ОСОБЕННОСТИ ОБРАБОТКИ НЕСТРУКТУРИРОВАННЫХ ДАННЫХ В ИНФОРМАЦИОННОЙ БАЗЕ НАУЧНЫХ ИССЛЕДОВАНИЙ ВУЗА

FEATURES OF PROCESSING OF UNSTRUCTURED DATA IN INFORMATION BASE OF SCIENTIFIC RESEARCH OF HIGHER EDUCATION INSTITUTION

B. Ukuev

Summary. Databases contain mainly the structured information, but the considerable volume of primary information is unstructured. Computer processing of unstructured information is very difficult. The main step on coercion of unstructured data to structured — determination of ontology of data domain. The efficiency of processing of unstructured information can be promoted by application of the object-oriented approach and, respectively, data storage in object-oriented databases.

Keywords: structured data, unstructured data, ontology, object-oriented approach, object-oriented DBMS.

Укуев Бейшенбек Такырбашович
Д.т.н., профессор, Кыргызский государственный университет строительства, транспорта и архитектуры им.Н.Исанова (г. Бишкек)
ukuevb@gmail.com

Аннотация. Базы данных содержат главным образом структурированную информацию, но значительный объем первичной информации является неструктурированной. Компьютерная обработка неструктурированной информации весьма затруднительна. Основной шаг по приведению неструктурированных данных к структурированным — определение онтологии предметной области. Эффективности обработки неструктурированной информации может способствовать применение объектно-ориентированного подхода и, соответственно, хранение данных в объектно-ориентированных базах данных.

Ключевые слова: структурированные данные, неструктурированные данные, онтология, объектно-ориентированный подход, объектно-ориентированная СУБД.

Структурированные и неструктурированные данные представляют собой два разных класса данных в информационном пространстве, которые необходимы для анализа каждого класса. Изучая закономерности типов данных в отдельности, можно выделить основные закономерности, присущие только конкретному классу. Например, используя структурированные данные о котировках акций той или иной компании, можно оценить динамику роста или падения на фондовой бирже. При обработке неструктурированных данных, таких как, например, публикации в средствах массовой информации, открывается возможность исследовать характер и фон вокруг той или иной компании и сформулировать общую оценку экспертного характера и определить влияния данных на котировки акций. Работая с данными о конкретном событии или процессе, можно осознанно и целенаправленно управлять этим процессом.

В общем представлении бизнес-анализа организация должна иметь возможность анализировать структурированные и неструктурированные данные. Определенный набор средств анализа позволят провести совместные исследования. Вместе с тем, сегодня наблюдается слабая интеграция систем анализа структурированных и неструктурированных данных — совместный анализ данных из различных источников пока возможен только при условии, что структуры этих данных перед применением инструментов анализа приведены к схожему виду. То есть неструктурированные данные должны быть структурированы, так как именно для структурированных данных наиболее развит математический и функциональный аппарат подготовки и анализа.

По данным специалистов, около 70 процентов внутрикорпоративного информационного пространства имеют неструктурированный или частично структурированный характер, к которому относятся файлы различных форматов (фото, аудио и видео, электронная почта), несущие в себе большой потенциал для анализа.

Неструктурированные данные характеризуются рядом признаков, затрудняющих их обработку средствами стандартного аналитического инструментария, но при этом как раз и составляющих уникальный потенциал для получения новых знаний. С одной стороны, она очень разнообразна, она неоднозначна — одинаковый набор данных может содержать разный смысл в зависимости от контекста, языковых и культурных особенностей. С другой, она динамична — со временем меняется структура данных, их значение. Кроме того, неструктурированные данные зачастую носят субъективный и эмоционально окрашенный характер. Таким образом, а также анализ не учтенных ранее данных, выделение дополнительных и неявных предметных областей, пересечение и взаимовлияние предметных областей являются на сегодняшний день основным предметом изучения аналитиков в сфере неструктурированных данных.

Для понимания о методах исследования неструктурированных данных очень важно понятие онтологии — совокупности схемы описания предметной области и правил

Таблица 1. Методы обработки неструктурированной информации

Способы	Характеристика
Осуществляется выделение онтологии– описания-схемы предметной части, характеризующейся конкретной структурой.	С применением семантического анализа данных, набора написанных лингвистических правил создается наполнение указанной онтологии (показание данных из информационного потока). Логическая форма представления структурирована, в связи с этим наполняющим ее данным используема реляционная алгебра.
Осуществляется поиск упоминаний, категоризация и извлечение фактов.	Использует поиск по ключевым фразам, выделение связанных с объектами поиска фактов и может быть применен как на данных онтологии, так и на неструктурированном тексте.
Выполняется выделение эмоциональной окраски, оценки интересов, отношения.	Является семантическим анализом на базе лингвистических правил, применяемых после выделения онтологии.
Выделение закономерностей — динамика и процесс изменения отношения, выделение общего, заимствования.	Основываются на выделенные в рамках онтологии концепты-сущности, их атрибуты и связи.

отнесения данных к этой предметной области. В качестве схемы она должна иметь концепты — сущности, атрибуты сущности и, в обязательном порядке, связи. При этом связи должны быть нагруженными, то есть включать основные и дополнительные атрибуты, которые позволяют отразить служебную информацию: эмоциональный оттенок отношения, предмет связи, способ и т.д. Для связей определяются критерии — правила отбора данных, удовлетворение которых позволяет отнести данные из неструктурированного потока информации к той или иной предметной области.

Рассмотрим возможности обработки неструктурированных данных (табл. 1). Указанные методы ограничивают влияние на способы обработки неструктурированных данных, они требуют обязательного участия человека, отвечающего за формирование запросов и схем предметных областей: онтологий, описаний грамматических правил, а также за обучение системы и настройку семантического анализа. Осмысленное определение предметной области (источники данных, критерии и особенности, обязательные сущности, атрибуты и связи) в настоящее время способен сделать только человек.

Определение структуры предметной области (онтологии) — это основной шаг по приведению неструктурированных данных к структурированному виду. Каждая самостоятельная предметная область — это только определенная часть неструктурированного набора данных.

Интеграция систем анализа структурированных и неструктурированных данных способствует организации обработки всех данных компании, проведению анализа перекрестного влияния различных сведений, обнаружению наложения и пересечения данных, скрывающих новые знания, влияющих на качество и обоснованность принимаемых решений.

В Кыргызском государственном университете строительства, транспорта и архитектуры им. Н. Исанова (КГУСТА) ведется подготовка специалистов для широкого спектра областей производства. При проведении научных исследований формируются группы из студентов и преподавателей различных направлений, что обеспечивает весь необходимый теоретический и практический потенциал для реализации проектов. В то же время информационное обеспечение научных исследований, а также учебного процесса требует создания банка знаний с большим объемом информации, поступающей из различных источников, автоматизация первичной обработки которой в настоящее время весьма затруднительна. Помимо текстовой информации обработке подлежит информация в графической форме, связанная с различными чертежами, эскизами и фотографиями. Обработка текстовой информации на государственном языке вносит дополнительные трудности.

В настоящее время специалисты Института новых информационных технологий КГУСТА работают над объединением имеющихся информационных ресурсов университета в единую информационно-поисковую систему и созданием интегрированной системы обработки данных. При обработке информации планируется широкое применение объектно-ориентированного подхода, обеспечивающего не только выделение определенных атрибутов, но и методов их обработки. Соответственно, рассматривается также возможность построения системы на базе одной из объектно-ориентированных СУБД.

ЛИТЕРАТУРА

1. А. Оганесян. Неструктурированные данные 2.0 // Открытые системы. СУБД, 2012, № 04
2. Б.Т. Укуев. Теория и методы моделирования управленческих и инженерных задач на базе новых информационных технологий. // Бишкек, Илим, 2014.— 220с.