Организация Объединенных Наций ECE/CES/2019/23

Экономический и Социальный Совет

Distr.: General
9 April 2019
Russian
Original: English

Европейская экономическая комиссия

Конференция европейских статистиков

Шестьдесят седьмая пленарная сессия
Париж, 26–28 июня 2019 года
Пункт 2 b) предварительной повестки дня
Новые источники данных – доступность и использование
Заседание 2: Навыки, необходимые для использования новых источников данных

Новые компетенции, необходимые для работы с большими данными – уроки, извлеченные к настоящему времени

Записка Статистического управления Словении

Резюме

В настоящем документе описываются уроки, извлеченные за последние шесть лет Статистическим управлением Словении в отношении компетенций, необходимых для работы с большими данными. Он начинается с перечня полученных компетенций, наиболее важными из которых являются компетенции, которые имеют отношение к характеристикам новых источников данных, передовым методам их обработки и коммуникации с держателями данных. Далее рассматриваются способы приобретения таких компетенций, обучение в процессе практической деятельности, что рассматривается в качестве наиболее важного метода, за которыми следует обмен опытом в рамках международного статистического сообщества. В заключение в докладе рассматриваются будущие вопросы и говорится о сохранении компетенций как о наиболее сложной задаче.

Данный документ представлен для обсуждения на семинаре Конференции европейских статистиков 2019 года по теме «Новые источники данных – доступность и использование», заседание 2 «Навыки, необходимые для использования новых источников данных».

GE.19-05941 (R) 110419 120419

1905941

I. Введение

1. Большие данные и их использование в последние годы стали одной из самых актуальных тем в официальной статистике. Статистическое управление Словении (СУС) занимается сбором, обработкой и анализом источников больших данных на протяжении более шести лет. За этот период был завершен один проект, в результате которого в регулярное статистическое производство был включен один из источников больших данных – источник данных сканирования цен. Другие проекты включают анализ возможностей использования данных о местоположении мобильных телефонов и совершенных с их помощью операциях, онлайновых объявлений о вакансиях, онлайновых цен, спутниковых изображений, данных о финансовых операциях и данных датчиков движения. При их осуществлении возникают одни и те же проблемы, такие как проблемы доступа, этических соображений, методологии, технологии и качества. Однако существует проблема, которая, как представляется, имеет наиболее всеобъемлющий и актуальный характер для всех проектов и этапов статистической деятельности, а именно: необходимые навыки и знания.

2. В этом документе подробно рассматриваются уроки, извлеченные из опыта СУС в отношении навыков и знаний, полученных при выполнении проектов в области больших данных. Он начинается с перечисления того, какие новые знания или навыки были получены, затем следует описание того, как они были получены, и завершается изложением перспектив на будущее. В настоящем документе для обозначения знаний и навыков используется термин «компетенция».

II. Какие уроки были извлечены

3. В ходе осуществления проектов с использованием больших данных в рамках проекта СУС были разработаны новые компетенции, при этом существующие компетенции были усовершенствованы или усилены. Некоторые компетенции были новыми для одних сотрудников и для некоторых областей статистики, но не для всех. Так, некоторые отраслевые статистики начали использовать более современные ИТ-инструменты в ходе осуществления этих проектов, в то время как некоторые эксперты по ИТ получили более полное представление о статистических методологиях. Одним из важных результатов этой работы стало распространение уже существующих компетенций в Управлении. Кроме того, многие компетенции не являются новыми для «внешнего мира», но являются таковыми для СУС; они существуют уже несколько десятилетий, но лишь недавно стали играть более заметную роль в официальной статистике Словении (например, моделирование данных). К наиболее важным новым или усовершенствованным компетенциям, приобретенным СУС при работе с большими данными, относятся следующие:

a) увеличение числа известных источников данных. Были обнаружены новые источники данных и изучены их характеристики (владельцы/хранители, частота, качество, содержание, доступность). Примеры включают информацию, полученную от розничных торговцев (цены, количество, коды проданной продукции), информацию, полученную в результате использования мобильных телефонов (местоположение, сделки), систем наблюдения Земли, информацию, полученную от счетчиков движения и т. д.;

b) увеличение объема знаний о законодательстве, регулирующем существование различных видов данных, главным образом о законодательстве, касающемся защиты личных данных, и о законодательстве, касающемся электронных сообщений;

c) увеличение числа используемых методов сбора данных. Например, данные сканирования цен собираются путем приема файлов данных непосредственно от предприятий. В прошлом такой подход использовался только для сбора данных от органов государственного управления, но не от предприятий. Кроме того, в настоящее время для получения данных, касающихся цен в интернет-магазинах и объявлений о вакансиях на веб-порталах занятости, используются методы извлечения данных из

веб-страниц. Для этого необходимо было приобрести знания о составе интернет-сайтов, функционировании браузеров, программировании специальных и общих «веб-пауков» и «веб-искателей», оптимизации поиска тегов с учетом HTML-структуры;

d) разработка новых подходов к взаимодействию с поставщиками данных. Традиционно от предприятий требовалось предоставлять данные в соответствии с установленной формой и содержанием данных. Теперь, используя данные сканирования цен, каждое предприятие предоставляет данные таким образом и в такой форме, которая максимально адаптирована к его возможностям и способу работы. Таким образом, требования СУС стали еще более удобными для исполнения поставщиками данных;

e) получение дополнительной информации о предприятиях: какова их бизнес-модель, что их беспокоит (раскрытие коммерческой тайны, раскрытие данных о клиентах и т. д.), как они видят государственное управление, свою деловую и законодательную среду, как они функционируют, что их интересует;

f) признание смены фундаментальной парадигмы подготовки статистических данных, которая постепенно принимается. Традиционно работа статистика начинается с конечного продукта. После определения потребностей пользователей и конечного продукта начинается процесс рассмотрения источников данных и оптимальных способов сбора и обработки данных (от конечного продукта к вводимым данным). В случае больших данных, как правило, применяется противоположный подход. Сначала изыскиваются сами источники данных, и только после этого статистик приступает к анализу того, имеют ли эти источники какую-либо ценность с точки зрения статистики (от вводимых данных к конечному продукту). Благодаря такому новому подходу к статистическому производству СУС получает дополнительные компетенции для отбора наиболее «показательных» данных, интеграции данных и принятия решения о направлении исследований;

g) распространение мышления в более широком контексте. Располагая источниками больших данных, СУС учится непредвзятости при рассмотрении возможных вариантов использования. Так, вопрос о возможном использовании спутниковых снимков в производстве статистических данных стал предметом широкого рассмотрения в рамках СУС с привлечением большого числа сотрудников, которым было предложено представить свои предложения;

h) освоение новых методологий. В условиях наличия данных сканирования цен была отменена методология составления индекса потребительских цен на продовольствие и напитки на основе репрезентативных товаров, и вместо нее стали использоваться данные по всей совокупности товаров. Благодаря этому сотрудники также ознакомились с методами составления новых видов индексов;

i) повышение уровня знаний о методах интеграции данных. Интеграция данных из различных источников не является чем-то новым для СУС, поскольку административные записи уже используются на протяжении нескольких десятилетий. Новизна заключается в интеграции вводимых данных в виде непосредственно собранных и больших данных;

j) освоение подходов в плане оптимизации программ обработки больших объемов данных и интенсивных операций обрабатывающих подразделений. Так, были приложены значительные усилия для освоения и подготовки программ в многопоточном порядке, что позволяет ускорить сбор, анализ, обработку и агрегирование данных. Многопоточные методы используются для анализа, корректировки, изменения и агрегирования данных счетчиков движения; в связи с объемом таких данных (десятки гигабайт) стандартные циклы занимали бы гораздо больше времени. Многопоточные программы используются также для извлечения данных из веб-страниц и хранения веб-сайтов (которые исчисляются сотнями тысяч);

k) приобретение знаний в области анализа текстовых данных и машинного обучения, используемые для структурирования неструктурированных данных и классификации данных с использованием различных алгоритмов. Проверяются

методы классификации потребительских товаров в соответствии с классификацией КИПЦ, сопоставления бизнес-единиц и структурирования объявлений о вакансиях, размещаемых в сети Интернет;

l) превращение моделирования данных наряду с машинным обучением прогнозированию и составлению статистических данных в обычный вид деятельности. Это включает в себя знания для понимания и разработки моделей, а также понимание ограничений и последствий комбинирования моделей. Так, введено прогнозирование сводных экономических показателей на основе линейной регрессии с использованием данных счетчиков движения и оборачиваемости. Разработана методика сбора информации по вакансиям с опорой на неструктурированные данные с использованием логистической регрессии, линейной регрессии, метода случайной подстановки ближайшего соседа и метода ансамбля AdaBoost;

m) внедрение новых и более широкое использование существующих средств программирования и библиотек, например технологий (.NET, ASP.NET, ML.NET frameworks), инструментов (MS Visual Studio, SQL Server Data Tools – SQL Server Integration Services), языков (C#, R, Python), баз данных (MS SQL Server, Oracle), библиотек Python (TensorFlow, Scikit-Learn, Keras) и др.;

n) в связи с высокой степенью нестабильности инструментов и бизнес-моделей их поставщиков (например, некоторые инструменты в течение некоторого времени были в открытом доступе, а затем внезапно стали полностью коммерческими; неожиданное исчезновение инструментов, которые еще вчера имелись в наличии) СУС стало более гибким и адаптивным к таким изменениям, чтобы они не вызывали серьезных сбоев в работе;

o) расширение знаний о координации и увязке всех вышеупомянутых компетенций и признание того, что наиболее полезным для сотрудников является наличие нескольких различных таких компетенций (например, в областях кодирования, математики, статистической методологии).

III. Как были извлечены эти уроки

4. С организационной точки зрения СУС подходило к большим данным так же, как и к другим разработкам. Участие в разработках рассматривается как неотъемлемая часть деятельности каждого статистика. Не существует специального подразделения, которое занималось бы только инновациями, прогрессом, улучшениями и т. д. СУС считает, что разделение разработок и текущей работы приводит к сегрегации между ними, к созданию такой ситуации, когда информация передается руководству, минуя промежуточные инстанции, к отчуждению «статистиков, которым поручаются разработки», от занятия практическими вопросами, «обычных статистиков» – от перспективных видов деятельности, а также к дополнительным проблемам при осуществлении изменений. То же самое справедливо и для работы с большими данными. В настоящее время она организуется в рамках официально утвержденных проектов с участием сотрудников различных структурных подразделений.

5. В частности, компетенции, необходимые для работы с большими массивами данных, приобретались (и продолжают приобретаться) следующими различными способами:

a) обучение на собственном опыте. Осуществление проектов и практическая работа – это, безусловно, самый важный способ приобретения компетенций. В состав проектных групп входят сотрудники различных подразделений (специалисты по методологии статистики, эксперты в области ИТ, математики), и их тесное сотрудничество имеет крайне важное значение для эффективного обмена идеями, знаниями и т. д. Кроме того, некоторые статистики участвуют в нескольких проектах, что позволяет использовать одни и те же решения для разных проектов;

b) участие в международных мероприятиях: участие в работе таких образований, как Целевая группа по большим данным ЕЭК ООН, Целевая группа по большим данным ЕСС (Европейская статистическая система), Целевая группа по большим данным ESSNet BigData1 (извлечение данных по вакансиям из веб-страниц, ранние оценки, методология), ESSnet BigData2 (вакансии, финансовые операции), Группа экспертов по данным сканирования ЕСС;

c) ознакомительные поездки в статистические управления, имеющие обширный опыт использования источников больших данных, с целью изучения их документов и практики. Такие поездки чаще всего были посвящены вопросам использования данных сканирования;

d) участие в международных конференциях, семинарах, курсах и т. д., как непосредственно, так и в виртуальной форме (вебинары, веб-курсы, семинары в Интернете). К числу относятся: «Большие данные: эффективная обработка и анализ сверхбольших и неструктурированных данных для официальной статистики»; «Использование исследований в официальной статистике»; «Может ли статистик стать ученым, занимающимся обработкой данных?»; «Общие сведения о больших данных и их инструментах»; «Источники больших данных – веб-источники, социальные сети и текстовый анализ»; «Современные источники больших данных: мобильные телефоны и другие датчики»; «Автоматизированный сбор онлайновых цен»; «Макроэкономическое прогнозирование»; «Наука о данных: линейная регрессия»; «Анализ символьных данных»; «Практикум для анализа данных официальной статистики»; «Машинное обучение и эконометрика»; глобальные конференции Организации Объединенных Наций по большим данным и т. д.;

e) самообучение с использованием ресурсов, имеющихся в Интернете (например, руководств, справочников, видеоматериалов, платформ для разработчиков программного обеспечения, форумов);

f) посещение лекций в университете. Сотрудники СУС прослушали несколько лекций в рамках очного обучения на факультете информатики по тематике обработки больших данных;

g) сотрудничество с институтами. СУС сотрудничает в основном с Институтом им. Йожефа Стефана, прежде всего в области современных алгоритмов классификации, основанной на языке текста;

h) кроме того, знания приобретаются в ходе обучения, т. е. до устройства на работу в СУС. В настоящее время наиболее востребованы специалисты в областях компьютерных наук, информатики, математики, физики, социальных наук и статистики. При наличии базовой подготовки и при том условии, что соответствующее лицо разбирается в программировании, объектном программировании и структурах данных, знания по машинному обучению, получению данных и анализу текстовых данных могут быть получены относительно быстро, по крайней мере в базовой форме.

6. Один из наиболее важных уроков, усвоенных в отношении обучения, заключается в том, что нет смысла накапливать знания на случай, если они кому-нибудь понадобятся. Ранее многие сотрудники посещали курсы по работе с большими данными, но затем либо не пользовались ими, либо уходили из Управления до того, как их можно было бы подключить к проектам в области больших данных. В настоящее время существует возможность получать знания на основе любого из вышеупомянутых методов в тех случаях, когда соответствующему лицу поручается работа, требующая таких знаний.

IV. Будущие шаги

7. Что касается будущих компетенций, то следует упомянуть следующие три вопроса.

8. Во-первых, недостающие компетенции. Наиболее насущной необходимостью является накопление компетенций в таких областях, как более сложный анализ

текстовых данных, понимание сути машинного обучения, моделирование данных и интеграция больших данных. Кроме того, возникла необходимость в приобретении компетенций в плане обработки данных с использованием распределенных вычислений на основе кластеризации, нейросетей, «дерева решений» и алгоритма «случайного леса». Такие компетенции будут приобретены всеми перечисленными выше способами.

9. Во-вторых, распределение компетенций. До сих пор лишь ограниченное число сотрудников были включены в работу с большими данными и проходили для этого требуемую подготовку. В будущем необходимо будет обеспечить распределение таких компетенций и среди других сотрудников. Этот процесс будет упрощаться с каждой успешной интеграцией нового источника данных в обычное статистическое производство, т. е. с каждым успешным проектом будет легче выполнять следующие шаги. Таким образом, для руководителей одной из самых больших проблем станет мотивация сотрудников к принятию изменений, которые цифровизация вносит в официальную статистику.

10. В связи с этим вопросом следует также отметить очень быстрый прогресс в этих науках. Отдельному лицу уже трудно быть осведомленным во всех отраслях машинного обучения и анализа текстовых данных. В частном секторе уже начался процесс специализации по этим направлениям. Тем не менее от эксперта по-прежнему требуется иметь хотя бы общее представление о всех элементах и следить за новыми разработками. Таким образом, следует ожидать, что и в официальной статистике уровень специализации ученых, занимающихся обработкой данных, будет расти. Это приведет к тому, что потребуется еще больше экспертов для охвата всех элементов работы с большими данными.

11. В-третьих, удержание компетенций. В связи с большим спросом на ученых, занимающихся обработкой данных, и обусловленным этим высоким вознаграждением в частном секторе довольно трудно удержать сотрудников с соответствующими компетенциями. По крайней мере, в среднесрочной перспективе удержание персонала будет одной из основных, если не главной, кадровой проблемой для СУС.

V. Заключение

12. Готовность решать новые вопросы и исследовать новые идеи, отсутствие страха перед неудачами, любознательность, позитивное отношение к постоянным изменениям и энтузиазм в работе с данными – вот те качества, которыми в оптимальном случае должны обладать специалисты, работающие с большими данными. Кроме того, для достижения результатов необходимо развивать соответствующие компетенции. Ниже излагаются два урока наиболее общего порядка в отношении компетенций, которые были извлечены СУС при выполнении проектов в области больших данных. Во-первых, работа с большими данными не так сильно отличается от любой другой осуществляемой деятельности; есть некоторые компетенции, которые несколько более специфичны для источников больших данных, однако большинство из них являются такими же. Во-вторых, приобретение компетенций не представляет собой серьезную проблему. Для этого требуются умственные усилия, временные и финансовые ресурсы, но при этом следует отметить наличие широких возможностей для приобретения знаний и навыков. Действительно серьезным вопросом как всегда является то, как изыскать подходящих новых и уже работающих сотрудников, как побудить их к приобретению новых компетенций и взятию на себя новых обязанностей, как удержать их для того, чтобы сохранить эти компетенции в Управлении.