ISSN 1512-1232

004.62

ИНСТРУМЕНТЫ И СИСТЕМЫ ДЛЯ ОБРАБОТКИ И АНАЛИЗА ИНФОРМАЦИИ ДАННЫХ В СИСТЕМАХ БИГ ДАТА

Кулиджанов Константин Борисович
Georgian American University, Georgia, Tbilisi, 0160, Merab Aleksidze str. 10

Аннотация

Наступление эры больших данных принесло новые вызовы и возможности в области анализа информации. В этой статье анализируются новые методы иисследования и анализа в среде биг дата, такие как анализ данных, визуализация, семантическая обработка и т.д. В статье рассматриваются некоторые новые инструменты, такие как Weka, Sitespace и т.д. Для развития новых методов анализа информации и применения их на практике, очень важно использовать современные инструменты в этой области. Необходимо также находить и применять новые модели для анализа информации.

Abstract

Era of big data systems has brought challenges and opportunities to intelligence research. This paper analyzes the emerging techniques of intelligence research under the big data environment, like data mining, visualization, semantic processing, etc. It also summarizes some new tools, such as Weka, Sitespace, etc. In order to promote the development of intelligence theory research and practice, it is vital and useful to explore the updating of intelligence research techniques and tools, and to discover the new model of intelligence analysis.

Keywords*: big data; intelligence research; big data techniques; big data tools.*

I. ВСТУПЛЕНИЕ

На фоне роста обьемов данных и с появлением систем big data, компьютерные технологии, такие как анализ данных и визуализация данных, обеспечивают огромную перспективу для научных исследований. Знания в области искуственного интеллекта, в свою очередь дают возможность для разработки новых методов для обработки и анализа данных. Соответственно, многие инструменты анализа данных систем биг дата в настоящее время широко используются в областях военной разведки, научно-технических исследованиях а также в анализе когнитивных процессов. Все это позволяет развивать разработку инструментов и систем анализа данных

В условиях быстрого роста объемов данных необходимы новые технологии автоматизации процессов. С помощью различных технологий, средств автоматизации и инструментов, необходимо максимально полностью анализировать данные и установить все возможные связи между данными различного типа, для того чтобы в дальнейшем избегать неправильного толкования и понимания полученной информации. [1]

ISSN 1512-1232

II. ПРИМЕНЕНИЕ СИСТЕМ БИГ ДАТА ДЛЯ АНАЛИЗЕ ДАННЫХ

Развитие систем и методов анализа данных, дало возможность проводить исследования не только методами качественного анализа или простой статистикой. [2] В наше время исследования в области технологий и методов анализа данных открывают перед нами новые возможности. В мае 2011 года McKinsey Global Institute опубликовал свой исследовательский отчет: Big data: следующий рубеж для инноваций, конкуренции и производительности. Этот доклад разделен на шесть частей. Особо важной является вторая часть, в которой обсуждались технологии применения систем Big data в трех аспектах: методы анализа информации расположенной в системах Big data, технологии и методы обработки данных и визуализация данных. В методах анализа данных систем Big Data перечислены 26 различных технологий анализа, подходящих для многих отраслей. Рассмотрены методы clustering analysis, crowdsourcing, data mining, natural language processing, network analysis, predictive modeling, regression, visualization и т.д. Большинство из этих методов уже существуют; С развитием глобальной информационной сети Интернет и растущим спросом на анализ больших обемов данных генерируемых в этой сети, возникла необходимость их обработки и анализа для принятия последующиз решений, поэтому некоторые из перечисленных методов были специально разработаны и оптимизированны для этого. Эти методы можно разделить на методы хранения данных, обработки больших обьемов данных расположенных в системах big data, методы анализа больших обьемов данных и визуализации результатов анализа. Среди них первые два метода являются основой для создания массивов big data, а последний наиболее часто используется в области искуственного интеллекта, и ему следует уделять наибольшее внимание. Аналитика данных Big data и визуализации больших данных в основном включают анализ собранных данных, аналитику, визуальную аналитику, представление полученных результатов а также семантический анализ.

A. Анализ данных

Анализ данных, как правило, сводится к процессу поиска скрытой информации в большом обьеме данных при помощи различных алгоритмов. Анализ данных связан с информатикой и с помощью различных методов, таких как статистика, аналитическая обработка, поиск информации, машинное обучение, распознавание образов, применение экспертных систем, представления знаний из баз данных. [3] Основная задача анализа данных - идентифицировать модели на основе больших объемов данных. В зависимости от поставленной задачи анализ данных можно разделить на множество типов; наиболее типичными примерами являются корреляционный анализ, классификационный анализ на основе дерева принятия решений или нейронной сети, кластерный анализ, анализ последовательностей и т. д.

Основой анализа информации расположенной в системах big data является алгоритм. Каждый алгоритм анализа данных, основанный на разнообразных типах, структурах и форматах данных, раскрывает их специфику и свойства. Кроме того, обработка данных в системах big data с применением различных распространенных алгоритмов анализа и обработки данных, позволяют получить внутреннюю, скрытую часть данных и позволяют получить скрытую информацию.

С точки зрения концепции анализа данных, она имеет естественную связь с теорией информации. С точки зрения методов анализа данных она содержит особые свойства и процессы реализации, которые могут быть использованы для решения задач исследования информации. Однако множество алгоритмов анализа данных используются только для решения простых задачь, таких как например подсчет статистики, подсчет общих слов и решения других задачь, основанных на текущих результатах исследования информации. В процессе анализа данных эти простые приложения являются только предварительной обработкой данных, далее требуется более глубокий анализ полученной информации.

ISSN 1512-1232

B. Визуальная аналитика и представление данных

Визуальная аналитика - это метод анализа взаимосвязей посредством интерактивной визуализации, которая облегчает конечным пользователям принятие решений. В результате создаются визуализации аналитических данных и таблицы, которые зависят от обьема, сложности и структуры данных. В визуализациях и таблицах представлена информация обо всех связанных событиях, результатах анализа данных и тренды. Визуальная аналитика отличается от визуализации информации, которая фокусируется на графическом представлении сгенерированной информации или отчетов а также на их разработке и дизайне.

Визуальная аналитика развивается на основе визуализации информации и фокусируется на выборе методов анализа и сочетании методов анализа с техниками визуализации, для достижения поставленных целей и последующего принятия решений. Визуальная аналитика, одна из актуальных и активных тем для исследований в области анализа информации, которая может значительно улучшить эффект процесса анализа данных. Системы визуализации информации помогают преодолеть недостатки традиционных методов анализа данных и позволяют проводить анализ информации на под другим ракурсом. Они позволяют раскрыть, глубоко понять и проанализировать скрытую информацию, которую трудно или невозможно найти и получить с помощью ранееиспользуемых стандартных методов. Метод визуализации позволяет генерировать ценные выводы для последующего принятия решений, которые значительно улучшают эффект и качество анализа информации. [4]

Пользователи систем big data как и стандартные пользователи имеют возможность экспорта данных и результатов анализа. Основным требованием пользователей является получение результатов в виде визуальной аналитики, так как визуальная аналитика способна напрямую представлять результаты обработки данных систем big data, а также по сравнению с таблицами и цифрами позволяет легко воспринимать результаты в визуальном виде.

C. Семантический анализ

Семантика - это наука о смысловых значениях. Семантический анализ проверяет, есть ли семантическая ошибка в исходном коде, чтобы собрать данные для фазы генерации кода. Процесс семантического анализа заключается в проверке контекста, типов для правильно структурированного исходного кода. Методы семантического анализа позволяют машине более эффективно понимать данные и код. Происходят процесы интеграции языка, индексации, методов работы с базами данных а так же других техник с целью облегчения процесса обработки информации, ее интеграции и повторного использования структурированных или неструктурированных данных.

Основные методы семантического анализа включают семантическую маркировку, выборку знаний, индексирование, моделирование, выводы и так далее. Семантические методы являются хорошей основой для глубокого анализа данных, дают возможность распознать потенциальные шаблоны которые содержат данные, посредством семантического процесса для различных типов информации и алгоритмов анализа данных для структурированных данных с извлеченной семантикой.

III. ПРИМЕНЕНИЕ СИСТЕМ БИГ ДАТА И РАЗЛИЧНЫХ ИНСТРУМЕНТОВ ДЛЯ АНАЛИЗА ДАННЫХ

Анализ данных - самая важная часть систем big data, на основании полученной информации он помогает принимать более обоснованные решения. Таким образом, на основе вышеуказанных технологий и платформ, применение инструментов анализа данных big data в исследованиях можно разделить на три типа: инструменты анализа данных, инструменты визуального анализа и инструменты семантического анализа.

ISSN 1512-1232

A. Инструменты анализа данных

Из-за ограниченного количества существующих инструментов предназначенных для исследований и анализа данных, в процессе анализа информации часто приходится применять инструменты из интердисциплинарных областей. Это приводит к выполнению дополнительных операций и дополнительной работе, так как используемые инструменты имеют различный функционал и их приходится использовать поочередно либо одновременно. Кроме того это создает проблемы, связанные с отсутствием целостности результатов анализа данных, что может привести к принятию неверных решений. Ниже будут рассмотренны два наиболее часто используемых инструмента анализа данных: Weka и RapidMiner.

FIGURE I. ИНТЕРФЕЙС СИСТЕМЫ WEKA

1) WEKA, программное обеспечение *с* открытым исходным кодом: WEKA - это сокращение от Waikato Environment for Knowledge Analysis, это бесплатное, некоммерческое программное обеспечение для машинного обучения и анализа данных, основанное на JAVA. WEKA работает как открытая платформа для анализа данных и объединяет в себе большое количество алгоритмов машинного обучения. Алгоритмы могут выполнять задачи анализа данных, включая их предварительную обработку, классификацию, регрессию, кластеризацию, правила ассоциации и визуализацию результатов анализа в новом интерактивном интерфейсе.

В августе 2005 года на 11-й международной конференции ACM SIGKDD команда WEKA из Университета Вайкато получила высшую награду в области анализа данных. Система WEKA признана одной из самых популярных систем анализа данных и машинного обучения. Это один из самых полных инструментов анализа информации, который загружается десятки тысяч раз в месяц.

ISSN 1512-1232

FIGURE II. ИНТЕРФЕЙС СИСТЕМЫ RAPIDMINER

2) RapidMiner, программное обеспечение с открытым исходным кодом для анализа текста и данных различных типов: RapidMiner – одно из ведущих в мире решение для анализа данных, владеющее передовыми технологиями. Его функционал охватывает широкий спектр, включая различные виды структуризации данных, которые могут упростить планирование и оценку процесса анализа данных.

RapidMiner предлагает бесплатную технологию анализа данных и базу данных, 100% Java код, интуитивно понятный процесс, широкие возможности. С помощью простого языка сценариев система может автоматически обрабатывать данные больших обьемов. [5] Она показывает многоуровневое представление данных для обеспечения эффективности обработки и анализа информации, а также обладает мощным механизмом визуализации и расширенным визуалным моделированием многомерных данных больших обьемов, поддерживает более 400 операторов анализа данных. Система успешно используется для решения различных задачах Йельского университета, включая анализ текстов, анализ мультимедийной информаци, функциональный дизайн, анализ потоков данных, интегрированную разработку и распределенный анализ данных.

B. Системы визуальной аналитики

Текущие системы анализа данных предлагают различные методы обработки информации а также позволяют представлять в удобной форме результаты анализа для принятия последующих решений. В то же время существующие инструменты анализа, в процессе работы с данными должны позволять применять всевозможные разнообразные параметры. Вместе с тем аналитик не имеет возможности применять когнитивный процесс анализа, что увеличивает сложность процессов анализа информации. Однако визуальная аналитика позволяет решить эту проблему, она объединяет методы применяемые во многих областях, включая анализ информации и научный анализ. Позволяет применять результаты из областей управления данными, представления знаний, статистического анализа, получения знаний для автоматзации процессов анализа. Она координирует связь между человеком и машиной, чтобы лучше представлять, облегчать понимание и анализ полученных результатов. Ниже мы рассмотрим несколько систем визуализации: Pajek, UCINET, Jigsaw и Citespace.

1) Pajek, программное обеспечение визуализации информации: Pajek - это специально разработанная программа сетевого анализа и визуализации для обработки больших обьемов

ISSN 1512-1232

данных. Это полезный инструмент, используемый для изучения всех видов сложных нелинейных сетей. [6] Pajek работает в среде Windows и используется для анализа и визуализации больших сетевых систем с тысячами или даже миллионами узлов. Pajek может обрабатывать одновременно несколько сетей, а также сети событий, дополнительно система содержит инструмент для лонгитюдинального анализа сети. [7]

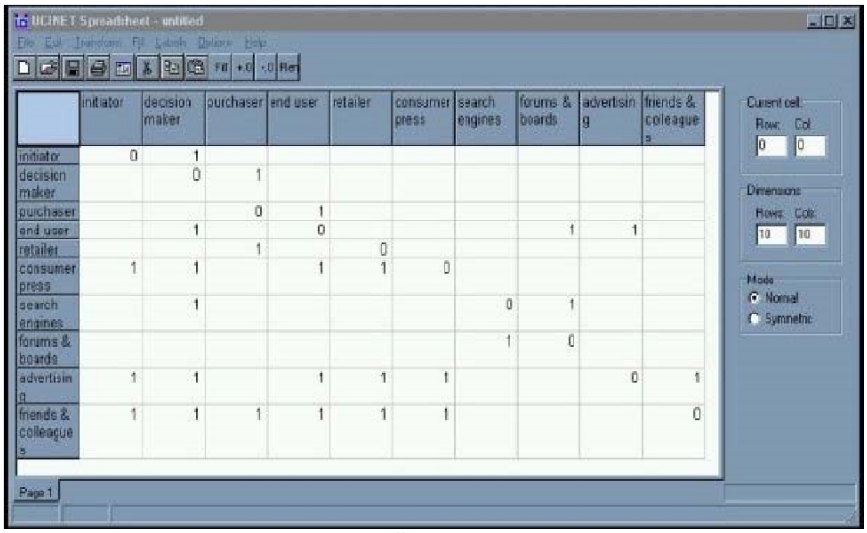

FIGURE III. ИНТЕРФЕЙС СИСТЕМЫ UCINET

2) UCINET, программное обеспечение для визуализации информации: UCINET на сегодняшний день является самым популярным программным обеспечением для анализа социальных сетей. Она содержит методы анализа социальных сетей, включая центральный анализ, анализ подгрупп, анализ характера, статистический анализ на основе смещения и т. Д. [8]

UCINET создан группой аналитиков сетей из Калифорнийского университета в Ирвине. Программное обеспечение для интеграции анализа сетей UCINET включает в себя программное обеспечение для одномерного и двумерного анализа данных NetDraw, программное обеспечение для анализа трехмерного отображения Mage, а также содержит бесплатное прикладное программное обеспечение для анализа сетей крупного масштаба. [7] Кроме того, пакет имеет мощные функции матричного анализа, такие как матричная алгебра и многомерный статистический анализ. Это одна из самых популярных программ для анализа социальных сетей, она проста в использовании и подходит для начинающих пользователей.

3) Jigsaw, система визуализации анализа документов: Джон Стаско и др. из Технологического института Джорджии, создали систему анализа визуализации под названием Jigsaw, основанную на выдвинутой Пиролли концептуальной модели анализа. Они применили ее к области академических и сетевых исследований, а также продемонстрировали возможность применения технологии визуализации анализа в своих исследования.

Jigsaw - это система, позволяющая визуализировать все виды документов с помощью алгоритма интеллектуального анализа текста. Она может генерировать диаграммы кластеризации документов, временные шкалы, диаграммы и т.д. Моделирование выполняется автоматически в зависимости от поставленных задачь. Также пользователи могут свободно настраивать параметры и внешний вид диаграмм.

ISSN 1512-1232

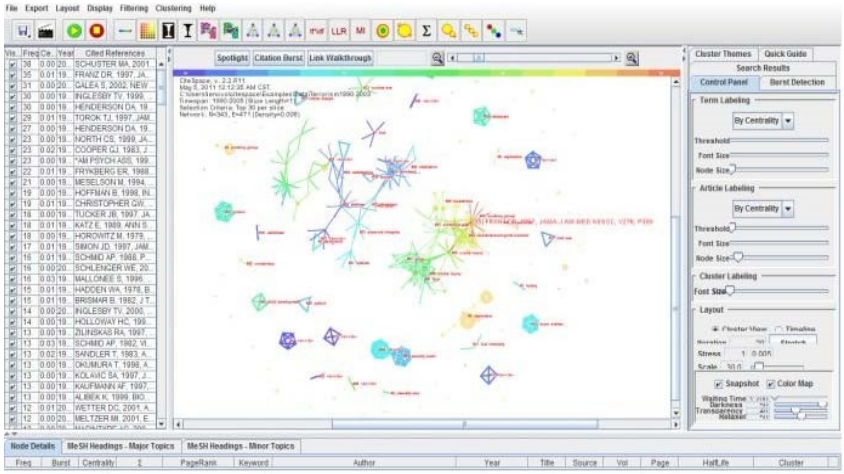

FIGURE IV. ИНТЕРФЕЙС СИСТЕМЫ

4) Citespace, следующий инструменты визуализации: Citespace и его обновленная версия инструмента визуализации в областях представления знаний, которые широко используются. [9] Инструмент разработан доктором Чаомей Чен, и его можно использовать бесплатно. Пользователи могут получить визуализации на основе набора данных, а результаты представления знаний являются стабильными, информативными и удобочитаемыми.

CiteSpace объединил в себе интеллектуальный анализ текстов, визуализацию информации и наукометрию, сформировал методы визуализации, пригодные для множественной визуализации, разделения времени, динамического анализа, и продвинул научные и технологические исследования на новый уровень на основе картографирования областей знаний и визуализации знаний.

C. Инструменты семантического анализа

Отсутствие анализа семантики – наиболее частая проблемма в существующих инструментах и алгоритмах анализа информации. В научных исследованиях, таких как научные статьи, патенты и т. д., многие применяемые инструменты анализа данных не имеют семантической поддержки. Для быстрого анализа объектов данных, таких как новости, блоги и т. д., анализ все еще проводится стандартными способами. В настоящее время он в основном проводится аналитиками, которые выявляют необходимую информацию и объединяют ее в структурированные данные. Наличие вышеуказанных проблем делает семантические технологии наиболее актуальными и их внедрение становится неизбежным.

1) Annotation.cn, платформа для крупномасштабных семантических исследований: Annotation.cn - это многодоменная система службы аннотаций, основными ее частями являются: модуль управления словарями, модуль тегов, модуль управления пользователями. Управление словарями включает четыре функции: менеджмент словарей, управление типами данных, управление концепциями, представление концепций; Модуль тегирования содержит две функции: управление документами и аннотации документов; Управление системой включает функции управления пользователями и их ролями.

IV. ЗАКЛЮЧЕНИЕ

Развитие систем биг дата является дополнительной возможностью для обработки и анализа данных, их организации, хранения и. Это позволяет присменять новые методы в

ISSN 1512-1232

области анализа данных, открывает новые возможности в областях управления, анализа, использования, обработки и хранения данных. С началом эры систем биг дата необходимы новые инновационные методы и инструменты для исследований и анализа данных. Новые инструменты и методы создаваемые для работы с системами биг дата открыли как дополнительные возможности, так и проблемы в области исследования и анализа данных. Исследователям и специалистам необходимо применять новые возможности, активно примененять новые инструменты и методы для интеграции, обработки, организации и использования данных систкм биг дата. Это позволит улучшить уровень хранения и обслуживания данных, способствовать росту эффекта управления и результатов анализа, получать новые знания и применять их.

Список использованной литературы:

[1] Li Guangjian, Yang Lin. Intelligence Analysis and Intelligence Technology in View of Big Data. Library & Information, 2012 (6): 1-8.

[2] Gu Tao. Research on Collaboration Analysis of Competitive Intelligence Based on Big Data. Information Science. 2013, 31 (12): 114-118

[3] Tang Zhixiong, Xian Donglai. The Implement Method of Analyzing Customer Retention by Data Mining Technology. Information & Communications. 2011 (2): 99-100.

[4] Tang Tianbo, Gao Feng. Case Study of Visual Analytics in Intelligence Research. Information Studies: Theory & Application. 2009,8 (32): 63-67.

[5] Liu Zhilong. Data Analysis and Data Mining Application in Statistics Industry. Statistics and consultation. 2014 (1): 36-38.

[6] Zhou Qingshan, Zhao Xue, Zhao Xuyao, Zhou Gefei. Knowledge Mapping Analysis in Digital Content Industry in China. Information Studies: Theory & Application. 2014, 35 (4): 56-61

[7] Fei Zhonglin, Wang Jingan. Social Network Analysis: Method and Perspective of Management Research. Science and Technology Management Research. 2010 (24): 216-219.

[8] Li Gang, Li Ang. Study on Coauthorship Based on Social Network Analysis. Journal of Information Resources Management. 2011 (3): 43-47.

[9] Liao Shengjiao. The Comparative Study on the Scientific Knowledge Mapping Tools: OSviewer and Citespace. Sci-Tech Information Development & Economy. 2011, 21(7): 137-139.

Article received: 2020-12-23