ПРОБЛЕМА ВІЗУАЛІЗАЦІЇ ДАНИХ ВЕЛИКОГО РОЗМІРУ

Танянський О.С.

Харьковский национальный университет радиоэлектроники

На перше місце у питаннях проблематики великих даних поставлений аналіз даних, тому що при обробці отриманих результатів припускають, що будуть застосовуватися вже існуючі методи у вигляді генерації звітів, а також побудови різного роду діаграм або графіків. Однак для перегляду результатів аналізу існуючі методи можуть бути незастосовні відразу з кількох причин. По-перше, велика кількість даних на вході породжує велику кількість результатів аналізу на виході – якщо раніше багато закономірностей виявлялися за межами статистичної похибки, то тепер вони чітко долають цей бар'єр. Може скластися враження, що цим можна знехтувати і для прийняття рішень обмежитися тільки ключовими закономірностями, але це не так. У випадку великих даних для досягнення максимальної ефективності прийнятих рішень потрібно враховувати навіть ледь помітні закономірності, інакше взагалі немає сенсу в обробці великих потоків найрізноманітніших відомостей. По-друге, значно ускладнюється концептуальна модель вихідної інформації. Якщо в інформаційних сховищах типовий звіт мав не більше десятка вхідних параметрів (наприклад, часовий зріз, регіон, тощо), а більш точна параметризація була не потрібна, оскільки звіт банально ставав виродженим і складається з порожніх рядків і нулів, то для великих даних така вирожденність зникає.

Задачі аналізу даних можна представити у вигляді рівнобедреного трикутника (рис. 1), в основі якого лежить обсяг вихідних даних, а будь-яка горизонтальна пряма, проведена на деякому рівні, показує, як багато результатів будуть давати відповідні методи аналізу.

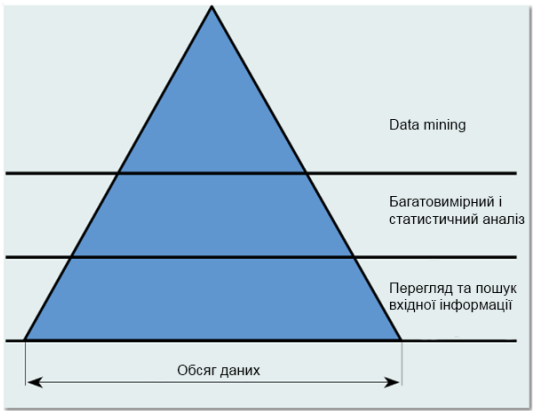

Рисунок 1 – Представлення задачі аналізу даних

По мірі зростання обсягу вихідних даних вершина трикутника спочатку проходить рівень простого пошуку і перегляду даних, потім багатовимірного і статистичного аналізу, і лише потім виходить на рівень Data Mining. Однак по мірі зростання обсягів вхідних даних і на рівні Data Mining стає занадто багато

вихідної інформації. Виходить, що якщо раніше для прийняття рішень необхідно було переглянути лише кілька листів звіту, то в разі великих даних це не так. На людину, що відповідає за прийняття рішень, звалюється купа даних, з якої необхідно виділити найбільш важливі. Цю проблему можна вирішити двома способами.

Перший полягає у використанні автоматизованих засобів, що дозволяють вибрати найбільш важливі звіти, наприклад, у разі моніторингу динаміки продажів відбирати звіти, в яких спостерігається найбільш різка зміна показників в порівнянні з попереднім періодом. Такий метод можна застосовувати не завжди, оскільки зміна показників може просто пояснюватися зовнішніми факторами, невідомими системі: наприклад, різке зростання попиту на газовану воду може пояснюватися спекою і зовсім не говорить про те, що на наступний період необхідно планувати такі ж обсяги продажів.

Другий спосіб боротьби з великою кількістю звітів полягає в реорганізації роботи: виділяються окремі люди, в обов'язки яких входить перегляд звітів і формування резюме, які направляються особам, які приймають рішення. Формування таких резюме в значній мірі відрізняється від існуючих засобів генерації звітів з наступних причин. По-перше, в резюме може входити сама різнорідна інформація. Тобто один і той же документ може включати в себе і показники продажів, і динаміку зростання цін, і зміни в штатному розкладі, і що завгодно ще. По-друге, всі резюме унікальні і навіть можуть не мати загального формату. По-третє, резюме завжди готуються для конкретної людини по конкретному випадку, а отже, враховують і специфіку випадку, і особливості цієї людини. Резюме повинно дозволяти максимально швидко і наочно отримати всю інформацію, необхідну для прийняття рішень.

Однак ці засоби і методи не такі прості, як здається на перший погляд, і сфокусовані на відбір і максимально наочне уявлення різнорідної інформації. Вирішення цього завдання вручну створює високе навантаження на персонал, а хоча б часткова автоматизація роботи вимагає міждисциплінарних наукових досліджень, що стосуються перебування найбільш ефективних способів подання відомостей в резюме. На даний момент подібні системи здаються надмірностями, але в міру зростання числа практичних завдань, пов'язаних з аналізом великих даних, швидка і зручна підготовка звітів стане життєво необхідна.