
11. Зеленков, Ю.Г. Сравнительный анализ методов определения нечетких дубликатов для Web-документов /Ю.Г. Зеленков, И.В. Сегалович // Труды IX Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL'2007.— 9 с.
12. Результат работы алгоритмов. Date Views 15.09.2021 drive.google.com/file/d/1d2srSZyzVJ0ZANAWe4eca0QTeKagmFyJ/view?usp=sharing.

г. Иркутск