

Програмне забезпечення для зберігання та обробки неструктурованих документів

Виконала студентка
КПІ ім. Ігоря Сікорського, ТЕФ,
Групи ТР-71мп
Масечко І. О.

Керівник: к.т.н. Михайлова І. Ю.

Проблема

- ▶ Значна частина усієї інформації, яка використовується підприємствами для щоденної роботи, зберігається у вигляді паперових документів або їх відсканованих копій. Це особливо стосується даних, отриманих із зовнішніх джерел. Тобто більшість вхідних даних компаній є неструктурованими.
- ▶ **Неструктуровані дані** – інформація, яка не має попередньої певної структури даних, або не організована в установленому порядку. Це призводить до труднощів аналізу, особливо у випадку використання традиційних програм, призначених для роботи зі структурованими даними.

Мета та основні завдання дисертації

Метою магістерської дисертації є розробка програмної системи для зберігання та обробки неструктурованих документів у мультимодельній СКБД Caché.

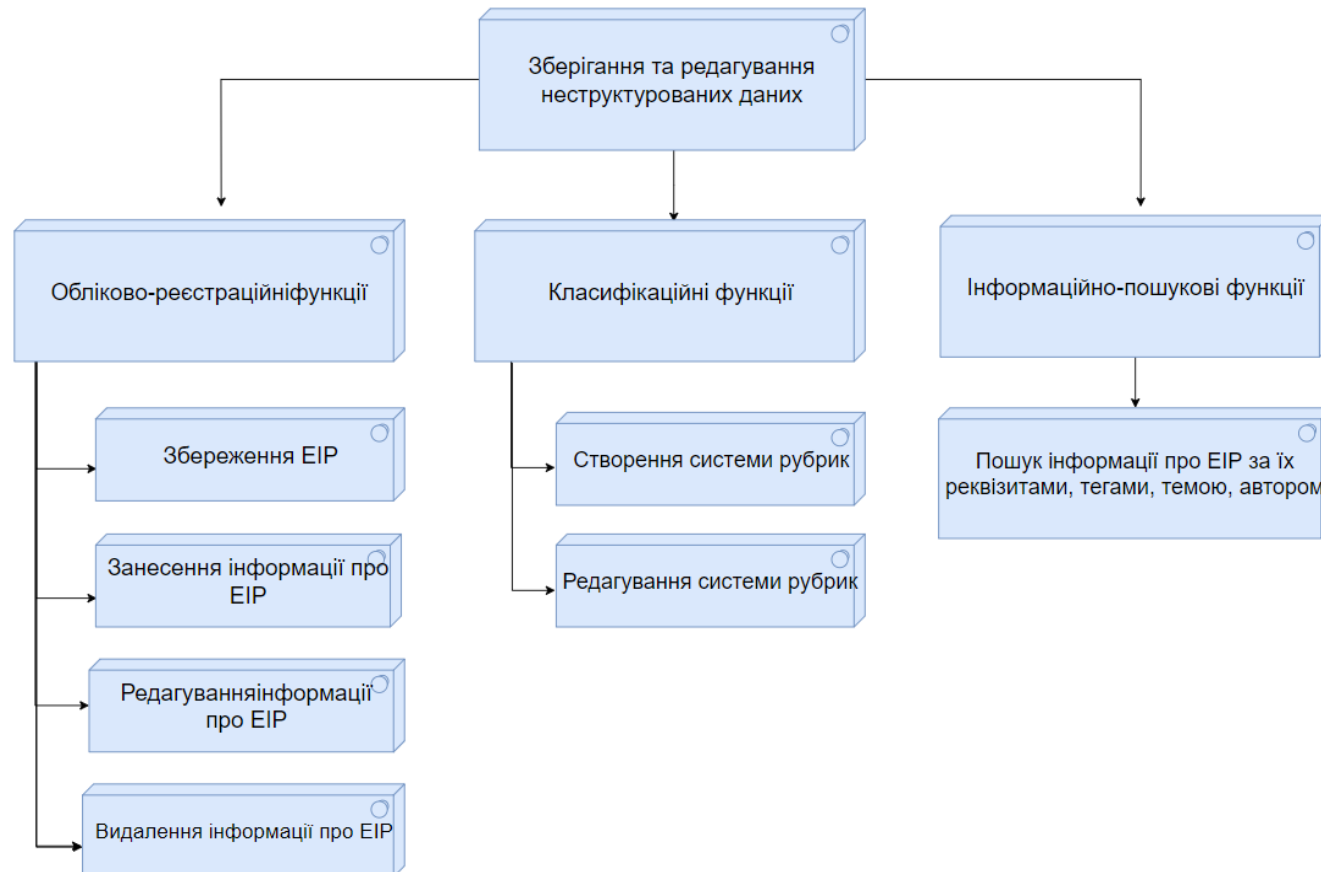
Для досягнення поставленої мети необхідно вирішити завдання:

- проаналізувати існуючі програмні системи для роботи з неструктурованими документами, визначити їх переваги та недоліки;
- обрати засоби реалізації програмної системи для роботи з неструктурованими документами;
- розробити архітектуру та структуру програмної системи для роботи з неструктурованими документами;
- розробити програмну систему для роботи з неструктурованими документами.

Програмна система для роботи з неструктурованими документами повинна бути розроблена з використанням СКБД InterSystems Caché та технології InterSystems iKnow.

Мета та основні завдання системи

Основною метою розробленої програми для управління неструктурованими документами є забезпечення обліково-реєстраційних, класифікаційних та інформаційно-пошукових функцій.



Програмна реалізація системи

Для ефективної реалізації зазначеної задачі було вирішено використати постреляційну базу даних InterSystems Caché та технологію роботи з неструктурованими даними iKnow для автоматичного визначення тегів текстових документів.

Система керування базами даних (СКБД) InterSystems Caché дозволяє швидко обробляти дані при реалізації складних систем, використовувати об'єктно-орієнтований підхід до проектування, створювати класи з властивостями різних типів даних та об'єкти, до яких, за необхідності, можна отримати реляційний доступ .

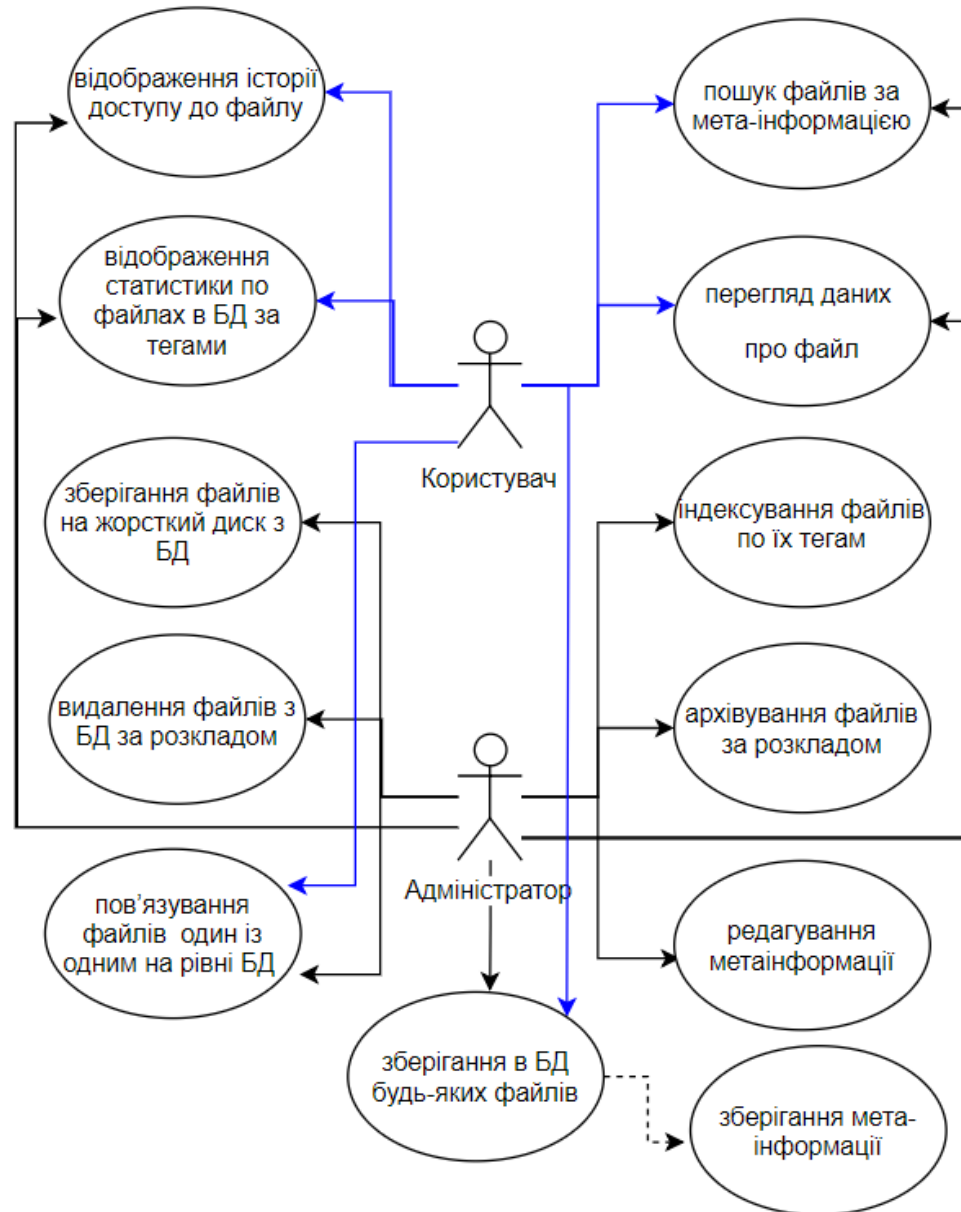
InterSystems iKnow – технологія, яка дозволяє розробникам створювати застосування для отримання інформації з неструктурованих даних . Технологія iKnow дозволяє знаходити потрібні концепти і зв'язки в неструктурованих даних.

Програмний продукт реалізований *мовою C#* засобами Visual Studio з використанням об'єктного доступу до даних, що зберігаються в СКБД.

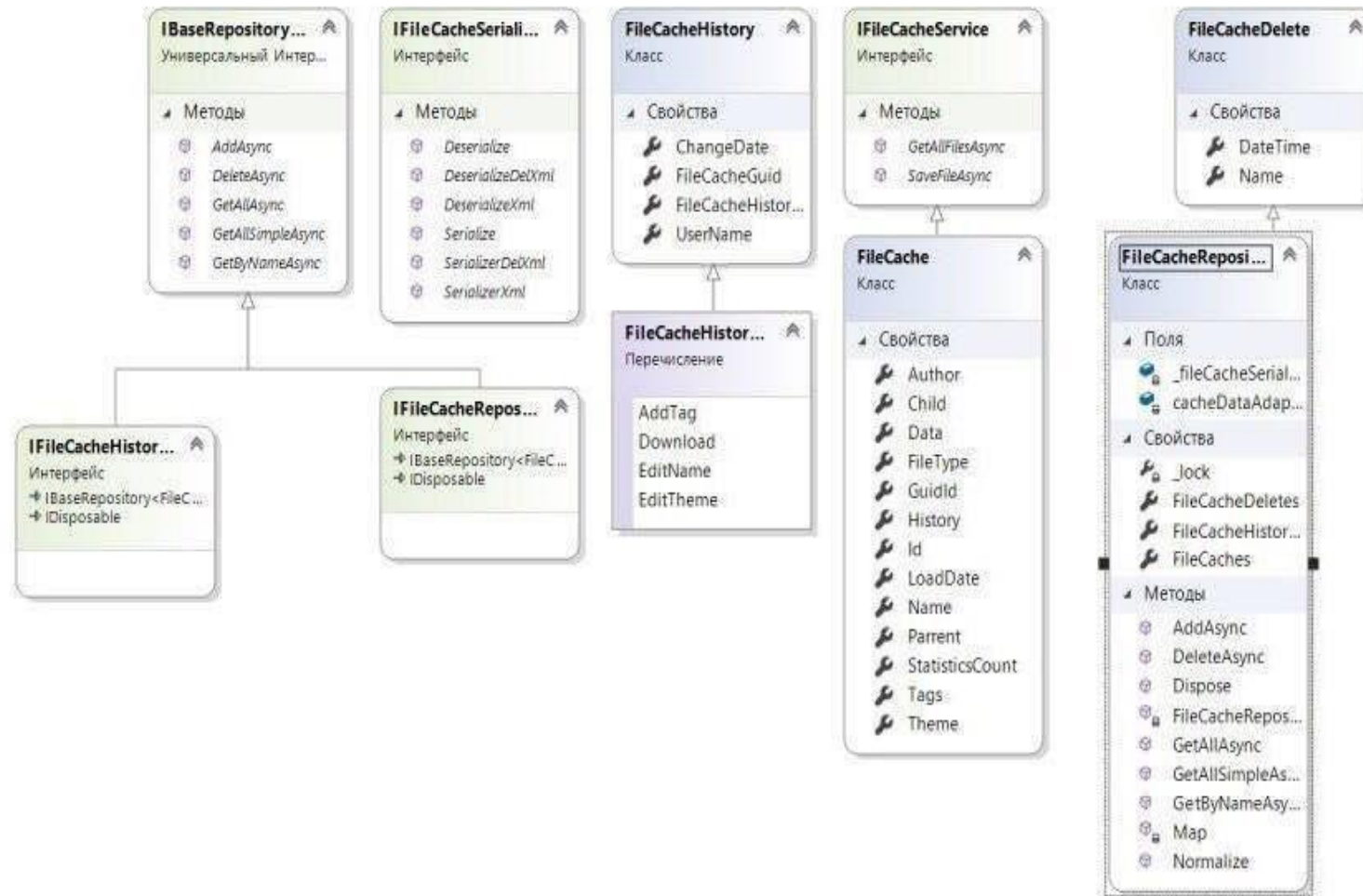
Архітектура ПЗ



Діаграма прецедентів



Структура програмної системи



Приклад роботи програми

The screenshot displays the CacheWriter application window. The title bar shows 'Form1' and standard window controls. The menu bar includes 'File', 'About', and 'Settings'. The main header area features the 'CacheWriter' logo and connection status: 'current connection: _SYSTEM', 'port: 1972', and 'connected: true'. Two buttons, 'Закрити' (Close) and 'Створити підключення' (Create connection), are present.

The main interface is divided into a left sidebar and a main content area. The sidebar, titled 'Cache', lists a directory of files, with 'Хохлова Я.' selected. The main content area has a tabbed interface with 'Обраний файл' (Selected file) active. It displays file details for 'Хохлова Я.':

- Назва: Хохлова Я.
- Тема: Трудовий договір
- Автор: DESKTOP-B0GED1R\Ira
- Тип файлу: doc
- Дата загрузу: 23.10.2018 00:00:00 +03:00
- Теги: трудовий договір, рекрутер, відділ 7

A 'History' table is also shown:

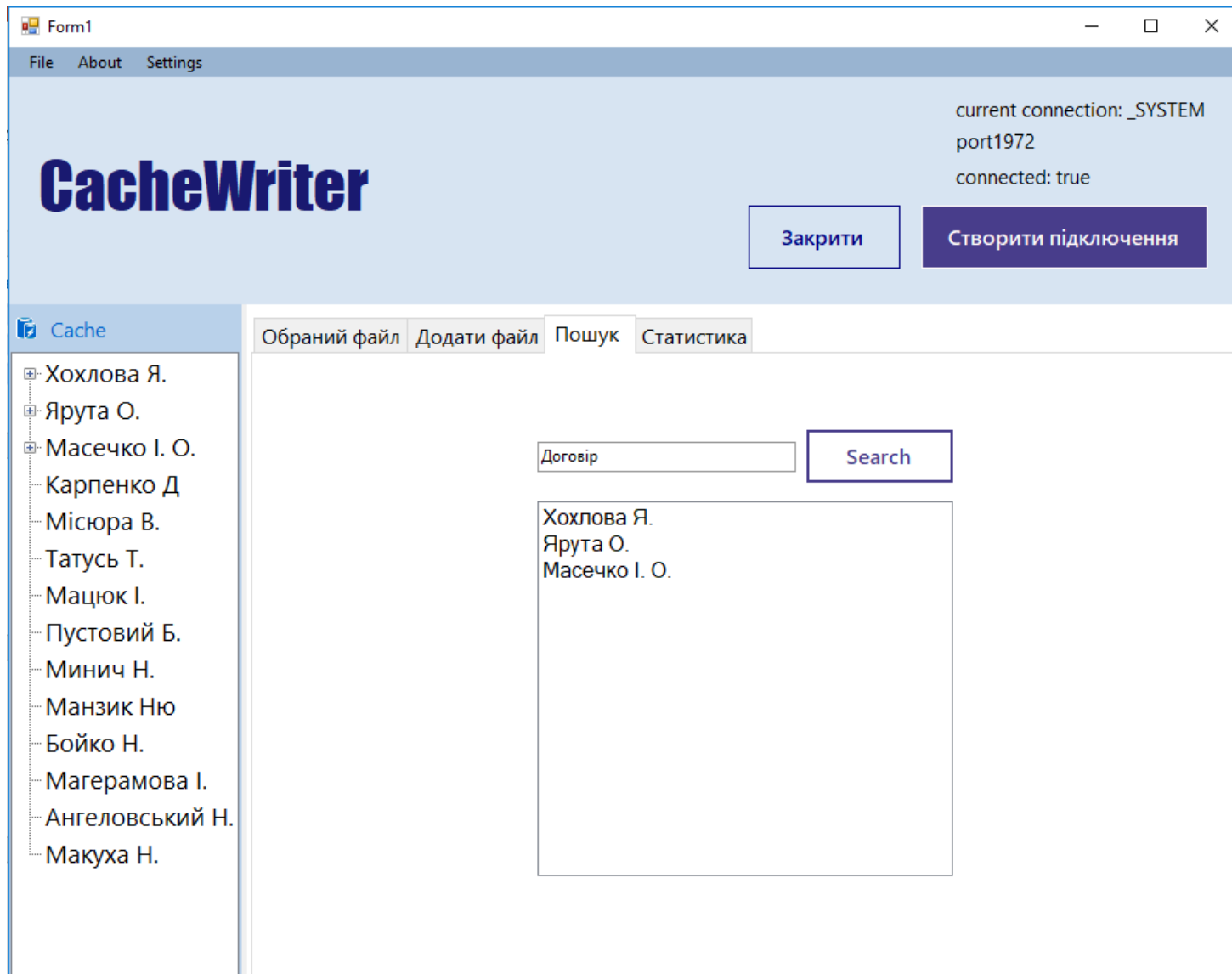
№	UserName	Date	History Type
1	EditName	23.10.20...	DESKTOP-B0GE...

Below the file details is a 'Редагувати файл' (Edit file) section with three input fields and buttons: 'Додати Тег' (Add Tag), 'Змінити Тему' (Change Theme), and 'Змінити Назву' (Change Name). To the right, there is a date selector (currently 'понедельник, 26 ноября 20') and buttons: 'Додати Child' (Add Child), 'Видалити файл за датою' (Delete file by date), 'Видалити' (Delete), 'Архівувати' (Archive), and 'Загрузити файл' (Upload file).

Приклад роботи програми

The screenshot displays the CacheWriter application window. The title bar shows 'Form1' and standard window controls. The menu bar includes 'File', 'About', and 'Settings'. The main header area features the 'CacheWriter' logo on the left and connection status on the right: 'current connection: _SYSTEM', 'port1972', and 'connected: true'. Below the status are two buttons: 'Закрити' (Close) and 'Створити підключення' (Create connection). The main interface is divided into a left sidebar and a central content area. The sidebar, titled 'Cache', contains a tree view of names: Хохлова Я., Ярута О., Масечко І. О., Карпенко Д., Місюра В., Татусь Т., Мацюк І., Пустовий Б., Минич Н., Манзик Ню, Бойко Н., Магерарова І., Ангеловський Н., and Макуха Н. The central content area has a tabbed interface with 'Обраний файл', 'Додати файл', 'Пошук', and 'Статистика'. The 'Додати файл' tab is active, showing a form titled 'Додати файл' with the subtitle 'Форма для додавання файлу в базу даних'. The form includes three input fields: 'Шлях до файлу' (with a folder icon), 'Назва файла', and 'Назва теми'. To the right, there is a 'Всі теги' (All tags) list area and a 'Новий тег' (New tag) input field with a 'Додати Тег' (Add Tag) button. A large green 'Додати' (Add) button is positioned at the bottom center of the form area.

Приклад роботи програми



Приклад роботи програми

The screenshot displays the CacheWriter application window. The title bar reads "Form1". The menu bar includes "File", "About", and "Settings". The main header area features the "CacheWriter" logo on the left and connection status on the right: "current connection: _SYSTEM", "port1972", and "connected: true". Below the header are two buttons: "Закрити" (Close) and "Створити підключення" (Create connection). The main interface has a "Cache" sidebar on the left with a tree view of folders containing names like "Хохлова Я.", "Ярута О.", etc. The main content area has tabs for "Обраний файл", "Додати файл", "Пошук", and "Статистика". The "Статистика" (Statistics) tab is active, showing a "Form for Statistics" table.

Name Tag	Count
скан	3
документи	3
паспорт	3
трудовий договір	3
документ	3
фото	2
працівник	2
відділ 5	2
дизайнер	2

Висновки

- ❑ Під час виконання магістерської дисертації була проаналізована та вивчена проблема роботи з неструктурованими документами. Було встановлено, що для вирішення завдання зберігання та обробки неструктурованих документів, недостатньо існуючих на сьогодні програмних рішень.
- ❑ Було проведено огляд методів і засобів розробки програмної системи для вирішення даного завдання.
- ❑ Розроблено архітектуру програмного забезпечення для зберігання та обробки неструктурованих даних. Розроблено загальну методику роботи користувача з програмою, що забезпечує швидку адаптацію користувача до особливостей роботи застосунку.
- ❑ Було спроектовано та реалізовано програмну систему роботи з неструктурованими даними з використанням технологій InterSystems, яка надає можливість з легкістю маніпулювати великою кількістю документів та виконувати пошук по ним.
- ❑ Розроблений програмний продукт виконано відповідно до заданих вимог, з використанням патерну Repository та сучасних підходів до розробки програмних застосунків.
- ❑ Система впроваджена та використовується працівниками компанії ТОВ "МОБІМІЛЛ СОФТ".

Дякую за увагу