У Д К 004.912

**В.В. Костенко, Д.І. Оболкін, В.І. Фрінцко**
Університет митної справи та фінансів, Україна

## ДЕЯКІ АСПЕКТИ УПРАВЛІННЯ ДАНИМИ В СИСТЕМАХ ОБРОБКИ ІНФОРМАЦІЇ

**V.V. Kostenko, D.I. Obolkin, V.I. Frintsko**

## SOME ASPECTS OF DATA MANAGEMENT IN INFORMATION PROCESSING SYSTEMS

Неструктурованими данними зазвичай вважається інформація, яка або не має наперед визначеної структури даних, або не організована в установленому порядку. Неструктуровані дані, як правило, представлені у вигляді тексту, який може містити такі дані, як дати, цифри і факти. А це одразу призводить до труднощів аналізу, особливо в разі використання традиційних програм, призначених для роботи зі структурованими даними.

Неструктуровані дані, як правило, зберігаються в зручних для сприйняття людиною форматах, хоча такі формати ускладнюють автоматичне управління даними.

На думку експертів, 80-85% всіх даних існують в неструктурованих форматах. Як приклад: службові записки, медичні записи, юридичні контракти, новини в соціальних мережах, повідомлення електронної пошти. В останніх двох зазвичай ще й використовуються "сленгові" слова та вирази. А це ще більше ускладнює їх структурування.

Якщо структуровані дані – це, як правило, джерело кількісних фактів, то в неструктурованих даних часто можна знайти більш цікаві і потенційно більш цінні експертні оцінки і висновки. У сучасному світі в умовах безпрецедентно швидкого створення величезних обсягів текстової інформації ймовірність того, що ці дані можна буде ефективно використовувати, без їх додаткової обробки, в тому числі автоматичними засобами, невелика. А отже, автоматизація аналізу тексту – непроста задача. Зазвичай для її вирішення потрібні: додаткові відомості про текст, створення потрібних словників або онтологій.

Управління неструктурованою інформацією набуває все більшого значення з трьох причин.

По-перше, з часом така інформація стає все більш структурованою. XML та інші засоби розмітки спрощують процес пошуку, класифікації, сортування і створення звітів для інформації, що зберігається в файлах, а не в структурованих базах даних.

По-друге, проблеми, пов'язані з доступом до файлів і їх збереженням, сьогодні стають все менш гострими завдяки не припиняється вже більше десяти років роботи з налагодження операційних систем і відкритих стандартів в області вилучення та зберігання даних.

По-третє, системи роботи з неструктурованою інформацією оснащуються все новими функціями, що полегшують використання цієї інформація для бізнес-цілей. Паралельно з цим зростає частка інформації, яку організації створюють і зберігають в електронній формі.

Управління неструктурованою інформацією складається з шести основних компонентів:

1. Системи управління документами.
2. Системи управління Web-контентом.
3. Управління архівами.
4. Управління цифровими правами.

5. Співпраця в галузі управління контентом.

6. Функції введення зображень.

Більш глибинна цінність управління неструктурованою інформацією проявляється, коли неструктуровані дані використовуються для створення або вдосконалення продуктів або послуг, для оптимізації системи прийняття рішень та виконавчих процесів.

Інтерес до роботи з неструктурованими даними виник ще в п'ятидесяті роки. Але, саме методи роботи з даними удосконалюються не з такою швидкістю, з якою зростають їхні обсяги. Це є результатами наукової роботи, однак виникнення проблеми великих об’ємів даних помітно прискорило хід подій.

Компанії, що спеціалізуються на роботі з неструктурованими даними, виникли приблизно 10-20 років тому. Каталізатором цього була зростаюча необхідність практичної роботи з такими даними.

Сьогодні все радикально змінилося – зросла необхідність роботи з неструктурованими даними (Unstructured Data Analysis, UDA).

Але виникає актуальне питання: що ж робити з неструктурованими даними?

Для вирішення проблеми такі дані потрібно класифікувати. Для початку потрібно створити ряд стандартних, чітких правил.

Наприклад:

1) У документації окремою процедурою виділяти пошук паспортних даних і номерів;

2) Відокремлювати визначення конфіденційних даних і даних для службового користування;

3) Відокремлювати ідентифікацію аудіо- та відео-записів.

Але і тут є свої складнощі. Як один з прикладів – синхронізація роботи пошукових запитів та фільтрів, адже ми не хочемо, щоб аналіз даних відбувався в години максимального навантаження на сервер.

Звісно, що для зручності можна скористатися результатами відповідних звітів, які наочно покажуть, що конкретно і наскільки часто зустрічається у файлах, і де ці файли знаходяться. Але такий підхід доречний тоді, коли до таких файлів звертаються дуже часто.

Необхідно отримати розуміння структури і повноцінний контроль над поширенням даних всередині установи або організації, автоматично вживати заходів щодо мінімізації ризиків при виникненні нових пошукових запитів.

Класифікація (а саме – детальна) даних є занадто важливим елементом контролю за неструктурованою інформації, щоб його можна було просто ігнорувати. Без нього просто неможливо бути впевненим, що дані знаходяться саме там, де вони і повинні знаходитися.

**Література**

1. Головянко М.В. Методи і модель верифікації знань для інтелектуалізації Web-контенту: Автореф. дис. канд. техн. наук: 05.13.23 / Марія Валентинівна Головянко – Х.: ХНУРЕ, 2011. – 19 с.

2. Ландэ Д.В. Поиск знаний в Internet. Профессиональная работа. / Д.В.Ландэ – СПб.: Диалектика, 2005. – 272 с.