Нестерук Р.В., магістр,
Скачков В.О., ст. викладач,
Житомирський державний технологічний університет

СИСТЕМА ЗБЕРІГАННЯ ДАНИХ З УГРУПОВАННЯМ ПО КАТЕГОРІЯХ

Дана робота присвячена темі зберігання різних даних і їх структуризації.

В даний час є проблема обмеженості способів опису, визначення даних/файлів в ієрархічній структурі, основні характеристики якої назва, місце в ієрархії (каталогів), метадані. Із цього випливають труднощі в знаходженні даних/файлів, оскільки пам'ять людини влаштована так, що характеризує інформацію не єдиним чином і потім при спробі згадати, як він її назвав раніше або де вона повинна бути, швидше за все, зазнає невдачі або на це буде потрібно додатковий час. Так само відбувається велика витрата часу на структурування даних у відповідну ієрархію, що по суті є зайвою роботою.

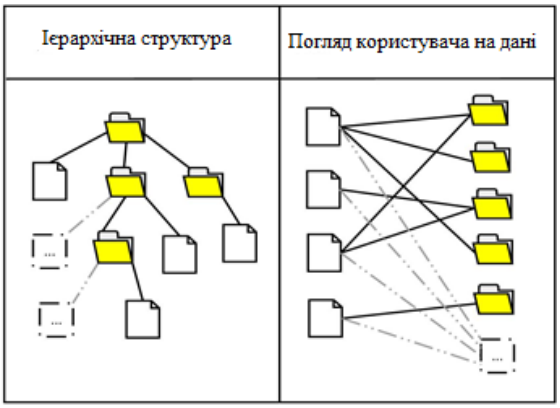

Рис.1. Підходи до структурування даних

Новий підхід, який вже досить популярний – додавання міток до даних (мітка або короткий опис будь-яких даних, який, на думку користувача, характеризує їх). Крім того наявність тих чи інших міток у даних може розділяти дані на різні категорії, за якими згодом можна здійснювати пошук.

Для системи зберігання даних з їх описами користувача важлива в першу чергу повнота опису цих даних і відповідно швидкість, адекватність їх пошуку, а також зручність у використанні.

Архітектура. Основною частиною архітектури системи зберігання даних є сховище даних. Саме сховище даних складається з бази даних (не ієрархічної структури), керуючого модуля для управління інформацією про дані і API, за допомогою якого здійснюється робота з системою зовні.

Віртуальна файлова система – модуль, який відповідає за імітацію звичайної ієрархічної файлової системи для зручності користувача і використання звичайними програмами. Шлях у віртуальній файлової системи відображається в запиті до сховища даних. Тобто, іншими словами, при переході в будь-який каталог віртуальної файлової системи, відбувається запит до бази даних сховища з критеріями пошуку, після чого там відображаються файли і підкаталоги, отримані як результат вибірки.

Аналізатор відповідає за автоматичне визначення можливих категорій/міток, до яких можна віднести дані. Аналізатор працює на основі такої інформації як вміст даних і набір міток, пов'язаний з цими даними. Аналізатор повинен мати можливість навчання в міру роботи з користувачем і при аналізі бази даних. Завдяки аналізатору можна спробувати отримати більш повний опис файлу, навіть якщо користувач сам не досить повно опише дані.

Різні спеціалізовані додатки, написані спеціально для роботи з даною системою, можуть звертатися до сховища даних безпосередньо через API. Звичайні ж додатка, які підтримують плагіни, аналогічно можуть працювати за допомогою цих плагінів. Якщо ж звичайне програма не вміє обробляти плагіни, воно може працювати з файлами зі сховища через віртуальну файлову систему.

В якості підходів до аналізу інформації про дані, за якими в подальшому можуть робитися ті чи інші висновки про приналежність даних до тієї чи іншої категорії, можуть бути взяті різні методи виявлення загального з уже іншими категоріями даними. Наприклад, логічний висновок (Особливо при використанні дедуктивної бази даних), а також і інші підходи, як, наприклад, аналіз кореляції слів які зустрічаються в текстах, байєсівська класифікація. Останній метод є досить відповідним в рамках даної задачі, оскільки враховує вміст текстових даних. Крім того теоретично аналогічні підходи можуть бути поширені і на аналіз нетекстових даних, де буде розглядатися тільки набір міток, пов'язаний з цими даними.