Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.422.83

Пухович
Светлана Валерьена

Организациябольшихмассивовиассоциативный поискинформации наосновесамоорганизующихсякарт Кохонена

АВТОРЕФЕРАТ

на соискание на соискание степени магистра технических наук
по специальности 1 - 40 80 03 "Вычислительные машины и системы"

_____

Научный руководитель
Татур Михаил Михайлович
профессор, доктор технических наук

_____

Минск 2016

ВВЕДЕНИЕ

Современной тенденцией является бурный рост обрабатываемой информации. Подобная закономерность является результатом развития компьютерных технологий: даже у рядового пользователя на данный момент больше возможностей хранения данных, чем было в недалеком прошлом, да и мощности вычислительной техники несравнимо ушли вперед, если обратить внимание на последние несколько десятилетий.

Было создано множество электронных сервисов по сбору данных и статистики, поскольку это служит целям современной рыночной экономики. В области хранения информации появились, так называемые, облачные сервисы, которые не только в промышленное, но и в личное пользование предоставляют хранилища информации, что подтверждает мысль об увеличении объемов данных. К тому же, человечество не стоит на месте и исследует все новые и новые области знания, которые требуют быстрого доступа к базовым сведениям и литературе.

Можно сказать, что большие массивы информации становятся нормой для нашего времени, соответственно растет необходимость в их организации и эффективном поиске информации. Некоторыми путями решения этих задач является классификация и ассоциация. Методами классификации и ассоциации являются нейронные сети, а также программный поиск ассоциативных правил.

Таким образом, можно отметить актуальность работы по двум критериям: современность проблемы и исследование передовых методов ее решения.

Одной из целей данной работы является исследование эффективности применения модели нейронной сети к организации и поиска данных в больших массивах информации, определение и обоснование принципов построения подобных структур данных. Также задачей этого исследования можно назвать авторскую реализацию некоторых методов поиска ассоциативных правил и оценку их эффективности. Кроме того, немаловажной частью работы является обзор и анализ существующих методов работы в заданной области, и, как следствие, определение недостатков и достоинств существующих систем обработки больших объемов информации.

Нейронные сети являются моделью, работа которой основывается на примере работы нервных клеток живого организма. Они используются в широком спектре задач: распознавание образов, прогнозирование, управление. Общим для них является то, что необходимо оперировать одновременно достаточно большим массивом информации, а также использовать обобщение, чего зачастую не добиться обычным программированием.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Объектом исследования являются методы организации больших массивов информации, а также способы ассоциативного поиска в больших массивах информации.

Целью работы стала реализация метода ассоциативного поиска информации при помощи самоорганизующихся карт Кохонена, а также исследование того, насколько данный метод будет эффективен в сравнении с некоторыми другими алгоритмами поиска ассоциативных правил.

Работа с большими массивами информации принимает все больший размах, поэтому их организация и ассоциативный поиск являются достаточно актуальными проблемами.

В ходе работы предложен метод ассоциативного поиска информации на основе самоорганизующихся карт Кохонена. На базе этого метода построен алгоритм ассоциативного поиска.

Реализован инструментарий для оценки данного метода по критериям эффективности, а также возможность визуального отображения результатов работы алгоритма поиска.

Некоторые результаты исследований были представлены на конференции «Молодежь в науке 2015» - Академия наук Республики Беларусь, Минск.

Сравнение технологий работы с большими массивами информации было представлено на конференции «Теория и практика современной науки 2014» - Москва, Россия, доклад опубликован.

СОДЕРЖАНИЕ РАБОТЫ

Диссертация состоит из четырех глав.

В первой главе рассматриваются проблемы работы с большими массивами информации, а также дается их определение. Описываются теоретические подходы и задачи, которые решаются при обработке больших данных. В этой главе выполняется постановка целей и задач.

Вторая глава посвящена теоретическим формальным основам. В этой главе математически описываются основы построения нейронных сетей, а также даются формальные алгоритмы различных методов поиска ассоциативных правил.

Третья глава рассматривает вопросы программной реализации модели на базе самоорганизующейся карты Кохонена. В данной главе описан процесс проектирования работы нейронной сети, рассматриваются и выбираются программные средства реализации, а также описывается модель реализованного программного средства.

В четвертой главе описывается процесс тестирования реализованного метода, основанный на его сравнении с уже существующими алгоритмами, реализованными для проверки эффективности.

ЗАКЛЮЧЕНИЕ

В работе затронута важная и актуальная на сегодняшний день тема обработки и организации больших массивов информации. Были проанализированы сложности оперирования таким типом данных, а также описаны основные технологии работы, которые имеют место быть на сегодняшний день.В диссертации также были затронуты темы нейронных сетей и машинного обучения. Данные методы позволяют решать различные задачи работы с большими массивами неструктурированных данных без вовлечения в их работу человека, то есть автоматически. Таким образом, была определена область применения алгоритмов.

В рамках исследования был спроектирован метод поиска ассоциативных правил на основе самоорганизующихся карт Кохонена, которые предназначены для того, чтобы отображать многомерные вектора на пространство с меньшей мерностью, то есть двумерное. В ходе разработки которого были отмечены также и другие алгоритмы, которые решают указанную задачу и на их основепроизводилось тестирование программного средства.

Программное средство обладает следующими достоинствами

– возможность визуализации результатов;

– автоматизация сложной эвристической задачи;

– переносимость;

– поддержка стандартных форматов файлов;

– простой интерфейс;

Область знаний, освещенная в данной работе, является достаточно актуальной и востребованной из-за постоянного роста объемов обрабатываемой информации, а также из-за совершенствования методов исследований в различных областях знания.

В качестве продолжения работы можно реализовать следующие идеи:

- использование не только данных из текстовых файлов,
- получение необходимой информации из сети Интернет,
- применение другого вида визуализации,
- экспериментирование с настройками нейронной сети.

СПИСОК ПУБЛИКАЦИЙ СОИСКАТЕЛЯ

[1-А] Пухович, С. В. Методы интеллектуального аналаиза больших массивов данных / С.В. Пухович // Теория и практика современной науки: материалы научной конференции – Москва, 2014 – С. 105-108