Обзор алгоритмов синтаксического анализа и марковской модели переменного порядка для задачи предсказания пользовательских действий

Аронов Александр Вениаминович – студент магистрант Брянского государственного технического университета.

Аннотация: Данная задача была поставлена и реализована в контексте платформы интерактивного микрообучения «HintEd», основной целью является помощь «потерявшимся» пользователям сценариями микрообучения. Для этого в данной статье был проведён обзор и сравнение результатов работы алгоритмов синтаксического анализа и марковской модели переменного порядка для задачи предсказания пользовательских действий на основе заранее записанных последовательностей. По полученным результатам не удалось однозначно выбрать алгоритм для поставленной задачи, однако был найден смешанный подход, который может вполне эффективно удовлетворить поставленные требования.

Ключевые слова: Микрообучение, e-Learning, распознавание действий пользователя, распознавание образов, сопоставление с образом, цепи Маркова, синтаксический анализ, макровская модель переменного порядка.

Ключевой особенностью программной платформы интерактивного микрообучения «HintEd» является возможность распознавания пользовательских действий на странице информационного веб-ресурса, по которым определяется шаблон действий для нахождения похожих сценариев микрообучения, которые будут отображаться в случае, если пользователь не может самостоятельно найти кратчайший путь до задачи. Для этого необходимо, основываясь на последовательности пользовательских событий (нажатие левой кнопки мыши на элементе А, ввод данных в элемент Б, наведение курсора на элемент В), искать среди существующих сценариев микрообучения тот, который соответствует последовательности действий. Последовательность действий может быть верна лишь частично из-за незнания пользователем информационной системы, поэтому алгоритм должен уметь предсказывать возможный следующий шаг на основе натренированных данных, которыми являются сценарии микрообучения.

Исходя из постановки задачи, необходимо формализовать входной формат в вид понятный алгоритму и применить для него один из алгоритмов, который предскажет выходной результат. Так как действия пользователя можно легко классифицировать вручную, то задачу можно решить как методами сопоставления с образом, так и методами распознавания и прогнозирования. Основным критерием для выбора алгоритмов является их простота и скорость работы, алгоритмы должны выдавать результат в реальном времени без запросов к серверу вычислений и без необходимости долгого обучения на тренировочных данных.

Обзор алгоритмов

Синтаксический анализ – процесс сопоставления линейной последовательности лексем (слов, токенов) естественного или формального языка с его формальной грамматикой. Результатом обычно является дерево разбора (синтаксическое дерево) [1].

В ходе синтаксического анализа исходный текст преобразуется в структуру данных, обычно – в дерево, которое отражает синтаксическую структуру входной последовательности и хорошо подходит для дальнейшей обработки.

Как правило, результатом синтаксического анализа является синтаксическое строение предложения, представленное либо в виде дерева зависимостей, либо в виде дерева составляющих, либо в виде некоторого сочетания первого и второго способов представления.

В математической теории стохастических процессов **Марковская модель переменного порядка** [2] (ММПМ) очень важный класс моделей, которые расширяют хорошо известные цепи Маркова. В отличии от моделей цепей Маркова, где каждая случайная величина в последовательности со свойством Маркова зависит на фиксированном количестве случайных переменных, в ММПМ это число случайных переменных может варьироваться в зависимости от специфики выбранной реализации.

Последовательность реализации также часто называют контекстом, а ММПМ контекстными деревьями. Гибкость в числе опциональных случайных переменных даёт

большое преимущество для множества направлений, таких как статистический анализ, классификация и предсказание.

В практических условиях у ММПМ редко бывает достаточно данных для точной оценки экспоненциально возрастающего числа компонентов условной вероятности при увеличении порядка цепи Маркова [3]. ММПМ предполагает, что существуют определенные реализации состояний (представленных контекстами), в которых некоторые прошлые состояния независимы от будущих состояния, соответственно «может быть достигнуто значительное сокращение количества параметров модели».

Сравнение результатов

В качестве решения задачи был реализован алгоритм, который разбивал дерево сценария микрообучения на поддеревья с максимальной глубиной дерева в 5. Каждый узел представляет из себя хеш-код действия пользователя, основанный на пользовательском действии (ввод данных, нажатие на клавишу, наведение на элемент), и элементе, на котором произошло событие. Все пользовательские события формируются аналогичным образом и имеют 5 временных буферов, состоящие от 1 до 5 последних действий, которые передаются моделям для получения результатов. В случае полного или частичного совпадения одного из буферов с веткой в одном из поддеревьев выдаётся результат совпадения с учётом уровня до которого дошла ветка в процентах.

Для алгоритма синтаксического анализа был обнаружен существенный недостаток. Так как основная цель – помощь пользователю, который ищет информацию, то и набор событий, получаемых от пользователя, который впервые познакомился с ресурсом, в 95% случаев не подходит под последовательности, с которыми происходит сравнение. Однако если пользователь уже хотя бы раз проходил сценарий микрообучения, то наблюдается обратная ситуация, последовательность действий пользователя в 64% случаев хотя бы частично соответствует сценарию микрообучения, что позволяет с большой точностью корректно отобразить подсказку, которую ожидает увидеть пользователь.

Подход, с очень малой вероятностью сможет помочь пользователям, которые впервые знакомятся с ресурсом, однако он крайне полезен для пользователей, которые хотя бы раз воспользовались сценарием микрообучения.

Результаты, полученные методом Марковской модели переменного порядка, оказались гораздо стабильнее. Входные данные для модели остались аналогичными алгоритму синтаксического анализа, модель выдавала правильный результат для действий пользователя в 33% случаев. Для того чтобы повысить точность варьировались максимальный порядок модели и длина контекста, однако не удалось однозначно установить улучшение первоначального результата, при этом производительность на слабых устройствах перестала быть удовлетворительной.

Выводы

В ходе статьи были проанализированы и реализованы некоторые из возможных алгоритмов, которые можно использовать для предсказания пользовательских действий, на основе существующих сценариев микроообучения. Среди рассмотренных алгоритмов не нашлось универсального решения проблемы, поэтому решено использовать смешанный подход, который позволит предсказывать действия пользователей с учётом ошибок в их последовательности, а также методы синтаксического анализа, для поиска полного соответствия последовательности на поддеревьях сценария микрообучения.

Список литературы

1. А. Ахо, Дж. Ульман. Теория синтаксического анализа, перевода и компиляции. Т. 1. Пер. с англ. В.Н. Агафонова под ред. В. М. Курочкина. М.: Мир, 1978. 614 с.
2. Rissanen, J. (Sep 1983). "A Universal Data Compression System". IEEE Transactions on Information Theory. 29 (5): 656–664.
3. Кельберт М. Я., Сухов Ю. М. Вероятность и статистика в примерах и задачах. Т. II: Марковские цепи как отправная точка теории случайных процессов и их приложения. — М.: МЦНМО, 2010. — 295 с. — ISBN 978-5-94057-252-7.