ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ: TREND ИЛИ APPLICATION?

Татур М. М., Демидчук А. И., Перцев Д. Ю., Искра Н. А., Самаль Д. И.
Кафедра электронных вычислительных машин, Белорусский государственный университет информатики и радиоэлектороники
Минск, Республика Беларусь
E-mail: {tatur, niskra, samal}@bsuir.com, {demidchuk.aleksey, dmitrypertsev}@gmail.com

В данной работе рассматриваются некоторые проблемы, связанные с анализом данных, и анонсируется авторская разработка интеллектуальной системы, позволяющей упростить задачи, стоящие перед исследователем.

Введение

В настоящее время повсеместно создаются процессинговые и Data-центры, в которые стекаются огромные потоки данных. В этих центрах осуществляется хранение и предоставление данных, поддержка в актуальном состоянии, статистическая обработка с использованием индустриальных программно-аппаратных технологий (SPARK, Hadoop, Cassandra). Примерами использования подобных систем являются:

1. Расплачиваясь в магазине, Вы оставляете данные о времени своего посещения, приобретенных товарах с соответствующими суммами. В случае, если Вы используете карту покупателя, эта транзакция персонифицируется (т.е. она может содержать дополнительные сведения, такие как возраст, пол, место проживания, день рождения и т.п.). Вся эта информация потенциально может быть использована для маркетинговых мероприятий по удержанию клиентов, promotion-акций и т.п.
2. Каждый звонок в экстренные службы 101, 102 и 103 фиксируется и инициирует сбор данных о времени, месте, причине происшествия и соответствующем реагировании, а также результатах реагирования или последствиях. Очевидно, что такие данные постоянно накапливаются и могут быть использованы для обобщения и формирования полезных на практике выводов.

В приведенных примерах осуществляется анализ уже структурированных данных, как правило, на уровне элементарных статистических подсчетов и фильтров. Например, несложно определить сколько посетителей побывало в магазине за отчетный период, какие товары пользовались наибольшим спросом, в каких районах складывалась наиболее криминогенная обстановка и т.п.

Обобщенная схема процесса сбора и анализа данных в процессинговом центре может быть представлена, как показано на рис. 1. В качестве главной, отражена функция формирования запроса Пользователя и получение ответа, при этом, технологические функции, такие как копирование, размещение данных на носителях и т.п. на рисунке не отражены.

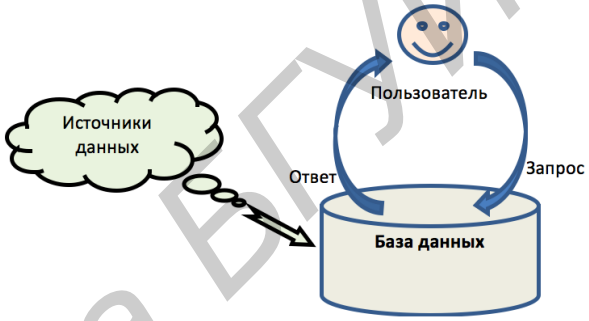

Рис. 1 – Обобщенная схема процесса сбора и анализа данных

В настоящей работе мы акцентируем внимание на некоторых проблемах, связанных с применением интеллектуальных систем анализа данных и анонсируем авторскую разработку подобной системы.

I. О проблемах, связанных с задачами анализа данных, и причинах, их вызывающих

В настоящее время Artificial Neural Networks, Machine Learning, Data Mining, Knowledge Discovery, Deep Learning, Big Data – стали наиболее часто употребляемыми терминами в контексте интеллектуального анализа данных, и находятся на гребне популярности в узком кругу IT-специалистов [1]. Однако, на практике довольно редко осуществляется глубокая (интеллектуальная) переработка данных с целью извлечения полезной информации в заданной предметной области, обнаружение скрытых (неочевидных) зависимостей, прогнозирование ситуаций и т.п. Например, с каким фактором (или совокупностью факторов) следует связать отток покупателей, насколько эффективны будут затраты на проведение профилактической работы по предупреждению чрезвычайных ситуаций и т.п.

Не претендуя на абсолютную истину, попытаемся назвать некоторые причины сложившей-

ся ситуации. Из объективных причин следует отметить:

1. Недостаточный уровень освоения теории интеллектуального анализа данных.
2. Обработка больших объемов данных требует специальных технологий распараллеливания [2, 3].
3. Отсутствие отечественного опыта создания и применения прикладных систем интеллектуального анализа данных [4].

Из субъективных причин отметим следующие:

1. Поставить (сформулировать) актуальную задачу анализа данных в конкретной предметной области должен руководитель (менеджер), заинтересованный в получении объективных результатов. Для решения задач подобного рода уже недостаточно формального применения СУБД. В таких случаях надо применять системы интеллектуального анализа данных, причем важная роль отводится специалистам по Data Science, которые должны корректно формализовать изначальную постановку задачи руководителем, корректно применить математический (алгоритмический) аппарат системы интеллектуального анализа данных и корректно интерпретировать полученный результат. С одной стороны, понятие «корректности» – расплывчато, что впрочем для интеллектуальной парадигмы является нормальным, с другой – взаимодействие руководителей и специалистов зачастую проблематично.
2. Начинающий разработчик интуитивно ищет доступ к реальным базам данных, как будто только с ними можно получить научный и/или инженерный результат, заслуживающий доверия и признания. По понятным причинам доступ к реальным базам данных ограничен.
3. Алгоритмическая сложность решения прикладной задачи интеллектуального анализа данных, как правило, выходит за рамки отдельно взятого алгоритма Data Mining (Machine Learning). В общем случае при решении задач из конкретной предметной области используется неформальная цепочка алгоритмов как из перечня Data Mining, так и обычных детерминированных.

II. Математические основы интеллектуального анализа данных

Анализируя математические методы и алгоритмы, лежащие в основе технологий интеллектуального анализа данных, можно заметить, что названные направления пересекаются. При этом, типовыми (или базовыми) являются задачи кластеризации, классификации, ранжирования, прогнозирования, ассоциативного поиска, регрессии и некоторые другие. Эти задачи могут быть решены ограниченным перечнем формальных алгоритмов (k-средних, SVM, сравнения с эталоном, k-ближайших соседей, деревьев решений и др.), рис. 2. Эти алгоритмы достаточно исследованы и уже реализованы в виде библиотек ряда систем и языковых сред (R, Weka, Wolfram Mathematics, Caffe, Tensorflow, cuDNN и др.).

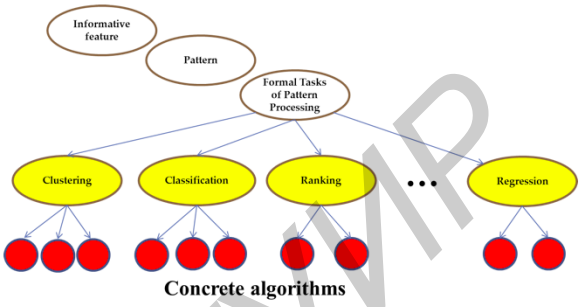

Рис. 2 – Ключевые термины и типовые задачи интеллектуального анализа данных

Тогда обобщенную архитектуру прикладной системы интеллектуального анализа данных можно представить в виде алгоритмического ядра определенных библиотек Data Mining (Machine Learning), поверх которого реализована сервисная оболочка интеллектуального анализа данных, включающая необходимые настройки, WEB-интерфейс, различные режимы визуализации процессов обработки, рис. 3

Рис. 3 – Мнемосхема архитектуры прикладной системы интеллектуального анализа данных

В отличие от стандартных библиотечных алгоритмов, функциональные возможности оболочки сервисов могут значительно отличаться от системы к системе. Тем не менее, можно выделить наиболее значимые функции:

– управление решением формальных задач кластеризации, классификации, ассоциативного поиска, ранжирования с использованием различных алгоритмов из Библиотеки DataMining;

– подготовка данных (нормализация, взвешивание, установление параметров решения задач, назначение критериев и т.п.);
– анализ решения задач на различных этапах, представление и визуализация результатов решения;
– оценка необходимых ресурсов (времени вычисления, производительности) для решения задач анализа данных определенной размерности;
– предоставление элементов визуального программирования.

Библиотеки алгоритмов, объединенных единой оболочкой, могут быть использованы в качестве инструментального средства для построения Системы интеллектуального анализа данных предметной области.

III. Техническая реализация оригинальной системы интеллектуального анализа данных

На кафедре ЭВМ Белорусского государственного университета информатики и радиоэлектроники разрабатываются элементы технологии для построения таких систем. В частности, разрабатывается вычислительная система, которая может служить:

1. целям проведения научных исследований и обучению основам интеллектуального анализа денных (DataMining, MachineLearning);
2. в качестве демонстрационной версии инструментального средства для разработки и построения интеллектуальной системы анализа данных в конкретной предметной области.

Система реализуется по модульному принципу. В качестве одного из модулей может использоваться вычислительный кластер с клиент-серверным принципом организации вычислений. Серверная часть предоставляет доступ к оболочке с библиотеками алгоритмов, одним из вариантов которой может быть WEB-интерфейс. Клиентская часть разворачивается на персональном компьютере пользователя.

Структурная схема оболочки с библиотеками алгоритмов представлена на рис. 4.

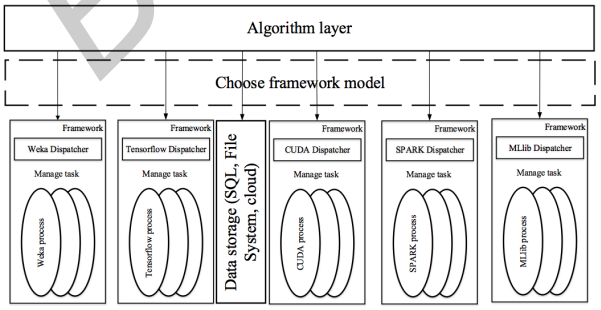

Рис. 4 – Структурная схема оболочки с библиотеками алгоритмов

Основными задачами данной оболочки являются:

– выбор подходящей версии алгоритма, с учетом особенностей входных параметров, целевого устройства исполнения;
– инициализация необходимой библиотеки алгоритмов (framework'a);
– создание интерфейса доступа для последующего применения алгоритма.

Структурно в представленной оболочке можно выделить следующие сущности:

– Algorithm layer – обеспечивает единый интерфейс доступа для всех поддерживаемых библиотек, скрывает от конечного пользователя особенности настройки и инициализации необходимых параметров, обобщает и систематизирует информацию по поддерживаемым алгоритмам от библиотек, в качестве клиента этого слоя может выступать web-интерфейс или совместимое ПО, развернутое на клиентской машине;
– Framework – библиотека алгоритмов от того или иного разработчика. В качестве базового набора библиотек алгоритмов возможно подключение Weka, CUDA, SPARK, MLlib. Кроме того, каждый компонент информирует клиента о поддерживаемых данной библиотекой алгоритмах и предоставляет интерфейс доступа на уровень «Algorithm layer»;
– хранилище данных предоставляет доступ на чтение и запись из поддерживаемых источников данных (например, файловая система, SQL база данных и т.д.).

Кроме того, планируется добавление слоя «Choose framework model», который будет выполнять автоматический выбор наиболее подходящей реализации того или иного алгоритма исходя из различных параметров (например, объем данных, место их расположения, особенности задачи).

1. Основы Data Science и Big Data. Python и наука о данных / С. Дэви [и др.]. – СПб.: Питер, 2017. – 336 с.
2. Татур, М. М. Особенности построения вычислителей интеллектуальной обработки данных/ М. М. Татур // Информатика. – Минск, 2015. – № 1(45). – С. 39–44.
3. Проявление закона Амдаля-Густавсона на примере реализации алгоритма k-средних / А. И. Демидчук [и др.] // Междунар. НПК «Big Data and Predictive Analytics. Использование Big Data для оптимизации бизнеса и информационных технологий». – Минск, 2015. – С. 151–154.
4. Применение методов DataMining и KnowledgeDiscovery в оперативно-розыскной деятельности / С. Н. Нефедов [и др.] // Материалы Республиканской научно-практической конференции «Актуальные проблемы оперативно-розыскной деятельности». – Минск, Академия МВД, 2017. – С. 70–72.