«Сейфуллин оқулары-18(2): «XXI ғасыр ғылымы – трансформация дәуірі» Халықаралық ғылыми-практикалық конференция материалдары = Материалы международной научно-практической конференции «Сейфуллинские чтения – 18(2): «Наука XXI века - эпоха трансформации» - 2022.- Т.I, Ч.III. - С.115-119.

ПРИМЕНЕНИЕ МЕТОДОВ КЛАСТЕРИЗАЦИИ И НОРМАЛИЗАЦИИ ДАННЫХ ДЛЯ ВОЗМОЖНОСТИ ПОСЛЕДУЮЩЕГО АНАЛИЗА

Турсумбаева А., , магистр

Казахский агротехнический университет им. С. Сейфуллина, г. Нур-Султан

В настоящее время существует огромное количество данных, доступных в социальных сетях и различных каналах социальных сетей. Проблема заключается в использовании этих данных и преобразовании их в полезную и действующую информацию, так как анализ данных из социальных сетей помогает обнаруживать и раскрывать модели взаимодействия между людьми.

В дополнение к обычным статистическим методам анализа данных, социальные сети исследуются с использованием определенных метрик. Это помогает понять зависимости между социальными объектами в данных, характеризуя их поведение и их влияние на сеть в целом и во времени.

Социальная сеть состоит из людей или организаций, связанных родственными отношениями, дружбой, убеждениями, общими интересами, финансовым обменом и многими другими вещами. Эти социальные отношения рассматриваются как граф, в котором сущности образуют узлы, а ребра — это связи между ними. Джордж Зиммель, на рубеже двадцатого века, был первым ученым, который непосредственно мыслил в терминах социальных сетей. Его очерки были сосредоточены на природе размера сети на взаимодействии и на вероятности взаимодействия в слабосвязанных сетях, а не в группах (Зиммель, 1908/1971). После перерыва, в начале двадцатого века, появились три основных традиции в социальных сетях. Дж. Л. Морено вел систематическую регистрацию и анализ социальных отношений в малых группах, а именно в рабочих группах и классных комнатах. Тем временем Гарвардская группа во главе с У. Ллойдом Уорнером и Элтоном Мейо исследовала межличностные связи на работе. В 1940 году Президентское обращение Рэдклиффа Брауна к британским антропологам подчеркнуло систематическое изучение сетей. В 1960-х и 1970-х годах все большее число исследователей работало над объединением различных традиций и направлений. Значительную независимую работу также проделали ученые из различных университетов, таких как Университет Калифорнии, Университет Чикаго, Университет Торонто и Университет штата Мичиган. Они подвергли критике методологический индивидуализм и групповой анализ, заявив, что

просмотр мира как социальных сетей дает больше аналитических преимуществ [1].

Социальная сеть состоит из конечного набора вершин и определенных для них отношений или связей. Установленные отношения могут носить личный или профессиональный характер и могут варьироваться от случайного знакомства до близких знакомых связей. Помимо социальных отношений, ссылки могут также представлять поток информации / товаров / денег, взаимодействия, сходства, среди прочего. Структура таких сетей обычно представлена графами. Поэтому сети часто рассматриваются как эквивалентные графам [2].

Когда сеть состоит из двух наборов узлов, она называется двухрежимной сетью. Некоторые примеры двухрежимных сетей включают сети пользовательских продуктов (Amazon, eBay и т. д.), Сети членства или афиляции (actor-movie (IMDB), группа пользователей (youtube), канал пользователя (youtube), проект пользователя (GitHub).), организация пользователей и т. д.), сети пользовательских предпочтений (Pinterest, Instagram, twitter), сети цитирования, инвестиции в акции пользователей. Эти двухрежимные сети могут быть преобразованы в одномодовые сети между единым набором узлов, как в приведенных выше примерах, и затем проанализированы. Общие задачи анализа социальных сетей включают в себя выявление наиболее влиятельных, престижных или центральных субъектов с использованием статистических измерений; идентификация хабов и авторитетов с использованием алгоритмов анализа ссылок; обнаружение сообществ, использование методов обнаружения сообщества, и как информация распространяется в сети, используя алгоритмы распространения. Задания очень полезны в процессе извлечения знаний из сетей и, следовательно, в процессе решения проблем. Из-за привлекательного характера таких задач и высокого потенциала, который открывает этот вид анализа, анализ социальных сетей стала популярным подходом в бесчисленных областях, от биологии до бизнеса. Например, некоторые компании используют анализ социальных сетей для того, чтобы максимизировать положительный слух о своих продуктах, ориентируясь на клиентов с более высоким значением сети (с более высоким влиянием и поддержкой) [3].

Проблема больших данных постепенно обсуждалась во всех областях с 2001 года, когда Gartner Co. выпустила свое описание больших данных «3 В» (объем, скорость и разнообразие). Это стало горячей темой за последние 4 года. Тем не менее, пока нет единого определения больших данных; его интерпретация варьируется от академических и деловых кругов. Национальный научный фонд США описывает большие данные как «большие, разнообразные, сложные, продольные и / или распределенные наборы данных, сгенерированные из инструментов, датчиков, транзакций в Интернете, электронной почты, видео, потоков кликов и / или всех других цифровых источников, доступных сегодня и в будущем » [2, стр. 233], в то время как Википедия говорит: «Большие данные - это всеобъемлющий

термин для любого набора наборов данных, настолько больших или сложных, что становится трудно обрабатывать их с использованием традиционных приложений обработки данных. Стоит отметить, что группа международных ученых провела мозговой штурм по двум определениям «больших данных» на сессии, посвященной науке о данных и большим данным на научной конференции в Сяншань в Пекине. В настоящее время большинство исследователей и специалистов обычно соглашаются использовать «4 В» - объем, скорость, разнообразие и достоверность - для описания основных характеристик больших данных [4].

Приведем некоторые идеи, которые можно обсудить в рамках данной работы. Во-первых, алгоритм больших данных должен быть алгоритмом, который может обрабатывать и анализировать большие данные при наличии доступных вычислительных ресурсов и выполнять их в разумные сроки. Большие данные, с которыми они могут работать, имеют, по крайней мере, одну из следующих характеристик: большие, неоднородные, распределенные, многоисточники, поток данных, большой размер и высокая неопределенность. Алгоритм может быть выполнен при соответствующей степени времени, сложности хранения и связи. Он также обладает некоторыми уникальными свойствами, такими как высокая отказоустойчивость, интеграция решений и возможность сборки. Во-вторых, ключевые идеи разработки алгоритма больших данных могут включать поддержание правильного соотношения выборки данных и совокупности; простое моделирование и простая процедура; низкая точность, сложность и теоретическая основа. Наконец, в дополнение к известным методам статистики или интеллектуального анализа данных, для построения высокоэффективного алгоритма больших данных могут использоваться другие вычислительные методы, такие как обработка на основе множеств, стохастические вычисления, онлайн-вычисления, распределенные и параллельные вычисления, облачные вычисления [5].

Кластеризация — это фундаментальный метод анализа данных. Например, при решении задач, связанных с переездом в другой город, можно попробовать объединить контакты в LinkedIn по географическим регионам, чтобы лучше оценить имеющиеся экономические возможности. При реализации решений задач кластеризации данных из LinkedIn или из других источников с двумя основными темами [6].

Даже при использовании очень хорошего API данные редко бывают предоставлены в нужном вам формате, — часто требуется нечто большее, чем простое преобразование, чтобы привести данные в форму, пригодную для анализа. Имея набор хорошо нормализованных элементов, вы можете пожелать оценить сходство любых двух из них, будь то названия должностей или компаний, описание профессиональных интересов, географические названия или любые другие поля, значения которых могут быть представлены произвольным текстом. Для этого вам нужно определить эвристику, оценивающую сходство двух любых значений. В некоторых

ситуациях определение сходства вполне очевидно, но в других может быть сопряжено с некоторыми сложностями.

Нормализация данных и определение сходства — это две главные темы, с которыми вы будете сталкиваться в кластеризации на абстрактном уровне. Но есть еще третья тема — сокращение размерности, которая становится актуальной, как только масштаб данных перестает быть тривиальным. Для группировки элементов в множестве с использованием метрики сходства в идеале желательно сравнить каждый элемент с каждым другим элементом. В этом случае, при наихудшем развитии событий, для множества из n элементов вам придется вычислить степень сходства примерно n2 раз, чтобы сравнить каждый из n элементов с n–1 другими элементами. В информатике эту ситуацию называют проблемой квадратичной сложности и обычно обозначают как O(n2); в разговорах ее обычно называют «проблемой квадратичного роста большого O». Проблемы O(n2) становятся неразрешимыми для очень больших значений n, и в большинстве случаев термин неразрешимые означает, что вам придется ждать «слишком долго», пока решение будет вычислено. «Слишком долго» — это могут быть минуты, годы или эпохи, в зависимости от характера задачи и ее ограничений [7].

Методы кластеризации являются основной частью арсенала инструментов любого специалиста по анализу данных, потому, что почти в любой отрасли — от военной разведки до банковского дела и ландшафтного дизайна — может потребоваться проанализировать по-настоящему огромный объем нестандартных реляционных данных, и рост числа вакансий специалистов по данным за предыдущие годы служит тому явным свидетельством.

Давайте попробуем стандартизировать названия компаний из вашей профессиональной сети. Как рассказывалось выше, извлечь данные из LinkedIn можно двумя основными способами: программным, с помощью LinkedIn API, или с использованием механизма экспортирования профессиональной сети в виде адресной книги, которая включает такие основные сведения, как имя, должность, компания и контактная информация. Представим, что у нас уже есть CSV-файл с контактами, экспортированный из LinkedIn, и теперь мы можем нормализовать и вывести выбранные сущности, как показано в примере. Как описывается в комментариях внутри примеров, вам нужно переименовать CSV-файл с контактами, который вы экспортировали из LinkedIn, следуя инструкциям в разделе «Загрузка файла с информацией о контактах в LinkedIn», и скопировать в определенный каталог, где его сможет найти программный код [8].

Простая нормализация сокращений в названиях компаний.

```
import os
import csv
from collections import Counter
from operator import itemgetter
from prettytable import PrettyTable
```

```
CSV_FILE = os.path.join("resources", "ch03-linkedin", 'Connections.csv')
transforms = [(', Inc.', ''), (', Inc', ''), (', LLC', ''), (', LLP', ''),
(' LLC', ''), (' Inc.', ''), (' Inc', '')]
companies = [c['Company'].strip() for c in contacts if c['Company'].strip() !
= '']
for i, _ in enumerate(companies):
for transform in transforms:
companies[i] = companies[i].replace(*transform)
pt = PrettyTable(field_names=['Company', 'Freq'])
pt.align = 'l'
c = Counter(companies)
[pt.add_row([company, freq]) for (company, freq) in
sorted(c.items(), key=itemgetter(1), reverse=True) if freq > 1]
print(pt)
```

Ниже приводятся результаты простого частотного анализа (рис.2) :

```
+--------------------------------+------+
| Company                        | Freq |
+--------------------------------+------+
| Digital Reasoning Systems      | 31   |
| O'Reilly Media                 | 19   |
| Google                         | 18   |
| Novetta Solutions              | 9    |
| Mozilla Corporation            | 9    |
| Booz Allen Hamilton            | 8    |
| ...                            | ...  |
+--------------------------------+------+
```

Рисунок 2 - Результаты простого частотного анализа

В языке Python поддерживается возможность передачи аргументов функциям путем разыменования списка и/или словаря, что иногда очень удобно, как показано в примере 4.4. Например, вызов f(*args, **kw) эквивалентен вызову f(1, 7, x=23), где args аргументов определяется как список [1,7] и kw — как словарь {'x': 23}. Имейте в виду, что для обработки более сложных ситуаций, например, для нормализации разных названий одной и той же компании, менявшихся с течением времени, таких как O’Reilly Media, вам потребуется написать более замысловатый код. В данном случае название этой компании может быть представлено как O’Reilly & Associates, O’Reilly Media, O’Reilly, Inc. или просто O’Reilly [9].

На данный момент разработка программ и приложений для обеспечения собственной безопасности является актуальным аспектом для развития государства. Страна, которая действительно серьезно относится к своей собственной безопасности, должна создавать свои собственные программные и аппаратные средства для обработки персональных данных. Это означает, что каждая страна должна создавать свои собственные приспособления для поиска, обработки и анализа данных, которые находятся в социальных сетях и других интернет ресурсах.

Список использованной литературы

1 Charu C. Aggarwal. Social Network Data Analytics. 2011. -520 p.

2 Milgram S. The Small World Problem // Psychology Today. -1967. -Vol. 2.- P. 60–67.

3 Granovetter M. S. The Strength of Weak Ties [Text] / American Journal of Sociology. -1973.-Vol. 78. -No. 6. -P. 1360–1380.

4 Kleinberg J. M. Authoritative Sources in a Hyperlinked Environment [Text] / J. ACM. -1999. -Vol. 46. -No. 5. -P. 604–632.

5 Gyöngyi Z., Garcia-Molina H., Pedersen J. Combating Web Spam with TrustRank [Text] / Proceedings of the International Conference on Very Large Data Bases. -2004. -Vol. -30. -P. 576.

6 Арустамов А.И., Васильев Е.П., Орешков В.И. Интеллектуальные платформы - современный инструмент анализа данных в экономике и бизнесе [Текст] / Сб. трудов Международной научно-практической конференции «Дни науки», Прага, 2012.

7 Башмаков А.И., Башмаков И.А. Интеллектуальные информационные технологии [Текст] : Учеб. пособие. - М.: Изд-во МГТУ им. Н.Э. Баумана, 2005. - 304 с : илл. — (Информатика в техническом университете).

8 Орешков В.И. Интеллектуальный анализ данных как современный инструмент поддержки управленческих решений [Текст] / Вестник Рязанского гос.

агротехнологического университета имени П.А. Костычева. Рязань: РГАТУ. - 2011.-№4.-С. 55-59.

9 Advances in data mining : applications in E-commerce, medicine, and knowledge management [Text] / Perner. P (ed.). - Berlin: Springer, 2002. - P. 108.