РОЗРОБКА МЕТОДУ ВИКОРИСТАННЯ ТЕХНОЛОГІЙ BIG DATA ТА DATA MINING В ІНТЕЛЕКТУАЛЬНИХ СИСТЕМАХ ОБРОБКИ НЕСТРУКТУРОВАНИХ ДАНИХ

Демченко О. Е.
Науковий керівник – к.т.н., проф. Ситніков Д. Е.
Харківський національний університет радіоелектроніки
(61166, Харків, пр. Науки,14, каф. Системотехніки, тел. (057) 702-13-06)
e-mail: oleksandr.demchenko@nure.ua

Many companies and organizations have difficulty trying to find out what employees or customers think about products or services, about a team or an internal organization, about working conditions or the quality of the equipment. However, if the company or organization is large enough, the process of interviewing and processing the information received becomes an extremely difficult problem. It will be best to use software systems to help solve these problems.

Багато компаній та організацій стикаються з труднощами при спробі дізнатися думку працівників або споживачів (клієнтів) про продукти або послуги, про колектив або внутрішню організацію. Але, якщо компанія досить велика, процес опитування та обробки отриманої інформації перетворюється в надзвичайно складну проблему. Оптимально буде використовувати програмні системи, які допоможуть вирішити ці проблеми.

Щодня дані генеруються, збираються у величезній кількості, але багато разів вони залишаються невикористаними без отримання корисної та змістовної інформації. Дані, що генеруються різними каналами, зберігаються у структурованих, а також неструктурованих формах. Неструктуровані дані розширюють можливість будь-якого бізнесу отримувати більшу інформацію з наборів даних. Неструктуровані дані – це найважливіша частина даних для будь-якого бізнесу. Інструменти можуть допомогти підприємствам використовувати ці дані з максимальним потенціалом. Ці неструктуровані дані потрібно перетворити на щось трохи корисніше.

Програмний засіб, що розробляється – засіб для опитування респондентів та подальшої обробка даних за допомогою Big Data та Data Mining. В ньому використовується NoSQL база даних MongoDB – це cross-платформна, документо-орієнтована система керування базами даних (СКБД), яка не потребує опису схеми таблиць. На відміну від традиційного SQL, в MongoDB є колекції. Колекції можуть містити найрізноманітніші об'єкти, що мають різну структуру і різний набір властивостей. Для обробки отриманих даних потрібно провести інтелектуальний аналіз. При розробці програмного продукту був використаний один з найбільш затребуваних методів – метод пошуку асоціативних правил. Даний метод призначений для виявлення взаємозв'язків між наборами даних з статистики. При розробці зупинилися на найбільш відомому алгоритмі – алгоритмі Apriori. Наведемо позначення, що використовуються в алгоритмі:

L_k – множина k-елементних наборів, чия підтримка не менше заданої користувачем, C_k – множина потенційно частих k-елементних наборів. Наш алгоритм пошуку асоціативних правил Apriori має наступний вигляд:

Крок 1. Присвоїти k = 1 і виконати відбір всіх 1-елементних наборів, у яких підтримка більше мінімально заданої користувачем Suppmin.

Крок 2. k = k + 1.

Крок 3. Якщо не вдається створити k-елементні набори, то завершити алгоритм, інакше виконати наступний крок.

Крок 4. Створити множину k-елементних наборів кандидатів з частих наборів. Для цього необхідно об'єднати в k-елементні кандидати (k-1) - елементному частому набору. Кожен кандидат буде формуватися шляхом додавання до (k-1) – елементного частому набору - p елемента з іншого (k-1) - елементного частого набору - q. Причому додається останній елемент набору q, який по порядку вище, ніж останній елемент набору p. При цьому всі k-2 елементи обох наборів однакові.

Крок 5. Для кожної транзакції T з множини D вибрати кандидатів C_t з множини C_k, присутніх в транзакції T. Для кожного набору з побудованого множини C_k видалити набір, якщо хоча б одне з його (k-1) підмножин не часто зустрічається тобто відсутнє в множині L_{k-1}.

Крок 6. Для кожного кандидата з C_k збільшити значення підтримки на одиницю.

Крок 7. Вибрати тільки кандидатів L_k з множини C_k, у яких значення підтримки більше заданої користувачем Suppmin. Повернутися до кроку 2.

Результатом роботи алгоритму є об'єднання всіх множин L_k для всіх k.

Головною особливістю використання MongoDB є те, що вона має вбудовану структуру MapReduce, яка добре піддається реалізації алгоритму Apriori. Реалізований модуль системи дає змогу виявити закономірності в даних для їх подальшої обробки.

Список використаних джерел:

1. Барсегян А.А., Куприянов М.С., Степненко В.В., Холод И.И. Технологии анализа данных. Data Mining, Visual Mining, Text Mining, OLAP. − Спб.: БХВ-Петербург, 2004. − 250 с.

2. 5th International Conference on System Modeling & Advancement in Research Trends, 2016 College of Computing Sciences & Information Technology, Teerthanker Mahaveer University, Moradabad , India

3. Офіційна сторінка документації MongoDB. URL: https://docs.mongodb.com/manual/