Дипломна робота

на тему: «Аналіз відповідності коду документа універсальному десятковому класифікатору»

Виконав:
студент 4 курса 2 групи
напрямку «Компьютерная инженерия»
Хоба Юрій
Викладач: ст. викладач
Трубіна Н. Ф.

Мета роботи

Метою даної дипломної роботи є створення системи, яка проводить аналіз відповідності коду документа універсальному десятковому класифікатору.

Для досягнення поставленої мети необхідно виконати наступні завдання:

- дослідити предметну область;
- проаналізувати наявність допоміжних засобів визначення УДК;
- сформулювати вимоги до створюваної системи;
- запропонувати деяку модель оцінки якості співвідношення документа до розділу УДК;
- провести проектування бази даних інформаційної системи;
- провести проектування програми;
- вибрати інструментальне середовище розробки і виконати програмну реалізацію системи;
- провести тестування розробленої програми.

УДК та його структура

УДК 528.8.04:[552.323.6+553.81.044](100)
ББК 26.31с+26.342с
С32

Серокуров, Юрий Николаевич.

Дистанционный прогноз кимберлитового магматизма = Remote of kimberlite magmatism / Ю. Н. Серокуров, В. Д. Калмыков, В. М. Зуев ; Ин-т дистанционного

- 0 ОБЩИЙ ОТДЕЛ. НАУКА И ЗНАНИЕ. ИНФОРМАЦИЯ. ДОКУМЕНТАЦИЯ. БИБЛИОТЕЧНОЕ ДЕЛО. ОРГАНИЗАЦИИ. ПУБЛИКАЦИИ В ЦЕЛОМ
  - 00 Общие вопросы науки и культуры
    - 001 Наука и знание в целом. Науковедение. Организация умственного труда
      - 001.1 Общее понятие о науке и знании
      - 001.32 Ученые, научные общества. Академии
      - 001.8 Общая методология. Научные и технические методы исследований, изучения, поисков и дискуссий. Научный анализ и синтез
        - 001.89 Организация науки и научно-исследовательских работ
      - 001.9 Распространение знаний: факты, фантазии и фальсификации. Ограничения в распространении знаний. Сохранение знаний в тайне
    - 002 Документация. Научно-техническая информация (НТИ). Печать в целом. Авторство
    - 003 Системы письма и письменности. Знаки и символы. Семиотика в целом. Коды. Графическое представление мысли
    - 004 Информационные технологии. Вычислительная техника. Обработка данных
      - 004.01/.08 Специальные определители для вычислительной техники
      - 004.2 Архитектура вычислительных машин
      - 004.3 Аппаратные средства. Техническое обеспечение
        - 004.3`1/`2 Специальные определители для аппаратных средств
        - 004.31 Блоки обработки данных. Процессоры
        - 004.32 Магистрали ЭВМ
        - 004.33 Блоки памяти. Накопители. Запоминающие устройства
        - 004.35 Периферия. Устройства ввода-вывода
        - 004.38 Виды компьютеров

Аналіз сервісів визначення УДК

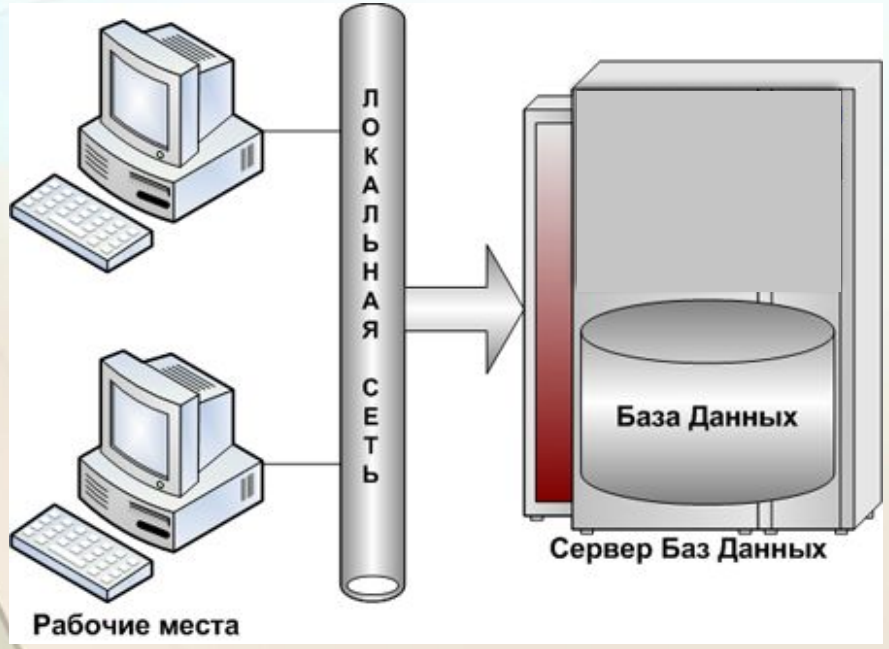
Код УДК	Описание
0	Наука и знание. Организация. Информационные технологии. Информация. Документация. Библиотечное дело. Учреждения. Публикации
1	Философия. Психология
2	Религия. Богословие
3	Общественные науки. Статистика. Политика. Экономика. Торговля. Право. Государство. Военное дело. Социальное обеспечение. Страхование. Образование. Фольклор
5	Математика и естественные науки
6	Прикладные науки. Медицина. Технология
7	Искусство. Развлечения. Зрелища. Спорт

Код УДК	Описание
0	Наука и знание. Организация. Информационные технологии. Информация. Документация. Библиотечное дело. Учреждения. Публикации
1	Философия. Психология
2	Религия. Богословие
3	Общественные науки. Статистика. Политика. Экономика. Торговля. Право. Государство. Военное дело. Социальное обеспечение. Страхование. Образование. Фольклор
5	Математика и естественные науки
6	Прикладные науки. Медицина. Технология
7	Искусство. Развлечения. Зрелища. Спорт
8	Язык. Языкознание. Лингвистика. Литература
9	География. Биографии. История

код УДК	описание	число кодов
00	Наука в целом (информационные технологии - 004)	1082
1	Философия. Психология	740
2	Религия. Теология	993
30	Теория и методы общественных наук	428
31	Демография. Социология. Статистика	748
32	Политика	328
33	Экономика. Народное хозяйство. Экономические науки	2964
34	Право. Юридические науки	4414
35	Государственное административное управление. Военное искусство. Военные науки	2428
36	Обеспечение духовных и материальных жизненных потребностей. Социальное обеспечение. Социальная помощь. Обеспечение жильем. Страхование	1400
37	Народное образование. Воспитание. Обучение. Организация досуга	1174
39	Этнография. Нравы. Обычаи. Жизнь народа. Фольклора	308
50	Общие вопросы математических и естественных наук	152

Сервіси НоваяТипография, Triumph и TeaCode.

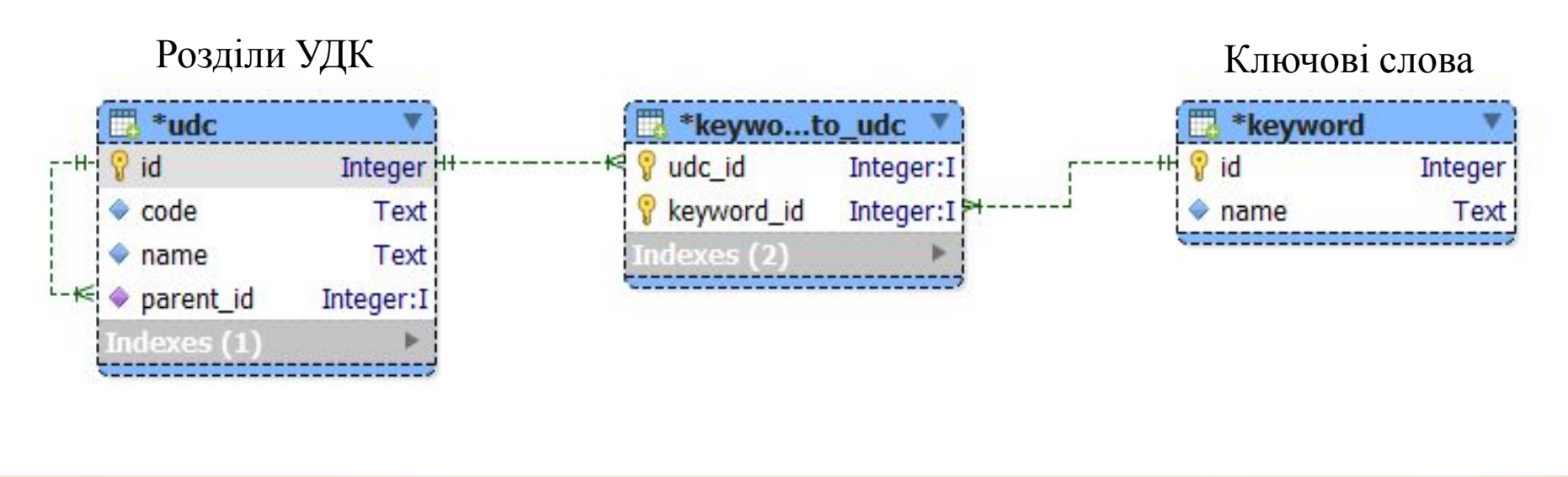
Архітектура системи й засоби розробки



Дворівнева архітектура

Засоби розробки

Схема Бази Даних



Модель оцінки якості співвідношення документа розділу УДК

Набір ключових слів розділу с кодом 0

0 Наука і знання. Організація. Інформаційні технології. Інформація. Документація. Бібліотечна справа. Установи. Публікації — 100%

00 Загальні питання науки та культури. Пропедевтика — 92%

004 Інформаційні технології. Обчислювальна техніка. Обробка даних — 86%

Модель оцінки якості співвідношення документа розділу УДК

1. Для пошуку в заданому тексті беруться всі ключові слова, які стосуються заданого розділу УДК і до його батьківських розділів.

0	Ключевое слово1, ключевое слово2, …
00	Ключевое слово3, ключевое слово4, …
004	Ключевое слово5, ключевое слово6, …
004.6	Ключевое слово7, ключевое слово8, …
004.65	Ключевое слово9, ключевое слово10, …

Модель оцінки якості співвідношення документа розділу УДК

2. Система проводить обчислення ваги кожного ключового слова на підставі того, наскільки високо воно знаходиться за ієрархією по відношенню до заданого розділу УДК.

$$W = 100 - X * l$$

де l – різниця між рівнем заданого розділу УДК та розділом, до якого належить знайдене ключове слово;
X – число, на яке зменшується вага з кожним рівнем.

Модель оцінки якості співвідношення документа розділу УДК

3. Система здійснює пошук ключових слів у тексті документа. Обчислення відсотка відповідності коду документа УДК відбувається за наступною формулою:

$$p = \frac{(W_1 + W_2 + \ldots + W_m)}{n}$$

де $W_1, W_2, \ldots, W_m$ – ваги знайдених у тексті ключових слів;
n – загальна кількість ключових слів заданого розділу і його батьківських розділів;
m – кількість знайдених у тексті ключових слів заданого розділу та його батьківських розділів.

Приклад роботи системи

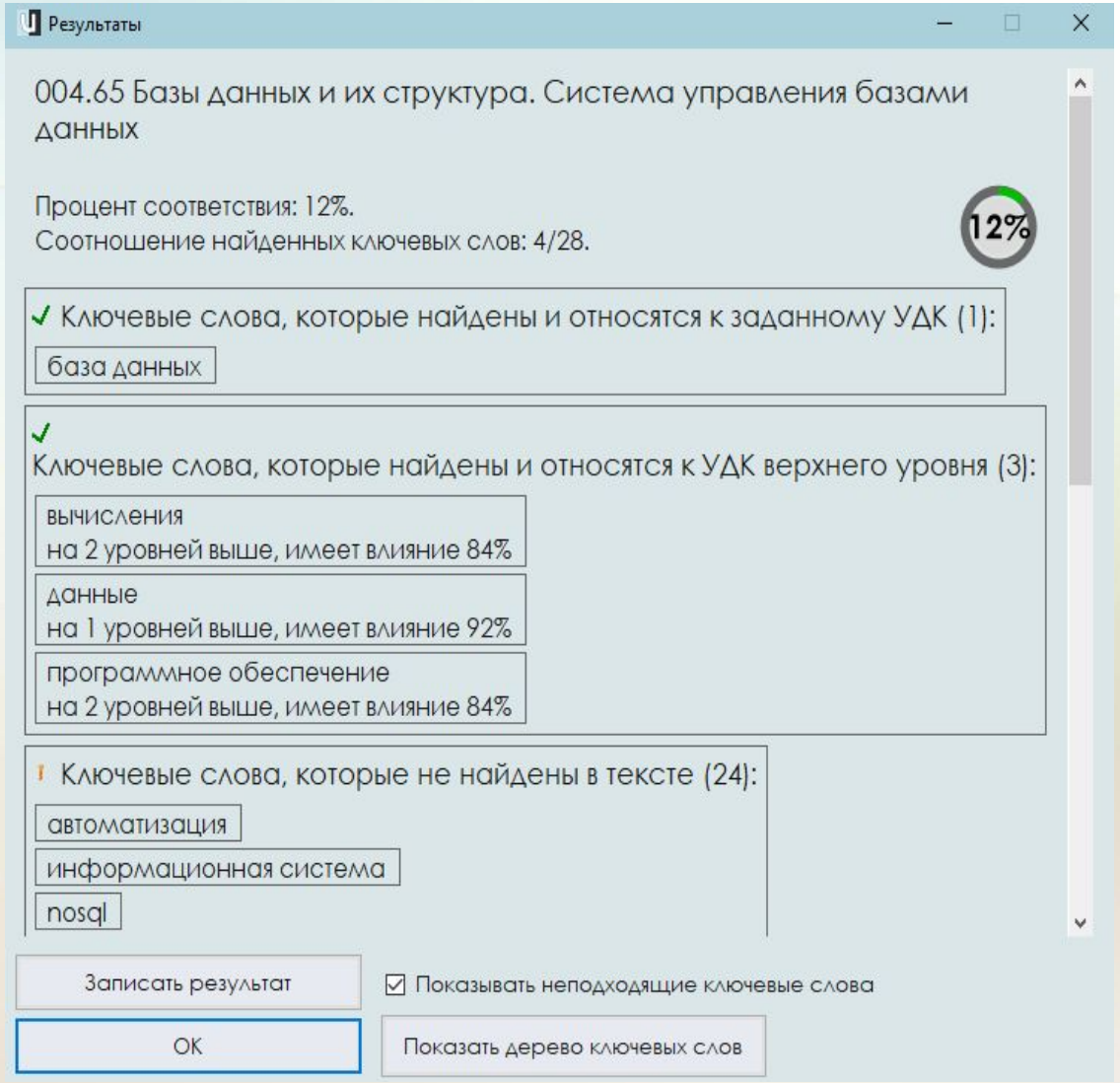
Аналіз відповідності тексту документа до УДК 004.65:

«За последние десять лет крупные компании вкладывали большие средства, заменяя свои системы обработки транзакций ERP-системами (Enterprise Resource Planning, ERP - системы Планирования ресурсов предприятия).

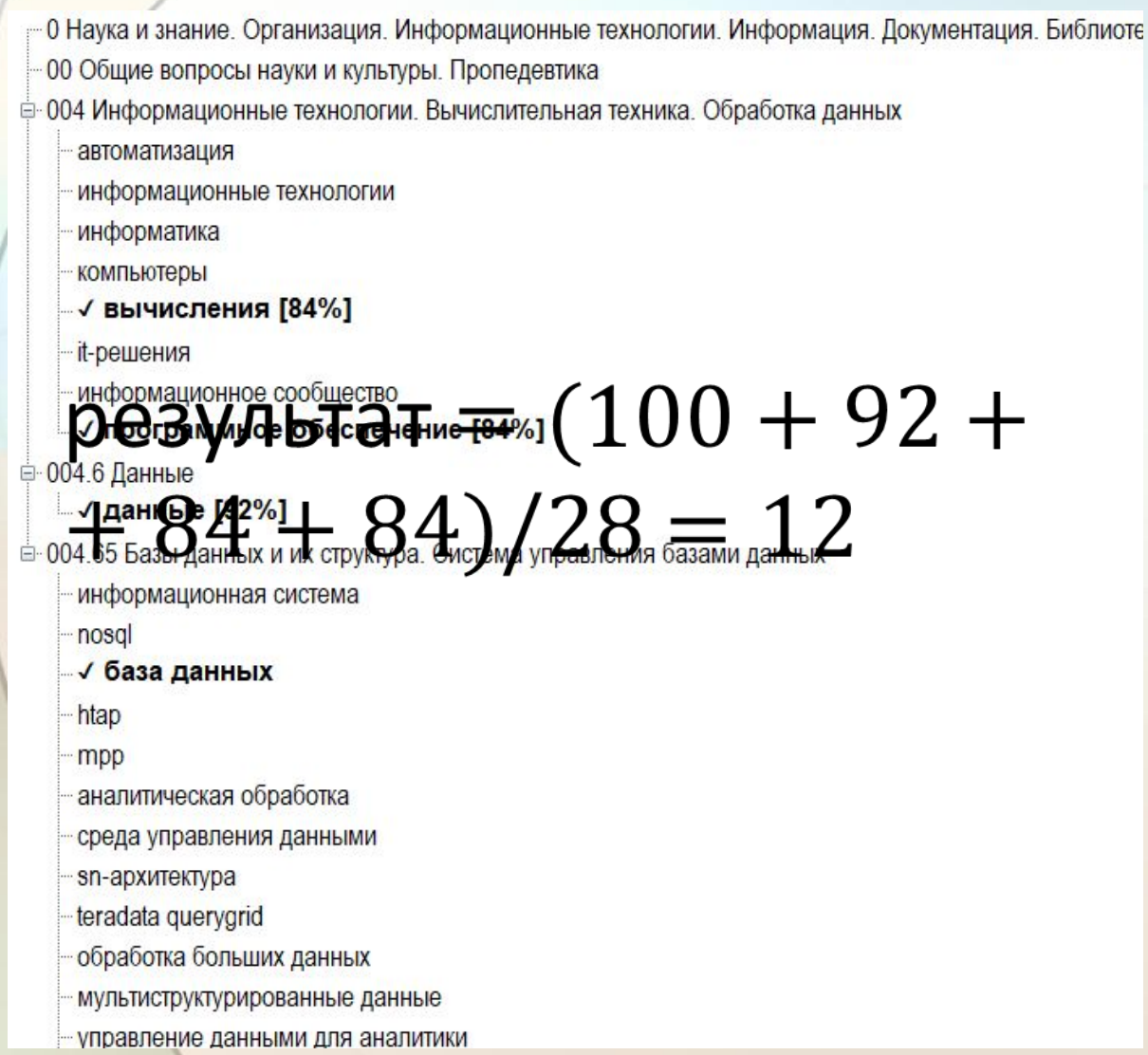
С другой стороны, большое значение уделялось разработке Хранилищ и/или витрин данных, позволяющих агрегировать и представлять данные и оказывающих существенную поддержку в принятии решений. Часто такие проекты выполнялись параллельно, однако при этом в ERP-системах никак не использовались возможности и преимущества Хранилищ данных.

В результате многие фирмы, потратившие миллионы долларов на ERP-решения, остались неудовлетворенными, так как не могут своевременно получить из систем обработки транзакций агрегированные данные. Следующее поколение ERP-систем должно учесть эту проблему.

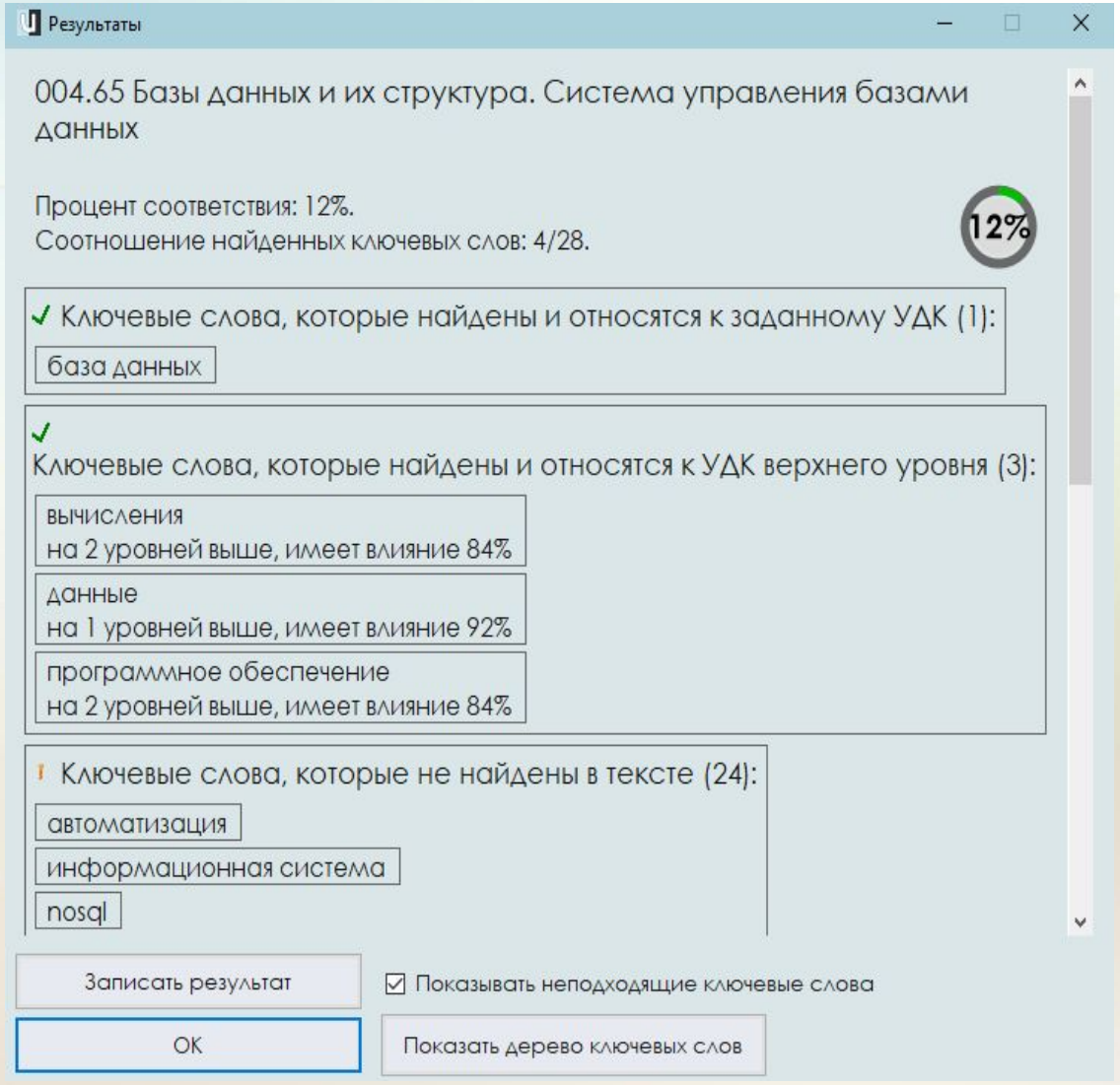
В этой статье рассматривается развитие ERP-систем, роль Хранилищ данных в информационной ERP-архитектуре, а также перспективы создания интегрированных сред, использующих возможности этих двух технологий. Кроме того, дается оценка достижений двух крупнейших фирм (SAP и People Soft), производителей ERP-систем в области разработки Хранилищ данных и аналитических приложений.»



Приклад роботи системи



$$результат = (100 + 92 + 84 + 84)/28 = 12$$



Висновки про проведені тести

1. Система практично ніколи не дає 100% результат відповідності.
2. Результат відповідності навіть в 10-15% для анотацій або документів досить невеликого обсягу при правильно накопиченої базі фахівцем, свідчать про те, що даний документ добре ставиться до заданого розділу УДК.
3. Для підвищення ефективності роботи системи, необхідно переконатися, що база наповнена достатньою кількістю правильних ключових слів і проводити аналіз документів, обсяг яких не є малим.

Висновки

- ✔ Була досліджена предметна область.
- ✔ Розглянуті ресурси за схожою тематикою.
- ✔ Сформульовані вимоги до розроблюваної системі.
- ✔ Спроектовані додатки і база даних.
- ✔ Запропонована власну модель оцінки якості відповідності документа до розділу УДК.
- ✔ Система реалізована і протестована.

Дякую за увагу!